PROTEINS: Structure, Function, and Genetics 50:392–399 (2003)

Strategies for Structural Proteomics of Prokaryotes: Quantifying the Advantages of Studying Orthologous Proteins and of Using Both NMR and X-Ray Crystallography Approaches

Alexei Savchenko,[1] Adelinda Yee,[1] Anna Khachatryan,[1] Tatiana Skarina,[1] Elena Evdokimova,[1] Marina Pavlova,[1] Anthony Semesi,[1] Julian Northey,[1] Steven Beasley,[1] Ning Lan,[2] Rajdeep Das,[2] Mark Gerstein,[2] Cheryl H. Arrowmith,[1,4] and Aled M. Edwards[1,3,4]*

[1]*Ontario Center for Structural Proteomics, University Health Network, Toronto, Ontario, Canada*
[2]*Molecular Biophysics and Biochemistry Department, Yale University, New Haven, Connecticut*
[3]*Banting and Best Department of Medical Research, Toronto, Ontario, Canada*
[4]*Ontario Cancer Institute and Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada*

ABSTRACT **Only about half of non-membrane-bound proteins encoded by either bacterial or archaeal genomes are soluble when expressed in *Escherichia coli* (Yee et al., Proc Natl Acad Sci USA 2002;99:1825–1830; Christendat et al., Prog Biophys Mol Biol 200;73:339–345). This property limits genome-scale functional and structural proteomics studies, which depend on having a recombinant, soluble version of each protein. An emerging strategy to increase the probability of deriving a soluble derivative of a protein is to study different sequence homologues of the same protein, including representatives from thermophilic organisms, based on the assumption that the stability of these proteins will facilitate structural analysis. To estimate the relative merits of this strategy, we compared the recombinant expression, solubility, and suitability for structural analysis by NMR and/or X-ray crystallography for 68 pairs of homologous proteins from *E. coli* and *Thermotoga maritima*. A sample suitable for structural studies was obtained for 62 of the 68 pairs of homologs under standardized growth and purification procedures. Fourteen (eight *E. coli* and six *T. maritima* proteins) samples generated NMR spectra of a quality suitable for structure determination and 30 (14 *E. coli* and 16 *T. maritima* proteins) samples formed crystals. Only three (one *E. coli* and two *T. maritima* proteins) samples both crystallized and had excellent NMR properties. The conclusions from this work are: (1) The inclusion of even a single ortholog of a target protein increases the number of samples for structural studies almost twofold; (2) there was no clear advantage to the use of thermophilic proteins to generate samples for structural studies; and (3) for the small proteins analyzed here, the use of both NMR and crystallography approaches almost doubled the number of samples for structural studies. Proteins 2003;50:392–399.**

Key words: protein expression; structural proteomics; nuclear magnetic resonance spectroscopy; X-ray crystallography

INTRODUCTION

The field of proteomics seeks to determine the biochemical and cellular functions and structure of proteins on a genome-wide scale. Proteomics includes, among other things, studies of protein-protein interactions,[3–6] protein structure,[7–9] protein abundance,[10,11] post-translational modification,[12–14] and protein fossils.[15] The biochemical analysis of purified proteins is referred to as functional proteomics. Structural proteomics aims to derive the three-dimensional structures for all proteins. The consensus strategy for structural proteomics is to determine the experimental structures for enough proteins such that remaining structures can be predicted accurately using computational approaches.[16]

One of the most significant challenges facing experimentally based functional or structural proteomics is the reality that a large proportion of proteins are insoluble when expressed in heterologous systems or when purified and concentrated to the levels required for structural techniques.[1,17] Although there continue to be improvements in the techniques for generating soluble versions of an individual protein, such as variation of fusion tags,[18] genetic screens,[19] or using computational methods[17] to engineer soluble variants, these strategies are time-consuming and the success rate remains poor.

An alternative approach, and one that is afforded by the wealth of genome sequence information, is to express and purify a series of orthologs from various species for a given protein of interest. The underlying hypothesis is that subtle differences in the surface properties and/or stability

*Correspondence to: Aled Edwards, Banting and Best Department of Medical Research, 112 College St., Toronto, Ontario, Canada. E-mail: aled.edwards@utoronto.ca

Received 18 June 2002; Accepted 19 August 2002

of a protein, which might arise from the sequence variation among orthologs, may contribute to altered solubility and perhaps more suitable properties for NMR or X-ray structure determination. There is anecdotal evidence that this strategy may improve the probability of crystallization or NMR structure determination for individual proteins, but the extent to which this strategy improves the structure determination process remains unclear.

There is also a prevailing belief that orthologous proteins from thermophilic organisms are more amenable to structural biology methods, presumably because they are predicted to have fewer disordered regions and a higher proportion of salt bridges on the surface.[20,21] As such, thermophilic proteins may be more stable, may crystallize more readily, and may also be more amenable to NMR analysis.

In this study, 68 different pairs of homologous bacterial proteins from *Escherichia coli* and *Thermotoga maritima*, respectively, were cloned into bacterial expression vectors for studies of expression, purification and crystallization, and NMR. The results of this study support the notion that the use of a series of sequence homologues will increase the probability of obtaining a structural sample for at least one member of a given protein family. However, the data do not demonstrate a clear advantage for the use of proteins from a thermophilic organism. The results also highlight the complementarity of NMR and X-ray crystallography approaches.

MATERIALS AND METHODS

Target Selection

An all-versus-all implementation of BLAST was used to identify pairs of orthologous proteins in *E. coli (EC)* and *T. maritima (TM)*. Candidate protein sequences were scanned for homologues in the Protein Database using liberal thresholds of $E < 0.01$ and ID $\sim$ 30% or higher. ORFs that had significant BLAST hits to any other protein in either organism within these thresholds were excluded from the candidate list. Proteins with transmembrane regions were identified using a hidden Markov model algorithm and excluded as targets. Proteins harboring secretory signal sequences were identified using a neural network approach and also excluded.[22] The remaining sequences from the *E. coli* and *T. maritima* genomes were then compared for homology against each other. In this case, BLAST hits ($E < 0.0001$ and ID $\sim$ 20% and higher) were included in the candidate list, resulting in a list of pairs of proteins, one from each genome. We further selected protein pairs that had similar lengths and functional annotations and, therefore, are likely to have similar structures and functions. This selection produced 68 pairs of *E. coli* and *T. maritima* orthologous proteins (Table I) that were used for further studies. All but two of the 136 proteins were smaller than 320 residues.

Cloning

Target genes were amplified from genomic DNA using primers designed to create *Nde*I and *Bam*HI sites upstream from the initiation and stop codons, respectively. In cases where genes contained either of these restriction sites, they were replaced by *Ace*I for *Nde*I and *Bgl*II for *Bam*HI. The PCR of the target samples was performed in a 96-well format using Pfx Polymerase (Invitrogen, La Jolla, CA). The amplification products were cloned as previously described.[23]

PCR reactions were optimized for each set of ortholog genes based on the results for an initial set of 10 genes. The open reading frames were cloned into a modified pET15b T7 RNA polymerase-based expression vector (Invitrogen) that provided an N-terminal hexahistidine fusion (Fig. 1). The sequence encoding the thrombin cleavage site (LVPR $\downarrow$ GS) in the pET15b cloning vector was replaced by sequence encoding the cleavage site (ENLYFQ $\downarrow$ G) for the TEV protease for two reasons. First, the TEV protease is available in recombinant form (Invitrogen), and second, it is available in a histidine-tagged version (Science Reagents), which facilitates the removal of the enzyme (see Zhang et al.[23]). The 3′ end of the coding region was also modified. Two consecutive ochre stop codons (TAATAA) were introduced immediately downstream from the 3′ BamHI restriction site, and the stop codon was omitted from the PCR-amplified coding region. This strategy provides an advantage in that the same PCR fragment can be cloned, if necessary, in a different expression vector that appends a C-terminal hexahistidine fusion. The disadvantage of using the single PCR fragment is that two additional amino acids are added onto the C-terminus of the N-terminally-tagged recombinant proteins because of the addition of a BamHI site in the coding region of the 3′ primer (Fig. 1).

Protein Expression, Solubility, and Purification

Clones were transformed into *E. coli* BL21-Gold (DE3) (Stratagene, La Jolla, CA), which harbor an extra plasmid (pMgk) encoding three rare tRNAs (AGG and AGA for Arg, ATA for Ile).[24] Two to three colonies of each clone were grown in a 24-well format at 37°C in 3 ml Luria Broth supplemented with kanamycin and ampicillin (0.1 mg/ml each). The cultures were grown (37°C, 220 rpm) until an $OD_{600} \sim 0.6$ was reached. Protein expression was induced by the addition of 0.4 mM IPTG followed by overnight growth at 15°C or 30°C. Two 300-μl aliquots of the culture were transferred to separate 96-well plates and centrifuged to obtain cell pellets (20 min at 3,000 rpm; Beckman Coulter Allegra 6R centrifuge). The cell pellets of one plate were resuspended in denaturing buffer and kept as the whole cell fraction. The fractions of the other plate were flash-frozen in liquid N_2 and the soluble protein was extracted by the addition of 100 μl of BugBuster (Novagen, Madison, WI) followed by centrifugation (20 min at 3,000 rpm). The resulting supernatant, representing the soluble protein fraction, was compared against the whole cell fraction by denaturing gel electrophoresis in order to determine the size and the relative expression level of each protein.

Large-scale expression and purification was performed as described in Zhang et al.[23] for the proteins destined for crystallization samples and as described in Yee et al.[1] for the proteins destined for NMR spectroscopy.

TABLE I. Screening of 68 Orthologous Protein Pairs for Structural Studies[†]

	Annotation	Short description	Length (aa)	ID (%)	Expression	Solubility	Crystal trials	HSQC
	gi|1789047		61		5	5	tr	gd
1	gi|4980762	Carbon storage regulator	83	49	5	5	tr	gd
	gi|1788239		77		3	2	tr	pr
2	gi|4981522	Conserved hypothetical protein	79	45	5	5	tr	gd
	gi|1789741		95		5	1	no	pr
3	gi|4981518	Conserved hypothetical protein	87	36	5	5	tr	gd
	gi|1790614		102		5	5	tr	pr
4	gi|4981039	Growth-related protein	92	48	2	1	tr	pr
	gi|1788247		104		4	1	no	pr
5	gi|4981926	Flagellar complex protein	94	31	2	2	tr	pr
	gi|1790579		112		5	5	tr	gd
6	gi|4981598	Divalent cation tolerance protein	101	35	5	5	cl	gd
	gi|1787471		117		5	5	cl	pr
7	gi|4981520	Conserved hypothetical protein	118	32	2	1	no	pr
	gi|1787810		125		0 (4)	0 (0)	no	pr
8	gi|4981234	Transcriptional regulator, MarR family	143	33	5	1	no	pr
	gi|1789755		134		5	3	no	pr
9	gi|4981456	Conserved hypothetical protein	138	40	3	2	cl	pr
	gi|1788234		136		2	1	no	pr
10	gi|4981728	Regulator of flagellar protein expression	137	32	3	3	tr	pr
	gi|1790491		138		4	1	no	pr
11	gi|4982458	Conserved hypothetical protein	132	38	2 (3)	0 (0)	no	pr
	gi|1790320		145		2	1	no	pr
12	gi|4981256	Conserved hypothetical protein	149	52	2	1	no	pr
	gi|1786899		148		5	5	tr	gd
13	gi|4980614	Ferric uptake regulation protein	121	23	4	2	tr	pr
	gi|1790528		149		3	2	cl	pr
14	gi|4981625	Sugar-phosphate isomerase	143	46	4	2	cl	pr
	gi|1786615		149		4	4	cl	pr
15	gi|4982284	Conserved hypothetical protein	156	50	3	1	tr	pr
	gi|1789561		152		5	5	no	gd
16	gi|4982356	Conserved hypothetical protein	150	33	0 (0)	0 (0)	no	pr
	gi|1786880		155		5	5	tr	gd
17	gi|4982075	Conserved hypothetical protein	150	39	5 (4)	0 (2)	cl	gd
	gi|1789103		159		5	5	cl	pr
18	gi|4981169	Conserved hypothetical protein	165	41	5	4	tr	pr
	gi|1789849		162		5	1	tr	pr
19	gi|4981094	Conserved hypothetical protein	179	35	4 (3)	0 (0)	no	pr
	gi|1786266		163		5	1	tr	pr
20	gi|4981064	Acetolactate synthase subunit	171	37	5 (3)	0 (2)	no	pr
	gi|1788621		166		5	3	tr	gd
21	gi|4981988	Hydrogenase subunit	164	31	3	2	tr	pr
	gi|1788196		167		4	3	tr	gd
22	gi|4981243	Purine-binding chemotaxis protein	152	36	1	1	tr	gd
	gi|1786341		179		5	1	cl	gd
23	gi|4982445	Conserved hypothetical protein	187	32	4	1	cl	pr
	gi|1788638		184		3	1	cl	
24	gi|4981646	Conserved hypothetical protein	189	38	3	2	tr	
	gi|1787557		185		5	2	tr	
25	gi|4981179	Conserved hypothetical protein	176	32	5 (0)	0 (0)	no	
	gi|1790590		188		5	5	tr	
26	gi|4982342	Translation elongation factor	185	40	5	5	cl	
	gi|1788926		191		5	5	tr	
27	gi|4981047	RNA polymerase sigma factor	193	27	4	1	tr	
	gi|1789127		191		5	1	tr	
28	gi|4981996	Anti-terminator regulatory protein	195	37	5	2	tr	
	gi|1788926		191		5	5	tr	
29	gi|4982169	RNA polymerase sigma factor	189	37	0 (0)	0 (0)	no	
	gi|1788334		196		4	3	tr	
30	gi|4981579	Amidotransferase	201	40	3	2	cl	
	gi|1789875		198		2	1	cl	
31	gi|4981949	Conserved hypothetical protein	175	35	1 (4)	0 (2)	tr	
	gi|1790296		199		0 (0)	0 (0)	no	
32	gi|4982031	Conserved hypothetical protein	195	41	4	2	cl	
	gi|1786258		201		5	5	tr	
33	gi|4980791	Isopropylmalate isomerase subunit	166	35	1 (2)	0 (2)	tr	
	gi|2367128		203		5	5	tr	
34	gi|4981576	Pyrophosphohydrolase	197	43	0 (0)	0 (0)	no	
	gi|1787735		205		5	1	tr	
35	gi|4980651	Actinorhodin polyketide dimerase-related protein	149	35	5	2	no	
	gi|1789452		207		2	1	tr	

TABLE I. (Continued)

	Annotation	Short description	Length (aa)	ID (%)	Expression	Solubility	Crystal trials	HSQC
36	gi|4980784	Conserved hypothetical protein	150	38	3	3	tr	
	gi|1788066		219		3	3	tr	
37	gi|4980985	Pyrazinamidase/nicotinamidase-related protein	214	32	4	3	tr	
	gi|1788624		220		5 (0)	0 (0)	no	
38	gi|4981767	NADH dehydrogenase	178	43	5	2	tr	
	gi|1790069		224		5	5	no	
39	gi|4982124	DNA repair protein	222	40	2 (3)	0 (0)	no	
	gi|1790431		225		3	2	tr	
40	gi|4982450	Endonuclease	225	44	2	1	tr	
	gi|1788920		226		5	2	tr	
41	gi|4981647	Ribonuclease	240	32	2	2	tr	
	gi|1788510		231		5	4	tr	
42	gi|4980760	16S pseudouridylate synthase	239	39	0 (0)	0 (0)	no	
	gi|1786393		235		5	2	tr	
43	gi|4980684	Conserved hypothetical protein	143	40	0 (5)	0 (5)	cl	
	gi|1788820		237		5	4	tr	
44	gi|4981798	SAICAR synthetase	230	40	5	5	cl	
	gi|1786409		243		5	2	tr	
45	gi|4981007	DNA polymerase subunit	189	34	2 (5)	0 (0)	no	
	gi|1790623		243		5	1	no	
46	gi|4982318	Conserved hypothetical protein	242	36	2 (0)	0 (0)	no	
	gi|2367289		246		3	1	tr	
47	gi|4981349	Lipopolysaccharide biosynthesis protein	274	32	2 (3)	0 (0)	no	
	gi|1786407		246		5	2	tr	
48	gi|4981851	Conserved hypothetical protein	207	32	5	1	cl	
	gi|1789858		247		5	2	cl	
49	gi|4982193	Conserved hypothetical protein	222	23	5	4	tr	
	gi|1786890		250		5	3	tr	
50	gi|4982319	nagD protein	259	32	5	2	cl	
	gi|2367307		251		5	1	no	
51	gi|4981280	Methyltransferase	229	33	4 (5)	0 (0)	no	
	gi|1786468		252		0	0	no	
52	gi|4980552	Transcriptional regulator, IclR family	246	31	5	3	cl	
	gi|1789829		252		3 (5)	0 (5)	tr	
53	gi|4981613	Transcriptional regulator, DeoR family	252	24	2	1	tr	
	gi|1789101		253		3	1	tr	
54	gi|4982237	Stationary phase survival protein	247	40	3	1	cl	
	gi|1788959		255		5	5	cl	
55	gi|4982138	tRNA methyltransferase	245	45	2	1	tr	
	gi|1790397		258		5	4	cl	
56	gi|4982363	Acetylglutamate kinase	282	34	4	1	no	
	gi|1786326		264		5	1	no	
57	gi|4982305	Hydroxymethyltransferase	270	48	5	1	no	
	gi|1787342		265		4	1	no	
58	gi|4981188	Conserved hypothetical protein	256	35	5 (5)	0 (0)	no	
	gi|1790035		273		5	4	tr	
59	gi|4981193	Acetyltransferase	220	38	0 (0)	0 (0)	no	
	gi|1786236		273		5	2	cl	
60	gi|4982002	Dimethyladenosine transferase	279	32	4	2	tr	
	gi|1789535		286		5	1	tr	
61	gi|4981233	Conserved hypothetical protein	222	43	0 (2)	0 (0)	no	
	gi|1786312		288		5	4	cl	
62	gi|4981177	Putrescine aminopropyltransferase	296	40	2	1	cl	
	gi|1789983		303		4	2	cl	
63	gi|4980714	glycyl-tRNA synthetase subunit	286	62	5	3	tr	
	gi|1789442		303		5	2	tr	
64	gi|4981054	Fumarate hydratase subunit	272	34	5	4	cl	
	gi|1788490		312		3	2	tr	
65	gi|4982029	Conserved hypothetical protein	285	38	0 (0)	0 (0)	no	
	gi|1786270		313		4	3	cl	
66	gi|4981407	Conserved hypothetical protein	299	41	2	1	tr	
	gi|1786616		367		5	2	tr	
67	gi|4980909	Deoxycytidylate deaminase	201	32	2	2	no	
	gi|1789385		375		3	1	no	
68	gi|4981549	Transcriptional regulator	299	23	4 (3)	0 (1)	tr	

[†]Paired orthologs are presented with *E. coli* results above *T. maritima* results, sorted by number of amino acids of *E. coli* orthologs. Annotation and short descriptions of the ortholog pairs are based on annotations available in NCBI database. Expression and solubility levels, as well as similarity between orthologs, were determined as described in Materials and Methods. Parentheses indicate results obtained at 30°C. The first 23 ortholog pairs were tested for structural studies using both crystal and NMR screening techniques. The final 45 pairs, with greater than 180 amino acids, were tested using crystal screens only. The results of the crystal screening: no, protein was not screened for crystallization due to precipitation or degradation during the purification procedure (see Results and Discussion); tr, protein was screened for crystallization, but no crystallization was detected; cl, initial crystallization conditions were determined. The HSQC results are annotated as pr for "poor" and gd for "good" HSQC profiles.

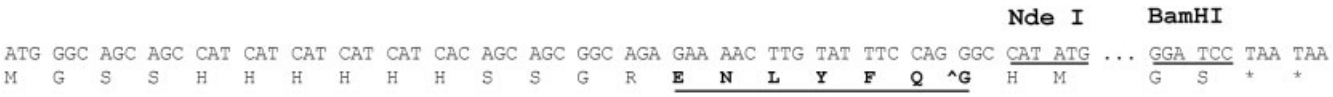

Fig. 1. Cloning site of the modified pET15b vector starting from the translation start point for the N-terminal tag fusion protein. DNA sequence coding for the NdeI and BamHI restriction sites are underlined. Amino acid sequence recognized by TEV protease is in bold and underlined. ∧, the point of cleavage.

Crystallization

The primary crystallization conditions were determined using a sparse crystallization matrix (Hampton Research kits) at room temperature using the sitting drop vapor diffusion technique in 96-well plates (Hampton Research). The protein samples were set up at a concentration of up to 10 mg/ml by mixing 2 μl of sample with 2 μl of the reservoir solution and equilibrated with 100 μl of the reservoir solution. Crystals selected for native and MAD data collection were flash-frozen in the crystallization buffer plus an empirically determined cryoprotectant. The diffraction data were collected at the Advanced Photon Source at Argonne National Laboratories (Argonne, Il).

NMR

Suitability of proteins for NMR analysis was evaluated as described by Yee et al.[1] Briefly, we employed a rapid batch purification of polyhistidine-tagged ^{15}N-labeled protein followed by a rapid "screening" of labeled proteins by ^{1}H-^{15}N heteronuclear correlation (HSQC) spectroscopy. The HSQC spectra were classified as "good" or "poor." The "good" spectra showed dispersion of peaks with roughly equal intensity and in the number expected from the sequence of the protein. These spectra indicated that the protein was readily amenable to structure determination by NMR methods.

RESULTS AND DISCUSSION

Expression and Solubility

All recombinant clones were tested for expression and solubility in *E. coli*. The levels of expression and solubility of the recombinant products were analyzed by comparing the total cell and soluble protein fractions obtained after small-scale (3 ml) growths using a standard expression protocol. Protein expression was induced overnight at 15°C in cells harboring a plasmid encoding three rare tRNAs. These generic expression conditions yielded the maximum expression of the greatest number of samples for structural studies in a 400-protein expression study of proteins from *Methanobacterium thermoautotrophicum*.[17] The levels of over-expression were graded from 0 (no detectable expression) to 5 (dominant protein in extract). The levels of solubility were also graded from 0 (completely insoluble) to 5 (completely soluble) (Table I).

Three *EC* genes and eight *TM* genes were not expressed under the standard expression conditions. Two more *EC* genes and fourteen *TM* genes were expressed, but in insoluble form. For *TM*, 12 of the 38 (32%) proteins under 200 residues and 10 of the 30 (33%) of the *TM* proteins over 200 residues were not expressed or expressed in insoluble form. For *EC*, two of the 32 (6%) proteins under 200 residues and three of 36 (8%) of the *EC* proteins over 200 residues were insoluble. From these analyses, *EC* proteins appear to be more likely to be expressed in soluble form compared with the orthologous proteins from *TM*.

We were concerned that the larger percentage of insoluble *TM* proteins might have been due to the relatively low temperature of induction (15°C), given that the optimal temperature for this organism is 80°C. The insoluble *TM* and *EC* proteins were, therefore, tested for induction at a higher (30°C) temperature (Table I). Four of 22 *TM* proteins and 1 of 5 *EC* proteins showed significantly higher expression and solubility levels when expressed at higher temperatures, and increasing the temperature of induction to 37°C did not further improve the expression and solubility of the proteins (data not shown). Although some improvement of solubility was seen after induction at 30°C, we conclude that the inability to express these specific *TM* proteins in soluble form is not due to the temperature of growth or induction. The recombinant proteins derived from *EC* genes were more highly expressed and more soluble than the corresponding proteins from *TM*, perhaps because the *EC* proteins were expressed in a homologous expression system.

In no instances were both orthologs insoluble (Table I, Fig. 2) showing that the addition of an ortholog does increase the probability of generating a soluble protein.

Protein Purification

The 64 soluble *EC* and 50 soluble *TM* proteins were grown on a larger scale for purification for structural studies. Of the 64 *EC* proteins, 53 (83%) could be purified in a form suitable for structural analysis, namely in sufficient yield (>2 mg/L of culture) and with no obvious precipitation at higher concentration (>2 mg/ml). Eleven *EC* proteins were either degraded or precipitated during

Fig. 2. Distribution of the ortholog clones according to the results of the test for expression and solubility in small scale. The recombinant proteins are classified as not expressed (I), expressed and soluble (II), and expressed, but insoluble (III). Each protein is represented as a number corresponding to the number of its ortholog pair in Table I. The clones that demonstrated increased solubility after over-expression at higher temperature (30°C) are shown in italic. The protein samples that were soluble during small-scale experiments but nevertheless did not generate samples suitable for structural studies due to precipitation or degradation are underlined. The orthologs from two genomes, for which similar results have been obtained, are marked in red.

Fig. 3. **A:** Distribution of the subset of soluble orthologous proteins according to results of screening for NMR samples by HSQC. **B:** Distribution of soluble orthologous proteins according to results of screening for crystallization conditions. The proteins are clustered by providing good HSQC and obtained initial crystallization conditions (I), or by providing poor HSQC and no crystallization in initial screens (II). Each protein is represented as a number corresponding to the number of its ortholog pair in Table I. The orthologs representing the same pair, for which similar results have been obtained, are marked in red.

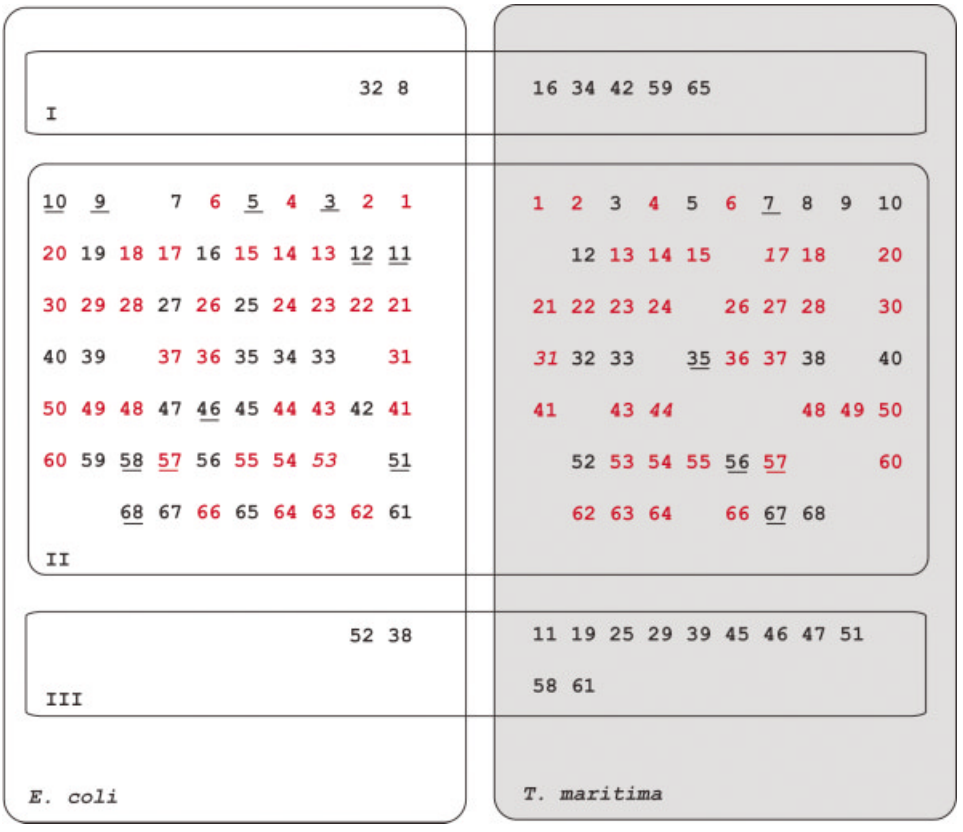

Figure 2.

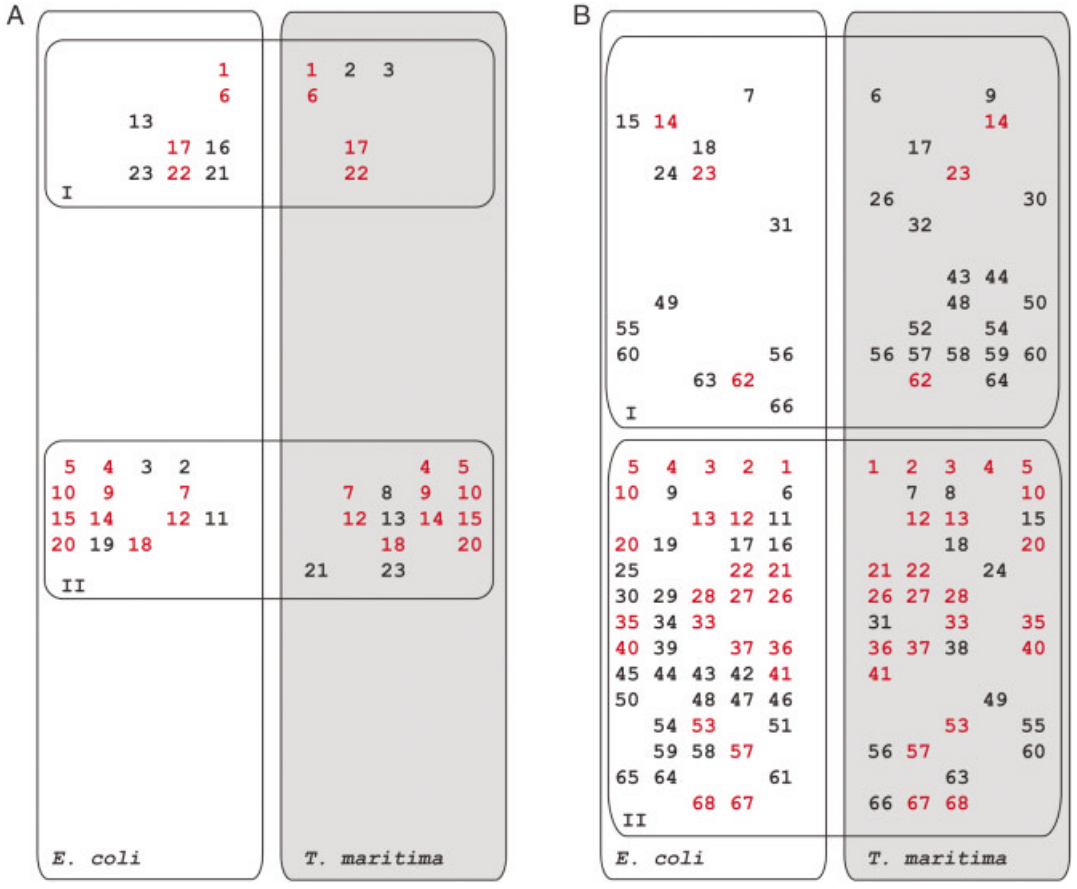

Figure 3.

the purification procedure. Of the 50 *TM* proteins, 46 (92%) could be purified and concentrated for structural studies. Thus, there did not appear to be a significant difference between the genomes in the proportion of soluble proteins that could be purified and concentrated for structural studies.

A total of 38 of the 68 pairs of orthologs could be purified from both *TM* and *EC* sources in a form suitable for structural studies, and six could not be purified from either source (Table I, Fig. 2). A total of 17 proteins could be purified and concentrated only from *EC* clones and 8 could be purified and concentrated only from *TM* clones. Thus, we observed a 25–50% increase in the number of soluble, concentratable samples by the addition of one ortholog of the protein of interest.

We did not observe a correlation between whether a protein could be purified from the *EC* gene and whether it could be purified and concentrated from the *TM* gene. In *EC*, 53 of 68 proteins (78%) could be purified and concentrated for structural studies. In *TM*, the corresponding number is 46 of 68 (68%). If the ability of a protein from one organism to be purified and concentrated is not related to that of its ortholog from another organism, then we would expect that the number of proteins soluble and able to be concentrated from both sources would be the product of the two probabilities ($0.78 \times 0.68 \times 68$ proteins), or 34 proteins, which is very close to 36, the number we observed. If this fact can be extended to orthologs from multiple organisms, then the probability P of deriving a purified, concentrated sample for at least one out of N protein samples would be:

$$P = 1 - \prod_i (1 - P_i) \quad (1)$$

where P_i is the probability of obtaining a good sample from organism i, which is 78% for *EC* and 68% for *TM*. Assuming in each organism we use the success rate of approximately 70%, the overall probability of obtaining at least one good sample would be 96% when we use orthologs from three organisms.

Interestingly, we did not observe a significant size-dependence on the ability to purify and concentrate a sample. Both smaller (<200 residues) and larger (>200 residues) proteins had similar ratios of proteins that were unable to be purified and concentrated.

Suitability for NMR

One of our aims was to explore the advantages of using different orthologs for deriving samples for structural biology. For soluble ortholog pairs under 180 amino acids (Table I), we assessed the suitability for NMR structure determination by employing a rapid batch purification of polyhistidine-tagged ^{15}N-labeled protein followed by a rapid "screening" of labeled proteins by ^{1}H-^{15}N heteronuclear correlation (HSQC) spectroscopy. The HSQC spectrum provides a diagnostic fingerprint of a protein, and the quality of the spectra can be used to assess the suitability for NMR structure determination.[1]. Twenty-three pairs of proteins were labeled with ^{15}N and targeted for NMR analysis. In 10 instances, both the *EC* and the *TM* protein yielded a poor NMR sample, and in four instances both the *EC* and *TM* orthologs had good HSQC spectra [Fig. 3(A)]. In four cases, a suitable NMR sample could be derived only from the *EC* ortholog and in another two instances one could only be obtained for the *TM* ortholog.

In this study, the analysis of proteins from either single genome generated eight (*EC*) or six (*TM*) NMR samples and the combination of genomes added only two or four samples to the total. This marginal increase in samples for structural studies for this small dataset may suggest that small homologous proteins may be more likely to share similar biophysical properties. This analysis also showed that, at least for this sample set, there was no clear advantage to using proteins from thermophiles for NMR spectroscopy in terms of obtaining a sample with favorable NMR properties at 25°C. However, the thermophilic orthologs may in certain cases provide an advantage for NMR data collection at higher temperatures, which may result in better sensitivity due to improved NMR relaxation properties.

Crystallization

All purified proteins, including those for which NMR spectra were collected, were screened for crystallization under a standard set of conditions. This corresponded to 53 *EC* proteins and 44 *TM* proteins [Fig. 3(B)]. Of the 53 *EC* proteins, 14 (26%) formed crystals. For *TM* proteins, 16/50, or 32%, crystallized. Thus, although the success rate of achieving crystals from the starting gene set was similar (14/68 for *EC* and 16/68 for *TM*), the proportion of purified and concentrated *TM* proteins that crystallized was higher. These data suggests that purified *TM* proteins might be easier to crystallize, although more difficult to obtain in soluble form, than their *EC* orthologs. In total, the combination of the two effects resulted in an equivalent number of samples for structural studies attained per gene cloned. However, one must be careful not to conclude that the effects we observe are solely a result of the thermal stability properties of the *TM* proteins. The differences observed between *EC* and *TM* proteins may simply arise because they represent the differences that might be observed between any two sources of orthologous proteins.

Overlap of NMR and Crystallization

We compared the effectiveness of NMR and crystallization to generate samples for structural studies of 46 of the smaller proteins (23 pairs of orthologs). We were able to generate either initial crystallization conditions or a good NMR spectrum for 24 of 46 proteins, corresponding to at least one member of 15 of 23 pairs. Of 34 proteins for which both crystallization and NMR data could be collected, only 3 proteins both crystallized and had good NMR properties. Crystals were obtained for seven proteins with poor NMR spectra and good NMR spectra could be obtained for ten proteins that failed to crystallize in our particular crystallization trials. It is evident that NMR methods and protein crystallization are complementary rather than redundant if the aim is to determine the structures of small proteins. In this sample set, the use of NMR alone would have generated 13 samples for structural studies. The use of

crystallization alone would have generated ten samples with defined initial crystallization conditions. However, the use of both methods increased the number of unique samples for structural studies to 20.

CONCLUSION

Strategy for Structural Proteomics of Small Proteins

The most efficient strategy for the structural proteomics of small proteins may favor the use of NMR. NMR has four clear advantages over crystallization approaches. First, after one has an expression clone that produces a soluble protein, it only takes a few days to characterize its suitability for NMR spectroscopy, whereas it could take days to months to grow a protein crystal. Second, the results of the NMR spectroscopy are usually decisive. Excellent samples can be identified immediately; poor samples can immediately be eliminated from the process. With crystal trials, it is very difficult to make an informed decision based on lack of a crystal. Third, the efficient determination of a crystal structure depends on the presence of methionine in the protein sequence. The smaller the protein, the lower the probability that the protein contains a methionine. Finally, in this study, although we found that an equal number of proteins generated excellent NMR spectra as did crystallize, it is not really accurate to equate these two metrics. An excellent NMR spectrum is highly correlated with the ability to determine its solution structure.[1] However, the growth of a crystal, while an important step towards determining a crystal structure and the most convenient and rapid parameter to measure, does not guarantee that a structure can be solved. A large percentage of protein crystals, in our hands over 50% are difficult to optimize, can be difficult to crystallize as a selenomethionine-labeled protein or are not well-enough ordered to diffract X-rays to high resolution. Therefore, we conclude that a coordinated combination of crystallography with NMR spectroscopy should provide the most efficient path to the structure determination of small proteins.

REFERENCES

1. Yee A, Chang X, Pineda-Lucena A, Wu B, Semesi A, Le B, Ramelot T, Lee GM, Bhattacharyya S, Gutierrez P, Denisov A, Lee CH, Cort JR, Kozlov G, Liao J, Finak G, Chen L, Wishart D, Lee W, McIntosh LP, Gehring K, Kennedy MA, Edwards AM, Arrowsmith CH. An NMR approach to structural proteomics. Proc Natl Acad Sci USA 2002;99:1825–1830.
2. Christendat, D, Yee A, Dharamsi A, Kluger Y, Gerstein M, Arrowsmith CH, Edwards AM. Structural proteomics: prospects for high throughput sample preparation. Prog Biophys Mol Biol, 2000;73:339–345.
3. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 2002;415:180–183.
4. Grant SG, Husi H. Proteomics of multiprotein complexes: answering fundamental questions in neuroscience. Trends Biotechnol 2001;19:S49–S54.
5. Simpson RJ, Dorow DS. Cancer proteomics: from signaling networks to tumor markers. Trends Biotechnol 2001;19:S40–S48.
6. Wojcik J, Schachter V. Protein-protein interaction map inference using interacting domain profile pairs. Bioinformatics 2001;17: S296–S305.
7. Gerstein M, Lin J, Hegyi H. Protein folds in the worm genome. Pac Symp Biocomput, 2000;30–41.
8. Jhoti H. High-throughput structural proteomics using x-rays. Trends Biotechnol 2001;19:S67–S71.
9. Qian J, Stenger B, Wilson CA, Lin J, Jansen R, Teichmann SA, Park J, Krebs WG, Yu H, Alexandrov V, Echols N, Gerstein M. PartsList: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information. Nucleic Acids Res 2001;29: 1750–1764.
10. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 1999;19: 1720–1730.
11. Greenbaum L, Gozlan Y, Schwartz D, Katcoff DJ, Malik Z. Nuclear distribution of porphobilinogen deaminase (PBGD) in glioma cells: a regulatory role in cancer transformation? Br J Cancer 2002;86:1006–1011.
12. Regnier FE, Riggs L, Zhang R, Xiong L, Liu P, Chakraborty A, Seeley E, Sioma C, Thompson RA. Comparative proteomics based on stable isotope labeling and affinity selection. J Mass Spectrom 2002;37:133–145.
13. Cohen AM, Rumpel K, Coombs GH, Wastling JM. Characterisation of global protein expression by two-dimensional electrophoresis and mass spectrometry: proteomics of Toxoplasma gondii. Int J Parasitol 2002;32:39–51.
14. Xing T, Ouellet T, Miki BL. Towards genomic and proteomic studies of protein phosphorylation in plant-pathogen interactions. Trends Plant Sci 2002;7:224–230.
15. Harrison PM, Hegyi H, Balasubramanian S, Luscombe NM, Bertone P, Echols N, Johnson T, Gerstein M. Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. Genome Res 2002;12:272–280.
16. Sali A. Comparative protein modeling by satisfaction of spatial restraints. Mol Med Today 1995;1:270–277.
17. Christendat D, Yee A, Dharamsi A, Kluger Y, Savchenko A, Cort JR, Booth V, Mackereth CD, Saridakis V, Ekiel I, Kozlov G, Maxwell KL, Wu N, McIntosh LP, Gehring K, Kennedy MA, Davidson AR, Pai EF, Gerstein M, Edwards AM, Arrowsmith CH. Structural proteomics of an archaeon. Nat Struct Biol 2000;7:903–909.
18. Hammarstrom, M, Hellgren N, van Den Berg S, Berglund H, Hard T. Rapid screening for improved solubility of small human proteins produced as fusion proteins in Escherichia coli. Protein Sci 2002;11:313–321.
19. Zhang P, Li MZ, Elledge SJ. Towards genetic genome projects: genomic library screening and gene-targeting vector construction in a single step. Nat Genet 2002;30:31–39.
20. Wesson L Eisenberg D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Prot Sci 1992;1:227–235.
21. Chakravarty S Varadarajan R. Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. Biochemistry 2002;41:8152–8161.
22. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int J Neural Syst 1997;8:581–599.
23. Zhang, RG, Kim Y, Skarina T, Beasley S, Laskowski R, Arrowsmith C, Edwards A, Joachimiak A, Savchenko A. Crystal structure of Thermotoga maritima 0065, a member of the IclR transcriptional factor family. J Biol Chem 2002;277:19183–19190.
24. Christendat D, Saridakis V, Dharamsi A, Bochkarev A, Pai EF, Arrowsmith CH, Edwards AM. Crystal structure of dTDP-4-keto-6-deoxy-D-hexulose 3,5-epimerase from Methanobacterium thermoautotrophicum complexed with dTDP. J Biol Chem 2000;275: 24608-24612.