Folded Alpha Helical Putative New Proteins from *Apilactobacillus kunkeei*

Weihua Ye [1,†], Phani Rama Krishna Behra [2,†], Karl Dyrhage [2], Christian Seeger [2], Joe D. Joiner [1], Elin Karlsson [1], Eva Andersson [1], Celestine N. Chi [1,‡,*], Siv G. E. Andersson [2,*] and Per Jemth [1,*]

*1 - **Department of Medical Biochemistry and Microbiology,** Uppsala University, BMC Box 582, 75123 Uppsala, Sweden*

*2 - **Department of Molecular Evolution,** Cell and Molecular Biology, Biomedical Centre, Science for Life Laboratory, Uppsala University, 75236 Uppsala, Sweden*

***Correspondence to Celestine N. Chi, Siv G.E. Andersson and Per Jemth:** celestine.chi@astrazeneca.com (C.N. Chi), siv.andersson@icm.uu.se (S.G.E. Andersson), Per.Jemth@imbim.uu.se (P. Jemth)*

https://doi.org/10.1016/j.jmb.2024.168490
Edited by Sarel Fleishman

Abstract

The emergence of new proteins is a central question in biology. Most tertiary protein folds known to date appear to have an ancient origin, but it is clear from bioinformatic analyses that new proteins continuously emerge in all organismal groups. However, there is a paucity of experimental data on new proteins regarding their structure and biophysical properties. We performed a detailed phylogenetic analysis and identified 48 putative open reading frames in the honeybee-associated bacterium *Apilactobacillus kunkeei* for which no or few homologs could be identified in closely-related species, suggesting that they could be relatively new on an evolutionary time scale and represent recently evolved proteins. Using circular dichroism-, fluorescence- and nuclear magnetic resonance (NMR) spectroscopy we investigated six of these proteins and show that they are not intrinsically disordered, but populate alpha-helical dominated folded states with relatively low thermodynamic stability (0–3 kcal/mol). The NMR and biophysical data demonstrate that small new proteins readily adopt simple folded conformations suggesting that more complex tertiary structures can be continuously re-invented during evolution by fusion of such simple secondary structure elements. These findings have implications for the general view on protein evolution, where *de novo* emergence of folded proteins may be a common event.

Introduction

New genes arise by duplication and divergence,[1] exon rearrangements and gene fusion/fission events in which protein domains encoded by already existing genes are reused for new or modified functions.[2,3] Expansions of functional repertoires by duplication and domain shuffling events are commonly observed in protein families involved in signal transduction pathways as well as in gene regulation and transport systems. More recently, it has been shown that new genes can arise *de novo* from non-coding sequences,[4–6] as first demonstrated for genes involved in male reproduction in fruit flies.[7,8] While the majority of genes encoding new proteins are lost because they do not confer a fitness advantage to the organism, some of the new proteins may be subject to positive selection and their genes retained in the genome of future generations. Furthermore, evolutionary experiments have demonstrated that randomly synthesized short open reading frames can confer new function, for example antibiotic resistance[9,10] or rescue of auxotrophy[11] in *Escherichia coli*.

However, despite these examples of gene birth from non-coding DNA, the mechanisms and frequencies with which new protein-coding genes are generated *de novo,* and in particular the structure and function of the new proteins, remain largely unknown. It is widely accepted that the origin of most of the recognized protein folds in present-day organisms are ancient, and were present in the last universal common ancestor.[12–14] However, there is a growing body of data showing that new proteins constantly emerge in living organisms.[15] It is conceptually very important to understand whether *de novo* emergence of protein domains and structural convergence[16,17] are common or rare events. A key question is whether *de novo* proteins can fold into well-defined structures and converge on existing protein folds, or if new proteins are predominately intrinsically disordered as suggested by predictions.[18]

We address this question by structural characterization of putative small new proteins expressed from small open reading frames (smORFs) in *Apilactobacillus kunkeii*, a defensive symbiont of honeybees that is highly abundant in the honey crop and the honeybee food products.[19–24] The *A. kunkeii* isolates have genome sizes ranging from 1.49 to 1.64 Mb (mean 1.57 Mb), and gene contents ranging from 1345 to 1504 genes (mean 1430).[25] A study of 104 closed genomes indicated that the population has an open pan-genome structure, meaning that the number of new chromosomal genes increases with the addition of every genome.[25] However, the mechanism whereby these new genes have originated is not known, nor is it known if the new genes are temporary residents in the genome or code for proteins that confer new functions to the bacterial cell.

We have here investigated in detail a subset of *A. kunkeii* smORFs from the perspective of protein structure. We find that these small and potentially new proteins are not intrinsically disordered but instead adopt simple alpha-helical folds, as shown by circular dichroism and NMR spectroscopy. Further evolution of such small structured proteins, for example by fusion of their genes, could generate larger folded proteins with implications for the general thinking about emergence of novel protein folds.

Results

Identification and expression of recently emerged open reading frames

The starting point for this analysis was the 1466 predicted genes in the genome of *A. kunkeei* strain A0901. A blastp search against the NR database (2023–01-10) showed that 137 genes had fewer than 100 hits to species outside *A. kunkeei* (E-value below 1e-3) (Figure S1a). Of these, 48 short open reading frames with a minimum length of 93 bp and an average length of ~ 337 bp showed no hits or only a few (max five) hits to hypothetical genes in bacteria with taxonomic names other than *A. kunkeei* (Table S1, Figure S1b), indicating that they may have emerged recently.

We attempted expression of 15 of these genes from *A. kunkeei* strain A0901 in an *E. coli* vector and were able to purify five of the encoded proteins. For simplicity, we refer to these open reading frames as well as their expressed proteins as smORFs: smORF5 (AKUA0901_04910), smORF7 (AKUA0901_04830), smORF8 (AKUA0901_04820), smORF9 (AKUA0901_04570), and smORF12 (AKUA0901_01190) (Table S1). Three smORFs (smORF5, smORF7, smORF8) are located within or near a putative phage region and code for very short proteins of 49 to 56 amino acids (Figure S1c). smORF9 is located near to the *cas* gene cassette upstream of the phage region and codes for a protein of 153 amino acids, while smORF12 is located elsewhere and potentially codes for a 66-residue protein (Figure S1c). In addition, we included a short, truncated gene of known origin that encodes a protein of similar size to the smORFs and is a fragment of an IS-element, smORF_IS (AKUA0901_13330). Thus, six proteins were selected for bioinformatics analyses and biophysical experiments.

To investigate whether the six smORFs are expressed in *A. kunkeei* under laboratory settings, we collected two proteomics datasets from *A. kunkeei* strain A0901. One dataset was obtained following sampling of bacterial cells during both exponential and stationary growth phases (below referred to as the “log-vs-stat” dataset). The other dataset was obtained following growth under stressful conditions, which were induced by different concentrations of ciprofloxacin that causes replication stalling (below referred to as the “CPX” dataset). Using LC-MS/MS based proteomics analyses, we experimentally identified 712 proteins in the log-vs-stat dataset and 864 proteins in the CPX-dataset. Inspection of the data for the 48 smORFs with no or few hits to other species showed that 11 proteins were identified in at least one dataset (Table S2). These included smORF7 and smORF8, which were identified in both datasets as well as smORF5 and smORF9, which were identified in the CPX-dataset. No protein could be detected for smORF12 in either dataset, but it was nonetheless selected for further studies because of its apparent non-phage origin. Since the sequence of the smORF_IS is present in multiple copies in the genome its expression pattern cannot be assessed, because the peptides obtained in the LC-MS/MS analyses do not distinguish between matches to smORF_IS and matches to all other copies of the IS-elements.

Taxonomic distribution profiles of the smORFs

We obtained a few hits to taxa with names other than *A. kunkeei* in the BLAST-based analyses (summarized in Table S3a). For example, both smORF9 and smORF12 have homologs in *Apilactobacillus zhangqiuensis* with sequence identity values in the range of 74 % to 95 % (E-values of 2.8e-39 for smORF12 and 5.56e-79 for smORF9 (Table S1). To infer the relatedness of these taxa to *A. kunkeei,* we calculated average nucleotide sequence identity values (ANI) for all pairwise genome comparisons (Table S3b). Most taxa for which hits were obtained showed ANI values above 97 % and should be considered subspecies of *A. kunkeei* (e.g., *Apilactobacillus nanyangensis, Apilactobacillus waqari, Apilactobacillus sp.* F1*, Lactobacillus sp. M0345* strain).[26] A few taxa, such as *A. zhanggiuensis*, showed ANI values of about 90–92 % and should be considered different species within the genus *Apilactobacillus.*

In attempt to identify more distantly related homologs to the smORFs that might have been missed in the BLAST-based analyses, we also used hidden Markov models to search for putative homologs in the UniProt and NR databases (Table S1). When smORF8 was used as the query, a few weak hits were obtained to proteins in *Lacticaseibacillus rhamnosus* (Chen et al. 2013) and *Lactiplantibacillus plantarum*. The putative homologs were of similar lengths as smORF8 and displayed 35–45 % amino acid sequence identity. The homolog in *L. rhamnosus* was located within a putative phage region, but no similarity to *A. kunkeei* genes was observed for any of the other genes in the phage region, suggesting that the similarity is not due to infections of the same phage. The search using hidden Markov models also revealed hits to metagenome sequences classified as *Staphylococcus epidermis* in the NR database when smORF5 and smORF9 were used as the query sequences*.* However, the metagenome sequences were obtained from the gut of *A. mellifera* (Meng et al. 2022) and the hits were nearly identical to the smORF5 and smORF9 gene sequences, respectively. It therefore seemed likely that the hits were derived from *A. kunkeei*, although the contigs on which they were located had been classified as *S. epidermis.* The search using hidden Markov models thus largely confirmed the BLAST-based analyses, based on which we conclude that the selected smORFs display a narrow taxonomic distribution pattern and are restricted to *A. kunkeei* or to species within the genus *Apilactobacillus*, with the only possible exception of smORF8, which may have homologs in a few other *Lactobacillus* genomes.

Next, we examined the phyletic distribution patterns of the five smORFs within the 104 *A. kunkeei* isolates for which closed genome data is available (Figure S2). To illustrate the results, we selected a smaller set of 34 isolates that represent the phylogenetic diversity of the 104 *A. kunkeei* isolates[25] (Figure 1). However, strain H3B1-11M, which was previously used as the reference strain for a clade of 12 isolates, did not contain any smORF and was substituted for strain H4B5-02X from the same clade, which contained both smORF5 and smORF7. We also included H3B2-03J from this clade because it contained both of these two genes as well as smORF8. Likewise, H3B2-06M, which was used as the reference strain for a clade of seven isolates, was substituted for strain H4B2-10M, which was the only isolate in that clade that contained both smORF7 and smORF8.

The analyses showed that the putative phage genes, smORF5 and smORF7, were present in 18 of the 34 representative strains, of which 4 strains also contained smORF8 (Figure 1; Figure S2). The strains containing the phage region were derived from the A, B, C and E phylogroups, suggesting multiple independent phage integration/excision events. This hypothesis was further supported by the finding that clades consisting of *A. kunkeei* isolates with otherwise identical genomes differed with regard to the occurrence of the phage genes. In contrast to these scattered presences, smORF9 was identified in all strains except *A. kunkeei* strains MP2 and A1404. Finally, smORF12 was identified in all of the 34 reference strains. Thus, the selected smORFs included genes that showed a scattered phyletic distribution pattern as well as genes that were broadly present in the *A. kunkeei* population.

Gene order structures and gene sequence evolution

To learn more about the forces and mechanisms that have driven the evolution of the selected smORFs, we examined the conservation of their protein sequences and gene order structures (Figure 2**;** Figures S3-S6). Notably, smORF7 represents the first gene in a long stretch of putative phage genes, including smORF5, all of which are located in the same transcriptional direction as smORF7 (Figure 2). As such, smORF7 is likely to correspond to the first gene in a phage operon. Interestingly, smORF8 is located immediately upstream of smORF7 but in the converse transcriptional direction, and is only present in four reference strains. A comparison of all reference genomes that contained the phage genes showed that the region upstream of smORF7 is highly variable in gene content in most genomes. However, smORF8 is flanked by the exact same set of genes in the few genomes that contain this gene. This suggests that smORF8 is located within a genomic segment that represents an insertion event that is distinct from the phage integration event.

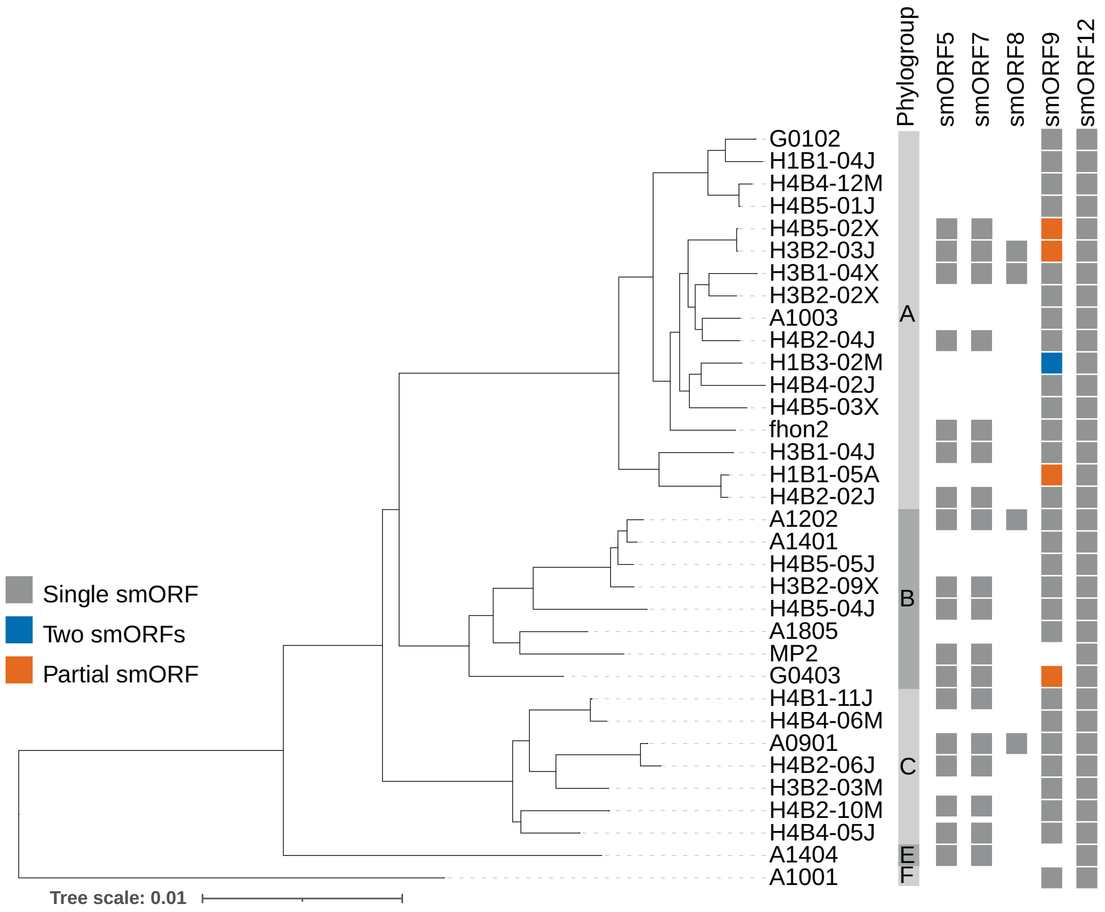


Figure 1. Phyletic distribution patterns of the selected smORFs. The presence and absence pattern of the selected smORFs is shown for 34 representative *A. kunkeei* isolates. Filled boxes represent the presence of a smORF homolog whereas empty boxes indicate that no sequence homolog could be detected in that particular strain. Phylogroup classifications of strains is indicated (Dyrhage et al. 2022).

The sequence alignments revealed only a few amino acid differences between the reference strains in the N- and C-termini of the smORF7 protein sequences. Likewise, only a few protein sequence variants were noted for smORF5 in the reference strains, with the exception of isolate H4B5-4J, which contained a longer gene at this site (Figure S3a). The smORF8 amino acid sequences were identical in all four reference strains, and these also contained a unique amino acid in the N-terminal end of the smORF7 protein which was not found in any of the other strains and did not correspond to another variable amino acid site at the C-terminal end (Figures S3b, S3c). This is indicative of an integration/recombination event that span across the smORF7 and smORF8 gene borders, with a break-point within the smORF7 gene.

Multiple sequence alignment of the smORF9 homologs revealed the presence of two different sequence variants that differed at multiple sites (Figure S4b). Protein variant I, which is present in *A. kunkeei* strain A0901, was additionally found in most strains of phylogroup C as well as those belonging to phylogroup F, whilst protein variant II was most common in isolates from phylogroups A and B. The start codon was AUG in the gene of variant I, but CUG in the gene of variant II, although the latter also contained an AUG codon located further downstream that may alternatively function as the start site. A few isolates that contained gene variant II contained a short deletion of eighteen nucleotides, multiple nucleotide substitutions and a single nucleotide deletion that disrupted the open reading frame. Another two isolates, Fhon2 and H1B1-05A contained a truncated C-terminal end of the smORF9 gene caused by an internal stop codon (Figure S4b). Thus, smORF9 is less conserved than the others and may be undergoing deterioration in some strains. smORF12, on the other hand, was highly conserved in amino acid sequence but gene order structures varied widely (Figure S5a, b). We also noted that sequences for IS-elements were present at the corresponding site as smORF_IS_A0901 in seven other isolates,

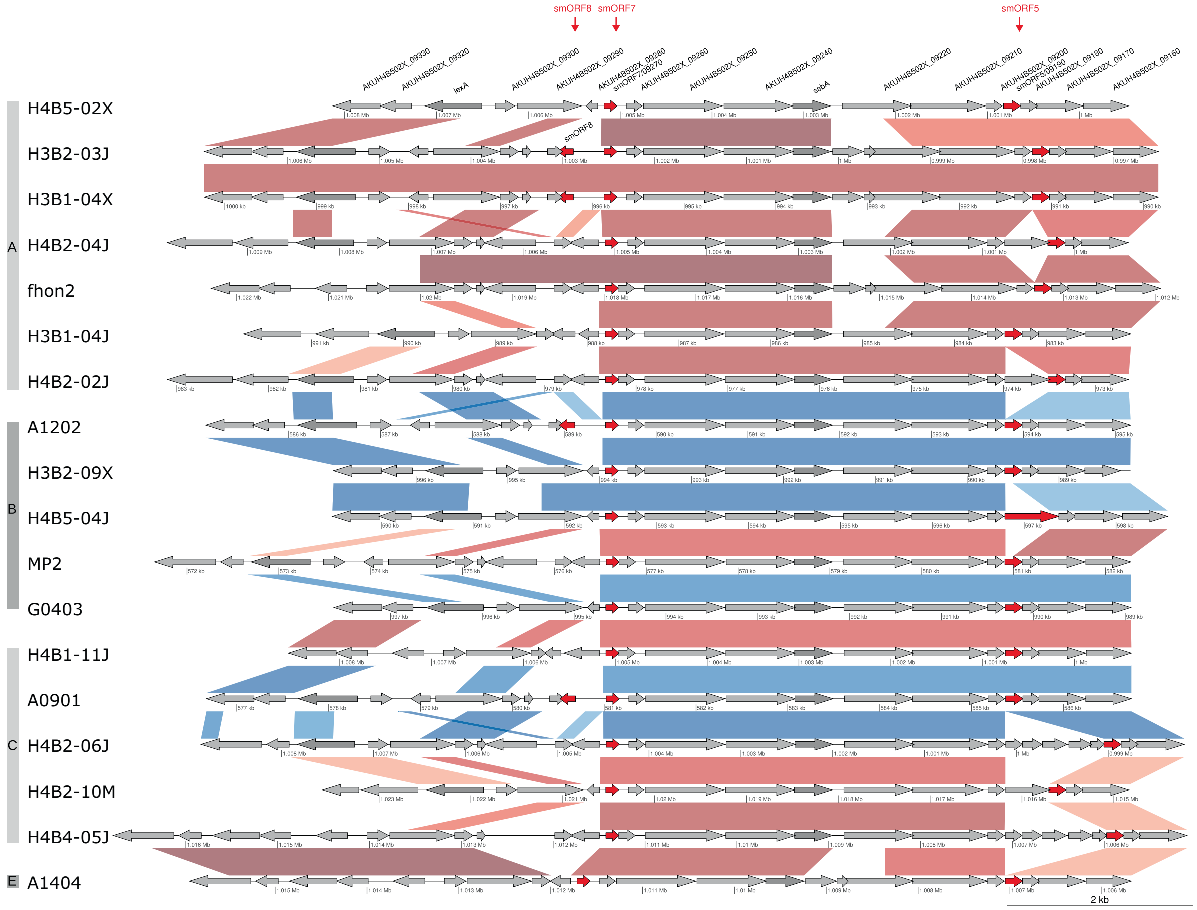


Figure 2. Gene order structures of smORFs located in a phage region. The figure shows a comparison of gene order structures for smORF5, smORF7 and smORF8 in a subset of the representative *A. kunkeei* strains. The smORFs are highlighted with red arrows, while genes marked in grey and dark grey represent hypothetical and functionally annotated protein-coding genes, respectively. The connecting vertical lines between any two strains show genes and segments with high levels of sequence similarity. The red and blue colors of these lines show segments located at the same or different chromosomal sites, respectively. Phylogroup classifications of strains is indicated (Dyrhage et al. 2022).

three of which were of similar lengths and had almost identical amino acid sequences as smORF_IS_A0901 (Figure S6a, b). Thus, the amino acid sequences of the selected smORFs are highly similar in all *A. kunkeei* isolates, in the range of 90 % identity, or 5 amino acid differences per 50 residues.

To determine whether the high levels of amino acid similarities reflect purifying selection, we estimated nonsynonymous (dN) and synonymous substitution (dS) frequencies for all smORFs in all possible pairwise strain combinations (Tables S4a-j). For this analysis, we selected smORFs that had been clustered into the same protein family by OrthoMCL and were in most cases encoded by full-length gene sequences. The dS values were consistently below 1 substitution per site, and thus showed no signs of being saturated for substitutions. Importantly, the dN/dS ratios were about or less than 0.1 for the smORFs, similar to the average dN/dS calculated for a set of 1154 core genes (Tables S4k-l), as expected for genes that evolve under selection for a function. However, some pairs of smORF were identical in nucleotide sequences although the core genes for the same pair of strains displayed considerable divergence at their synonymous sites. For example, we observed no nucleotide differences between any of the smORF8 gene sequences although these were derived from genetically divergent strains that belong to three different phylogroups. We interpret these results to suggest that recombination events have homogenized the nucleotide sequences of some smORFs.

The new proteins possess secondary structure

To assess the secondary structure of the proteins, we performed far-UV circular dichroism

(CD) experiments at 25 °C (Figure 3a). smORF5, smORF7, smORF12 and smORF_IS displayed a typical alpha helical signature with two minima around 210 and 222 nm, whilst smORF8 and smORF9 displayed CD spectra consistent with mixed secondary structure (Table S5). One of the proteins, smORF12, appeared as partially unfolded as judged by the CD spectrum. CD spectra at 90 °C were consistent with a lower degree of secondary structure as compared to 25 °C, suggesting that all six proteins contained secondary structure distinct from a random coil under physiological temperature. smORF9, which is the largest of the proteins, displayed the lowest molar ellipticity, suggesting it may contain more regions lacking secondary structure as compared to the other five proteins.

To further investigate whether the smORF proteins adopt a folded structure, we determined their thermodynamic stability with increasing denaturant concentration (urea or guanidinium chloride, GdmCl) (Figure 3b) or temperature (Figure 3c), respectively. Proteins with well-defined tertiary structures and a hydrophobic core typically unfold in a cooperative, sigmoidal fashion. In such cases, the slope of the transition

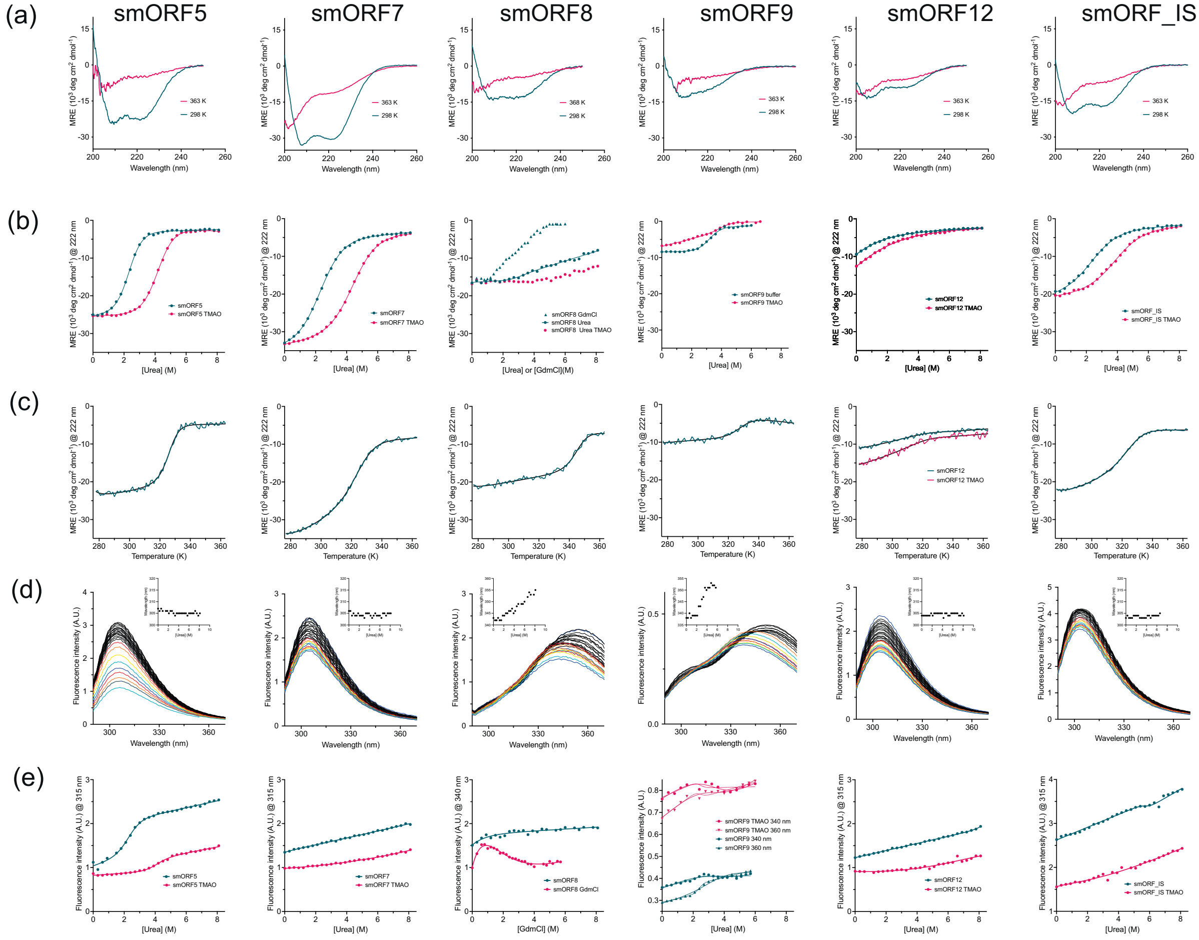


Figure 3. Secondary structure and stability of the new proteins. (a) Far-UV circular dichroism spectra show that the smORF proteins predominantly adopt helical structures. (b) Denaturation experiments using urea or GdmCl monitored by circular dichroism at 222 nm, where alpha helical secondary structure gives a strong negative signal. A two-state model was fitted to the experimental data. (c) Heat denaturation experiments monitored by circular dichroism at 222 nm. (d) Fluorescence emission spectra following excitation at 274 nm or 280 nm (smORF8 and smORF9) at different urea concentrations. The arrows indicate the red shift of the emission maximum as the protein unfolds when the urea concentration is increased (smORF8 and smORF9). The insets show how the maximum wavelength increases with urea concentration. Note the linear increase for smORF8 consistent with the non-cooperative urea denaturation shown in the next panel (e). The shoulder around 310 nm for smORF9 is the Raman peak. (e) Denaturation experiments using urea or GdmCl monitored by fluorescence emission at the indicated emission wavelengths. A sigmoidal shape indicates cooperative (un)folding of the protein in panels (b), (c) and (e). Fitted parameters from all experiments are shown in Table S5. Experiments were performed in 50 mM sodium phosphate, pH 7.4. TMAO, trimethylamine *N*-oxide, a stabilizer of tertiary structure.

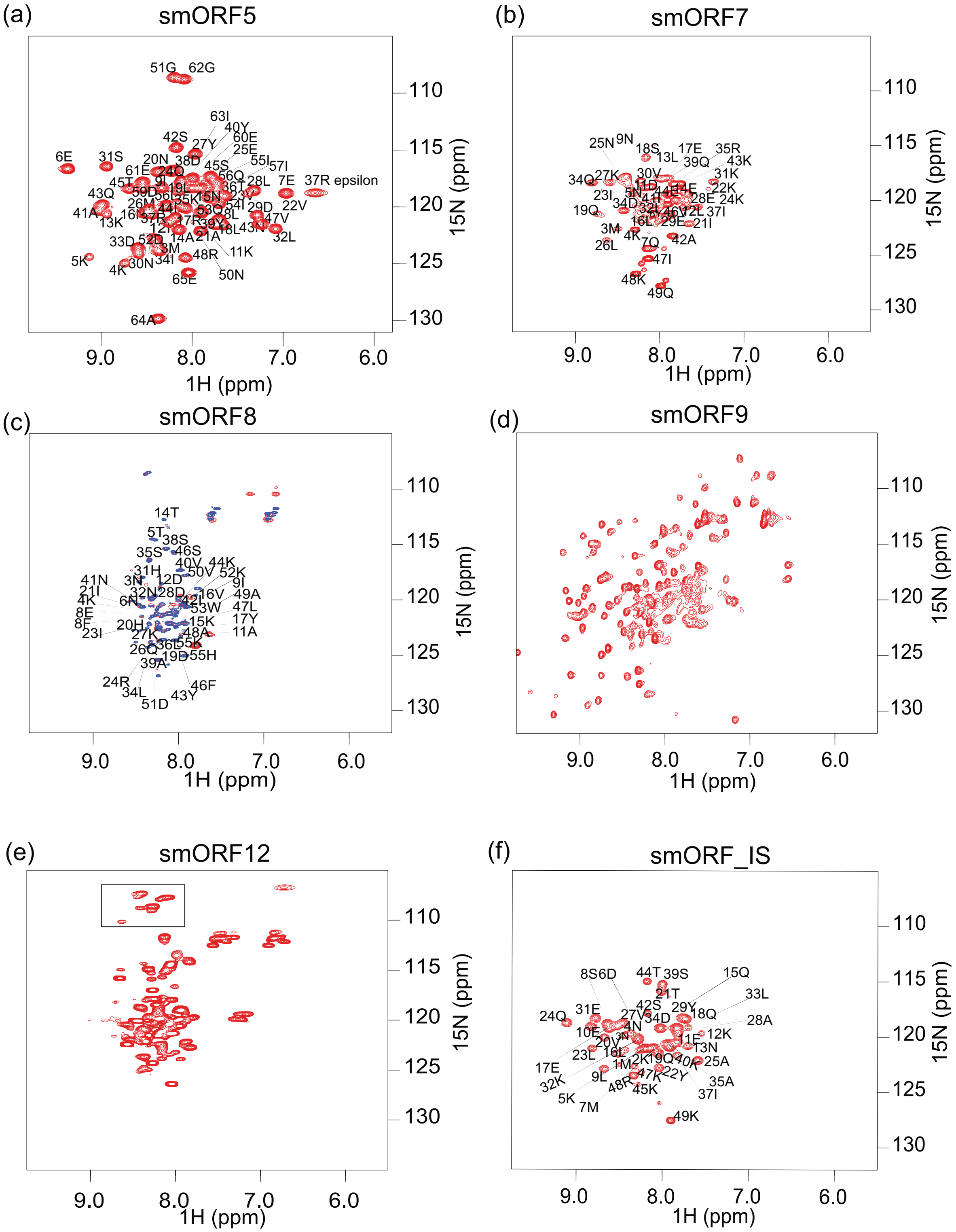
region ($m_{\text{D-N}}$ value) is related to the difference in solvent accessible surface area upon unfolding. Thus, a larger protein will generally have a higher $m_{\text{D-N}}$ value than a smaller protein. Moreover, if the unfolding transition is reversible, the thermodynamic stability of the protein can be estimated by assuming an apparent two-state scenario, where the native folded state and the denatured unfolded state are the only dominating molecular species under all conditions (temperature or denaturant concentration). Urea denaturation experiments, performed both in the presence and absence of the stabilizing molecule trimethylamine *N*-oxide (TMAO), and monitored by far-UV CD at 222 nm, showed clear cooperative unfolding transitions for five of the six proteins tested, corroborating the presence of secondary structure under native conditions (Figure 3b). The transition of smORF12 did not contain a native baseline, confirming that the protein is partially unfolded at 25 °C. The urea denaturation transition of smORF8 was very broad and incomplete, and therefore, we used the stronger denaturant GdmCl, which resulted in a non-cooperative broad unfolding transition. This behavior is consistent with a multi-step unfolding mechanism involving intermediates present at equilibrium. While smORF9 displayed a single cooperative transition in buffer without TMAO, addition of this stabilizing agent resulted in a broad transition suggesting that a more complex (un)folding process was induced by TMAO. Temperature denaturation experiments recapitulated the denaturant experiments, further demonstrating the presence of secondary structure in the native state (Figure 3c). Here, smORF8 displayed an apparent cooperative transition with a relatively high thermal midpoint (72 °C) for such a small protein domain.

Next, we performed urea denaturation experiments and monitored fluorescence, which probes gross changes in tertiary structure from the changes in the environment of Trp and Tyr residues upon denaturation. Typically, a change from hydrophobic to solvent exposed environment results in a red shift of the maximum emission wavelength. However, only smORF8 and smORF9 displayed a clear shift in emission maximum, suggesting that the aromatic residues in the other proteins do not experience an increase in solvent exposure upon denaturation and, thus, are likely not part of a hydrophobic core (Figure 3d). Consistent with this result, smORF7, smORF12 and smORF_IS did not display any visible transition when the urea denaturation was monitored at 315 nm (the emission wavelength of Tyr residues). This result could either be interpreted as a lack of globular structure in the native state, or that the Tyr residues are solvent exposed in the native state (as is the case for smORF7, see below). smORF5 displayed a cooperative transition, whereas smORF8 again showed a non-cooperative behavior upon addition of GdmCl consistent with a multi-step unfolding (emission at 340 nm due to presence of one Trp). smORF9 displayed a cooperative transition and, like for the CD data, addition of TMAO complicated the urea dependences with a switch from increase to decrease of the fluorescence upon unfolding at high (360 nm) but not low (340 nm) emission wavelength. The presence of two Trp residues in the sequence of smORF9 underlies the increase or decrease of fluorescence at different emission wavelengths.

Thermodynamic stabilities of the new proteins are in the typical range of small domains

Naturally occurring protein domains are usually marginally stable with free energies of folding ($\Delta G_{\text{D-N}}$) in the range 2–5 kcal mol^{-1}. In order to assess the thermodynamic stability of the *A. kunkeii* smORFs, the urea and temperature denaturation data were analyzed according to a two-state (un)folding scenario where the native and denatured states are assumed to be in equilibrium. From denaturant-induced unfolding data, $\Delta G_{\text{D-N}}$ is calculated from two parameters obtained from fitting a two-state mechanism: the midpoint of denaturation ([Urea]$_{50\%}$) and the $m_{\text{D-N}}$

Figure 4. ^{1}H-^{15}N HSQC spectra and backbone assignments of the new proteins. For spectra with amino acid assignment, the side-chain amide resonances of Asn and Gln are omitted for clarity. (a) Complete assignment of all ^{1}H-^{15}N pairs for smORF5. (b) Complete assignment of all ^{1}H-^{15}N pairs except for proline residues. (c) The assignment of smORF8 was performed for the acid-denatured state at pH 4.0 (blue). The spectrum of native smORF8 at pH 6.0 (red) did not show resonances for all amino acids. However, most of the resonances at pH 4.0 were also visible at pH 6.0 indicating that a state, similar to the acid-denatured one, is present under physiological conditions. (d) The ^{1}H-^{15}N HSQC spectrum of smORF9 shows that it is well folded and has beta strands because of the spread of the resonances. (e) The ^{1}H-^{15}N spectrum of smORF12 indicates that it displays structural heterogeneity. 50 % of the amino acids were visible and some contain multiple resonances. For example, the resonances above 8 ppm (^{1}H) or below 110 ppm (^{15}N) (denoted by the rectangle) correspond to the three HN-N pairs of three glycines present in the protein. In a single conformational state only three resonances are supposed to be visible but here six are visible. (f) Complete ^{1}H-^{15}N assignment of all amino acids in smORF_IS except prolines. The spectrum shows that most peaks exist only in a single state, suggesting one main conformation of smORF_IS.


(a)
smORF5
51G
62G
63I
40Y
60E
42S
27Y
25E
6E
31S
20N
38D
45S
55I
57I
61E
24Q
56Q
28L
7E
37R epsilon
45T
9I
36T
43Q
59D
19I
23V
56L
35K
15N
26M
44P
54I
29D
22V
41A
53Q
8L
16N
37R
47V
13K
17R
39Y
12I
18L
43N
32L
33D
52D
14A
21A
3M
11K
5K
48R
34I
30N
4K
50N
65E
64A
1H (ppm)
15N (ppm)
9.0
8.0
7.0
6.0
110
115
120
125
130
(b)
smORF7
25N
9N
18S
13L
17E
35R
39Q
43K
34Q
27K
30V
31K
23I
5N
11D
41I
14E
44E
22K
28E
24K
34D
32I
6V
46V
12L
37I
19Q
16L
29E
21I
3M
4K
42A
26L
7Q
47I
48K
49Q
1H (ppm)
15N (ppm)
9.0
8.0
7.0
6.0
110
115
120
125
130
(c)
smORF8
14T
5T
38S
35S
46S
40V
44K
31H
50V
52K
41N
3N
12D
9I
21I
32N
28D
49A
4K
6N
42I
16V
8E
53W
47L
8F
20H
15K
17Y
23I
27K
48A
11A
26Q
36L
55K
19D
55H
39A
24R
46F
34L
43Y
51D
1H (ppm)
15N (ppm)
9.0
8.0
7.0
6.0
110
115
120
125
130
(d)
smORF9
1H (ppm)
15N (ppm)
9.0
8.0
7.0
6.0
110
115
120
125
130
(e)
smORF12
1H (ppm)
15N (ppm)
9.0
8.0
7.0
6.0
110
115
120
125
130
(f)
smORF_IS
8S6D
44T
39S
15Q
21T
33L
42S
29Y
24Q
31E
27V
34D
18Q
28A
4N
10E
3N
11E
12K
20V
13N
23L
16L
1M
2K
19Q
40K
25A
17E
9L
47K
22Y
32K
48R
35A
45K
5K
37I
7M
49K
1H (ppm)
15N (ppm)
9.0
8.0
7.0
6.0
110
115
120
125
130

value, which is related to the change in solvent accessible surface area upon unfolding. In temperature denaturation experiments, $\Delta G_{\text{D-N}}$ is calculated from three fitted parameters: the thermal midpoint (T_m), the change in enthalpy of unfolding ($\Delta H_{\text{D-N}}$) and the change in heat capacity (ΔC_p) (Table S5). While midpoints can usually be determined accurately, the other parameters may have large errors for small proteins where the transitions are broad and baselines not well defined. (In particular, ΔC_p is very uncertain, but $\Delta G_{\text{D-N}}$ is not very sensitive to errors in this parameter.) Nevertheless, data from the different experiments were overall consistent and show that the smORF proteins display thermodynamic stabilities that are in the same range as typical protein domains of a similar size (Table S5). $\Delta G_{\text{D-N}}$ values derived from CD-monitored urea and temperature-induced denaturation agreed fairly well for smORF5 (2.5–3.3 kcal mol^{-1}), smORF7 (1.4–1.9 kcal mol^{-1}), smORF9 (4.3–4.7 kcal mol^{-1}) and smORF_IS (1.4–1.7 kcal mol^{-1}). smORF8 displayed non-cooperative transitions in denaturant-induced experiments but a clear cooperative transition in temperature denaturation ($\Delta G_{\text{D-N}}$ = 3.2 ± 1.6 kcal mol^{-1}). smORF9 displayed a transition towards non-cooperative unfolding in presence of TMAO, as monitored both by CD and fluorescence. The thermodynamic parameters ($m_{\text{D-N}}$ and $[\text{Urea}]_{50\%}$) from fluorescence-monitored denaturation were particularly non-consistent for smORF9, with and without TMAO. Association of monomers into dimers or higher order quaternary structure may underlie the observed non-cooperativity, both for smORF8 and smORF9. smORF12 was the least stable of the proteins and populated the native state to only around 50 % at 25 °C ($\Delta G_{\text{D-N}}$ = -0.6 to 1.2 kcal mol^{-1}).

The new proteins display simple tertiary structure

To obtain more detailed structural information we expressed and purified the six proteins as single (^{15}N) or double (^{13}C/^{15}N) labeled samples and subjected them to nuclear magnetic resonance (NMR) spectroscopy experiments. ^{1}H-^{15}N heteronuclear quantum coherence (HSQC) spectra (Figure 4) were well-dispersed, allowing for high resolution NMR spectroscopy analysis. From the NMR experiments, including dihedral angles, ^3J couplings constants and distance information obtained from nuclear Overhauser effects (NOEs), we were able to perform complete backbone and sidechain NMR assignments and determine the 3D structures of smORF5, smORF7 and smORF_IS at 25 °C in 50 mM sodium phosphate pH 6.5 (Figures 4 and 5**)**. The conformations of smORF8 and smORF12 were heterogeneous under these conditions, *i.e.*, the proteins populate more than one conformation, which is consistent with thermodynamic data. smORF9 was poorly expressed in minimal medium and we could not get sufficient amounts for 3D NMR experiments. Nevertheless, our NMR experiments revealed that the smORFs mainly contain helical structures in agreement with the CD data.

To probe the nature of the helices, we determined the hydrogen bond lengths between the amide proton (HN) and carbonyl oxygen (CO) atom for smORF5, smORF7 and smORF_IS using chemical shifts of HN from the ^{1}H-^{15}N HSQC experiment (Figure 4). In alpha helices, a hydrogen bond is formed between HN of amino acid *i* with the carbonyl oxygen of an amino acid at the *i* + 4 position. These bond lengths range between 2.8–3.2 Å. Hydrophobic pairs accumulate in the interior of the protein resulting in shorter hydrogen bond lengths while hydrophilic pairs on the exterior have longer hydrogen bond lengths. This arrangement will result in a 3–4 periodicity repeat and a curvature of the helix with the longer hydrogen bonds outside and the shorter ones inside. On the other hand, a mixture of hydrophobic-hydrophilic pairs will result in an average bond length similar across the chain. We found that the helices in smORF5, smORF7 and smORF_IS displayed the 3–4 repeat periodicity and the bond lengths alternating from short to long and longer stretches of shorter hydrogen bond lengths indicating bent structures.[27] smORF5 folds into two kinked parallel helices with the N- and the C-termini very close to each other. We observed several long-range NOEs (Figure S7 and S8), making the structure of smORF5 the most well defined of the three we determined. smORF7 appears disc-shaped with a diameter of approximately 20 Å. Only a few long-range NOEs were observed for smORF7 (Figure S9 and S10), which is typical for a circular helical structure.[28] On the other hand, smORF_IS folds into a single helix with intrinsically disordered N- and C-termini. Interestingly, smORF_IS contains several long-range NOEs between the ordered region and both the disordered N-terminus and C-terminus (Figure S11 and S12). More specifically, the ordered helix stretches between residues Glu17-Lys32. We observed NOEs between Tyr29 and Gln38, Tyr29 and Glu31, Asn13 and Ala35, Asn3 and Lys32, Lys2 and Lys32, Leu30 and Asp34, Tyr29 and Leu33, and Leu9 and Lys32, suggesting that the disordered N- and C-termini make transient interactions with the ordered central helix.

NMR relaxation parameters harbor several indicators of the behavior of the protein such as flexibility, shape and size. For example, NMR relaxation parameters such as longitudinal and transverse times, T1 and T2, respectively, give dynamic information in the ps to ms timescale but can also be used to estimate the molecular size. The ratio of the relaxation times (T1/T2) gives an

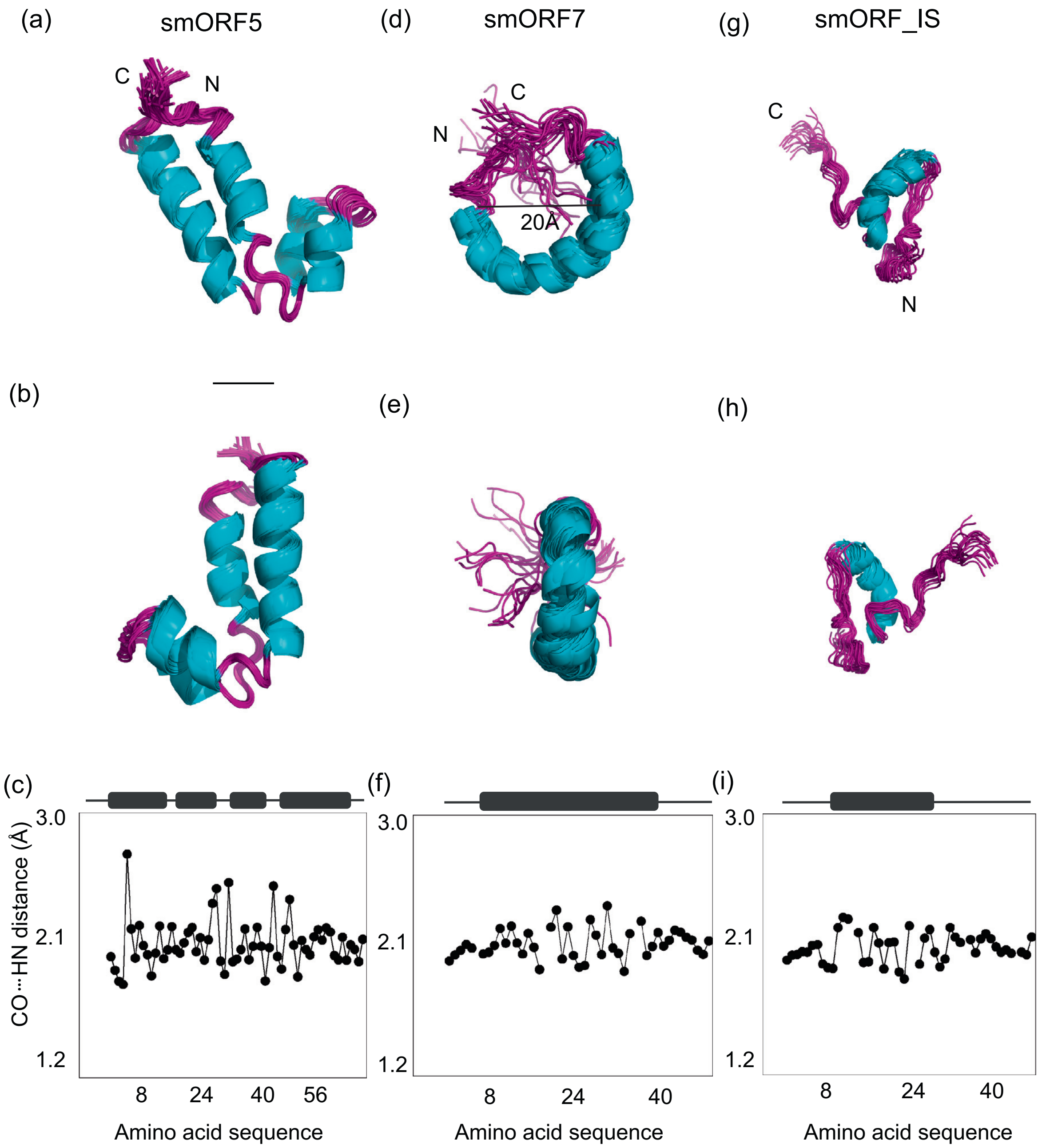


Figure 5. Three dimensional structures of three smORFs and hydrogen bond lengths in their helices. Helices are shown as cyan and loops in purple. (a) and (b) Cartoon representation of smORF5 shows a four-helix bundle including a helix-loop-helix motif, with the N and C termini close to each other. (c) amide (HN) - carbonyl oxygen (CO) bond length determined from ^{1}H-^{15}N chemical shifts of smORF5. Firstly, the bond lengths follow a repetitive pattern of four amino acids, a feature typical for helices. Secondly, the bond lengths are not uniform along the sequence indicating that the helices are not straight but bent. (d) and (e) The NMR data of smORF7 are consistent with a disc-like fold. Firstly, the data indicate that the protein is well folded, but we observed only a few long-range NOEs. Secondly, the secondary structure propensity predicted from the chemical shifts indicates a single helix. Thirdly, the CO-HN bond lengths determined from HN-H spectra (f) suggest that the helix is not straight. Together these observations point to a circular structure, similar to that of membrane associated proteins of so-called nano discs, and consistent with short-range, but not long-range, NOEs. Furthermore, T1/T2 relaxation data suggest that smORF7 is a dimer. Finally, the N- and C-termini are not well defined. This could be partly due to the few structural restraints, but also to a high flexibility in this region. The diameter of the disc is measured to 20 Å. (g) and (h) Cartoon representation of smORF_IS. This protein adopts a single short helix, which makes long-range interactions with parts of the unstructured N- and C-termini. (i) The CO-HN bond pattern for smORF_IS indicates that the folded helix is not straight.

estimate of the total rotational correlation time (τ_C), which is directly proportional to molecular size and can give a good estimate of molecular mass (Figure 6). The theoretical τ_C values calculated from the molecular masses approximated to 4.9 ns (smORF5, 7.4 kDa), 3.9 ns (smORF7, 5.9 kDa) and 6.6 ns (smORF_IS, 9.9 kDa). The experimental τ_C values were 10.9 ns (smORF5), 10.4 ns (smORF7) and 6.3 ns (smORF_IS). Thus, the T1/T2 relaxation data are most consistent with dimeric quaternary structures of smORF5 and smORF7, and a monomeric smORF_IS.

NMR data collected for smORF8 at pH 6.0 indicated that the protein populated at least two states, where one might be the denatured state under physiological conditions. However, in light of the non-cooperative unfolding transitions, as monitored by CD (Figure 3b) and fluorescence (Figure 3e), the observed heterogeneity may result from two or more folded states in equilibrium. CD experiments showed that smORF8 is unfolded at pH 4.0. Consistent with this, NMR data at pH 4.0 suggested a single state representing the acid-denatured state of the protein. For smORF12, the observed heterogeneity in the NMR experiments is most likely due to its marginal stability, with around 50 % denatured state under native conditions, as suggested by the CD-monitored denaturation data, which show the second part of an apparently cooperative unfolding transition (Figure 3b).

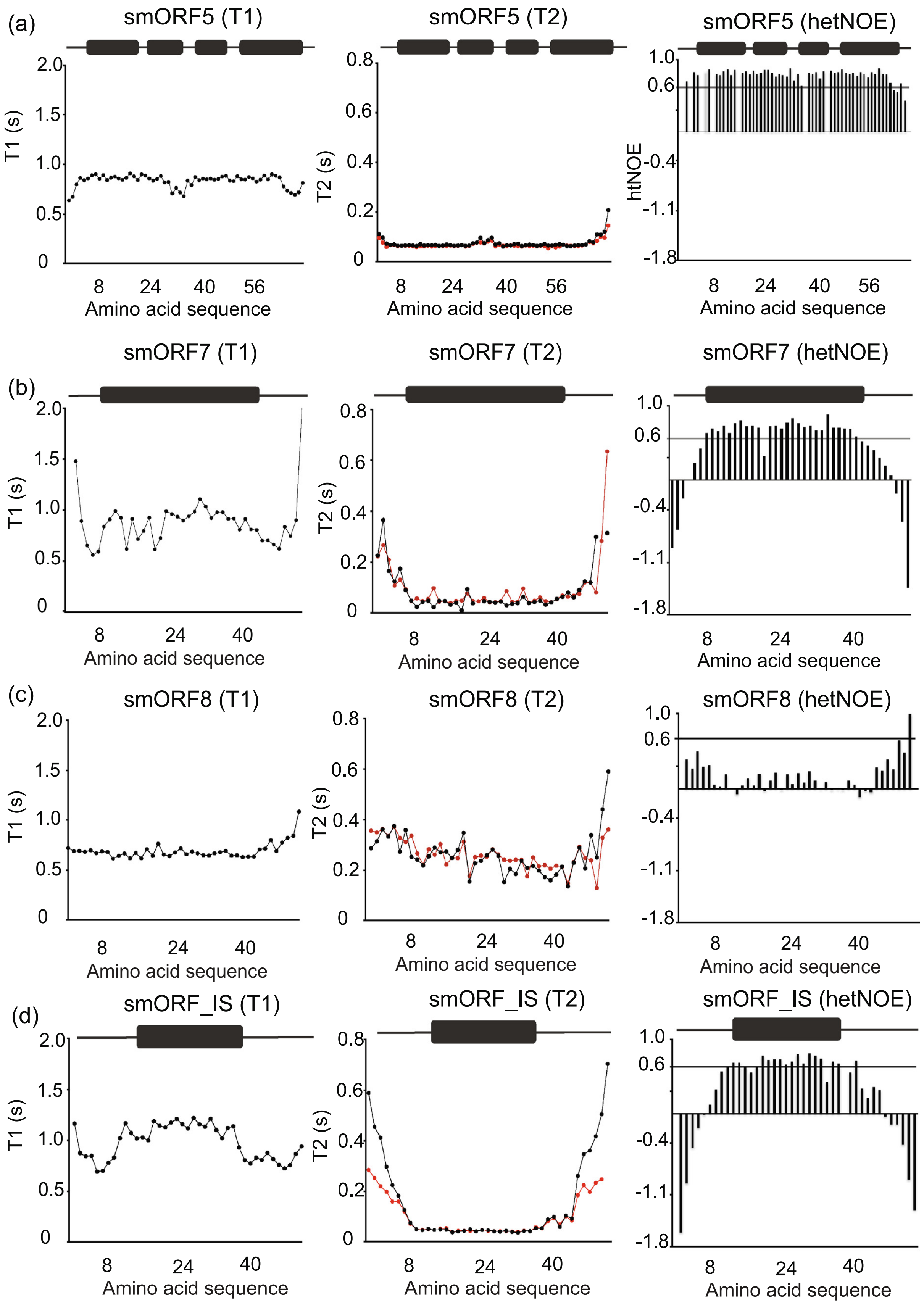
Discussion

Convergent evolution of phenotypic traits is ubiquitous in nature.[29] On the other hand, molecular evolution is generally viewed as a result of divergence as shown for the ancient nucleotide-binding Rossmann fold.[30] In the latter view, contingency is important since new function arises from point mutation or recombination of existing genes encoding a particular unique protein state.[31,32] Expressed proteins with a new or modified function would then be based on ancestral protein folds, and fine-tuned by adaptation according to changes in selection pressure over evolutionary timescales. However, four key findings during the last decades suggest that emergence of entirely new proteins may be more common than previously thought: (*i*) Intrinsically disordered proteins without a well-defined 3D structure play pivotal roles in the cell[33] and are suggested to be more prevalent in young genes.[18] (*ii*) Non-coding regions are frequently transcribed and translated resulting in small proteins with unknown functions.[4] (*iii*) New proteins from random libraries can provide a selective advantage in experimental *in vivo* systems.[9–11] (iv) Globular protein domains can emerge from fusion of small structural motifs.[14,34–41] Thus, a consensus is developing that new proteins constantly emerge from novel open reading frames in living organisms.[15,42] Usually, the genes encoding these new proteins erode and become non-functional with time, but sometimes they prove beneficial for the organism, such that they are retained in the genome by positive selection. From a protein structure–function point-of-view, a fundamental question is what these new proteins look like. Is it possible that protein folds appear *de novo* such that convergent evolution occur on a molecular structure level, as observed for organismal phenotypes? Predictions suggest that new proteins are largely intrinsically disordered,[18] but there is a paucity of experimental data regarding their structures.

To shed light on this question we investigated in detail six small proteins from *A. kunkeii* strain A0901, and where five appear to represent young and putative new proteins. The genes were randomly sampled from a larger subset of genes that could be expressed in *E. coli* and for which no homologs were identified outside the genus *Apilactobacillus* based on a blast search against the NR database. Two genes, smORF9 and smORF12, were successfully identified in all but two *A. kunkeei* strains, suggesting that they represent novel, functional genes, as also supported by significantly lower nonsynonymous compared to synonymous substitution frequencies.

The other smORFs showed a more scattered phyletic distribution pattern. For example, smORF8 was only identified in 11 *A. kunkeei* genomes. The surrounding gene order structure was conserved in these genomes, with the smORF8 gene being flanked by phage genes on one side. Searches based on hidden Markov models yielded hits to *Lacticaseibacillus*

Figure 6. NMR relaxation parameters for the new proteins. The relaxation times and heteronculear NOEs (hetNOE) for (a) smORF5, (b) smORF7, (c) smORF8 and (d) smORF_IS. T1 represents the longitudinal relaxation times while T2 represents the transverse relaxation times. hetNOE is the heteronuclear NOE between the amide (HN) hydrogen and the directly bonded nitrogen atom. The combined behavior of these relaxation parameters gives information about ps-ms timescale motion of the protein. The ratio of T1/T2 is unitless, but can be used as an approximation of the total rotational correlation time τ_C in nanoseconds, which in turn is related to the molecular mass of the protein.


(a)
smORF5 (T1)
smORF5 (T2)
smORF5 (hetNOE)
(b)
smORF7 (T1)
smORF7 (T2)
smORF7 (hetNOE)
(c)
smORF8 (T1)
smORF8 (T2)
smORF8 (hetNOE)
(d)
smORF_IS (T1)
smORF_IS (T2)
smORF_IS (hetNOE)
T1 (s)
T2 (s)
htNOE
2.0
1.5
1.0
0.5
0
0.8
0.6
0.4
0.2
-0.4
-1.1
-1.8
8
24
40
56
Amino acid sequence

rhamnosus and a few other species when smORF8 was used as the query, raising the possibility that this gene and perhaps a few of the nearby genes may have been acquired by horizontal gene transfer. Interestingly, the smORF8 genes were identical in sequence in the four reference strains, displaying no substitutions at either nonsynonymous or synonymous sites. We interpret these observations to suggest that smORF8 has spread via recombination within the *A. kunkeei* population rather recently. One of the breakpoints may be located within smORF7, since this protein contains a unique amino acid replacement in the N-terminal end in all strains that also contained smORF8. This specific residue (proline) was not present in any other strain and a single amino acid replacement at the C-terminal end of the protein showed a different phyletic distribution pattern.

Another two smORFs that displayed a scattered phyletic distribution pattern were located within a previously identified prophage region of 35–50 kb. Of these, smORF7 was located at the 5' end of the phage region, near to smORF8, while smORF5 was located further downstream within the phage region. Different regions of the bacterial genome may have different likelihood to give birth to new genes, depending on differences in mutation rates, recombination frequencies, proximity to transposons and promoter sequences, etc. For all of these reasons, we believe that prophage regions may serve as bacterial testbeds for the evolution of novel genes. Thus, it may not be a coincidence that several of the genes in *A. kunkeei* for which no homologs were found were located within or near to phage regions. In the future, large-scale, in-depth comparative studies of bacterial and bacteriophage genomes are needed to determine the role that phages play for the origin of novel genes in bacteria.

We find that the new smORF proteins and smORF_IS from *A. kunkeii* populate simple, mainly alpha-helical folds. Note that there may be an experimental bias here, where alpha helical proteins express better than other ones and therefore dominate our selection of six proteins. Five of the proteins are marginally stable, like numerous small protein domains found across organismal proteomes. (Note that the temperature of the natural habitat of *A. kunkeii*, a bee gut, may greatly vary over the course of a day and over seasons, and it is not known if and how such variation in temperature affects protein evolution and thermodynamic stability.) Two of the proteins, smORF5 and smORF7, are likely present as dimers based on NMR relaxation data. smORF12 is partially unfolded under physiological temperatures, but displays a cooperative transition to a fully unfolded state. smORF7 displays clear cooperative unfolding indicative of a small globular protein, but it has very dynamic N- and C-termini, which nonetheless contribute to the tertiary structure. smORF8 displays rather peculiar biophysical characteristics including a relatively high thermal stability and structural heterogeneity, unlike globular folds of similar size. smORF9 behaves as a two-state folder with less helical content than the other proteins and addition of TMAO induces folding heterogeneity, perhaps by stabilizing cryptic non-native conformational states. Interestingly, we note that none of the five new smORF proteins or smORF_IS behave as a *bona fide* intrinsically disordered protein, but instead display well-defined secondary and simple tertiary structure. A recent high-throughput study, which used limited proteolysis to assess structure in putative *de novo* human and *Drosophila melanogaster* proteins, found evidence for both disordered and ordered proteins in the data set.[43] Our small set of six smORFs may be strongly biased, but there could also be a difference between eukaryotic and bacterial proteins since disorder is less prevalent among the latter. We note that small *de novo* proteins selected for function from random libraries are predicted to be helical,[9–11] but that the yeast *de novo* protein Bsc4 contains β sheets as judged by circular dichroism.[44]

Small new proteins such as the ones investigated here appear to be prevalent across life.[45] Moreover, the symmetry of many extant protein folds suggest that fusion of small structural motifs is a common pathway for emergence of larger proteins.[37] The potential mechanism and underlying selective forces were investigated for a lectin β-propeller protein, which consist of ancestral motifs shorter than 50 amino acid residues. Interestingly, a reconstruction of the evolutionary trajectory suggested that constraints related to folding governed the process,[46] which raises the possibility of structural convergence in new proteins. The new proteins from *A. kunkeii* are based on helical and possibly mixed alpha/beta secondary structure and they display variation in stability, dynamics and quaternary structure. Intrinsic disorder is less prevalent than might be expected *a priori*. Our data suggest that foldable structural motifs may arise continuously in living cells. Such motifs could act as intermediates on a trajectory towards larger *de novo* globular proteins by duplication and fusion of the genes encoding the motifs. Indeed, evolution of protein domains from accretion of smaller sub domains represents a likely mechanism for invention of new domains during origin of life.[14,16,47,48] Our smORFs may represent such sub domains in an ongoing emergence of *de novo* protein domains, which incidentally could converge on ancient folds. We find it likely that selection for foldability, as part of functionality, may evolve new proteins into similar structures as ancient protein folds. Thus, our present findings and previous data[37,46] suggest that convergent evolution of protein structure is a realistic possibility.

Methods

Bioinformatic analyses

The encoded protein sequences of all 1466 predicted genes in *A. kunkeei* strain A0901 were used as queries in blastp searches (version 2.11.0 +) against the NCBI NR database (2023-01-10), using an E-value of 0.001 and default parameters,[25,49,50] excluding self-hits to NCBI-taxonomy IDs “148814”, “1419324”, “1423768”. The 48 genes with the least hits were then used as queries in more sensitive searches for distantly related homologs based on hidden Markov models, using two different methods.[51] First, the phmmer method was used to search the UniProtKB database[52] with the aid of the online hmmer web-server version 2.41.2[53,54] using default parameters. In this search, we selected all taxa (all organisms) in the database with the exception of *A. kunkeei* and set the E-value cutoff to 0.001. Secondly, the hmm-search method was used to search the NCBI NR database (2023-06-30)[55] with the aid of the MPI Bioinformatics toolkit[51,56,57] using default parameters. Pairwise amino acid identities were calculated between smORF sequences and putative homologs found with hmmer using the tool infoalign (EMBOSS version 6.6.0.0).[58]

Identification of orthologous protein families was performed by using the selected protein sequences in *A. kunkeei* strain A0901 as queries in tblastn searches against all other *A. kunkeei* genomes. Manual inspection of gene order structures helped identify additional related sequences, most of which contained nonsense-mutations and had therefore not been predicted as protein coding genes. Multiple sequence alignments of full-length and partial protein sequences were performed with the aid of ClustalW version 2.1[59] and visualized with the tool Jalview version 2.11.2.6.[60] For calculations of substitution frequencies, only smORFs coding for proteins that clustered together with OrthoMCL were considered.[61] The protein sequences were aligned with MAFFT version 7.520[62,63] with the –auto parameter, and the results were used for codon-based nucleotide alignments with PAL2NAL version 14.[64] Nonsynonymous (dN) and synonymous (dS) substitution frequencies were calculated using yn00 from the PAML software suite version 4.10.6.[65,66] The chromosome representation of *A. kunkeei* A0901 strain was generated using Circos version 0.69-9.[67] The gene synteny plots were generated with the R-package genoPlotR version 0.8.11.[68]

Proteomic analysis of A. kunkeei A0901 under conditions of replication stalling

We collected two proteomics datasets from *A. kunkeei* strain A0901 to investigate whether the predicted smORF genes are expressed under laboratory conditions. The first dataset was obtained by comparing protein expression during exponential and stationary growth phase (“log-vs-stat”). The second dataset was collected during exponential growth phase under the influence of increasing concentrations of ciprofloxacin (“CPX”).

For the log-vs-stat dataset, *A. kunkeei* A0901 was cultivated in MRS broth (Sigma Aldrich) supplemented with Tween-80 and D-Fructose to final concentrations of 0.1 % and 0.5 %, respectively (referred to as fMRS). Cells were cultivated in biological triplicates at 35 °C, 5 % CO2 under static conditions and samples were harvested during exponential (Optical density at 600 nm, $OD_{600} \approx 0.2$, 3.5 h) and stationary growth ($OD_{600} \approx 1.6$, 6 h). Cells were pelleted by centrifugation (3000 *g*, 10 min, 4 °C), the supernatant was discarded and the pellet was washed twice in 50 mM Tris, pH 8.0. For protein extraction, the pellets were re-suspended in 25 mM Tris, pH 8.0, 10 μg/mL lysozyme (Sigma Aldrich), 1x Sigma Fast Protease inhibitors (Sigma Aldrich) to achieve a final cell density of $OD_{600} \approx 10$ and incubated at 37 °C for 5 h under gentle orbital shaking. Additional lysis was achieved by sonication using a Sonics VCX 130 sonicator (Sonics & Materials Inc., 2 mm tip, 4 x 5 s, 5 s pause, on ice). The cell suspension was cleared by centrifugation (16 000 *g*, 10 min, 4 °C) and the supernatant was transferred to 1.5 mL reaction tubes. Total protein concentration was determined by Bradford analysis (Thermo Scientific).

For the CPX-dataset, *A. kunkeei* A0901 was cultivated in fRMS broth in the presence of 0.0, 3.1, 6.3 and 12.5 μg/mL ciprofloxacin (CPX, Sigma Aldrich). Batch-cultivation of biological triplicates per condition was performed under static conditions at 35 °C, 5 % CO_2 and samples were harvested by centrifugation (4500 *g*, 10 min, 4 °C). Pellets were washed twice in HyClone (Cytiva). For protein extraction, washed pellets were re-suspended in 25 mM Tris, pH 8.0, 15 μg/mL Lysozyme and 1x SigmaFast protease inhibitors and incubated at 37 °C under gentle orbital shaking for 2.5 h followed by sonication (VCX 130 sonicator, 4 x 5 s, 5 s pause, 2 mm tip, on ice). The suspension was cleared by centrifugation (16 000 *g*, 10 min, 4 °C) and the supernatant was transferred to 1.5 mL reaction tubes. Protein concentration was determined by using the Bradford assay (Thermo Scientific) with BSA as the standard.

Proteomic analysis of the log-vs-stat and CPX-samples was essentially performed as described previously.[69] Aliquots corresponding to 20 μg protein were taken out for digestion. The proteins were reduced, alkylated and in-solution digested by trypsin according to a standard operating procedure. Thereafter the samples were purified by Pierce C18 Spin Columns (Thermo Scientific) and dried. Dried peptides were resolved in 60 μL of 0.1 % FA and further diluted 4 times (CPX) and 5 times (log-vs-stat) prior to nano-LC-MS/MS. The resulting peptides were separated in reversed-phase on a C18-column, applying a 90 min long gradient, and electrosprayed on-line to a QEx-Orbitrap mass spectrometer (Thermo Finnigan). Tandem mass spectrometry was performed applying HCD. Database searches were performed in the MaxQuant software (version 1.5.1.2).[70,71] Proteins were identified by searching against the annotated genome of *A. kunkeei* A0901.[16] Fixed modification was carbamidomethyl (C), and variable modifications were oxidation (M), and deamidation (NQ). A decoy search database, including common contaminants and a reverse database, was used to estimate the identification false discovery rate (FDR). An FDR of 1 % was accepted for peptides and protein identification. The criteria for protein identification were set to at least two identified peptides per protein. The MS data has been deposited in PRIDE (accession number PDX037237).

Heterologous protein expression and purification

The DNA sequences of all smORF proteins were synthesized and subcloned into a modified pRSET vector by GenScript (Hong Kong). The construct was tailored to have an N-terminal hexa-histidine-tagged lipoyl fusion protein followed by a thrombin cleavage site and the respective smORF protein. The DNA sequences were confirmed by Sanger sequencing (Eurofins Genomics, Uppsala). The smORF8 (AKUA0901_04820) construct encoded only the His-tag and the protein due to problems with thrombin cleavage and purification. *Escherichia coli* BL21(DE3) pLysS cells (Invitrogen) were transformed with the plasmid and grown in 2 × TY medium containing 100 μg/mL ampicillin at 37 °C. At an OD_{600} of around 0.6, expression was induced by 1 mM isopropyl-β-thio galactopyranoside. The cells were then grown overnight at 18 °C, spun down in a centrifuge at 4 °C and resuspended in 20 mM Tris (pH 8.0), 500 mM NaCl. Cells were disrupted by ultrasonication and cell debris was removed by centrifugation at 20,000 *g* for 60 min followed by filtration (0.2 or 0.45 μm). The general purification strategy was as follows: (i) Nickel (II) affinity chromatography (Ni Sepharose 6 Fast Flow, GE Healthcare) in 25 mM Tris (pH 8.0), 500 mM NaCl and 20 mM Imidazole. Bound proteins were eluted by 300 mM or 500 mM imidazole in 25 mM Tris (pH8.0). (ii) Dialysis into 25 mM Tris (pH 8.0), 150 mM NaCl followed by thrombin (GE Healthcare) digestion to remove the lipoyl domain. (iii) A second nickel (II) affinity chromatography step, with 20 mM imidazole included in the washing buffer and where the lipoyl domain binds and the smORF protein is collected in the unbound fraction. (iv) A final purification step involving either ion exchange, size exclusion or reversed phase chromatography. smORF5 (AKUA0901_04910), smORF8 (AKUA0901_04820) (with His tag), and smORF12 (AKUA0901_01190) were purified using a reversed-phase chromatography column (Vydac C8, Grace Davison Discovery Sciences) as the final purification step. The column was equilibrated with 0.1 % trifluoroacetic acid and bound proteins were eluted with a 0–100 % gradient of acetonitrile (0.1 % trifluoroacetic acid). The thrombin-digested smORF7 (AKUA0901_04830) and smORF_IS (AKUA0901_13330) samples were dialyzed against 25 mM Tris (pH 8.0) and loaded onto a Q column (HiTrap Q Fast Flow, GE Healthcare) equilibrated with the same buffer. The smORF proteins eluted in the unbound fraction before the start of a gradient 0–600 mM NaCl in 25 mM Tris (pH 7.5). Although expected to bind an S column, a cation exchanger, neither smORF7 (AKUA0901_04830) or smORF_IS (AKUA0901_13330) did; Q was then used since it bound more impurities. The thrombin-digested smORF9 (AKUA0901_04570) was dialyzed against 25 mM Tris (pH 7.5), 150 mM NaCl, 4 mM DTT, concentrated using Vivaspin columns (Sartorius) and loaded onto a size exclusion chromatography column (S-100, GE Healthcare). Protein purity was checked by SDS-PAGE and the identity of the purified proteins was confirmed by MALDI-TOF mass spectrometry.

Circular dichroism and fluorescence spectroscopy

Circular dichroism (CD) and fluorescence spectroscopy experiments were carried out on a JASCO J-1500 spectrophotometer and with a Peltier temperature control system at temperatures indicated in the figures. All the CD experiments were performed using a 1 mm quartz cuvette. Protein concentrations were 6–25 μM and the buffer 50 mM sodium phosphate, pH 7.4, unless otherwise indicated. CD spectra were averages of three to five individual spectra. For both chemical (urea or GdmCl) and thermal unfolding experiments, the CD signal was monitored at 222 nm, and a scan speed of 1 K min^{-1} was used for thermal denaturation. Fluorescence spectroscopy experiments were performed with protein in 50 mM sodium phosphate, pH 7.4 in a 10 mm quartz cuvette at 25 °C. Emission spectra were recorded with an excitation wavelength of 276 nm for smORF5 (35 μM), smORF7 (50 μM), smORF12 (35 μM) and smORF_IS (35 μM), which lack any Trp residues. An excitation wavelength of 280 nm was used for smORF8 (25 μM) and smORF9 (2.5 μM), which contain one and two Trp residues, respectively. In urea and GdmCl-induced unfolding experiments, the emission at 315 nm (Tyr) or 340 nm (Trp, smORF8 and smORF9) was plotted versus denaturant concentration. The (un)folding experiments were analyzed with GraphPad Prism using the standard equations based on a two-state assumption with only native and denatured state significantly populated at equilibrium.[72]

NMR spectroscopy

NMR spectroscopy experiments were performed on a Bruker 600 MHz NeoAdvance HD spectrometer equipped with a cryogenic TCI probe (CRPHe TR-1H and 19F/13C/15 N 5 mm-EZ). smORF proteins were either single (^{15}N) or double (^{13}C,^{15}N) labeled for NMR experiments, by expression in *E. coli* in M9 minimal medium supplemented with 1 g ^{13}C glucose and/or 1 g ^{15}N ammonium chloride per liter medium. Purification of the labeled proteins were as described for unlabeled proteins. The protein concentration for assignment and subsequent structure determination ranged from 0.5 mM to 2 mM. For smORF5, smORF7, smORF8 and smORF_IS, the following NMR experiments were recorded for backbone and side-chain assignment on a double-labeled ^{13}C-^{15}N protein sample: ^{1}H-^{15}N HSQC, ^{1}H-^{13}C HSQC, HNCACB, HNCoCACB, HBHACoHN, HNCA, HNCoCA, ^{1}H-^{1}H ^{13}C resolved HCCH-TOCSY. For smORF9 only ^{1}H-^{15}N HSQC was recorded on a ^{15}N labeled sample. While for smORF12 ^{1}H-^{15}N HSQC was recorded on a ^{13}C-^{15}N labeled sample. The following were the buffer composition and measuring temperatures for the different protein samples: smORF5, smORF7, smORF9 and smORF_IS were measured in 50 mM sodium phosphate pH 6.5 at 298 K, smORF8 was measured in 50 mM sodium acetate pH 4.0 at 298–315 K and smORF12 in 20 mM sodium phosphate, pH 6.0, at 298 K. Protein samples were supplemented in 10 % D_2O, 0.1 % sodium azide. ^{15}N and ^{13}C resolved ^{1}H-^{1}H NOESY (28 (^{15}N or ^{13}C) $\times$ 256 (^{1}H) $\times$ 2048 (^{1}H, direct)) were measured with mixing times ranging from 70 to 120 ms and used for distance estimation during structure determination. $^{3}J_{\mathrm{HNHA}}$ couplings used for structure calculations were measured with a 3D HNHA type experiment. Phi-Psi dihedral angles were estimated using TALOSN. Structure calculation was done with CYANA 3.98 by simulated annealing in 10,000 steps. A total of 100 conformers were calculated and 20 with the lowest target function were selected for analysis. For relaxation experiments, T1, T2, and heteronuclear NOE (hetNOE), were estimated using standard Bruker pulse programs using randomized relaxation delays of 7–10 durations. The D1 delay was set to 3–5 s. The rotational correlation time τ_C was estimated from the ratio of T1/T2. Data were evaluated with Topspin and the Bruker program DynamicCenter version 2.8.01. All other experiments were processed with Topspin version 4 series. Assignments were performed in the CcpNmr analysis software.

Accession numbers

smORF5: PDB ID 8QNJ, BMRB ID 34864; smORF7: PDB ID 8QNT, BMRB ID 34865; smORF_IS: PDB ID 8QNV, BMRB ID 34866.

CRediT authorship contribution statement

Weihua Ye: Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis, Data curation. **Phani Rama Krishna Behra:** Writing – review & editing, Visualization, Investigation, Formal analysis, Data curation. **Karl Dyrhage:** Investigation, Formal analysis, Data curation, Writing – review & editing. **Christian Seeger:** Investigation, Formal analysis, Data curation. **Joe D. Joiner:** Writing – review & editing, Investigation, Formal analysis. **Elin Karlsson:** Investigation, Formal analysis. **Eva Andersson:** Methodology, Investigation. **Celestine N. Chi:** Writing – review & editing, Visualization, Investigation, Formal analysis, Data curation. **Siv G.E. Andersson:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Conceptualization. **Per Jemth:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Funding acquisition, Conceptualization.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Swedish Research Council (2020-04395 to PJ and 2018-4135 to SGEA), and Knut and Alice Wallenberg foundation (2015.0069 to SGEA and PJ; and 2018.044 and 2020.0305 to SGEA). We used the NMR Uppsala infrastructure, which is funded by the Department of Chemistry - BMC and the Disciplinary Domain of Medicine and Pharmacy. The computations were performed on resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at UPPMAX under project NAISS 2023/22-319 partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2024.168490.

Received 28 September 2023;
Accepted 8 February 2024;
Available online 13 February 2024

Keywords:
De novo protein;
protein evolution;
protein fold;
protein folding;
protein Stability

† Equal contribution.

References

[1]. Soskine, M., Tawfik, D.S., (2010). Mutational effects and the evolution of new protein functions. *Nature Rev. Genet.* **11**, 572–582. https://doi.org/10.1038/nrg2808.

[2]. Andersson, D.I., Jerlström-Hultqvist, J., Näsvall, J., (2015). Evolution of new functions de novo and from preexisting genes. *Cold Spring Harb. Perspect. Biol.* **7**, https://doi.org/10.1101/cshperspect.a017996 a017996.

[3]. Oakley, T.H., (2017). Furcation and fusion: the phylogenetics of evolutionary novelty. *Dev. Biol.* **431**, 69–76. https://doi.org/10.1016/j.ydbio.2017.09.015.

[4]. Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., et al., (2012). Proto-genes and de novo gene birth. *Nature* **487**, 370–374. https://doi.org/10.1038/nature11184.

[5]. Schmitz, J.F., Ullrich, K.K., Bornberg-Bauer, E., (2018). Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nature Ecol. Evol.* **2**, 1626–1632. https://doi.org/10.1038/s41559-018-0639-7.

[6]. Bornberg-Bauer, E., Hlouchova, K., Lange, A., (2021). Structure and function of naturally evolved de novo proteins. *Curr. Opin. Struct. Biol.* **68**, 175–183. https://doi.org/10.1016/j.sbi.2020.11.010.

[7]. Begun, D.J., Lindfors, H.A., Thompson, M.E., Holloway, A.K., (2006). Recently evolved genes identified from Drosophila yakuba and D. erecta accessory gland expressed sequence tags. *Genetics* **172**, 1675–1681. https://doi.org/10.1534/genetics.105.050336.

[8]. Begun, D.J., Lindfors, H.A., Kern, A.D., Jones, C.D., (2007). Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. *Genetics* **176**, 1131–1137. https://doi.org/10.1534/genetics.106.069245.

[9]. Knopp, M., Gudmundsdottir, J.S., Nilsson, T., König, F., Warsi, O., Rajer, F., Ädelroth, P., Andersson, D.I., (2019). De Novo Emergence of Peptides That Confer Antibiotic Resistance. *MBio* **10** https://doi.org/10.1128/mBio.00837-19.

[10]. Knopp, M., Babina, A.M., Gudmundsdóttir, J.S., Douglass, M.V., Trent, M.S., Andersson, D.I., (2021). A novel type of colistin resistance genes selected from random sequence space. *PLoS Genet.* **17**, e1009227. https://doi.org/10.1371/journal.pgen.1009227.

[11]. Babina, A.M., Surkov, S., Ye, W., Jerlström-Hultqvist, J., Larsson, M., Holmqvist, E., Jemth, P., Andersson, D.I., Knopp, M., (2023). Rescue of Escherichia coli auxotrophy by de novo small proteins. *Elife* **12**, e78299. https://doi.org/10.7554/eLife.78299.

[12]. Caetano-Anollés, G., Kim, H.S., Mittenthal, J.E., (2007). The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *PNAS* **104**, 9358–9363. https://doi.org/10.1073/pnas.0701214104.

[13]. Alva, V., Remmert, M., Biegert, A., Lupas, A.N., Söding, J., (2010). A galaxy of folds. *Protein Sci.* **19**, 124–130. https://doi.org/10.1002/pro.297.

[14]. Alva, V., Söding, J., Lupas, A.N., (2015). A vocabulary of ancient peptides at the origin of folded proteins. *Elife* **4**, e09410.

[15]. Van Oss, S.B., Carvunis, A.-R., (2019). De novo gene birth. *PLoS Genet.* **15**, e1008160. https://doi.org/10.1371/journal.pgen.1008160.

[16]. Lupas, A.N., Ponting, C.P., Russell, R.B., (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.* **134**, 191–203. https://doi.org/10.1006/jsbi.2001.4393.

[17]. Gough, J., (2005). Convergent evolution of domain architectures (is rare). *Bioinformatics* **21**, 1464–1471. https://doi.org/10.1093/bioinformatics/bti204.

[18]. Wilson, B.A., Foy, S.G., Neme, R., Masel, J., (2017). Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nature Ecol. Evol.* **1**, 0146. https://doi.org/10.1038/s41559-017-0146.

[19]. Olofsson, T.C., Vásquez, A., (2008). Detection and identification of a novel lactic acid bacterial flora within the honey stomach of the honeybee Apis mellifera. *Curr. Microbiol.* **57**, 356–363. https://doi.org/10.1007/s00284-008-9202-0.

[20]. Vásquez, A., Olofsson, T.C., (2009). The lactic acid bacteria involved in the production of bee pollen and bee bread. *J. Apic. Res.* **48**, 189–195. https://doi.org/10.3896/IBRA.1.48.3.07.

[21]. Anderson, K.E., Sheehan, T.H., Mott, B.M., Maes, P., Snyder, L., Schwan, M.R., Walton, A., Jones, B.M., et al., (2013). Microbial ecology of the hive and pollination landscape: bacterial associates from floral nectar, the alimentary tract and stored food of honey bees (Apis mellifera). *PLoS One* **8**, e83125. https://doi.org/10.1371/journal.pone.0083125.

[22]. Endo, A., Irisawa, T., Futagawa-Endo, Y., Takano, K., du Toit, M., Okada, S., Dicks, L.M.T., (2012). Characterization and emended description of Lactobacillus kunkeei as a fructophilic lactic acid bacterium. *Int. J. Syst. Evol. Microbiol.* **62**, 500–504. https://doi.org/10.1099/ijs.0.031054-0.

[23]. Tamarit, D., Ellegaard, K.M., Wikander, J., Olofsson, T., Vásquez, A., Andersson, S.G.E., (2015). Functionally structured genomes in lactobacillus kunkeei colonizing the honey crop and food products of honeybees and stingless bees. *Genome Biol. Evol.* **7**, 1455–1473. https://doi.org/10.1093/gbe/evv079.

[24]. Anderson, K.E., Ricigliano, V.A., (2017). Honey bee gut dysbiosis: a novel context of disease ecology. *Curr. Opin.*

Insect Sci. **22**, 125–132. https://doi.org/10.1016/j.cois.2017.05.020.

[25]. Dyrhage, K., Garcia-Montaner, A., Tamarit, D., Seeger, C., Näslund, K., Olofsson, T.C., Vasquez, A., Webster, M.T., et al., (2022). Genome evolution of a symbiont population for pathogen defense in honeybees. *Genome Biol. Evol.* **14**, evac153. https://doi.org/10.1093/gbe/evac153.

[26]. Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., Aluru, S., (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Commun.* **9**, 5114. https://doi.org/10.1038/s41467-018-07641-9.

[27]. Zhou, N.E., Zhu, B.Y., Sykes, B.D., Hodges, R.S., (1992). Relationship between amide proton chemical shifts and hydrogen bonding in amphipathic.alpha.-helical peptides. *J. Am. Chem. Soc.* **114**, 4320–4326. https://doi.org/10.1021/ja00037a042.

[28]. Bibow, S., Polyhach, Y., Eichmann, C., Chi, C.N., Kowal, J., Albiez, S., McLeod, R.A., Stahlberg, H., et al., (2017). Solution structure of discoidal high-density lipoprotein particles with a shortened apolipoprotein A-I. *Nature Struct. Mol. Biol.* **24**, 187–193. https://doi.org/10.1038/nsmb.3345.

[29]. Conway Morris, S., (2010). Evolution: like any other science it is predictable, Phil. *Trans. R. Soc. B.* **365**, 133–145. https://doi.org/10.1098/rstb.2009.0154.

[30]. Laurino, P., Tóth-Petróczy, Á., Meana-Pañeda, R., Lin, W., Truhlar, D.G., Tawfik, D.S., (2016). An ancient fingerprint indicates the common ancestry of rossmann-fold enzymes utilizing different ribose-based cofactors. *PLoS Biol.* **14**, e1002396.

[31]. Starr, T.N., Flynn, J.M., Mishra, P., Bolon, D.N.A., Thornton, J.W., (2018). Pervasive contingency and entrenchment in a billion years of Hsp90 evolution. *PNAS* **115**, 4453–4458. https://doi.org/10.1073/pnas.1718133115.

[32]. Xie, V.C., Pu, J., Metzger, B.P., Thornton, J.W., Dickinson, B.C., (2021). Contingency and chance erase necessity in the experimental evolution of ancestral proteins. *Elife* **10**, e67336.

[33]. Wright, P.E., Dyson, H.J., (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nature Rev. Mol. Cell Biol.* **16**, 18–29. https://doi.org/10.1038/nrm3920.

[34]. Eck, R.V., Dayhoff, M.O., (1966). Evolution of the structure of ferredoxin based on living relics of primitive amino Acid sequences. *Science* **152**, 363–366. https://doi.org/10.1126/science.152.3720.363.

[35]. Remmert, M., Biegert, A., Linke, D., Lupas, A.N., Söding, J., (2010). Evolution of outer membrane beta-barrels from an ancestral beta beta hairpin. *Mol. Biol. Evol.* **27**, 1348–1358. https://doi.org/10.1093/molbev/msq017.

[36]. Farías-Rico, J.A., Schmidt, S., Höcker, B., (2014). Evolutionary relationship of two ancient protein superfolds. *Nature Chem. Biol.* **10**, 710–715. https://doi.org/10.1038/nchembio.1579.

[37]. Balaji, S., (2015). Internal symmetry in protein structures: prevalence, functional relevance and evolution. *Curr. Opin. Struct. Biol.* **32**, 156–166. https://doi.org/10.1016/j.sbi.2015.05.004.

[38]. Zhu, H., Sepulveda, E., Hartmann, M.D., Kogenaru, M., Ursinus, A., Sulz, E., Albrecht, R., Coles, M., et al., (2016). Origin of a folded repeat protein from an intrinsically disordered ancestor. *Elife* **5**, e16761. https://doi.org/10.7554/eLife.16761.

[39]. Berezovsky, I.N., (2019). Towards descriptor of elementary functions for protein design. *Curr. Opin. Struct. Biol.* **58**, 159–165. https://doi.org/10.1016/j.sbi.2019.06.010.

[40]. Kolodny, R., Nepomnyachiy, S., Tawfik, D.S., Ben-Tal, N., (2021). Bridging themes: short protein segments found in different architectures. *Mol. Biol. Evol.* **38**, 2191–2208. https://doi.org/10.1093/molbev/msab017.

[41]. Qiu, K., Ben-Tal, N., Kolodny, R., (2022). Similar protein segments shared between domains of different evolutionary lineages. *Protein Sci.* **31**, e4407.

[42]. Andrews, S.J., Rothnagel, J.A., (2014). Emerging evidence for functional peptides encoded by short open reading frames. *Nature Rev. Genet.* **15**, 193–204. https://doi.org/10.1038/nrg3520.

[43]. Heames, B., Buchel, F., Aubel, M., Tretyachenko, V., Loginov, D., Novák, P., Lange, A., Bornberg-Bauer, E., et al., (2023). Experimental characterization of de novo proteins and their unevolved random-sequence counterparts. *Nature Ecol. Evol.* **7**, 570–580. https://doi.org/10.1038/s41559-023-02010-2.

[44]. Bungard, D., Copple, J.S., Yan, J., Chhun, J.J., Kumirov, V.K., Foy, S.G., Masel, J., Wysocki, V.H., et al., (2017). Foldability of a natural de novo evolved protein. *Structure* **25**, 1687–1696.e4. https://doi.org/10.1016/j.str.2017.09.006.

[45]. Storz, G., Wolf, Y.I., Ramamurthi, K.S., (2014). Small proteins can no longer be ignored. *Annu. Rev. Biochem* **83**, 753–777. https://doi.org/10.1146/annurev-biochem-070611-102400.

[46]. Smock, R.G., Yadid, I., Dym, O., Clarke, J., Tawfik, D.S., (2016). De novo evolutionary emergence of a symmetrical protein is shaped by folding constraints. *Cell* **164**, 476–486. https://doi.org/10.1016/j.cell.2015.12.024.

[47]. Salem, G.M., Hutchinson, E.G., Orengo, C.A., Thornton, J.M., (1999). Correlation of observed fold frequency with the occurrence of local structural motifs. *J. Mol. Biol.* **287**, 969–981. https://doi.org/10.1006/jmbi.1999.2642.

[48]. Fernandez-Fuentes, N., Dybas, J.M., Fiser, A., (2010). Structural characteristics of novel protein folds. *PLoS Comput. Biol.* **6**, e1000750.

[49]. Boratyn, G.M., Camacho, C., Cooper, P.S., Coulouris, G., Fong, A., Ma, N., Madden, T.L., Matten, W.T., et al., (2013). BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.* **41**, W29–W33. https://doi.org/10.1093/nar/gkt282.

[50]. NCBI Resource Coordinators, (2012). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **41**, D8–D20. https://doi.org/10.1093/nar/gks1189.

[51]. Eddy, S.R., (2011). Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

[52]. The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., et al., (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489. https://doi.org/10.1093/nar/gkaa1100.

[53]. Finn, R.D., Clements, J., Eddy, S.R., (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37. https://doi.org/10.1093/nar/gkr367.

[54]. Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R., Finn, R.D., (2018). HMMER web server: 2018 update. *Nucleic Acids Res.* **46**, W200–W204. https://doi.org/10.1093/nar/gky448.

[55]. Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., et al., (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26. https://doi.org/10.1093/nar/gkab1112.

[56]. Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., et al., (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J. Mol. Biol.* **430**, 2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007.

[57]. Gabler, F., Nam, S.-Z., Till, S., Mirdita, M., Steinegger, M., Söding, J., Lupas, A.N., Alva, V., (2020). Protein sequence analysis using the MPI bioinformatics toolkit. *Curr. Protoc. Bioinformatics* **72**, e108. https://doi.org/10.1002/cpbi.108.

[58]. Rice, P., Longden, I., Bleasby, A., (2000). EMBOSS: the European molecular biology open software suite. *Trends Genet.* **16**, 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2.

[59]. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I. M., et al., (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948. https://doi.org/10.1093/bioinformatics/btm404.

[60]. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., Barton, G.J., (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191. https://doi.org/10.1093/bioinformatics/btp033.

[61]. Li, L., Stoeckert, C.J., Roos, D.S., (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189. https://doi.org/10.1101/gr.1224503.

[62]. Katoh, K., Misawa, K., Kuma, K., Miyata, T., (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066. https://doi.org/10.1093/nar/gkf436.

[63]. Katoh, K., Standley, D.M., (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780. https://doi.org/10.1093/molbev/mst010.

[64]. Suyama, M., Torrents, D., Bork, P., (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612. https://doi.org/10.1093/nar/gkl315.

[65]. Yang, Z., (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556. https://doi.org/10.1093/bioinformatics/13.5.555.

[66]. Yang, Z., (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591. https://doi.org/10.1093/molbev/msm088.

[67]. Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A., (2009). Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645. https://doi.org/10.1101/gr.092759.109.

[68]. Guy, L., Roat Kultima, J., Andersson, S.G.E., (2010). genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334–2335. https://doi.org/10.1093/bioinformatics/btq413.

[69]. Seeger, C., Dyrhage, K., Mahajan, M., Odelgard, A., Lind, S.B., Andersson, S.G.E., (2021). The subcellular proteome of a planctomycetes bacterium shows that newly evolved proteins have distinct fractionation patterns. *Front. Microbiol.* **12**, https://doi.org/10.3389/fmicb.2021.643045 643045.

[70]. Cox, J., Mann, M., (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnol.* **26**, 1367–1372. https://doi.org/10.1038/nbt.1511.

[71]. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., Mann, M., (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805. https://doi.org/10.1021/pr101065j.

[72]. Fersht, A., (1999). Structure and Mechanism in Protein Science: a Guide to Enzyme Catalysis and Protein Folding. Macmillan.