

Functional β -propeller lectins by tandem duplications of repetitive units

Itamar Yadid and Dan S. Tawfik¹

Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel

¹To whom correspondence should be addressed.
E-mail: tawfik@weizmann.ac.il or dan.tawfik@weizmann.ac.il

Received July 15, 2010; revised July 15, 2010;
accepted July 17, 2010

Edited by Valerie Daggett

Internal symmetry in proteins is likely to be the footprint of evolution by gene duplication and fusion. Like other symmetrical proteins, β -propellers, which are made of 4–10 β -sheet units (blades) circularly arranged around a central tunnel, have probably evolved by duplication and fusion of a rudimentary repetitive unit. However, reproducing the evolution of functional β -propellers by duplication and fusion of repeated units remains a challenge, in particular, because the repeated units must jointly pack to form one hydrophobic core while maintaining intact active sites. As model for generating repeat propellers, we chose tachylectin-2—a highly symmetrical five-bladed β -propeller lectin with five sugar-binding sites. We report the engineering of folded and functional lectins by duplication and fusion of repetitive sequence modules taken from tachylectin-2. The repeated modules comprise three strands of one blade plus one strand of the next blade, thus enabling the closure of the propeller's ring via strand–strand Velcro-like interactions. Duplication and fusion of five modules with the same sequence gave rise to a highly aggregated protein, yet its soluble fraction exhibited lectin function. Subsequently, a library of diversified sequence modules fused in tandem was selected by phage display for glycoprotein binding. A range of new lectins were isolated with binding and biophysical properties that resemble those of wild-type tachylectin-2. These results demonstrate the ability to construct folded and functional globular repeat proteins, and support the role of duplication and fusion of elementary modules in the evolutionary routes that led to the β -propellers fold.

Keywords: beta-propeller/evolutionary precursors/gene duplication/protein evolution/symmetrical folds

Introduction

β -Propellers are made of repetitive structural units, or modules, dubbed blades. The blades are tightly packed in a circular array to form the intact propeller. Despite their remarkable structural similarity—all blades comprise a four-stranded anti-parallel β -sheet—the sequence similarity between blades of the same propeller, and between blades of different propellers, can vary from near identity to almost no similarity

(Chaudhuri *et al.*, 2008). β -Propellers are abundant in all *taxa*, and are highly diverse in function, being represented among enzymes, receptors and cell cycle regulators (Jawad and Paoli, 2002). The modular origin of the β -propeller fold is also manifested in the variation in the number of blades (structures with 4 to 10 blades have been reported), and has the potential to form functional propellers via homo-oligomerization of monomers comprising only two blades (Kostlanova *et al.*, 2005; Yadid and Tawfik, 2007). The high structural symmetry observed in β -propellers is attributed to their evolutionary origin by gene duplication and fusion of small rudimentary units that initially assembled via oligomerization to give a functional protein (Vellieux *et al.*, 1989; Murzin, 1992; Nikkhah *et al.*, 2006; Chaudhuri *et al.*, 2008). This evolutionary scenario was supported by the isolation of smaller fragments of an existing β -propeller that spontaneously assembled into functional pentamers (Yadid and Tawfik, 2007; Yadid *et al.*, 2010). Previous reports also described the construction of propellers by duplication and fusion of a consensus repetitive unit based on the WD motif to yield soluble, yet non-functional, proteins (Nikkhah *et al.*, 2006). We therefore sought to construct functional homologues of an existing β -propeller through duplication and fusion of identical, or nearly identical, sequence modules.

We used tachylectin-2 as our starting point. Tachylectin-2 is a 236 amino acid, five-bladed β -propeller that is part of the innate immune system of *Tachypleus tridentatus*—the Japanese Horse Shoe Crab (Okino *et al.*, 1995). The five blades of tachylectin-2 are essentially identical in structure, and five binding sites for *N*-acetyl-D-Glucosamine (GlcNAc), or *N*-Acetyl-D-Galactosamine (GalNAc), are located between the blades (Beisel *et al.*, 1999). The multiple binding sites mediate multivalent binding of glycans or glycoproteins, and thereby promote cell agglutination. Five sequence repeats are observed in tachylectin-2 that do not directly coincide with the five blades, but instead are shifted by one β -strand (Fig. 1A). This shift results in the annealing of the C- and N-termini into one blade, thus forming the 'Velcro' closure of the circular propeller fold as observed in most β -propellers (Fulop and Jones, 1999) (Fig. 1B). The sequence similarity between the sequence repeats of tachylectin-2 is exceptionally high, possibly due to the presence of five identical binding sites. Indeed, sequence similarity is higher at the sugar-binding sites (57–71% identity), although the β -propeller scaffold also shows relatively high similarity (34–65% identity). At the level of primary sequence, having the binding sites encoded in their entirety within individual sequence repeats makes the assembly of functional propellers easier. However, at the structure level, the binding sites reside at the interfaces between blades, and the lectin's function therefore depends on the correct assembly of the blades within a correctly folded propeller.

Given the high internal sequence homology, tachylectin-2 could be classified as a 'repeat protein' (Andrade *et al.*, 2001). However, most repeat proteins adopt elongated non-globular

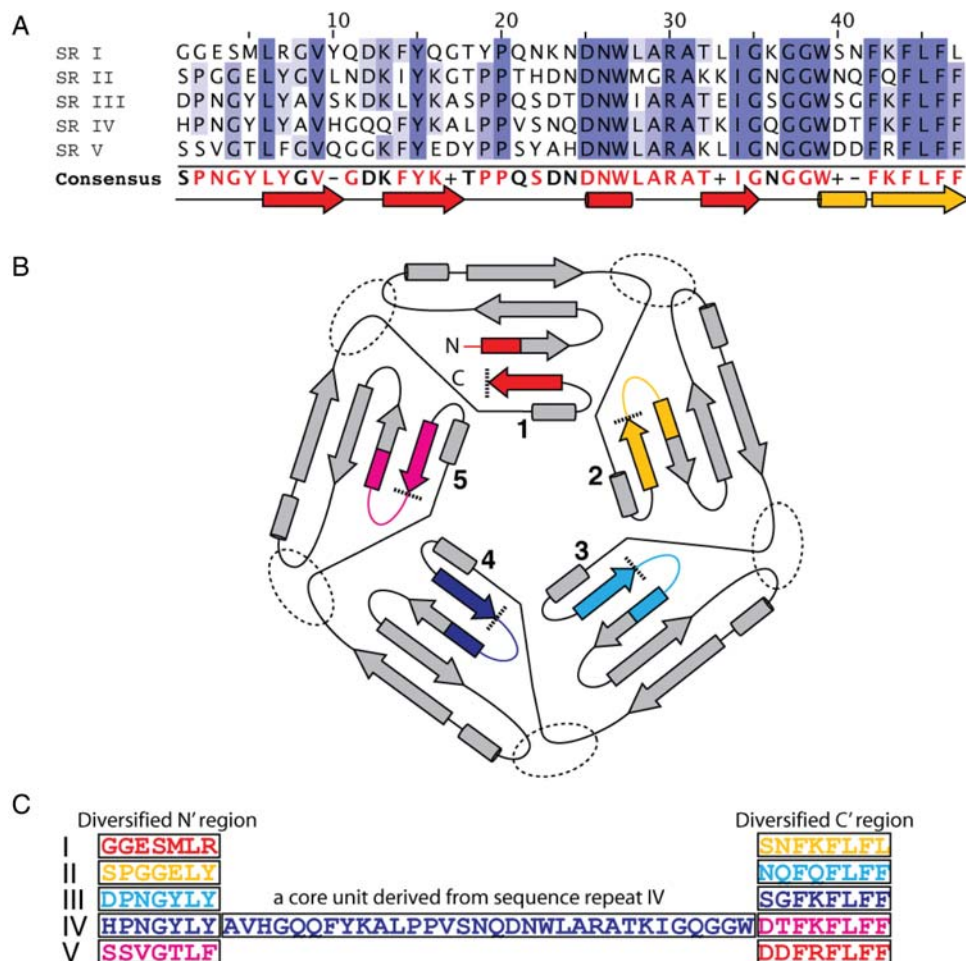


Fig. 1. Library design. (A) Alignment of the five sequence repeats from tachylectin-2. In red, consensus residues that are identical to those in sequence repeat IV. The ribbon diagram represents the secondary elements of the first blade of tachylectin-2. (B) Secondary structure representation of the structure of tachylectin-2 (Beisel *et al.*, 1999). The diversified regions are coloured according to blades. Blades are numbered from 1 to 5. Dashed circles indicate the five sugar-binding sites, and dashed lines indicate the ends of the sequence repeats. (C) Sequences of the library repeats. The central rectangle corresponds to the sequence of the conserved core taken from repeat IV. To the left and right are the flanking regions taken from repeats I to V (repeats I, IV and V were used to construct Lib, and all five repeats were used to construct Lib*). The diversified regions are coloured according to the structural repeats of tachylectin-2 as in (B).

structures (Main *et al.*, 2005), whereas in tachylectin-2, and in other β -propellers, the repeating units form long-range interactions to create a single hydrophobic core. The reconstruction of non-globular repeat proteins by fusing various numbers of individual modules was extensively explored, and the evolutionary and biotechnological aspects have been addressed (Binz *et al.*, 2003; Main *et al.*, 2003; Stumpp *et al.*, 2003; Kajander *et al.*, 2006; Wetzel *et al.*, 2008). The construction of globular repeat proteins that are both folded and functional was demonstrated for the $(\beta/\alpha)_8$ -barrel fold by recombination of two different $(\beta/\alpha)_4$ modules (half barrels) taken from natural enzymes (Claren *et al.*, 2009). Described below is the first step towards the construction of functional β -propellers where sugar-bonding lectins were generated via tandem duplication and fusion of five 47-amino acid sequence repeats derived from tachylectin-2.

Materials and methods

Expression of recombinant tachylectin-2 in *E. coli*

Wild-type gene from *Tachypleus tridentatus* (accession no. D45909) was optimized for translation in *E. coli* using the

Wada codon usage table (Wada *et al.*, 1991). The signal sequence of 19 amino acids was removed from the N-terminus and replaced with an ATG codon (coding for methionine) and an NcoI restriction site. No modifications were done at the C-terminus, except for the addition of a NotI restriction site after the original stop codon. The gene was synthesized by Entelecon GmbH and inserted under NcoI and NotI restriction sites into pET32 (Novagen) vector from which the Trx fusion protein and peptide tags had been truncated (pET32-tr), resulting in an expression of the unaltered protein (DNA and amino acid sequences are provided as Supplementary information). The new vector containing the tachylectin-2 gene was named pET32tac.

BL21 (DE3) competent *E. coli* cells were transformed with pET32tac by electroporation and spread on agar plates containing (100 μ g/ml) ampicillin and (1%) glucose. A starter was grown from a single colony in LB medium (5 ml) supplemented with glucose (1%) and ampicillin (100 μ g/ml) at 37°C over-night. Starter cultures were transferred into LB (500 ml) supplemented with ampicillin, and grown to an $OD_{(600nm)}$ of 0.6. Isopropyl β -D-1-thiogalactopyranoside (IPTG) was added to a final concentration of 0.1 mM and the

culture shaken for 4.5 h at 37°C. Cells were harvested at 2500×g for 20 min and the bacteria pellet resuspended in lysis buffer [50 ml; 20 mM Tris pH 7.5, 150 mM NaCl (TBS), plus one tablet of complete protease inhibitor (Roche)]. Cells were then sonicated at 40% of the maximum intensity (Branson Sonicator Cell Disrupter model 130, Branson Ultrasonics, Danbury, CT, USA) three times for 30 s with 30 s intervals on ice, and then centrifuged 20 min at 18 000×g at 4°C, keeping the supernatant. GlcNAc Agarose (Sigma; 1 ml) was added, and the supernatant was agitated gently at 4°C for 16 h. The agarose beads were loaded on a column and rinsed with 25 bead volumes of TBS (50 mM Tris pH 7.5, 150 mM NaCl). The bound lectins were eluted with aliquots of TBS (1 ml) containing 250 mM GlcNAc added to the beads, incubated at room temperature for 30 min and subsequently collected. Additional aliquots of TBS containing 250 mM GlcNAc were added until no protein was observed by absorption at 280 nm. Protein containing fractions were pooled and dialysed six times against TBS to remove the GlcNAc. The activity of the protein was determined by agglutination and/or enzyme linked lectin assay (ELLA) as described below. The above protocol was used for the expression and purification of wild-type tachylectin-2 and all its engineered variants.

Library design and construction

The gene encoding the perfectly symmetrical repeat lectin was constructed by assembling oligos 1.4, 2, 3.4 and 4 listed in Supplementary data, Fig. S1. Each oligo had an 18-nucleotide region complementary to the next oligo. Correct assembly of the four oligos resulted in a gene segment corresponding to the fourth sequence repeat of tachylectin-2. The 5' of the repeat contained an NcoI restriction site followed by a BsaI site. The 3' contained a BsmBI site followed by a NotI restriction site. The type II-S restriction enzymes (BsaI, BsmBI) that cut away from their recognition site allowed us to form a series of stepwise ligations up to the full length of five repeats. For library construction, two sets of four oligos each, corresponding to the two libraries, Lib and Lib*, were used to assemble the library sequence repeats as described in Supplementary data, Fig. S1.

For assembly, four oligos (100 pmol from each) were mixed with dNTPs (0.05 mM each) and Ex *Taq*TM (Takara, 1 unit) in the enzyme reaction buffer, and subjected to 20 cycles of PCR (30 s at 94°C, 30 s at 55°C and 30 s at 72°C) in a Mastercycler PCR machine (Eppendorf). The product was further amplified by nested PCR using 25 cycles (30 s at 94°C, 30 s at 58°C and 30 s at 72°C) and primers For-type-2 (5'-ACTGCTAGCTGTAGCTGAGC-3') and Universal (5'-GGATGGCAGCAGAACCATG-3'). The resulting PCR product was digested with NcoI and NotI, and cloned into a modified pTZ18 (Hamersham) containing the NcoI–NotI restriction sites (pTZn), to give pTZn-1SR (one sequence repeat). Following transformation and plasmid isolation, the pTZn-1SR plasmid was digested with BsmBI and NotI to produce the accepting vector and the digested nested PCR product (digested with BsaI and NotI) was used as a donor in the ligation [200 units of T4 ligase (NEB) in the enzyme reaction buffer for 16 h at 16°C]. The resulting vector, pTZn-2SR, contained two in-frame sequence repeats. By repeating the ligation step, the sequence repeats were duplicated five times to give the Lib and Lib5* libraries (Fig. 2).

The 2 + 3lib and 3 + 2lib libraries were constructed as follows. Wild-type sequences with a BsmBI restriction site at the 3'-end of the second or third repeat, respectively, were synthesized by Entelecon GmbH. The BsmBI site allowed removal of the stop codon and the creation of sticky ends that were complementary to the library repeats described above. The sequence repeats were digested with NotI and BsaI and subsequently cloned into NotI and BsmBI digested pTZn3bWT to give the plasmid pTZn3 + 2lib. For expression and screening, the library was recloned into pET32-tr using NcoI and NotI restriction sites.

Haemagglutination

Human type A red blood cells (2 ml, obtained from the Israeli blood bank) were rinsed three times with TBS (10 ml) by centrifugation for 5 min at 600×g at 4°C, and diluted to a final volume of 40 ml TBS. Cells (9 ml) were supplemented with 1 ml of 1% w/v Bovine Trypsin (Sigma) in TBS, and incubated for 1 h at 37°C. The cells were then washed five times with TBS (10 ml) and diluted to a final volume of 10 ml. Purified protein samples, or bacterial lysates (100 μ l), at 2-fold serial dilutions in TBS were mixed with 1% trypsinized cell suspension (100 μ l) in V-shaped-bottom microtiter plates (Nunc). Haemagglutination was indicated by the maintenance of cell suspension after 1 h incubation at room temperature (in oppose to a cell pellet at the bottom of the control wells). The titre was defined as the highest endpoint dilution that could induce agglutination. To test for inhibitors, a set of carbohydrate or glycoprotein samples in TBS (50 μ l) was pre-mixed with the lectin samples (50 μ l, at twice the minimum lectin concentration required for agglutination) for 15 min before adding the trypsinized red blood cells (100 μ l). Cited is the minimum inhibitor concentration required for complete inhibition of the haemagglutination activity. Large variations in the agglutination response were observed between cell batches due to different blood donors (up to 10-fold, in the most variable cases). Nevertheless, the relative titres for different samples, e.g. wild-type various engineered variants, were reproducible.

Enzyme linked lectin assay

Cell lysates were incubated on 96 MaxiSorp plates (Nunc) coated with mucin (0.1 ml of 5 μ g/ml porcine stomach mucin type II, Sigma, in PBS) and blocked with BSA (2% for 1 h). Following 3 rinses with PBS-0.5% tween, the bound lectins were detected using a polyclonal rabbit anti tachylectin-2 polyclonal serum (produced at our institute, by immunization with recombinant tachylectin-2), and goat-anti-rabbit HRP antibodies (Jackson).

Selection of libraries by phage display

The assembled genes were cloned into a modified ampicillin-resistant phagemid (Sidhu *et al.*, 2000) named pDAarI5B, where an AarI restriction site was added to allow cloning of the libraries in frame with a truncated pIII protein at the C-terminus, and with a periplasmic secretion signal at the N-terminus. For the production of phage-lectin particles for the first round of selection, pDAarI5Blib and pDAarI5Blib*, were transformed into *E.coli* TG1 cells at an efficiency of $\sim 10^7$ transformants per library. Cells were grown in 2XYT supplemented with 100 μ g/ml ampicillin and 1% glucose to OD₆₀₀ \approx 0.4. At this stage, 2×10^{11} of M13KO7 helper

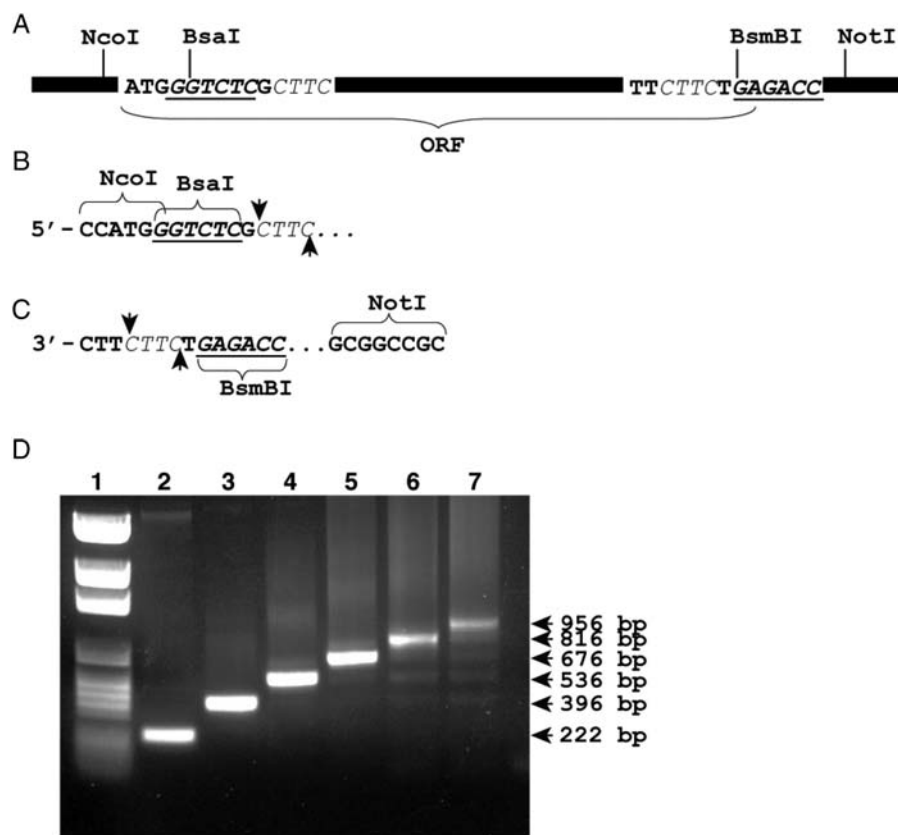


Fig. 2. Construction of the sequence repeat libraries. (A) The modular units, with restriction sites as noted, and open reading frames that correspond to one sequence repeat of tachylectin-2 (for their construction by assembly PCR, see Materials and Methods section (Supplementary data, Fig. S1)). (B) The 5'-end of the modular units, showing the overlapping NcoI and BsaI restriction sites. The BsaI site is underlined, and the two black arrow tips mark the cut site. Restriction leaves a 'sticky end' for ligation to the plasmid. Ligation reproduces the ATG site at the 5'-end, and inserts the 3'-end in-frame with the next unit. (C) The 3'-end of the modular unit, showing NotI and BsmBI restriction sites. The BsmBI site is underlined, and the black arrow tips mark the cut site. Restriction removes the TGA stop codon and leaves a sticky end for ligation of the next unit (the last unit inserts in-frame with a TGA stop codon appended by the plasmid). (D) PCR products amplified by primers that anneal to the pTZn plasmid up- and down-stream to the cloned library repeats (see Materials and Methods) analysed on 1% agarose gel and stained with ethidium bromide. Lane designation: 1, DNA marker (pGEM, NEB); 2, amplification from pTZn (the 'empty' plasmid used for library construction); 3, amplification from pTZn carrying a library encoding one sequence repeats, two sequence repeats (lane 4); three sequence repeats (lane 5); four sequence repeats (lane 6) and five sequence repeats of library 5Lib (lane 7).

phages (NEB) were added to 50 ml of *E. coli* culture and incubated at 37°C in a water bath. After 30 min, cells were harvested by centrifugation at 3300×g for 10 min and the pellet was resuspended in 250 ml of 2XYT supplemented with 100 µg/ml ampicillin, 50 µg/ml kanamycin and 0.1% glucose. The cells were grown at 20°C and 250 rpm. After 48 h, cells were harvested and PEG/NaCl (20% PEG 6000 2.5 M NaCl) in a 1:5 ratio was added to the supernatant. The mixture was left on ice for 1 h and phage particles were collected by centrifugation at 3300×g for 30 min. The pellet was resuspended in 4 ml of PBS and centrifuged at 12 000×g for 10 min to remove cell debris. Phage concentration (CFUs) was determined by titration and by absorption at 268 nm. For the selection of functional lectins, maxisorp plates (Nunc) were coated with 5 µg/ml mucin in PBS, washed and blocked with 2% milk in PBS. Approximately 10¹¹ phage-lectin particles were incubated on the plate at room temperature for 1 h and washed thoroughly with PBS and 0.05% Tween. Phage particles were eluted from the plate with 0.5 M GlcNAc to transfect mid-log phase TG1 *E. coli* cells, and prepare phages for the next round of selection. These panning cycles were conducted three times before individual clones were taken for further characterizations.

Fluorescence measurements

Emission spectra were measured following excitation at 280 nm of lectin variants [0.15 µM in TBS, placed in a 1 cm path quartz cuvette (Hellma)] in a Varian Cary Eclipse fluorescence spectrophotometer at 25°C. In the sugar-binding assay, the fraction of protein molecules bound with the ligand was determined from the shift in fluorescence between the spectrum in the absence of ligand and the spectrum at saturating GlcNAc concentrations (Okino *et al.*, 1995). The ratio between the fluorescence intensities at 330 and 320 nm was used, and the data for different sugar concentrations were fitted using KaleidaGraph to a standard binding isotherm (Copeland, 2000).

$$B = \frac{1}{1 + K_d/[L]}; \quad (1)$$

where B is the fraction of lectin molecules occupied with ligand, K_d is the dissociation constant and $[L]$ is the sugar concentration.

Guanidinium hydrochloride (GdnHCl) denaturation

Unfolding of lectin variants was determined by the change in fluorescence emission at 326 nm after excitation at 285 nm, under the same conditions as above. Seven molar GdnHCl equilibrated to pH 7.5 in TBS was used as a stock solution to prepare the desired GdnHCl concentrations. After addition of protein, the reactions were incubated at room temperature for 15 min, prior to the measuring of fluorescence (no further changes in fluorescence were observed at incubation longer than 15 min). Buffer and GdnHCl fluorescence were subtracted, and the intensity in arbitrary units was plotted against the GdnHCl concentration. Data were fitted by the linear extrapolation method by Santoro and Bolen (1988), as described in Equation (2), where F is the observed fluorescence signal at a given GdnHCl concentration, α_N the signal at 0 M GdnHCl (fully folded protein), β_N the slope for the folded state of the graph and α_D and β_D the corresponding quantities for the denatured state.

$$F = \frac{(\alpha_N + \beta_N[D]) + (\alpha_D + \beta_D[D]) \exp[m_{D-N}([D] - D^{50\%})/RT]}{1 + \exp[m_{D-N}([D] - D^{50\%})/RT]}$$

Thermal stability

Far-UV CD spectra at each temperature were obtained with a circular dichroism spectropolarimeter (Aviv instruments, Model 202) in a quartz cell with a path length of 0.1 cm (260–200 nm, sampling every 1 nm and an averaging time of 3 s). Protein samples were dialysed against 100 mM sodium phosphate buffer pH 7.4, and diluted into a final concentration of 7 μ M in the same buffer. The change in signal at 222 nm related to the fraction of unfolded lectin and data were fit as described (Fersht, 1999; Greenfield, 2006). We also determined the residual binding activity after incubation of the protein at different temperatures at the same concentration as the CD experiments. Protein samples in phosphate buffer were incubated in a PCR cycler at the designated temperatures for 1 min at each temperature, and then placed on ice. The samples were subsequently diluted, and analysed for activity by ELLA on mucin-coated plates at twice the concentration required for 50% ELLA signal. Residual activity was defined as the fraction of the ELLA signal as compared to the signal at 30°C and data were fitted to Equation (2) using residual activity rather than fluorescence values.

Results

The perfectly symmetric repeat protein

As in most propellers, the first blade in tachylectin-2 is not encoded as one, continuous sequence stretch and the sequence repeats are therefore shifted by one strand (Fig. 1A). Since our objective was duplication and fusion at the gene level, we chose the sequence repeats and not the blades to serve as the modular, repetitive element of our design.

Alignment of the five 47-amino acid sequence repeats of wild-type tachylectin-2 (49–68% identity) indicated that the third and fourth sequence repeats represent the consensus best (Fig. 1A). We chose the fourth repeat to serve as the core element for library design. To mimic the simplest

evolutionary scenario that results in a perfectly symmetric and repetitive protein, we constructed a five-bladed protein by fusing in tandem five identical copies of the fourth repeat. The resulting protein was active as shown by an enzyme-linked lectin assay (ELLA; using the glycoprotein mucin that is avidly bound by tachylectin-2 coated on plates, and detection of binding with polyclonal anti-tachylectin-2 antibodies; Supplementary data, Fig. S2). Although it could be affinity purified on GlcNAc-Agarose beads, this perfectly symmetrical repeat protein gave mostly insoluble inclusion bodies (>99%; Supplementary data, Fig. S3) and showed high tendency for aggregation after purification.

Library design and making

To obtain better folded lectins, we constructed a library by diversifying two regions of the fourth repeat—a block of seven amino acids at the N-terminus and a block of eight amino acids at the C-terminus (Fig. 1C). In wild-type tachylectin-2, these regions connect the sequence repeats and mediate interactions between them, yet they do not include the sugar-binding site. The diversification of the sequence repeats was therefore expected to increase the likelihood of obtaining correctly folded and functional lectins, also owing to the effect of reducing sequence symmetry (Wright *et al.*, 2005).

Initially, we created library repeats encoding blocks from the flanking regions of three different repeats (I, IV, V), thus generating nine different combinations per repeat, and an overall diversity of 9^5 ($\sim 6 \times 10^4$) assembled genes. These libraries are marked as 'Lib'. Later, we created library repeats encoding blocks from all five repeats (marked as Lib*), thus generating 25 different combinations per repeat, and an overall diversity of 25^5 ($\sim 10^7$) assembled genes. Four different libraries were made. The first two made were $3 + 2Lib$ where two diversified library repeats were fused to the original tachylectin-2 sequence at the C-terminus of its third sequence repeat, and $2 + 3Lib$ with three diversified repeats fused to the tachylectin-2 sequence at the C-terminus of its second sequence repeat. These two libraries were designed to test the tolerance of the protein to sequence perturbations in a stepwise manner, and examine the level of cooperativity between replaced sequence repeats. Once the tolerance of tachylectin-2 to these replacements was confirmed, two additional libraries were built. Libraries $5Lib$ and $5Lib^*$ comprised of five library repeats (Supplementary data, Fig. S4). The diversified library repeats were assembled by PCR from a set of complementary oligonucleotides (Supplementary data, Fig. S1) as described in the Materials and Methods section, and cloned into a plasmid vector. The assembled genes were then ligated and fused to give the desired number of repeats using recursive cycles of digestion and ligation (Fig. 2).

Functional lectins from sequence repeat libraries

Several hundred individual clones from the $3 + 2Lib$ library were grown in 96-well plates, lysed and assayed for agglutination of human red blood cells. Approximately 20% of the clones were found to be active, and many exhibited expression levels and binding properties (agglutination titre and sugar inhibition patterns) comparable with wild-type tachylectin-2 (e.g. variants et2P2F1 and et2P1H8, Table I). This suggests that nearly half of the diversified sequence

Table I. Function and stability characteristics of representative lectin variants

	Tachylectin-2		3 + 2Lib variants		5Lib variant	5Lib* variants	
	Tachylectin-2 ^a	Recombinant Tachylectin-2	et2P2F1	et2P1H8	G3-3	H1R3	D1R4
Haemagglutination							
Minimum concentration required for agglutination ($\mu\text{g/ml}$)	1.6	0.25	0.39	1.56	12.5	n.d ^b	n.d ^b
GlcNAc ^c	0.097 mM	0.25 mM	0.031 mM	0.062 mM	0.125 mM	n.d ^b	n.d ^b
GalNAc	0.78 mM	5 mM	1.25 mM	1.25 mM	2.5 mM	n.d ^b	n.d ^b
Glucosamine	NI ^d	NI ^d	12.5 mM	NI ^d	NI ^d	n.d ^b	n.d ^b
Galactosamine	NI ^d	NI ^d	50 mM	NI ^d	NI ^d	n.d ^b	n.d ^b
Mucin	1.95 $\mu\text{g/ml}$	1.56 $\mu\text{g/ml}$	0.78 $\mu\text{g/ml}$	3.125 $\mu\text{g/ml}$	25 $\mu\text{g/ml}$	n.d ^b	n.d ^b
Lactoferrin	–	NI ^e	0.78 $\mu\text{g/ml}$	NI ^e	NI ^e	n.d ^b	n.d ^b
Conalbumin	–	NI ^e	NI ^e	NI ^e	NI ^e	n.d ^b	n.d ^b
Ovomucoid	–	NI ^e	NI ^e	NI ^e	NI ^e	n.d ^b	n.d ^b
ELISA							
50% ELLA signal ^f	–	3.5 \pm 0.4 ng/ml	7.7 \pm 1 ng/ml	–	9.0 \pm 4 ng/ml	15.4 \pm 4 ng/ml	13.3 \pm 3 ng/ml
T ₅₀ activity ^g	–	74.7 \pm 0.7°C	–	–	60.6 \pm 0.5°C	56.3 \pm 0.3°C	54.5 \pm 0.2°C
T ₅₀ activity +10 mM GlcNAc	–	83.9 \pm 0.3°C	–	–	65.8 \pm 0.2°C	61.9 \pm 0.1°C	60.8 \pm 0.1°C
Fluorescence							
K _d (GlcNAc)	–	0.12 \pm 0.02 mM	0.11 \pm 0.02 mM	0.19 \pm 0.02 mM	0.56 \pm 0.06 mM	0.77 \pm 0.1 mM	0.9 \pm 0.1 mM
D ₅₀ GdnHCl	–	3.9 \pm 0.04 M	–1.8 \pm 0.02 M	–2.8 \pm 0.02 M	3.0 \pm 0.03 M	2.16 \pm 0.06 M	2.14 \pm 0.04 M
D ₅₀ GdnHCl ^h + GlcNAc	–	4.76 \pm 0.03 M (1 mM)	–3.1 \pm 0.02 M (1 mM)	–3.7 \pm 0.05 M (1 mM)	3.24 \pm 0.02 M (5 mM) / 4.13 \pm 0.01 M (50 mM)	3.04 \pm 0.02 M (5 mM)	2.86 \pm 0.02 M (5 mM)
CD	T ₅₀ CD ⁱ	72.6 \pm 0.2°C	–	–	59.0 \pm 0.3°C	57.6 \pm 0.3°C	57.0 \pm 0.4°C

^aData derived from Okino *et al.* (1995).

^bThe combination of high agglutination titre and very low solubility prevented the detection of agglutination with these variants, and therefore, of inhibition of agglutination.

^cNoted is the minimum concentration required for complete inhibition of haemagglutination (these values may differ from one batch of red blood cells to another, see Materials and Methods section).

^dNI, no inhibition observed at \leq 50 mM glucosamine or galactosamine.

^eNI, no inhibition observed at \leq 50 $\mu\text{g/ml}$ lactoferrin, conalbumin or ovomucoid.

^fProtein concentration in which 50% of the maximal ELLA binding signal is observed.

^gT₅₀ activity is the temperature at which 50% thermal denaturation is observed as determined by residual binding activity.

^hGlcNAc concentrations applied for D₅₀ measurements in the presence of GdnHCl: wild-type, 1 mM; G3-3, 50 mM; 5Lib* and 'short segments' variants, 5 mM.

ⁱT₅₀ CD is the temperature at which 50% denaturation is observed as determined by CD signal at 222 nm.

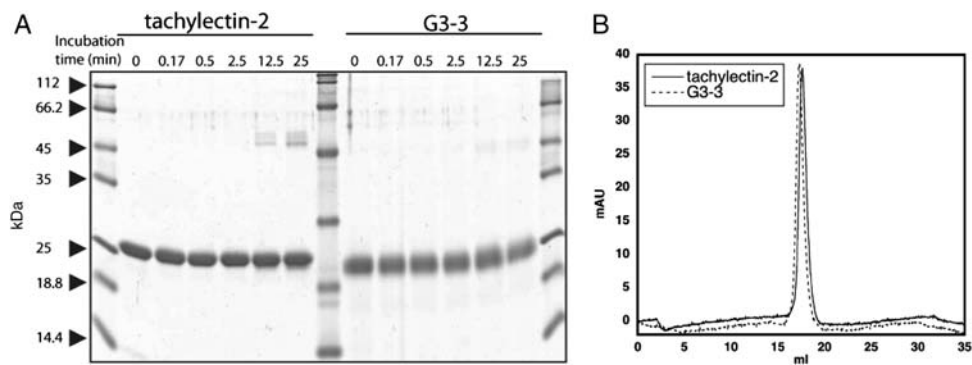


Fig. 3. The monomeric state of the selected library variants. (A) Wild-type tachylectin-2 and duplicated variant G3-3 were cross-linked with 0.01% (v/v) glutaraldehyde for the depicted time points in the presence of 5 mM GlcNAc. The reaction was stopped by the addition of NaBH₄ to a final concentration of 0.01 M, where the protein was analysed by SDS-PAGE –15% (w/v). A single band corresponding to the molecular mass of the monomer (~25 kDa) present across all time points indicated that the duplicated variant as well as tachylectin-2 are monomeric. Black arrow tips indicate the molecular size marker. (B) The oligomerization state was also verified by gel filtration. The elution pattern of wild-type tachylectin-2 and the duplicated variant G3-3 indicate that both proteins are monomeric.

repeats in this library were functional ($0.45^2 \approx 0.2$). It was therefore expected that 9% of the Lib2 + 3 library variants would be active ($0.45^3 \approx 0.09$). However, screening by agglutination indicated less than 0.5% active clones. It seems therefore that in contrast to non-globular repeat proteins (Main *et al.*, 2003; Stumpp *et al.*, 2003), the fusion of modules that interact with each other to form a tightly packed core is highly restricted, and modules (sequence repeats) that are folded and functional on their own become non-functional when placed in the context of other modules.

Given this sharp decline in the proportion of active variants, the libraries composed of five repeats (5Lib, 5Lib*) were selected using high throughput phage display rather than screened by agglutination. Selection was performed by panning of the phage libraries on the glycoprotein mucin as described in Materials and Methods. After three rounds of selection, ELLA indicated that ~20% of the phage clones displayed mucin-binding proteins. The selected pool was re-cloned and expressed in *E.coli* with no fusion partner. PAGE analysis of 25 arbitrarily chosen variants revealed a high occurrence of proteins (~60%) with a molecular weight corresponding to five repeats (25 kDa wild-type size). Three of the 15 full-length variants were subsequently characterized and were all found to be active lectins with properties similar to wild-type tachylectin-2. Indeed, variants G3-3, H1R3 and D1R4 were randomly picked from the pool of full-length ELLA positive clones (Table I). Although these variants varied in their solubility levels (H1R3 and D1R4 exhibited very low soluble expression, whereas G3-3 expressed at comparable levels to wild-type tachylectin-2), they exhibited gel filtration and glutaraldehyde cross-linking patterns indicating that they are compact, monomeric proteins (Fig. 3). The CD and UV spectra of G3-3 and wild type were also essentially identical, and the analysis of the former indicated ~90% beta-sheet content. While all the analysed full-length variants were found to be monomeric, five-bladed, wild-type-like lectins, fragments with fewer than five blades were also selected. These are likely to be functional via oligomer formation, as previously observed with tachylectin-2-derived fragments (Yadid and Tawfik, 2007), and with other protein fragments selected by phage display (de Bono *et al.*, 2005).

Table II. The flanking regions identified in sequence repeats of 3 + 2lib variants

Non-functional variants					Functional variants				
Variable position	IV	IV	V	V	Variable position	IV	IV	V	V
	N'	C'	N'	C'		N'	C'	N'	C'
Variant name					Variant name				
et2P1E5	I	V	IV	V	et2P1B3	IV	I	V	IV
et2P2D3	V	IV	IV	IV	et2P1A8	IV	IV	V	IV
et2P2D4	I	I	I	IV	et2P1F8	IV	IV	IV	I
et2P2F5	V	I	IV	V	et2P1H8	IV	IV	V	V
et2P2A6	I	I	I	IV	et2P1C12	IV	I	IV	V
et2P2G9	I	IV	I	V	et2P2F1	V	I	IV	I
					et2P2H2	IV	I	IV	V
					et2P2B7	IV	V	I	V
					et2P2C7	IV	V	IV	V
%I	66.6	50	50	0		0	33	11	22
%IV	0	33.3	50	50		89	33	55	22
%V	33.3	16.6	0	50		11	33	33	55

Sequence composition of the repeat lectins

Sequence analysis of 3 + 2lib functional variants revealed that the N-terminus of the fourth duplicated repeat (the first diversified repeat) was almost exclusively (88.8%) the N-terminus of the wild-type fourth sequence repeat (Table II). Concomitantly, in the variants that lacked haemagglutination activity, the N-terminus was taken from sequence repeats V or I, but not from repeat IV. At the C-terminus of the selected sequence repeats, there was no clear trend since the fragments were distributed almost equally between the C-terminus of repeats I, IV and V. At the N-terminus of the fifth duplicated sequence repeat, there was no clear trend either, although the fragment from the wild-type repeat I was present only once. At the C-terminus of the fifth sequence repeat, there were no significant biases although the flanking regions taken from repeats IV and V were overrepresented. The presence of the N-terminus of repeat IV in the beginning of the fourth duplicated repeat actually completes the original sequence all the way down to the C-terminus of the fourth blade. This N-terminus probably interacts optimally with the C-terminus from the third repeat

Table III. The flanking regions identified in sequence repeats of 5Lib variants (functional variants isolated from the third round of phage-display selection)

Name	SR									
	I		II		III		IV		V	
	N'	C'	N'	C'	N'	C'	N'	C'	N'	C'
G3-1	I	V	IV	IV	IV	V	IV	IV	V	IV
G3-3	I	IV	IV	V	IV	V	IV	V	IV	IV
G3-5	I	IV	IV	IV	IV	IV	V	V	IV	V
G3-6	V	IV	IV	V	IV	IV	IV	IV	I	IV
G3-101	V	V	IV	I	IV	V	IV	V	IV	V
G3-103	V	IV	IV	V	IV	IV	IV	IV	I	IV
G3-104	V	IV	IV	V	IV	V	IV	V	I	IV
G3-108	I	V	IV	V	IV	V	IV	V	IV	I
%I	50	0	0	12.5	0	0	0	0	37.5	12.5
%IV	0	62.5	100	25	100	37.5	87.5	37.5	50	62.5
%V	50	37.5	0	62.5	0	62.5	12.5	62.5	12.5	25

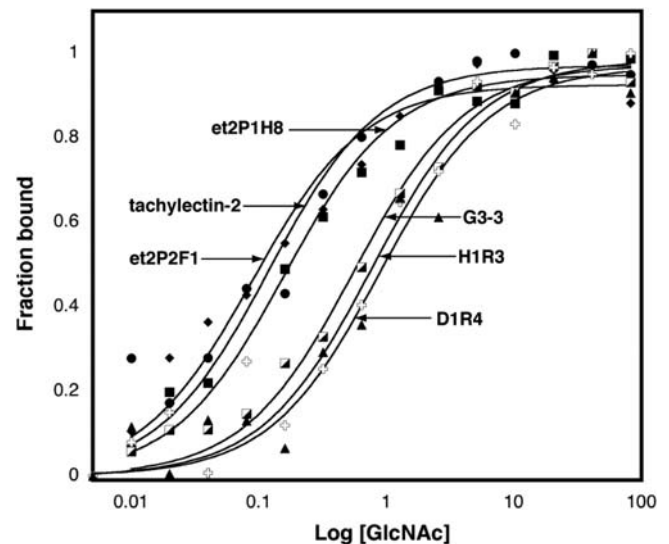
Table IV. Flanking regions identified in sequence repeats in 5Lib* variants (functional variants isolated from the third and fourth rounds of phage display selections)

Name	SR									
	I		II		III		IV		V	
	N'	C'	N'	C'	N'	C'	N'	C'	N'	C'
C5r3	I	I	IV	V	III	V	III	III	II	V
H1r3	II	IV	III	IV	IV	II	III	IV	II	II
D1r4	III	IV	IV	II	III	IV	III	II	III	I
A11r3 ^a	III	III	II	V	IV	V	III	IV	II	V
F10r3 ^a	I	I	III	V	III	V	IV	IV	III	V
D9r3	I	IV	III	II	III	I	III	V	IV	III
D2r3	V	I	III	V	II	V	III	IV	III	IV
C8r3	I	III	II	V	II	V	IV	V	IV	III
%I	37.5	37.5	0	0	0	12.5	0	0	0	12.5
%II	12.5	0	25	25	25	12.5	0	12.5	37.5	12.5
%III	12.5	25	50	0	50	0	75	12.5	37.5	25
%IV	0	37.5	25	12.5	25	12.5	25	50	25	12.5
%V	12.5	0	0	62.5	0	62.5	0	25	0	37.5

^aThe first sequence repeats of these variants contained frame shifts (a double base deletion after 111 bp in A11r3, and after 115 bp in F10r3). Assuming the nearest start codon, the mature proteins are ≥ 37 amino acids shorter. Nonetheless, phage particles encoding these variants exhibited mucin binding, possibly via oligomerization as observed with shorter tachylectin-2 segments (Yadid and Tawfik, 2007), and as observed with other phage-display proteins (de Bono *et al.*, 2005).

in the wild type, and thereby completes the third blade. The same consideration probably applies to the C-terminus of the fifth repeat. This flanking region interacts with the N-terminus enabling the 'Velcro' ring closure. Nevertheless, alternatives such as the C-terminus region of repeat IV were also tolerated at the C-terminus.

Sequencing of functional variants from the 5Lib and 5Lib* libraries revealed similar trends (Table III and Table IV, and Supplementary data, Tables S1 and S2). The N-terminus of the internal repeats (second to fourth repeat) was taken mostly from the wild-type repeat IV in 5Lib, and repeats II, III and IV in 5Lib* (note that repeats II and III were not present in 5Lib). Thus, as largely observed with 3 + 2Lib, the N-terminus flanking regions from repeats I and V were excluded. The regions comprising the 'Velcro' closure (the

**Fig. 4.** Sugar-binding isotherms. GlcNAc binding isotherms on a semi-log scale. The data were derived from the change in fluorescence between the unbound form of recombinant tachylectin-2 and the duplicated variants, and the fluorescence of the proteins at GlcNAc saturation (>80 mM). Data were fit to a binding isotherm [Equation (1)]. The resulting affinities (K_d values) are summarized in Table I.

N-terminus flanking region of the first repeat, and the C-terminus flanking region of the fifth repeat) show no distinct preferences apart from the exclusion of the fourth flanking region as observed in Lib3 + 2. While the above observations are clear, more subtle trends regarding the preference of certain termini could exist but their detection requires the analysis of many more active and inactive library variants.

Binding characteristics

A detailed analysis of the biophysical and binding properties of the duplicated variants was performed to confirm that these are folded, stable and functional proteins that resemble wild-type tachylectin-2. The binding characteristics of the reconstructed variants were analysed by ELLA for mucin binding, and by haemagglutination of red blood cells, both of which are efficiently executed by wild-type tachylectin-2. Agglutination is entirely dependent on multivalent binding, and could therefore report the presence of multiple binding sites on individual protein variants. Inhibition of both agglutination and mucin binding by a range of saccharides and glycoproteins enabled the comparison of the specificity of the duplicated variants with that of tachylectin-2. Fluorescence spectra were also recorded to determine the ligand affinity of individual sites. Following excitation at 280 nm, both tachylectin-2 and the duplicated variants produced a typical fluorescence emission spectrum with a single maximum peak at 330 nm. This peak blue-shifted to 320 nm upon binding of GlcNAc (Okino *et al.*, 1995). The fraction of bound ligand corresponds to the shift measured for each GlcNAc concentration, and the resulting dissociation constants (K_d) were derived by fitting the data to a standard binding isotherm (Copeland, 2000) (Fig. 4). These tests indicated that most of the selected library variants exhibit binding affinities (for individual sites, as measured by fluorescence shifts), avidities (indicated by haemagglutination and ELLA titres) and specificities (indicated by the

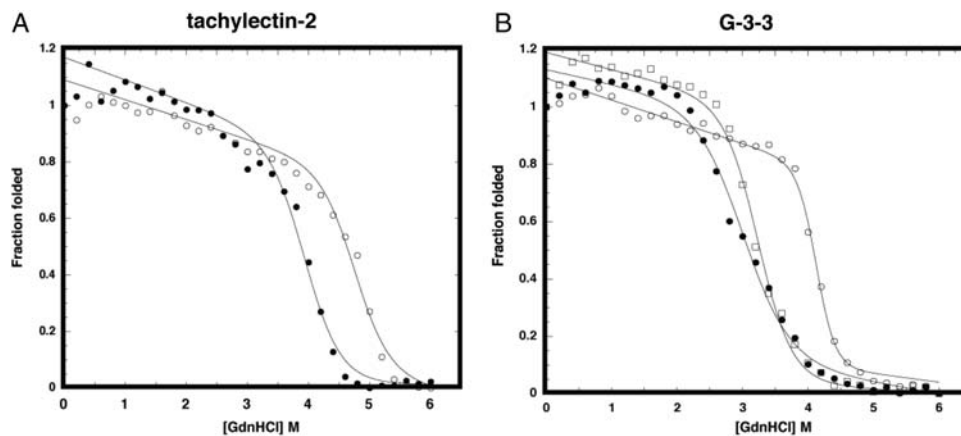


Fig. 5. Chemical denaturation curves. Presented is the fraction of folded protein versus GdnHCl (Guanidinium hydrochloride) concentration. Data were fitted to a two-state model equation [Equation (2), in Materials and Methods]. (A) Melting profiles for recombinant tachylectin-2 (closed circle) and tachylectin-2 in the presence of 1 mM GlcNAc (open circle). (B) Melting profiles of the 5Lib variant G3-3 (closed circle), and variant G3-3 in the presence of 5 mM (open square), or 50 mM GlcNAc (open circle).

inhibition by various saccharides and glycoproteins) that are similar to those of wild-type tachylectin-2 expressed in *E. coli* (Table I). The latter showed close but not identical values to those reported for tachylectin-2 purified directly from *Tachypleus tridentatus* (Okino *et al.*, 1995).

Distinct exceptions were present in variants selected from libraries 5Lib and 5Lib*. These variants seem to be systematically inferior, both in terms of folding and binding. The latter could be attributed to the lower affinity of individual binding sites to GlcNAc, but also to the existence of less than five functional sites. In addition, variants from Lib5* could not be concentrated enough for the agglutination assays (see footnotes in Table I).

Stability measurements

Equilibrium unfolding of wild-type tachylectin-2 and its duplicated variants was followed by monitoring the protein's fluorescence at different guanidinium hydrochloride (GdnHCl) concentrations. Data were analysed by linear extrapolation (Fersht, 1999) [Equation (2)]. Because the denaturation plots of both tachylectin-2 and its duplicated variants deviated systematically from the standard two-state model (Fig. 5), we only derived the apparent D_{50} values from these fits (Table I).

All the duplicated variants exhibited lower D_{50} values than tachylectin-2, although the addition of the ligand GlcNAc induced higher stability of both tachylectin-2 and the duplicated variants. Following complete denaturation and dilution into buffer with no denaturant, wild-type tachylectin-2 refolded to its native state with 80% yield. In contrast, much smaller fractions of the duplicated variants ($\leq 50\%$) were refolded (Fig. 6). Interestingly, the addition of ligand (GlcNAc) during refolding markedly improves the process.

Two methods were used to monitor thermal denaturation: circular dichroism (CD) signal change at 222 nm and residual binding activity. In agreement with the chemical denaturation values, the T_m values obtained by both methods indicated the lower stability of the duplicated variants (Supplementary data, Fig. S5 and Table S1). It should be noted, however, that T_m value was measured by residual activity. The transition of the former was also unusually abrupt. Since the binding was measured after cooling the sample to ambient temperature, it is possible that a fraction

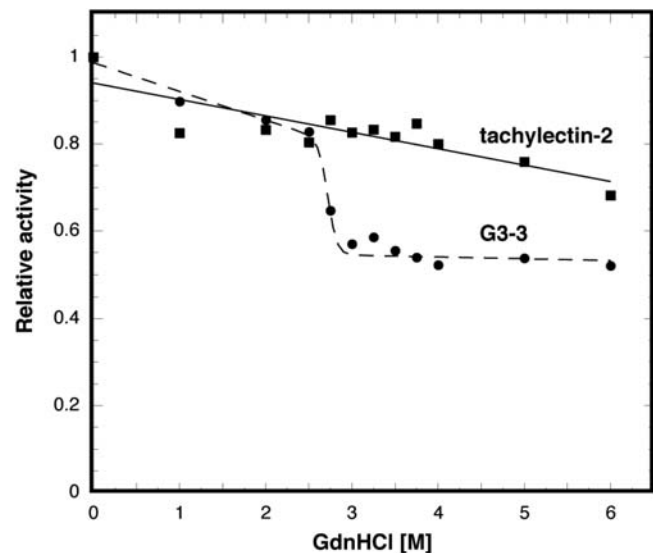


Fig. 6. Refolding to the native state. Wild-type tachylectin-2 and the duplicated variant G3-3 isolated from Lib5 were unfolded using various concentrations of GdnHCl. Refolding was induced by dilution into buffer with no denaturant. The fraction of properly refolded protein was determined by measuring the residual mucin-binding activity by ELLA. Upon complete denaturation, wild-type tachylectin-2 gave $\geq 80\%$ refolding, while G3-3 gave $\sim 50\%$ refolding.

of protein molecules at the transition point were able to refold (in particular wild type; the differences for the library variants were insignificant). Complete thermal denaturation (both with and without GlcNAc) led to irreversible unfolding of both wild-type and the duplicated variants.

Discussion

This article and previous works [for examples and reviews see (Nikkhah *et al.*, 2006; Peisajovich *et al.*, 2006; Vogel and Morea, 2006; Arnold *et al.*, 2007; Yadid and Tawfik, 2007; Bharat *et al.*, 2008; Edwards *et al.*, 2008; Huang *et al.*, 2008; Claren *et al.*, 2009; Richter *et al.*, 2010)] explore one of Nature's most abundant and oft-used evolutionary routes—the generation of new protein topologies and folds via the truncation, duplication and fusion of gene segments that

encode discrete structural modules (Patthy, 1999; Grishin, 2001; Abraham *et al.*, 2009; Koide, 2009; Yadid *et al.*, 2010). An obvious obstacle in reconstructing the evolutionary intermediates that may have led to contemporary proteins is the extensive changes in sequence, and sometimes in structure, that occurred since these proteins emerged. Interestingly, tachylectin-2 is found in the horseshoe crab *Tachypleus tridentatus*—an organism that is believed to date back 500 million years, and has been declared a ‘living fossil’ (Graur and Li, 2000). Although the sequence of tachylectin-2 is likely to have drifted considerably, a relatively high degree of internal symmetry has been maintained, making tachylectin-2 an ideal model. Our results indicate that putative evolutionary progenitors of tachylectin-2 can be reconstructed via the duplication and tandem fusion of five, identical, or nearly identical modules. Notably, duplication and fusion of the very same module (repeat IV) gave rise to a functional protein, although it was mostly aggregated. The strong aggregation tendency seen also in wild-type tachylectin-2, and in the library variants, is probably the outcome of high β -strand content, but also of having such high sequence symmetry (Wright *et al.*, 2005). Our results suggest that braking the symmetry by introducing diversity at the repeats’ ends improves folding, soluble expression and binding activity. Indeed, the library variants show higher sequence identity between repeats than in wild-type tachylectin-2 (75% on average, versus 55%), but lower than 100% identity as the case with variant made by fusion of five identical repeats. The solubility and binding function of the latter could also be improved by screening of a library of random mutations (Yadid *et al.*, unpublished results).

The duplicated and fused module, or unit, applied here is a 47-amino acid sequence repeat observed in tachylectin-2’s sequence (Fig. 1A) (Beisel *et al.*, 1999). These repeats are systematically shifted by one β -strand relative to the five structural modules of the propeller (i.e. its five blades; Fig. 1A). This arrangement of tachylectin-2’s sequence repeats guided our library design—the core of the library’s sequence repeats was identical and therefore retained the essential sugar-binding site. The sequence diversification was restricted to the N- and C-termini (Fig. 1C). Indeed, we observed a great deal of sequence diversity in the termini of all five repeats (Tables II–IV). Overall, sequence differences of 31–37% were observed between wild-type tachylectin-2 and the functional variants arising from Lib5 and Lib5*. As expected, these differences are mostly in the N- and C-regions flanking the highly conserved sugar-binding domain (at these regions, the sequence divergence from wild-type tachylectin-2 reached 61%). Despite these large sequence variations and the fact that the binding site is located between two duplicated repeats, the engineered proteins retain similar functional and biophysical properties to those of wild-type tachylectin-2 (Table I). This sequence divergence is in agreement with the generally low sequence similarities observed between different propellers and between different blades of the same propeller (Chaudhuri *et al.*, 2008).

The observed variability at the N-terminus of repeat I and the C-terminus of repeat V (Table IV) is particularly surprising given that these regions mediate the ‘Velcro’ closure of the entire chain and, as such, are deemed essential for the stability of the propeller’s fold (Neer and Smith, 1996).

However, recent findings suggest that closure of the N- and C-termini is not as crucial as packing at the interfaces between blades. Indeed, the overall structure of the β -propeller domain of Tup1 could be retained despite removal of the N- and C-terminus blades that mediate this closure (Zhang *et al.*, 2002). The existence of a naturally occurring (Kostlanova *et al.*, 2005) and laboratory evolved (Yadid *et al.*, 2010) oligomeric propellers that assemble via blade-to-blade interactions is also indicative of the fact that the Velcro closure provides additional stability but is not as crucial for packing as the blade interfaces.

Our study therefore identifies putative evolutionary routes leading to functional β -propeller lectins via the attachment of smaller modules—i.e. by fusion of duplicated gene segments. That such modules can be functional on their own is indicated by the fact that \sim 100-amino acid long fragments truncated from tachylectin-2 spontaneously assemble into pentamers while retaining the sugar-binding function (Yadid and Tawfik, 2007). In length, these fragments correspond to two modules. In topology, they overlap with blades in one case, and with another topology that comprises neither blades nor sequence repeats. Further, the subunits were found to adopt multiple backbone structures (metamorphism), and thus the pentameric forms are homo-oligomers in sequence but hetero-oligomers in structure (Yadid *et al.*, 2010). Oligomerization is therefore one way by which modules or repetitive units seen in today’s monomeric single domain proteins could assemble to give a functional protein (D’Alessio, 1999; Richter *et al.*, 2010). Duplication and fusion followed by various permutations (new starts and stops) could then lead to the highly symmetrical folds such as β -propellers and β/α -TIM barrels we see today (Vogel and Morea, 2006; Weiner and Bornberg-Bauer, 2006; Abraham *et al.*, 2009).

Our study also suggests the feasibility of engineering globular repeat proteins. The engineering of repeat proteins has thus far been limited to the fusion of modules that comprise stand-alone domains, or the attachment of elongated non-globular proteins (Binz *et al.*, 2003; Main *et al.*, 2003; Stumpp *et al.*, 2003). The engineering of repeat proteins may be expanded to the fusion of interacting modules, although, as shown here, the joining of such modules is more restricted. This was manifested in the exponential decline in the number of functional variants following the increase in the number of fused modules. As demonstrated with armadillo or ankyrin repeat proteins (Milovnik *et al.*, 2009; Parmeggiani *et al.*, 2008), the type of repeat protein studies here could be applied towards the construction of novel proteins with altered or possibly new binding specificities. Indeed, several of the variants selected here showed binding specificities that differ from wild-type tachylectin-2, including some cross-reactivity with new ligands and increased affinity for mucin (Table I). Future efforts would therefore be directed towards the recruitment of the tandem duplication strategy towards the evolution of lectins with new binding specificities. The longer-term challenge regards the generation of more complex proteins in which the active sites are not individually encoded by single sequence repeats.

Supplementary data

Supplementary data are available at *PEDS* online.

Acknowledgements

D.S.T. is the incumbent of the Nella and Leon Benozziyo Professorship. We are grateful to Moti Saban, Liat Rockah and Korina Goldin for their valuable contributions and to Nathan Sharon for providing sugars and glycoproteins, and for inspiring discussions.

Funding

Financial support by the Sasson and Marjorie Peress Foundation and the Wolgin Prize are gratefully acknowledged.

References

- Abraham, A.-L., Pothier, J. and Rocha, E.P.C. (2009) *J. Mol. Biol.*, **394**, 522–534.
- Andrade, M.A., Perez-Iratxeta, C. and Ponting, C.P. (2001) *J. Struct. Biol.*, **134**, 117–131.
- Arnold, T., Poynor, M., Nussberger, S., Lupas, A.N. and Linke, D. (2007) *J. Mol. Biol.*, **366**, 1174–1184.
- Beisel, H.G., Kawabata, S., Iwanaga, S., Huber, R. and Bode, W. (1999) *EMBO J.*, **18**, 2313–2322.
- Bharat, T.A., Eisenbeis, S., Zeth, K. and Hocker, B. (2008) *Proc. Natl Acad. Sci. USA*, **105**, 9942–9947.
- Binz, H.K., Stumpp, M.T., Forrer, P., Amstutz, P. and Pluckthun, A. (2003) *J. Mol. Biol.*, **332**, 489–503.
- Chaudhuri, I., Soding, J. and Lupas, A.N. (2008) *Proteins*, **71**, 795–803.
- Claren, J., Malisi, C., Hocker, B. and Sterner, R. (2009) *Proc. Natl Acad. Sci. USA*, **106**, 3704–3709.
- Copeland, R.A. (2000) *Enzymes: A Practical Introduction to Structure, Mechanism, and Data Analysis*, 2nd edn. J. Wiley, New York; Chichester England.
- D'Alessio, G. (1999) *Eur. J. Biochem. FEBS*, **266**, 699–708.
- de Bono, S., Riechmann, L., Girard, E., Williams, R.L. and Winter, G. (2005) *Proc. Natl Acad. Sci. USA*, **102**, 1396–1401.
- Edwards, W.R., Busse, K., Allemann, R.K. and Jones, D.D. (2008) *Nucleic Acids Res.*, **36**, e78.
- Fersht, A. (1999) *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. W.H. Freeman, New York.
- Fulop, V. and Jones, D.T. (1999) *Curr. Opin. Struct. Biol.*, **9**, 715–721.
- Graur, D. and Li, W.-H. (2000) *Fundamentals of Molecular Evolution*, 2nd edn. Sinauer Associates, Sunderland, Mass.
- Greenfield, N.J. (2006) *Nat. Protoc.*, **1**, 2527–2535.
- Grishin, N.V. (2001) *J. Struct. Biol.*, **134**, 167–185.
- Huang, J., Koide, A., Makabe, K. and Koide, S. (2008) *Proc. Natl Acad. Sci. USA*, **105**, 6578–6583.
- Jawad, Z. and Paoli, M. (2002) *Structure*, **10**, 447–454.
- Kajander, T., Cortajarena, A.L. and Regan, L. (2006) *Methods Mol. Biol.*, **340**, 151–170.
- Koide, S. (2009) *Curr. Opin. Biotechnol.*, **20**, 398–404.
- Kostlanova, N., Mitchell, E.P., Lortat-Jacob, H., Oscarson, S., Lahmann, M., Gilboa-Garber, N., Chambat, G., Wimmerova, M. and Imberty, A. (2005) *J. Biol. Chem.*, **280**, 27839–27849.
- Main, E.R., Jackson, S.E. and Regan, L. (2003) *Curr. Opin. Struct. Biol.*, **13**, 482–489.
- Main, E.R., Stott, K., Jackson, S.E. and Regan, L. (2005) *Proc. Natl Acad. Sci. USA*, **102**, 5721–5726.
- Milovnik, P., Ferrari, D., Sarkar, C.A. and Pluckthun, A. (2009) *Protein Eng. Des. Sel.*, **22**, 357–366.
- Murzin, A.G. (1992) *Proteins*, **14**, 191–201.
- Neer, E.J. and Smith, T.F. (1996) *Cell*, **84**, 175–178.
- Nikkhah, M., Jawad-Alami, Z., Demydchuk, M., Ribbons, D. and Paoli, M. (2006) *Biomol. Eng.*, **23**, 185–194.
- Okino, N., Kawabata, S., Saito, T., Hirata, M., Takagi, T. and Iwanaga, S. (1995) *J. Biol. Chem.*, **270**, 31008–31015.
- Parmeggiani, F., Pellarin, R., Larsen, A.P., Varadamsetty, G., Stumpp, M.T., Zerbe, O., Caffisch, A. and Pluckthun, A. (2008) *J. Mol. Biol.*, **376**, 1282–1304.
- Patthy, L. (1999) *Gene*, **238**, 103–114.
- Peisajovich, S.G., Rockah, L. and Tawfik, D.S. (2006) *Nat. Genet.*, **38**, 168–174.
- Richter, M., Bosnali, M., Carstensen, L., Seitz, T., Durchschlag, H., Blanquart, S., Merkl, R. and Sterner, R. (2010) *J. Mol. Biol.*, **398**, 763–773.
- Santoro, M.M. and Bolen, D.W. (1988) *Biochemistry*, **27**, 8063–8068.
- Sidhu, S.S., Lowman, H.B., Cunningham, B.C. and Wells, J.A. (2000) *Methods Enzymol.*, **328**, 333–363.
- Stumpp, M.T., Forrer, P., Binz, H.K. and Pluckthun, A. (2003) *J. Mol. Biol.*, **332**, 471–487.
- Vellieux, F.M., Huitema, F., Groendijk, H., Kalk, K.H., Jzn, J.F., Jongejan, J.A., Duine, J.A., Petratos, K., Drenth, J. and Hol, W.G. (1989) *EMBO J.*, **8**, 2171–2178.
- Vogel, C. and Morea, V. (2006) *Bioessays*, **28**, 973–978.
- Wada, K., Wada, Y., Doi, H., Ishibashi, F., Gojobori, T. and Ikemura, T. (1991) *Nucleic Acids Res.*, **19**(suppl.), 1981–1986.
- Weiner, J., 3rd and Bornberg-Bauer, E. (2006) *Mol. Biol. Evol.*, **23**, 734–743.
- Wetzel, S.K., Settanni, G., Kenig, M., Binz, H.K. and Pluckthun, A. (2008) *J. Mol. Biol.*, **376**, 241–257.
- Wright, C.F., Teichmann, S.A., Clarke, J. and Dobson, C.M. (2005) *Nature*, **438**, 878–881.
- Yadid, I. and Tawfik, D.S. (2007) *J. Mol. Biol.*, **365**, 10–17.
- Yadid, I.K., Sharon, N.M., Dym, O. and Tawfik, D.S. (2010) *Proc. Natl Acad. Sci. USA*, **107**, 7287–7292.
- Zhang, Z., Varanasi, U., Carrico, P. and Trumbly, R.J. (2002) *Arch. Biochem. Biophys.*, **406**, 47–54.