**Supporting information:**

# An unusual cell penetrating peptide identified using a plasmid display-based functional selection platform

Shan Gao[1], Melissa J. Simon[2], Christopher D. Hue[2], Barclay Morrison III[2] and Scott Banta[1]

[1]Department of Chemical Engineering, Columbia University, 500 West 120th Street, New York, NY, USA, 10027

[2]Department of Biomedical Engineering, Columbia University, 500 West 120th Street, New York, NY, USA, 10027

Corresponding Author:

Scott Banta
820 Mudd, MC4721
500 West 120th Street
New York, NY, USA
10027
Phone: (212) 854-7531
Fax: (212) 854-3054
Email: sbanta@columbia.edu

**Materials** Oligonucleotides were obtained from Integrated DNA Technologies (Coralville, IA). Restriction endonucleases and Phusion polymerase for DNA cloning and manipulation were purchased from New England Biolabs (Ipswich, MA). iQ SYBR Green Supermix and Plasmid Midiprep kit was obtained from Bio-Rad Laboratories (Hercules, CA), The gel extraction kit, miniprep kit and Ni-NTA Superflow were purchased from Qiagen (Valencia, CA). SDS-PAGE gels, trypsin/EDTA, CellTracker™ Red CMTPX and 7-AAD were obtained from Invitrogen (Carlsbad, CA). PC12 cells and F12K medium were obtained from American Type Culture Collection (Manassas, VA). M1061 *E. coli* cells were kindly provided by Dr. Patrick Daugherty at University of California, Santa Barbara. GC5 competent *E. coli* cells, BL21 competent *E. coli* cells and all other chemicals used in the study were from Sigma-Aldrich (St. Louis, MO).

**Vector construction** The dsDNA fragment encoding the peptide library was generated by a sense oligonucleotide: 5'-GGG AGC TCG $(NNS)_{14}$ GGC TAA GGT CTC GTA AGA AGC GT-3' and a reverse primer: 5'-ACG CTT CTT ACG AGA CCT TAG CC -3'. The ligation reaction was performed at 16°C overnight using T4 DNA ligase before it was terminated by adding 350µl N3 buffer from the Miniprep kit into the 300µl ligation reaction and then the DNA was purified using Miniprep columns. After washing with the PE buffer, the DNA was eluted by MilliQ $H_2O$. The purified ligation vector was then transformed into M1061 electrocompetent cells through electroporation. Ten micro liters transformation culture from the total volume of 40ml was inoculated onto an LB agar plate containing 50µg/mL kanamycin and 100µg/mL streptomycin to evaluate the diversity of the library. The rest of the culture, ~40ml, was inoculated into 600ml LB supplemented with 0.2% glucose, 50µg/mL kanamycin and 100µg/mL streptomycin. When

$OD_{600}$ reached ~1.5, 10µl of the culture was sampled and inoculated onto plates with serial dilutions. Glycerol stocks were prepared from the rest of the culture. Five colonies from the plate inoculated after electroporation and 15 colonies from the plates inoculated with saturated culture were sequenced to verify the diversity of the library.

Doubled stranded DNA fragments containing a *Kpn*I site and a *Sph*I site were designed to encode for selected peptides (SG1, SG2, SG3) obtained from the selection experiments and they cloned into pRSET-S65T plasmid to create pGFP-SG1, pGFP-SG2 and pGFP-SG3 similarly to the methods reported for the creation of the pGFP-TAT vector (1). The following primers were used to create the new fusion proteins:

CPP-SG3-f 5' - GAT GAA CTA TAC AAA TTT CGG TTG TCG GGC ATG AAC GAG GTG CTG TCG TTC AGG TGG TTG GGC TAA GGT AC - 3';
CPP-SG3-r 5' - CTT AGC CCA ACC ACC TGA ACG ACA GCA CCT CGT TCA TGC CCG ACA ACC GAA ATT TGT ATA GTT CAT CCA TG - 3';
CPP-SG2-f: 5' - GAT GAA CTA TAC AAA TTT TAC AAC AAG CAC GAG GGG ACC ACA GGC GGC AGA ACC GAG ATC GGC TAA GGT AC - 3';
CPP-SG2-r: 5' - CTT AGC CGA TCT CGG TTC TGC CGC CTG TGG TCC CCT CGT GCT TGT TGT AAA ATT TGT ATA GTT CAT CCA TG - 3'.

CPP-SG1-f: 5' - GAT GAA CTA TAC AAA TTT GTG AAA CGG CTG ATG AGG TGG GGG CAG GAG TTG GGG CGG TGC GGC TAA GGT AC - 3';
CPP-SG1-r: 5' - CTT AGC CGC ACC GCC CCA ACT CCT GCC CCC ACC TCA TCA GCC GTT TCA CAA ATT TGT ATA GTT CAT CCA TG - 3';

**Expression and purification of recombinant fluorescent proteins** The vectors pGFP-SG1, pGFP-SG2 and pGFP-SG3 were transformed into BL21 *E. coli* cells for expression of the fluorescent fusion proteins. The fusion proteins were expressed and purified following the protocol developed for the purification of PGT using nickel ion affinity chromatography and size exclusion chromatography as previously described (1). The concentration of purified protein was measured using a Bradford assay kit.

**References:**


1. Gao S, Simon MJ, Morrison B, 3rd, & Banta S (2009) Bifunctional chimeric fusion proteins engineered for DNA delivery: optimization of the protein to DNA ratio. *Biochim Biophys Acta* 1790(3):198-207.
2. Combet C, Blanchet C, Geourjon C, & Deleage G (2000) NPS@: network protein sequence analysis. (Translated from eng) *Trends Biochem Sci* 25(3):147-150 (in eng).

**STable I**. Predicted peptide secondary structures using the Hierarchical Neural Network method (2).

| Peptide | Amino acid sequence<br>Secondary structure prediction* |
|---|---|
| TAT | YGRKKRRQRRR<br>ccchhhchccc |
| Penetratin | RQIKIWFQNRRMKWKK<br>cceeeeeechhchccc |
| SG2 | YNKHEGTTGGRTEI<br>cccccccccccccc |
| SG3 | RLSGMNEVLSFRWL<br>ccchhheeeeecc |
| BH3 | MGQVGRQLAIIGDDINRRY<br>cccccchheeeeccccccccc |
| BH3-TAT | MGQVGRQLAIIGDDINRRYNNGYGRKKRRQRRR<br>cccccchheeeeccchhhhcchcchhhhhhccc |
| BH3-SG2 | MGQVGRQLAIIGDDINRRYYNKHEGTTGGRTEI<br>cccccchheeeeccchhhhhhcccccccccccccc |
| BH3-SG3 | MGQVGRQLAIIGDDINRRYRLSGMNEVLSFRWL<br>cccccchheeeechchchheecccccheeeeeeec |

* h - Alpha helix; c - Random coil; e - Extended beta strand.

**STable II**. Calculated isoelectric point and charges of peptides and proteins at pH=7.4.

| | isoelectric point (pI) | charges@pH=7.4 |
|---|---|---|
| BH3-SG3 | 10.56 | 1.8 |
| BH3-TAT | 11.92 | 8.8 |
| BH3-SG2 | 8.69 | 0.9 |
| SG2 | 7.19 | -0.1 |
| BH3 | 8.88 | 0.8 |
| SG3 | 10 | 0.8 |
| TAT | 12.3 | 7.8 |
| GFP-TAT | 6.65 | -1.6 |
| GFP-SG3 | 6.33 | -8.6 |
| GFP | 5.756 | -9.6 |

**STable III**. Results of peptide sequencing before and after each round of selection with the PD system

Initial library (Unselected)
(60 clones sequenced, 7 failed sequencing results (failure to prime), 53 listed below including 4 with an initial stop codon, sequences with more than 14 randomized amino acids occurred due to frame shifts, 5 frame-shifted sequences contained a portion of the TAT sequence which is due to the fact that the oligonucleotides for library creation were inserted into the doubly digested pTAT vector – the presence of the TAT sequence indicates incomplete removal of the TAT sequence from the PD vector)

*  (initial stop codon found 4 times)
GA*
WG*
LL*
EAT*
TGWL*
FPLTA*
MLSLAA*
RSRVRGE*
FSGSWWSRWS*
IYSICPPCDQDG*
WWFGYACWRATG*
LPNEHERNKRVS*
LEYAKNKHSSDS*
WRLALQRYMSLWL*
WNHRQYMEADMEV*
TKKTQNNKGRKTK*
SVCARDKITERDR*
EQRMMNLLGRNQKKG*
WGLGPAWWGPVVVVG*
WTGRGGLWRGRWESG*
AGWVGTMDGFEGWQG*
WGLVLMPLHFLWSIG*
GMQKRGQRGGCRKKG*
PNPFGRRWSGSTMSG*
PNEMQNQKKNSRKRG*
HTGRFPLSPPSPTTG*
KVKKNKTEEEEAQGG*
PSWGCARGRVMAWMG*
SNGGNKGKGLAEDRG*
GASMPMVMAQVTVRG*
GGHQQGTMDGIPGLG*
MMLTGRRKGLVGVAG*
SALQKRRNRYNGTYG*
RRLGCGAVGLVGSTG*
PKSVRRRGKNNEWEG*

EGQMKINYREPRWEG*
AYAVKGSDSTGLLGG*
TMGGMKDTDKKENKG*
SGLLNRRCWWCRWAKVS*
VLMTAWGAVWVRGWAKVS*
ASYGGATGTGQGCAAKVS*
GAGLPRGWATASAGRAKVS*
DRRRPGGGRGSSGGGAKVS*
NNGYGLVRSVVSVVVAKISCLSE*
KEDRERPREEGVGLRSRKKRRQRRRG*
RRWGGWGWEGVGAGGLRSRKKRRQRRRG*
VERAGEEEDGEGEEGLRSRKKRRQRRRG*
QPRARPAVLGGPLVRLRSRKKRRQRRRG*
ASVGGADARDARVEGLRSRKKRRQRRRG*

Library 1 (After 1 round of selection)
(20 sequenced colonies, 7 failed sequences (failure to prime), 13 listed below including 1 initial stop codon, 3 sequences with more than 14 randomized amino acids occurred due to frame shifts, 2 sequences contained a portion of the TAT sequence as described above)

* (initial stop codon found 1 time)
NW*
ANGTV*
FNNMQKKNRK*
KRNGERKRKGG*
EKDEGHMAGNSKVEG*
VFKLLMFFPRMRVHG*
SILMVLVCVSMVLRG*
CVSVVTRDQLMGVLG*
TNWSGKPEEGEKRHG*
RGRTSGRAGGGRTGAKVS*
KGVQADVVEQQREGGLRSRKKRRQRRRG*
VQGGAVGVGAGGGRRLRSRKKRRQRRRG*

Library 2 (After 2 rounds of selection)
(20 sequenced colonies, 20 listed below including 2 frame-shift sequences, one of which was found 4 times)

ISAG*
PELRWTRAMWVAGGG*
IPGVVREKGGWYRKG*
RWLWMVKGGEAERSG*
SDDPVDVGVESVPSG*
GHGNVSLFPLCNVVG*
RHPTPGHSADLTRWG*
MMCAVRVWQCSVMGG*

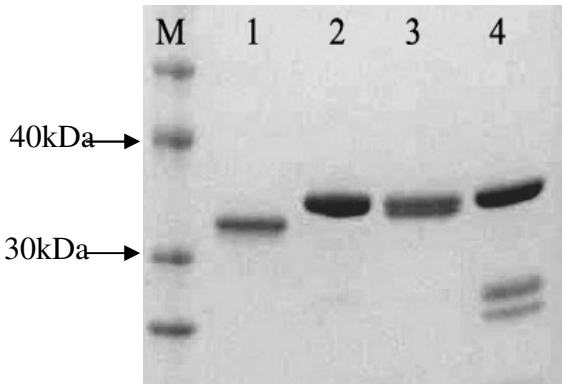
WMGPMQPNGEAAGNG*
SALLLVGFSVTGVRG*
FLLGWGGFQVPRVSG*
ALVWLATTPVVKDPG*
YNKHEGTTGGRTEIG*
KGGHSRRGEVVGETG*
VKRLMRWGQELGRCG*(SG1)
SHGGGASGGWARGARAKVS*
NNGYGLVRSVVSVVVAKISCLSE* (frame-shift sequence was found repeated 4 times)

Library 3 (After 3 rounds of selection)
(20 sequenced colonies, 7 failed sequencing results (failure to prime), 13 listed below)

MLVVFAR*
RAFVHGEGTK*
VLRKLEAWLPWK*
WMDALQKMKGSKLRG*
YERQVQQAQRTDLRG*
IGLGGGTFEEGNPLG*
HCTFFGLAPHMSWLG*
RPWRSLPCLSFSFQG*
SVLGYRRLMRSGWMG*
VKRLMRWGQELGRCG*(SG1)
YNKHEGTTGGRTEIG*(SG2) (this sequence was found repeated 2 times)
RLSGMNEVLSFRWLG*(SG3)

Library 4 (After 4 rounds of selection)
(20 sequenced colonies, 13 failed sequencing results (failure to prime), 7 sequences listed below, 4 had initial stop codons)

* (initial stop codon found 4 times)
SW*
QAWV*
RLSGMNEVLSFRWLG*(SG3)




**SFigure 1** SDS-PAGE of the purified fusion fluorescent proteins (Lane M: Molecular weight markers; Lane 1: GFP; Lane 2: GFP-TAT, Lane 3: GFP-SG3; Lane 4: GFP-SG2.) The contaminating bands in Lane 4 are likely from some GFP-SG2 that has been proteolyzed during the purification process. This was not observed with GFP-TAT or GFP-SG3.