Supporting Information for: Programmed protein self-assembly driven by genetically encoded intein-mediated native chemical ligation.

Joseph A. Harvey[1], Laura S. Itzhaki[2] and Ewan R.G. Main[1]*

[1]School of Biological and Chemical Sciences
Queen Mary, University of London,
Mile End Road
London E1 4NS, U.K

[2]Department of Pharmacology
University of Cambridge
Tennis Court Road
Cambridge CB2 1PD, U.K.

*To whom correspondence should be addressed, email: e.main@qmul.ac.uk

S.I. Materials & Methods

Construction of fusion protein genes & vectors:

Anchor: GST – modified TEV cleavage site – module 1 - Strep Tactin.

The module1-anchor is based on the N-terminal GST-tagged pGEX-4T-3 vector (GE Healthcare). The pGEX-4T-3 was altered to remove the existing thrombin cleavage site and replaced with the modified TEV cleavage site - ENLYFQ↓C. The multi cloning site then allows cloning of whichever protein module and C-terminal "anchor" affinity tag is required. Here, a gene encoding CTPR3ΔS with a C-terminal StrepTactin tag (synthesised by Invitrogen Gene Art) was sub-cloned into the modified vector to produce the CTPR3ΔS Spacer-anchor fusion protein.

Linkers: His Tag – modified TEV cleavage site - module2 – Mxe GyrA intein – CBD.

Previous work carried out by the Main group produced a CTPR3ΔS polymerisation construct by cloning the sequence corresponding to Factor Xa cleavage Site – CTPR3ΔS – MxeGyrA intein – chitin binding domain (CBD) into the Invitrogen expression vector pTrc His A [1]. Importantly when designed, the construct contained unique restriction sites after the 6XHis tag and before the Mxe GyrA intein. Thus differing protease cleavage and oligomerisation modules could then be swapped directly into the vector when required. In this manner genes encoding modified TEV cleavage site – and either the "spacer" (CTPR3ΔS) or a "SpacerBinder" CTPR6ΔS (synthesised by Invitrogen Gene Art) and were sub-cloned into the modified vector to produce differing module2 fusion variant proteins.

Protein expression & purification of fusion proteins

All constructs were transformed into BL21 OverExpress C41(DE3) Electrocompetent Cells and grown in 2L flasks of 2xYT media at 37 °C until an OD600 of ~0.8 was reached. They were induced to overexpress by addition of IPTG (final concentration 0.2 mM and 0.5 mM for pGEX-4T-3 and pTrc His TOPO vectors, respectively), transferred to 30 °C and grown for a further 5 hours. Cells were harvested by centrifugation (5,000 RPM, 10 mins, 4 °C), the pellets were resuspended in 50 mM Tris-HCl, 150 mM NaCL, 5 % Glycerol, pH 8.

Cells were homogenised by freeze-thawing in liquid nitrogen followed by sonication on ice (70 % amplitude, 15 sec on, 45 sec off, 10 mins). Cell debris was removed by centrifugation

(18,000 RPM, 30 mins, 4 °C) and supernatant was filtered and purified by GST / Ni^{2+} / Chitin affinity chromatography according to manufacturer's instructions (Pierce and Sigma respectively). If required, the fusion proteins were further purified by Size Exclusion Chromatography on an ÄKTA pure system (GE Heathcare) using a HiLoad™ 26/600 Superdex™ 30 prep grade column. Samples were concentrated to 100 µM, determined spectrophotometrically at 280 nm from the calculated molar extinction coefficients.

Protein storage

All proteins were stored at 4 °C for use up to 24 hours. Where it was necessary to store for more than 24 hours, 1 mL protein aliquots were flash-frozen in liquid nitrogen in an appropriate buffer for downstream applications at a concentration of 100 µM. Flash-frozen protein was stored at -80 °C.

Confirmation of products & reaction yields from Native Chemical Ligations

All purification steps, cleavage and ligation reactions were monitored and confirmed by denaturing SDS-PAGE electrophoresis and either MALDI or Electrospray Mass Spectrometry. Final protein ligation products secondary structure was confirmed by Far UV circular dichroism (C.D.)

(i) Denaturing SDS-PAGE gels: Samples were denatured in loading buffer at 95 °C for 2 minutes (SDS, DTT, glycerol and Coomassie dye) and run on either 15 % or 18 % SDS-PAGE gels. Protein bands were visualised using Coomassie dye and yields calculated using an Odyssey LI-COR in 800 nm imaging channel.

(ii) Mass Spectrometry: Samples were either analysed with a MALDI-TOF or Electrospray Mass Spectrometer. When using MALDI, excess salts and chemicals were removed from protein samples using EMD Millipore Zip-Tip® pipette tips according to the manufacturer's instructions. Protein was eluted from the tip with 10 µl of 0.1 % Trifluoracetic acid (TFA), 60 % Acetonitrle (ACN) in ddH_2O and concentrated by speed-vacuum to 2 µl volume. Protein was then diluted in 0.1 % TFA to adjust the concentration to 5 – 10 µM. 1 µl of sample was mixed 1:1 with saturated sinapinic acid (sinapic acid dissolved in 50% aqueous acetonitrile, with the addition of 0.5% trifluoroacetic acid), loaded onto a Bruker MALDI-TOF/MS steel

plate and allowed to air dry. Mass spectrometry was performed on an autoflex™ MALDI-TOF instrument (Bruker). Protein Standard I and II (Bruker) were used as calibration standards. Using positive ion mode, the electronic gain and laser power were varied until an optimal signal to noise ratio was produced. Peaks were visualised using the Bruker Flex Analysis software.

Electrospray Ionisation Mass Spectrometry (ESI/MS) was used where possible in place of MALDI-TOF/MS to produce a more accurate and higher resolution spectrum. Samples were prepared by denaturation in 6 M Guanidine Hydrochloride (GuHCl), pH 4 followed by twice desalting using a PD Spin G-25 desalting Column (GE Healthcare) into 10 mM Tris-HCl, 1 mM TCEP pH 4. 0.1 % Formic acid was added to samples prior to loading 1 µl of a 2 µM protein sample to a LC-MS, comprising of an 1100 Series LC and SL Ion Trap MSD (Agilent). The sample was loaded via a C18 Reverse Phase HPLC column and eluted over a 0 - 100 % ACN, 0.1 % formic acid gradient at 0.1 ml minute^{-1} flow rate. Raw spectrum data was deconvoluted using the MaxEnt algorithm in Bruker's Analysis software.

(iii) Circular Dichroism (CD): Protein concentrations of 2 – 10 µM in 20 mM Tris-HCl, 1 mM TCEP, pH 8 were analysed in a 5 mm path length cuvette using a Chirascan™ CD Spectrometer (Applied Photophysics Ltd, UK). For each sample a spectrum from 200 – 300 nm was recorded with points taken at 0.5 nm intervals and 0.1 sec per point scanning time. The averaged spectrum of 3 repeats was taken for each sample. Data was converted to molar ellipticity using Equation 1:

Equation 1 : $\theta_{molar} = 100 \times \theta_{obs} \times d \times M$

where θ_{molar} is the molar ellipticity in deg cm^2 mol^{-1}, θ_{obs} is the observed CD signal in milli-degrees, d is the path length in cm and M is the molar concentration of protein.

N-terminal activation by differing proteases

Three different proteases (Thrombin, Factor Xa and TEV) were trialled to investigate their activation efficiency and effect on ligation yield. This was achieved by firstly sub cloning

CTPR3ΔS – Mxe GyrA intein – CBD fusion gene into either the original pGEX-4T-3 vector (thrombin cleavage site) or a modified pGEX-4T-3 vectors that encoded a TEV or a Factor Xa cleavage site. These genes produced three protein fusions: (i) GST – Thrombin - CTPR3ΔS – Mxe GyrA intein – CBD, (ii) GST – TEV - CTPR3ΔS – Mxe GyrA intein – CBD and (iii) GST – Factor Xa - CTPR3ΔS – Mxe GyrA intein – CBD. Each construct was purified by GST affinity chromatography to > 95 %.

N-terminal activation efficiency: 100 µM of each of the inactive monomer constructs, still containing the C-terminal intein fusion, was cleaved with their respective proteases in 50 mM Tris-HCl, 150 mM NaCl, 5 % Glycerol, pH 8. 5mM TCEP was added to the reaction mixture for TEV cleavage and 0.1 mM TCEP for both Thrombin and Factor Xa. 10 U / mg of protease to substrate were added and cleavage was carried out at 25 °C for 24 hours. Cleavage was analysed by SDS-PAGE after 24 hours (S.I Figure 2). Cleavage with both TEV and Thrombin protease went to > 90 % completion and a single N-terminal activated monomer was produced. However, although the Factor Xa mutant was again cleaved > 90 %, the cleavage was not specific to the IEGR amino acid recognition sequence and promiscuous cleavage produced multiple protein species. The N-terminal GST tag contains multiple sites resembling recognition sites accepted by Factor Xa. Therefore, the same experiment was repeated for the His tagged fusion: His - Factor Xa - CTPR3ΔS – Mxe GyrA intein – CBD (named H-F-3-I) [S.I Figure 2D & E]. Two cleavage products were observed corresponding to the correctly cleaved product and a secondary cleaved product within the N-terminal poly-histidine tag (confirmed by ESI mass spectrometry as the sequence QMGR' between amino acids 20-23). This explains the lower reaction yield and the multiple protein bands visible in SDS-PAGE analysis of fibre formation reported by Phillips *et. al* [1].

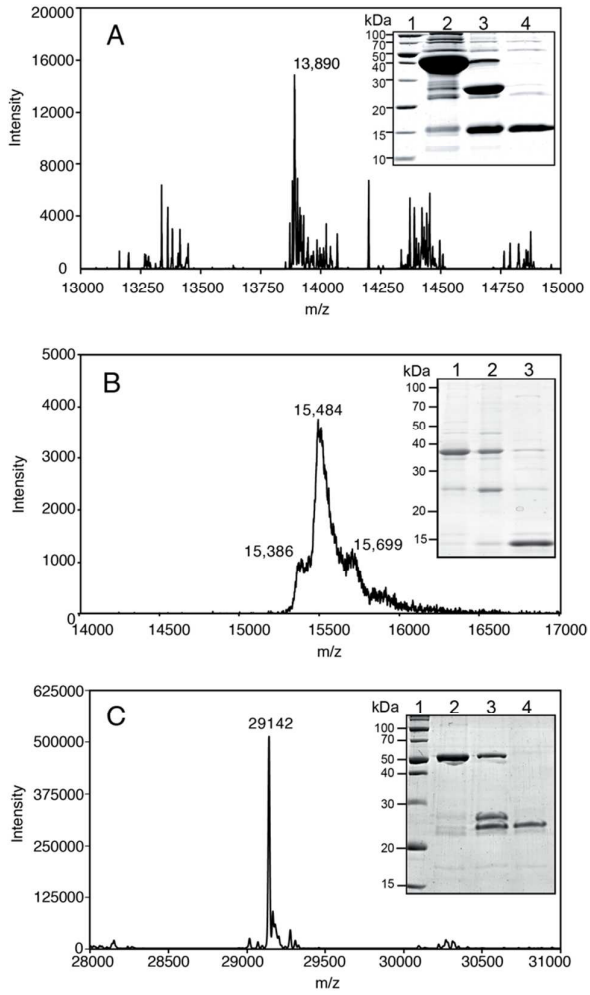
NCL oligomerisation of CTPR3ΔS: Oligomerisation of each cleavage mutant was used to assess the effectiveness of each protease for use within the sequential ligation system. Activated C-terminal thioester containing G-T-3-Thio, G-Thr-3-Thio and H-F-3-Thio were cleaved with their corresponding protease to liberate an N-terminal cysteine and thus elicit oligomerisation. The experiments were performed with 100 µM protein samples in 50 mM Tris-HCl, 150 mM NaCl, 25 mM MESNA, 1 mM TCEP, pH 8 and were incubated with 10 U / mg of protein of their respective protease at 25 ˚C. SDS-PAGE samples were taken after 8,

16 and 24 hours (S.I. Figure 3). In all cases oligomerisation was observed. However, with Thrombin and Factor Xa cleavage the need for reducing environment slows the polymerisation rate substantially. Moreover, Factor Xa's promiscuity lowered ligation yield due to the high levels of nonspecific cleavage. For the Thrombin cleaved protein ligated oligomers were seen to degrade after 24 hours (due to the ligated oligomers forming a new Thrombin cleavage site). In contrast, TEV protease is the most successful and rapid protease to use in our system. It rapidly cleaves to produce N-terminal cysteine as it is not inhibited by the MESNA and TCEP reducing agents needed for successful protein ligation. This leads to a distinct increase in oligomerisation compared to using Factor Xa. Moreover, it is highly specific to its cleavage recognition site. Thus, there is no promiscuous cleavage of CTPR3ΔS monomer or oligomers and thus the termini of oligomers remained active and available for ligation for longer.

References

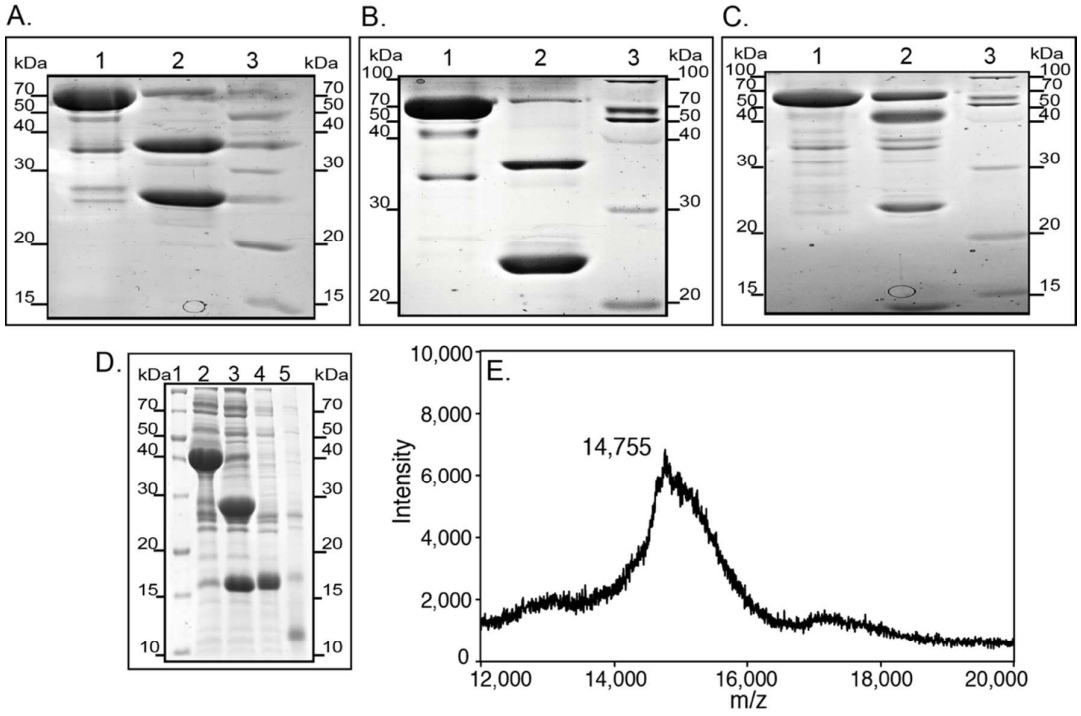
[1] Phillips JJ, Millership C, Main ER. Fibrous nanostructures from the self-assembly of designed repeat protein modules. Angewandte Chemie. 2012;51:13132-5.

S.I. Figures



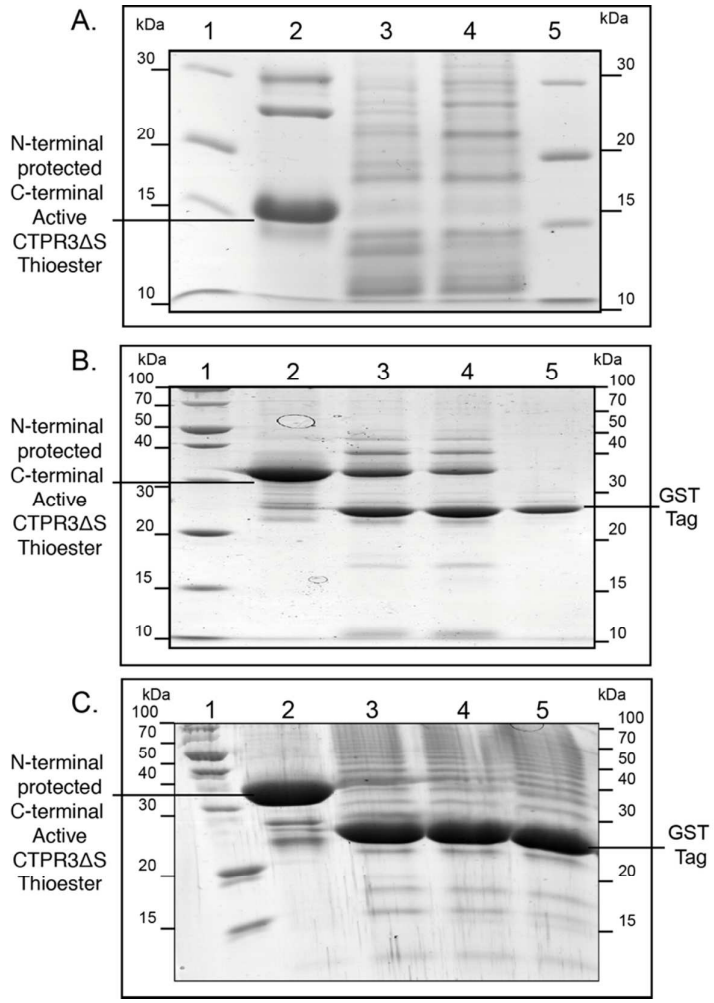
S.I Figure 1: Mass Spectrometry & SDS PAGE analyses (insets) showing the purification & activation of CTPR3ΔS "spacer" anchor fusion, activated CTPR3ΔS "spacer" linker and activated CTPR6ΔS "binder" linker.

(A) ESI Mass Spectrometric analysis of purified TEV protease activated CTPR3ΔS anchor (lane 4 of SDS PAGE analysis inset). Calculated = 13,894 Da, Observed = 13,890 Da. Inset SDS PAGE analysis: lane 1 = molecular weight markers, lane 2 = affinity purified fusion protein, lane 3 = activation via TEV cleavage of fusion protein & lane 4 = purified activated protein.
(B) MALDI-TOF analysis of thioester activated CTPR3ΔS containing linker (lane 3 of SDS PAGE analysis inset). Calculated mass = 15,540 Da, Observed = 15,484 Da. Inset SDS PAGE analysis: lane 1 = fusion protein attached to chitin resin, lane 2 = activation via intein-mediated cleavage of fusion protein from chitin resin & lane 3 = elution from chitin resin of pure activated protein.
(C) MALDI-TOF of thioester activated CTPR6ΔS containing linker (lane 4 of SDS PAGE analysis inset). Calculated size = 29,292 Da, observed = 29,142 Da. Inset SDS PAGE analysis: lane 1 = molecular weight markers, lane 2 = fusion protein attached to chitin resin, lane 3 = activation via intein-mediated cleavage of fusion protein from chitin resin & lane 4 = elution from chitin resin of pure activated protein.
Note, masses of the CTPR proteins seem smaller than expected on the SDS PAGE gels due to "gel shifting" (their high charge causes them to migrate faster than proteins of similar molecular weight).



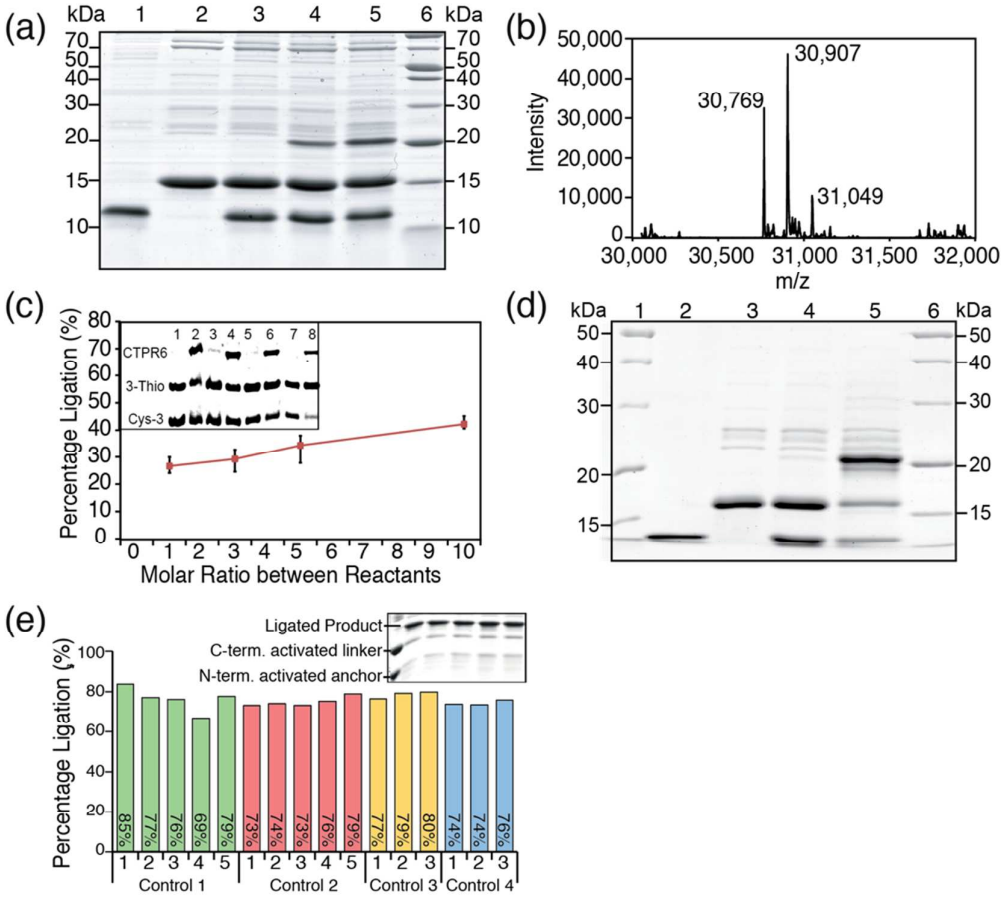
S.I Figure 2: SDS-PAGE analysis of N-terminal activation by differing proteases (24 hours cleavage).

(A) GST-TEV-CTPR3ΔS-INTEIN-CBD fusion construct cleaved by TEV protease. Lane 1: Non-cleaved G-T-3-I (67.7 kDa), Lane 2: TEV cleaved G-T-3-I to produce Cys-3-I (40.4kDa) & Lane 3: protein marker.
(B) GST-THROMBIN-CTPR3ΔS-INTEIN-CBD fusion construct cleaved by Thrombin protease. Lane 1: Non-cleaved G-Thr-3-I (66.7 kDa), Lane 2: Thrombin cleaved G-Thr-3-I to produce Cys-3-I (40.5kDa) & Lane 3: protein marker.
(C) GST-Factor Xa-CTPR3ΔS-INTEIN-CBD fusion construct cleaved by Factor Xa protease. Lane 1: Non-cleaved G-F-3-I (66.7 kDa), Lane 2: Factor Xa cleaved G-F-3-I to produce Cys-3-I (40.5kDa) & Lane 3: protein marker.
(D) HIS-Factor Xa-CTPR3ΔS-INTEIN-CBD fusion construct cleaved by Factor Xa protease. Lane 1: protein marker, Lane 2: H-F-3-I (44.9 kDa), Lane 3: Intein cleavage with DTT to produce H-F-3-DTT (17.2 kDa), Lane 4: H-F-3-DTT Chitin Purified (17.2 kDa) & Lane 5: Factor Xa cleaved H-F-3-DTT (≈ 12.6 kDa for Cys-3-DTT).
(E) MALDI-TOF Mass Spectrometry of the cleavage of the HIS-Factor Xa-CTPR3ΔS fusion construct by Factor Xa. Calculated mass = 14,755 Da corresponds to cleavage at the sequence QMGR' between amino acids 20-23 (upstream of the correct protein cleavage site IEGR').



S.I Figure 3: SDS PAGE analysis of one-pot polymerisation reaction with CTPR3ΔS modules after N-terminal activation with differing proteases.

(A) HIS-Factor Xa-CTPR3ΔS-Thioester activated by Factor Xa cysteine liberation. Lane 1, Protein marker; Lane 2, C-terminal thioester containing HIS-Factor Xa-CTPR3ΔS-Thioester (17 kDa); Lane 3, 24 hours after Factor Xa addition; Lane 4, 48 hours post Factor Xa addition.
(B) GST-Thrombin-CTPR3ΔS-Thioester activated by Thrombin cysteine liberation. Lane 1, Protein marker; Lane 2, C-terminal thioester containing GST-Thrombin-CTPR3ΔS-Thioester (38.8 kDa); Lane 3, 16 hours after thrombin addition; Lane 4, 24 hours after thrombin addition; Lane 5, 48 hours after thrombin addition
(C) GST-TEV-CTPR3ΔS-Thioester activated by TEV cysteine liberation. Lane 1, Protein marker; Lane 2, C-terminal thioester containing GST-TEV-CTPR3ΔS-Thioester (39.5 kDa); Lane 3, 16 hours after TEV addition; Lane 4, 24 hours after TEV addition; Lane 5, 48 hours after TEV addition.



S.I Figure 4: SDS-PAGE analysis, ESI Mass Spectrometry, molar ratio & intra/inter-assay repeatability of ligation experiments.

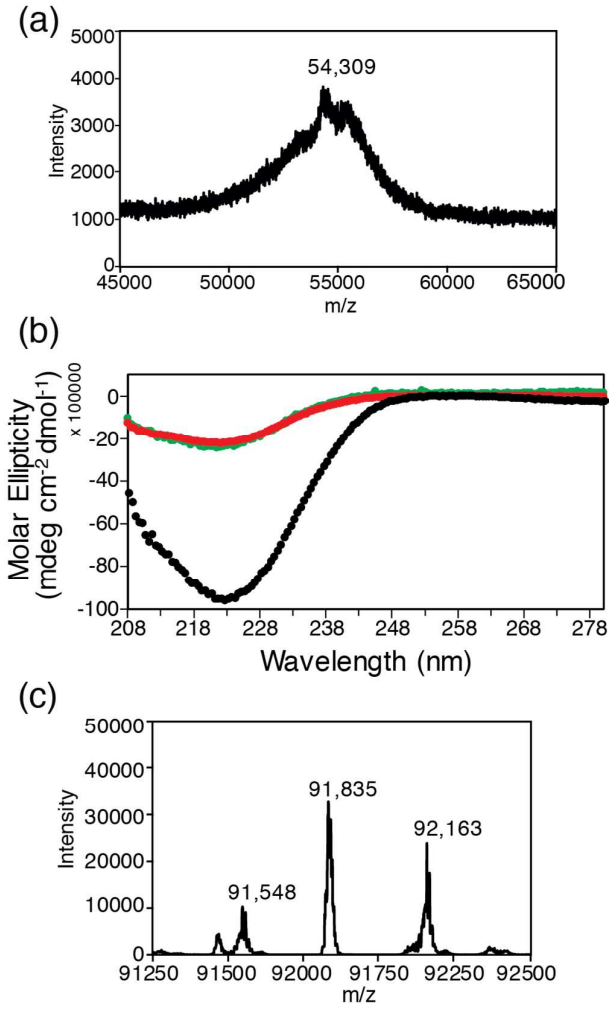
(A) SDS-PAGE analysis of initial NCL reaction between N-terminal cysteine activated spacer CTPR3ΔS anchor with C-terminal activated spacer CTPR3ΔS linker thioester. Lane 1, activated Cys-CTPR3ΔS anchor; Lane 2, activated spacer CTPR3ΔS linker thioester; Lane 3, 0 hour ligation; Lane 4, 24 hour ligation; Lane 5, 48 hour ligation; Lane 6, protein marker. Ligated product can be seen at 20 kDa, due to the high charge of the CTPR protein.

(B) ESI Mass Spectrum of the ligation reaction between N-terminal cysteine activated spacer CTPR3ΔS anchor with C-terminal activated spacer CTPR3ΔS linker thioester after 24 hours incubation. Calculated Mass of ligation product of 30,920 Da, observed mass of 30,907 Da.

(C) Quantification from SDS Page gels of the NCL reaction yield between N-terminal activated CTPR3ΔS anchor (Cys-3) and C-terminal activated CTPR3ΔS linker (3-Thio) over a time course of 48 hours whilst varying molar ratio .

(D) SDS-PAGE analysis of optimised NCL reaction between N-terminal cysteine activated spacer CTPR3ΔS anchor with C-terminal activated spacer CTPR3ΔS linker thioester. Lane 1, protein marker; Lane 2, activated Cys-CTPR3ΔS anchor; Lane 3, activated spacer CTPR3ΔS linker thioester; Lane 4, 0 hours ligation; Lane 5, 24 hours ligation; Lane 6, protein marker.

(E) Comparison of the intra-assay repeatability of 4 optimised ligation experiments. Ligations were performed at 30 °C for 24 hours. Standard deviations for experiments 1-4 were as follows: 5.4, 2.2, 1.4 & 1.1 %. Standard deviation of the entire population was 3.7 %. Mean ligation yield for the entire population was 76.0 %. (Inset) Example SDS-PAGE of assay. Lane 1, 0 hour ligation time point; Lanes 2 – 6, Repeat ligations at 30 °C under optimised conditions.



S.I Figure 5: Mass Spectrometry & Far-UV circular diochroism of nanostructures formed from 3 native chemical ligations with either spacer CTPR3ΔS modules or binder CTPR6ΔS modules.

(A) MALDI-TOF mass spectrum of the ligation product CTPR12ΔS obtained after three NCL rounds using spacer CTPR3ΔS modules. Calculated mass = 54,546 Da, observed mass = 54,309 Da, percentage difference = 0.4 %.
(B) Far UV Circular Dichroism of N-terminal activated spacer CTPR3ΔS anchor, C-terminal activated spacer CTPR3ΔS linker and ligation product CTPR12ΔS obtained after three NCL rounds using spacer CTPR3ΔS modules. The spectra show the average Molar Ellipticity vs Wavelength from 3 repeats. Absorbance was measured from 200 – 280 nm. Measured ellipticity was converted to Molar Ellipticity.
(C) MALDI-TOF mass spectrum of the ligation product CTPR21ΔS obtained after three NCL rounds using spacer CTPR3ΔS anchor and 3 binder CTPR6ΔS linker modules. Calculated mass = 92,406 Da, observed mass = 91,835 Da, percentage difference = 0.6 %