Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Protein design by fusion: implications for protein structure prediction and evolution

Katarzyna Skorupka,[a] Seong Kyu Han,[b] Hyun-Jun Nam,[b] Sanguk Kim[b]* and Salem Faham[a]*

[a]Department of Molecular Physiology and Biological Physics, University of Virginia School of Medicine, Charlottesville, VA 22093, USA, and [b]Division of Molecular and Life Science, Pohang University of Science and Technology, Pohang, Republic of Korea

Correspondence e-mail: sukim@postech.ac.kr, sf3bb@virginia.edu

Received 5 June 2013
Accepted 12 August 2013

PDB Reference: FliS-FliC fusion, 4iwb

Domain fusion is a useful tool in protein design. Here, the structure of a fusion of the heterodimeric flagella-assembly proteins FliS and FliC is reported. Although the ability of the fusion protein to maintain the structure of the heterodimer may be apparent, threading-based structural predictions do not properly fuse the heterodimer. Additional examples of naturally occurring heterodimers that are homologous to full-length proteins were identified. These examples highlight that the designed protein was engineered by the same tools as used in the natural evolution of proteins and that heterodimeric structures contain a wealth of information, currently unused, that can improve structural predictions.

1. Introduction

Protein domains are discrete units with stable three-dimensional structures (Wetlaufer, 1973). Algorithms for the identification of domain boundaries generally agree in cases of continuous domains; however, for multidomain proteins with discontinuous domains different algorithms can yield different results (Siddiqui & Barton, 1995). Nevertheless, there is an overall agreement that protein domains are compact structural entities that can exist and fold independently of the rest of the protein (Jaenicke, 1987). Protein domains can be used as building blocks to produce more complex multidomain proteins by fusion. This evolutionary mechanism can easily be mimicked experimentally, where protein domains are successfully fused in a myriad of laboratory settings. On the other hand, although evolution has formed a wide array of unique protein domains, engineering novel protein domains experimentally is a challenging task. Indeed, one important milestone achieved in the field of protein design was the production of a new protein fold by purely computational means (Kuhlman *et al.*, 2003).

A variety of examples of protein fusions are available in the PDB. For example, a number of structures of fusions to maltose-binding protein (MBP; Smyth *et al.*, 2003) or to glutathione *S*-transferase (GST; Zhan *et al.*, 2001) can be found. These types of fusions are used to aid in protein solubility and crystallization. The final structures obtained with these fusions were not pre-designed. There are also examples of heterodimers that have been fused successfully in order to aid in crystallization and structure determination (Ye *et al.*, 2006; Hennecke *et al.*, 2000; Yu *et al.*, 2010). These fusions do not generate new distinct domains and similarly were not pre-designed with a specific final structural target. Elegant examples of designed protein fusions include a lysozyme insertion into the β_2 adrenergic receptor (Rosenbaum *et al.*, 2007) and a fusion based on the alignment of the N- and C-terminal helices of different domains using a helical linker

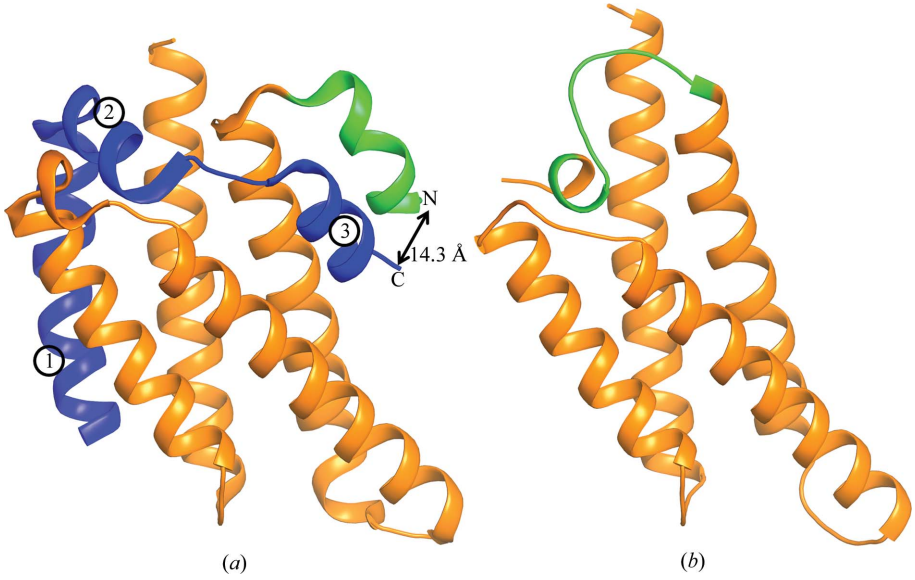


Figure 1
Design of the FliC·FliS fusion construct. (*a*) The structure of the FliS-FliC heterodimer (PDB entry 1ory). The helical bundle of FliS is shown in orange and the N-terminal region is shown in green; FliC is shown in blue. The three FliC helices are numbered. The gap between the C-terminus of FliC and the N-terminus of FliS is indicated by the double-headed arrow and the distance of 14.3 Å that separates them is shown. (*b*) The structure of FliS by itself (PDB entry 1orj). The helical bundle of FliS is shown in orange and the N-terminal region is shown in green. The N-terminal region of FliS clearly adopts two different conformations.

(Padilla *et al.*, 2001). These examples produced the desired multidomain proteins. Fusion has also successfully been used to produce a single domain, although not with a novel fold. By joining two $(\beta\alpha)_4$ half-barrels, a full $(\beta\alpha)_8$ barrel was obtained (Höcker *et al.*, 2004). Similarly, a chimeric design simulating a recombination event resulted in the reproduction of the desired $\beta\alpha$ barrel with an unexpected additional β-strand (Bharat *et al.*, 2008).

The proximity of N- and C-termini has been taken advantage of in cases of fusion-based protein engineering. For example, tandem repeats of protein domains have been designed effectively by fusing the proximal N- and C-termini of neighboring molecules. This type of duplication event results in repetition of the same fold (Brucker, 2000; Hytönen *et al.*, 2006; Nauli *et al.*, 2007; Zhou *et al.*, 2008). It has also been shown that when the N- and C-termini of the same protein are proximal, these termini can be linked together and new termini can be introduced in a number of different possible locations (Luger *et al.*, 1989; Hennecke *et al.*, 1999; Graf & Schachman, 1996; Iwakura *et al.*, 2000). This circular-permutation approach has been used to engineer a number of proteins (Yu & Lutz, 2011; Shui *et al.*, 2011; Lo *et al.*, 2009).

Here, we examine experimentally whether heterodimers can be used for protein design. To obtain the desired structure, we take advantage of the proximity of the N-and C-termini of the heterodimer. The flagellar export chaperone FliS from *Aquifex aeolicus* forms a heterodimeric complex with the flagellin FliC. They form part of the multi-component flagella-assembly system. In *A. aeolicus*, FlgE, FlgK and FlgL form the basal component, FliD makes up the filament cap and FliC polymerizes to generate the tail. However, FliC needs to be secreted first, and the export chaperone FliS is required for this step.

The structure of FliS is available both in complex with a fragment of FliC (PDB entry 1ory) and by itself (PDB entry 1orj) (Evdokimov *et al.*, 2003). The FliS·FliC[1] complex structure shows a distance of 14.3 Å between the N-terminus of FliS and the C-terminus of FliC (Fig. 1*a*). The crystal structure of FliS in the absence of FliC contains four molecules in the asymmetric unit. In two of these molecules the N-terminal region of FliS adopts a conformation substantially different from the conformation found in the structure of the FliS·FliC complex (Fig. 1*b*), while in the other two molecules the N-terminal helix is missing. Taken together, these observations indicate that the N-terminal region of FliS is structurally flexible. The FliS protein has a helical fold composed of a four-helix bundle and a flexible N-terminal region that is partly helical. The FliC fragment is made up of three helices that wrap around the FliS helical bundle. Helix 1 and helix 3 of FliC are on opposite sides of the FliS helical bundle, and helix 2 is sandwiched at the top of the FliS helical bundle (Fig. 1*a*). We predicted that the fusion of FliC with FliS should produce a single domain. We determined the structure of the FliC-FliS fusion protein. Fold-recognition programs failed to recognize the new protein in its entirety, demonstrating that heterodimeric structures contain information that is not currently used for structural predictions. We identified examples in nature of heterodimers that are homologous to full-length proteins, highlighting that the fission of full-length proteins to generate heterodimers or the fusion of heterodimers to produce full-length proteins is consistent with evolutionary mechanisms. We suggest that heterodimeric structures contain a significant amount of unnoticed data that can be extracted and used for structural predictions.

2. Materials and methods

2.1. Fusion-protein design

The crystal structure of the FliS·FliC complex shows that some of the residues at the termini are disordered. The crystallized FliS is composed of 130 amino acids; however, only 119 are ordered in the crystal. Similarly, although the size of the crystallized FliC fragment is 55 amino acids, only 40 are visible in the crystal structure. For our fusion we used what is ordered and visible in the crystal structure of the FliS·FliC complex in addition to a small loop linking the two

[1] FliS·FliC represents the heterodimer and FliC-FliS represents the fusion.

polypeptide chains and a hexahistidine tag for purification. We recognized that the N-terminus of FliS is flexible. In order not to exacerbate this flexibility, we chose to link these two polypeptide chains with a short two-amino-acid linker (Gly-Ala), even though based on the C^{α}–C^{α} distance of 14.3 Å a longer linker would have been required. We reasoned that the flexible N-terminal region can adjust to accommodate this distance such that a two-amino-acid linker would be sufficient. As a result, the overall size of our construct is 170 amino acids (Supplementary Table S3[2]). We tested whether or not our designed fusion results in a novel domain using the structure-recognition program *VAST* (Gibrat *et al.*, 1996) and the results indicated that the fusion would generate a novel protein fold.

2.2. Protein expression, purification and characterization

A synthetic DNA sequence corresponding to the FliC-FliS fusion protein was purchased from GENEWIZ and was cloned into a pBAD vector. The expression vector was transformed into *Escherichia coli* Top10 competent cells and the protein was expressed at 310 K in LB Miller medium supplemented with 100 µg ml^{-1} ampicillin. Once the cells had reached an OD_{600} of 0.7, expression was induced by the addition of arabinose to a final concentration of 0.02%. Protein expression was carried out for 3 h. The cells were harvested by centrifugation. Next, the cells were resuspended in lysis buffer (50 m*M* Tris–HCl pH 8.0, 150 m*M* NaCl, 5% glycerol) with the addition of 1 m*M* PMSF and were lysed using a microfluidizer (three passes at 138 MPa). The soluble fraction (supernatant) was loaded onto Ni–NTA metal-affinity resin (Qiagen). The resin was washed with lysis buffer with an increasing concentration of imidazole (from 10 to 50 m*M*). The protein was eluted with 50 m*M* Tris–HCl pH 8.0, 150 m*M* NaCl, 5% glycerol, 250 m*M* imidazole, 2 m*M* β-mercaptoethanol. The purified protein was concentrated using an Amicon Ultra centrifugation unit with a 10K cutoff. The buffer was exchanged to 20 m*M* Tris–HCl pH 8.0, 50 m*M* NaCl. The protein was shown to be monomeric by size-exclusion chromatography (Superdex 75 column from GE Healthcare equilibrated with 20 m*M* Tris–HCl pH 8.0, 50 m*M* NaCl).

2.3. Crystallization and structure determination

Crystals were obtained by the hanging-drop method at 290 K. They were formed in 38% PEG 500 MME, 0.1 *M* NaCl buffered with Tris–HCl pH 9.0. For data collection, crystals were flash-cooled at 100 K in the crystallization solution. Diffraction data were obtained on the 22-ID beamline at the Advanced Photon Source (Argonne National Laboratory, Argonne, Illinois, USA). The data were indexed, integrated and scaled using the *HKL*-2000 package (Otwinowski & Minor, 1997). The protein crystals were determined to belong to space group $P2_12_12_1$. The structure was solved by molecular replacement using *Phaser* (McCoy *et al.*, 2007) with PDB entry 1ory as a search model. We identified two molecules in the asymmetric unit. Model building was performed using *Coot* and refinement was performed using *REFMAC* (Murshudov *et al.*, 2011; Winn *et al.*, 2011). Noncrystallographic symmetry (NCS) restraints were used in the early stages of refinement. The Ramachandran plot of the final model shows 98.7, 1.3 and 0% of the residues in the preferred, allowed and disallowed regions, respectively. Structure analysis was performed using *PROCHECK* (Laskowski *et al.*, 1993) and the *MolProbity* server (Chen *et al.*, 2010). Molecular-graphics images were prepared using *PyMOL* (Schrodinger; http://www.pymol.org).

2.4. Heterodimer selection

For our initial heterodimer data set, we performed a text search using the PDB server with the keyword 'heterodimer'. Structures with more than 90% sequence homology were excluded. 268 PDB entries were identified. We inspected all of them and removed entries that were not heterodimeric, entries containing DNA or RNA and entries in which either of the two chains was smaller than 25 amino acids. This produced a set of 160 PDB entries of nonredundant heterodimers. These structures were all visually inspected and the distances between their N- and C-termini were determined. 12 entries with N- to C-termini distances of less than 15 Å were identified (Supplementary Table S2). This protocol of identifying PDB entries with heterodimeric proteins is similar to that of Sowmya *et al.* (2011).

For a more comprehensive heterodimeric database, we first integrated two heterodimer databases, protein–protein interface data sets (Mintz *et al.*, 2005) and the 3D-dimer template library (Lo *et al.*, 2010), and obtained 2506 heterodimer structures that included PDB coordinates released up to 2006. Next, we searched and compiled 3181 newly added heterodimer structures from the PDB library. We searched all two-chain complexes from the PDB library released after 2006. Heterodimer structures were selected by interfacial contacts that have more than ten residue pairs and a distance between C^{α} atoms of residue pairs of lower than 5 Å. To reduce the redundancy, we used a 90% sequence-identity cutoff with 90% aligned length coverage for unique heterodimer structures. We excluded modeled structures and structures with a resolution lower than 3.5 Å.

2.5. Identification of structural homology between heterodimers and single-chain proteins

Heterodimer structures were stitched into single-chain structures and compared with 81 553 PDB structures using the structure-alignment tool *TM-align* (Zhang & Skolnick, 2005). Specifically, the residue numbers of each chain were reassigned to be recognized as single chains by the structure-alignment program. Both directions ($A \rightarrow B$ and $B \rightarrow A$) of the two chains were used for stitching if both termini distances were below 15 Å.

The local alignment provided by *TM-align* was particularly useful to find the aligned region spanning the N- and C-termini of each chain in the heterodimer. A TM score of >0.5 and an r.m.s.d. of <3.5 were used to find the structural

[2] Supplementary material has been deposited in the IUCr electronic archive (Reference: YT5059). Services for accessing this material are described at the back of the journal.

homology between heterodimer fusion and single-chain structures.

2.6. Identification of sequence homology between heterodimers and single-chain proteins

To investigate the potential utility of heterodimeric complexes for the prediction of homologs of their fusions, we carried out sequence analysis of heterodimeric complexes that have N- to C-termini distances of 15 Å or less. We generated 284 nonredundant set of fusion sequences with N- to C-termini linkages that satisfied this distance requirement. Some heterodimers have both possible fusion sequences generated, where both N- and C-termini are separated by less than 15 Å. A decoy set of 284 stitched heterodimers was built by shuffling native protein pairs to compare the performance of identifying fused homolog sequences. A *BLAST* search was then carried out to identify potential homologs of these fusion sequences. An initial set of homologs was generated by selecting protein sequences that showed an aligned sequence identity with a *BLAST E*-value of $\leq$0.001 and a sequence identity of $\geq$30% in each chain. We removed homolog hits that were from post-translational proteolysis.

2.7. Threading and domain indentification

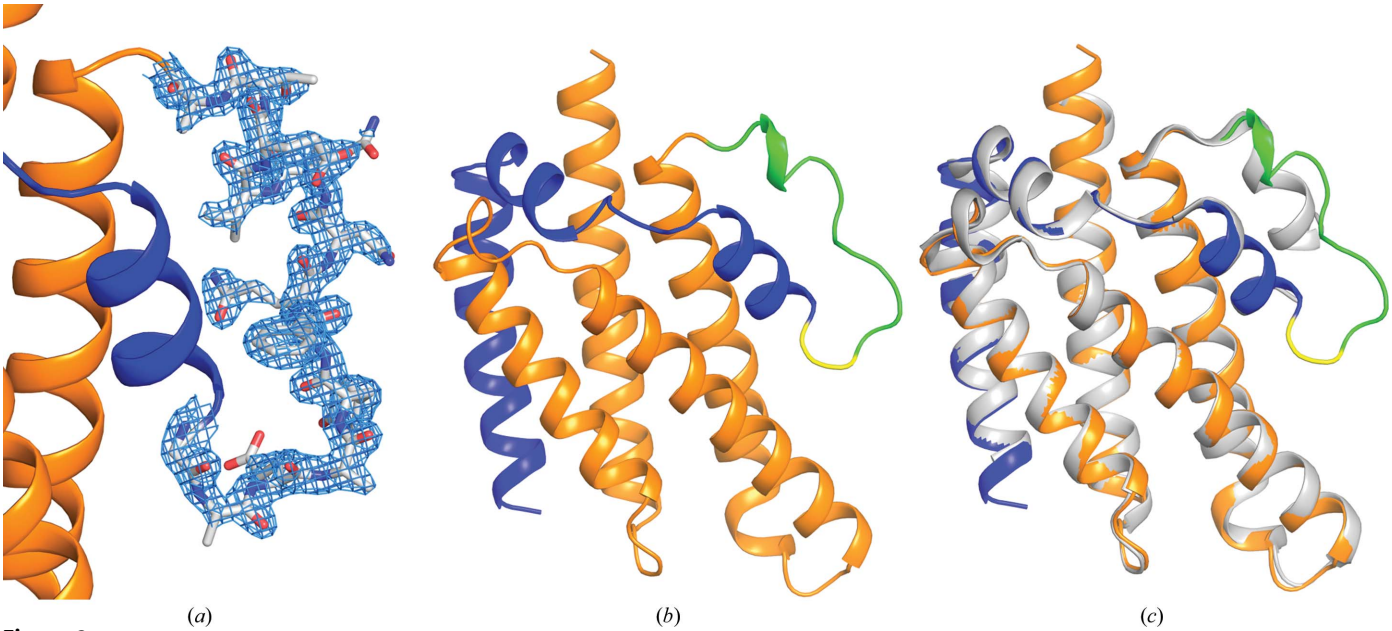
The FliC-FliS fusion sequence (Supplementary Table S3) was submitted to a number of threading-based structure-prediction servers. We used http://zhanglab.ccmb.med.umich.edu/I-TASSER/ for *I-TASSER*, http://zhanglab.ccmb.med.umich.edu/MUSTER/ for *MUSTER*, http://www.sbg.bio.ic.ac.uk/phyre2/ for *Phyre*2, http://raptorx.uchicago.edu/ for *RaptorX*, http://sparks.informatics.iupui.edu/sparks-x/ for *SPARKS*^{X} and http://ser-loopp.tc.cornell.edu/loopp_old.html for *LOOPP*. Structure alignments of the output PDB structures were performed using *Coot* (Emsley & Cowtan, 2004).

Domain identification was carried out using the *Protein Peeling* server at http://www.dsimb.inserm.fr/dsimb_tools/peeling3/index.html, the *DIAL* server at http://caps.ncbs.res.in/DIAL/DIALserver.html and the *Domain* 3*D* server at http://www.ibi.vu.nl/programs/domain3Dwww/.

3. Results

3.1. Structural characterization

We determined the structure of our FliC-FliS fusion protein to 1.75 Å resolution using X-ray crystallographic methods (Supplementary Table S1). Two molecules were identified in the asymmetric unit. The electron density at the site of the engineered linker for one of the two molecules (chain *A*) was clearly visible (Fig. 2*a*). This allowed us to build this entire region with high confidence. In the second molecule (chain *B*), the ten-amino-acid region that corresponds to the flexible N-terminal region of FliS was disordered, along with the two-amino-acid insertion. We built 165 residues for the first molecule (chain *A*) without any interruptions in the chain and 156 residues for the second molecule (chain *B*). The fusion protein maintained the same overall structure of the FliS·FliC heterodimer as predicted (Figs. 2*b* and 2*c*). The largest structural deviation between the *A* chain of our structure and



(*a*) (*b*) (*c*)

Figure 2
Structure of the FliC·FliS fusion. (*a*) Electron density covering the region that corresponds to the flexible N-terminal region of FliS and the two-amino-acid linker. The $2F_o - F_c$ electron density is contoured at 1.3σ and calculated using (FWT and PHIWT) values output by the refinement program *REFMAC*. (*b*) The structure of the FliC·FliS fusion protein. The region that corresponds to the helical bundle of FliS is colored orange, the two-amino-acid linker is colored yellow, the flexible N-terminal region of FliS is colored green and the portion that corresponds to the three FliC helices is colored blue. (*c*) Superposition of the FliC·FliS fusion with the FliS-FliC heterodimer. The FliC·FliS fusion is shown in the same colors as in (*b*) and the FiS-FliC heterodimer is shown in gray.

PDB entry 1ory is at the site corresponding to the flexible FliS N-terminal region and the two-amino-acid insertion. Not surprisingly, the region corresponding to the N-terminus of FliS proved to be flexible. The corresponding region in the *B* chain did not have clear electron density and was not included in the model. In the *A* chain, the N-terminal region of FliS adopts a new extended conformation that connects it to the C-terminus of FliC. The C^{α} r.m.s.d. between the *A* chain of our FliC-FliS fusion and the FliS·FliC complex is 0.62 Å over 159 residues and 0.45 Å over 149 residues on removing the region corresponding to the flexible N-terminal end of FliS and the two-amino-acid insertion. The r.m.s.d. between chains *A* and *B* is 0.47 Å over 156 residues. To assess whether the fusion represents a single domain or two distinct domains, we analyzed the final structure using the following domain-detection methods: *Protein Peeling* (Gelly *et al.*, 2006), *DIAL* (Sowdhamini & Blundell, 1995) and *Domain* 3*D* (Taylor, 1999). All three programs indicated that the fusion forms a single domain. We also found experimentally that the FliC-FliS fusion protein was monomeric based on size-exclusion chromatography and was highly soluble to at least 23 mg ml^{-1}.

3.2. Homology search

We evaluated the final structure using a number of structural databases and programs, including CATH (Greene *et al.*, 2007), *VAST* (Gibrat *et al.*, 1996), *DALI* (Holm & Rosenström, 2010) and *PDBeFold* (Krissinel & Henrick, 2004) tested both against SCOP categories (Murzin *et al.*, 1995) as well as all PDB entries. *DALI* and *PDBeFold* identified the *A* chain of PDB entry 1ory (FliS) as the closest structural homolog, missing the three helices of the FliC component. Similarly, *VAST* identified PDB entry 1orj (FliS) as the closest homolog, again missing the FliC component. If ranked by extent of coverage, the closest structural homolog identified by *VAST* is the *C* chain of bovine heart cytochrome *c* oxidase

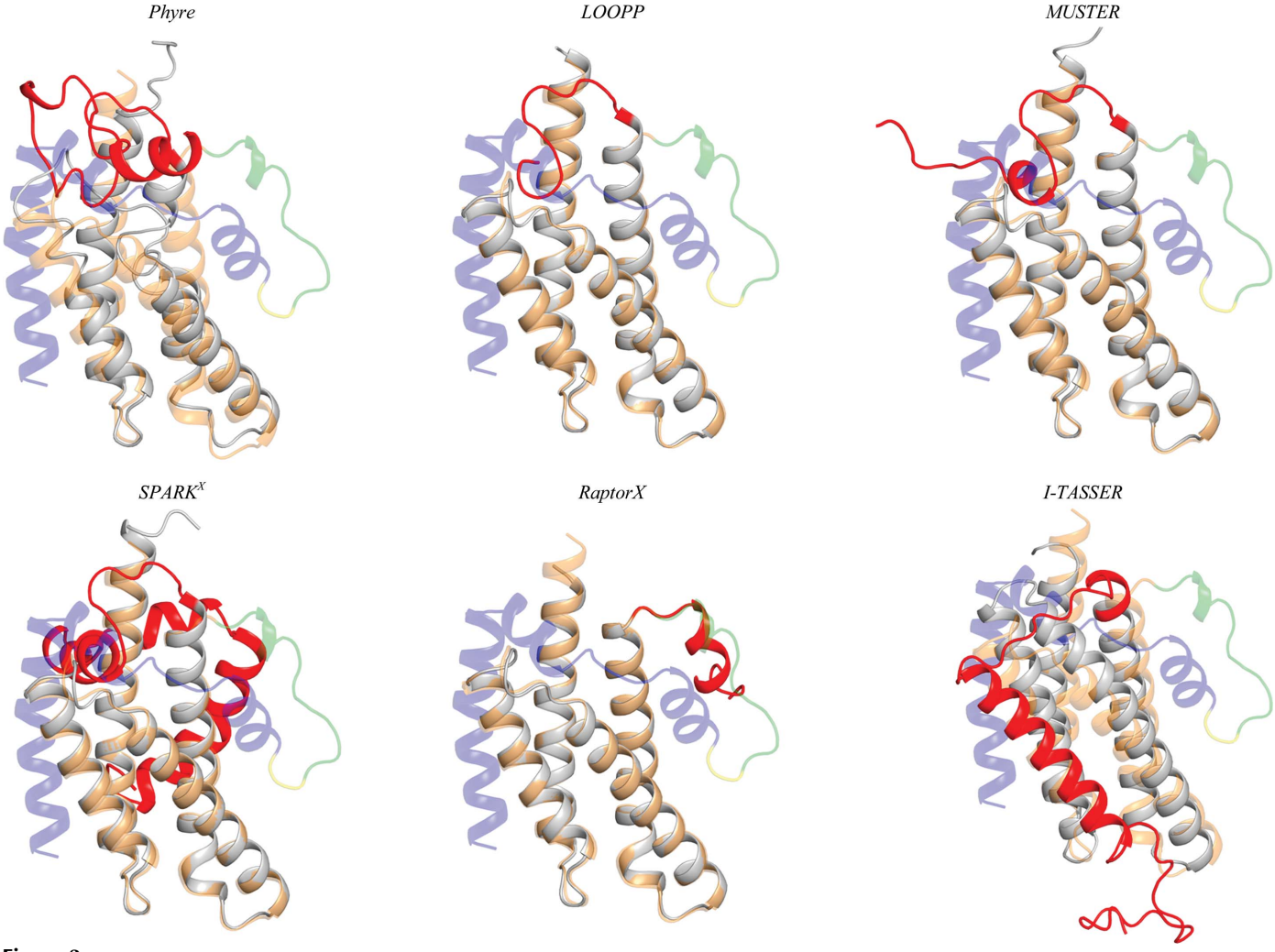


Figure 3
Threading results. Superposition of the FliC·FliS fusion protein with the top model produced by different threading programs (*Phyre, LOOPP, MUSTER, SPARK*^{X}*, RaptorX* and *I-TASSER*). The fusion protein is shown as semi-transparent. The color scheme for the fusion protein is the same as in Fig. 2(*b*). The outputs of the threading programs are colored in gray and red. The gray portion covers the region that corresponds to the helical bundle of the FliS protein, which was properly predicted. The red part covers the region that was not modeled properly, including the two-amino-acid helical linker, the flexible N-terminal region of the FliS protein and the entire FliC segment.

(Supplementary Fig. S1), in which transmembrane helices 3–7 resemble the first helix of FliC along with the four helices of the FliS helical bundle. CATH also identified the *C* chain of cytochrome *c* oxidase as the closest structural homolog. To examine whether we can identify proteins with homology to both the FliC and FliS chains in the same order as our fusion, we searched the GenBank nonredundant database by performing several rounds of *PSI-BLAST* (Altschul *et al.*, 1997) using the full length of the FliC-FliS fusion sequence. In each round we only selected the top three hits with the largest coverage. While many homologs corresponding to each chain independently can be obtained, this search did not yield any apparent homologs that included portions corresponding to both chains in the same protein.

3.3. Threading

Threading programs have been shown to be very effective for fold recognition even in cases of low sequence similarity (Bowie *et al.*, 1991). However, in order for threading programs to integrate the structural information of heterodimeric structures they would have to be able to 'stitch' heterodimers into single chains. To test whether current threading programs perform a 'stitching' function, we examined their ability to create a valid model of the FliC-FliS fusion protein. Using the sequence of our designed fusion protein, we tested the following threading programs: *Phyre*2 (Kelley & Sternberg, 2009), *LOOPP* (Vallat *et al.*, 2009), *MUSTER* (Wu & Zhang, 2008), *SPARK*^{X} (Yang *et al.*, 2011), *RaptorX* (Peng & Xu, 2011) and *I-TASSER* (Zhou & Skolnick, 2012). While the

Table 1
Full-length proteins that are structurally homologous to heterodimeric complexes.

	Heterodimer				R.m.s.d (Å)	
	A	*B*	Single chain	Distance between N- and C-termini (Å)	*A*	*B*
PDB code	1gx7 chain *A*	1gx7 chain *D*	1feh chain *A*	13.44 (397*A*→36*D*)	2.57	2.25
Chain length	371	88	574			
Protein description	Periplasmic [Fe]-hydrogenase large subunit	Periplasmic [Fe]-hydrogenase small subunit	Iron hydrogenase 1			
PDB code	1krl chain *B*	1krl chain *A*	1eqk chain *A*	6.42 (148*B*→101*A*)	2.47	2.60
Chain length	48	44	102			
Protein description	Monellin chain B	Monellin chain A	Oryzacystatin 1			
PDB code	2b9s chain *A*	2b9s chain *B*	1a31 chain *A*	4.47 (456*A*→211*B*)	1.28	1.29
Chain length	426	52	457			
Protein description	Topoisomerase 1B large subunit	Topoisomerase 1B small subunit	DNA topoisomerase 1			

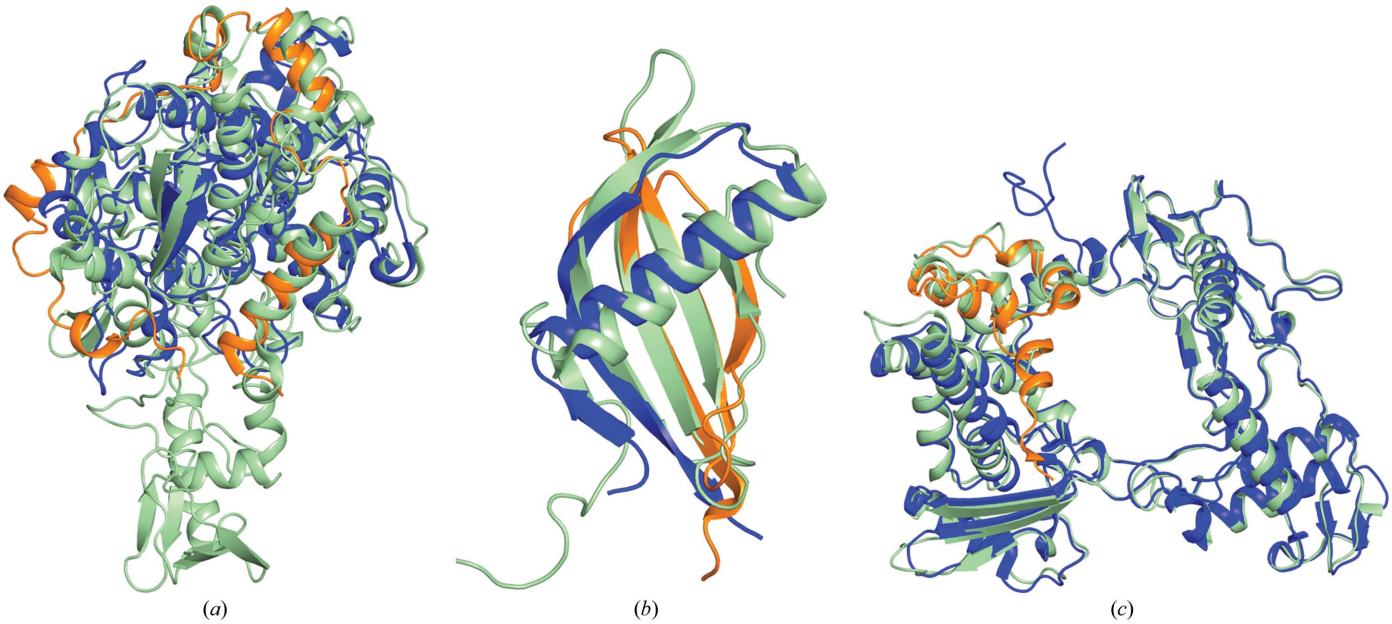


Figure 4
Structural homology between heterodimers and single-chain proteins. Three examples of heterodimers that have the same overall strucuture as a full-length protein homolog. (*a*) Periplasmic cytochrome *c* [Fe]-hydrogenase large subunit (PDB entry 1gx7 chain *A*), small subunit (PDB entry 1gx7 chain *D*) and cytoplasmic [Fe]-only hydrogenase (PDB entry 1c4a chain *A*). (*b*) Monellin B-chain (PDB entry 1krl chain *B*), A-chain (PDB entry 1krl chain *A*) and orizacystatin (PDB entry 1eqk chain *A*). (*c*) Topoisomerase 1 large subunit (PDB entry 2b9s chain *A*), small subunit (PDB entry 2b9s chain *B*) and topoisomerase 1 (PDB entry 1tl8 chain *A*).

Table 2
Proteins homologous to both components of heterodimeric complexes with proximal N- and C-termini.

PDB code of heterodimer	NCBI ID of identified sequence	Protein
1gcq *BC*	XP_003438306.1	Intersectin 1
1hqm *CD*	YP_006235825.1	RNA polymerase
1nfi *CE*	CAG05087.1	Unnamed
1pk6 *AC*	XP_003465201.1	Complement c1q TNF
1spp *BA*	XP_003363452.1	Hypothetical protein
3cdg *EF*	XP_002169267.1	Tyrosine kinase receptor
2b9s *AB*	AAA61207.1	Topoisomerase
1gx7 *AD*	YP_461142.1	Iron-only hydrogenase
4mon *BA*	1iv7 *A*	Monellin

threading programs were successful in identifying chain *A* of PDB entries 1ory or 1orj (corresponding to FliS alone), separate hits identified the *B* chain of 1ory (corresponding to FliC). None were able to put these two chains together and generate a reasonable model of the fusion protein (Fig. 3).

3.4. Heterodimer database analysis

The distance between the N- and C-termini of the FliS·FliC complex is 14.3 Å; therefore, we examined how often heterodimers have N- to C-termini distances of less than 15 Å. Firstly, an initial heterodimer data set was prepared by carrying out a simple keyword search in the PDB. We identified an initial set of 160 nonredundant (<90% sequence identity) heterodimeric entries from the PDB (Supplementary Table S2) and found 13 (7.5%) structures with N- to C-termini distances of 15 Å or less. For a more comprehensive heterodimeric library, previous databases (Mintz *et al.*, 2005; Lo *et al.*, 2010) were integrated along with additional structures by performing a search based on interfacial contact distances. After redundant entries were excluded, a set of 5687 heterodimer structures was obtained. When the N- to C-termini distances were measured for the more comprehensive data set, this library showed that 380 of 5687 (6.7%) heterodimers have N- to C-termini distances below 15 Å (Supplementary Fig. S2).

3.5. Homology of heterodimers to full-length proteins

To examine whether examples can be found in nature of heterodimers that adopt the same structures as full-length proteins, we searched the PDB for single-chain proteins that are structurally homologous to both components of heterodimeric complexes. Using our library of 5687 heterodimeric structures, we searched the PDB for single-chain proteins that are structurally homologous to at least 35 amino acids in each chain of a heterodimeric complex. We identified three examples: an iron hydrogenase (Peters *et al.*, 1998), oryzacystatin 1 (Nagata *et al.*, 2000) and a topoisomerase (Redinbo *et al.*, 1998). In all three examples the single-chain proteins align well with both chains of the heterodimeric complexes, with a C^{α} r.m.s.d. below 3.5 Å (Fig. 4 and Table 1). Interestingly, the N- to C-termini distances in all three examples that we identified are below 15 Å (13.44, 6.42 and 4.47 Å, respectively).

Given the relatively small size of the PDB compared with protein-sequence databases, it is reasonable to expect that many more examples of full-length proteins that are structurally homologous to heterodimers are present in these larger databases. Some of these full-length proteins may be homologous to heterodimers with known structures. However, as we have demonstrated, these examples are not detected by current threading methods. Hence, we carried out a detailed sequence search to look for examples of proteins whose structure predictions may benefit from the information provided by heterodimeric structures. Domain fusions are a common occurrence throughout evolution, but a fusion may not maintain the structure of a related heterodimer (Kim *et al.*, 2006). Therefore, to maximize the predictive potential of heterodimeric structures, only heterodimers with proximal N- and C-termini (less than 15 Å) were selected. Also, only the matching hits that had the two chains linked directly in the proper N- to C-terminal direction were accepted. Using our heterodimeric database, we identified 284 heterodimers with proximal N- and C-termini. Of these 284 heterodimers, nine were identified to have homologs to their fusions in the GenBank nonredundant database, whereas for a set of 284 decoy heterodimers we were not able to identify any fused homologs (Fig. 5 and Table 2).

4. Discussion

4.1. Threading and structure predictions

Protein-design methods and protein structure-prediction techniques are interrelated. The structure of the FliC-FliS fusion suggests that heterodimeric structures can be used for the prediction of structures of fusion proteins. We showed that threading programs do not extract information from heterodimeric structures beyond the single chain to single structure relationship. Therefore, a wealth of structural information is under-utilized. To improve threading-based predictions, a 'stitching' function may need to be introduced. Such functionality should allow the algorithms to recognize how to connect the end of one chain to the beginning of another.

Naturally occurring fusion proteins have already been applied very effectively to identify functionally linked protein domains, leading to improved functional predictions (Marcotte *et al.*, 1999). These functional linkages include proteins that do not form direct structural interactions, such as proteins that are functionally linked by participating in the same metabolic pathway. Here, we have taken a structural point of view and suggest that the structures of heterodimers can be used for structural predictions of fused proteins.

Structurally, fusions of heterodimers have been classified into two possible categories (Kim *et al.*, 2006): (i) genuine fusions, in which the fusion maintains the structure of the heterodimer, and (ii) non-genuine fusions, in which the fusion adopts a structure that differs from that observed in the heterodimeric complex. Here, we started with heterodimeric structures and searched for related fused proteins. We presumed that heterodimers with proximal N- and C-termini are more likely to produce genuine fusions when connected by short linkers than heterodimers with distant N- and C-termini.

Therefore, we limited our search to heterodimers with proximal N- and C-termini. In essence, we treated the gaps in three dimensions that separate the N- and C-termini of heterodimers similarly to gaps in traditional sequence alignments, which are given a larger penalty for longer distances. Since we do not have sufficient data to establish a reliable weight, we simply used a cutoff of 15 Å. This 15 Å distance was based on the 14.3 Å distance between the termini of the FliS and FliC heterodimer. Through sequence analysis of our heterodimeric data set, we identified nine hits (Table 2) that represent examples where the incorporation of data from heterodimeric structures can improve structural predictions.

The reliability of predictions based on heterodimeric data is likely to depend on a number of factors including the overall quality of the sequence alignment, the distances between the termini, the strength of the interaction between the polypeptide chains and the degree of conservation of the sequence corresponding to the interface region. Further research may be required to correlate these parameters with a meaningful reliability index. Gene fusion and gene fission are commonly accepted themes in protein evolution (Kummerfeld & Teichmann, 2005; Peisajovich *et al.*, 2006; Marsh & Teichmann, 2010; Pasek *et al.*, 2006). Both of these mechanisms can produce structural homologies between single-chain proteins and heterodimeric complexes.

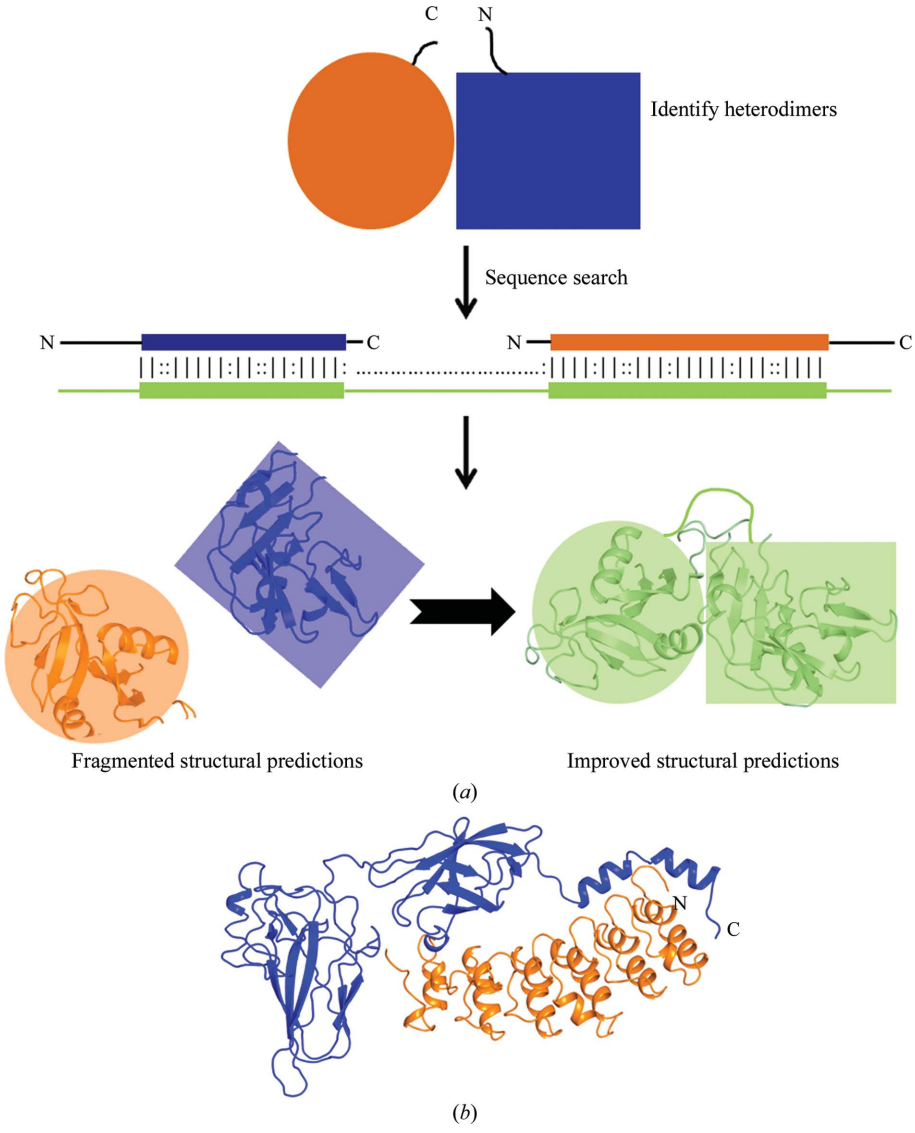


Figure 5
Structural predictions. (*a*) A scheme in three steps illustrating the approach we used to identify proteins for which structures can be more readily predicted by taking advantage of known heterodimeric structures. (i) Identify heterodimers with proximal N- and C-termini; (ii) generate the fusion sequence and identify homologs to the fusion that show homology to both domains of the heterodimer; (iii) more accurate prediction of the full-length protein is now possible. (*b*) An example of a full-length protein sequence with homology to a heterodimeric complex with proximal N- and C-termini. The protein from puffer fish with GenBank ID CAG05087.1 displays sequence homology to both components of the IκBα–NF-κB heterodimeric complex (PDB entry 1nfi). As a result, the structure of this protein can now be more easily predicted by taking advantage of heterodimeric structural information.

4.2. New domain?

To examine whether the FliC-FliS fusion protein represents a new domain, we tested the fusion structure with a number of programs and databases. Threading programs produced fragmented hits for the FliS and FliC portions, but were unable to produce a complete match to the fusion. Similar results were obtained by testing with CATH (Greene *et al.*, 2007), *VAST* (Gibrat *et al.*, 1996), *DALI* (Holm & Rosenström, 2010) and *PDBeFold* (Krissinel & Henrick, 2004). These results showed that our FliC-FliS fusion represents a novel structure. To determine whether the fusion is classified as a single-domain or a multidomain protein by automated methods, a number of domain-identification programs were tested: *Protein Peeling* (Gelly *et al.*, 2006), *DIAL* (Sowdhamini & Blundell, 1995) and *Domain 3D* (Taylor, 1999). All three identified the entire fusion as a single domain. Typically, the fusion of two domains results in a multidomain protein. Our fusion is between a protein domain (FliS) and a protein fragment (FliC). The FliC segment is a protein fragment that does not represent a protein domain. Protein domains are defined to have stable three-dimensional structures (Wetlaufer, 1973) and to be able to fold independently (Jaenicke, 1987). The tertiary structure of FliC as seen in the FliS·FliC complex is extended and does not have its own hydrophobic core, and its structure

appears to be greatly influenced by interactions with the FliS protein (1861.5 Å^2 buried area; Krissinel & Henrick, 2007). FliC on its own would not be expected to maintain the same three-dimensional structure as is observed when it is bound to FliS. Many examples of intrinsically disordered regions in proteins are available that behave similarly to FliC by only achieving their final structure on binding to their physiological partners (Wright & Dyson, 2009). Whether the FliC-FliS fusion is a new fold or a new domain depends on the exact definition that is used for protein domains (Grishin, 2001). We can state that out of 165 ordered residues, fold-recognition programs failed to predict the first 54, corresponding to missing three of the seven secondary-structural elements of the FliC-FliS fusion protein.

4.3. Evolutionary implications

The designed FliC-FliS fusion protein provides an example of how evolution may produce novel protein folds and how more complex structures can be created from simpler units such as peptide fragments. Interestingly, the construct that we prepared does not have any readily identifiable homologs within the sizeable GenBank (nonredundant) database or in the PDB, even though the design is simple and protein fusion is an accepted mechanism of protein evolution (Kummerfeld & Teichmann, 2005; Peisajovich *et al.*, 2006). This suggests that no natural fusions of FliS-FliC are known, which agrees with the concept that the structural integrity of a protein fold plays only a minor role in how frequently it is found in nature

It is accepted that protein domains act as modules that can be fused to generate multidomain proteins. However, a fundamental question that remains unanswered is: 'how do new protein domains originate?' Several mechanisms, such as insertion, fusion, deletion, recombination, duplication or oligomerization, circular permutation and rearrangements, have been proposed to play a role in protein evolution (Grishin, 2001; Söding & Lupas, 2003; Lupas *et al.*, 2001). It has also been proposed that peptide segments, although unable to fold on their own, can play a role in the evolution and assembly of novel protein domains (Riechmann & Winter, 2006). FliC is an example of such a protein fragment that is unlikely to maintain the same tertiary structure on its own, since all of its tertiary interactions are with its partner FliS. Therefore, not only does this example show that fusion is a mechanism that can create novel protein domains, but also that protein fragments can contribute to the generation of new domain folds. Similarly, the examples that we have identified of full-length proteins that are homologous to heterodimeric complexes support the concept that fusion of heterodimers is a viable mechanism for the natural expansion of protein diversity.

We thank Cameron Mura, Peter Horanyi, Doug Rees and James Bowie for critical reading of the manuscript.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). *Nucleic Acids Res.* **25**, 3389–3402.
Bharat, T. A., Eisenbeis, S., Zeth, K. & Höcker, B. (2008). *Proc. Natl Acad. Sci. USA*, **105**, 9942–9947.
Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991). *Science*, **253**, 164–170.
Brucker, E. A. (2000). *Acta Cryst.* D**56**, 812–816.
Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst.* D**66**, 12–21.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Evdokimov, A. G., Phan, J., Tropea, J. E., Routzahn, K. M., Peters, H. K., Pokross, M. & Waugh, D. S. (2003). *Nature Struct. Biol.* **10**, 789–793.
Gelly, J.-C., de Brevern, A. G. & Hazout, S. (2006). *Bioinformatics*, **22**, 129–133.
Gibrat, J.-F., Madej, T. & Bryant, S. H. (1996). *Curr. Opin. Struct. Biol.* **6**, 377–385.
Graf, R. & Schachman, H. K. (1996). *Proc. Natl Acad. Sci. USA*, **93**, 11591–11596.
Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., Sillitoe, I., Yeats, C., Thornton, J. M. & Orengo, C. A. (2007). *Nucleic Acids Res.* **35**, D291–D297.
Grishin, N. V. (2001). *J. Struct. Biol.* **134**, 167–185.
Hennecke, J., Carfi, A. & Wiley, D. C. (2000). *EMBO J.* **19**, 5611–5624.
Hennecke, J., Sebbel, P. & Glockshuber, R. (1999). *J. Mol. Biol.* **286**, 1197–1215.
Höcker, B., Claren, J. & Sterner, R. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 16448–16453.
Holm, L. & Rosenström, P. (2010). *Nucleic Acids Res.* **38**, W545–W549.
Hytönen, V. P., Hörhä, J., Airenne, T. T., Niskanen, E. A., Helttunen, K. J., Johnson, M. S., Salminen, T. A., Kulomaa, M. S. & Nordlund, H. R. (2006). *J. Mol. Biol.* **359**, 1352–1363.
Iwakura, M., Nakamura, T., Yamane, C. & Maki, K. (2000). *Nature Struct. Biol.* **7**, 580–585.
Jaenicke, R. (1987). *Prog. Biophys. Mol. Biol.* **49**, 117–237.
Kelley, L. A. & Sternberg, M. J. (2009). *Nature Protoc.* **4**, 363–371.
Kim, W. K., Henschel, A., Winter, C. & Schroeder, M. (2006). *PLoS Comput. Biol.* **2**, e124.
Krissinel, E. & Henrick, K. (2004). *Acta Cryst.* D**60**, 2256–2268.
Krissinel, E. & Henrick, K. (2007). *J. Mol. Biol.* **372**, 774–797.
Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). *Science*, **302**, 1364–1368.
Kummerfeld, S. K. & Teichmann, S. A. (2005). *Trends Genet.* **21**, 25–30.
Laskowski, R. A., Moss, D. S. & Thornton, J. M. (1993). *J. Mol. Biol.* **231**, 1049–1067.
Lo, W.-C., Lee, C.-C., Lee, C.-Y. & Lyu, P.-C. (2009). *Nucleic Acids Res.* **37**, D328–D332.
Lo, Y.-S., Chen, Y.-C. & Yang, J.-M. (2010). *BMC Genomics*, **11**, Suppl. 3, S7.
Luger, K., Hommel, U., Herold, M., Hofsteenge, J. & Kirschner, K. (1989). *Science*, **243**, 206–210.
Lupas, A. N., Ponting, C. P. & Russell, R. B. (2001). *J. Struct. Biol.* **134**, 191–203.
Marcotte, E. M., Pellegrini, M., Ng, H.-L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). *Science*, **285**, 751–753.
Marsh, J. A. & Teichmann, S. A. (2010). *Genome Biol.* **11**, 126.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
Mintz, S., Shulman-Peleg, A., Wolfson, H. J. & Nussinov, R. (2005). *Proteins*, **61**, 6–20.
Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.
Nagata, K., Kudo, N., Abe, K., Arai, S. & Tanokura, M. (2000). *Biochemistry*, **39**, 14753–14760.

Nauli, S., Farr, S., Lee, Y.-J., Kim, H.-Y., Faham, S. & Bowie, J. U. (2007). *Protein Sci.* **16**, 2542–2551.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Padilla, J. E., Colovos, C. & Yeates, T. O. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 2217–2221.
Pasek, S., Risler, J. L. & Brézellec, P. (2006). *Bioinformatics*, **22**, 1418–1423.
Peisajovich, S. G., Rockah, L. & Tawfik, D. S. (2006). *Nature Genet.* **38**, 168–174.
Peng, J. & Xu, J. (2011). *Proteins*, **79**, Suppl. 10, 161–171.
Peters, J. W., Lanzilotta, W. N., Lemon, B. J. & Seefeldt, L. C. (1998). *Science*, **282**, 1853–1858.
Redinbo, M. R., Stewart, L., Kuhn, P., Champoux, J. J. & Hol, W. G. J. (1998). *Science*, **279**, 1504–1513.
Riechmann, L. & Winter, G. (2006). *J. Mol. Biol.* **363**, 460–468.
Rosenbaum, D. M., Cherezov, V., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kobilka, T. S., Choi, H.-J., Yao, X.-J., Weis, W. I., Stevens, R. C. & Kobilka, B. K. (2007). *Science*, **318**, 1266–1273.
Shui, B., Wang, Q., Lee, F., Byrnes, L. J., Chudakov, D. M., Lukyanov, S. A., Sondermann, H. & Kotlikoff, M. I. (2011). *PLoS One*, **6**, e20505.
Siddiqui, A. S. & Barton, G. J. (1995). *Protein Sci.* **4**, 872–884.
Smyth, D. R., Mrozkiewicz, M. K., McGrath, W. J., Listwan, P. & Kobe, B. (2003). *Protein Sci.* **12**, 1313–1322.
Söding, J. & Lupas, A. N. (2003). *Bioessays*, **25**, 837–846.
Sowdhamini, R. & Blundell, T. L. (1995). *Protein Sci.* **4**, 506–520.
Sowmya, G., Anita, S. & Kangueane, P. (2011). *Bioinformation*, **6**, 137–143.
Taylor, W. R. (1999). *Protein Eng.* **12**, 203–216.
Vallat, B. K., Pillardy, J., Májek, P., Meller, J., Blom, T., Cao, B. & Elber, R. (2009). *Proteins*, **76**, 930–945.
Wetlaufer, D. B. (1973). *Proc. Natl Acad. Sci. USA*, **70**, 697–701.
Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.
Wright, P. E. & Dyson, H. J. (2009). *Curr. Opin. Struct. Biol.* **19**, 31–38.
Wu, S. & Zhang, Y. (2008). *Proteins*, **72**, 547–556.
Yang, Y., Faraggi, E., Zhao, H. & Zhou, Y. (2011). *Bioinformatics*, **27**, 2076–2082.
Ye, Q., Li, X., Wong, A., Wei, Q. & Jia, Z. (2006). *Biochemistry*, **45**, 738–745.
Yu, A., Xing, Y., Harrison, S. C. & Kirchhausen, T. (2010). *Structure*, **18**, 1311–1320.
Yu, Y. & Lutz, S. (2011). *Trends Biotechnol.* **29**, 18–25.
Zhan, Y., Song, X. & Zhou, G. W. (2001). *Gene*, **281**, 1–9.
Zhang, Y. & Skolnick, J. (2005). *Nucleic Acids Res.* **33**, 2302–2309.
Zhou, Z., Feng, H., Hansen, D. F., Kato, H., Luk, E., Freedberg, D. I., Kay, L. E., Wu, C. & Bai, Y. (2008). *Nature Struct. Mol. Biol.* **15**, 868–869.
Zhou, H. & Skolnick, J. (2012). *Proteins*, **80**, 352–361.