INFECTION AND IMMUNITY, Aug. 2003, p. 4563–4579 Vol. 71, No. 8
0019-9567/03/$08.00+0 DOI: 10.1128/IAI.71.8.4563–4579.2003

Genome-Based Bioinformatic Selection of Chromosomal *Bacillus anthracis* Putative Vaccine Candidates Coupled with Proteomic Identification of Surface-Associated Antigens

N. Ariel,[1]* A. Zvi,[1] K. S. Makarova,[2] T. Chitlaru,[1] E. Elhanany,[1] B. Velan,[1] S. Cohen,[1] A. M. Friedlander,[3] and A. Shafferman[1]*

Israel Institute for Biological Research, Ness Ziona 74100, Israel[1]; National Center for Biotechnology Information, National Library of Medicine, National Institute of Health, Bethesda Maryland 20894[2]; and United States Army Medical Research Institute of Infectious Diseases, Fort Detrick, Maryland 21702[3]

Received 14 January 2003/Returned for modification 19 March 2003/Accepted 1 May 2003

Bacillus anthracis **(Ames strain) chromosome-derived open reading frames (ORFs), predicted to code for surface exposed or virulence related proteins, were selected as *B. anthracis*-specific vaccine candidates by a multistep computational screen of the entire draft chromosome sequence (February 2001 version, 460 contigs, The Institute for Genomic Research, Rockville, Md.). The selection procedure combined preliminary annotation (sequence similarity searches and domain assignments), prediction of cellular localization, taxonomical and functional screen and additional filtering criteria (size, number of paralogs). The reductive strategy, combined with manual curation, resulted in selection of 240 candidate ORFs encoding proteins with putative known function, as well as 280 proteins of unknown function. Proteomic analysis of two-dimensional gels of a *B. anthracis* membrane fraction, verified the expression of some gene products. Matrix-assisted laser desorption ionization–time-of-flight mass spectrometry analyses allowed identification of 38 spots cross-reacting with sera from *B. anthracis* immunized animals. These spots were found to represent eight in vivo immunogens, comprising of EA1, Sap, and 6 proteins whose expression and immunogenicity was not reported before. Five of these 8 immunogens were preselected by the bioinformatic analysis (EA1, Sap, 2 novel SLH proteins and peroxiredoxin/AhpC), as vaccine candidates. This study demonstrates that a combination of the bioinformatic and proteomic strategies may be useful in promoting the development of next generation anthrax vaccine.**

Anthrax is a zoonotic disease, caused by the spore forming bacterium *Bacillus anthracis* (58). The disease occurs mostly in wild or domestic mammals and may occur in humans when exposed to infected animals, or upon direct exposure (1). *B. anthracis* is considered to be one of the most likely biological warfare agents due to the ability of its spores to be transmitted by the respiratory route, the high mortality associated with inhalation anthrax, and spore stability. Primary and secondary aerosolization of *B. anthracis* spores have been recently shown to be of major concern in bioterrorist acts involving deliberate release of *B. anthracis* spores (34, 70).

Fully virulent forms of *B. anthracis* carry two large plasmids: pXO1 and pXO2. The plasmids are considered major virulence determinants, as strains lacking either one are attenuated in animal hosts (58, 67, 68). Most of the documented *B. anthracis* virulence factors identified so far are encoded by plasmid derived genes including the two major virulence factors: the tripartite toxin and the antiphagocytic capsule, respectively. The tripartite toxin is encoded by the genes *pagA*, *lef*, and *cya*, which code for protective antigen (PA) and the lethal and edema factors, respectively (located on pXO1), whereas the genes encoding for the antiphagocytic capsule are located on pXO2. A battery of as-yet-undefined virulence factors probably resides on the *B. anthracis* chromosome (8, 9, 12, 58), the sequence of which has been recently completed (78).

The licensed human vaccine consists of the PA component of the anthrax toxin as the principal protective immunogen. However, for effective long-term protection, multiple immunizations are required. Studies in experimental animal models indicated that the efficacy of PA vaccines is far below that of the Sterne live spore vaccine (105). Therefore, it has been suggested that additional somatic antigens and/or cellular immunity may be required for full protection. Recent studies report that immune response directed against spore antigens, either through live vaccines (17) or by supplementing PA-based vaccine with formalin-inactivated spores (11), is indeed involved in enhanced protection. Identification of spore antigens or additional vegetative antigens as possible enhancers of vaccine efficacy, could permit development of improved vaccines for human use (8, 9, 17, 58).

Until recently, the major barrier to target-based screening of potential vaccine candidates has been the limited number of cloned and characterized bacterial genes. The currently available genomic sequences of numerous human pathogens facilitates the identification, analysis and cloning of genes of interest. Genome-based selection of vaccine candidates has been recently coined "reverse vaccinology" (73–75). In silico gene selection, in combination with functional genomics studies, have been applied towards novel vaccine generation for several human pathogens (e.g., *Neisseria meningitis* (69), *Streptococcus*

* Corresponding authors. Mailing address for Naomi Ariel: Department of Biochemistry and Molecular Genetics, Israel Institute for Biological Research, P. O. Box 19, Ness Ziona 74100, Israel. Phone: 972-8-9381529. Fax: 972-8-9401404. E-mail ariel@iibr.gov.il. Mailing address for Avigdor Shafferman: Israel Institute for Biological Research, P. O. Box 19, Ness Ziona 74100, Israel. Phone: 972-8-9381595. Fax: 972-8-9401404. E-mail avigdor@iibr.gov.il.

pneumoniae (109), and *Chlamydia pneumoniae* (59) and recently also to the selection of potential vaccine candidates from the *B. anthracis* virulence plasmid pXO1 (7).

Here we describe results of a genome-based bioinformatic screening of the entire *B. anthracis* draft chromosome (sequenced by The Institute for Genomic Research [TIGR], Rockville, Md.; Feb. 2001 version, 460 contigs, 98% coverage [9], following its translation and function assignments), for putative vaccine candidates. This screening process comprised of search for gene products with recognizable sequence or structural features characteristic of proteins, which either confer protective immunity (consisting mostly of surface exposed or exported proteins), or are similar to documented microbial virulence factors. The in silico approach allowed for identification of 520 *B. anthracis* open reading frame (ORF) products (240 with putative function and 280 hypothetical proteins or proteins of unknown function), mostly exhibiting features of surface exposed or exported proteins. Proteomic analysis of a *B. anthracis* membrane fraction allowed for validation of the selection process, verifying the expression of several membrane-associated candidate ORF products, and assessment of their immunoreactivity (by immunoblotting with *B. anthracis* immune animal sera). The employed complementary strategies could provide the basis for subsequent experimental evaluation of a future generation of anthrax vaccines.

MATERIALS AND METHODS

Computational analyses. Computational analyses were conducted on an SGI server (Silicon Graphics, Inc.) at the National Center for Biotechnology Information (NCBI) and on a Sun Enterprise 420R server equipped with a Sun Solaris operating system at the Israel Institute for Biological Research (IIBR). The DNA draft sequence of the *B. anthracis* chromosome (Ames strain, version of February 2001, 460 contigs) was downloaded upon agreement from TIGR (http://www.tigr.org).

Translation and assignment of ORFs was accomplished using the Wimklein module (SEALS package; NCBI [103]) a naive translator based on intrinsic properties of the sequence. A minimal length of 75 amino acids (aa) per ORF was imposed. The translation of the February version yielded 5,045 ORFs, with various lengths ranging from 75 to 5,017 aa.

Sequence similarity searches were conducted by running Blast analyses (3, 4), using the splishpgp module (SEALS package; NCBI), against the following databases: nonredundant (NCBI), unfinished microbial genomes (NCBI), *Clostridium acetobutylicum* genome (prior to its publication [65]), *Bacillus halodurans* genome (upon its publication [96]), and *Bacillus cereus* ATCC 14579 draft genome (Integrated Genomics, Ill.). The filtering program SEG (110) was applied to mask sequence segments exhibiting low compositional complexity. However, for ORFs longer than 200 aa, a nonfiltered analysis was carried out as well. Blast results were mapped according to taxonomy using the module tax_collector (SEALS, NCBI). Paralogs were identified by Blast analysis of each ORF against the *B. anthracis* genome database. ORFs were defined as paralogs, in cases where sequence similarity extends over 80% mutual coverage and the expectation value of the alignment is smaller than e^{-10}. The Blast results were tabulated by the Btab program (NCBI) and further parsed by in-house Perl scripts.

Analyses of protein domains was carried out by searching against the Pfam (89), SMART (85), and CDD databases (107), using the HMM program installed on the NCBI SGI server. Orthology was assigned through analyses of the 5,045 ORFs (February version) against the COGs database (100) (carried out by R. Tatusov [NCBI]).

Cellular localization predictions, for each ORF, were carried out as follows: prediction of presence and location of signal peptides in the N-terminal 70 aa of an ORF, using the program SignalP (63) (run locally on the NCBI server; prediction of membrane-spanning regions, using the program Tmpred [29], run via the BCM Launcher batch client [87]); recognition of lipoprotein signatures by the Lipop program of the PSORT package (60) (installed locally on the IIBR server); identification of the presence of a typical sequence signature (32, 92), as well as identification of gram-positive specific anchoring motifs (13, 62), using the GREF module (SEALS, NCBI) and in-house scripts. Anchoring motifs probed include the tripartite sortase motif (32), a refined motif with increased sensitivity and specificity which includes the traditional LPXTG motif, iron-dependent sortase motif (48), choline-binding motifs (identified using the Pfam HMM PF01473), PKD (Pfam HMM PF00801), LysM (Pfam HMM PF01476) and SLH domains (Pfam HMM PF00395), (13). The draft chromosome-derived ORFs generated in this study were compared to ORFs derived from the full sequence of the *B. anthracis* Ames strain chromosome, obtained from TIGR (78), and to the recently deposited draft sequence of the *Bacillus anthracis* A2012 strain (79) (GenBank accession number NC_003995, gi|21397375), by Blast analysis. Table 1 incorporates the equivalent of TIGR ORF number and annotation, as well as the gi number of the A2012 strain ORF equivalent, for of each of our selected draft chromosome ORFs. Throughout the text, ORF products are referred to as the draft chromosome ORF number, and listed in Table 1.

Bacterial cultures and sample preparation. *B. anthracis* cells (ATCCΔ14185, pXO1$^-$, pXO2$^-$ strain [17]) were cultured to late stationary stage in Luria broth medium for 48 h at 37°C with vigorous agitation. Preliminary sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis and Western blot analysis (data not shown) of cell membranes taken from various culture growth phases (6, 12, 24, and 48 h) established that the 48-h cells contain the largest number of polypeptides which cross-react with *B. anthracis* immune sera (see below). This preparation was therefore selected for the subsequent proteomic analysis of membrane proteins. The protein signature of the *B. anthracis* strain under study (pXO1$^-$, pXO2$^-$) reflects the expression of chromosomal ORFs only. However, one should keep in mind that expression of some chromosomal genes may be affected by pXO1 or pXO2 encoded regulators, as demonstrated recently for the *sap* gene (54). Cells were collected by centrifugation and washed twice with PBS and once in 50 mM Tris base (pH 8), followed by resuspension in 2 ml of 50 mM Tris base buffer (pH 8) at a concentration of 10^{10} cells/ml and sonication for 30 s. The insoluble material was collected by Eppendorf centrifugation (15 min), washed twice with Tris base buffer (pH 8) and incubated in a tube rotator for 30 min at room temperature in an extraction solution consisting of 8 M urea, 4% (wt/vol) 3-[(cholamidopropyl)-dimethylammonio]-1-propanesulfonate (CHAPS), 40 mM Tris, and 0.2 (wt/vol) Bio-Lyte 3/10 (Bio-Rad). Following centrifugation, the supernatant representing a fraction enriched in membrane proteins was stored at −70°C until use. The protein concentration of this fraction was 1.5 mg/ml.

Two-dimensional gel electrophoresis (2-DE) and serological analysis. Four hundred micrograms of the membrane protein mixture was separated first by isoelectric focusing (IEF) on ready-made 17-cm, pH 3 to 10 nonlinear immobilized pH gradient (IPG) strips (Immobiline DryStrips, Amersham Pharmacia Biotech) and applied to a Protean IEF cell (Bio-Rad). IEF was carried out at 10,000 V to a total 50,000 V · h, initiated by a slow step at 250 V for 30 min. Strips were then processed for the second dimension separation by a 10-min incubation in a solution containing 6 M urea, 2% SDS, 0.375 M Tris-HCl (pH 8.8), 20% glycerol, and 2% (wt/vol) dithiothreitol, and this was followed by a 10-min incubation in a similar solution in which the dithiothreitol was replaced by 2% iodoacetamide. Strips were applied to 12.5% polyacrylamide SDS gels and electrophoresis was carried out on an Ettan DALT II System (Amersham Pharmacia Biotech). Gels were stained with Coomassie blue and spots detected and analyzed by scanning on a Bio-Rad GS-800 Calibrated Densitometer assisted by the PDQuest 2-D Gel Analysis Software (Bio-Rad). Western blots were probed with immune guinea pig serum taken 9 weeks post-first immunization, from animals inoculated with four doses (two weeks apart) of 5×10^7 *B. anthracis* MASC-10 spores (17). The ELISA-determined anti *B. anthracis* (total cellular extract) titer of the immune serum, was 1:12,800.

In-gel trypsin digestion and matrix-assisted laser desorption ionization–time-of-flight mass spectrometry (MALDI-TOF MS) protein identification. Protein spots were cut from 2-DE gels and destained for 1 h in 30% acetonitrile (CH_3CN)–50 mM ammonium bicarbonate and subjected to in-gel overnight digestion with 10 μl of trypsin (6.25 μg/ml; Promega). Gel fragments were washed twice with double-distilled water followed by peptide elution with 1% trifluoroacetic acid (TFA) (20 min, room temperature) and 50% CH_3CN (20 min, room temperature). Finally, eluted tryptic peptides were dried under vacuum and resuspended in 10 μl of 25% CH_3CN–0.1% TFA. Two microliters was mixed with an equal volume of 10-mg/ml α-cyano solution containing 50% ethanol, 25% CH_3CN, and 0.1% TFA, applied to a MALDI-TOF MS target and allowed to crystallize in vacuum. Mass spectra were obtained on Micromass TofSpec 2E in positive ion reflectron mode, using source voltage of 20,000 V, pulse voltage of 2,600 to 3,000, and laser intensity of 20%. External calibration using standard peptides was applied. Spectra were compared to theoretical tryptic digestion fragments of all ORFs derived from the *B. anthracis* (pXO1$^-$, pXO2$^-$) Ames strain (78) genomic DNA sequence. Identification of proteins

TABLE 1. List of selected *B. anthracis* potential vaccine candidates and virulence factors[a]

Protein category and ORF no.	Protein name	TIGR's equivalent ORF	A2012 strain equivalent ORF (GenBank)
S-layer homology domain proteins			
1	SLH amidase_2	[+]/BA0872	gi|21398829
2	S-layer protein Sap	[+]/BA0885	gi|21398844
3	S-layer protein (EA1)	[+]/BA0887	gi|21398845
4	SLH amidase_3	[+]/BA0898	
5	SLH Transglutaminase/protease-like	[+]/BA0981/S-layer protein, putative	gi|21398925
6	NEAT domain SLH protein	[+]/BA1093/S-layer protein, putative	gi|21399019
7	SLH protein	[+]/BA1127	gi|21399050
8	SLH protein	[+]/BA1129	gi|21399051
9	SLH protein	[+]/BA1130	gi|21399052
10	SLH amidase_3	[+]/BA1817	
11	SLH amidase_4	[+]/BA1818	gi|21399697
12	SLH AAA ATPase	[+]/BA1926	gi|21399805
13	SLH protein	[+]/BA2315	gi|21400193
14	SLH amidase_3	[+]/BA2528	gi|21400397
15	SLH metallo-beta lactamase	[+]/BA2803/S-layer protein, putative	gi|21400699
16	SLH protein	[+]/BA3332	gi|21401209
17	SLH protein	[+]/BA3338	gi|21401213
18	SLH protein	[+]/BA3695	
19	SLH amidase_2	[+]/BA3737	gi|21401593
20	SLH protein	[+]/BA5054	gi|21402845
Adhesins/lipoproteins/ autotransporters:			
21	ABC transporter, substrate binding protein	[+]/BA0175	gi|21398133
22	Lipoprotein_9	[+]/BA0314	gi|21398260
23	ABC transporter, substrate-binding protein, SAND domain	[+]/BA0382	gi|21398333
24	Glutamine ABC transporter	[+]/BA0642/amino-acid ABC transporter, permease protein	gi|21398601
25	Oligopeptide ABC transporter	[+]/BA0656	gi|21398617
26	Phosphate ABC transporter, phosphate binding protein PstS	[+]/BA0715	gi|21398678
27	Putative amino-acid binding transporter	[+]/BA0855	gi|21398810
28	Probable collagen adhesin LPXTG anchored repeat protein	[+]/BA0871/LPXTG-motif cell wall anchor domain protein	gi|21398828
29	SBP_BAC_5 peptide binding protein	[+]/BA1191	gi|21399103
30	Spermidine/putrescine ABC transporter	[+]/BA1300	gi|21399202
31	Periplasmic binding protein-like II	[+]/BA1735	gi|21399623
32	Periplasmic binding protein-like I ANF-receptor binding domain	[+]/BA1931/branched-chain amino acid ABC transporter	gi|21399809
33	SBP_bac_5 amino-acid binding protein	[+]/BA2848/oligopeptide ABC transporter	gi|21400746
34	Bacterial periplasmic substrate-binding protein sulfur utilization	[+]/BA2921/sulfonate ABC transporter	gi|21400805
35	Putative ABC superfamily (peri_perm) sugar transporter	[+]/BA2975	gi|21400856
36	PsaA-like protein	[+]/BA3189/adhesion lipoprotein	gi|21401074
37	SPB_BAC_3/alkylphosphonate ABC tranporter	[+]/BA3750	gi|21401605
38	Fibronectin/fibrinogen-binding protein	[+]/BA4013	gi|21401859
39	Helical backbone metal receptor/ferrichrome ABC transporter	[+]/BA4597	gi|21402412
40	Iron(III) dicitrate transporter(fhud)	[+]/BA4766/iron compound ABC transporter, iron binding	
41	ABC transporter, ATP-binding protein	[+]/BA5003	
42	Outer membrane protein/Lipoprotein_9	[+]/BA5220/ABC transporter, substrate binding protein, putative	gi|21397467
43	Probable collagen adhesin anchored repeat protein	[−]/BA5258/cell wall surface anchor family protein	gi|21397496
44	Iron(III) ABC transporter, periplasmic binding protein	[+]/BA5330	gi|21397566
45	AfuA-like protein, iron(III) ABC transporter, periplasmic	[+]/BA5501/lipoprotein	gi|21397724

Continued on following page

TABLE 1—*Continued*

Protein category and ORF no.	Protein name	TIGR's equivalent ORF	A2012 strain equivalent ORF (GenBank)
Repeat proteins:			
46	Anchored repeat protein	[+]/BA0397/LPXTG-motif cell wall anchor domain protein	gi|21398347
47	Family similar to phage tail protein	[+]/BA0477/tape measure protein, putative	gi|21398428
48	Internalin-like protein	[+]/BA0552	
49	Secreted ANK repeat protein	[+]/BA0564	gi|21398517
50	Repeat protein	[−]/BA0881/hypothetical protein	gi|21398840
51	Repeat domain Tol-B C-terminal domain	[+]/BA1311/lipoprotein, putative	gi|21399212
52	LRR repeat protein SLH protein	[−]/BA1346/internalin, putative	
53	Bacterial neuraminidase repeat protein	[+]/BA1755	gi|21399642
54	Fibronectin-binding protein, putative	[+]/BA1840	gi|21399720
55	Contractile-protein like/enterotoxin A	[+]/BA1887	gi|21399764
56	Enterotoxin B	[+]/BA1888	gi|21399765
57	SH3_bac NLP/P60 family protein	[+]/BA1952/NLP/P60 family protein	gi|21399829
58	Secreted repeat protein	[−]/BA2029/hypothetical protein	
59	TPR repeat protein	[−]/BA2179/hypothetical protein	
60	TPR repeat protein	[+]/BA2180	gi|21400051
61	TPR repeat protein	[+]/BA2193	
62	Repeat protein	[−]/BA3065/hypothetical protein	gi|21400947
63	Bacterial neuraminidase repeat protein	[+]/BA3445	gi|21401316
64	Ankyrin repeat protein, similar to ankyrin repeat domain of bcl-3	[+]/BA3478/ankyrin repeat domain protein	gi|21401353
65	Large repeat protein	[−]/BA3725/conserved repeat domain protein	gi|21401582
66	Collagen-like protein	[−]/BA3841/conserved hypothetical protein	
67	VrrA	[−]/BA4510/missing sequence, authentic frameshift	gi|21402333
68	Repeat protein	[−]/BA4659/LysM domain protein,	gi|21402472
69	Collagen-repeat and CBD domain protein	[−]/BA4764/hypothetical protein	gi|21402569
70	NEAT-anchored protein	[−]/BA4787/conserved hypothetical protein	gi|21402590
71	Secreted NEAT protein	[−]/BA4788/conserved domain protein	gi|21402591
72	Secreted ANK repeat protein	[+]/BA4922/ankyrin repeat domain protein	
73	Collagen triple-helix domain protein	[+]/BA4978/collagen triple helix repeat domain protein	
74	VrrB	[+]/BA4989	
75	Secreted repeat protein	[−]/BA5058/lipoprotein, putative	gi|21402849
76	Secreted repeat protein	[−]/BA5326/lipoprotein, putative	gi|21397561
Enzymes			
77	Peptidase_S9 prolyl oligopeptidase family	[+]/BA0165	gi|21398122
78	FAA hydrolase/lyase	[+]/BA0241	gi|21398183
79	GDSL-like lipase/acetylhydrolase	[−]/BA0242/homogentisate 1,2-dioxygenase, putative	
80	ALDH-like/delta-1-pyrroline-5-carboxylate dehydrogenase	[+]/BA0309	gi|21398255
81	Polysaccharide deacetylase-like protein	[+]/BA0331	gi|21398276
82	Alkyl hydroperoxide reductase, subunit C	[+]/BA0345	gi|21398291
83	Transglycosylase with Fn3 and cellulose-binding domains	[+]/BA0385/chitinase B	gi|21398336
84	Endolysin	[+]/BA0485/glycosyl hydrolase, family 25, putative	gi|21398435
85	Peptidase M9, collagenase	[+]/BA0555	gi|21398504
86	PLDc	[+]/BA0625	gi|21398583
87	Immune inhibitor A metalloprotease	[+]/BA0672	gi|21398634
88	Phospholipase C/cereolysin	[+]/BA0677	
89	Sphingomyelinase C	[+]/BA0678	
90	Putative sortase	[+]/BA0688	
91	Spore germination protein GerLC	[+]/BA0711	gi|21398674
92	Purple acid-phosphatase Fn3 and laminin-binding domains	[+]/BA0815/purple-acid phosphatase	gi|21398776

Continued on following page

TABLE 1—*Continued*

Protein category and ORF no.	Protein name	TIGR's equivalent ORF	A2012 strain equivalent ORF (GenBank)
93	Proline racemase	[+]/BA0901	gi|21398858
94	5-Methycytosine-specific restriction-related enzyme	[−]/BA0928/conserved domain protein	gi|21398882
95	Phospholipase A2 family protein /PAF acetyl hydrolase	[−]/BA0950/conserved hypothetical protein	gi|21398901
96	Dihydroxyacetone kinase family protein	[+]/BA0965	gi|21398915
97	Alcohol dehydrogenase, zinc containing	[+]/BA1004	gi|21398944
98	α/β-hydrolase	[+]/BA1019	
99	PrsA	[+]/BA1041	gi|21398974
100	Isocitrate lyase	[+]/BA1132	
101	Catalase	[+]/BA1159	gi|21399079
102	PrsA	[+]/BA1169	gi|21399086
103	Metalloproteases ("zinctins") peptidase M3 family	[+]/BA1206/oligoendopeptidase F	gi|21399117
104	Probable *O*-methyltransferase	[+]/BA1221	gi|21399132
105	Immune inhibitor A	[+]/BA1295	gi|21399197
106	α/β-hydrolase PhaC(polyhydroxyalkanoic acid synthase)	[+]/BA1331	gi|21399229
107	PLP-dependent transferases	[+]/BA1341/aminotransferase, class V	gi|21399241
108	Peptidase M3 family	[+]/BA1353/oligoendopeptidase F, putative	
109	Diaminopimelate decarboxylase	[+]/BA1438	gi|21399334
110	Secreted peptidase M23/M37 LysM domain	[+]/BA1500/peptidase, M23/M37 family	
111	Amidase 2/autolysin	[+]/BA1514	gi|21399412
112	Putative metallo-dependent phosphatase/CapA	[−]/BA1855/conserved hypothetical protein	
113	α/β Hydrolase	[+]/BA1860	gi|21399739
114	Putative lipase/haloperoxidase	[−]/BA1866/conserved hypothetical protein	gi|21399746
115	Secreted peptidase, M23/M37	[+]/BA1879	
116	Multicopper oxidase/Cupredoxins	[+]/BA1902	gi|21399781
117	Secreted peptidase, M23/M37 with SH3b domains	[+]/BA1903/peptidase, M23/M37 family	gi|21399782
118	PLDc	[+]/BA2054	gi|21399928
119	ScdA	[+]/BA2147	
120	Aspartate/glutamate racemase	[+]/BA2223	gi|21400101
121	Had-like hydrolase	[+]/BA2231	gi|21400110
122	Penicillin-binding protein 2x and transpeptidase	[+]/BA2238	
123	Secreted FtsK/SpoIIIE family protein	[+]/BA2321	gi|21400199
124	Dependent dsDNA exonuclease SbcC	[+]/BA2360/exonuclease, SbcC family, putative	
125	Inosine-uridine preferring nucleoside hydrolase family	[+]/BA2400	gi|21400272
126	Secreted Metallo-hydrolase/oxidoreductase betalactamase	[+]/BA2472	gi|21400341
127	P-loop containing nucleotide triphosphate hydrolase family	[−]/BA2556/hypothetical protein	gi|21400433
128	α/β-hydrolases/lipase	[+]/BA2607	gi|21400493
129	Metallo-hydrolase/oxidoreductase	[−]/BA2688/conserved hypothetical protein	gi|21400582
130	Aromatic-L-amino-acid decarboxylase	[+]/BA2724/decarboxylase, pyridoxal-dependent	gi|21400621
131	Secreted metalloproteinase/thermolysin-like	[−]/BA2730/neutral protease	gi|21400627
132	Spore cortex lytic enzyme	[+]/BA2748	gi|21400649
133	Kynureninase (PLP dependent transferase)	[+]/BA2753	gi|21400654
134	Endolysin [Phage bastille]	[+]/BA2805/glycosyl hydrolase, family 25, putative	gi|21400701
135	Putative diaminopimelate epimerase	[−]/BA2833/hypothetical protein	gi|21400729
136	B-cell mitogen	[+]/BA2834	gi|21400730
137	X-Pro dipeptidyl-peptidase (S15 family)/cocaine esterase	[+]/BA2860/hydrolase, CocE/NonD family	gi|21400756
138	Fumble protein/Panthetoate kinase	[−]/BA2901/conserved hypothetical protein	gi|21400792
139	Putative glutathionylspermidine synthase	[+]/BA2932	
140	Polysaccharide deacetylase	[+]/BA2944	gi|21400827

Continued on following page

TABLE 1—*Continued*

Protein category and ORF no.	Protein name	TIGR's equivalent ORF	A2012 strain equivalent ORF (GenBank)
141	Secreted SH3b domain protein, putative metallohydrolase	[−]/BA2967/conserved domain protein	gi|21400849
142	Histidinol-phosphate aminotransferase	[−]/BA2995/pyrroline-5-carboxylate reductase	
143	Alkaline phosphatase	[+]/BA3001	gi|21400884
144	Peptidase	[+]/BA3033	gi|21400917
145	Probable methyl transferase	[−]/BA3085/conserved hypothetical protein	gi|21400965
146	Pyrrolidone carboxyl peptidase	[+]/BA3090	gi|21400970
147	Aminoglycoside phosphotransferase	[+]/BA3111	gi|21400988
148	Pyrroline-5-carboxylate reductase	[+]/BA3143	gi|21401019
149	Isochorismatase family protein	[+]/BA3166	gi|21401049
150	Isochorismatase family protein	[+]/BA3167	gi|21401050
151	Lactamase class C domain (PBPX family)	[+]/BA3214/penicillin-binding protein, putative	gi|21401102
152	Metallo-hydrolase/oxidoreductase	[−]/BA3247/conserved hypothetical protein	gi|21401131
153	3-beta hydroxysteroid dehydrogenase/isomerase family protein	[+]/BA3248	gi|21401132
154	Secreted peptidase_P9 (collagenase) PKD domain	[+]/BA3299	
155	Isochorismatase family protein	[+]/BA3315	gi|21401191
156	Secreted beta-lactamase/D-ala carboxypeptidase	[+]/BA3346/6-aminohexanoate-dimer hydrolase, putative	gi|21401221
157	Perfringolysin O	[+]/BA3355/thiol-activated cytolysin	gi|21401229
158	Protein-tyrosine phosphatase-like protein	[+]/BA3394	gi|21401268
159	Peptidase M4 (metalloproteinase)	[−]/BA3442/neutral protease	
160	Viral enhancin [metalloproteinase]	[+]/BA3443/enhancin family protein	gi|21401314
161	Cyclopropane-fatty-acyl-phospholipid synthase	[+]/BA3460	gi|21401332
162	Metallo-beta-lactamase family protein	[+]/BA3514	gi|21401388
163	Metalloprotease family protein	[+]/BA3553/oligoendopeptidase F	gi|21401418
164	Microbial collagenase PKD domain	[+]/BA3584	
165	Oxidoreductase family NAD binding Rossman fold	[+]/BA3655/oxidoreductase, Gfo/Idh/MocA family	gi|21401519
166	Serine protease PDZ domain	[+]/BA3660/serine protease	gi|21401523
167	Anaerobic ribonucleotide triphosphate reductase phage	[+]/BA3663	gi|21401527
168	Spore peptidoglycan hydrolase with LysM domains	[+]/BA3668/Glycosyl hydrolase, family 18	gi|21401531
169	Amidase_2 and peptidoglycan recognition domain	[+]/BA3698/*N*-acetylmuramoyl-L-alanine amidase	gi|21401556
170	Amidase_2 with SH3b domains	[+]/BA3767/*N*-acetylmuramoyl-L-alanine amidase, family 2	
171	Secreted exochitinase Chi36	[+]/BA3854	gi|21401702
172	α/β-hydrolase	[+]/BA3877	gi|21401726
173	Phosphoinositol specific phospholipase C	[+]/BA3891/1-phosphatidylinositol phosphodiesterase	gi|21401737
174	Secreted peptidase S8 (subtilisin-like)	[+]/BA3892	
175	Secreted LysM repeats protein	[+]/BA3893/cell wall hydrolase, putative	gi|21401739
176	Thermostable dipeptidase (Zn dependent)	[+]/BA3911/renal dipeptidase family protein	gi|21401756
177	Amidase 2	[+]/BA4073	gi|21401919
178	Clp protease family protein	[+]/BA4092	gi|21401939
179	Carbon-nitrogen hydrolase	[+]/BA4253	gi|21402087
180	Peptidyl-prolyl cis-trans isomerase (PPlase)	[+]/BA4283/peptidyl-prolyl cis-trans isomerase, cyclophilin-ty	gi|21402116
181	Metallo-dependent phosphatase secreted and LPXTG anchored	[+]/BA4346/2′,3′-cyclic-nucleotide 2′-phosphodiesterase	gi|21402178
182	Prolyl 4-hydroxylase alpha subunit homologue	[+]/BA4459/prolyl 4-hydroxylase-alpha subunit domain	gi|21402285
183	Pterin-4a-carbinolamine dehydratase	[+]/BA4587	gi|21402401
184	FKBP-like prolyl isomerase	[+]/BA4705/trigger factor	gi|21402519
185	Class I glutamine amidotransferase-like protein	[+]/BA4720/ThiJ/PfpI family protein	gi|21402528
186	Spore cortex lytic enzyme	[+]/BA3668/Glycosyl hydrolase	gi|21401531

Continued on following page

TABLE 1—*Continued*

Protein category and ORF no.	Protein name	TIGR's equivalent ORF	A2012 strain equivalent ORF (GenBank)
187	Metallo-dependent hydrolase/putative deaminase	[+]/BA4727/*N*-acyl-D-amino-acid deacylase family	
188	Secreted oxidoreductase/putative acid phosphatase	[+]/BA4746/acid phosphatase	gi|21402552
189	D-Alanyl-D-alanine-carboxypeptidase family protein	[+]/BA4750	gi|21402556
190	Lysophospholipase homolog	[−]/BA5009/conserved hypothetical protein	gi|21402804
191	LPXTG-site transpeptidase family protein	[+]/BA5069	
192	Cu,Zn superoxide dismutase secreted	[+]/BA5139	gi|21397388
193	Phosphoglycerate mutase-like protein	[+]/BA5203	gi|21397451
194	Acid phosphatase	[+]/BA5238/PAP2 family protein	gi|21397479
195	Secreted peptidase M4 bacillolysin	[+]/BA5282/neutral protease B	gi|21397519
196	NLP/P60 family peptidase M27/37	[+]/BA5427/endopeptidase LytE, putative	gi|21397656
197	UDP-*N*-acetyl-D-mannosamine dehydrogenase Cap5O	[+]/BA5512	
198	Heparinase III protein homolog	[−]/BA5513/conserved domain protein	
199	NADH dehydrogenase I, H subunit	[+]/BA5538	gi|21397767
200	Cell-wall associated hydrolase	[+]/BA2849/NLPP60 family protein	gi|21400747
201	Secreted phospholipase D/nuclease	[+]/BA5591/cardiolipin synthase domain protein	gi|21397822
202	UDP-glycosyltransferase/glycogen phosphorylase	[+]/BA5670/glycosyl transferase 1, group 1	gi|21397904
203	Methyl-accepting chemotaxis protein	[+]/BA5673	gi|21397907
204	Secreted metalloprotease	[+]/BA5706/oligoendopeptidase F, putative	gi|21397939
Others			
205	Secreted putative α/β-hydrolase	[−]/BA0177/lipoprotein, putative	gi|21398135
206	Transglycosylase	[−]/BA0543/penicillin-binding domain protein	gi|21398490
207	Outer surface protein	[−]/BA0789/conserved hypothetical protein	
208	Outer surface protein	[−]/BA0795/conserved hypothetical protein	gi|21398756
209	Sensor histidine kinase/response regulator	[+]/BA0996	gi|21398939
210	Siderophore biosynthesis protein	[+]/BA1981	gi|21399857
211	Large anchored repeat protein with FN3 and CBM domains	[+]/BA1094/wall-associated protein, putative	gi|21399020
212	Wall-associated domain protein	[+]/BA1098	
213	PlcR	[+]/BA1163/DNA-binding protein	
214	Spore coat-associated protein, putative	[+]/BA1290	gi|21399194
215	Flagellar motor switch protein, putative/SpoA	[+]/BA1662	gi|21399552
216	Flagellar hook-associated FliD	[+]/BA1671	gi|21399562
217	Flagellar basal-body rod protein	[+]/BA1674/FlgB	gi|21399565
218	FligG/C	[+]/BA1679/flagellar motor switch protein FliG	gi|21399570
219	Similar to flagellar hook assembly protein	[+]/BA1685	gi|21399576
220	Flagellin/IL-4-like	[+]/BA1706	gi|21399593
221	SpoA/flagellin B chain	[−]/BA1710/flagellar motor switch protein FliN, putative	gi|21399598
222	Microcin immunity protein MccF	[+]/BA1949	gi|21399826
223	Bacterial SH3 domain homolog	[+]/BA5481/conserved domain protein	gi|21397703
224	Siderophore biosynthesis protein	[+]/BA1982	gi|21399858
225	ErfK family protein	[−]/BA3232/conserved hypothetical protein	gi|21401118
226	LPXTG-anchored coiled coil protein	[+]/BA3254/LPXTG motif cell wall anchor domain protein	gi|21401137
227	Secreted repeat protein	[+]/BA3366/lipoprotein, putative	gi|21401240
228	LPXTG-anchored repeat protein similar to VLPE prolipoprotein	[+]/BA3367/LPXTG motif cell wall anchor domain protein	gi|21401241
229	TPR repeat/HTH 3 protein	[+]/BA3373/DNA-binding protein	gi|21401248
230	Toxin secretion/phage lysis holin	[+]/BA4074	gi|21401920
231	Wall-associated domain protein	[+]/BA5303/Conserved domain protein	gi|21397539

Continued on following page

TABLE 1—*Continued*

Protein category and ORF no.	Protein name	TIGR's equivalent ORF	A2012 strain equivalent ORF (GenBank)
232	CheY-homologous receiver domain/spo0A	[+]/BA4394/stage 0 sporulation protein A	gi|21402223
233	Wall-associated domain protein	[+]/BA4742	gi|21402548
234	NEAT domain anchored protein	[−]/BA4789/LPXTG motif cell wall anchor domain protein	gi|21402592
235	HlyD family secretion protein, probably MDR protein	[−]/BA4813/conserved hypothetical protein	gi|21402613
236	PhoR	[+]/BA4832/sensory box histidine kinase PhoR	gi|21402631
237	Cell surface anchor family protein	[+]/BA5070	gi|21402863
238	Putative membrane protein	[+]/BA5324	gi|21397559
239	VWA	[−]/BA0623/hypothetical protein	gi|21398581
240	Probable hemolysin	[+]/BA5701/channel protein, hemolysin III family	gi|21397934

[a] Putative candidate vaccine ORFs are grouped in categories, as described in the text. Preliminary functional annotation of the chromosome draft sequence-derived ORF products, is compared to the complete chromosome homologs (78). For easy future reference, ORF products are numbered sequentially (rather than by contig allocation, column 1) and listed side by side with draft genome-based annotation (column 2). Agreement between the chromosome draft sequence-based annotation and TIGR's data is marked in column 3, where identical annotations are marked as [+]. In cases of either discrepancy or disagreement in annotation (marked as [−]), TIGR's annotation is provided as well (column 3). The gi numbers of the A2012 strain equivalent ORF products are given in column 4 (79).

was based on peptide coverage of more than 30% and peptide mass deviation between observed and calculated values of less than 100 ppm.

RESULTS

In silico selection strategy. Preliminary function assignment ("rough" annotation) to the 5,045 ORFs of the *B. anthracis* chromosome (draft sequence, February 2001 version) was the first step towards the selection of vaccine candidate genes. The steps employed were: sequence similarity searches against public databases as well as against genomic databases of interest (*B. cereus* ATCC 14579 draft genome, *C. acetobutylicum* annotated genome [prior to its publication]), and *B. halodurans* genome domain assignments and protein localization prediction studies (see Materials and Methods). The rationale for candidate selection was based on a rational approach comprising of the following criteria: (i) surface-exposed proteins, whether secreted and/or membrane-anchored (see below), as potential immunogens; (ii) ORF products with putative known function, similar to documented virulence factors in other bacteria or containing motifs known to affect virulence, such as repeats; (iii) ORF products with no sequence similarity to proteins present in a nonpathogenic bacteria, particularly bacilli-related, as proteins potentially responsible for the unique attributes of *B. anthracis* virulence.

Sequence similarity searches and taxonomical classification. Blast analysis against the nonredundant database (NCBI) resulted in a preliminary categorization of the 5,045 ORF products into three groups. The first group comprised of ~3,000 ORF products (Fig. 1, right lane), for which a putative function could be assigned based on significant similarity (expectation values smaller than e^{-3}) to a known protein in the databank. The second group comprised of ~800 ORF products assigned as hypothetical, uncharacterized or putative proteins (Fig. 1 middle lane). The third group included ~1,200 ORF products for which no function could be assigned (including ORF products similar to a protein with an unknown function as well as *B. anthracis* unique ORFs, Fig. 1 left lane).

As the first step toward defining chromosome-derived vaccine candidates, the Blast results of the 5,045 ORF products were mapped according to the taxon of hits with expectation value $<e^{-3}$. This mapping enabled us to perform a first round of subtraction of ORF products for which the similarity to a protein in a nonpathogenic organism (e.g., *B. subtilis* and *B. halodurans*) and/or an organism unrelated in terms of evolution (e.g., *Aquificales*, *Thermotogales*, and *Cyanobacteria*) was more significant than the hits to other taxa (expectation values at least 100 times lower than that of the nearest competitor). However, ORF products with hits to pathogenic bacteria from unrelated taxa or from eukaryotic organisms were retained.

Selection of potential vaccine candidates and virulence factors, from the ORF product set with putatively assigned function. Taxonomy-based subtraction of *B. anthracis* ORFs exhibiting sequence similarity to proteins from nonpathogenic bacteria, from the 3,000 "known" ORFs (first group) resulted in ~2,000 ORF products with sequence similarity to known proteins from pathogenic or eukaryotic organisms. We initially scanned the list of the 2,000 ORFs (Fig. 1) for putative housekeeping genes; all genes representing ribosomal proteins, phage proteins and fragmented genes, were subsequently removed. Due to possible future experimental restrictions, which may be imposed by the presence of more than one copy of a selected gene (complementation), ORF products with more than two paralogs in the genome were excluded as well. Similarly, in order to avoid possible cloning problems, putative proteins with more than four predicted trans-membrane segments were also removed. ORF products predicted by protein localization algorithms to code for surface associated or secreted components were selected, as well as proteins with sequence similarity to those described as surface exposed proteins in other bacteria, independent of the in silico prediction. ORF products resembling virulence-associated proteins, irrespective of their cellular location, were retained. The resulting ~450 ORF products were subjected to comparative genomics analysis, removing ORF products with significant similarity

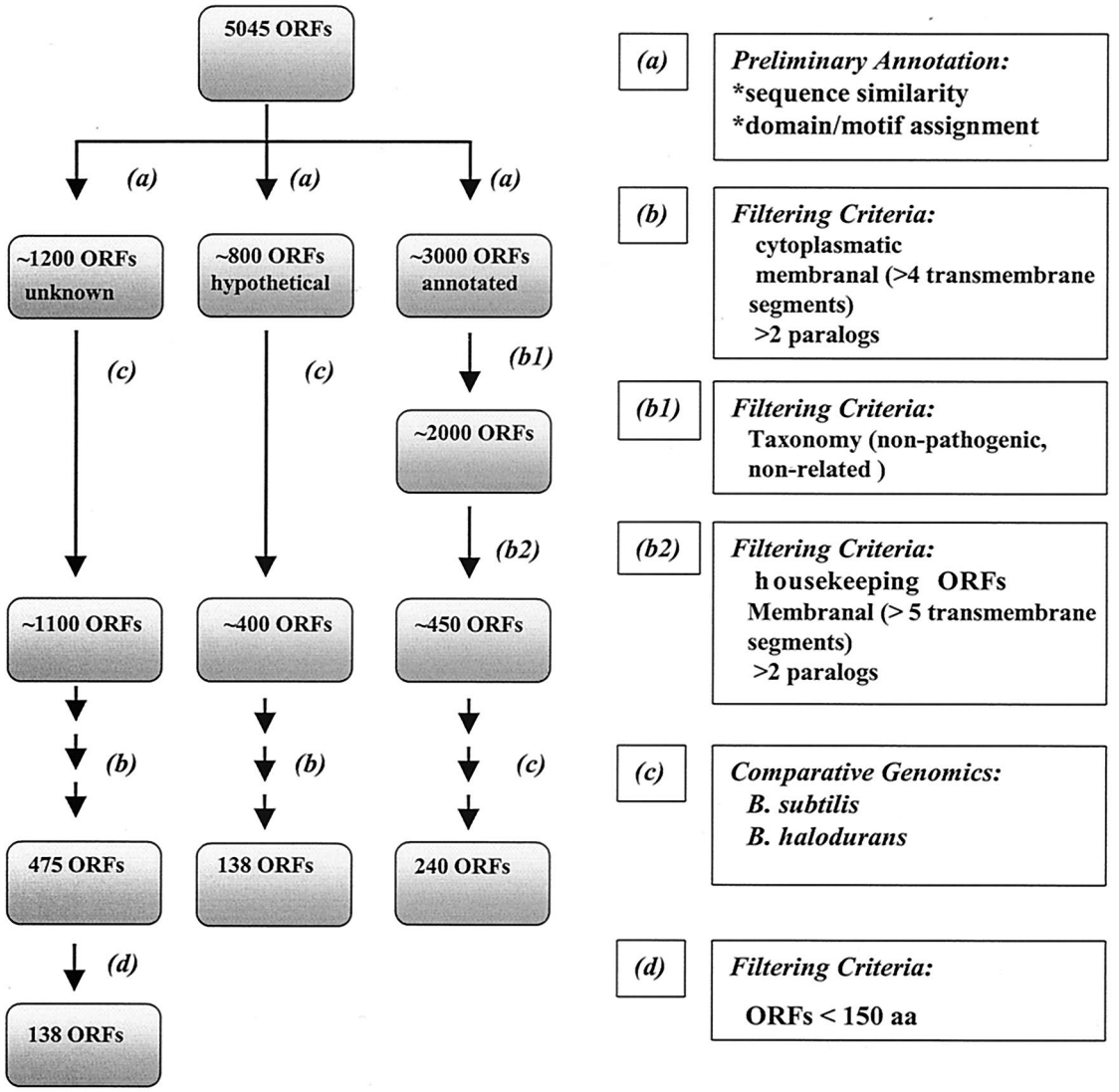

FIG. 1. Strategy for reductive selection of vaccine candidates from the *B. anthracis* chromosome draft sequence (version of February 2001). Flowchart of the computational analysis and the filtering steps, of the *B. anthracis* chromosome candidate selection (left panel). Filtering steps are noted by letters (a through d, as detailed in the right panel), and the resulting number of ORF products at each step is noted in a shaded box.

(not necessarily as first Blast hit) to proteins from the nonpathogenic bacilli *B. subtilis* and *B. halodurans* (upon completion and publication of the genome sequence). The remaining ORF products were inspected individually by manual curation, leaving 240 putative ORF products with potential vaccine and/or virulence relevance as listed in Table 1. For a schematic representation of the reductive strategy, see Fig. 1.

The *B. anthracis* chromosome-derived list of putative vaccine candidates or virulence factors (Table 1) includes, as expected, multiple virulence-related subfamilies, namely, toxins, S-layer homology domain proteins, repeat proteins, adhesions/colonization factors, lytic enzymes, and zinc proteases, etc. All subfamilies were implicated in microbial pathogenesis in different organisms.

S-layer homology domain proteins. The *B. anthracis* cell surface, in the vegetative nonencapsulated state as well as the capsulated state, is covered by a cell wall polymer, known as the surface layer (or S-layer [23, 50, 52, 84]). Various functions have been assigned to the S-layer, ranging from shape maintenance to virulence, host recognition evasion, cell adhesion and resistance to phagocytosis (52, 58, 84). *B. anthracis* is known to synthesize two surface layer (S-layer) proteins, EA1 (extractable antigen 1) and Sap (surface array protein), which account for 5 to 10% of total cellular proteins (52). Both proteins contain a standard signal-peptide followed by three SLH (S-layer homology) anchoring motifs (50). Both proteins are considered major surface antigens and vaccine carriers in vivo (50, 52, 53). Mock and Fouet (58) were the first to report that the *B. anthracis* genome harbors additional genes coding for SLH proteins (other than the two SLH proteins on pXO1 previously identified by Okinaka [66]) which may constitute potential vaccine candidates (7); an amidase on pXO2 (51) and several unidentified genes located on the bacterial chromosome (58)). Inspection of the *B. anthracis* draft version of the

chromosome for ORFs containing at least one SLH domain, revealed the presence of 20 putative S-layer homology domain proteins (including Sap and EA1, products of ORFs 2 and 3). These include cell surface-targeted enzymes such as *N*-acetylmuramoyl-L-alanine-amidases (e.g., products of ORFs 4 and 10) and proteins of unknown function (Table 1, 1st category).

Adhesins. Adhesins, bacterial surface proteins which interact with receptors on the eukaryotic cell, have been studied as targets for vaccine development for many years, since blocking the primary stages of infection could be an effective strategy to prevent bacterial infections (19, 37, 108). Whole-genome sequence of a pathogen could allow for identification of putative novel adhesins based on sequence and/or structural properties shared by bacterial and intracellular human adhesins. Examples for known adhesin families include fibronectin or fibrinogen-binding proteins, collagen adhesins, etc. The *B. anthracis* draft chromosome was found to contain several putative fibronectin-binding proteins, ranging in size from 213 to 1,102 aa (Table 1). The product of ORF 38 is the only member of the group identified as COG 1293 (COGs database [99]) fibronectin-binding proteins and is similar to a fibronectin-binding protein from *Streptococcus pyogenes*, reported to confer protective immunity in mice (35). It also exhibits sequence similarity to putative fibronectin-binding proteins from *B. subtilis*, *B. halodurans* as well as the *S. pneumoniae* adherence and virulence protein A (42% identity). Streptococcal fibronectin-binding proteins (in particular *pavA* from *S. pneumoniae*) were shown to be essential for extra-cellular targeting and efficient cellular invasion (2, 30, 47, 49, 97, 102, 104).

The *B. anthracis* genome harbors one ORF product (ORF 43), which represents another type of putative adhesin: collagen-adhesin, a peptidoglycan anchored protein. Vaccination with a recombinant fragment of the *S. aureus* collagen adhesin, and passive transfer of collagen adhesin-specific antibodies was shown to protect mice against sepsis-induced death (64). Recently it has been reported that human-derived antibodies to the collagen adhesin, such as the *ace*-encoded protein from *Enterococcus faecalis*, expressed during infection in humans, block microbial adherence (61).

Lipoprotein adhesins and autotransporters. A different type of so-called adhesins are lipoproteins, implicated in modulation of the immune system (93). A prototypical adhesin belonging to this group is ORF 36, encoding a 311-aa protein with both secretory and lipoprotein signals and a typical metal-binding site. This ORF product was identified as belonging to COG0830—a zinc-binding lipoprotein of the ABC type (surface adhesin A), involved in zinc uptake. This ORF exhibits extensive sequence similarity to streptococcal adhesins PsaA (pneumococcal surface antigen A) and colonization factors of other gram-positive organisms. The streptococcal adhesin PsaA was shown to play an essential role in virulence (10, 98). Induction of antibody response by immunization with purified PsaA protein or as a DNA vaccine, correlates with protection of mice against otherwise fatal infection with *S. pneumoniae* (56). Although PsaA was originally considered to be an adhesin, it was subsequently shown that the *psa* operon encodes a manganese permease complex (21). Recently, Marra et al. (46) have identified the *psa* promoter, which drives the expression of the *psaBCA* operon, as one of the promoters expressed during lung infection of mice (with *S. pneumoniae*). A similar operon appears to be present in the *B. anthracis* chromosome. In vivo analysis of the *psa* genes demonstrated the importance of this manganese transporter to *S. pneumoniae* virulence. Based on its properties, the product of ORF 36 may be considered a *B. anthracis* vaccine candidate.

Repeat-containing proteins. Many surface-proteins of gram-positive bacteria contain tandem repeat domains that can vary in size from several amino acids to several hundred amino acids. The importance of repeats in understanding protein function, resides not only in their ability to confer multiple binding and structural roles on proteins (6), but also in their possible role in antigenic variation, phase variation (documented mostly for gram-negative organisms and resulting in differential gene expression) and subsequent immune escape (6, 16, 20, 31, 38, 45). The *B. anthracis* genome harbors several proteins (Table 1) with tandem repeats such as TPR-like (e.g., the product of ORF 59), Ankyrin (e.g., the product of ORF 49), Collagen-like (e.g., the product of ORF 73), LRR (leucine rich repeats, e.g., the product of ORF 48) and diverse internal repeats (Table 1). Although certain repeat proteins may have a structural role, others could be involved in *B. anthracis* pathogenesis. Four *B. anthracis* ORF products harbor collagen-like repeats (multiple copies of G-X-X units that form a right-handed triple helix). Of these, the product of ORF 73 was also recently identified as a collagen-like surface glycoprotein and a structural component of the *B. anthracis* exosporium (95). This ORF product, like that of ORF 66, exhibits general sequence similarity to the *Streptococcal* collagen-like protein SclB (76, 106) but also include unique regions. *Streptococcal* collagen-like proteins were shown to participate in adherence to host cells and soft tissue pathology (43, 44, 76, 77). Collagen-like proteins may also be one of the examples for molecular mimicry of host proteins, exploited by many pathogens as a means for the manipulation of host response (90). In eukaryotes they act as membrane bound defense proteins, and are part of the macrophage scavenger receptor and of soluble proteins such as C1q and collectins (101).

Enzymes. The vaccine candidate list includes several enzymes (lytic enzymes, racemases, etc., Table. 1). One example is immune inhibitor A (InhA), a zinc metalloprotease, a putative virulence factor in *B. thuringiensis* known to inhibit the immune system of insects, by specifically cleaving antibacterial proteins produced by the insect host (41, 42). The draft sequence of the *B. anthracis* chromosome contains 2 paralogs (products of ORFs 87 and 105, Table 1), both secreted proteins, annotated as inhA. Blast analysis also shows significant sequence similarity to immune inhibitor-like proteins from *Bacillus stearothermophilus*, *Vibrio cholera* O1, *C. acetobutylicum*, and *Streptomyces coelicolor* (expect values ranging from 0 to e^{-79}). Putative functional significance of this enzyme in *B. anthracis*-related bacilli, was first described by Charlton et al. (15), who reported its presence in the exosporium of *B. cereus* strain ATCC 10876. InhA gene was also shown to exist in a majority of the 23 *B. cereus* strains and the 3 *B. anthracis* strains (26). Regulation of *inhA* expression depends on the transition state regulator AbrB recently shown to be responsible for the timing of toxin expression in *B. anthracis* (83). InhA's presence on the spore surface in bacilli of the *B. cereus* group, suggests that it could play a role in bacterial survival in the host, as reported for other microbial zinc proteases of human microbial

pathogens (57). In a recent study, one of the two *Bacillus thuringiensis inh* genes was shown to be required for pathogenicity via the oral route.

Another class of enzymes are prolyl racemases. Proline racemases were implicated in the virulence of organisms from several genera: *Legionella* (Mip macrophage infectivity potentiator [81]), *Salmonella* (SurA mutants are effective attenuated live oral vaccines [94]), and *Trypanosoma* (80). The *B. anthracis* draft sequence contains three ORF products with sequence similarity to a eukaryoric (*Trypanosoma cruzi*) B-cell mitogen and microbial (*Clostridium difficile*, *Clostridium stricklandi*, and *Pseudomonas aeruginosa*) proline racemases. The product of the ORF 93 is the closest sequence neighbor of the trypanosomal enzyme, maintaining the residues necessary for both catalytic and B-cell mitogenic activity of the *Trypanosoma*. In *Trypanosoma*, this enzyme is considered as a target for the development of vaccines/anti-Chagas' disease drugs (36). Mutants show attenuated virulence and reduced adhesion, and antibodies raised against the *Trypanosoma* enzyme reduce infectivity. The same family of enzymes has been reported as having chaperone activity facilitating preferentially maturation of outer membrane proteins and thus, proline racemases may play a dual role (71). Although these enzymes do not harbor any signal/anchoring motifs, they have been reported to be present on the surface and hence immuno-accessible (80). Their precise putative contribution to *B. anthracis* pathogenesis (whether direct or via a chaperone-like putative activity) needs to be evaluated experimentally.

Yet another example of enzymes with documented evidence of immunogenicity, are autolysins. Autolysins are members of a widely distributed group of enzymes that naturally digest the cell wall peptidoglycan backbone of bacterial organisms, resulting in cell lysis, death and release of inflammatory cell-wall components and cytoplasmic bacterial proteins. These enzymes are located in the cell envelope and are also presumed to play a role in a variety of cellular functions (40). According to their hydrolytic bond specificity, they are classified as muraminidases, glucosaminidases, *N*-acetylmuramoyl-L-alanine-amidases, amidases, and endopeptidases (88). The *B. anthracis* genome contains several ORFs with sequence similarity to autolysins and amidases (Table 1). The product of ORF 11 is the only amidase/autolysin that resembles a documented virulence associated autolysin (lytB *S. pneumoniae* autolysin). This is a 459-aa putative protein with three S-layer homology (SLH) domain anchoring domains in its N terminus (aa 1 to 201), followed by a LytB-like C-terminal domain. Most *B. anthracis* amidases seem to be anchored via SLH domains (either at their N terminus or C terminus [Table 1]) including a recently identified autolysin in pXO2; others are probably anchored via peptidoglycan recognition motifs (e.g., ORF 169) or SH3b domains (e.g., ORF 170); and for some, no anchoring signals were identified (e.g., ORF 84). There is growing evidence of the contribution of autolysins to microbial virulence (55). For example, LytA amidase from *Streptococcus pneumoniae* (82), was shown to induce a protective response when inoculated into the lungs of mice (33). The three *Streptococcal* cell wall hydrolases (LytA, LytB, and LytC), anchored to the membrane via teichoic acid residues, were recently shown to affect colonization in the nasopharynx in *S. pneumoniae*. Results of a genome-based approach to identify vaccine molecules affording protection against *S. pneumoniae* revealed that two out of the six proteins conferring protection are autolysins (LytB and LytC [109]).

Vaccine candidates selected from ORF products with unknown function. The blast analysis carried out on the 5,045 ORF products resulted in ~2,000 ORF products for which a function could not be assigned (Fig. 1 left and middle lanes). This finding is similar to observations in other bacterial genomes, where in spite of the increasing number of sequenced genomes, the assignment of a function to a sequence remains in many cases a challenge: ~20% of the predicted ORF products in a bacterial genome do not match any entry in the databases and an additional 15 to 20% are similar to genes with no known function (24, 25).

Of the 2,000 ORF products with no clues as to their function, ~1,200 had no matches in the databanks whatsoever ("totally unknown") and ~800 were similar to proteins annotated in the databanks as hypothetical, uncharacterized or putative proteins. The first step towards reduction of number of candidate ORF products was identification of genes common to *B. anthracis* and to its taxonomically related yet nonpathogenic *B. halodurans*, irrespective of their putative function. This step resulted in removal of ~500 ORFs, 100 from the category of unknown ORFs and 400 from the category of hypothetical ORFs (see Fig. 1). In an attempt to further reduce the number of candidates to a tangible number, filtering criteria were applied (as described above for ORF products exhibiting sequence similarity to known proteins): selection of candidates was targeted to genes encoding surface-exposed and/or secreted proteins, with less than two paralogs in the chromosome and up to four *trans*-membrane segments. These reductions resulted in a total of 475 unknown and 138 hypothetical ORF products satisfying the above criteria (see Fig. 1). Considering the difficulty in distinguishing short noncoding ORF products from real genes, and given the bias towards short proteins in data sets of ORF products having no matches to proteins in databases (86), a further reduction of the unknown ORF product group, based on ORF length, was carried out. A total of 138 unknown ORF products remain upon filtering out ORF products shorter than 150 aa. The resulting number of unannotated ORF products is still rather large (~280 ORF products); therefore, further reduction is necessary prior to their experimental evaluation (Fig. 1).

Serological proteomic analysis of *B. anthracis* membranes. The proteomic analysis of a *B. anthracis* membrane-associated fraction was employed as a means to verify expression of the predicted membrane-associated candidates. A partial 2-DE proteomic map, representing the membrane-associated chromosomally encoded protein repertoire of a *B. anthracis* late stationary culture, derived from a strain devoid of the two virulence plasmids, was generated (see Materials and Methods and Fig. 2). Close to 100 protein spots (appearing either as unique or as multiple isoforms) were extracted from the gel, digested with trypsin, analyzed by MALDI-TOF MS and identified by comparing their MALDI-TOF spectra with the hypothetical tryptic digests of the *B. anthracis* chromosomal (draft sequence) ORFs data set. The comprehensive proteomic study, including detailed information pertaining to the proteinous composition of the *B. anthracis* cell membrane, will be reported elsewhere (Chitlaru et al., submitted for publication).

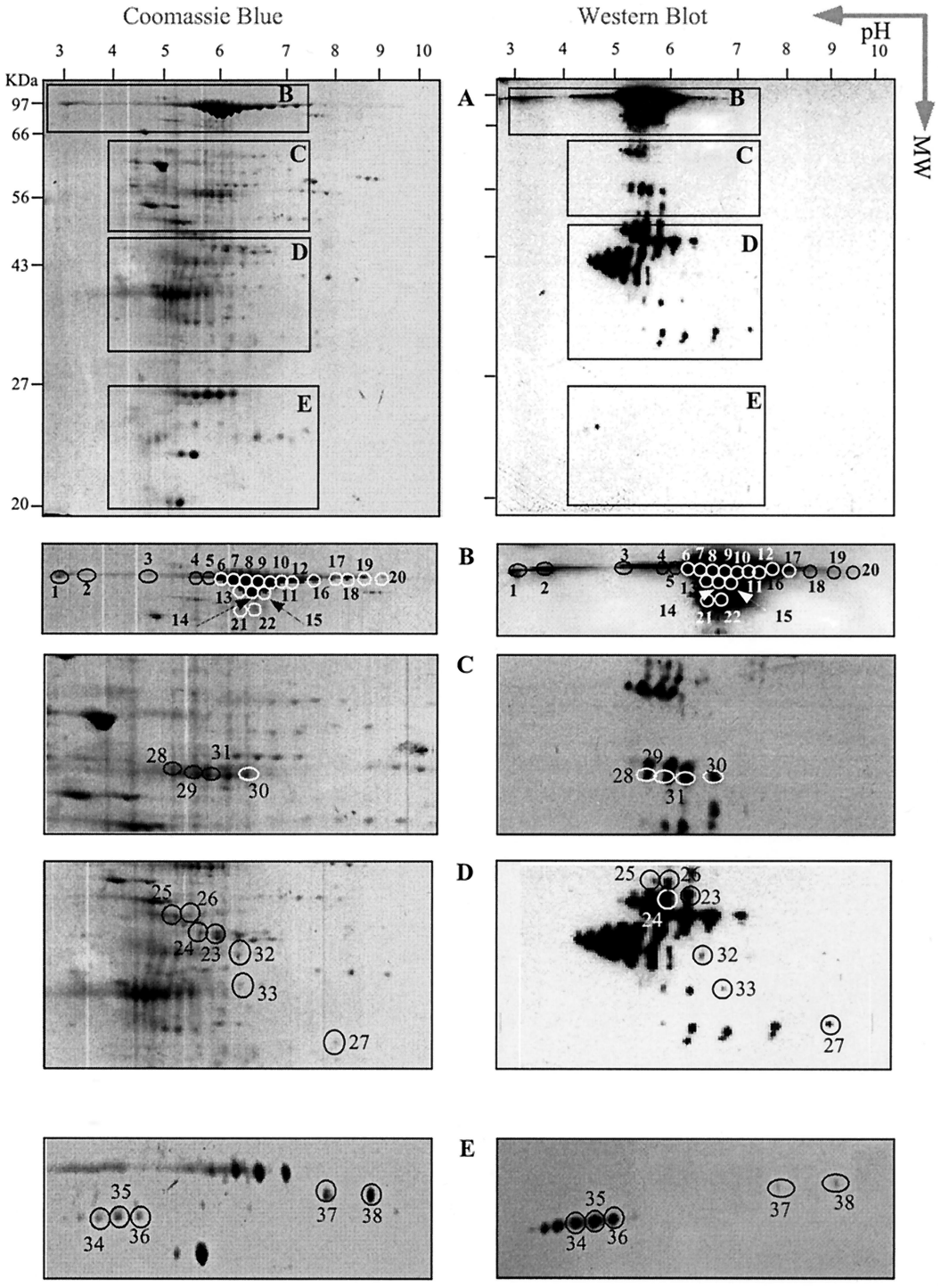

FIG. 2. Serological proteome analysis of *B. anthracis* membranal proteins. *B. anthracis* ATCC Δ14185 (pXO1$^-$, pXO2$^-$) membranal proteins were separated by 2-DE (IEF on a pH 3 to 10 IPG strip). The gel was stained with Coomassie blue, for total protein spot detection. Twin gels were transferred to nitrocellulose membranes and probed with guinea-pig anti-*B. anthracis* immune sera (whole-cell lysate titer of 1:12,800). The Coomassie blue stain is shown on the left and the respective Western blot, on the right. The complete 2-DE gel (A) and enlarged sections (B through E) are shown. Western blots in panels A and regions B through D were probed with 1:1,000 diluted antiserum. For better resolution, the region depicted in E is taken from a 2-DE gel run on a pH 4 to 7 IPG strip (first dimension) and its respective Western blot developed with 1:300-diluted antiserum. The seropositive protein spots, are identified by running numbers. See Table 2 for the complete list of identified seropositive proteins detected in these experiments.

TABLE 2. Seroreactive proteins identified by MALDI-TOF-MS in the *B. anthracis* membranal fraction

Spot no.[a]	Protein name	Draft genome ORF no.[c]	Function[d]
1–15, 21–27	EA1[b]	3	S-layer protein
16–20	Sap	2	S-layer protein
28–30	Dhas		Aldehyde dehydrogenase (BA3609, gi|21401476)
31	MmgE/PrpD		Methylcitrate dehydratase (BA2349, gi|21400220)
34–36	AhpC	82	Hydroxyperoxide reductase
32		19	Novel SLH protein-amidase_2
33	GlpX		Fructose 1–6 biphosphatase classII (BA5576, gi|21397809[e])
37, 38		8	Novel SLH protein-unknown function

[a] Spot numbers are as marked in the gels shown in Fig. 2.
[b] Protein EA1 appears in multiple isoelectric forms as well as in several molecular weight variants. All forms share the same N-terminal amino acid sequence.
[c] Draft genome ORF numbers are as in Table 1.
[d] All functions are noted based on our draft genome annotation. For proteins that were not pre-selected by the bioinformatic approach (Table 1), we provide both TIGR's identification number and A2102 strain gi number.
[e] Appears as a truncated ORF product in the A2102 strain draft sequence.

In order to assess the immunogenic potential of the verified gene products and to identify immunorelevant antigens, a Western blot analysis of the two-dimensional gels with immune anti-*B. anthracis* antiserum was carried out (Fig. 2). By comparing the Coomassie blue-stained 2-DE gels with their respective immunoblots, it appears that at least 38 protein spots (numbered 1 to 38 in Fig. 2) are recognized by the immune sera. Following MALDI-TOF MS analysis of their tryptic digest fragments, it was found that these 38 spots represent isoforms of proteins encoded by 8 distinct ORFs, as detailed in Table 2. The seropositive proteins include four SLH proteins (products of ORF 2, ORF 3, ORF 8, and ORF 19) and four enzymes, one of which (AhpC, ORF 82) exhibits a putative membranal localization lipobox signal. Apart from the two S-layer proteins EA1 and Sap, none of the other seropositive proteins (Table 2) were previously described in *B. anthracis.* Furthermore, with the exception of the two S-layer proteins mentioned above, which were previously reported to cross-react with *B. anthracis* immune sera (58), none of the proteins distinguished by the present serological proteomic analysis, were shown to elicit an immune response in *B. anthracis* exposed animals (see Discussion). Most notably, five out of the eight seropositive proteins (products of ORF 2, ORF 3, ORF 8, ORF 19, and ORF 82 [Table 2]) were predicted to be potential immunogens (Table 1) by the present bioinformatic analysis.

DISCUSSION

The efforts directed towards the development of an improved human anthrax vaccine have been, so far, limited to vaccines based on anthrax toxin components. Availability of the *B. anthracis* chromosomal draft sequence (Ames strain chromosome), allows for in silico identification of putative antigens selected from the complete protein repertoire, thus overcoming preceding limitations, in terms of antigen number and nature, of target based screening for vaccine candidates.

In the study reported herein, *B. anthracis* putative vaccine candidates, representing proteins likely to be surface exposed, and/or similar to documented virulence related proteins, and/or contain sequence motifs characteristic of virulence factors or immunogens, were selected by a multistep computational analysis (Fig. 1) of the draft version of the *B. anthracis* Ames strain chromosome (February 2001, 460 contigs). Integration of the results, together with careful manual curation, resulted in identification of 520 potential antigenic proteins (240 proteins with putatively assigned function and 280 unknown/hypothetical proteins). As shown in Table 1, the 240 proteins with putatively assigned function and relevance to virulence/pathogenicity, could be grouped to several functional categories: SLH proteins, adhesins, repeat proteins, enzymes and "others." Obviously, the assignment of an ORF product to a specific group is not always unequivocal, as larger proteins frequently comprise of more than a single functional domain. However, grouping of the candidates by their putative function, facilitates selection of candidate ORF products from each group, as representatives of families documented to be involved in microbial pathogenesis. For example, the protein products of ORF 1 and ORF 4 are amidases, anchored to the peptidoglycan layer of the bacterial membrane via their SLH domains. These ORF products could be grouped either as SLH proteins or as enzymes. The rationale behind choosing these putative proteins as vaccine candidates, lies in the fact that amidases and autolysins have been shown to be involved in virulence, confer protective immunity and/or act as adhesins (e.g., see reference 109). Moreover, amidases and autolysins are modular enzymes, which probably make use of different membrane anchoring modalities (SLH and choline binding, etc.) as a means of adaptation to a particular biological niche and thus may also act as adhesins (55).

Certain microbial pathogens produce virulence factors expressed only during infection (phase variation), which harbor in their primary sequence patterns known collectively as tandem repeats (Table 1). This group of repeat proteins contains both surface anchored proteins and proteins without obvious secretion and/or anchoring signals (e.g., the most recently documented group of gram-positive virulence determinants named anchorless adhesins [16]). In addition to their reported participation in adhesion, invasion or immune-evasion, repeat proteins have been recently implicated to be involved in Fe^{3+} siderophore regulation (NEAT [near transporter repeat] repeat proteins [5]) and thus may affect the survival of the bacteria within the host. Examples of such putative Fe^{3+} siderophore regulatory proteins are the *B. anthracis* NEAT protein products of ORF 70, ORF 71, ORF 234, and probably

also ORF 6. In particular, the products of ORFs 70 and 71, which are anchored proteins harboring several copies of the NEAT domain, are the best matches to the siderophore regulatory proteins described by Andrade et al. (5), to be present in gram-positive organisms (mostly pathogenic). As mentioned by Andrade et al., in the *B. anthracis* chromosome, these ORFs are indeed located adjacent to iron ABC transporters.

The candidate gene list also includes proteins of unknown function, which are anchored to the bacterial membrane by diverse gram-positive specific anchoring modes. Representatives include sortase-anchored proteins (both iron dependent and independent [13, 18, 48]). Sortases are membrane proteins, which cleave the polypeptide chain between two amino acids within a characteristic C-terminal motif and subsequently catalyze the formation of an amide bond between the carboxyl group of the cleaved polypeptide and the amino group of peptidoglycan cross-bridges. Since most sortase-anchored proteins are considered essential for bacteria to establish successful infection (39), such proteins could be relevant candidates. An example for this group is the product of ORF 228, an anchored repeat protein, relatively unique to *B. anthracis*, exhibiting weak similarity to a protein involved in immune evasion of Mycoplasma.

The remaining large number of candidates, selected by the bioinformatic analyses (520 putative proteins), necessitates implementation of additional filtering strategies. Reducing the number of candidates to be evaluated experimentally, to a manageable number, would involve application of a high throughput biological screening system (such as proteomic-based analysis of in vivo-expressed immunogens or in vitro-in vivo expression systems [7, 14, 25]); and/or application of additional computational steps directed toward selection of microorganism-specific genes.

B. anthracis, *B. cereus*, and *B. thuringiensis* are considered essentially one genetic species and members of the *B. cereus* group of bacteria (28). In spite of the overall genetic similarity, the fact that these species differ significantly in pathogenesis may imply the presence of *B. anthracis*-specific virulence determinants. Subtraction of genes common to *B. anthracis* and *B. cereus* 14579 (gapped genome, Integrated Genomics, Inc.) could reveal the presence of *B. anthracis*-specific genes. Preliminary subtraction, from the group of putative proteins with assigned functions (known proteins), of ORF products exhibiting significant overall sequence similarity to *B. cereus* orthologs, results in ~80 *B. anthracis*-specific ORF products, leaving out most of the classical *B. cereus* group virulence factors. As for the unknown protein subgroup, ~100 ORF products did not exhibit extensive sequence similarity to *B. cereus* 14579 proteins. Such subtraction was not performed only on the basis of threshold values (mutual coverage of 85%, expectation values smaller than e^{-10}), but included other considerations regarding specific sequence variations (e.g., extent of insertions, divergence etc). In view of the fact that differences may also be ascribed to differences in genomic context, and in view of the limitations imposed by the quality of draft genomes versus complete genomes, this reduction should be carried out more carefully once sequencing of the more closely related *B. cereus* 10987 is finished. One should also keep in mind that effective immunogens may not necessarily be *B. anthracis* specific; thus, this type of subtraction should be probably applied only as an optional reductive measure.

In order to demonstrate the expression and cellular location of the in silico selected chromosomal gene products and in order to expedite the identification of *B. anthracis* immunogenic membrane and/or outer surface proteins, a direct proteomic inspection of *B. anthracis* membrane protein fraction was carried out. The proteomic analysis involved separation of a *B. anthracis* Δ14185 (pXO1$^-$, pXO2$^-$) subcellular membranal fraction by 2-DE, and identification of the most abundant protein spots by MALDI-TOF MS analysis of their fingerprint tryptic digestion products. Close to 100 spots from the 2-DE gel were analyzed and found to represent 32 proteins (detailed results of the analysis are documented in another report [Chitlaru et al., submitted]). In interpreting the proteomic data, one should take into account the fact that the proteomic approach is characterized by inherent underestimation of gene products due to (i) bias toward identification of abundant proteins (this effect is even more pronounced when one particular protein species prevails in the preparation as is the case in the membrane fraction of *B. anthracis* and the S-layer protein EA1 (e.g., Fig. 1, box B); (ii) differential proteins expression, depending on the origin of the membrane fraction (culture conditions and in vitro versus in vivo gene expression); (iii) sample preparation procedure. Membrane proteins are notoriously difficult to separate by 2-DE due to solubilization constraints and may not be represented in the two-dimensional map, despite their abundance. It should be noted that the bioinformatic approach may circumvent the above limitations and therefore may result in identification of gene candidates representing the complete gene repertoire of each organism, which following further individual exploration of their immunogenic potential will aid in developing improved protective and therapeutic measures.

Here we report on a serological proteome analysis carried out in order to address the issue of in vivo immunogenicity of the *B. anthracis* membrane proteins identified. Thirty-eight spots were found to cross-react with sera from *B. anthracis* infected animals (Fig. 2). The analysis also established that the cross-reactive spots, which represent the products of 8 ORFs, are indeed expressed in vivo (in guinea pigs) during exposure to *B. anthracis*, and are able to elicit an immune response (Fig. 2 and Table 2). Most notably, five out of these eight proteins (four SLH proteins and the AhpC/peroxiredoxin) were predicted to be potentially antigenic by the present independent in silico survey. It is worth noting that although antibodies against S-layer proteins EA1 and Sap (ORFs 3 and 2, respectively) were described before in infected animals (22, 58), neither the expression nor the in vivo immunogenicity of the other 2 novel SLH proteins (ORFs 8 and 19) was noted before. Humoral response against AhpC (ORF 82) was reported in other virulent bacterial systems such as *Legionella pneumophila* and *Helicobacter pylori* (27, 72, 91) but not for *B. anthracis* or *B. cereus*. The seropositive methylcitrate dehydrates MngE/PrpD (gi|21400220, Table2) was not previously invoked as a potential immunogen in other systems, yet it was shown to be necessary for survival of *Legionella* in the macrophage (72, 91). It is worth noting that about 50% of the immunogenic proteins identified in the Western blot of the 2-DE gel, are S-layer homology domain proteins (Table 2). Nevertheless, comparison of the

Coomassie blue-stained 2-DE gels and their respective Western blots (Fig. 2), appears to reveal a differential order of immunopotencies among these seropositive proteins. For example, AhpC (spots 34 to 36) appears to be an exceptionally strong immunogen, since in the Coomassie blue-stained gel it appears as a weak signal while in the Western blot it appears as an intense signal.

In conclusion, as demonstrated in this study, combining bioinformatic chromosome screening with serological proteome analysis allows for judicious selection of in vivo immunogens. While the bioinformatic strategy resulted in identification of 240 vaccine candidates with putative functions (out of the 5,045 assigned and annotated ORFs derived from the chromosome *B. anthracis* draft sequence), the serological proteome analysis enables to focus on putative anthrax vaccine candidate genes by confirming their in vivo expression and antigenicity.

ACKNOWLEDGMENTS

We thank E. V. Koonin for fruitful scientific discussion and D. Lipman, S. Bryant, A. Marchler-Bauer, A. Panchenko, and J. Fleshman from the NCBI for their interest and support. We are deeply indebted to T. D. Read (TIGR, Rockville, Md.) for his cooperation from the beginning of this project and for providing access (prior to publication) to TIGR's complete genome derived chromosomal ORFs sequence (78). We also thank Menachem Lion and Dror Bar-Natan from the IIBR for their help in genome-based tryptic peptide fingerprint screening and for performing local and remote computational analyses.

REFERENCES

1. **Abramova, F. A., L. M. Grinberg, O. V. Yampolskaya, and D. H. Walker.** 1993. Pathology of inhalational anthrax in 42 cases from the Sverdlovsk outbreak of 1979. Proc. Natl. Acad. Sci. USA **90:**2291–2294.
2. **Ahmed, S., S. Meghji, R. J. Williams, B. Henderson, J. H. Brock, and S. P. Nair.** 2001. *Staphylococcus aureus* fibronectin binding proteins are essential for internalization by osteoblasts but do not account for differences in intracellular levels of bacteria. Infect. Immun. **69:**2872–2877.
3. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
4. **Altschul, S. F., and E. V. Koonin.** 1998. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. Trends Biochem. Sci. **23:**444–447.
5. **Andrade, M. A., F. D. Ciccarelli, C. Perez-Iratxeta, and P. Bork.** 2002. NEAT: a domain duplicated in genes near the components of a putative Fe^{3+} siderophore transporter from Gram-positive pathogenic bacteria. Genome Biol. **3:**1–5.
6. **Andrade, M. A., C. Perez-Iratxeta, and C. P. Ponting.** 2001. Protein repeats: structures, functions, and evolution. J. Struct. Biol. **134:**117–131.
7. **Ariel, N., A. Zvi, H. Grosfeld, O. Gat, Y. Inbar, B. Velan, S. Cohen, and A. Shafferman.** 2002. Search for potential vaccine candidate open reading frames in the *Bacillus anthracis* virulence plasmid pXO1: in silico and in vitro screening. Infect. Immun. **70:**6817–6827.
8. **Baillie, L.** 2001. The development of new vaccines against Bacillus anthracis. J. Appl. Microbiol. **91:**609–613.
9. **Baillie, L., and T. D. Read.** 2001. Bacillus anthracis, a bug with attitude! Curr. Opin. Microbiol. **4:**78–81.
10. **Berry, A. M., and J. C. Paton.** 2000. Additive attenuation of virulence of *Streptococcus pneumoniae* by mutation of the genes encoding pneumolysin and other putative pneumococcal virulence proteins. Infect. Immun. **68:**133–140.
11. **Brossier, F., M. Levy, and M. Mock.** 2002. Anthrax spores make an essential contribution to vaccine efficacy. Infect. Immun. **70:**661–664.
12. **Brossier, F., and M. Mock.** 2001. Toxins of Bacillus anthracis. Toxicon **39:**1747–1755.
13. **Cabanes, D., P. Dehoux, O. Dussurget, L. Frangeul, and P. Cossart.** 2002. Surface proteins and the pathogenic potential of Listeria monocytogenes. Trends Microbiol. **10:**238–245.
14. **Chakravarti, D. N., M. J. Fiske, L. D. Fletcher, and R. J. Zagursky.** 2000. Application of genomics and proteomics for identification of bacterial gene products as potential vaccine candidates. Vaccine **19:**601–612.
15. **Charlton, S., A. J. Moir, L. Baillie, and A. Moir.** 1999. Characterization of the exosporium of Bacillus cereus. J. Appl. Microbiol. **87:**241–245.
16. **Chhatwal, G. S.** 2002. Anchorless adhesins and invasins of Gram-positive bacteria: a new class of virulence factors. Trends Microbiol. **10:**205–208.
17. **Cohen, S., I. Mendelson, Z. Altboum, D. Kobiler, E. Elhanany, T. Bino, M. Leitner, I. Inbar, H. Rosenberg, Y. Gozes, R. Barak, M. Fisher, C. Kronman, B. Velan, and A. Shafferman.** 2000. Attenuated nontoxinogenic and nonencapsulated recombinant *Bacillus anthracis* spore vaccines protect against anthrax. Infect. Immun. **68:**4549–4558.
18. **Cossart, P., and R. Jonquieres.** 2000. Sortase, a universal target for therapeutic agents against gram-positive bacteria? Proc. Natl. Acad. Sci. USA **97:**5013–5015.
19. **Cunningham, M. W.** 2000. Pathogenesis of group A streptococcal infections. Clin. Microbiol. Rev. **13:**470–511.
20. **De Bolle, X., C. D. Bayliss, D. Field, T. van de Ven, N. J. Saunders, D. W. Hood, and E. R. Moxon.** 2000. The length of a tetranucleotide repeat tract in Haemophilus influenzae determines the phase variation rate of a gene with homology to type III DNA methyltransferases. Mol. Microbiol. **35:**211–222.
21. **Dintilhac, A., G. Alloing, C. Granadel, and J. P. Claverys.** 1997. Competence and virulence of Streptococcus pneumoniae: Adc and PsaA mutants exhibit a requirement for Zn and Mn resulting from inactivation of putative ABC metal permeases. Mol. Microbiol. **25:**727–739.
22. **Fouet, A., and S. Mesnage.** 2002. Bacillus anthracis cell envelope components. Curr. Top. Microbiol. Immunol. **271:**87–113.
23. **Fouet, A., S. Mesnage, E. Tosi-Couture, P. Gounon, and M. Mock.** 1999. Bacillus anthracis surface: capsule and S-layer. J. Appl. Microbiol. **87:**251–255.
24. **Fraser, C. M.** 2000. Microbial genome sequencing: new insights into physiology and evolution. Novartis Found. Symp. **229:**54–62.
25. **Grandi, G.** 2001. Antibacterial vaccine design using genomics and proteomics. Trends Biotechnol. **19:**181–188.
26. **Guttmann, D. M., and D. J. Ellar.** 2000. Phenotypic and genotypic comparisons of 23 strains from the Bacillus cereus complex for a selection of known and putative B. thuringiensis virulence factors. FEMS Microbiol. Lett. **188:**7–13.
27. **Haas, G., G. Karaali, K. Ebermayer, W. G. Metzger, S. Lamer, U. Zimny-Arndt, S. Diescher, U. B. Goebel, K. Vogt, A. B. Roznowski, B. J. Wiedenmann, T. F. Meyer, T. Aebischer, and P. R. Jungblut.** 2002. Immunoproteomics of Helicobacter pylori infection and relation to gastric disease. Proteomics **2:**313–324.
28. **Helgason, E., O. A. Okstad, D. A. Caugant, H. A. Johansen, A. Fouet, M. Mock, I. Hegna, and Kolsto.** 2000. *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis*: one species on the basis of genetic evidence. Appl. Environ. Microbiol. **66:**2627–2630.
29. **Hoffman, K., and W. Stoffel.** 1993. TMbase—a database for membrane spanning protein segments. Biol. Chem. Hoppe-Seyler **374:**166–186.
30. **Holmes, A. R., R. McNab, K. W. Millsap, M. Rohde, S. Hammerschmidt, J. L. Mawdsley, and H. F. Jenkinson.** 2001. The pavA gene of Streptococcus pneumoniae encodes a fibronectin-binding protein that is essential for virulence. Mol. Microbiol. **41:**1395–1408.
31. **Hood, D. W., M. E. Deadman, M. P. Jennings, M. Bisercic, R. D. Fleischmann, J. C. Venter, and E. R. Moxon.** 1996. DNA repeats identify novel virulence genes in Haemophilus influenzae. Proc. Natl. Acad. Sci. USA **93:**11121–11125.
32. **Janulczyk, R., and M. Rasmussen.** 2001. Improved pattern for genome-based screening identifies novel cell wall-attached proteins in gram-positive bacteria. Infect. Immun. **69:**4019–4026.
33. **Jedrzejas, M. J.** 2001. Pneumococcal virulence factors: structure and function. Microbiol. Mol. Biol. Rev. **65:**187–207.
34. **Jernigan, J. A., D. S. Stephens, D. A. Ashford, C. Omenaca, M. S. Topiel, M. Galbraith, M. Tapper, T. L. Fisk, S. Zaki, T. Popovic, R. F. Meyer, C. P. Quinn, S. A. Harper, S. K. Fridkin, J. J. Sejvar, C. W. Shepard, M. McConnell, J. Guarner, W. J. Shieh, J. M. Malecki, J. L. Gerberding, J. M. Hughes, and B. A. Perkins.** 2001. Bioterrorism-related inhalational anthrax: the first 10 cases reported in the United States. Emerg. Infect. Dis. **7:**933–944.
35. **Kawabata, S., E. Kunitomo, Y. Terao, I. Nakagawa, K. Kikuchi, K. Totsuka, and S. Hamada.** 2001. Systemic and mucosal immunizations with fibronectin-binding protein FBP54 induce protective immune responses against *Streptococcus pyogenes* challenge in mice. Infect. Immun. **69:**924–930.
36. **Kelly, J. M.** 2000. A B-cell activator in Chagas disease. Nat. Med. **6:**865–866.
37. **Klemm, P., and M. A. Schembri.** 2000. Bacterial adhesins: function and structure. Int. J. Med. Microbiol. **290:**27–35.
38. **Kobe, B., and A. V. Kajava.** 2001. The leucine-rich repeat as a protein recognition motif. Curr. Opin. Struct. Biol. **11:**725–732.
39. **Lee, S. G., V. Pancholi, and V. A. Fischetti.** 2002. Characterization of a unique glycosylated anchor endopeptidase that cleaves the LPXTG sequence motif of cell surface proteins of Gram-positive bacteria. J. Biol. Chem. **277:**46912–46922.
40. **Lopez, R., M. P. Gonzalez, E. Garcia, J. L. Garcia, and P. Garcia.** 2000. Biological roles of two new murein hydrolases of Streptococcus pneumoniae representing examples of module shuffling. Res. Microbiol. **151:**437–443.
41. **Lovgren, A., C. R. Carlson, K. Eskils, and A. B. Kolsto.** 1998. Localization

of putative virulence genes on a physical map of the bacillus thuringiensis subsp. gelechiae chromosome. Curr. Microbiol. **37:**245–250.
42. **Lovgren, A., M. Zhang, A. Engstrom, G. Dalhammar, and R. Landen.** 1990. Molecular characterization of immune inhibitor A, a secreted virulence protease from Bacillus thuringiensis. Mol. Microbiol. **4:**2137–2146.
43. **Lukomski, S., K. Nakashima, I. Abdi, V. J. Cipriano, R. M. Ireland, S. D. Reid, G. G. Adams, and J. M. Musser.** 2000. Identification and characterization of the *scl* gene encoding a group A *Streptococcus* extracellular protein virulence factor with similarity to human collagen. Infect. Immun. **68:**6542–6553.
44. **Lukomski, S., K. Nakashima, I. Abdi, V. J. Cipriano, B. J. Shelvin, E. A. Graviss, and J. M. Musser.** 2001. Identification and characterization of a second extracellular collagen-like protein made by group A *Streptococcus*: control of production at the level of translation. Infect. Immun. **69:**1729–1738.
45. **Marcotte, E. M., M. Pellegrini, T. O. Yeates, and D. Eisenberg.** 1999. A census of protein repeats. J. Mol. Biol. **293:**151–160.
46. **Marra, A., J. Asundi, M. Bartilson, S. Lawson, F. Fang, J. Christine, C. Wiesner, D. Brigham, W. P. Schneider, and A. E. Hromockyj.** 2002. Differential fluorescence induction analysis of *Streptococcus pneumoniae* identifies genes involved in pathogenesis. Infect. Immun. **70:**1422–1433.
47. **Massey, R. C., M. N. Kantzanou, T. Fowler, N. P. Day, K. Schofield, E. R. Wann, A. R. Berendt, M. Hook, and S. J. Peacock.** 2001. Fibronectin-binding protein A of Staphylococcus aureus has multiple, substituting, binding regions that mediate adherence to fibronectin and invasion of endothelial cells. Cell. Microbiol. **3:**839–851.
48. **Mazmanian, S. K., H. Ton-That, K. Su, and O. Schneewind.** 2002. An iron-regulated sortase anchors a class of surface protein during Staphylococcus aureus pathogenesis. Proc. Natl. Acad. Sci. USA **99:**2293–2298.
49. **Meehan, M., Y. Lynagh, C. Woods, and P. Owen.** 2001. The fibrinogen-binding protein (FgBP) of Streptococcus equi subsp. equi additionally binds IgG and contributes to virulence in a mouse model. Microbiology **147:**3311–3322.
50. **Mesnage, S., T. Fontaine, T. Mignot, M. Delepierre, M. Mock, and A. Fouet.** 2000. Bacterial SLH domain proteins are non-covalently anchored to the cell surface via a conserved mechanism involving wall polysaccharide pyruvylation. EMBO J. **19:**4473–4484.
51. **Mesnage, S., and A. Fouet.** 2002. Plasmid-encoded autolysin in *Bacillus anthracis*: modular structure and catalytic properties. J. Bacteriol. **184:**331–334.
52. **Mesnage, S., E. Tosi-Couture, M. Mock, and A. Fouet.** 1999. The S-layer homology domain as a means for anchoring heterologous proteins on the cell surface of Bacillus anthracis. J. Appl. Microbiol. **87:**256–260.
53. **Mesnage, S., E. Tosi-Couture, M. Mock, P. Gounon, and A. Fouet.** 1997. Molecular characterization of the Bacillus anthracis main S-layer component: evidence that it is the major cell-associated antigen. Mol. Microbiol. **23:**1147–1155.
54. **Mignot, T., M. Mock, and A. Fouet.** 2003. A plasmid-encoded regulator couples the synthesis of toxins and surface structures in Bacillus anthracis. Mol. Microbiol. **47:**917–927.
55. **Milohanic, E., R. Jonquieres, P. Cossart, P. Berche, and J. L. Gaillard.** 2001. The autolysin Ami contributes to the adhesion of Listeria monocytogenes to eukaryotic cells via its cell wall anchor. Mol. Microbiol. **39:**1212–1224.
56. **Miyaji, E. N., W. O. Dias, M. Gamberini, V. C. Gebara, R. P. Schenkman, J. Wild, P. Riedl, J. Reimann, R. Schirmbeck, and L. C. Leite.** 2001. PsaA (pneumococcal surface adhesin A) and PspA (pneumococcal surface protein A) DNA vaccines induce humoral and cellular immune responses against Streptococcus pneumoniae. Vaccine **20:**805–812.
57. **Miyoshi, S., and S. Shinoda.** 2000. Microbial metalloproteases and pathogenesis. Microbes Infect. **2:**91–98.
58. **Mock, M., and A. Fouet.** 2001. Anthrax. Annu. Rev. Microbiol. **55:**647–671.
59. **Montigiani, S., F. Falugi, M. Scarselli, O. Finco, R. Petracca, G. Galli, M. Mariani, R. Manetti, M. Agnusdei, R. Cevenini, M. Donati, R. Nogarotto, N. Norais, I. Garaguso, S. Nuti, G. Saletti, D. Rosa, G. Ratti, and G. Grandi.** 2002. Genomic approach for analysis of surface proteins in *Chlamydia pneumoniae*. Infect. Immun. **70:**368–379.
60. **Nakai, K., and P. Horton.** 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem. Sci. **24:**34–36.
61. **Nallapareddy, S. R., X. Qin, G. M. Weinstock, M. Hook, and B. E. Murray.** 2000. Enterococcus faecalis adhesin, ace, mediates attachment to extracellular matrix proteins collagen type IV and laminin as well as collagen type I. Infect. Immun. **68:**5218–5224.
62. **Navarre, W. W., and O. Schneewind.** 1999. Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. Microbiol. Mol. Biol. Rev. **63:**174–229.
63. **Nielsen, H., J. Engelbrecht, S. Brunak, and G. von Heijne.** 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. **10:**1–6.
64. **Nilsson, I. M., J. M. Patti, T. Bremell, M. Hook, and A. Tarkowski.** 1998. Vaccination with a recombinant fragment of collagen adhesin provides protection against Staphylococcus aureus-mediated septic death. J. Clin. Investig. **101:**2640–2649.
65. **Nolling, J., G. Breton, M. V. Omelchenko, K. S. Makarova, Q. Zeng, R. Gibson, H. M. Lee, J. Dubois, D. Qiu, J. Hitti, Y. I. Wolf, R. L. Tatusov, F. Sabathe, L. Doucette-Stamm, P. Soucaille, M. J. Daly, G. N. Bennett, E. V. Koonin, and D. R. Smith.** 2001. Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. J. Bacteriol. **183:**4823–4838.
66. **Okinaka, R., K. Cloud, O. Hampton, A. Hoffmaster, K. Hill, P. Keim, T. Koehler, G. Lamke, S. Kumano, D. Manter, Y. Martinez, D. Ricke, R. Svensson, and P. Jackson.** 1999. Sequence, assembly and analysis of pX01 and pX02. J. Appl. Microbiol. **87:**261–262.
67. **Pezard, C., P. Berche, and M. Mock.** 1991. Contribution of individual toxin components to virulence of *Bacillus anthracis*. Infect. Immun. **59:**3472–3477.
68. **Pezard, C., E. Duflot, and M. Mock.** 1993. Construction of Bacillus anthracis mutant strains producing a single toxin component. J. Gen. Microbiol. **139:**2459–2463.
69. **Pizza, M., V. Scarlato, V. Masignani, M. M. Giuliani, B. Arico, M. Comanducci, G. T. Jennings, L. Baldi, E. Bartolini, B. Capecchi, C. L. Galeotti, E. Luzzi, R. Manetti, E. Marchetti, M. Mora, S. Nuti, G. Ratti, L. Santini, S. Savino, M. Scarselli, E. Storni, P. Zuo, M. Broeker, E. Hundt, B. Knapp, E. Blair, T. Mason, H. Tettelin, D. W. Hood, A. C. Jeffries, N. J. Saunders, D. M. Granoff, J. C. Venter, E. R. Moxon, G. Grandi, and R. Rappuoli.** 2000. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science **287:**1816–1820.
70. **Plotkin, S. A., P. S. Brachman, M. Utell, F. H. Bumford, and M. M. Atchison.** 2002. An epidemic of inhalation anthrax, the first in the twentieth century. I. Clinical features. Am. J. Med. **112:**4–12. [Reprint of Am. J. Med. **29:**992–1001, 1960.]
71. **Ramm, K., and A. Pluckthun.** 2001. High enzymatic activity and chaperone function are mechanistically related features of the dimeric E. coli peptidyl-prolyl-isomerase FkpA. J. Mol. Biol. **310:**485–498.
72. **Rankin, S., Z. Li, and R. R. Isberg.** 2002. Macrophage-induced genes of *Legionella pneumophila*: protection from reactive intermediates and solute imbalance during intracellular growth. Infect. Immun. **70:**3637–3648.
73. **Rappuoli, R.** 2001. Conjugates and reverse vaccinology to eliminate bacterial meningitis. Vaccine **19:**2319–2322.
74. **Rappuoli, R.** 2000. Reverse vaccinology. Curr Opin. Microbiol. **3:**445–450.
75. **Rappuoli, R.** 2001. Reverse vaccinology, a genome-based approach to vaccine development. Vaccine **19:**2688–2691.
76. **Rasmussen, M., and L. Bjorck.** 2001. Unique regulation of SclB: a novel collagen-like surface protein of Streptococcus pyogenes. Mol. Microbiol. **40:**1427–1438.
77. **Rasmussen, M., A. Eden, and L. Bjorck.** 2000. SclA, a novel collagen-like surface protein of *Streptococcus pyogenes*. Infect. Immun. **68:**6370–6377.
78. **Read, T. D., S. N. Peterson, N. Tourasse, L. W. Baillie, I. T. Paulsen, K. E. Nelson, H. Tettelin, D. E. Fouts, J. A. Eisen, S. R. Gill, E. K. Holtzapple, O. A. Okstad, E. Helgason, J. Rilstone, M. Wu, J. K. Kolonay, M. J. Beanan, R. J. Dodson, L. M. Brinkac, M. Gwinn, R. T. DeBoy, R. Madpu, S. C. Daugherty, A. S. Durkin, D. H. Haft, W. C. Nelson, J. D. Peterson, M. Pop, H. M. Khouri, D. Radune, J. L. Benton, Y. Mahamoud, L. Jiang, I. R. Hance, J. F. Weidman, K. J. Berry, R. D. Plaut, A. M. Wolf, K. L. Watkins, W. C. Nierman, A. Hazen, R. Cline, C. Redmond, J. E. Thwaite, O. White, S. L. Salzberg, B. Thomason, A. M. Friedlander, T. M. Koehler, P. C. Hanna, A. B. Kolsto, and C. M. Fraser.** 2003. The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. Nature **423:**81–86.
79. **Read, T. D., S. L. Salzberg, M. Pop, M. Shumway, L. Umayam, L. Jiang, E. Holtzapple, J. D. Busch, K. L. Smith, J. M. Schupp, D. Solomon, P. Keim, and C. M. Fraser.** 2002. Comparative genome sequencing for discovery of novel polymorphisms in Bacillus anthracis. Science **296:**2028–2033.
80. **Reina-San-Martin, B., W. Degrave, C. Rougeot, A. Cosson, N. Chamond, A. Cordeiro-Da-Silva, M. Arala-Chaves, A. Coutinho, and P. Minoprio.** 2000. A B-cell mitogen from a pathogenic trypanosome is a eukaryotic proline racemase. Nat. Med. **6:**890–897.
81. **Riboldi-Tunnicliffe, A., B. Konig, S. Jessen, M. S. Weiss, J. Rahfeld, J. Hacker, G. Fischer, and R. Hilgenfeld.** 2001. Crystal structure of Mip, a prolylisomerase from Legionella pneumophila. Nat. Struct. Biol. **8:**779–783.
82. **Ronda, C., J. L. Garcia, E. Garcia, J. M. Sanchez-Puelles, and R. Lopez.** 1987. Biological role of the pneumococcal amidase. Cloning of the lytA gene in Streptococcus pneumoniae. Eur. J. Biochem. **164:**621–624.
83. **Saile, E., and T. M. Koehler.** 2002. Control of anthrax toxin gene expression by the transition state regulator *abrB*. J. Bacteriol. **184:**370–380.
84. **Sara, M., and U. B. Sleytr.** 2000. S-layer proteins. J. Bacteriol. **182:**859–868.
85. **Schultz, J., R. R. Copley, T. Doerks, C. P. Ponting, and P. Bork.** 2000. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res. **28:**231–234.
86. **Skovgaard, M., L. J. Jensen, S. Brunak, D. Ussery, and A. Krogh.** 2001. On the total number of genes and their length distribution in complete microbial genomes. Trends Genet. **17:**425–428.
87. **Smith, R. F., B. A. Wiese, M. K. Wojzynski, D. B. Davison, and K. C.**

Worley. 1996. BCM Search Launcher—an integrated interface to molecular biology data base search and analysis services available on the World Wide Web. Genome Res. **6:**454–462.
88. **Smith, T. J., S. A. Blackman, and S. J. Foster.** 2000. Autolysins of Bacillus subtilis: multiple enzymes with multiple functions. Microbiology **146:**249–262.
89. **Sonnhammer, E. L., S. R. Eddy, E. Birney, A. Bateman, and R. Durbin.** 1998. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res. **26:**320–322.
90. **Stebbins, C. E., and J. E. Galan.** 2001. Structural mimicry in bacterial virulence. Nature **412:**701–705.
91. **Stone, B. J., A. Brier, and Y. A. Kwaik.** 1999. The Legionella pneumophila prp locus; required during infection of macrophages and amoebae. Microb. Pathog. **27:**369–376.
92. **Sutcliffe, I. C., and D. J. Harrington.** 2002. Pattern searches for the identification of putative lipoprotein genes in Gram-positive bacterial genomes. Microbiology **148:**2065–2077.
93. **Sutcliffe, I. C., and R. R. Russell.** 1995. Lipoproteins of gram-positive bacteria. J. Bacteriol. **177:**1123–1128.
94. **Sydenham, M., G. Douce, F. Bowe, S. Ahmed, S. Chatfield, and G. Dougan.** 2000. *Salmonella enterica* serovar Typhimurium *surA* mutants are attenuated and effective live oral vaccines. Infect. Immun. **68:**1109–1115.
95. **Sylvestre, P., E. Couture-Tosi, and M. Mock.** 2002. A collagen-like surface glycoprotein is a structural component of the Bacillus anthracis exosporium. Mol. Microbiol. **45:**169–178.
96. **Takami, H., K. Nakasone, Y. Takaki, G. Maeno, R. Sasaki, N. Masui, F. Fuji, C. Hirama, Y. Nakamura, N. Ogasawara, S. Kuhara, and K. Horikoshi.** 2000. Complete genome sequence of the alkaliphilic bacterium Bacillus halodurans and genomic sequence comparison with Bacillus subtilis. Nucleic Acids Res. **28:**4317–4331.
97. **Talay, S. R., A. Zock, M. Rohde, G. Molinari, M. Oggioni, G. Pozzi, C. A. Guzman, and G. S. Chhatwal.** 2000. Co-operative binding of human fibronectin to Sfbl protein triggers streptococcal invasion into respiratory epithelial cells. Cell. Microbiol. **2:**521–535.
98. **Talkington, D. F., B. G. Brown, J. A. Tharpe, A. Koenig, and H. Russell.** 1996. Protection of mice against fatal pneumococcal challenge by immunization with pneumococcal surface adhesin A (PsaA). Microb. Pathog. **21:**17–22.
99. **Tatusov, R. L., M. Y. Galperin, D. A. Natale, and E. V. Koonin.** 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. **28:**33–36.
100. **Tatusov, R. L., D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin.** 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. **29:**22–28.
101. **Tenner, A. J.** 1999. Membrane receptors for soluble defense collagens. Curr. Opin. Immunol. **11:**34–41.
102. **Terao, Y., S. Kawabata, E. Kunitomo, J. Murakami, I. Nakagawa, and S. Hamada.** 2001. Fba, a novel fibronectin-binding protein from Streptococcus pyogenes, promotes bacterial entry into epithelial cells, and the fba gene is positively transcribed under the Mga regulator. Mol. Microbiol. **42:**75–86.
103. **Walker, D. R., and E. V. Koonin.** 1997. SEALS: a system for easy analysis of lots of sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. **5:**333–339.
104. **Wann, E. R., S. Gurusiddappa, and M. Hook.** 2000. The fibronectin-binding MSCRAMM FnbpA of Staphylococcus aureus is a bifunctional protein that also binds to fibrinogen. J. Biol. Chem. **275:**13863–13871.
105. **Welkos, S. L., and A. M. Friedlander.** 1988. Comparative safety and efficacy against Bacillus anthracis of protective antigen and live vaccines in mice. Microb. Pathog. **5:**127–139.
106. **Whatmore, A. M.** 2001. Streptococcus pyogenes sclB encodes a putative hypervariable surface protein with a collagen-like repetitive structure. Microbiology **147:**419–429.
107. **Wheeler, D. L., D. M. Church, A. E. Lash, D. D. Leipe, T. L. Madden, J. U. Pontius, G. D. Schuler, L. M. Schriml, T. A. Tatusova, L. Wagner, and B. A. Rapp.** 2001. Database resources of the National Center for Bio/Technology Information. Nucleic Acids Res. **29:**11–16.
108. **Wizemann, T. M., J. E. Adamou, and S. Langermann.** 1999. Adhesins as targets for vaccine development. Emerg. Infect. Dis. **5:**395–403.
109. **Wizemann, T. M., J. H. Heinrichs, J. E. Adamou, A. L. Erwin, C. Kunsch, G. H. Choi, S. C. Barash, C. A. Rosen, H. R. Masure, E. Tuomanen, A. Gayle, Y. A. Brewah, W. Walsh, P. Barren, R. Lathigra, M. Hanson, S. Langermann, S. Johnson, and S. Koenig.** 2001. Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. Infect. Immun. **69:**1593–1598.
110. **Wootton, J. C., and S. Federhen.** 1996. Analysis of compositionally biased regions in sequence databases. Methods Enzymol. **266:**554–571.

Editor: D. L. Burns