DOI 10.1002/pmic.200800061 *Proteomics* 2008, *8,* 4772–4781

RESEARCH ARTICLE

Structural features and the persistence of acquired proteins

Hema Prasad Narra, Matthew H. J. Cordes and Howard Ochman

Department of Biochemistry & Molecular Biophysics, University of Arizona, Tucson, AZ, USA

ORFan genes can constitute a large fraction of a bacterial genome, but due to their lack of homologs, their functions have remained largely unexplored. To determine if particular features of ORFan-encoded proteins promote their presence in a genome, we analyzed properties of ORFans that originated over a broad evolutionary timescale. We also compared ORFan genes to another class of acquired genes, heterogeneous occurrence in prokaryotes (HOPs), which have homologs in other bacteria. A total of 54 ORFan and HOP genes selected from different phylogenetic depths in the *Escherichia coli* lineage were cloned, expressed, purified, and subjected to circular dichroism (CD) spectroscopy. A majority of genes could be expressed, but only 18 yielded sufficient soluble protein for spectral analysis. Of these, half were significantly α-helical, three were predominantly β-sheet, and six were of intermediate/indeterminate structure. Although a higher proportion of HOPs yielded soluble proteins with resolvable secondary structures, ORFans resembled HOPs with regard to most of the other features tested. Overall, we found that those ORFan and HOP genes that have persisted in the *E. coli* lineage were more likely to encode soluble and folded proteins, more likely to display environmental modulation of their gene expression, and by extrapolation, are more likely to be functional.

Received: January 21, 2008
Revised: June 10, 2008
Accepted: June 17, 2008

Keywords:
E. coli / Genome evolution / Lateral gene transfer / ORFans / Protein folding

1 Introduction

Most free-living bacteria have obtained and incorporated alien DNA, with the result that a large fraction of the variation in the size and contents of microbial genomes, even among the strains of same species, is attributable to horizontally acquired sequences [1, 2]. Due to the very large number of sequenced genomes that are currently available, it is often possible to apply phylogenetic or similarity-based approaches to pinpoint the source and origins of acquired DNA [3, 4]. Whereas, many foreign genes have counterparts in other organisms, bacterial genomes frequently contain lineage-, species-, strain-, or genome-specific genes, termed ORFans, that have no recognizable homologs in the current sequence databases [5, 6]. For example, the presence of over 500 strain-specific ORFan genes accounts for most of the difference in genome size between *Escherichia coli* CFT073 and *E. coli* K12 [7, 8]. ORFan genes have been detected in virtually all sequenced organisms, and they can constitute up to 14% of the annotated genes in a bacterial genome [9].

Several hypotheses have been put forward to account for the presence of ORFan genes in bacterial lineages [5, 10–12]. One such hypothesis is the recurrent insertion of palindromic elements or domains into existing ORFs resulting in genes that, due to complex sequence rearrangements, are beyond the LODs of existing computational approaches [13, 14]. But, based on the sequence characteristics, it was proposed that ORFans represent genes principally of viral origin [6], and a more recent study found that a small fraction of ORFans have viral gene homologs [8]. Moreover, the pres-

Correspondence: Dr. Howard Ochman, 233 Life Sciences South, University of Arizona, Tucson, AZ 85721, USA
E-mail: hochman@email.arizona.edu
Fax: +1-520-621-3709

Abbreviations: CD, circular dichroism; **HOPs,** heterogeneous occurrence in prokaryotes

ence of mycobacterial ORFans in several mycobacteriophages further suggests that phage serve as vehicles for the origin, transfer and integration of these genes into bacterial genomes [15].

Alignments of the ORFan gene homologs that are confined to the bacterial clade containing *E. coli* and *Salmonella* revealed that most ORFans represent functional protein-coding regions, as apparent from their low K_a/K_s values [6]. Many ORFan genes from *E. coli* and other species have been shown to be transcribed [16–18]. Recent studies on mimivirus proteomes revealed that more than 10% of the ORFan-encoded proteins display antigenic properties suggesting that many are functional [19]. The occurrence of putative transmembrane domains and repetitive sequences in several viral orphan proteins also indicates a possible role in metabolic exchange and nucleic acid interactions during host infections [20]. Moreover, orphan proteins contain large numbers of aromatic amino acids (phenylalanine, tryptophan, and tyrosine) resulting, in part, from the A+T-rich base composition in their coding sequences [21]. Interestingly, aromatic amino acids have been hypothesized to play some role in generating new protein functions as well as to stabilize protein structures by creating long hydrophobic aromatic side chains [22–25]. But, because bacterial ORFans rarely exhibit significant sequence similarity to known proteins, and genome-wide mutagenesis studies have typically failed to assign phenotypes to the vast majority of ORFan genes [26], their functions have remained elusive.

Considerable progress toward the functional assignment of ORFan-encoded proteins has been made by computational approaches, including the analysis of conserved residues or motifs and of 3-D structural models [27–29] and by structural characterization *via* NMR and X-ray crystallography [30, 31]. Since ORFans encode unique proteins, and are therefore likely to confer novel functions, it is perhaps not surprising that many of the novel domain folds deposited in the databases are derived from ORFans [32–34].

To understand the structural similarities between ORFans and characterized proteins, and to begin resolving new structures that might be unique to ORFans, we expressed and purified ORFan-encoded proteins, which were then subjected to structural analysis by circular dichroism (CD) spectroscopy. By analyzing ORFans that originated at different times during the evolution of the gammaproteobacterial lineage leading to *E. coli*, we show that ORFans genes that persist in bacterial genomes are more likely to encode soluble and folded proteins with significant secondary structure content.

2 Materials and methods

2.1 Selection of target genes

Genes were selected for analysis based on their phylogenetic distributions in the lineage leading to *E. coli* MG1655, as reported in Daubin and Ochman [6]. For each gene in the *E. coli* MG1655 genome, authors distinguished clade-specific (i) ORFans, which are genes assigned parsimoniously to a particular ancestral lineage and have no homologs outside of the descendents of this lineage, and (ii) heterogeneous occurrence in prokaryotes (HOPs), which are genes that were assigned parsimoniously to a particular ancestral lineage and have homolog(s) in a bacterial genome outside of the descendents of this lineage (suggesting an origin by horizontal transfer). Because ORFans and HOPs can originate at any time over the evolutionary history of a lineage, their relative ages can be inferred from their phylogenetic distributions. For the present study, the target genes were selected from clades of different phylogenetic depths – clade n_0, restricted to the *E. coli* MG1655 genome; n_1, ancestral to all *E. coli* strains; n_2, ancestral to *E. coli* and *Salmonella enterica*; n_3, ancestral to all enteric bacteria – each representing a different origination time and duration within the gammaproteobacterial lineage leading to *E. coli* (See Daubin and Ochman [6] for additional details). We selected genes ranging from approximately 400–800 bp in length and encoding protein products of 15–30 KDa, under the assumption that shorter genes, although conducive to CD spectral analysis, were more likely to represent annotation artifacts. Additionally, the context of genes was taken into consideration; we selected target genes that had different strand orientations and/or combinations of ORFans, HOPs, or ancestral genes as neighbors.

2.2 Amplification and cloning of target genes

Target genes were amplified from *E. coli* MG1655 genome using gene-specific primers that incorporated a 5′-tagged recognition site for either *Nde*I or *Xho*I. The amplified PCR products were purified using Qiaex II (Qiagen), digested with the appropriate restriction endonucleases (New England Biolabs) and directionally ligated into pET21b expression vector (Novagen) at the *Nde*I and *Xho*I sites. This procedure leads to the C-terminal addition of a hexa-histidine tag, which allows purification of the recombinant proteins by Ni^{2+}-NTA chromatography. Ligated plasmid constructs were transformed into XL10-Gold® Ultracompetent cells (Stratagene), and the inserts of selected recombinants were confirmed by sequencing.

2.3 Expression and purification of target proteins

Target proteins were expressed in and purified from *E. coli* BL21(λDE3) competent cells (Stratagene). Briefly, transformants were inoculated in 40 mL of Luria–Bertani broth and grown at 37°C to an $OD_{600} = 0.5$. Cultures were induced by the addition of IPTG to a final concentration of 1 mM and grown at either 30 or 37°C for 3 h to allow expression of the target gene. After the induction period, cells were harvested by centrifugation, and the resulting

pellet was suspended in Buffer A [100 mM NaH_2PO_4 (pH 8.0), 10 mM Tris, 6 M guanidine hydrochloride] and lysed by mechanical inversion for 1 h. The lysate was then adjusted to 10 mM imidazole and centrifuged at 12 000 × *g* for 15 min. Clear guanidine lysates were passed through Ni^{2+}-NTA columns (Qiagen), which were washed twice with Buffer A supplemented with 10 mM imidazole to prevent the nonspecific binding of proteins. Target proteins were eluted with 0.6 mL of Buffer F (200 mM acetic acid, 6 M guanidine hydrochloride). The purified proteins were refolded by dialysis into 100 mM NaH_2PO_4 buffer (pH 7.0) supplemented with 3 mM β-mercaptoethanol to reduce the formation of disulfide bonds. Dialyzed proteins were centrifuged at 12 000 × *g* for 15 min to separate soluble protein from precipitates. A minimum of 600 μL of supernatant fraction containing soluble refolded protein were recovered after renaturation by dialysis for each sample. The amount of soluble refolded protein in the supernatant fractions ranged from 0.03 mg/mL (yabP) to 1.43 mg/mL (paaD). When necessary, proteins were concentrated using centrifugal filters with a cutoff value of 5 kDa (Millipore). Relative amounts of precipitated and soluble protein were assessed by running volume-equalized pellet and supernatant samples on 12% SDS-PAGE gels stained with CBB. The concentration of soluble protein was measured on a CARY 50 spectrophotometer. Molar extinction coefficients (ε_{280}) were calculated based on the number of tryptophan residues (5559 cm^{-1} *per* residue) plus the number of tyrosine residues (1197 cm^{-1} *per* residue). Soluble protein concentration was then estimated by dividing the A_{280} values by the ε_{280} extinction coefficient.

2.4 Circular dichroism (CD) spectroscopy

Far-UV CD spectra of soluble proteins were obtained with an Olis DSM20 CD spectrophotometer. Baseline calibrations were generated from the wavelength scans obtained with dialysis buffer (100 mM NaH_2PO_4, 3 mM β-mercaptoethanol). Wavelength scans for each *in vivo* soluble protein were obtained from an average of three scans at 1 nm increments at 20°C in a 1 mm path length cuvette with an integration time of 2 s. Protein concentrations ranging from 8.0 to 10.3 μM were used for obtaining CD spectra. Three independent scans were recorded for each protein. Mean residue ellipticity (Θ) was calculated as follows: $(\Theta) = \Theta \times 100/l \times n \times c$, where "Θ" is the ellipticity in millidegrees, "*l*" the cell path length in centimeters, "*n*" the number of amino acids in the protein sequence, and "*c*" is the millimolar concentration of the protein. Percent helicity was calculated as % helicity = $100 \times [(\Theta_{222} + 2340)/30\,300]$, where "$\Theta_{222}$" is the mean residual ellipticity at 222 nm [35]. Based on the CD spectra and percent helicity, proteins were characterized as significantly α-helical, predominantly β-sheet, or intermediate/indeterminate. Representative CD spectra of ORFan and HOP proteins are presented in Fig. 1.

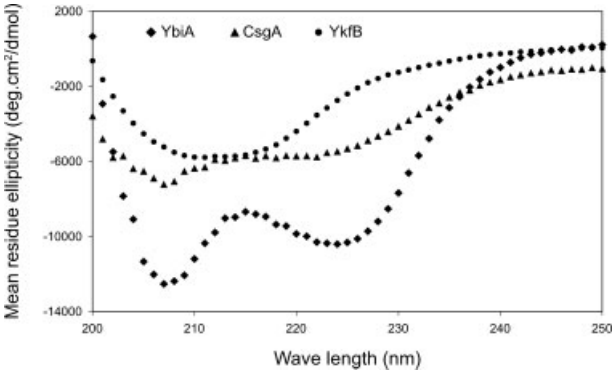

Figure 1. Representative CD spectra of ORFan and HOP proteins. Proteins were classified as follows: YbiA, significantly α-helical; CsgA, intermediate/indeterminate; YkfB predominantly β-sheet.

3 Results

3.1 Expression, denaturing purification, and refolding of ORFans and HOPs

Of 30 ORFan genes initially cloned and expressed, 16 yielded amounts of protein that were detectable on SDS-PAGE gels following denaturing purification and refolding by dialysis. Eight of these proteins precipitated as an insoluble fraction during refolding, whereas others yielded varying levels of soluble protein in supernatant fractions (Table 1). In three cases (YdaT from clade n_2, CrcA from clade n_3, and DnaT from clade n_3), solubility of the expressed protein was higher when induced at 30°C than at 37°C (Table 1). Proteins were recovered from 7 of 17 of the more recently originated ORFan genes (those from clades n_0 and n_1) *versus* 9 of 13 ORFans genes from the n_2 and n_3 clades. Three of 17 of the n_0 and n_1 genes yielded soluble refolded proteins, with only two generating amounts sufficient (>8 μM) for further characterization. By contrast, 5 of 13 of the n_2 and n_3 genes produced sufficient amounts of soluble refolded protein. Thus, a larger proportion of the phylogenetically older ORFans produced adequate yields of soluble refolded proteins in these expression experiments.

Of 24 HOP genes cloned, 15 encoded proteins that yielded detectable amounts of expressed protein following denaturing purification, which is only a slightly larger fraction than observed with ORFan genes (Table 1). Among the proteins expressed by HOP genes, however, a significantly smaller proportion was insoluble than among those expressed by ORFan genes: only three were insoluble (and these were all from clade n_1), three were partially soluble, and nine were soluble upon renaturation. Unlike the ORFans, phylogenetically younger HOP genes, *i.e.*, those from the n_0 and n_1 clades, did not show lower expression levels. However, there were clade-specific differences in refolding behavior: 5 of 13 of the n_0 and n_1 genes gave some soluble refolded protein, but only four yielded amounts sufficient (>8 μM) for further

Table 1. Characteristics of genes cloned and expressed for structural analysis

Gene	Size (bp)	Type[a)]	Clade[a)]	Gene context[b)]	Protein recovery[c)]				Solubility
					37°C Pellet/supernatant		30°C Pellet/supernatant		
ybfP	495	ORFan	n_0	A → *for* ← H n_0	++	+++	–	–	Soluble
ycdV	414	ORFan	n_0	O n_0 → *for* → A	–		–	–	
C3003	504	ORFan	n_0	H n_3 ← *for* → O n_1	+++	–	–	–	Insoluble
yefJ	474	ORFan	n_0	H n_0 → *rev* → H n_0	–	–	–	–	
yhcE-1	480	ORFan	n_0	A → *for* ← H n_0	++	–	–	–	Insoluble
paaD	504	HOP	n_0	H n_0 → *for* → A	+++	++	+++	+++	Soluble
paaI	423	HOP	n_0	A → *for* → A	–	–	–	–	
yegJ	462	HOP	n_0	H n_1 ← *for* ← H n_1	++	++	–	–	Soluble
yieJ	588	HOP	n_0	O n_1 → *for* ← H n_4		–	–	–	–
fecI	522	HOP	n_0	H n_0 → *rev* ← O n_3		–	–	–	–
htgA	591	ORFan	n_1	H n_2 ← *for* ← O n_2	–	–	–	–	
yabP	651	ORFan	n_1	A → *for* → O n_0	+	+	–	–	Soluble
yafT	786	ORFan	n_1	A → *for* ← O n_0	–	–	–	–	
yafO	399	ORFan	n_1	O n_4→*for*→A	+	–	++	–	Insoluble
ykfB	468	ORFan	n_1	O n_1 → *rev* → H n_1	+	+++	–	–	Soluble
ybcN	456	ORFan	n_1	A → *for* → O n_1	–	–	–	–	
Z0753	462	ORFan	n_1	A → *for* ← A	–	–	–	–	
ybfC	570	ORFan	n_1	A → *for* → H n_1	–	–	–	–	
ymfL	570	ORFan	n_1	O n_0 → *for* → O n_0	+++	–	+++	–	Insoluble
sieB	612	ORFan	n_1	O n_2 ← *for* → O n_1	–	–	–	–	
ydcD	483	ORFan	n_1	A → *for* → H n_1	–	–	–	–	
yeeP	711	ORFan	n_1	A → *for* → A	–	–	–	–	
yafX	459	HOP	n_1	A ← *rev* → O n_1	–	–	–	–	
insO-1	426	HOP	n_1	H n_1 → *for* → H n_1	+	–	–	–	Insoluble
cynS	471	HOP	n_1	A → *for* → H n_2	+	+++	–	–	Soluble
yaiX	444	HOP	n_1	O n_1 → *rev* ← A	+++	+	–	–	Partially soluble
ybiA	483	HOP	n_1	A ← *rev* ← A	+++	+++	–	–	Soluble
ychG	591	HOP	n_1	A ← *rev* → A	+++	–	+++	–	Insoluble
wbbJ	591	HOP	n_1	H n_0 → *rev* → H n_0	+	–	+++	–	Insoluble
yigE	486	HOP	n_1	H n_1 → *rev* ← A		–	–	–	–
yaiV	669	ORFan	n_2	A → *for* ← A	+	–	+++	–	Insoluble
ylbF	816	ORFan	n_2	O n_2 → *for* → A	+	–	–	–	Insoluble
ybhQ	411	ORFan	n_2	H n_2 ← *for* ← A	–	–	–	–	
csgA	456	ORFan	n_2	H n_2 → *for* → O n_2	++	+++	–	–	Soluble
ydaT	423	ORFan	n_2	H n_1 → *for* → H n_2	+++	+	+++	++	Partially soluble
ydaW	612	ORFan	n_2	A → *for* → O n_3	–	+++	–	–	Soluble
allA	483	HOP	n_2	A ← *for* → A	+++	+	++	++	Partially soluble
nohB	546	HOP	n_2	O n_1 → *for* → H n_3	+++	++	++	+++	Soluble
mntR	468	HOP	n_2	O n_1 → *for* → A	+++	+++	–	–	Soluble
eutP	480	HOP	n_2	H n_2 → *rev* → H n_2	+++	+	++	++	Partially soluble
hyfJ	477	HOP	n_2	A → *for* → A	–	–	–	–	
yhaM	567	HOP	n_2	O n_2 ← *rev* → H n_4	–	–	–	–	
crcA	561	ORFan	n_3	H n_4 → *for* → A	++	++	+	+++	Soluble
ompX	516	ORFan	n_3	A ← *for* ← A	+++	–	+++	–	Insoluble
ybjN	477	ORFan	n_3	A → *for* → A	–	–	–	–	
ybjO	489	ORFan	n_3	A → *for* → A	+++	–	+++	–	Insoluble
yniB	537	ORFan	n_3	H n_4 ← *rev* ← A	–	–	–	–	
yicN	480	ORFan	n_3	A → *rev* → A	–	–	–	–	
dnaT	540	ORFan	n_3	H n_3 → *rev* → H n_4	–	++	–	+++	Soluble
yccU	495	HOP	n_3	O n_4 ← *for* ← H n_3	+++	+++	–	–	Soluble

Table 1. Continued

Gene	Size (bp)	Type[a)]	Clade[a)]	Gene context[b)]	Protein recovery[c)]				Solubility
					37°C Pellet/supernatant		30°C Pellet/supernatant		
ydeR	504	HOP	n_3	O n_1 → *rev* → A	+	+	++	++	Soluble
nrdI	411	HOP	n_3	A → *for* → A	–	++	–	+++	Soluble
yqjF	483	HOP	n_3	O n_3 → *for* → A	–	–	–	–	
yiaA	441	HOP	n_3	H n_4 ← *rev* → H n_3	–	–	–	–	

a) Gene and clade assignments follow those of Daubin and Ochman [6].

b) The strand orientation of the cloned gene, forward (*for*) or reverse (*rev*), in the *E. coli* MG1655 genome is noted, and the orientation of adjacent genes relative to that of the cloned gene are denoted by the direction of the arrows. Gene notations are as followed: A, ancestral gene; O, ORFan gene; H, HOP gene, followed by the clade assignments in the case of ORFans and HOPs.

c) Sample amounts after dialysis are as follows: "+" (<5 μM), "++" (5–10 μM), and "+++" (>10 μM). Total dialysate volume was 600 μL. For soluble proteins, concentrations were measured on a CARY 50 spectrophotometer; for insoluble fractions, amounts were quantified visually after staining on 12% SDS-PAGE.

characterization. By contrast, all seven of the expressed n_2 and n_3 genes gave significant yields of soluble protein (Table 1). Thus, with the HOPs, as with the ORFans, fewer of the "younger" genes gave refoldable and soluble proteins.

Overall, of the 54 genes cloned, 31 expressed amounts of protein that were detectable after denaturing nickel affinity chromatography when induced at 30 or 37°C, of these, 13 yielded soluble proteins. For ten genes (five ORFans, five HOPs), protein production was higher when expression was induced at 30°C (as opposed to 37°C), and such enhancements at lower temperatures were observed principally among genes from the older clades, n_2 and n_3.

3.2 Secondary structure content of soluble ORFan and HOP proteins

The secondary structure of each of the soluble (or partially soluble) refolded proteins was analyzed by far-UV CD spectroscopy. CD spectra and general structural classifications were obtained for the 7 ORFans and 11 HOPs that yielded sufficiently high concentrations (>8 μM) of soluble protein after renaturation by dialysis (Table 2). All of these proteins appeared to have significant secondary structure content and thus appear to be folded rather than disordered in solution. In particular, all proteins lacked a strong negative signal in the 200–205 region, which would be present in a protein dominated by unfolded, random coil structure. Representative CD spectra for each of the secondary structure category observed for ORFan and HOP proteins are shown in Fig. 1.

Based on their CD spectral curves (Fig. 1 of Supporting Information), three ORFan-encoded proteins (YbfP, YdaW, and CrcA) were significantly α-helical, YkfB and DnaT were characterized as β-sheet proteins, and CsgA and YdaT were of intermediate/indeterminate structure (Fig. 2). The percent helicity (*i.e.*, roughly the proportion of helical residues in the protein) in the α-helical proteins was twice that observed in the predominantly β-sheet proteins (40% *vs.* 19%) (Table 2).

Among the HOP proteins for which we obtained CD spectra, a majority were significantly α-helical, with an average helicity of 38% (Fig. 2 and Table 2). The spectra of the four HOP proteins (YegJ, PaaD, EutP, and NrdI) clearly exhibited a fold corresponding to intermediate/indeterminate structures (Fig. 2 of Supporting Information). Only a single HOP protein (AllA) showed a CD spectrum consistent with that of a β-sheet protein (Fig. 2), and this protein had lowest helicity (11%) among all proteins tested (Table 2).

4 Discussion

Because the functions of most proteins in most genomes are recognized only through sequence or structural similarities to previously characterized proteins, inferring the functions of ORFans (*i.e.*, genome- or lineage-specific proteins), which lack homologs, presents a particular challenge. Although, ORFans were originally thought to correspond to the nonfunctional or poorly annotated regions of a genome, there is growing evidence that these sequences encode functional proteins [34]. In the present study, we applied a multifactorial approach that considered numerous characteristics (gene clonability and expression; protein refoldability, solubility, and secondary structure) of ORFans of different ages in the *E. coli* lineage in order to identify factors that promote their presence and maintenance in bacterial genomes.

The initial steps in the physical characterization of these proteins rely on the cloning and expression of ORFan genes. Although, there are large numbers of genes that cannot be cloned or expressed at high copy-numbers because they are deleterious to their *E. coli* host [36–38], all of the ORFan genes we tested were already constituents of the *E. coli* MG1655 genome, and the expression of each native locus has been observed under some growth condition. Using quantitative PCR, we verified that several of the ORFan genes generated high levels of mRNA when grown at different

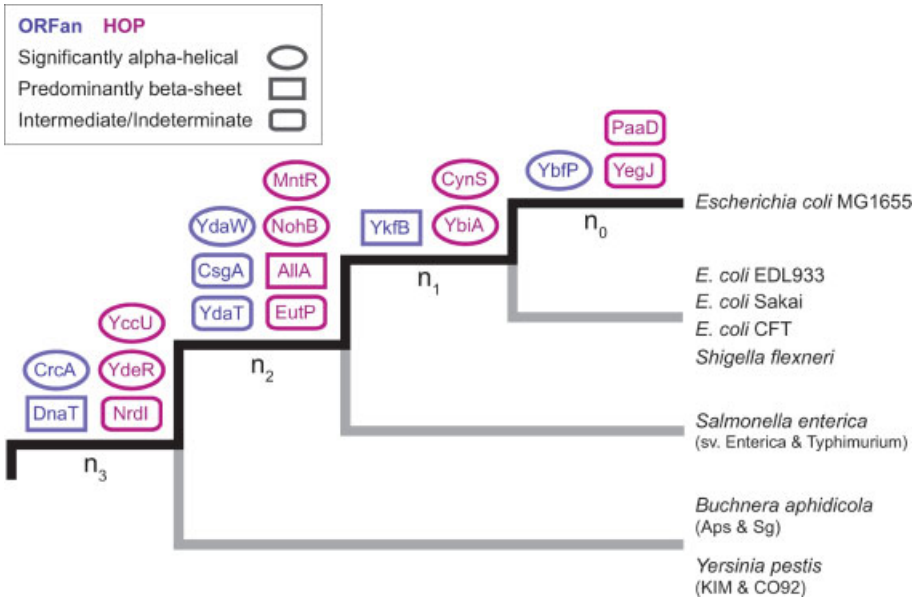

Figure 2. Phylogenetic distribution and structural classification of ORFan and HOP proteins with resolved CD spectra.

Table 2. Structural characteristics of *E. coli* ORFan and HOP proteins

Clade	Protein name	Protein length	Θ (mdeg)	Θ_{222}	Helical residues	Percent helicity
A. ORFans (genome/clade-specific proteins)						
n_0	YbfP	164	−16.2	−9903.3	66	40.4
n_1	YkfB	155	−5.5	−3548.4	30	19.4
n_2	CsgA	151	−8.7	−5723.7	40	26.6
n_2	YdaW	203	−40.8	−9570.7	80	39.3
n_2	YdaT	140	−4.8	−4033.6	29	21.0
n_3	CrcA	186	−18.5	−9946.2	75	40.5
n_3	DnaT	179	−5.9	−3296.1	33	18.6
B. HOPs (genome- or clade-specific proteins with homologs in distant bacteria)						
n_0	YegJ	153	−2.2	−2879.5	26	17.2
n_0	PaaD	167	−12.7	−7620.4	55	32.9
n_1	YbiA	160	−16.5	−10 306.0	67	41.5
n_1	CynS	156	−15.0	−9520.2	61	39.1
n_2	MntR	155	−19.9	−12 838.7	78	49.8
n_2	EutP	159	−2.0	−1382.3	20	12.3
n_2	AllA	160	−1.7	−1031.6	18	11.1
n_2	NohB	181	−14.7	−11 438.8	82	45.5
n_3	NrdI	136	−4.9	−3714.4	27	20.0
n_3	YdeR	167	−18.1	−10 731	72	43.1
n_3	YccU	164	−15.1	−9207.3	63	38.1

conditions. Nevertheless, in surveying the conditions that might regulate the expression of ORFan genes, a survey of available microarray data shows that 5 of the 14 ORFan genes that did not yield proteins never showed differential expression under any condition tested [39, 40]. In contrast, those ORFan genes for which we were able to isolate proteins had a higher probability of being differentially expressed, indicating that they are likely produced (and useful) during some stage of the *E. coli* life cycle. It should be noted that over half of the ORFan genes examined were induced during adaptation to glucose-limiting conditions [41, 42]. Thus, we were more likely to recover proteins from ORFan genes whose regulation had previously been shown to be modulated by some environmental variable.

Because bacterial genes are often arranged in operons, gene context – *i.e.*, the order and orientation of adjacent genes – can also influence patterns of gene expression. We examined the nearest neighbors of each of the cloned ORFan genes with the thought that ORFans that reside next to and transcribed in the same strand orientation as an ancestral gene would be more likely to produce stable proteins because they do not require their own promoter to drive transcription. Neither the ancestry of adjacent genes nor the orientation of an ORFan gene relative to its neighbors provided clues about our ability to isolate protein from a particular gene. Those ORFans situated immediately downstream of ancestral genes were as likely as not to produce proteins in our system; and although most ORFan genes were aligned in the same orientation as their upstream genes, in three of the eight cases where they were in opposite direction, we were able to recover proteins for structural analysis (Table 1).

Only about 60% of the ORFans and HOPs tested in our study displayed detectable levels of protein after overexpression from the pET21b vector grown in *E. coli* BL21 (λDE3). However, Yee *et al.* [43] using a similar expression system, found that nearly 80% of the *E. coli* native proteins of under 23 kDa (and which lacked predicted transmembrane domains) could be expressed. Thus, the expression of ORFan genes, as a group, may differ substantially from the rest of the genome, and the difference in expressability may be attributable to the fact that (i) the genes in our study were chosen without regard to their status as membrane proteins, (ii) the unusual base compositions of ORFans genes might result in an increased high metabolic cost to the host, and (iii) some ORFan proteins could not be extracted by traditional lysis procedures (6 M guanidine).

Although, we have no information about their *in vivo* solubility, 58% of the expressed ORFan and HOP proteins could be refolded into a soluble form by dialysis out of guanidine, similar to that obtained in a previous study that reported that about 55% of the *in vivo* insoluble proteins in the size range of 6–18 kDa could be refolded successfully out of guanidine (in contrast to soluble proteins in which nearly 90% of could be refolded) [44]. Because larger proteins are more difficult to refold, the overall lower rate of refolding success of ORFan and HOP proteins (58% in the present study *vs.* 55–88% observed by Maxwell *et al.* [44]) might be attributable to differences in the size class of the proteins tested (15–30 kDa *vs.* 6–18 kDa, respectively).

The ORFans and HOPs which were refoldable and had significant secondary structure content seem likely to represent functional proteins. It must also be noted, however, that expressed proteins which fail to refold in a soluble form from denaturant solutions may also be perfectly functional *in vivo*. First, some proteins that are insoluble in physiological buffers may be associated with cell membranes and in the two cases for which functions have been assigned to ORFan genes producing insoluble proteins, one (*yafO*) encodes a toxin, and the other (*ompX*) encodes an outer member protein implicated in adhesion and virulence [45–47]. Second, some proteins may be soluble in their native form, but undergo non-native aggregation and precipitation in competition with correct refolding during dialysis from guanidine. Failure of soluble proteins to refold correctly *in vitro* may be due to the absence of protein folding modulators, such as chaperones and proteases, that may be necessary to ensure proper folding, stabilization, and solubilization *in vivo* [48–51]. In short, refoldability is one observation in favor of a protein being functional, but it is by no means a necessary precondition.

It is also likely that some of proteins that were insoluble by our procedures are soluble in their native state but undergo precipitation in competition with correct refolding during dialysis. Ventura and Villaverde [52] reported that inclusion bodies often contain well-folded and biologically active proteins, and studies on β-galactosidase of *E. coli* [53] and endoglucanase D of *Clostridium thermocellum* [54] detected active proteins in inclusion bodies, further suggesting that protein solubility may be due to the conditions used for protein extraction and refolding, but not the functional quality, of the expressed protein.

The majority of soluble and refolded ORFan and HOP proteins yielded CD spectra indicative of significant secondary structure content, with a prevalence of α-helical structures. A previous study focusing on insoluble hypothetical proteins from *Thermotoga maritima* and *Saccharomyces cerevisiae* reported that approximately 50% of the proteins with well-resolved CD spectra corresponded to an α-helical structure, whereas only 9–25% corresponded to β-sheet configurations [44]. A broad survey of the SP175 CD spectra reference database (http://pcddb.cryst.bbk.ac.uk) showed that proteins with β-sheet structures are twice as common as α-helical proteins [55]. However, Wolf *et al.* [56] and Hegyi *et al.* [57] note that β-sheet proteins are most commonly involved in regulation and signal transduction, and are generally more highly abundant in eukaryotes than in bacteria. Thus, the dearth of predominantly β-sheet protein structures in our study could be attributable to the size class of proteins that we investigated, the emphasis of the present study on prokaryotic proteins, and even an increased probability for α-helical proteins to refold from guanidine [58]. It should also be noted that some of the proteins classed as α-helical in our study may also contain substantial β-sheet structure that is masked by the helical signal, which tends to dominate the spectrum. For example, protein YbiA (categorized as significantly α-helical in our study) for which the NMR structure has been solved (PDB code: 2B3W), has some β-sheet though it is mainly dominated by α-helix structure. Furthermore, our CD spectral observation on protein AllA (predominantly β-sheet) correlated with crystallographic studies which revealed the protein to be of β-sheet structure [59].

Among those ORFans for which we were able to isolate proteins, there is an association between the age (*i.e.*, phylogenetic distribution) of a gene and its ability to refold out of guanidine. Previous microarray studies assessing gene expression levels under various growth conditions revealed

that about half of our cloned ORFan genes selected from younger clades (n_0 and n_1) were transcriptionally variable as compared to 92% from older clades (n_2 and n_3) [39–42].

It was much more common for older ORFans to yield substantial amounts of soluble refolded protein with highly resolved CD spectra when compared to the proteins encoded by younger ORFans. This presence of folded ORFan proteins with significant secondary structure in deeper clades is perhaps not surprising given that nonfunctional genes are eventually lost from bacterial genomes [60, 61]. However, the preponderance of well-folded proteins among the older ORFans indicates that genes that encode proteins with well-folded (as opposed to unfolded) structures are more likely to be functional and retained.

Though the sources of ORFans remain obscure, most are likely to be of viral origin [8]. To determine if ORFans differ from other classes of acquired genes, we also examined the same features in genes acquired from distantly related species, with the thought that many of these sequences (termed HOPs) were already functioning in another bacterial genome prior to their arrival in *E. coli*. As a class, HOPs encode longer proteins (Daubin and Ochman [6]), but in the present study, we selected HOPs that averaged approximately the same length as ORFans.

We were able to obtain proteins for 63% of the 24 HOP genes cloned, but only 13% resulted in insoluble fractions, about half of that observed for ORFans. In addition, recovery of soluble and partially soluble HOP proteins for structural studies was much higher (73% *vs.* 44%). But for most of the other features that we tested, HOPs resembled ORFans. Like ORFans, gene context and orientation had no apparent affect on the expression of cloned HOPs. Furthermore, of the HOP proteins for which the CD spectra were obtained, the majority were from older clades (n_2 and n_3). Finally, structural analysis revealed that approximately half of the HOP proteins that yielded CD spectra had significant amounts of α-helical structure, which agrees with that observed for ORFan proteins and for proteins of this size class in general [44].

In summary, we examined and determined the expression, solubility, refolding ability, and secondary structure of ORFan and HOP genes present in *E. coli* MG1655 genome. Of the total of 54 genes cloned, nearly 60% could be expressed and 18 produced refoldable, soluble proteins that yielded CD spectra resembling structures of folded proteins. Our data support the notion that many ORFans are indeed real genes and encode folded and probably functional proteins, despite their lack of clear conserved homologs. A recent study focusing on the poorly conserved paralogs in *Halobacterium* sp. NRC-1 found that nearly 80% of the genes tested were transcriptionally active, indicating that these genes might encode true proteins [18]. Interestingly, it seems that many ORFans, even those that have persisted for long evolutionary time spans, are likely to encode proteins that have truly novel folded structures with unique functions or to correspond to rapidly evolving known protein folds of similar functions [5, 62]. Large-scale initiatives aimed at revealing the structure and function of proteins present in *Mycoplasma* have not only assigned structure-based functions for the majority of the hypothetical proteins but also revealed that nearly 50% of the ORFans tested had unique folds [63].

Recent studies focusing on proteins encoded by the SARS coronavirus revealed that 50% of the proteins for which the structure have been resolved had distinct and novel folds [64]. And one such protein, nsp1, which is presumed to be involved in mRNA degradation, was shown to have an irregular β-barrel fold that was previously unknown among proteins involved in this biological function [65]. Thus, it could be that ORFans also have a role in basic cellular processes but perform these functions by unique mechanisms. Although, it is not possible to define and characterize new structures based solely on CD spectral analysis, future studies focusing on NMR and X-ray crystallography of ORFan proteins, in association with these resolved spectra, can potentially yield distinctive novel folds.

The authors thank Becky Nankivell for help with the preparation of the figures. We acknowledge Frank Aylward for his assistance in screening the clones. This work is supported by Award No. GM-74738 from the National Institutes of Health.

The authors have declared no conflict of interest.

5 References

[1] Hayashi, T., Makino, K., Kurokawa, K., Ishii, K. *et al.*, Complete sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.* 2001, *8*, 11–22.

[2] Lerat, E., Daubin, V., Ochman, H., Moran, N. A., Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol.* 2005, *3*, e130.

[3] Koonin, E. V., Makarova, K. S., Aravind, L., Horizontal gene transfer in prokaryotes: Quantification and classification. *Annu. Rev. Microbiol.* 2001, *55*, 709–742.

[4] Daubin, V., Moran, N. A., Ochman, H., Phylogenetics and the cohesion of bacterial genomes. *Science* 2003, *301*, 829–832.

[5] Fischer, D., Eisenberg, D., Finding families for genomic ORFans. *Bioinformatics* 1999, *15*, 759–762.

[6] Daubin, V., Ochman, H., Bacterial genomes as new gene homes: The genealogy of ORFans in *E. coli*. *Genome Res.* 2004, *14*, 1036–1042.

[7] Welch, R. A., Burland, V., Plunkett, G., III, Redford, P. *et al.*, Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 2002, *99*, 17020–17024.

[8] Yin, Y., Fischer, D., On the origin of microbial ORFans: Quantifying the strength of the evidence for viral lateral transfer. *BMC Evol. Biol.* 2006, *6*, 63.

[9] Siew, N., Fischer, D., Twenty thousand ORFan microbial protein families for the biologists? *Structure* 2003, *11*, 7–9.

[10] Amiri, H., Davids, W., Andersson, S. G., Birth and death of orphan genes in rickettsia. *Mol. Biol. Evol.* 2003, *20*, 1575–1587.

[11] Charlebois, R. L., Clarke, G. D., Beiko, R. G., St Jean, A., Characterization of species-specific genes using a flexible, web-based querying system. *FEMS Microbiol. Lett.* 2003, *225*, 213–220.

[12] Domazet-Loso, T., Tautz, D., An evolutionary analysis of orphan genes in *Drosophila*. *Genome Res.* 2003, *13*, 2213–2219.

[13] Russell, R. B., Domain insertion. *Protein Eng.* 1994, *7*, 1407–1410.

[14] Claverie, J. M., Ogata, H., The insertion of palindromic repeats in the evolution of proteins. *Trends Biochem. Sci.* 2003, *28*, 75–80.

[15] Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T. *et al.*, Origins of highly mosaic mycobacteriophage genomes. *Cell* 2003, *113*, 171–182.

[16] Alimi, J. P., Poirot, O., lopez, F., Claverie, J. M., Reverse transcriptase-polymerase chain reaction validation of 25 "orphan" genes from *Escherichia coli*. *Genome Res.* 2000, *10*, 959–966.

[17] Shimomura, S., Shigenobu, S., Morioka, M., Ishikawa H., An experimental validation of orphan genes of Buchnera, a symbiont of aphids. *Biochem. Biophys. Res. Commun.* 2002, *292*, 263–267.

[18] Shmuely, H., Dinitz, E., Dahan, I., Eichler, J. *et al.*, Poorly conserved ORFs in the genome of the archaea *Halobacterium* sp. NRC-1 correspond to expressed proteins. *Bioinformatics* 2004, *20*, 1248–1253.

[19] Renesto, P., Abergel, C., Decloquement, P., Moinier, D. *et al.*, Mimivirus giant particles incorporate a large fraction of anonymous and unique gene products. *J. Virol.* 2006, *80*, 11678–11685.

[20] Ogata, H., Claverie, J. M., Unique genes in giant viruses: Regular substitution pattern and anomalously short size. *Genome Res.* 2007, *17*, 1353–1361.

[21] Pascal, G., Médigue, C., Danchin, A., Universal biases in protein composition of model prokaryotes. *Proteins* 2005, *60*, 27–35.

[22] Clark, E. H., East, J. M., Lee, A. G., The role of tryptophan residues in an integral membrane protein: Diacylglycerol kinase. *Biochemistry* 2003, *42*, 11065–11073.

[23] Pascal, G., Médigue, C., Danchin, A., Persistent biases in the amino acid composition of prokaryotic proteins. *Bioessays* 2006, *28*, 726–738.

[24] Zhu, K., Jutila, A., Tuominen, E. K., Patkar, S. A. *et al.*, Impact of the tryptophan residues of *Humicola lanuginosa* lipase on its thermal stability. *Biochem. Biophys. Acta* 2001, *1547*, 329–338.

[25] Koshi, J. M., Bruno, W. J., Major structural determinants of transmembrane proteins identified by principal component analysis. *Proteins* 1999, *34*, 333–340.

[26] Tang, C. M., Moxon, E. R., The impact of microbial genomics on antimicrobial drug development. *Annu. Rev. Genomics. Hum. Genet.* 2001, *2*, 259–269.

[27] Siew, N., Saini, H. K., Fischer, D., A putative novel alpha/beta hydrolase family in bacillus. *FEBS Lett.* 2005, *579*, 3175–3182.

[28] Saini, H. K., Fischer, D., Structural and functional insights into mimivirus ORFans. *BMC Genomics* 2007, *8*, 115.

[29] Al-Shahib, A., Breitling, R., Gilbert, D., Predicting protein function by machine learning on amino acid sequences – a critical evaluation. *BMC Genomics* 2007, *8*, 78.

[30] Xu, Q., Krishna, S. S., McMullan, D., Schwarzenbacher, R. *et al.*, Crystal structure of an ORFan protein (TM1622) from *Thermotoga maritima* at 1.75 A resolution reveals a fold similar to the Ran-binding protein Mog1p. *Proteins* 2006, *15*, 777–782.

[31] Chin, K. H., Ruan, S. K., Wang, A. H. J., Chou, S. H., XC5848, an ORFan protein from *Xanthomonas campestris*, adopts a novel variant of Sm-like motif. *Proteins* 2007, *68*, 1006–1010.

[32] Portugaly, E., Linial, M., Estimating the probability for a protein to have a new fold: A statistical computational model. *Proc. Natl. Acad. Sci. USA* 2000, *97*, 5161–5166.

[33] Siew, N., Fischer, D., Structural biology sheds light on the puzzle of genomic ORFans. *J. Mol. Biol.* 2004, *342*, 369–373.

[34] Shin, D. H., Hou, J., Chandonia, J. M., Das, D. *et al.*, Structure-based inference of molecular functions of proteins of unknown function from berkeley structural genomics center. *J. Struct. Funct. Genomics* 2007, *8*, 99–105.

[35] Chen, Y. H., Yang, J. T., Martinez, H. M., Determination of the secondary structures of proteins by circular dichroism and optical rotary dispersion. *Biochemistry* 1972, *11*, 4120–4131.

[36] May, G., Dersch, P., Haardt, M., Middendorf, A., Bremer, E., The *osmZ* (*bglY*) gene encodes the DNA-binding protein H-NS (H1a), a component of the *Escherichia coli* K-12 nucleoid. *Mol. Gen. Genet.* 1990, *224*, 81–90.

[37] Boyd, D., Weiss, D. S., Chen, J. C., Beckwith, J., Towards single-copy gene expression systems making gene cloning physiologically relevant: Lambda InCh, a simple *Escherichia coli* plasmid-chromosome shuttle system. *J. Bacteriol.* 2000, *182*, 842–847.

[38] Sorek, R., Zhu, Y. W., Creevey, C. J., Francino, M. P. *et al.*, Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 2007, *318*, 1449–1452.

[39] Selinger, D. W., Cheung, K. J., Mei, R., Johansson, E. M. *et al.*, RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nat. Biotechnol.* 2000, *18*, 1262–1268.

[40] Osterhout, R. E., Figueroa, I. A., Keasling, J. D., Arkin, A. P., Global analysis of host response to induction of a latent bacteriophage. *BMC Microbiol.* 2007, *7*, 82.

[41] Franchini, A. G., Egli, T., Global gene expression in *Escherichia coli* K-12 during short-term and long-term adaptation to glucose-limited continuous culture conditions. *Microbiology* 2006, *152*, 2111–2127.

[42] Bergholz, T. M., Wick, L. M., Weihong, Qi., Riordan, J. T. *et al.*, Global transcriptional response of *Escherichia coli* O157:H7 to growth in glucose minimal medium. *BMC Microbiol.* 2007, *7*, 97.

[43] Yee, A., Chang, X., Pineda-Lucena, A., Wu, B. *et al.*, An NMR approach to structural proteomics. *Proc. Natl. Acad. Sci. USA* 2002, *99*, 1825–1830.

[44] Maxwell, K. L., Bona, D., Liu, C., Arrowsmith, C. H., Edwards, A. M., Refolding out of guanidine hydrochloride is an effective approach for high-throughput structural studies of small proteins. *Protein Sci.* 2003, *12*, 2073–2080.

[45] Vogt, J., Schulz, G. E., The structure of the outer membrane protein OmpX from *Escherichia coli* reveals possible mechanisms of virulence. *Structure* 1999, *7*, 1301–1309.

[46] Otto, K., Hermansson, M., Inactivation of *ompX* causes increased interactions of type I fimbriated *Escherichia coli* with abiotic surfaces. *J. Bacteriol.* 2004, *186*, 226–234.

[47] Motiejunaite, R., Armalyte, J., Markuckas, A., Suziedeliene, E., *Escherichia coli dinJ-yafO* genes acts as toxin–antitoxin module. *FEMS Microbiol. Lett.* 2007, *268*, 112–119.

[48] Rinas, U., Bailey, J. E., Overexpression of bacterial hemoglobin causes incorporation of preβ-lactamase into cytoplasmic inclusion bodies. *Appl. Environ. Microbiol.* 1993, *59*, 561–566.

[49] Carrio, M. M., Villaverde, A., Construction and deconstruction of bacterial inclusion bodies. *J. Biotechnol.* 2002, *96*, 3–12.

[50] Baneyx, F., Mujacic, M., Recombinant protein folding and misfolding in *Escherichia coli*. *Nature Biotechnol.* 2004, *22*, 1399–1408.

[51] Gonzalez-Montalban, N., Carrio, M. M., Cuatrecasas, S., Aris, A. *et al.*, Bacterial inclusion bodies are cytotoxic *in vivo* in absence of functional chaperones DnaK or GroEL. *J. Biotechnol.* 2005, *118*, 406–412.

[52] Ventura, S., Villaverde, A., Protein quality in bacterial inclusion bodies. *Trends Biotechnol.* 2006, *24*, 179–185.

[53] Worrall, D. M., Gross, N. H., The formation of biologically active beta-galactoside inclusion bodies in *Escherichia coli*. *Aust. J. Biotechnol.* 1989, *3*, 28–32.

[54] Tokatlidis, K., Dhurjati, P., Millet, J., Beguin, P. *et al.*, High activity of inclusion bodies formed in *Escherichia coli* overproducing *Clostridium thermocellum* endoglucanase D. *FEBS Lett.* 1991, *282*, 205–208.

[55] Lees, J. G., Miles, A. J., Wien, F., Wallace, B. A., A reference database for circular dichroism spectroscopy covering fold and secondary structure space. *Bioinformatics* 2006, *22*, 1955–1962.

[56] Wolf, Y. L., Brenner, S. E., Bash, P. A., Koonin, E. V., Distribution of protein folds in the three superkingdoms of life. *Genome Res.* 1999, *9*, 17–26.

[57] Hegyi, H., Lin, J., Greenbaum, D., Gerstein, M., Structural genomics analysis: Ch aracteristics of atypical, common, and horizontally transferred folds. *Proteins: Struct. Funct. Genet.* 2002, *47*, 126–141.

[58] Jackson, S. E., How do small single-domain proteins fold? *Fold Des.* 1998, *3*, R81–R91.

[59] Raymond, S., Tocili, A., Ajamian, E., Li, Y. *et al.*, Crystal structure of ureidoglycolate hydrolase (AllA) from *Escherichia coli* O157:H7. *Proteins* 2005, *61*, 454–459.

[60] Mira, A. H., Ochman, H., Moran, N. A., Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 2001, *10*, 589–596.

[61] Pallen, M. J., Wren, B. W., Bacterial pathogenomics. *Nature* 2007, *449*, 835–842.

[62] Yin, Y., Fischer, D., Identification and investigation of ORFans in the viral world. *BMC Genomics* 2008, *9*, 24.

[63] Kim, S. H., Shin, D. H., Liu, J., Oganesyan, V. *et al.*, Structural genomics of minimal organisms and protein fold space. *J. Struct. Funct. Genomics* 2005, *6*, 63–70.

[64] Bartlam, M., Xu, Y., Rao, Z., Structural proteomics of the SARS coronavirus: A model response to emerging infectious diseases. *J. Struct. Funct. Genomics* 2007, *8*, 85–97.

[65] Almeida, M. S., Johnson, M. A., Herrmann, T., Geralt, M., Wuthrich, K., Novel β-barrel fold in the nuclear magnetic resonance structure of the replicase nonstructural protein 1 from the severe acute respiratory syndrome coronavirus. *J. Virol.* 2007, *81*, 3151–3161.

