PROTEINS: Structure, Function, and Bioinformatics 60:787–796 (2005)

Structural Analysis of a Set of Proteins Resulting From a Bacterial Genomics Project

J. Badger,* J.M. Sauder, J.M. Adams, S. Antonysamy, K. Bain, M.G. Bergseid, S.G. Buchanan, M.D. Buchanan, Y. Batiyenko, J.A. Christopher, S. Emtage, A. Eroshkina, I. Feil, E.B. Furlong, K.S. Gajiwala, X. Gao, D. He, J. Hendle, A. Huber, K. Hoda, P. Kearins, C. Kissinger, B. Laubert, H.A. Lewis, J. Lin, K. Loomis, D. Lorimer, G. Louie, M. Maletic, C.D. Marsh, I. Miller, J. Molinari, H.J. Muller-Dieckmann, J.M. Newman, B.W. Noland, B. Pagarigan, F. Park, T.S. Peat, K.W. Post, S. Radojicic, A. Ramos, R. Romero, M.E. Rutter, W.E. Sanderson, K.D. Schwinn, J. Tresser, J. Winhoven, T.A. Wright, L. Wu, J. Xu, and T.J.R. Harris

Structural GenomiX Inc., San Diego, California

ABSTRACT **The targets of the Structural GenomiX (SGX) bacterial genomics project were proteins conserved in multiple prokaryotic organisms with no obvious sequence homolog in the Protein Data Bank of known structures. The outcome of this work was 80 structures, covering 60 unique sequences and 49 different genes. Experimental phase determination from proteins incorporating Se-Met was carried out for 45 structures with most of the remainder solved by molecular replacement using members of the experimentally phased set as search models. An automated tool was developed to deposit these structures in the Protein Data Bank, along with the associated X-ray diffraction data (including refined experimental phases) and experimentally confirmed sequences. BLAST comparisons of the SGX structures with structures that had appeared in the Protein Data Bank over the intervening 3.5 years since the SGX target list had been compiled identified homologs for 49 of the 60 unique sequences represented by the SGX structures. This result indicates that, for bacterial structures that are relatively easy to express, purify, and crystallize, the structural coverage of gene space is proceeding rapidly. More distant sequence-structure relationships between the SGX and PDB structures were investigated using PDB-BLAST and Combinatorial Extension (CE). Only one structure, SufD, has a truly unique topology compared to all folds in the PDB. Proteins 2005;60:787–796.** © 2005 Wiley-Liss, Inc.

Key words: X-ray crystallography; novel fold; protein knots; Protein Data Bank

INTRODUCTION

In the year 2000, a bacterial structural genomics project was initiated at Structural GenomiX Inc. (SGX) to determine the structures of a set of novel bacterial proteins (i.e., proteins with no sequence homolog to structures available from the Protein Data Bank (PDB)[1] at that time). The selected proteins were potential anti-infective drug targets that had either been shown to be essential for bacterial growth or were highly conserved among numerous species. A considerable proportion of the early effort lay in establishing the laboratory, computational and procedural infrastructure required for high throughput protein crystal structure determination and analysis. The first structures were completed in December 2000 and the program ended in mid-2002, with structure determinations from most remaining diffraction data sets completed by September 2002. A total of 80 structures, covering 60 different sequences, were determined in this project. If inter-species sequence variations are discounted, structures corresponding to 49 different gene names were determined (Table I). All 80 structures, together with the associated diffraction data and (where available) experimentally determined phases, were subsequently submitted to the Protein Data Bank for public dissemination. Working procedures for this project were generally aimed at maximizing the number of novel structures. However, closely related structures were solved opportunistically if, for example, diffraction quality crystals in multiple space groups arose during early crystallization trials or crystallization trials across multiple orthologs yielded crystals for more than one protein. In a few instances, suitable molecular replacement models became available in the Protein Data Bank during the course of this project. Structures were determined if data were recorded to better than 3-Å resolution with adequate experimental phasing for either manual or automated map interpretation. The two significant exceptions to this structural genomics pipeline approach, where a more focused effort was made to obtain additional orthologs or cocrystals, were a set of six LuxS structures[2] and a set of three ArnB Aminotransferase[3] structures.

The aim of this paper is to document the set of structures now available in the public domain as a result of this project. The systematic structure validation procedures and automated annotation methods developed at SGX to streamline Protein Data Bank depositions are also de-

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat

*Correspondence to: John Badger, Active Sight and Molecular Images, 4045 Sorrento Valley Blvd., San Diego, CA 92121. E-mail: jbadger@active-sight.com

Received 15 October 2004; Revised 24 January 2005; Accepted 28 January 2005

Published online 14 July 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20541

TABLE I. Catalog of SGX Bacterial Structures Deposited with the Protein Data Bank and Listed According to GenBank Gene Names and Accession Numbers[†]

Gene	GB Acc No.	PDB code	Resolution (Å)
alkH	NP_438220	1VHC	1.9
aroE	NP_416207	1VI2	2.1
aroK	NP_281577	1VIA	1.6
ArnB	AAM92146	(1MDO, 1MDX, 1MDZ)	1.7
coaD	NP_389385	1O6B	2.2
comA	NP_415129	(1VH5, 1VI8), 1VH9	1.3
cutE	NP_228862	1VHF	1.5
dapE	NP_27453	1VGY	1.9
dph5	NP_069217	1VHV	1.8
elbB	NP_417676	1VHQ	1.7
fliS	NP_391413	1VH6	2.5
frwX	NP_228854	1VHO	1.9
gbsB	NP_228728	1VHD	1.6
his1	NP_228848	1O63, 1O64	2.0
his6	NP_228842	1VH7	1.9
kdsA	NP_439706	1O60	1.8
kdsB	NP_415438	1VH1, (1VH3, 1V1C)	1.8
kimE	NP_248080	1VIS	2.7
luxS	NP_296108	(1INN, 1VJE, 1J6V, 1VH2, 1VGX), 1J6W, 1J6X	1.6
nudE	NP_417856	(1VHG, 1VHZ)	2.3
panB	NP_273911	(1O66, 1O68)	1.8
pdxY	NP_416153	1VI9	2.0
pepT	NP_415645	1VIX	2.5
plsX	NP_389471	1VI1	3.0
rsuA	NP_439399	1VIO	1.6
sufD	NP_416196	1VH4	1.8
wlaK	AAD09304	(1O61, 1O62, 1O69)	1.8
yacE	NP_285799	(1VHL, 1VHT, 1VIY)	1.6
Dus	NP_227912	1VHN	1.6
ybhB	NP_415294	1VI3	1.8
yckF	NP_388227	1VIV, 1VIM	1.4
yer0	NP_390733	1VI0	1.7
yerB	NP_416346	1VHM	2.1
yffH	NP_416962	1VIU, 1VIQ	2.4
ygbP	NP_417227	(1VGT, 1VGU), (1VGW, 1VGZ)	1.8
ygbB	NP_438831	(1VH8, 1VHA)	2.4
yigZ	NP_418290	1VI7	2.8
YiiM	NP_418346	(1O65, 1O67)	2.3
ysdC	NP_390760	1VHE	1.9
yqeU	NP_390442	1VHK, 1VHY	1.9
yqgF	NP_390617	1VHX	2.0
yvyH	NP_391446	1O6C, 1VGV	2.3
yydA	NP_370548	1VH0, 1O6D	1.7
ywnH	NP_391537	1VHS	1.8
pcrB	NP_388542	1VIZ	1.9
rraA1	NP_231996	1VI4	1.9
deoD	NP_231977	1VHJ, 1VHW	1.5
rps2p	NP_069962	(1VI5, 1VI6)	2.0
AF1521	NP_070350	1VHU	1.3

[†]The Protein Data Bank identification codes contained within parenthesis have identical sequences. GenBank gene names could not be assigned for the final five protein sequences. Where there are multiple matching structures, the highest resolution is quoted.

scribed. Sequence comparisons of the SGX structures with other structures deposited in the Protein Data Bank during the course of the project provide an indication of the rate at which the structural coverage of unique genes for bacterial structures is being extended. Analysis of the sequence and structural homologies between the SGX structures and other structures in the Protein Data Bank structures provides examples where structure comparisons strengthen existing knowledge of protein functional roles and indicate relationships that were previously unknown or considered tentative.

MATERIALS AND METHODS

Structure Determination

The set of proteins that are the material for the analysis reported in this paper were all cloned and expressed in *Escherichia coli*. Standard procedures for protein expression and purification were as follows: 1–2-L *E. coli* cultures of the C-terminally hexa-his-tagged target proteins were expressed in ZYP5052 medium, induced at OD600 0.6–0.8, and grown overnight at room temperature. The cell pellets were resuspended in 50 mM Tris-HCl (pH 7.5), 20 mM Imidazole, 0.1% Tween 20, 500 mM NaCl, and sonicated. The clarified supernatant was then loaded onto a 5-ml affinity Nickel column (Qiagen), washed in 50mM Tris HCl (pH 7.8), 500 mM NaCl, 10% Glycerol, 10 mM Imidazole,10 mM Methionine, and eluted with 50 mM Tris HCl (pH 7.8), 500 mM NaCl, 10% Glycerol, 500 mM Imidazole, 10 mM Methionine. The eluate was run on a Superdex 200 column (Pharmacia) in 10 mM Hepes (pH 7.5), 150 mM NaCl, 10 mM Methionine, 10% Glycerol, 5 mM DTT. Finally, the elution peak fractions were combined and concentrated to at least 10 mg/ml.

A production approach involving parallel expression and purification across several species per gene was used in this project, with a cessation of effort once a representative structure for that gene had been solved. Overall (including duplicate orders), 2069 clones were made available for small scale expression and solubility testing and passed to fermentation, 1752 fermentations were passed to purification and 937 purifications were passed to crystallization. For the subset of genes for which structures were eventually obtained, 301 clones were constructed. Crystals were grown by hanging drop vapor diffusion methods with conditions obtained from a variety of commercial and internally developed screens. Native protein was used for crystal screening and optimization. Crystallizations with protein incorporating Se-Met were undertaken only after adequate growth conditions had been demonstrated by the observation of diffraction patterns extending to beyond 3 Å in native crystals. All crystals were frozen prior to data collection.

Almost all diffraction data were collected at the COMCAT beam line (sector 32ID) at the Advanced Photon Source during its commissioning period, with a typical utilization of $\sim$ 2 days/month. Initially, structures were solved using a multi-wavelength MAD data phasing methodology where, in order to preserve the crystal over the collection of the 3–4 required data sets, it was sometimes

only possible to collect minimal data at each wavelength. The standard data collection protocol then shifted towards the measurement of highly redundant and complete data sets (usually via 180° rotations of the crystal) at the Se edge for SAD phasing, followed by a second data set with wavelength adjusted to the high-energy remote position if the crystal retained any useful diffraction. A total of 45 structures were determined by experimental phasing from protein incorporating Se-Met, from SAD (20 examples), MAD (22 examples) or SIRAS (three examples) data. The remaining structures were solved by molecular replacement, usually from another structure within this set, and in a few cases from a Protein Data Bank structure that became available over the course of the project.

The SGX Structure Solution System was developed during the course of this project to provide a robust framework that allows structure determination tasks to be carried out through command-line operations and/or editable script wrappers. Within this framework, data integration is carried out by MOSFLM,[4] with subsequent merging and reduction steps performed by CCP4/SCALA,[5] and CCP4/TRUNCATE.[6] For the bacterial structure determinations that required experimental phase determination, the Se sites were determined with SnB[7] with subsequent site refinement performed by either CCP4/MLPHARE[8] or SHARP.[9] Following the initial Se-site determinations, 1–3 passes of site refinement were usually performed, with modifications of the Se site constellations to eliminate bogus sites and model any additional sites revealed by SAD residual difference maps. CCP4/MLPHARE was found to be a very rapid and effective program for structure determinations involving SAD data, since issues of nonisomorphism and unbiased utilization of multiple data sets are not present for this case. Density modification was usually performed with CCP4/SOLOMON[10] because postmortem evaluations for several early structures showed that, when the initial phase determinations were provided by SAD data, this program gave more accurately refined phases than CCP/DM[11] when run with default protocols. CCP4/DM was used for calculations applying noncrystallographic symmetry averaging. However, the majority of electron density maps were of sufficiently high resolution and quality that symmetry averaging was rarely considered desirable. If a data set was available in which the resolution extended to beyond 2.3 Å the initial model building was carried out using arp/wARP.[12] For structures determined by molecular replacement, the CCP4/MOLREP[13] and EPMR[14] programs were used to provide the initial model placement. The majority of structure refinements were performed using CCP4/REFMAC[15] with interactive model building using XtalView/Xfit.[16] Outside the Structure Solution System, some data sets were processed using DENZO/SCALEPACK[17] and refined using CNX.[18] Data processing and refinement statistics for the SGX structures are recorded in the Protein Data Bank coordinate files (see below) and provided as supplementary materials to this paper (Supplement 1). Using the Structure Solution System, many structures were experimentally phased and largely built by automated methods within 24 hours of data collection; several structures were also fully refined and uploaded into the SGX database within that time frame.

The average resolution for this set of 80 structures was 2.1 Å and the resolution was better than 2.3 Å for 55 structures. The average number of amino acids per crystal asymmetric unit was 575.6. Only 18 of the 80 structures contained a monomer in the crystal asymmetric unit, with two protein copies per asymmetric unit as the predominant crystal assembly, occurring in 37 of the 80 structures.

Structure Validation

Prior to deposition with the Protein Data Bank, the structures were validated using a set of automated checks built into an evolving in-house quality-control system and uploaded into a local database. The validation system provides a convenient mechanism for executing standard structure validation programs and parsing information from the resulting output files into more convenient lists of global quality scores and putative local errors.[19] R-factors were calculated using CCP4/REFMAC5[15] using the Babinet bulk solvent correction, with SFCHECK[20] providing supplementary analysis of the diffraction data. Percentages of amino acids lying in the core of the Ramachandran plot (A, B, and L areas[21]), counts of abnormally close protein contacts and counts of abnormal side chain rotamers were obtained with PROCHECK.[22] Data for the display of electron density maps was precomputed in convenient forms for use with the XtalView/Xfit[18] molecular graphics program.

Regression analysis of quality metrics for this set of structures gave the following suggested *lower* bounds for resolution (d) dependent global quality criteria

Maximum $R_{free} = -0.02d^2 + 0.13d + 0.11$

Maximum $R_{work} - R_{free} = -0.01d^2 + 0.065d - 0.02$

Minimum percentage of residues in Ramachandran core = $100 \times (-0.04d + 0.96)$

Maximum number of abnormal $\chi1 - \chi2$ angles/100 residues = $0.075d + 0.75$

Maximum number of short contacts/100 residues = $2.8571d - 5.5714$ d < 2.3 Å = 1.0 d > 2.3 Å

These bounds update our previously published calculation methods and values.[19]

Prior to transfer to the SGX database a crystallographer responsible for quality control reviewed all structures in the context of their associated electron density maps. Particular emphasis was placed on checking amino acids appearing in putative "error lists" to ensure that any detected abnormalities were justifiable. Amino acids appeared as probable errors if (1) density correlations for main or side chains in likelihood-weighted maps were less than 0.4, (2) main-chain torsion angles corresponded to disallowed regions of the Ramachandran plot or nonpropyl cis peptides, (3) side-chain $\chi1 - \chi2$ angles deviate significantly from expected rotamer values, (4) "flipping" of Asn, Gln, or His side chains resulted in improved H-bonding interactions, or (5) covalent bonds and angles were severely strained. Additional checks were also implemented

to detect any large volumes of electron density that were not accounted for by the atomic model and flag large features in final difference maps.

Structure Deposition to the Protein Data Bank

With the exception of seven structure entries described in earlier publications,[2,3] which were deposited to the RCSB Protein Data Bank using standard GUI-driven ADIT interface,[13] the deposition of the SGX structures was expedited by the development of a semi-automated command-line system. This software runs a set of operations to (1) parse data processing diagnostics from standard output files in the SGX structure repository, (2) calculate structure quality diagnostics, and (3) read additional gene/structure annotation required by the RCSB PDB from a standard file created from internal SGX information. This information and the associated atomic coordinates are gathered together and written to a special PDB deposition file developed in conjunction with staff at the RCSB Protein Data Bank. This deposition file employs mmCIF tags from the current mmCIF dictionary and the PDB/mmCIF data item correspondence dictionary.[23,24] Other than providing a simplification of the deposition process and labor reduction, the advantages of this system over manual data entry are that the deposition will usually contain more complete and accurate information. The mmCIF deposition file is compliant with current operating procedures at the Protein Data Bank (i.e., it includes all required items and can be parsed by procedures built in to the ADIT deposition tool). Examples of these deposition files, which might serve as templates for workers in other projects wishing to develop similar systems, are available upon e-mail request to jbadger@active-sight.com and are provided as supplementary material to this paper (Supplement 2).

GenBank[25] gene codes were provided to the Protein Data Bank for all sequences for which they could be determined. Experimental sequencing was carried out on protein samples for all structures to ensure that the cited sequences (i.e., those appearing in the SEQRES records of the final Protein Data Bank coordinate files) were correct. Discrepancies between sequences in the solved structure and the GenBank sequences are the result of cloning artifacts (N- and C-terminal tags), the product of protein engineering (usually Se-Met substitution to increase phasing power for SAD/MAD structure determination), naturally occurring mutants, or sequencing errors from genome sequencing projects. Based on reliable sequence annotations, 41 of the solved structures were classified by Enzyme Commission numbers extending to three or more digits (Table II). Descriptive protein names (contained in TITLE records in the resulting PDB files) were assigned to the structures where a classification was possible.

Sequence and Structure Comparison

Sequence comparisons for the 60 unique sequences represented by the structures determined at SGX with other structures in the Protein Data Bank were carried out in late August 2004 via BLAST searches[26] with E-value cutoffs of 0.001 (Table III). Although more sensitive sequence comparison methods are available, BLAST was used for this analysis because it employs a simple well-defined search algorithm and the intention was to detect convincing sequence matches, rather than weak sequence similarities with uncertain relevance.

Structure comparisons for the most novel structures (i.e., those structures for which no BLAST hit was obtained with this cutoff) were made with the CE algorithm[27] via the CE server at the San Diego Supercomputer Center (http://cl.sdsc.edu/ce.html). These calculations were run using default settings, which report structure matches for which Z-scores exceed 3.7 and cover all "representative structures" in the Protein Data Bank. For those structure comparisons in which the crystal asymmetric unit of the SGX structure contained multiple molecules, the A-chain molecule was used.

TABLE II. The 41 SGX Bacterial Structures for Which Enzyme Commission (EC) Numbers were Assigned Based on Gene Annotations†

Category	PDB code
1. Oxidoreductase	1VHD, 1VI2
2. Transferase	1O6B, 1VH1, (1VH3, 1VIC), (1VGT, 1VGU), 1VGW, 1VGZ, (1O66, 1O68), 1VHV, 1VIS, 1VI9, (1VHL, 1VHT, 1VIY), 1VIA, (1VHJ, 1VHW), (1O63, 1O64)
3. Hydrolase	(1INN, 1VJE, 1J6V, 1VH2, 1VGX), 1J6W, 1J6X, 1VHX, (1VHG, 1VHZ), 1VIQ, 1VIX, 1VIU
4. Lyase	(1VH8, 1VHA), 1VH7, 1VIO
5. Isomerase	1O6C, 1VGV
6. Ligase	No examples

†Structures corresponding to Protein Data Bank identification codes that are contained within parenthesis have identical sequences.

RESULTS AND DISCUSSION

Sequence and Structure Comparisons to Other Structures in the Protein Data Bank

At the time that the SGX bacterial structures target list was assembled (early 2000), there were no strong sequence homologies between proteins on the target list and structures already available through the Protein Data Bank—this was one of the criteria for inclusion in the target set. The majority of the SGX structures were deposited to the PDB in late Fall 2003 and the BLAST analysis of these sequences against all non-SGX structures deposited in the PDB was performed in August 2004. The results of these searches (Table III) show that of the 60 independent sequences, only 11 were not matched by any structure in the Protein Data Bank over these $\sim$ 4.5 years (and only one of these has a truly novel fold). Given that most of the SGX structure depositions were not made until late Fall 2003, few if any PDB structures from other groups are likely to have been determined using information from the SGX structures. These results indicate that, at least for targets relatively easy to purify, express and crystallize, avoiding duplication of effort in the publicly-funded structural genomics efforts is extremely important.[28,29]

TABLE III. BLAST Comparison of SGX Bacterial Structures with Structures Outside of This Set and Deposited with the Protein Data Bank Before August 26, 2004[†]

Gene	SGX/PDB code	Homolog PDB code	Percent id	Percent pos
alkH	1VHC	1FQO	37	61
aroE	1VI2	1O9B	99	99
aroK	1VIA	1KAG	31	54
ArnB	1MDO, 1MDX, 1MDZ	1B9I	31	48
coaD	1O6B	1OD6	50	69
comA	1VH5, 1VI8	1O0I	52	68
comA	1VH9	1O0I	48	66
cutE	1VHF	1O5J	99	99
dapE	1VGY	—	—	—
dph5	1VHV	1CBF	31	47
elbB	1VHQ	1OYI	97	97
fliS	1VH6	1ORY	24	48
frwX	1VHO	—	—	—
gbsB	1VHD	1O2D	99	99
his1	1O63, 1O64	1H3D	28	47
his6	1VH7	1THF	99	99
kdsA	1O60	1G6O	81	90
kdsB	1VH1	1H7T	45	60
kdsB	1VH3, 1VIC	1H7T	42	58
kimE	1VIS	1KKH	99	99
luxS	1INN, 1VJE, 1J6V, 1VH2, 1VGX	1JOE	48	69
luxS	1J6X	1JVI	47	63
luxS	1J6W	1JOE	100	100
nudE	1VHG, 1VHZ	—	—	—
panB	1O66, 1O68	1M3U	53	69
pdxY	1VI9	1LHR	30	47
pepT	1VIX	1FNO	92	96
plsX	1VI1	—	—	—
rsuA	1VIO	1KSV	57	74
sufD	1VH4	—	—	—
wlaK	1O61, 1O62, 1O69	1B9I	28	48
yacE	1VHL, 1VHT, 1VIY	1N3B	98	98
Dus	1VHN	—	—	—
ybhB	1VI3	1FJJ	98	99
yckF	1VIV	1M35	99	99
yckF	1VIM	1JEO	41	61
yer0	1VI0	1JUS	23	46
yerB	1VHM	1F5M	39	60
yffH	1VIU	1KHZ	28	50
yffH	1VIQ	1KHZ	99	99
ygbP	1VGT, 1VGU	1INJ	100	100
ygbP	1VGW	1H3M	42	57
ygbP	1VGZ	1H3M	44	59
ygbB	1VH8, 1VHA	1JN1	99	99
yigZ	1VI7	—	—	—
YiiM	1O65, 1O67	—	—	—
ysdC	1VHE	—	—	—
yqeU	1VHK	1NXZ	30	51
yqeU	1VHY	1NXZ	100	100
yqgF	1VHX	—	—	—
yvyH	1O6C	1F6D	55	68
yvyH	1VGV	1F6D	100	100
yydA	1VH0	1NS5	30	53
yydA	1O6D	1NS5	31	53
ywnH	1VHS	1UFH	40	56
pcrB	1VIZ	—	—	—
rraA1	1VI4	1Q5X	44	65
deoD	1VHJ, 1VHW	1K95	79	89
rps2p	1VI5, 1VI6	1PNX	25	43
AF1521	1VHU	1HJZ	100	100

[†]Sequentially distinct structures with the same GenBank gene names are listed as separate entries. BLAST searches were carried out using E-value cutoffs of 0.001. The Protein Data Bank identification code for the structure homolog that gave the best match is listed together with the percentage of identical residues and percentage of residues yielding a positive score.

TABLE IV. Comparison of Novel Structures as of August 2004 Resulting From the SGX Bacterial Genomics Project to Other Structures in the Protein Data Bank Using the CE algorithm.[27†]

SGX/PDB code	Homolog code	RMSD	Percent id	Aligned	Z-score
1VGY	1CG2:A	3.2	18	358/393	7.4
1VHG	1G0S:A	2.3	21	173/209	6.2
1VHN	2DOR:A	2.6	15	214/311	6.1
1VHE	1FT7:A	2.8	21	227/291	6.0
1VIZ	1PII:-	3.3	9	147/152	5.3
1VHO	1FT7:A	3.1	13	105/291	5.2
1VI1	1DR8:B	2.9	14	182/344	5.2
1VHX	1HJR:C	3.3	15	122/158	5.0
1VI7	1JQM:B	3.9	8	155/691	5.0
1VH4	1DAB:A	4.7	6	148/539	4.7
1O65	1PKY:C	2.1	9	74/470	3.7

†The SGX bacterial genomics project concluded September 2002.
The homolog code column contains the entry and chain identification for the structures in the PDB that showed the greatest similarity to the SGX structure, the RMSD column contains the root-mean- square deviation (Å) between overlapped CA positions, the percent id column shows the percentage sequence identity after structure-based alignment, the aligned column shows the fraction of aligned amino acids and the Z-score column contains the CE Z-score. Results are ordered by Z-score.

Four structures from two different families were of special interest because they contained deep trefoil knots,[30] both of which are classified at the fold and superfamily level in the SCOP database[31] as "alpha/beta knot." These structures were 1ybeA/yydA (1VH0, *Staphylococcus aureus*; 1O6D, *Thermatoga maritima*) and yggJ (1VHY, *Haemophilus influenzae*; 1VHK, *Bacillus subtilis*). Other PDB structures in this same SCOP superfamily, such as 1MXI and 1UAL contain bound ligands in the active sites that are found at the location of the knot.

Of the 11 unique structure-sequences in Table III for which BLAST sequence comparison revealed no strong homologies to the other structures already present in the PDB, structure comparisons were carried out versus representative structures in the Protein Data Bank using PDB-BLAST[32] and CE (http://cl.sdsc.edu/ce.html). PDB-BLAST uses PSI-BLAST to build a positive-specific score matrix (PSSM) for the target protein by searching the GenBank nonredundant sequence database, then using the PSSM to search a database of PDB sequences. The results of these searches (Table IV) show that 10/11 structures contain folds that are significantly similar to folds found in structures already in the PDB. Only SufD (1VH4) unambiguously has a novel fold.

Description of SGX Structures Without Strong BLAST Hits in the Protein Data Bank 1VGY (dapE)

The dapE protein from *Neisseria meningitidis* is required for diaminopimelate biosythesis, a critical component of cell wall and lysine biosynthesis. This gene encodes the protein succinyl diaminopimelate desuccinylase. Like carboxypeptidase G2 (1CG2; E.C. 3.4.17.11), which has a similar structure as detected by PSI-BLAST and CE, dapE [Fig. 1(A)] has a catalytic domain (residues 1–179 and 295–381) interrupted by a dimerization domain (180–294). By analogy to 1CG2, dapE residues His68, Asp101, Glu136, Glu164, and His350 are likely involved in binding two zinc atoms, although these were not observed in the electron density.

1VHG/1VHZ (nudE)

The nudE protein in *E. coli* is a nudix hydrolase family member active against ADP ribose, NADH, AP2A and AP3A[33] and is classified as a hydrolase (E.C. 3.6.1.−) based on previous gene annotations. The CE search with 1VHG [Fig. 1(B)] revealed structure similarity to 1G0S, a hypothetical 23.7-kDa protein in the Icc–Tolc intergenic region (ADP-ribose pyrophosphatase) and, with a somewhat lower score (Z = 5.6, RMSD = 3.4 Å, 14% sequence identity, 141/190 residues aligned), entry 1HZT, a isopentenyl diphosphate delta isomerase. The two SGX structures correspond to apo- and adenosine 5′-diphosphoribose (APR) bound forms of the protein. The crystal asymmetric unit contains a dimer in which the APR molecule is bound in a site with contacts from amino acids from both molecules.

1VHN (DUS)

This protein (*T. maritima* protein TM0096) is homologous to tRNA-dihydrouridine synthase (DUS; formerly called yacF in *B. subtilis*).[34] DUS homologs are well conserved among eubacteria but were previously without a known function. PDB-BLAST and CE searches revealed structure similarity to dihydroorotate dehydrogenase (HDOD) A & B (1EP1, 2DOR, E.C. 1.3.3.1). Our structure, 1VHN[35] [Fig. 1(C)], and the DHOD structures both contain a bound flavin molecule and function as oxidoreductases. The TIM-barrel fold (5–237) of DUS has an unusual C-terminal four helix bundle (238–309). This helical extension may have originated from an ancestral proteobacterial NtrC transcriptional regulatory protein,[36] allowing the protein to bind the dihydrouridine loop of tRNA.[34] Since *T. maritima* thrives in high temperature environments (~ 90°C), it is not surprising that DUS might have the capability to bind and reduce uridine to

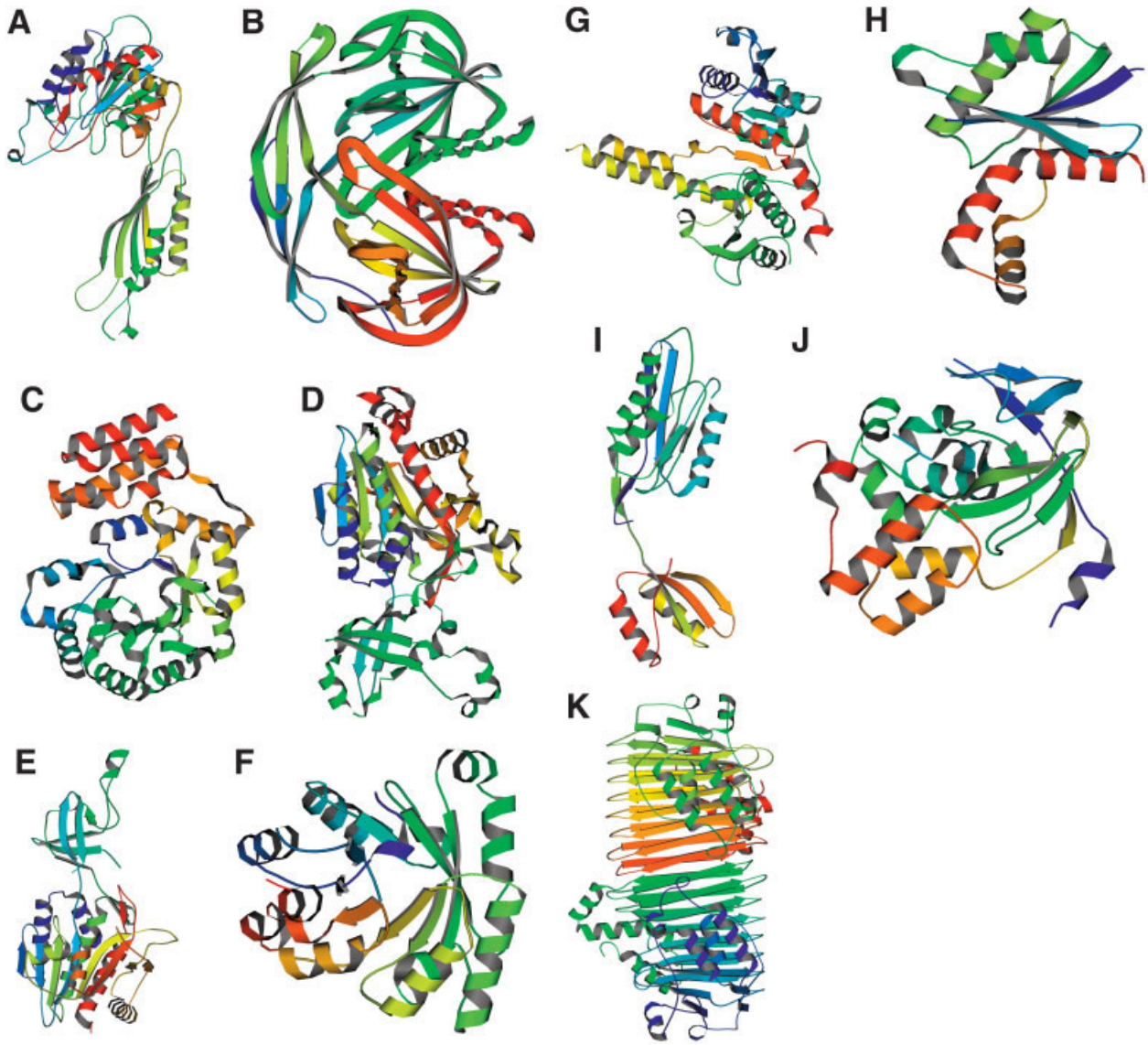

Fig. 1. Ribbon diagrams[54] of the eleven structures described in the Results and Discussion section: (**A**) monomer from the dapE structure (1VGY), (**B**) homodimer from the nudE structure (1VHG), (**C**) monomer from the DUS structure (1VHN), (**D**) monomer from the ysdC structure, 1VHE, (**E**) monomer from the frwX structure, 1VHO, (**F**) monomer from the perB structure (1VIZ), (**G**) monomer from the plsX structure (1VI1), (**H**) monomer from the yqgF structure (1VHX), (**I**) monomer from the yigZ structure (1VI7), (**J**) monomer from the YiiM structure (1O65), (**K**) the novel sufD structure (1VH4) with the homodimer interface in the center.

5,6-dihydrouridine, an adaptation that stabilizes RNA at high temperatures.[37]

1VHE (ysdC)

YsdC from *B. subtilis* is a putative deblocking aminopeptidase from the M42 family. This gene is conserved in a number of thermophiles, archaea and pathogenic bacterial species. Only one metal cation was seen bound in the active site, defined by residues H68, D182, E214, E215, D237, and H325; a second cation was not observed but two divalent metal cations are probably required for activity. It was modeled as zinc in the structure, but the anomalous signal suggests that it is probably not zinc. Mutation of the aspartic or glutamic acid residues has been shown to have an adverse effect on the function of an aminopeptidase from *Pyrococcus horikoshii*,[38] which requires two cobalt cations for activity. An unusual cis-peptide bond is found between D182 and N183, highlighting its role at the active site. There is one ysdC molecule per asymmetric unit in the crystal, but the protein forms a dimer with a symmetry related molecule, burying 2700 Å^2 in surface area, predominantly at the smaller dimerization subdomain. PDB-BLAST and CE searches with 1VHE [Fig. 1(D)] showed structure similarity to 1CG2 (carboxypeptidase G2) and 1FT7 (leucyl aminopeptidase), and to other SGX structures, including 1VHO (frwX; 34% identity), 1VGY (dapE; 15% identity), and 1VIX (pepT; 15% identity).

1VHO (frwX)

The closest structural homolog to *T. maritima* frwX (TM1048) [Fig. 1(E)] is 1VHE (ysdC; 34% identity), described above. The closest homologs of frwX in GenBank are annotated as either cellulases or endoglucanases; the enzyme is probably involved in polysaccharide biosynthesis or degradation.

1VIZ (pcrB)

PcrB is a TIM-barrel [Fig. 1(F)] of unknown function. PDB-BLAST detects similarity (~ 14% identity) to 1GEQ (tryptophan synthase; E.C. 4.2.1.20) and 1TQJ (ribulose-phosphate-3-epimerase; E.C. 5.1.3.1) As expected, the structural matches found by CE include TIM-barrels with a wide variety of activities, representing at least the first five enzyme classification categories. As the functions of proteins with TIM-barrel folds are so diverse, pcrB will have to await biochemical analysis to elucidate its function.

1VI1 (plsX)

The genes encoding several essential enzymes involved in fatty acid biosynthesis are clustered in *B. subtilis* in the order plsX-fabD-fabG-acpP.[39] We predict that plsX [Fig. 1(G)] is a glycerol 3-phosphate acyltransferase and catalyzes the first step in the biosynthesis of phospholipids, the attachment of a fatty-acid chain to a hydroxyl group of glycerol 3-phosphate (similar to plsB[40]). *E. coli* contains an additional gene, fabH, following plsX. The lack of fabH in *B. subtilis* explains the unusual amino acid composition of plsX in *B. subtilis* compared to *E. coli*.[39,41] PDB-BLAST identified similarity (~ 15% identity) with two phosphotransacetylases (1R5J; 1QZT, E.C. 2.3.1.8). An orthologous structure of plsX from *Enterococcus faecalis* (1U7N) was deposited in the PDB during preparation of this manuscript. The *E. faecalis* and *B. subtilis* proteins share 50% sequence identity.

1VHX (yqgF)

YqgF (YrrK in *B. subtilis*) [Fig. 1(H)] is conserved in bacterial pathogens and is an essential protein in *E. coli*[42] and *H. influenzae*.[43] The protein likely acts as a Holliday junction resolvase during DNA recombination.[44] A CE search using 1VHX revealed structural similarity (3.5 Å; 14% identity) to 1HJR (RuvC resolvase), a Holliday junction-specific endonuclease (E.C. 3.1.22.4). BLAST easily identifies the orthologous yqgF structures from *E. coli* (1OVQ, 1NMN, 1NU0; 32% identical to 1VHX).

1VI7 (yigZ)

YigZ[45] is a conserved protein of unknown function from *E. coli*. No significant structure similarity was found for 1VI7 [Fig. 1(I)] by the CE search reported here or in an earlier study[46] using Dali,[47] although there are weak similarities to several of the ribosomal proteins, with the CE search giving 1JQM, ribosomal protein L11 as the strongest match. PDB-BLAST detects weak similarity (15% identity) to residues 698–792 of *S. cerevisiae* translation elongation factor 2 (eEF2; 1N0V),[48] an ADP-ribosylated ribosomal translocase. Structural alignment of this second subdomain gives an RMSD of 1.6 Å and 17% sequence identity. Alignment of the first domain (3-138) of yigZ with residues 562–726 of 1N0V gives an RMSD of 3.4 Å (with essentially random sequence identity (7.5%).

1O65/1O67 (YiiM)

These two crystal structures of yiiM [Fig. 1(J)] differ in their exact cell dimensions and in that Se-Met is incorporated in the protein in 1O67. YiiM is a conserved *E. coli* protein of unknown function. PDB-BLAST detects homology to 1ORU (*B. subtilis* yuaD) and the CE alignment has an RMSD of 2.47 Å (17% identical). The protein contains a MOSC domain, which mediates sulfur transport using a strictly conserved cysteine residue to be used in the biosynthesis of metal-sulfur clusters.[49] The structure of YiiM has an electropositive cleft that likely binds a positively charged substrate; the active site residues are predicted to be H60, E96, N97, R127, and C130.

1VH4 (sufD)

SufD is part of the SufABCDSE operon, which is involved in [Fe-S] cluster assembly. The SufBCD protein complex is involved in iron acquisition,[50] and it acts synergistically with SufE (1MZG) in modulating the cysteine desulfurase activity of SufS.[51] The exact role of SufB and SufD is unknown, but they share almost 20% sequence identity and likely share a similar fold and function. The novel structure of SufD is a flattened right-handed beta-helix of nine turns with two strands per turn; the N- and C-termini form helical subdomains [Fig 1(K)]. Homodimerization of SufD doubles the length of the beta-helix (to ~ 80 Å); two highly conserved residues, P347 and H360, interact at the dimer interface (the H360 NE2 atoms from each molecule are 3.3 Å apart). There are several highly conserved residues in the C-terminal subdomain (Y374, R378, G379, A385, F393), but their role is unknown; all the residues mentioned are conserved in SufB, further supporting the hypothesis that it has a very similar function and is able to homodimerize in a similar manner to SufD. It is possible that in vivo SufB and SufD form a functional heterodimer analogous to the SufD homodimer.

CONCLUSION

Once the SGX structure determination platform was developed, several *new* structures were solved each month based on ~ 2 days of Se-Met crystal data collection. Post-mortem tests on the set of experimentally phased structures showed that the currently available automated model-building programs would build ~ 90% of the main chain traces when experimental phasing data was available and the resolution of the data extended to better than 2.3 Å.[52] This result implies that, if bottlenecks and costs involved in preparing protein crystals incorporating Se-Met can be overcome, the majority of structure determinations will not be rate-limited by the need to trace and fit density maps ab initio. Based on results achieved in this project, we would anticipate that it should be possible for

an adequately funded and organized structural genomics project to solve several hundred structures per year.

Once established, the SGX Structure Solution System provided an environment in which 1–2 individuals could, within ~ 24 hours, process, phase, and (where applicable) auto-build structure models into all useful data sets resulting from a 2-day data collection trip. A side benefit of an early conversion from a 3–4 wavelength MAD phasing methodology to a SAD/2-wavelength MAD phasing methodology was that this greatly reduced and simplified the number of possible structure determination pathways. SIRAS phasing (i.e., from combination of a Se-Met and a native data set) was only found effective in three cases, presumably because the effect of nonisomorphism between crystals often outweighs the signal obtained from the S–Se exchange. The structure finalization process for many proteins was inhibited by the presence of poorly ordered loop densities, as modeling these portions with the available interactive model-building programs is a relatively slow and uncertain process. In addition, electron density maps for several structures contained "mystery densities," relatively large and potentially important endogenous cofactors or ligands that had been carried through the purification, and the identification of these entities was not always immediate. The development of a complete LIMS system, capable of tracking *and linking* all steps in the structure determination process, from purification to structure annotation required a major development effort as well as some practical experience, and was not fully completed prior to the conclusion of the bacterial genomics project. Nevertheless, convenient access to data on crystallization conditions and the functional background of the protein is potentially useful as it provides a context for reliable density map interpretation. Beyond the structure determination process, the task of providing structure–function annotation and background material at the level of a typical journal publication article appears to be unavoidably time-consuming and is a potential cause of delay in exposing structure results.

The history of this project suggests that the gene space of conserved bacterial proteins amenable to rapid structure determination is quickly being filled out with structural data. For this reason, up-to-date information on structure determination progress must be maintained on publicly accessible target lists to avoid duplicated effort in the publicly funded structural genomics initiatives. Clearly, the use of homologous structures to provide a structure determination route through the molecular replacement method will increasingly eliminate the experimental costs of phase determination with anomalous scattering and isomorphous replacement methods.

In several of the structure examples resulting from this project, the family relations and functional role of the new structure was only fully revealed by three-dimensional comparisons[53] to other previously solved structures. For this reason, genes with putative annotations are particularly good targets for structural genomics projects since it may often be possible to quickly obtain new functional information from structure analysis.

ACKNOWLEDGMENTS

We thank John Westbrook and Kyle Burkhardt of the RCSB PDB for their time, help and patience in developing the structure deposition format. The CE resource at SDSC (http://cl.sdsc.edu/ce.html) was used for the structure comparison of the novel structures with the PDB. Use of the Advanced Photon Source was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. W-31-109-Eng-38. This project was entirely funded by Structural GenomiX, Inc.

REFERENCES

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
2. Lewis HA, Furlong EB, Laubert B, Eroshkina GA, Batiyenko Y, Adams JM, Bergseid MG, Marsh CD, Peat TS, Sanderson WE, et al. A structural genomics approach to the study of quorum sensing: crystal structures of three LuxS orthologs. Structure (Camb) 2001;9:527–537.
3. Noland BW, Newman JM, Hendle J, Badger J, Christopher JA, Tresser J, Buchanan MD, Wright TA, Rutter ME, Sanderson WE, et al. Structural studies of Salmonella typhimurium ArnB (PmrH) aminotransferase: a 4-amino-4-deoxy-L-arabinose lipopolysaccharide-modifying enzyme. Structure (Camb) 2002;10:1569–1580.
4. Powell HR. The Rossmann Fourier autoindexing algorithm in MOSFLM. Acta Crystallogr D Biol Crystallogr 1999;55:1690–1695.
5. Collaborative Computational Project, Number 4. The CCP4 Suite: programs for protein crystallography. Acta Crystallogr D Biol Crystallogr 1994;50:760–763.
6. French S, Wilson K. On the treatment of negative intensity observations. Acta Crystallogr A 1978;34:517–525.
7. Weeks CM, Miller R. Optimizing Shake-and-Bake for proteins. Acta Crystallogr D Biol Crystallogr 1999;55:492–500.
8. Otwinowski Z. Maximum likelihood refinement of heavy atom parameters. In: Wolf W, Evans PR, Leslie AGW, editors. A conference proceedings: isomorphous replacement and anomalous scattering. Warrington, IK: Daresbury Laboratory; 1991. p 80–86.
9. de la Fortelle E, Bricogne G. Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. Methods Enzymol 1997;276:472–494.
10. Abrahams JP, Leslie AGW. Methods used in the structure determination of bovine mitochondrial F_1 ATPase. Acta Crystallogr D Biol Crystallogr 1996;52:30–42.
11. Cowtan K. "DM": an automated procedure for phase improvement by density modification. Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography 1994;31:34–38.
12. Perrakis A, Morris RJ, Lamzin VS. Automated protein model building combined with iterative structure refinement. Nat Struct Biol 1999;6:458–463.
13. Vagin A, Teplyakov A. MOLREP: an Automated Program for Molecular Replacement. J Appl Cryst 1997;30:1022–1025.
14. Kissinger CR, Gehlhaar DK, Fogel DB. Rapid automated molecular replacement by evolutionary search. Acta Crystallogr D Biol Crystallogr 1999;55:484–491.
15. Murshudov GN. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 1997;53:240–255.
16. McRee DE. XtalView/Xfit—A versatile program for manipulating atomic coordinates and electron density. J Struct Biol 1999;125:156–165.
17. Otwinowski Z, Minor W. Processing of x-ray diffraction data collected in oscillation mode. Methods Enzymol 1997;276:307–326.
18. Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszeweski J, Nilges M, Pannu NS, et al. Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 1998;54:905–921.
19. Badger J, Hendle J. Reliable quality-control methods for protein

crystal structures. Acta Crystallogr D Biol Crystallogr 2002;58: 284–291.
20. Vaguine AA, Richelle J, Wodak SJ. SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Crystallogr D Biol Crystallogr 1999;55:191–205.
21. Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. Stereochemical quality of protein structure coordinates. Proteins 1992; 12:345–364.
22. Westbrook J, Feng Z, Burkhardt K, Berman HM. Validation of protein structures for protein data bank. Methods Enzymol 2003; 374:370–385.
23. Westbrook JD, Fitzgerald PM. The PDB format, mmCIF, and other data formats. Methods Biochem Anal 2003;44:161–179.
24. mmCIF dictionary. http://pdb.rutgers.edu/mmcif/dictionaries/mmcif_pdbx.dic/Index/index.html
25. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank: update. Nucleic Acids Res 2004;32 Database issue:D23–26.
26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
27. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
28. Berman HM, Westbrook JD. The impact of structural genomics on the protein data bank. Am J Pharmacogenomics 2004;4:247–252.
29. Chen L, Oughtred R, Berman HM, Westbrook J. TargetDB: a target registration database for structural genomics projects. Bioinformatics 2004;20:2860–2862.
30. Taylor WR. A deeply knotted protein structure and how it might fold. Nature 2000;406:916–919.
31. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
32. Sauder JM, Arthur JW, Dunbrack RL Jr. Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins 2000;40:6–22.
33. Gabelli SB, Bianchet MA, Bessman MJ, Amzel LM. The structure of ADP-ribose pyrophosphatase reveals the structural basis for the versatility of the Nudix family. Nat Struct Biol 2001;8:467–472.
34. Bishop AC, Xu J, Johnson RC, Schimmel P, de Crecy-Lagard V. Identification of the tRNA-dihydrouridine synthase family. J Biol Chem 2002;277:25090–25095.
35. Park F, Gajiwala K, Noland B, Wu L, He D, Molinari J, Loomis K, Pagarigan B, Kearins P, Christopher J, et al. The 1.59-Å resolution crystal structure of TM0096, a flavin mononucleotide binding protein from Thermotoga maritima. Proteins 2004;55:772–774.
36. Morett E, Bork P. Evolution of new protein function: recombinational enhancer Fis originated by horizontal gene transfer from the transcriptional regulator NtrC. FEBS Lett 1998;433:108–112.
37. Dalluge JJ, Hamamoto T, Horikoshi K, Morita RY, Stetter KO, McCloskey JA. Posttranscriptional modification of tRNA in psychrophilic bacteria. J Bacteriol 1997;179:1918–1923.
38. Onoe S, Ando S, Ataka M, Ishikawa K. Active site of deblocking aminopeptidase from Pyrococcus horikoshii. Biochem Biophys Res Commun 2002;290:994–997.
39. Morbidoni HR, de Mendoza D, Cronan JE Jr. Bacillus subtilis acyl carrier protein is encoded in a cluster of lipid biosynthesis genes. J Bacteriol 1996;178:4794–4800.
40. Lewin TM, Wang P, Coleman RA. Analysis of amino acid motifs diagnostic for the sn-glycerol-3-phosphate acyltransferase reaction. Biochemistry 1999;38:5764–5771.
41. Zhang Y, Cronan JE Jr. Transcriptional analysis of essential genes of the Escherichia coli fatty acid biosynthesis gene cluster by functional replacement with the analogous Salmonella typhimurium gene cluster. J Bacteriol 1998;180:3295–3303.
42. Freiberg C, Wieland B, Spaltmann F, Ehlert K, Brotz H, Labischinski H. Identification of novel essential Escherichia coli genes conserved among pathogenic bacteria. J Mol Microbiol Biotechnol 2001;3:483–489.
43. Zalacain M, Biswas S, Ingraham KA, Ambrad J, Bryant A, Chalker AF, Iordanescu S, Fan J, Fan F, Lunsford RD, et al. A global approach to identify novel broad-spectrum antibacterial targets among proteins of unknown function. J Mol Microbiol Biotechnol 2003;6:109–126.
44. Aravind L, Makarova KS, Koonin EV. Survey and summary: holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. Nucleic Acids Res 2000;28:3417–3432.
45. Hagiwara Y, Hirai M, Nishiyama K, Kanazawa I, Ueda T, Sakaki Y, Ito T. Screening for imprinted genes by allelic message display: identification of a paternally expressed gene impact on mouse chromosome 18. Proc Natl Acad Sci USA 1997;94:9249–9254.
46. Park F, Gajiwala K, Eroshkina G, Furlong E, He D, Batiyenko Y, Romero R, Christopher J, Badger J, Hendle J, et al. Crystal structure of YIGZ, a conserved hypothetical protein from Escherichia coli k12 with a novel fold. Proteins 2004;55:775–777.
47. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
48. Jorgensen R, Ortiz PA, Carr-Schmid A, Nissen P, Kinzy TG, Andersen GR. Two crystal structures demonstrate large conformational changes in the eukaryotic ribosomal translocase. Nat Struct Biol 2003;10:379–385.
49. Anantharaman V, Aravind L. MOSC domains: ancient, predicted sulfur-carrier domains, present in diverse metal-sulfur cluster biosynthesis proteins including Molybdenum cofactor sulfurases. FEMS Microbiol Lett 2002;207:55–61.
50. Nachin L, Loiseau L, Expert D, Barras F. SufC: an unorthodox cytoplasmic ABC/ATPase required for [Fe-S] biogenesis under oxidative stress. Embo J 2003;22:427–437.
51. Outten FW, Wood MJ, Munoz FM, Storz G. The SufE protein and the SufBCD complex enhance SufS cysteine desulfurase activity as part of a sulfur transfer pathway for Fe-S cluster assembly in Escherichia coli. J Biol Chem 2003;278:45713–45719.
52. Badger J. An evaluation of automated model-building procedures for protein crystallography. Acta Crystallogr D Biol Crystallogr 2003;59:823–827.
53. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–603.
54. Kraulis PJ. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 1991;24: 946–950.