Supplementary Materials

Engineering a thermostability-enhanced active *Clostridium thermocellum* cellobiose phosphorylase by a combination of rational design and directed evolution

Xinhao Ye, Chenming Zhang, Y.-H. Percival Zhang

I. Primers

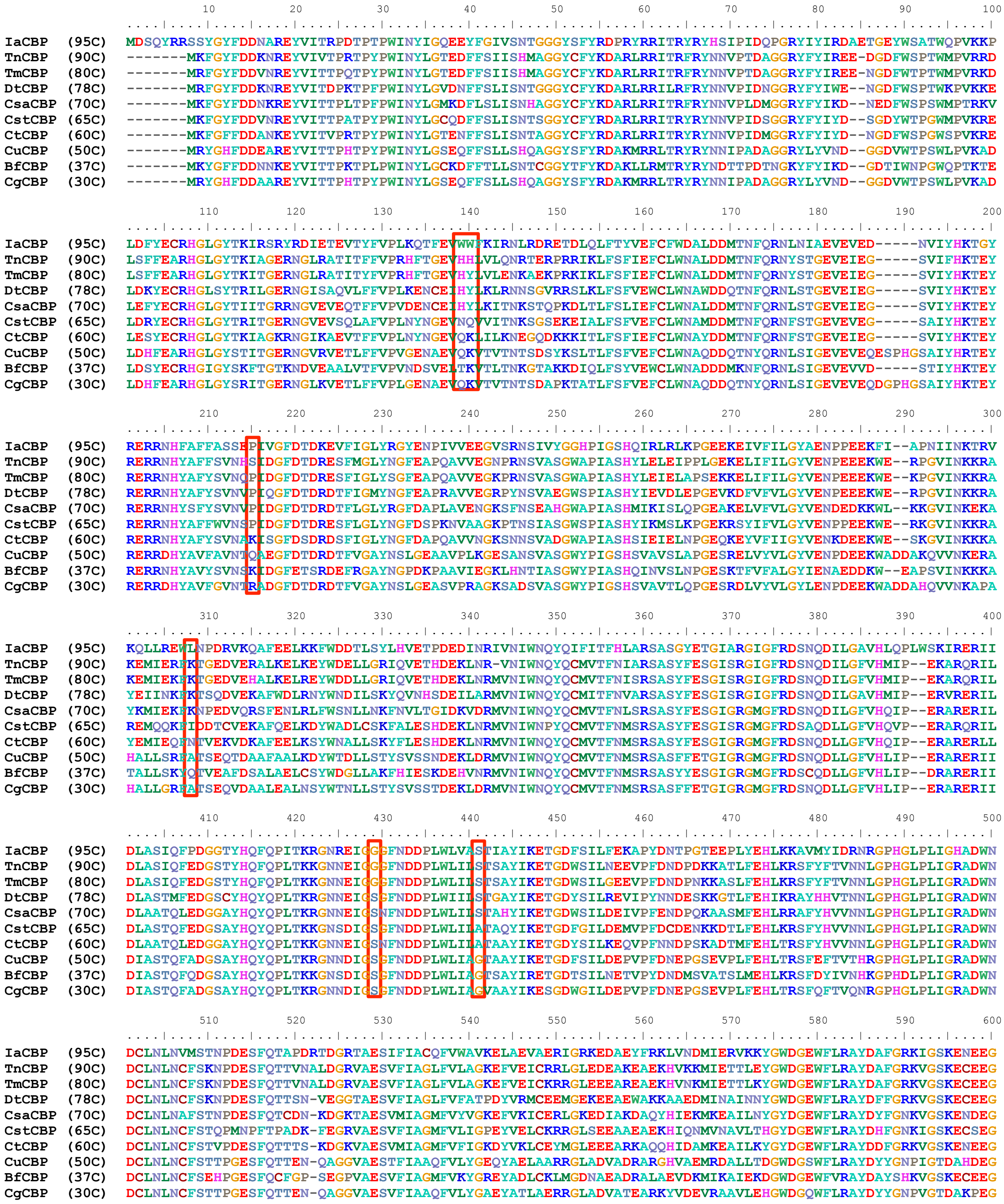
Table 1. Primers used in this study

Name	Sequence
For plasmid construction	
P1	5' GAAAGGCTGCAGCATGAAGTTCGGTTTTTTTG 3'
P2	5' CCTTTTGGATCCTTATCCCATAATTACTTCAAC 3'
P3	5’ AAACTGCTCGAGGGCTCTTCCATGAAGTTCGGTTTTTTTG 3’
For rational design	
P1_130/131	5' GTGAAGTCCA**TT**A**T**CTTATATTGAAGAATGAAG 3'
P2_130/131	5' CATTGTAGTTTAACGGAACGAAGAAAG 3'
P1_201	5' AAATGCA**CCG**ATCAGCGGATTTG 3'
P2_201	5' ACAGAATAGAATGCGTAATGGTTTCTGC 3'
P1_292	5' GATAGAGCAGTTCAA**G**ACTGTTG 3'
P2_292	5' ATTTCATAAGCTTTTTTCTTGTTGATG 3'
P1_411	5' AATGAAATCGGA**G**GCAACTTCAAC 3'
P2_411	5' GTTACCTTTTTTGGTAAGAGGCTGATAC 3'
P1_423	5' GCTGATTCTT**AGC**ACTGCTGCATATATTAAG 3'
P2_423	5' CACAACGGGTCATCGTTGAAGTTG 3'
P1_781	5' GTATCAAAAGGTGTG**AAA**AAAATTACTGTTGAC 3'
P2_781	5' ATGGTTCGGATTCTTCACAGTGATTTC 3'
For directed evolution	
P4	5' CGCCCAATACGCAAACC 3'
P5	5' CGCCATTCGCCATTCAGG 3'
P1_48	5' GGATATTGCTTTTACA**A**GGATGCAAGG 3’
P1_rs_48*	5’ GGATATTGCTTTTACA**G**GGATGCAAGG 3’
P2_48	5’ GCCTGCGGTATTCGAAATGAGTGAG 3’
P1_142	5’ GGACAGGACAAAA**G**GAAAATAACTC 3’
P1_rs_142*	5’ GGACAGGACAAAA**A**GAAAATAACTC 3’
P2_142	5’ TTCATTCTTCAATATAAGCTTTTGGACTTC 3’
P1_189	5’ ACAGAGTACAGAGAGC**T**CAGAAACC 3’
P1_rs_189*	5’ ACAGAGTACAGAGAGC**G**CAGAAACC 3’
P2_189	5' CTTGTGATAGATAACCGAGCCTTCAATC 3'
P1_526	5' CGGAAAAGACTATG**C**GAAGCTTTGC 3'
P1_rs_526*	5' CGGAAAAGACTATG**T**GAAGCTTTGC 3'
P2_526	5' ATGAACACAAACATTCCGGCAATCATAAC 3'

*primers used for reverse mutation

II. Alignment of the CBP homology set

55 unique CBP sequences, identified by PSI-BLAST of the amino acid sequence of *Ct*CBP, were aligned together by the program ClustalW from BioEdit (Carlsbad, CA). Nine CBP sequences, which exhibited the highest identities with the *Ct*CBP (56.4-73.9%), were chosen for homology analysis, as shown below.



```
                      10        20        30        40        50        60        70        80        90       100
              ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
IaCBP  (95C)  MDSQYRRSSYGYFDDNAREYVITRPDTPTPWINYIGQEEYFGIVSNTGGGYSFYRDPRYRRITRYRYHSIPIDQPGRYIYIRDAETGEYWSATWQPVKKP
TnCBP  (90C)  -------MKFGYFDDKNREYVIVTPRTPYPWINYLGTEDFFSIISHMAGGYCFYKDARLRRITRFRYNNVPTDAGGRYFYIREE-DGDFWSPTWMPVRRD
TmCBP  (80C)  -------MRFGYFDDVNREYVITTPQTPYPWINYLGTEDFFSIISHMAGGYCFYKDARLRRITRFRYNNVPTDAGGRYFYIREE-NGDFWTPTWMPVRKD
DtCBP  (78C)  -------MRFGYFDDKNREYVITDPKTPFPWINYLGVDNFFSLISNTGGGYCFYKDARLRRILRFRYNNVPIDNGGRYFYIWE--NGDFWSPTWKPVKKE
CsaCBP (70C)  -------MKFGYFDDNKREYVITTPLTPFPWINYLGMKDFLSLISNHAGGYCFYKDARLRRITRFRYNNVPLDMGGRYFYIKD--NEDFWSPSWMPTRKV
CstCBP (65C)  -------MKFGYFDDVNREYVITTPATPYPWINYLGCQDFFSLISNTSGGYCFYRDARLRRITRYRYNNVPIDSGGRYFYIYD--SGDYWTPGWMPVKRE
CtCBP  (60C)  -------MKFGFFDDANKEYVITVPRTPYPWINYLGTENFFSLISNTAGGYCFYRDARLRRITRYRYNNVPIDMGGRYFYIYD--NGDFWSPGWSPVKRE
CuCBP  (50C)  -------MRYGHFDDEAREYVITTPHTPYPWINYLGSEQFFSLLSHQAGGYSFYRDAKMRRLTRYRYNNIPADAGGRYLYVND--GGDVWTPSWLPVKAD
BfCBP  (37C)  -------MKYGFFDDNNKEYVITTPKTPLPWINYLGCKDFFTLLSNTCGGYTFYKDAKLLRMTRYRYNDTTPDTNGKYFYIKD--GDTIWNPGWQPTKTE
CgCBP  (30C)  -------MRYGHFDDAAREYVITTPHTPYPWINYLGSEQFFSLLSHQAGGYSFYRDAKMRRLTRYRYNNIPADAGGRYLYVND--GGDVWTPSWLPVKAD

                     110       120       130       140       150       160       170       180       190       200
              ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
IaCBP  (95C)  LDFYECRHGLGYTKIRSRYRDIETEVTYFVPLKQTFEVWWFKIRNLRDRETDLQLFTYVEFCFWDALDDMTNFQRNLNIAEVEVED-----NVIYHKTGY
TnCBP  (90C)  LSFFEARHGLGYTKIAGERNGLRATITFFVPRHFTGEVHHLVLQNRTERPRRIKLFSFIEFCLWNALDDMTNFQRNYSTGEVEIEG-----SVIFHKTEY
TmCBP  (80C)  LSFFEARHGLGYTKITGERNGLRATITYFVPRHFTGEVHYLVLENKAEKPRKIKLFSFIEFCLWNALDDMTNFQRNYSTGEVEIEG-----SVIYHKTEY
DtCBP  (78C)  LDKYECRHGLSYTRILGERNGISAQVLFFVPLKENCEIHYLKLRNNSGVRRSLKLFSFVEWCLWNAWDDQTNFQRNLSTGEVEIEG-----SVIYHKTEY
CsaCBP (70C)  LEFYECRHGLGYTIITGRRNGVEVEQTFFVPVDENCEIHYLKITNKSTQPKDLTLFSLIEFCLWNALDDMTNFQRNLSTGEVEIEG-----SVIYHKTEY
CstCBP (65C)  LDRYECRHGLGYTRITGERNGVEVSQLAFVPLNYNGEVNQVVITNKSGSEKEIALFSFVEFCLWNAMDDMTNFQRNFSTGEVEVEG-----SAIYHKTEY
CtCBP  (60C)  LESYECRHGLGYTKIAGKRNGIKAEVTFFVPLNYNGEVQKLILKNEGQDKKKITLFSFIEFCLWNAYDDMTNFQRNFSTGEVEIEG-----SVIYHKTEY
CuCBP  (50C)  LDHFEARHGLGYSTITGERNGVRVETLFFVPVGENAEVQKVTVTNTSDSYKSLTLFSFVEFCLWNAQDDQTNYQRNLSIGEVEVEQESPHGSAIYHRTEY
BfCBP  (37C)  LDSYECRHGIGYSKFTGTKNDVEAALVTFVPVNDSVELTKVTLTNKGTAKKDIQLFSYVEWCLWNADDDMKNFQRNLSIGEVEVVD-----STIYHKTEY
CgCBP  (30C)  LDHFEARHGLGYSRITGERNGLKVETLFFVPLGENAEVQKVTVTNTSDAPKTATLFSFVEFCLWNAQDDQTNYQRNLSIGEVEVEQDGPHGSAIYHKTEY

                     210       220       230       240       250       260       270       280       290       300
              ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
IaCBP  (95C)  RERRNHFAFFASSEPIVGFDTDKEVFIGLYRGYENPIVVEEGVSRNSIVYGGHPIGSHQIRLRLKPGEEKEIVFILGYAENPPEEKFI--APNIINKTRV
TnCBP  (90C)  RERRNHYAFFSVNHSIDGFDTDRESFMGLYNGFEAPQAVVEGNPRNSVASGWAPIASHYLELEIPPLGEKELIFILGYVENPEEEKWE--RPGVINKKRA
TmCBP  (80C)  RERRNHYAFYSVNQPIDGFDTDRESFIGLYSGFEAPQAVVEGKPRNSVASGWAPIASHYLEIELAPSEKKELIFILGYVENPEEEKWE--KPGVINKKRA
DtCBP  (78C)  RERRNHYAFYSVNVPIQGFDTDRDTFIGMYNGFEAPRAVVEGRPYNSVAEGWSPIASHYIEVDLEPGEVKDFVFVLGYVENPEEEKWE--RPGVINKKRA
CsaCBP (70C)  RERRNHYSFYSVNVPIDGFDTDRDTFLGLYRGFDAPLAVENGKSFNSEAHGWAPIASHMIKISLQPGEAKELVFVLGYVENDEDKKWL--KKGVINKEKA
CstCBP (65C)  RERRNHYAFFWVNSPIDGFDTDRESFLGLYNGFDSPKNVAAGKPTNSIASGWSPIASHYIKMSLKPGEKRSYIFVLGYVENPPEEKWE--RKGVINKKRA
CtCBP  (60C)  RERRNHYAFYSVNAKISGFDSDRDSFIGLYNGFDAPQAVVNGKSNNSVADGWAPIASHSIEIELNPGEQKEYVFIIGYVENKDEEKWE--SKGVINKKKA
CuCBP  (50C)  RERRDHYAVFAVNTQAEGFDTDRDTFVGAYNSLGEAAVPLKGESANSVASGWYPIGSHSVAVSLAPGESRELVYVLGYVENPDEEKWADDAKQVVNKERA
BfCBP  (37C)  RERRNHYAVYSVNSKIDGFETSRDEFRGAYNGPDKPAAVIEGKLHNTIASGWYPIASHQINVSLNPGESKTFVFALGYIENAEDDKW--EAPSVINKKKA
CgCBP  (30C)  RERRDHYAVFGVNTRADGFDTDRDTFVGAYNSLGEASVPRAGKSADSVASGWYPIGSHSVAVTLQPGESRDLVYVLGYLENPDEEKWADDAHQVVNKAPA

                     310       320       330       340       350       360       370       380       390       400
              ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
IaCBP  (95C)  KQLLREWLNPDRVKQAFEELKKFWDDTLSYLHVETPDEDINRIVNIWNQYQIFITFHLARSASGYETGIARGIGFRDSNQDILGAVHLQPLWSKIRERII
TnCBP  (90C)  KEMIERFKTGEDVERALKELKEYWDELLGRIQVETHDEKLNR-VNIWNQYQCMVTFNIARSASYFESGISRGIGFRDSNQDILGFVHMIP--EKARQRIL
TmCBP  (80C)  KEMIEKFKTGEDVEHALKELREYWDDLLGRIQVETHDEKLNRMVNIWNQYQCMVTFNISRSASYFESGISRGIGFRDSNQDILGFVHMIP--EKARQRIL
DtCBP  (78C)  YEIINKFKTSQDVEKAFWDLRNYWNDILSKYQVNHSDEILARMVNIWNQYQCMITFNVARSASYFESGISRGIGFRDSNQDILGAVHMIP--ERVRERIL
CsaCBP (70C)  YKMIEKFKNPEDVQRSFENLRLFWSNLLNKFNVLTGIDKVDRMVNIWNQYQCMVTFNLSRSASYFESGIGRGMGFRDSNQDILGFVHQIP--ERARERIL
CstCBP (65C)  REMQQKFIDDTCVEKAFQELKDYWADLCSKFALESHDEKLNRMVNIWNPYQCMVTFNMSRSASYFESGISRGMGFRDSAQDLLGFVHQVP--ERARQRIL
CtCBP  (60C)  YEMIEQFNTVEKVDKAFEELKSYWNALLSKYFLESHDEKLNRMVNIWNQYQCMVTFNMSRSASYFESGIGRGMGFRDSNQDLLGFVHQIP--ERARERLL
CuCBP  (50C)  HALLSRFATSEQTDAAFAALKDYWTDLLSTYSVSSNDEKLDRMVNIWNQYQCMVTFNMSRSASFFETGIGRGMGFRDSNQDLLGFVHLIP--ERARERII
BfCBP  (37C)  TALLSKYQTVEAFDSALAELCSYWDGLLAKFHIESKDEHVNRMVNIWNQYQCMVTFNMSRSASYYESGIGRGMGFRDSCQDLLGFVHLIP--DRARERII
CgCBP  (30C)  HALLGRFATSEQVDAALEALNSYWTNLLSTYSVSSTDEKLDRMVNIWNQYQCMVTFNMSRSASFFETGIGRGMGFRDSNQDLLGFVHLIP--ERARERII

                     410       420       430       440       450       460       470       480       490       500
              ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
IaCBP  (95C)  DLASIQFPDGGTYHQFQPITKRGNREIGGGFNDDPLWLVASTIAYIKETGDFSILFEKAPYDNTPGTEEPLYEHLKKAVMYIDRNRGPHGLPLIGHADWN
TnCBP  (90C)  DLASIQFEDGSTYHQFQPLTKKGNNEIGGGFNDDPLWLILSTSAYIKETGDWSILNEEVPFDNDPDKKATLFEHLKRSFYFTVNNLGPHGLPLIGRADWN
TmCBP  (80C)  DLASIQFEDGSTYHQFQPLTKKGNNEIGGGFNDDPLWLILSTSAYIKETGDWSILGEEVPFDNDPNKKASLFEHLKRSFYFTVNNLGPHGLPLIGRADWN
DtCBP  (78C)  DLASTMFEDGSCYHQYQPLTKRGNNEIGSGFNDDPLWIILSTGAYIKETGDYSILREVIPYNNDESKKGTLFEHIKRAYHHVTNNLGPHGLPLIGRADWN
CsaCBP (70C)  DLAATQLEDGGAYHQYQPLTKRGNNEIGSNFNDDPLWLILSTAHYIKETGDWSILDEIVPFENDPQKAASMFEHLRRAFYHVVNNLGPHGLPLIGRADWN
CstCBP (65C)  DLASTQFEDGSAYHQYQPLTKKGNSDIGSGFNDDPLWLILATAQYIKETGDFGILDEMVPFDCDENKKDTLFEHLKRSFYHVVNNLGPHGLPLIGRADWN
CtCBP  (60C)  DLAATQLEDGGAYHQYQPLTKKGNNEIGSNFNDDPLWLILATAAYIKETGDYSILKEQVPFNNDPSKADTMFEHLTRSFYHVVNNLGPHGLPLIGRADWN
CuCBP  (50C)  DIASTQFADGSAYHQYQPLTKRGNNDIGSGFNDDPLWLIAGTAAYIKETGDFSILDEPVPFDNEPGSEVPLFEHLTRSFEFTVTHRGPHGLPLIGRADWN
BfCBP  (37C)  DIASTQFQDGSAYHQYQPLTKKGNSDIGSGFNDDPLWLIAGTSAYIRETGDTSILNETVPYDNDMSVATSLMEHLKRSFDYIVNHKGPHDLPLIGRADWN
CgCBP  (30C)  DIASTQFADGSAYHQYQPLTKRGNNDIGSGFNDDPLWLIAGVAAYIKESGDWGILDEPVPFDNEPGSEVPLFEHLTRSFQFTVQNRGPHGLPLIGRADWN

                     510       520       530       540       550       560       570       580       590       600
              ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
IaCBP  (95C)  DCLNLNVMSTNPDESFQTAPDRTDGRTAESIFIACQFVWAVKELAEVAERIGRKEDAEYFRKLVNDMIERVKKYGWDGEWFLRAYDAFGRKIGSKENEEG
TnCBP  (90C)  DCLNLNCFSKNPDESFQTTVNALDGRVAESVFIAGLFVLAGKEFVEICRRLGLEDEAKEAEKHVKKMIETTLEYGWDGEWFLRAYDAFGRKVGSKECEEG
TmCBP  (80C)  DCLNLNCFSKNPDESFQTTVNALDGRVAESVFIAGLFVLAGKEFVEICKRRGLEEEAREAEKHVNKMIETTLKYGWDGEWFLRAYDAFGRKVGSKECEEG
DtCBP  (78C)  DCLNLNCFSKNPDESFQTTSN-VEGGTAESVFIAGLFVFATPDYVRMCEEMGEKEEAEWAKKAAEDMINAINNYGWDGEWFLRAYDFFGRKVGSKECEEG
CsaCBP (70C)  DCLNLNAFSTNPDESFQTCDN-KDGKTAESVMIAGMFVYVGKEFVKICERLGKEDIAKDAQYHIEKMKEAILNYGYDGEWFLRAYDYFGNKVGSKENDEG
CstCBP (65C)  DCLNLNCFSTQPMNPFTPADK-FEGRVAESVFIAGMFVLIGPEYVELCKRRGLSEEAAEAEKHIQNMVNAVLTHGYDGEWFLRAYDHFGNKIGSKECSEG
CtCBP  (60C)  DCLNLNCFSTVPDESFQTTTS-KDGKVAESVMIAGMFVFIGKDYVKLCEYMGLEEEARKAQQHIDAMKEAILKYGYDGEWFLRAYDDFGRKVGSKENEEG
CuCBP  (50C)  DCLNLNCFSTTPGESFQTTEN-QAGGVAESTFIAAQFVLYGEQYAELAARRGLADVADRARGHVAEMRDALLTDGWDGSWFLRAYDYYGNPIGTDAHDEG
BfCBP  (37C)  DCLNLNCFSEHPGESFQCFGP-SEGPVAESVFIAGMFVKYGREYADLCKLMGDNAEADRALAEVDKMIKAIEKDGWDGEWFVRAYDAYSHKVGSKECEEG
CgCBP  (30C)  DCLNLNCFSTTPGESFQTTEN-QAGGVAESVFIAAQFVLYGAEYATLAERRGLADVATEARKYVDEVRAAVLEHGWDGQWFLRAYDYYGNPVGTDAKPEG
```

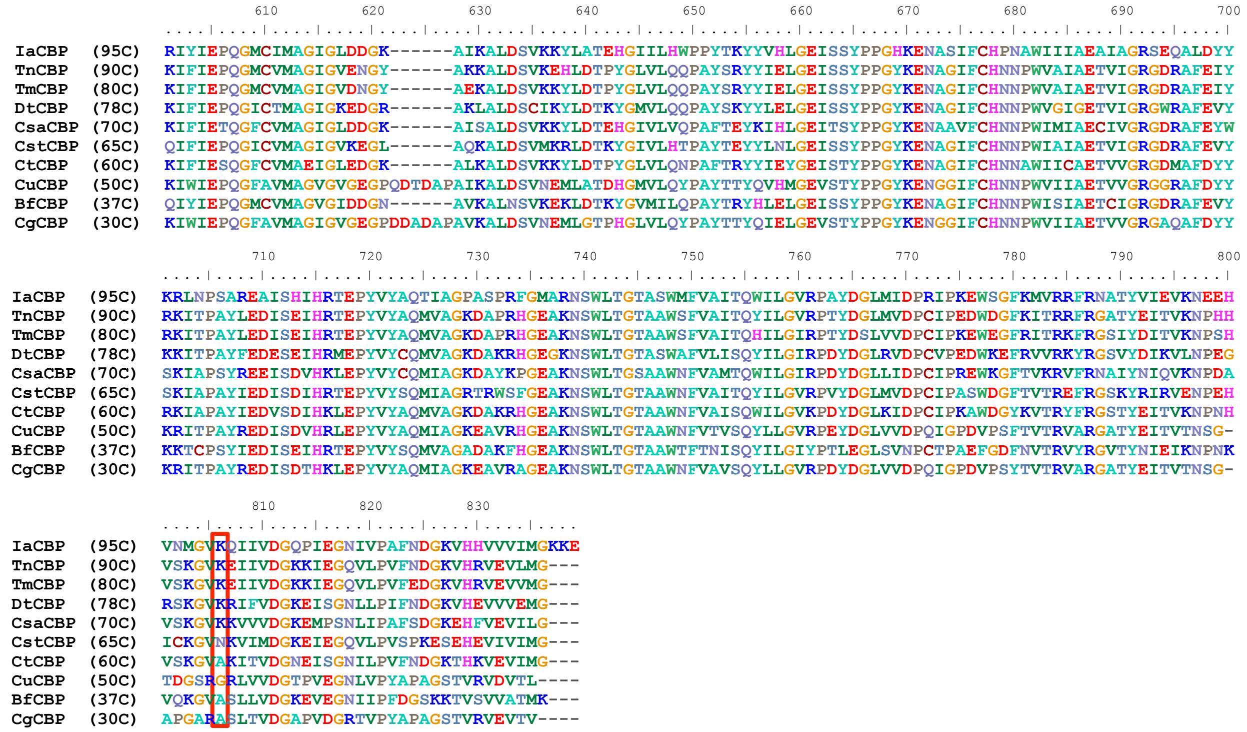

610 620 630 640 650 660 670 680 690 700
IaCBP (95C) RIYIEPQGMCIMAGIGLDDGK------AIKALDSVKKYLATEHGIILHWPPYTKYYVHLGEISSYPPGHKENASIFCHPNAWIIIAEAIAGRSEQALDYY
TnCBP (90C) KIFIEPQGMCVMAGIGVENGY------AKKALDSVKEHLDTPYGLVLQQPAYSRYYIELGEISSYPPGYKENAGIFCHNNPWVAIAETVIGRGDRAFEIY
TmCBP (80C) KIFIEPQGMCVMAGIGVDNGY------AEKALDSVKKYLDTPYGLVLQQPAYSRYYIELGEISSYPPGYKENAGIFCHNNPWVAIAETVIGRGDRAFEIY
DtCBP (78C) KIFIEPQGICTMAGIGKEDGR------AKLALDSCIKYLDTKYGMVLQQPAYSKYYLELGEISSYPPGYKENAGIFCHNNPWVGIGETVIGRGWRAFEVY
CsaCBP (70C) KIFIETQGFCVMAGIGLDDGK------AISALDSVKKYLDTEHGIVLVQPAFTEYKIHLGEITSYPPGYKENAAVFCHNNPWIMIAECIVGRGDRAFEYW
CstCBP (65C) QIFIEPQGICVMAGIGVKEGL------AQKALDSVMKRLDTKYGIVLHTPAYTEYYLNLGEISSYPPGYKENAGIFCHNNPWIAIAETVIGRGDRAFEVY
CtCBP (60C) KIFIESQGFCVMAEIGLEDGK------ALKALDSVKKYLDTPYGLVLQNPAFTRYYIEYGEISTYPPGYKENAGIFCHNNAWIICAETVVGRGDMAFDYY
CuCBP (50C) KIWIEPQGFAVMAGVGVGEGPQDTDAPAIKALDSVNEMLATDHGMVLQYPAYTTYQVHMGEVSTYPPGYKENGGIFCHNNPWVIIAETVVGRGGRAFDYY
BfCBP (37C) QIYIEPQGMCVMAGVGIDDGN------AVKALNSVKEKLDTKYGVMILQPAYTRYHLELGEISSYPPGYKENAGIFCHNNPWISIAETCIGRGDRAFEVY
CgCBP (30C) KIWIEPQGFAVMAGIGVGEGPDDADAPAVKALDSVNEMLGTPHGLVLQYPAYTTYQIELGEVSTYPPGYKENGGIFCHNNPWVIIAETVVGRGAQAFDYY
710 720 730 740 750 760 770 780 790 800
IaCBP (95C) KRLNPSAREAISHIHRTEPYVYAQTIAGPASPRFGMARNSWLTGTASWMFVAITQWILGVRPAYDGLMIDPRIPKEWSGFKMVRRFRNATYVIEVKNEEH
TnCBP (90C) RKITPAYLEDISEIHRTEPYVYAQMVAGKDAPRHGEAKNSWLTGTAAWSFVAITQYILGVRPTYDGLMVDPCIPEDWDGFKITRRFRGATYEITVKNPHH
TmCBP (80C) RKITPAYLEDISEIHRTEPYVYAQMVAGKDAPRHGEAKNSWLTGTAAWSFVAITQHILGIRPTYDSLVVDPCIPKEWEGFRITRKFRGSIYDITVKNPSH
DtCBP (78C) KKITPAYFEDESEIHRMEPYVYCQMVAGKDAKRHGEGKNSWLTGTASWAFVLISQYILGIRPDYDGLRVDPCVPEDWKEFRVVRKYRGSVYDIKVLNPEG
CsaCBP (70C) SKIAPSYREEISDVHKLEPYVYCQMIAGKDAYKPGEAKNSWLTGSAAWNFVAMTQWILGIRPDYDGLLIDPCIPREWKGFTVKRVFRNAIYNIQVKNPDA
CstCBP (65C) SKIAPAYIEDISDIHRTEPYVYSQMIAGRTRWSFGEAKNSWLTGTAAWNFVAITQYILGVRPVYDGLMVDPCIPASWDGFTVTREFRGSKYRIRVENPEH
CtCBP (60C) RKIAPAYIEDVSDIHKLEPYVYAQMVAGKDAKRHGEAKNSWLTGTAAWNFVAISQWILGVKPDYDGLKIDPCIPKAWDGYKVTRYFRGSTYEITVKNPNH
CuCBP (50C) KRITPAYREDISDVHRLEPYVYAQMIAGKEAVRHGEAKNSWLTGTAAWNFVTVSQYLLGVRPEYDGLVVDPQIGPDVPSFTVTRVARGATYEITVTNSG-
BfCBP (37C) KKTCPSYIEDISEIHRTEPYVYSQMVAGADAKFHGEAKNSWLTGTAAWTFTNISQYILGIYPTLEGLSVNPCTPAEFGDFNVTRVYRGVTYNIEIKNPNK
CgCBP (30C) KRITPAYREDISDTHKLEPYVYAQMIAGKEAVRAGEAKNSWLTGTAAWNFVAVSQYLLGVRPDYDGLVVDPQIGPDVPSYTVTRVARGATYEITVTNSG-
810 820 830
IaCBP (95C) VNMGVKQIIVDGQPIEGNIVPAFNDGKVHHVVVIMGKKE
TnCBP (90C) VSKGVKEIIVDGKKIEGQVLPVFNDGKVHRVEVLMG---
TmCBP (80C) VSKGVKEIIVDGKKIEGQVLPVFEDGKVHRVEVVMG---
DtCBP (78C) RSKGVKRIFVDGKEISGNLLPIFNDGKVHEVVVEMG---
CsaCBP (70C) VSKGVKKVVVDGKEMPSNLIPAFSDGKEHFVEVILG---
CstCBP (65C) ICKGVNKVIMDGKEIEGQVLPVSPKESEHEVIVIMG---
CtCBP (60C) VSKGVAKITVDGNEISGNILPVFNDGKTHKVEVIMG---
CuCBP (50C) TDGSRGRLVVDGTPVEGNLVPYAPAGSTVRVDVTL----
BfCBP (37C) VQKGVASLLVDGKEVEGNIIPFDGSKKTVSVVATMK---
CgCBP (30C) APGARASLTVDGAPVDGRTVPYAPAGSTVRVEVTV----

Fig. 1. Alignment of the CBP homology set, where IaCBP is from *Ignisphaera aggregans* (optimal growth temperature: 95 ^{o}C), TnCBP is from *Thermotoga neapolitana* (optimal growth temperature: 90 ^{o}C), TmCBP is from *Thermotoga maritime* (optimal growth temperature: 80 ^{o}C), DtCBP is from *Dictyoglomus thermophilum* (optimal growth temperature: 78 ^{o}C), CsaCBP is from *Caldicellulosiruptor saccharolyticus* (optimal growth temperature: 70 ^{o}C), CstCBP is from *Clostridium sterocorarium* (optimal growth temperature: 65 ^{o}C), CuCBP is from *Cellulomonas uda* (optimal growth temperature: 50 ^{o}C), BfCBP is from *Butyrivibrio fibrisolvens* (optimal growth temperature: 37 ^{o}C), and CgCBP is from *Cellvibrio gilvus* (optimal growth temperature: 30 ^{o}C). The candidates for substitution were selected based on the structural knowledge and highlighted with red frames. Ruler was designed in correspondence to the residue number of the first CBP sequence, IaCBP.

III. Stabilization centers

Stabilization center refers the residues that involved in cooperative long-range contacts, and are expected to stabilize proteins structures by preventing their decay with their cooperative long range interactions (Dosztányi et al. 1997). The stabilization centers of *Ct*CBP were determined by SCider (http://www.enzim.hu/scide), in response to its homology modeling (Dosztányi et al. 2003).

Table 2. Stabilization centers of *Ct*CBP

Resid. No.	Resid. Name	List of stabilization Centers
3	PHE	16
4	GLY	14,15
12	GLU	99
13	TYR	98
14	VAL	4, 97, 98
15	ILE	4, 96
16	THR	3
22	TYR	691
41	ALA	69, 70
42	GLY	68, 69, 70
43	GLY	68, 69
45	CYS	55
49	ASP	400
52	LEU	156
54	ARG	155
55	ILE	45, 153
56	THR	152, 153
57	ARG	152
68	GLY	42, 43
69	GLY	41, 42, 43
70	ARG	41, 42
73	TYR	146
74	ILE	145,146
82	SER	92
92	LEU	82
93	GLU	109
96	GLU	15, 106
97	CYS	14
98	ARG	13, 14
99	HIS	12
104	THR	118
105	LYS	118
106	ILE	96
109	LYS	93
113	ILE	135, 136
114	LYS	134
118	THR	104, 105
119	PHE	129
128	GLU	261
129	VAL	119, 260, 261
130[a]	**GLN**	**259, 260**
131[a]	**LYS**	**259**
134	LEU	114
135	LYS	113, 253, 254
136	ASN	113
145	THR	74
146	LEU	73, 74, 245
147	PHE	244
150	ILE	241, 242
152	PHE	56, 57
153	CYS	55, 56
155	TRP	54
156	ASN	52
234	VAL	207, 208
239	ALA	208
241	ILE	150
242	ALA	150
244	HIS	147, 204
245	SER	146, 202, 203, 204
253	GLY	135
254	GLU	135
259	VAL	130, 131, 198
260	PHE	129, 130, 197
261	ILE	128, 129, 196
264	TYR	193, 194
265	VAL	193
315	TYR	329, 758
316	PHE	757
317	LEU	756
329	ASN	315
332	ASN	725
335	GLN	718
337	MET	372
338	VAL	717
343	SER	361
356	GLY	398
357	MET	398
361	ASP	343
363	ASN	381
367	LEU	723
372	GLN	337
381	LEU	363
388	GLN	443, 444
396	HIS	406
398	TYR	356, 357
400	PRO	49
406	ASN	396
420	LEU	460
421	ILE	457
439	GLU	452
440	GLN	451, 452
441	VAL	451, 453
443	PHE	388
444	ASN	388
451	ASP	440, 441
452	THR	439, 440
453	MET	441
457	LEU	421
460	SER	420
472	GLY	551
474	PRO	510, 511, 514
475	LEU	510, 514
476	ILE	510, 511, 514
486	ASN	634, 635
510	GLU	474, 475, 476
511	SER	474, 476
622	LEU	648
623	GLN	647
624	ASN	582, 583
625	PRO	581
635	GLY	486
643	GLY	705
647	ASN	623
648	ALA	622
649	GLY	700, 701
650	ILE	700, 701
651	PHE	700
659	ILE	724
661	ALA	592
691	LYS	22
696	VAL	715, 716
697	TYR	715, 716, 718
698	ALA	714, 715, 716, 718, 720
699	GLN	713, 714
700	MET	649,650, 651
701	VAL	649, 650
705	ASP	643
713	LYS	699
714	ASN	698, 699
715	SER	696, 697, 698
716	TRP	696, 697, 698
717	LEU	338
718	THR	335, 697, 698
720	THR	698
723	TRP	367
724	ASN	659
725	PHE	332
733	LEU	748
741	GLY	795
742	LEU	794
743	LYS	794
748	ILE	733, 776
750	LYS	775, 776
752	TRP	772, 775
753	ASP	772
756	LYS	317
757	VAL	316, 767
758	THR	315
763	GLY	800
765	THR	802, 803
766	TYR	804
767	GLU	757
769	THR	807
770	VAL	808
771	LYS	809
772	ASN	752, 753
775	HIS	750, 752
776	VAL	748, 750, 810
779	GLY	810

179	VAL	196	514	ILE	474, 475, 476	780	VAL	810
193	TYR	264, 265	518	PHE	548	**781**[b]	**ALA**	**809, 810**
194	ALA	264	548	MET	518	782	LYS	808
196	TYR	179, 261	551	ALA	472	783	ILE	807
197	SER	260	560	GLU	605, 606	794	ILE	742, 743
198	VAL	259	561	TRP	574, 575	795	LEU	741
202	ILE	245	563	LEU	574	800	ASP	763
203	SER	245	574	GLY	563, 563	802	LYS	765
205	PHE	231, 232	575	SER	561	803	THR	765
206	ASP	231, 232, 233	581	GLY	625	804	HIS	766
207	SER	233, 234	582	LYS	624	807	GLU	769, 783
208	ASP	235, 234, 239	583	ILE	624	808	VAL	770, 782
231	ASN	205, 206	592	VAL	661	809	ILE	771, 781
232	ASN	205, 206	605	LYS	560	810	MET	776, 779, 780, 781
233	SER	206, 207, 208	606	ALA	560			

[a] The resides Q130 and K131 were located in the stabilization centers in pair, both interacting with V259 and/or Y258. Mutation Q130H and K131Y generated strong π-π interactions and expanded the stabilization centers of both resid 130 and 131 to F119 and F260. Therefore, they were chosen for site-directed mutagenesis, resulted in a 1.7-fold increase of the inactivation halftime at 70 °C.

[b] The residues A781 was located in the stabilization centers. The substitution of A781 with Lysine didn't change its own stabilization center partners, whereas increased the stabilizing partners of T765 with T803, Y766 with P796, and V767 with H804. Therefore, the mutation A781K expanded the stabilization centers and, practically, increased the inactivation halftime to 15.3 min.

IV. Protein contact map

A protein contact map represents the distance between all possible residue pairs of a 3-D protein structure. The contact map of *Ct*CBP was established by Contact Map plugin of VMD 1.8.6 (Humphrey et al. 1996).

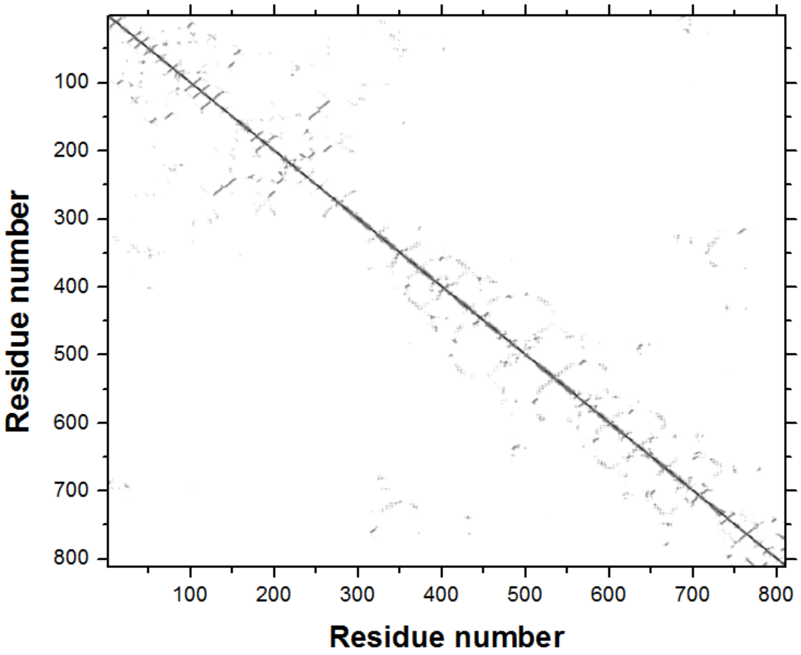


Fig.2 Contact map of *Ct*CBP. A graph square is colored black at 0.0 Angstrom distance, to a linear gray scale between 0.0 and 10.0 Angstroms, and white when equal to or greater than 10.0 Angstroms.

V. Optimal mutation rate for the random mutagenesis of CtCBP

Given a choice of protein scaffold, libraries of fixed sizes, and no reliable knowledge for rational engineering, a simple assessment of library optimality is the number of valuable clones they contain (Daugherty et al. 2000). Therefore, it is hard to tell if we have reasonable mutation rates before qualitative and quantitative evaluation of both low- and high- frequency mutagenesis libraries.

Although the optimal conditions for making mutant libraries are unclear, it is well believed an optimal mutation rate exists that balances diversity and retention of the function (Miura and Sonigo 2001; Zaccolo and Gherardi 1999). In general, low-error-rate mutagenesis results in a high probability of functional sequences with a low probability of beneficial mutations, while high-error-rate mutagenesis brings out a high probability of lethal mutations with a high probability of unique sequences. In the past decades, there have been many attempts to determine the optimum mutation rates for directed evolution theoretically and experimentally (Clune et al. 2008; Kimura 1960; Pritchard et al. 2005; Shafikhani et al. 1997; Zaccolo et al. 1996). Daugherty *et al.* quantitatively analyzed the effect of mutation frequency on the affinity maturation of single chain Fv antibodies (2000). At the low to moderate mutation frequencies with an average mutation rate of $m \leq 8$, the functional fraction of clones decreased exponentially, but the most highly mutated library ($m = 22.5$) had significantly more active clones than expected relative to this trend. The results indicated a preferred mutation frequency for functional improvement of scFv may persist between $m = 4$ and $m = 2$, under which the library includes a larger fraction of active mutants and is more likely to yield improved mutants. Miura and Sonigo proposed a simple model to determine the optimal mutation rate for random mutagenesis (2001). They linked the optimal rate with the number of simultaneous mutations required for possible beneficial and lethal changes. As a result, the model predicted the optimum is a mutation rate that induces at least 63% ($1 - e^{-n}$, n represents the required number of positive mutations and $n \geq 1$) of the cloned gene in the library to be non-functional.

Recently, a more inclusive model was presented by Drummond *et al.* (2005a). In their work, experimental data firstly proved that the mutations do not follow the Poisson distribution under the high-error-rate random mutagenesis, but rather a previously proposed distribution derived from a PCR model (Sun 1995). The PCR-distributed mutations were then modeled to investigate the effect of mutation rates on functionally improved fractions in the random mutagenesis libraries. It was found that the optimal mutation rates depend on the number of tranformants sampled, the PCR protocol used, the wild-type protein being mutated, and other parameters. Considering it's the most realistic model of error-prone PCR by far, we followed Sun's and Drummond's model to estimate the possible optimum in our work as below.

Assume the *cbp* gene would be amplified by error-prone PCR with a constant efficiency λ. Here λ is set as 0.6 (Drummond et al. 2005a). The thermal cycles are n, resulting in $d = n\lambda$ DNA doublings. Based on Sun's model (Sun 1995), the average of nucleotide mutations per sequence, say $E(m_{nt})$, is defined as

$$E(m_{nt}) = \frac{n\lambda\mu G}{1+\lambda} \qquad (1)$$

where μ and G represent mutation rate (mutations per base per PCR cycle) and gene length, respectively. Set

$$x = \mu G \tag{2}$$

which actually presents another type of mutation rate (mutations per PCR cycle).

Then the mutational distribution can be computed as follows,

$$\begin{aligned} P(M = m_{nt}) &= \sum_{k=0}^{n} P(M = m | K = k) P(K = k) \\ &= \sum_{k=0}^{n} \frac{(kx)^{m_{nt}} e^{-kx}}{m_{nt}!} \frac{\begin{pmatrix} n \\ k \end{pmatrix} \lambda^k}{(1+\lambda)^n} = \frac{x^{m_{nt}}}{(1+\lambda)^n m_{nt}!} \sum_{k=n}^{n} \begin{pmatrix} n \\ k \end{pmatrix} \lambda^k k^{m_{nt}} e^{-kx} \end{aligned} \tag{3}$$

The probability a nucleotide mutation produces a non-synonymous change is assumed to be binomial, with parameter $p_{ns.}$ Generally, p_{ns} is around 0.8 (Daugherty et al. 2000). Note that non-synonymous variations include the changes of amino acid encoded as well as the others that truncated or inactivated the protein (e.g. insertions, deletions, and mutations to stop codons). The latter constitutes a fraction of mutation with the probability p_{tr}. The p_{tr} is around 0.05-0.07 (Drummond et al. 2005b) and assumed to be 0.06 here. Then the probability a sequence with m_{nt} nucleotide mutations retains function can be defined as

$$\begin{aligned} P(f \mid m_{nt}) &= \sum_{m_{ns}=0}^{m_{nt}} \begin{pmatrix} m_{nt} \\ m_{ns} \end{pmatrix} P(non\, trunc. \mid m_{ns}) \times P(f \mid m_{ns}\ a\mathrm{min}o\ acid\ changes) \\ &= (1 - (1 - v(1 - p_{tr}/p_{ns}))p_{ns})^{m_{nt}} \end{aligned} \tag{4}$$

where m_{ns} denotes non-synonymous mutation, and ν represents the average fraction of functional one-mutant neighors on the protein-sequence-space network. By combining eq.(3) and (4), and assuming gene length $L \to \infty$ since $\langle m_n \rangle << L$, we find the probability a sequence from the library will retain function is:

$$P(f) = \sum_{m_{nt}=0}^{\infty} P(f \mid m_{nt}) P(m_{nt}) = \left(\frac{1 + \lambda \left(- \frac{\langle m_{nt} \rangle (1+\lambda)}{n\lambda} (1 - v(1 - p_{tr}/p_{ns}))p_{ns} \right)}{1+\lambda} \right)^n \tag{5}$$

Now ν is required to predict the functional fractions. In this work, since the surviving colonies in selection plates represent the functional mutants, the probability *P(f)* would approximate the selection power of each library. Using the $<m_{nt}>$ vs. selection power data, we then estimate the random mutagenesis of *Ctcbp* with ν ≈ 0.91 (see SOM-Fig. 3). Intriguingly, the ν value is much bigger than previous reports (ν ≈ 0.2 for the antibody binding task, and ν ≈ 0.65 for the subtilisin data), thanks to the larger sequence of *Ctcbp*. It suggests a low probability to inactivate *Ct*CBP by one-mutant-per-step spacewalk, which is helpful to produce functional mutants, but limits the unique fractions.

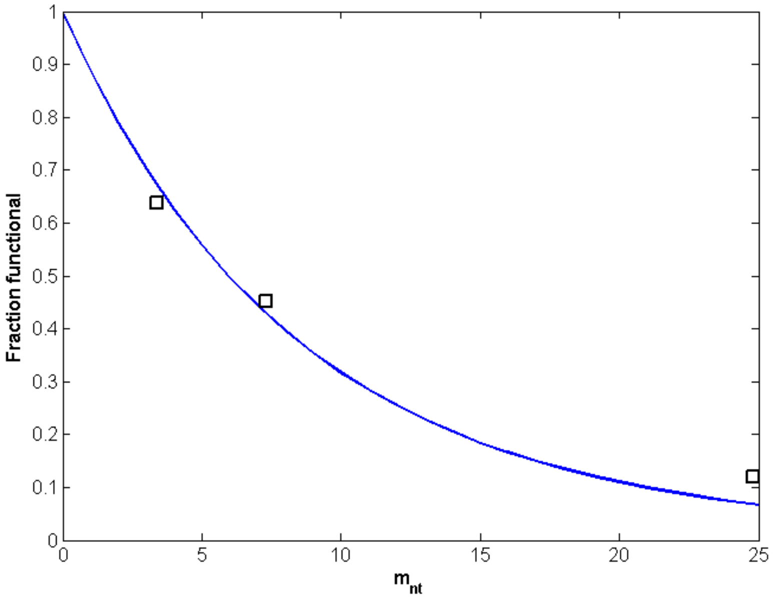


Fig. 3 Regression analysis of the experimental data (m_{nt} vs. selection power) to estimate ν. The blue line fits the data ('□') in Matlab 2009a (MathWorks, Natick, MA), based on the equation (5).

Next, the probability that a non-truncated sequences with m_{nt} substitutions can be calculated by

$$P(non-truncated \mid M = m_{nt}) = (1 - p_{tr}/p_{ns})^{m_{nt}} \qquad (6)$$

Suppose there are N transformants in the epPCR library. On average,

$$N_m = N \cdot P(M = m_{nt})P(non-truncated \mid M = m_{nt})$$

are non-truncated proteins with *m* nucleotide mutations.

If with one nucleotide mutation per codon, an average of 5.7 amino acid substitutions (out of a maximum of 19) is accessible due to the conservatism of the genetic code, the *m* nucleotide mutations will conduce to

$$M_m = \binom{L/3}{m} 5.7^m \qquad (7)$$

the average number of unique proteins, where L is the length of the gene in nucleotides.

Give N_m samples, the expected number of unique sequences produced by equal-probably sampling M_m sequences is

$$U_m = M_m - M_m(1-1/M_m)^{N_m} \approx M_m\left(1-e^{-N_m/M_m}\right) \tag{8}$$

Consequently, the total number of unique sequences in a library is the sum over all unique sequences with a specific number of substitutions:

$$U = \sum_{m=0}^{L/3} U_m \tag{9}$$

and the number of unique sequences that retain at least wide-type function is

$$U_f = \sum_{m=0}^{L/3} U_m v^m \tag{10}$$

Finally, the fraction of unique sequences (*P(U)*) and unique functional sequences(*P(U$_f$)*) are

$$P(U) = U/N \tag{11}$$

$$P(U_f) = U_f/N \tag{12}$$

respectively. The relationship between mutation rates and fraction unique sequences is shown in SOM-Fig. 4. It is clear the unique sequences are enriched in the high-frequency mutagenesis. However, the maximum fraction of unique sequences is relatively small (*P(U) < 0.5*) even as <m$_{nt}$> = 6, because sequences truncated by fameshifts and stop codons accumulate at increasing levels as the mutation rate is increased (Drummond et al. 2005a; Miura and Sonigo 2001).

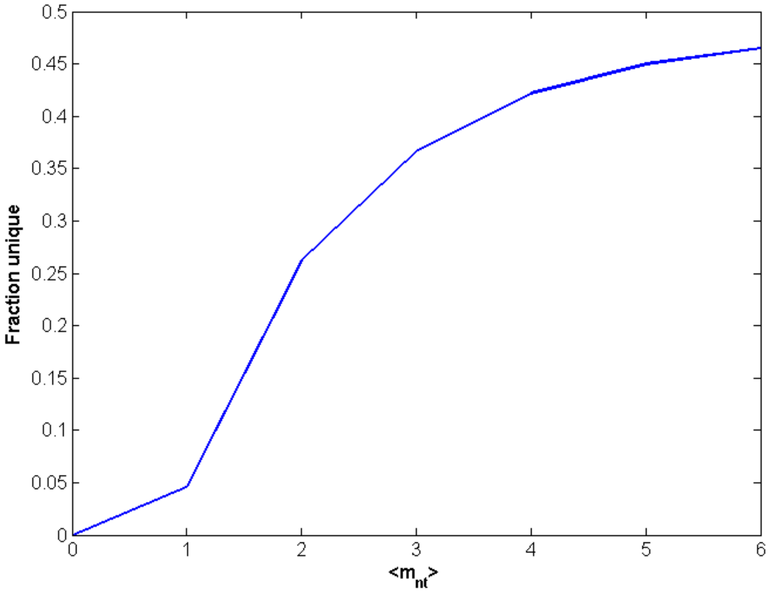


Fig. 4 The relationship between error rates and fraction of unique sequences. The line is simulated by eq. (11) with *n* = 25 thermal cycles, efficiency *λ* = 0.6, and library size *N* = 10,000 clones.

SOM-Fig. 5 illustrates the effects of mutation rates on the fraction of unique functional sequences $P(U_f)$. The results suggest optimal mutation rates are both protocol and protein-dependent. Under current circumstances with n = 25 thermal cycle, Taq DNA polymerase (λ = 0.6) and the structural plasticity (v = 0.91, p_{ns} = 0.8, p_{tr} = 0.07), we predict the optimal mutation rate for random mutagenesis of *Ctcbp* is 6 mutations per gene (SOM-Fig.4), which theoretically yield 4500+ unique mutants and more than 3,000 unique & functional CBP mutants over the 10, 000 clones.

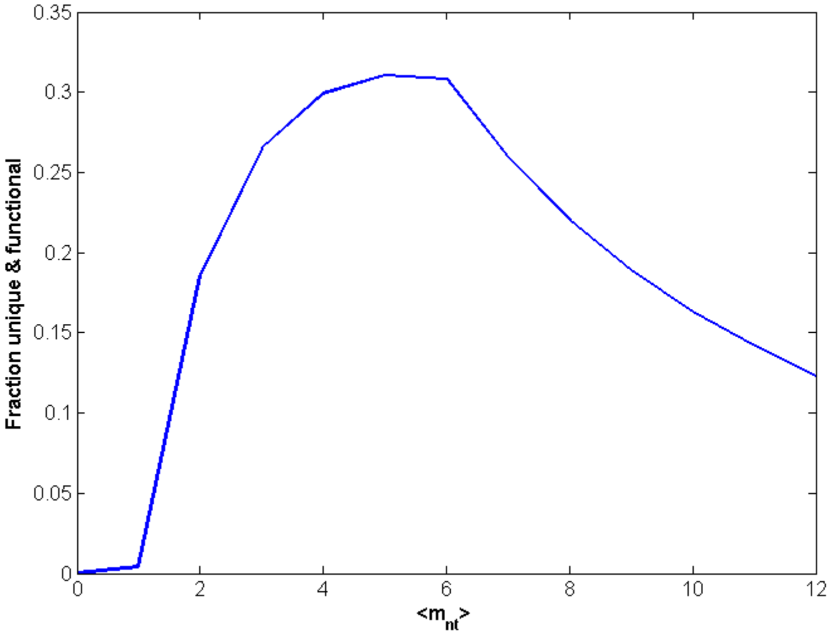


Fig. 5 The effect of mutagenesis frequencies on fraction of unique and functional sequences. The blue line is estimated by eq. (12) with n = 25 thermal cycles, PCR efficiency λ = 0.6, p_{ns} = 0.8, p_{tr} = 0.07, and library size N = 10,000 clones.

Based on previous work in our lab (Liu et al. 2009), we can achieve such a mutation rate (~ 0.25%) under the condition with regular PCR buffer with 5 mM $MgCl_2$ plus 3 mM $MnCl_2$, 0.2 mM dATP, 0.2 mM dGTP, 1 mM dCTP, and 1 mM dTTP. Practically, it resulted to a mutant library (Library O) with an average mutation rate 0.28%.

References:

Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. 2005. Thermodynamic prediction of protein neutrality. Proc Nat Acad Sci USA 102(3):606-611.

Clune J, Misevic D, Ofria C, Lenski RE, Elena SF, Sanjuán R. 2008. Natural Selection Fails to Optimize Mutation Rates for Long-Term Adaptation on Rugged Fitness Landscapes. PLoS Comput Biol 4(9):e1000187.

Daugherty PS, Chen G, Iverson BL, Georgiou G. 2000. Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. Proc Nat Acad Sci USA 97(5):2029-2034.

Dosztányi Z, Fiser A, Simon I. 1997. Stabilization centers in proteins:Identification, characterization and predictions. J. Mol. Biol. 272(4):597-612.

Dosztányi Z, Magyar C, Tusnády G, Simon I. 2003. SCide: identification of stabilization centers in proteins. Bioinformatics 19(7):899-900.

Drummond DA, Iverson BL, Georgiou G, Arnold FH. 2005a. Why high-error-rate random mutagenesis libraries are enriched in functional and improved proteins. J. Mol. Biol. 350(4):806-816.

Drummond DA, Silberg JJ, Meyer MM, Wilke CO, Arnold FH. 2005b. On the conservative nature of intragenic recombination. Proc Nat Acad Sci USA 102(15):5380-5385.

Humphrey W, Dalke A, Schulten K. 1996. VMD: Visual molecular dynamics. J. Mol. Graphics 14(1):33-38.

Kimura M. 1960. Optimum mutation rate and degree of dominance as determined by the principle of minimum genetic load. J Genet 57(1):21-34.

Liu W, Hong J, Bevan DR, Zhang Y-HP. 2009. Fast identification of thermostable beta-glucosidase mutants on cellobiose by a novel combinatorial selection/screening approach. Biotechnol. Bioeng. 103(6):1087-1094.

Miura T, Sonigo P. 2001. A mathematical model for experimental gene evolution. J. Theor. Biol. 209(4):497-502.

Pritchard L, Corne D, Kell D, Rowland J, Winson M. 2005. A general model of error-prone PCR. J. Theor. Biol. 234(4):497-509.

Shafikhani S, Siegel R, Ferrari E, Schellenberger V. 1997. Generation of large libraries of random mutants in Bacillus subtilis by PCR-based plasmid multimerization. BioTechniques 23(2):304-310.

Sun F. 1995. The polymerase chain reaction and branching processes. J. Comput. Biol. 2(1):63-86.

Zaccolo M, Gherardi E. 1999. The effect of high-frequency random mutagenesis on *in vitro* protein evolution: a study on TEM-1 β-lactamase. J. Mol. Biol. 285(2):775-783.

Zaccolo M, Williams DM, Brown DM, Gherardi E. 1996. An Approach to Random Mutagenesis of DNA Using Mixtures of Triphosphate Derivatives of Nucleoside Analogues. J. Mol. Biol. 255(4):589-603.