Bioorganic & Medicinal Chemistry Letters 21 (2011) 2372–2375

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry Letters

journal homepage: www.elsevier.com/locate/bmcl

Artificial leucine rich repeats as new scaffolds for protein design

Hemda Baabur-Cohen [†], Subashini Dayalan [†], Inbal Shumacher, Rivka Cohen-Luria, Gonen Ashkenasy [*]

Department of Chemistry, Ben Gurion University of the Negev, Beer Sheva, Israel

ARTICLE INFO

Article history:
Received 16 January 2011
Revised 16 February 2011
Accepted 21 February 2011
Available online 21 March 2011

Keywords:
Leucine rich repeat proteins
Native chemical ligation
Peptides
Protein design

ABSTRACT

The leucine rich repeat (LRR) motif that participates in many biomolecular recognition events in cells was suggested as a general scaffold for producing artificial receptors. We describe here the design and first total chemical synthesis of small LRR proteins, and their structural analysis. When evaluating the tertiary structure as a function of different number of repeating units (1–3), we were able to find that the 3-repeats sequence, containing 90 amino acids, folds into the expected structure.

The de-novo design of new protein structures may shine light on the basic principles that govern protein folding, protein–protein interactions, and interactions with other biomolecules. It may thus facilitate the development of devices for biotechnology applications, such as sensors and catalysts. Of special interest is the design of proteins that contain large and shallow solvent exposed surfaces readily available for multiple intermolecular interactions. Towards this aim, we have studied the design of consensus sequence Leucine Rich Repeat (LRR) proteins, and their total chemical synthesis. By evaluating their tertiary structure as a function of different number of repeating units, we were able to find a short sequence of 90 amino acid (aa) that folds into the correct structure.

Several research groups have applied the consensus design methodology to produce repeat proteins that fold correctly, and some proteins were also found to be good receptors for native, as well as non-native, ligands.[1–4] These studies have usually been carried out by biological expression of the proteins, and the most studied scaffolds were based on the ankyrin repeat proteins[5–7] and tetratricopeptide repeat proteins.[8,9] The LRR motif was chosen as the model for this study, since it has been associated with a large variety of molecular-recognition events in cells,[10–13] where it plays a role in such diverse processes as signal transduction, DNA repair, cell adhesion and more. Native LRR motifs are highly regular, with each repeat containing a short β-strand and a helix oriented in an anti-parallel manner. The hydrophobic residues, mainly Leu, and one Asn from each repeat, form an elongated core that stabilizes the 3D structure. The side-by-side association of the repeats builds an arch, with the β-strands being packed more closely together than the opposing helices. The arch's interior (concave) thus forms an extended ligand-binding surface.[10,14,15]

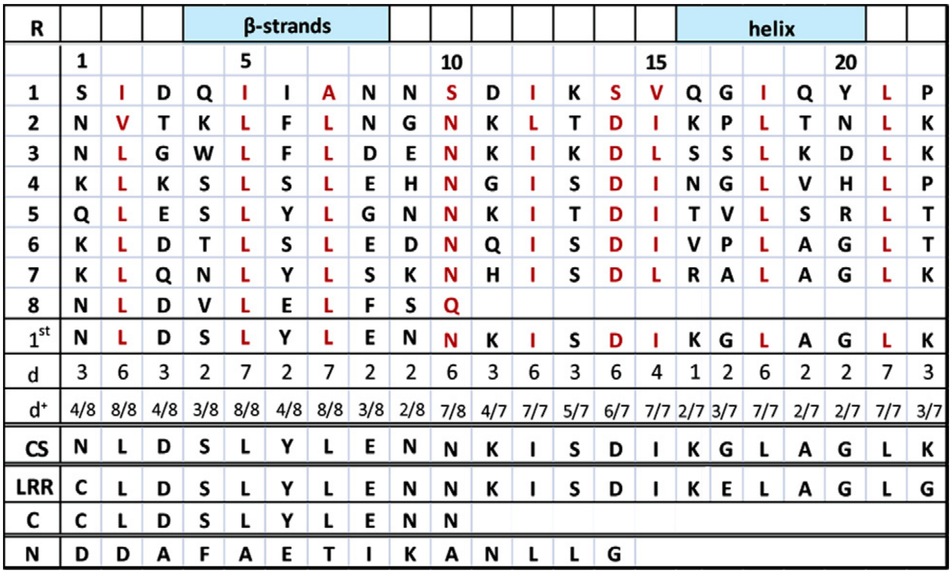
Our actual molecular design was based on the LRR domain of internalin B (InlB) protein from the bacterium *Listeria monocytogenes* for which the 3D structure is known.[16,17] This LRR domain is monomeric and folds correctly without the rest of the protein,[14,16,18] and the repeating unit is made of 22 amino acids, thus accessible by chemical synthesis. By aligning the 7.5 repeats of Inl B and counting the frequency of each residue at each position, we found a 22 amino acids consensus sequence (CS) that can serve as a building block for making the LRR by modular synthesis (Fig. 1). The resulted CS is very similar to sequences that were deduced from bioinformatics analysis of a number of known internalin proteins.[14] Nine conserved residues in each repeat (marked in *red* in Fig. 1) were kept for structural reasons, while residues at non-conserved positions were chosen by also taking into account the appearance of other residues with similar chemical characteristics (e.g., N and Q). Three additional changes were introduced to the calculated CS. Gly residue at position 17 of each repeat was replaced for Glu, to facilitate electrostatic interactions with Lys residue at position 16 in consecutive LRRs and promoting helix–helix interactions. Cys and Gly residues were introduced to the N- and C-termini, respectively, to facilitate repeat coupling through the native chemical ligation.

In order to search for the shortest protein sequences that can fold into the typical LRR structure we have designed three proteins that possess different number of repeating units (1–3), stabilized by short N- and C-termini sequences. The proteins were named **N1C**, **N2C** and **N3C**, and made of 46, 68 and 90 aa, respectively. The N-terminal sequence was kept almost as is from the native InlB, containing 14 residues that create an α-helix cap that shields the hydrophobic core from the aqueous media. The C-terminus is

[*] Corresponding author. Tel.: +972 8 6461637; fax: +972 8 6472943.
E-mail address: gonenash@bgu.ac.il (G. Ashkenasy).

[†] These two authors contributed equally to the paper.

0960-894X/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2011.02.093



R				β-strands												helix						
	1				5					10					15					20		
1	S	I	D	Q	I	I	A	N	N	S	D	I	K	S	V	Q	G	I	Q	Y	L	P
2	N	V	T	K	L	F	L	N	G	N	K	L	T	D	I	K	P	L	T	N	L	K
3	N	L	G	W	L	F	L	D	E	N	K	I	K	D	L	S	S	L	K	D	L	K
4	K	L	K	S	L	S	L	E	H	N	G	I	S	D	I	N	G	L	V	H	L	P
5	Q	L	E	S	L	Y	L	G	N	N	K	I	T	D	I	T	V	L	S	R	L	T
6	K	L	D	T	L	S	L	E	D	N	Q	I	S	D	I	V	P	L	A	G	L	T
7	K	L	Q	N	L	Y	L	S	K	N	H	I	S	D	L	R	A	L	A	G	L	K
8	N	L	D	V	L	E	L	F	S	Q												
1^{st}	N	L	D	S	L	Y	L	E	N	N	K	I	S	D	I	K	G	L	A	G	L	K
d	3	6	3	2	7	2	7	2	2	6	3	6	3	6	4	1	2	6	2	2	7	3
d*	4/8	8/8	4/8	3/8	8/8	4/8	8/8	3/8	2/8	7/8	4/7	7/7	5/7	6/7	7/7	2/7	3/7	7/7	2/7	2/7	7/7	3/7
CS	N	L	D	S	L	Y	L	E	N	N	K	I	S	D	I	K	G	L	A	G	L	K
LRR	C	L	D	S	L	Y	L	E	N	N	K	I	S	D	I	K	E	L	A	G	L	G
C	C	L	D	S	L	Y	L	E	N	N												
N	D	D	A	F	A	E	T	I	K	A	N	L	L	G								

Figure 1. Deriving the consensus sequence (CS) from sequence repeats in the *L. monocytogenes* Inl B protein. The nine structurally conserved residues are marked in red. R is the repeat number in the native LRR domain; "1^{st}" shows the chosen amino acid at each position, and d and d* the number of times this amino acid or similar one(s) occupy the specific position along the repeat. LRR, C, and N are the actual sequences chosen for chemical synthesis of the artificial LRR proteins.

made of the first 10 amino acids of the CS (Fig. 1). Computational analysis of the protein structures was performed using the I-TASSER program[19,20] that searches databases to identify similar aa sequences and uses the matched fragments as a template for modeling the structure of the query sequence. The lowest energy replica obtained for the three proteins showed that indeed uniquely folded proteins can be formed from the designed sequences, having anti-parallel arrangements of short helices and β-strands in each repeat and in the C-domain, and a more elongated α-helix in the N-domain (Fig. 2 and Supplementary data). These simulations suggest that the larger protein, **N3C**, would occupy a stable structure that better resembles that of the native Inl B, as observed from the fact that longer β-strands and more well-folded helices were obtained in its lowest energy replica, as compared to **N1C** and **N2C**. Furthermore, in accord with the recent reports that the existence of an N-cap on LRR motifs is crucial for both correct and rapid folding,[21,22] we observed that **N2C** is more stable in the native LRR structure, relative to its N-cap deficient analog **2C** (Supplementary data).

The modular chemical syntheses of **N1C**, **N2C** and **N3C** were achieved by first synthesizing their corresponding building blocks on solid phase and then consecutive attachments by the native chemical ligation[23] (Fig. 3). The C-terminal building blocks, containing free Cys residues, were made using the Fmoc methodology, while the thioester containing N-terminal fragments by the Boc based synthesis using the in situ neutralization protocol. All peptides were purified by preparative HPLC and their identity and purity confirmed by analytical HPLC and LCMS (Supplementary data).

The synthesis of **N1C** (Fig. 3a) was completed after ligation of mM concentrations of the **N1** fragment, which is made of 36 aa composing the N-terminus and one repeating unit, with the 10 aa C-terminal fragment **C** (see peptide sequences in Fig. 1 and Supplementary data). Likewise, **N2C** was prepared by a single ligation of **N1** with the 32 aa **1C** fragment of the C-terminus attached to one repeating unit (Fig. 3b). The synthesis of **N3C** required two ligation steps, processed from the C- to N-termini using two different methods (Figs. 3c and d). In the first method (Fig. 3c), a single-repeat peptide having its Cys residue protected with Acm group

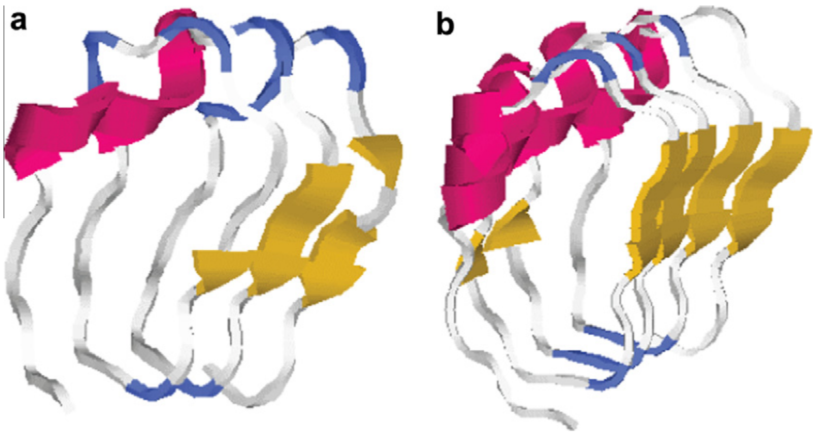


Figure 2. Lowest energy structures obtained by computational study of **N2C** (a) and **N3C** (b) using the I-TASSER program. The c-score for these lowest temperature replica were 0.89 (**N2C**) and 0.93 (**N3C**), respectively. The less stable structures were also usually similar to the presented models (data not shown).

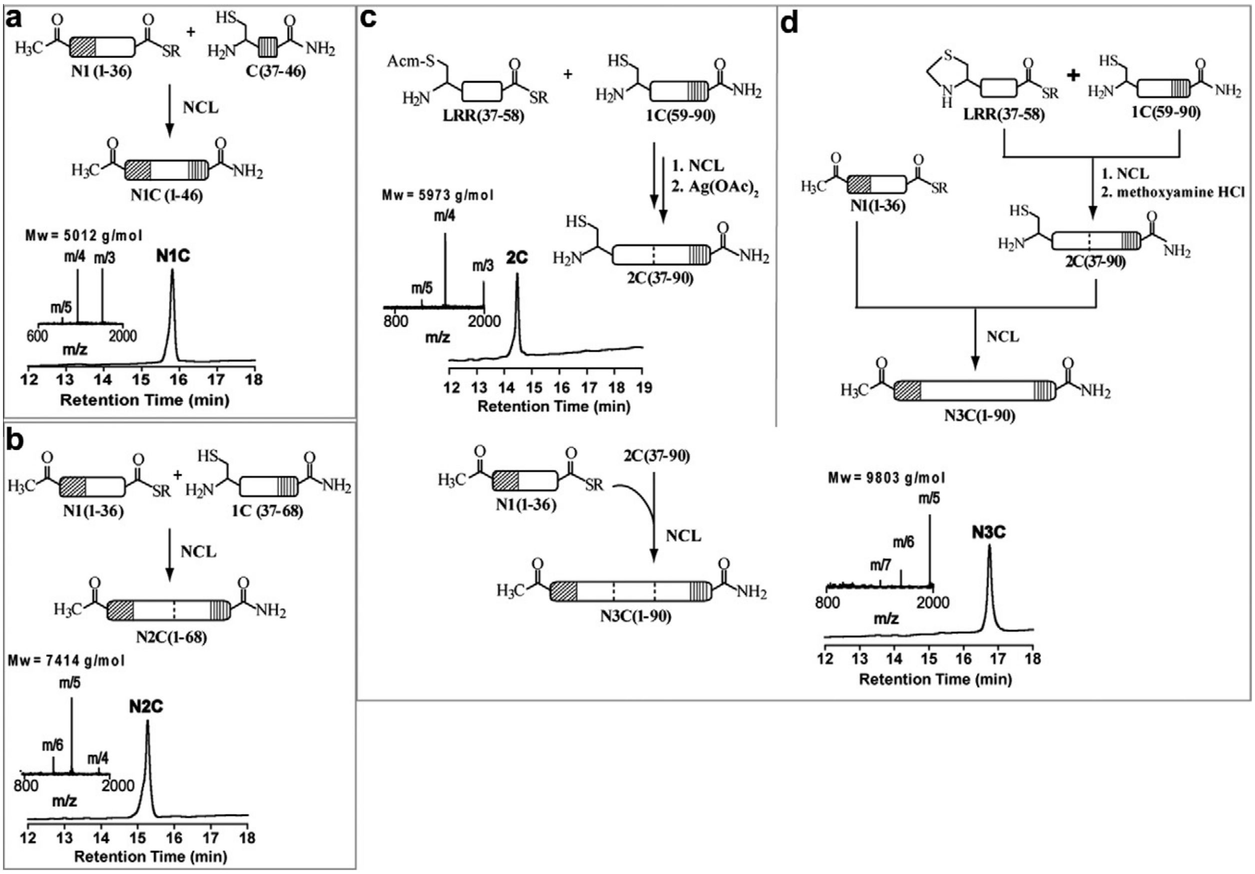


Figure 3. Synthesis and characterizations of **N1C** (a), **N2C** (b) and **N3C** (two different methods, c and d, respectively). Native chemical ligations (NCL) were performed using 3–5 mM concentration of each reactant, in unfolding conditions in MOPS buffer at pH 7.7 saturated with 6 M GnHCl, at 37 °C, and in the presence of thiophenol and benzyl mercaptan (4% in volume each). The intermediate product (**2C**) and all final products were purified by preparative HPLC, and characterized by analytical HPLC (230 nm) and LCMS, as shown next to each scheme. The measured MW in all cases was found to be ±3 molecular units from the calculated MW. After synthesis, **N2C** and **N3C** were further subjected to Cys desulfurization, re-purification and HPLC and MS analysis (Supplementary data). In all cases, all-L amino acids peptides have been used.

(LRR) first reacted with **1C** to yield the intermediate compound **2C**. The Acm group was then removed, and **2C** was purified and further ligated with **N1**. In the second method, a Thz protected LRR reacted with **1C** to yield **2C**, and then deprotection and further ligation with **N1** were performed in situ, as previously described.[24] This latter method required careful adjustment of the mole ratio of reactants at the first ligation to be exactly 1:1, since no purification takes place after this step. After achieving such optimization, the overall process afforded higher yields than obtained using the 1st method (with the Acm-protected LRR), and the detection of only minute amounts (typically less than 10%) of undesired side products. The observed reaction conversions during the ligation steps in all four syntheses were similar to previously reported, allowing to obtain $\sim$50% yields of purified compounds. The reactions with the LRR building blocks were found to be somewhat slower than those observed for shorter peptides, thus required 24–48 h to reach completion. After the synthesis, the free thiol side chains of Cys residues that were used for ligation were subjected to desulfurization, resulting in the CH$_3$ side chains of Ala at the attachment positions (Supplementary data). All the full-length proteins were purified by HPLC, and analyzed by HPLC and LCMS (Fig. 3).

Folding of the synthetic LRR proteins into structures having the expected secondary motifs was characterized by circular dichroism (CD), as shown in Figure 4. Thus, 50–100 μM protein solutions were allowed spontaneous folding by equilibrating in phosphate buffer at pH 7, for $\geqslant$0.5 h. The spectrum of **N1C** shows that this protein occupies mainly random coil conformations, as observed from the minimum at $\sim$200 nm. The red shift of this minimum towards 205 nm and the appearance of an additional minimum ('hump') at 226 nm in the spectrum of **N2C** reveal that part of the latter folded into a helix structure ($\sim$20% helicity). Interestingly, the spectrum of the cyclic LRR unit **1**$_{cyc}$—synthetically obtained by simple N to C cyclization via NCL (see sequence and characterization in Supplementary data)—also showed these features, suggesting that cyclization can help in getting the desired structure for the short sequence too. As can be expected from the structural features observed in Inl B (Fig. 1), and as was found using the I-TASSER modeling above and in previous studies,[18] only about a quarter of the amino acids—5 in each repeat and 10 in the N-cap—are expected to fold into helical structure, resulting in such small peaks. Most interestingly, the structure of **N3C** shows in addition to the helix characteristics minima (205 and 226 nm; 27% helicity), another minimum at 215 nm, which is a clear signature of b-sheet formation (25%). This data also correlates well with the I-TASSER prediction, where only **N3C** gave a reasonable amount of arrangement of the β-strands into small sheets (Fig. 2). We have further characterized the unfolding stability of **N2C** and **N3C**, by monitoring the changes in the 226 nm CD signal at different temperatures (Fig. 4b). The results show that both proteins undergo one-step conformational change from folded to unfolded structure. Furthermore, it was observed that while **N2C** completely unfolded during the measurements, allowing to calculate Tm of $\sim$50 °C, **N3C** was found to be more stable and did not completely unfold even at 90 °C.

We have shown in this paper the first chemical synthesis of LRR proteins, and demonstrated that small proteins can fold to form 3D

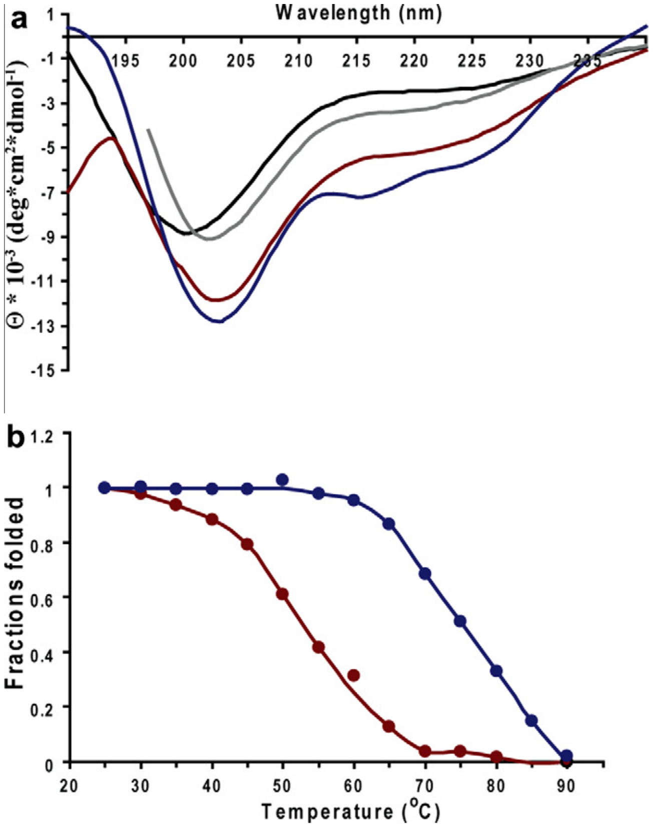


Figure 4. Structural characterization of the artificial LRR proteins. (a) CD spectra obtained for 50–100 μM solution of **N1C** (*black*), cyclic **N1C** (*gray*), **N2C** (*red*) and **N3C** (*blue*), after equilibration in phosphate buffer at pH 7. (b) Thermal denaturation of **N2C** (*red*) and **N3C** (*blue*) as deduced by following the 226 nm CD minimum at different temperatures.

structures that are similar to the structure of the larger native protein Inl B. Specifically, it was found that **N3C** that consists of the N, 3-repeat and C fragments can form quite stable structure, presumably useful for future use as an artificial receptors. It was observed for several LRR models that these proteins can bind other biomolecules via interactions of solvent exposed residues on the β-strands. Since three such residues are present on each strand (Fig. 1), binding of **N3C** to potential ligand may be mediated by the interactions of up to nine side chains. Testing this property would be the target of our future research. Furthermore, as was hypothesized before, the relatively easy way to form these LRR proteins chemically may facilitate the introduction into the sequence of non-native amino acids which may be further used to control folding and binding properties of such proteins.

Acknowledgements

This work was supported by a grant from the Israel–USA binational fund (2006358). The authors thank Dr. Dima Lukatsky with the help of I-TASSER modeling, and Prof. Yitzhak Tor (UCSD) for scientific discussions.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmcl.2011.02.093.

References and notes

1. Main, E. R. G.; Jackson, S. E.; Regan, L. *Curr. Opin. Struct. Biol.* **2003**, *13*, 482.
2. Tripp, K. W.; Barrick, D. *Structure* **2003**, *11*, 486.
3. Forrer, P.; Binz, H. K.; Stumpp, M. T.; Plueckthun, A. *ChemBioChem* **2004**, *5*, 183.
4. Kajander, T.; Cortajarena, A. L.; Regan, L. *Methods Mol. Biol.* **2006**, *340*, 151.
5. Kohl, A.; Binz, H. K.; Forrer, P.; Stumpp, M. T.; Pluckthun, A.; Grutter, M. G. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 1700.
6. Binz, H. K.; Amstutz, P.; Kohl, A.; Stumpp, M. T.; Briand, C.; Forrer, P.; Gruetter, M. G.; Plueckthun, A. *Nat. Biotechnol.* **2004**, *22*, 575.
7. Barrick, D. A. C. S. *Chem. Biol.* **2009**, *4*, 19.
8. Main, E. R. G.; Xiong, Y.; Cocco, M. J.; D'Andrea, L.; Regan, L. *Structure* **2003**, *11*, 497.
9. Main, E. R. G.; Stott, K.; Jackson, S. E.; Regan, L. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 5721.
10. Kobe, B.; Kajava, A. V. *Curr. Opin. Struct. Biol.* **2001**, *11*, 725.
11. Bell, J. K.; Mullen, G. E. D.; Leifer, C. A.; Mazzoni, A.; Davies, D. R.; Segal, D. M. *Trends Immunol.* **2003**, *24*, 528.
12. Matsushima, N.; Enkhbayar, P.; Kamiya, M.; Osaki, M.; Kretsinger, R. H. *Drug Design Rev.* **2005**, *2*, 305.
13. McEwan, P. A.; Scott, P. G.; Bishop, P. N.; Bella, J. *J. Struct. Biol.* **2006**, *155*, 294.
14. Marino, M.; Braun, L.; Cossart, P.; Ghosh, P. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 8784.
15. Stumpp, M. T.; Forrer, P.; Binz, H. K.; Pluckthun, A. *J. Mol. Biol.* **2003**, *332*, 471.
16. Marino, M.; Braun, L.; Cossart, P.; Ghosh, P. *Mol. Cell* **1999**, *4*, 1063.
17. Shen, Y.; Naujokas, M.; Park, M.; Ireton, K. *Cell* **2000**, *103*, 501.
18. Freiberg, A.; Machner, M. P.; Pfeil, W.; Schubert, W.-D.; Heinz, D. W.; Seckler, R. *J. Mol. Biol.* **2004**, *337*, 453.
19. Zhang, Y. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 108.
20. Zhang, Y. *BMC Bioinf.* **2008**, *9*, 40.
21. Truhlar, S. M. E.; Komives, E. A. *Structure* **2008**, *16*, 655.
22. Courtemanche, N.; Barrick, D. *Structure* **2008**, *16*, 705.
23. Dawson, P. E.; Muir, T. W.; Clark-Lewis, I.; Kent, S. B. *Science* **1994**, *266*, 776.
24. Bang, D.; Kent, S. B. H. *Angew. Chem., Int. Ed.* **2004**, *43*, 2534.