Supporting Information

Protein Sequence Selection Method That Enables Full Consensus Design of Artificial L-Threonine 3-Dehydrogenases With Unique Enzymatic Properties

Tomoharu Motoyama[a, 1], Nozomi Hiramatsu[a, 1], Yasuhisa Asano[b], Shogo Nakano[a, *] and Sohei Ito[a]

a. Graduate School of Integrated Pharmaceutical and Nutritional Sciences, University of Shizuoka, 52-1 Yada, Suruga-ku, Shizuoka 422-8526, Japan

b. Biotechnology Research Center and Department of Biotechnology, Toyama Prefectural University, 5180 Kurokawa, Imizu, Toyama 939-0398, Japan

1: These authors contributed equally to this work

*Correspondence to Shogo Nakano (snakano@u-shizuoka-ken.ac.jp)

Contents

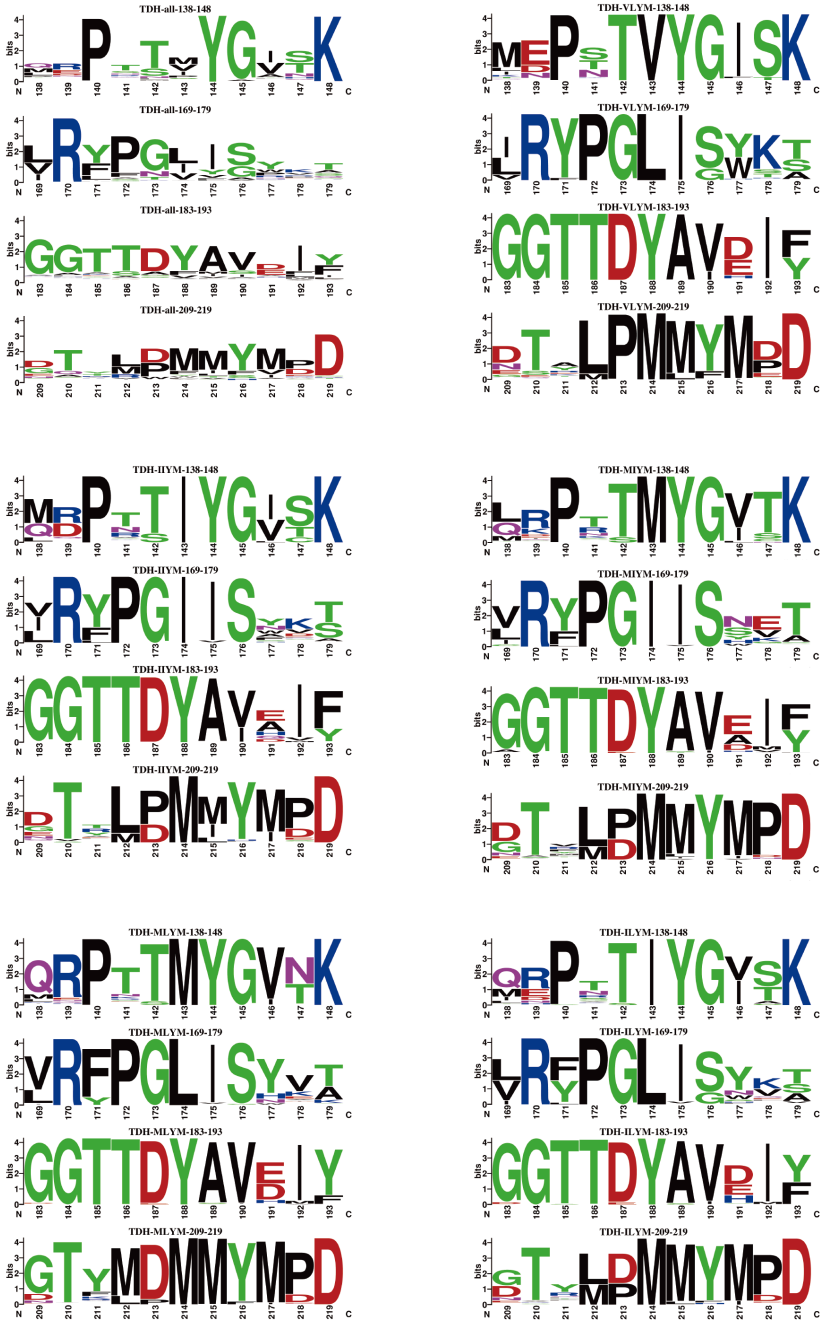


Figure S1. Sequence logo of non-curated (upper left), VLYM (upper right), IIYM (middle left), MIYM (middle right), MLYM (down left) and ILYM (down right) libraries of TDHs. Sequence logo was generated utilizing WebLogo tools (*1*).

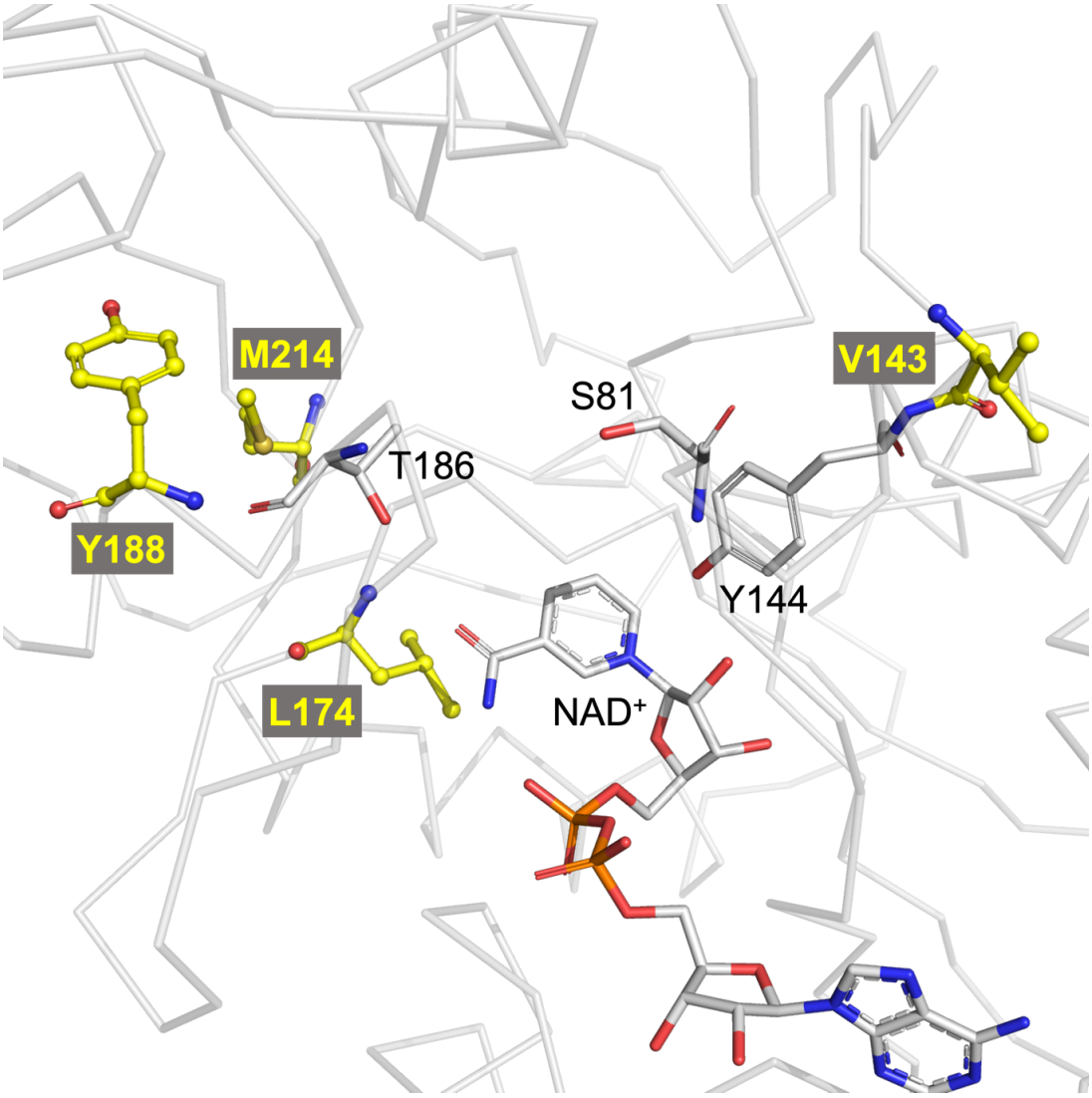


Figure S2. Structural position of active site residues and key residues on CnTDH structure. The active site residues (S81, Y144 and T186) and key residues (V143, L174, Y188, M214) were colored by gray and yellow, respectively.

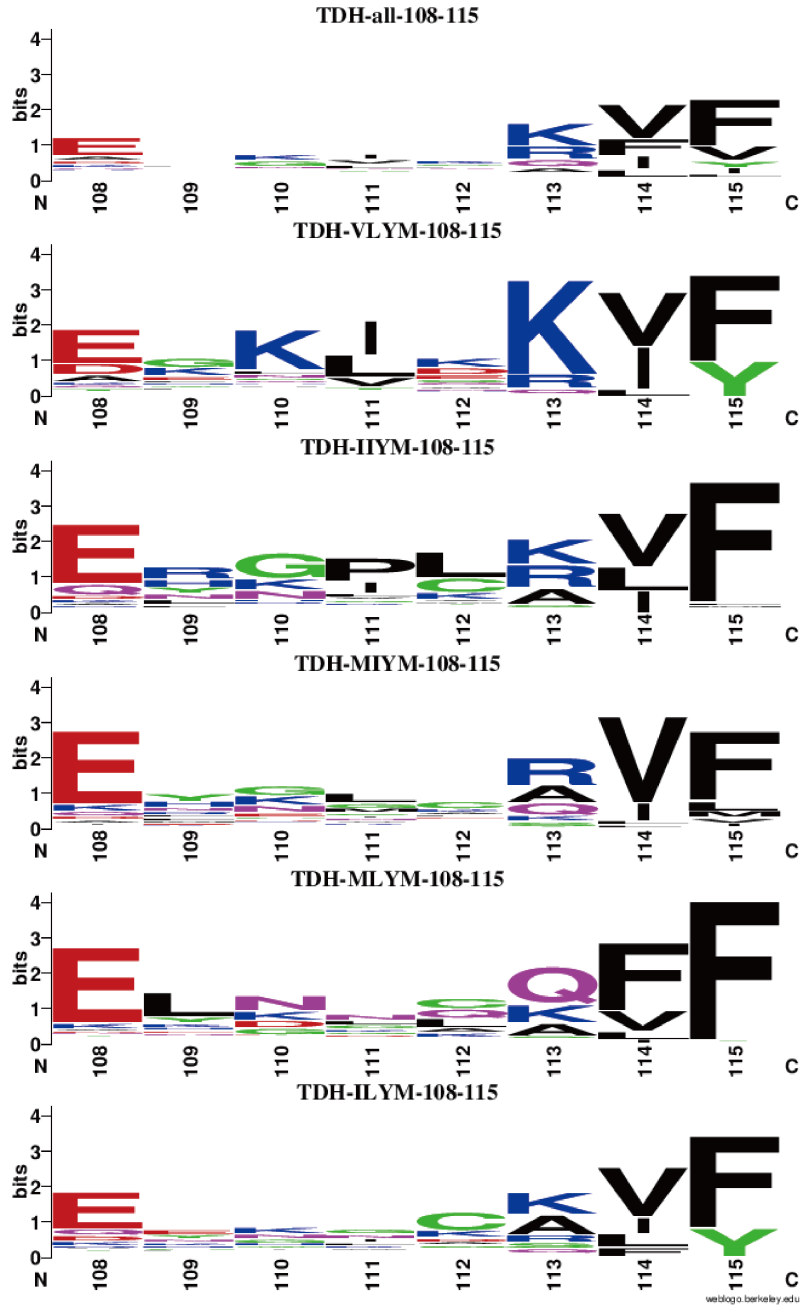


Figure S3. Sequence logo at 108-115 region of non-curated (1st row), VLYM (2nd row), IIYM (3rd row), MIYM (4th row), MLYM (5th row) and ILYM (6th row) libraries of TDHs. Sequence logo was generated utilizing WebLogo tools (*1*).

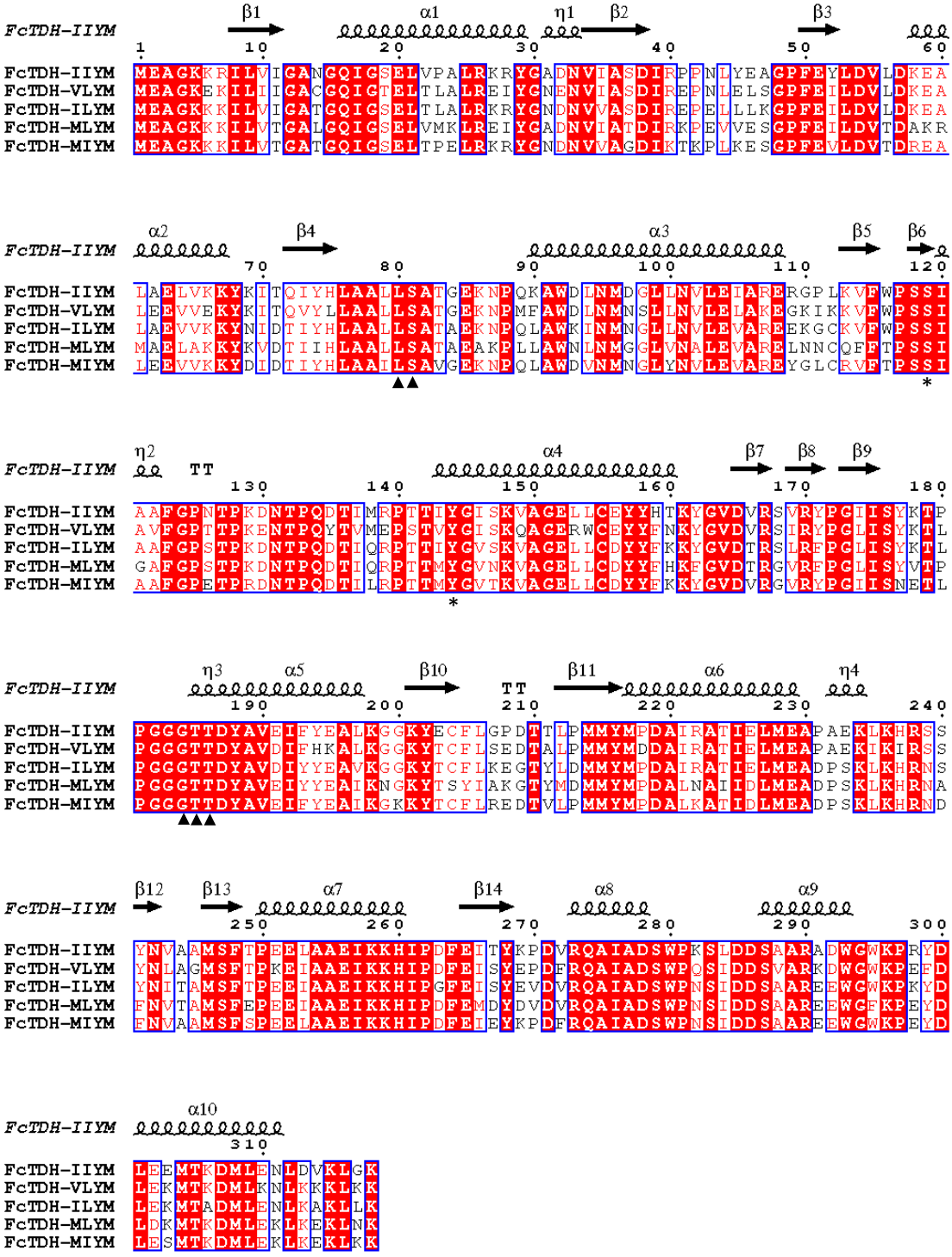
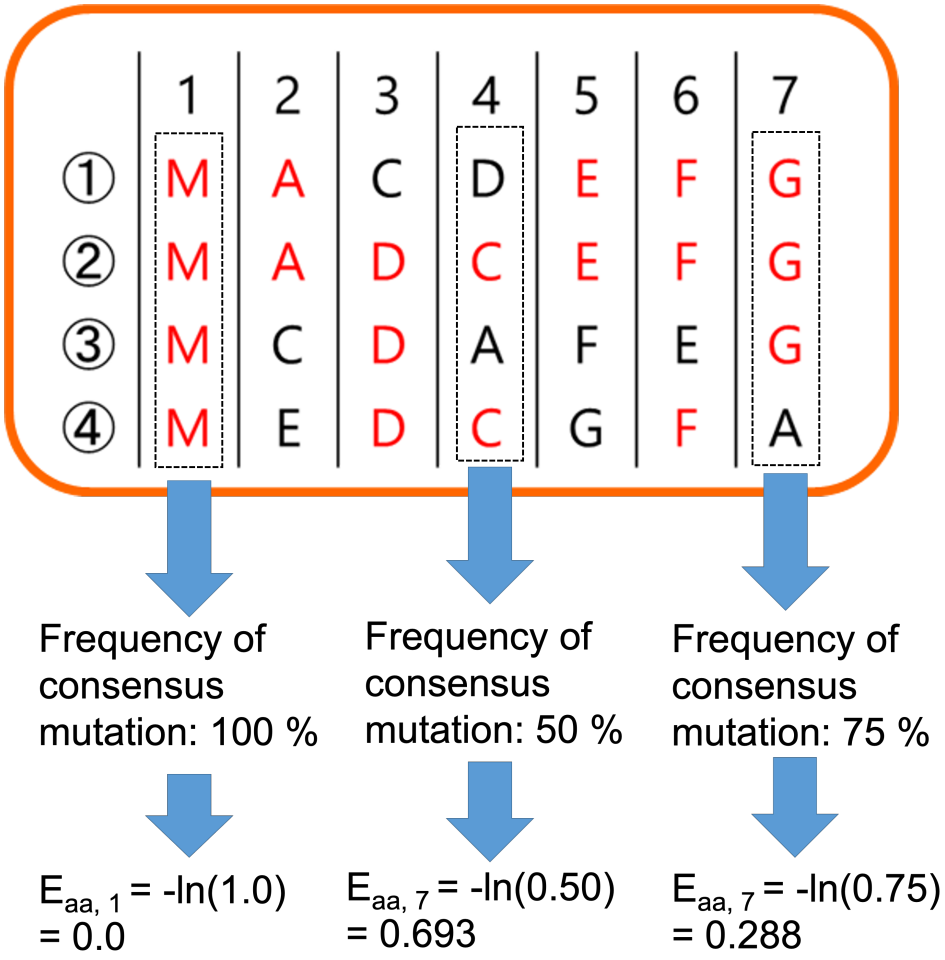


Figure S4. Multiple sequence alignment of five designed L-threonine 3-dehydrogenases (TDHs) by full consensus protein design method. The sequences of the designed TDHs were listed in Table S1. Catalytic residues are indicated by *. Alignment was performed using CLUSTALW, and the figure was generated by ESPript.



$$E_c = E_{aa,1}+E_{aa,2}+\ldots+E_{aa,7} = 1*0.0+ 3*0.693 + 3*0.288 = 2.94$$

Figure S5. Details how to calculate $E_{aa,i}$ and E_c value from the result of multiple sequence alignment. The $E_{aa,i}$ values were calculated for each residues (from 1 to 7th residues in this figure). The E_c value could be calculated by summing all E_{aa} residues (from $E_{aa,1}$ to $E_{aa,7}$ in this figure). The $E_{aa,i}$ and E_c value is near zero if the mutation was assigned accurately.

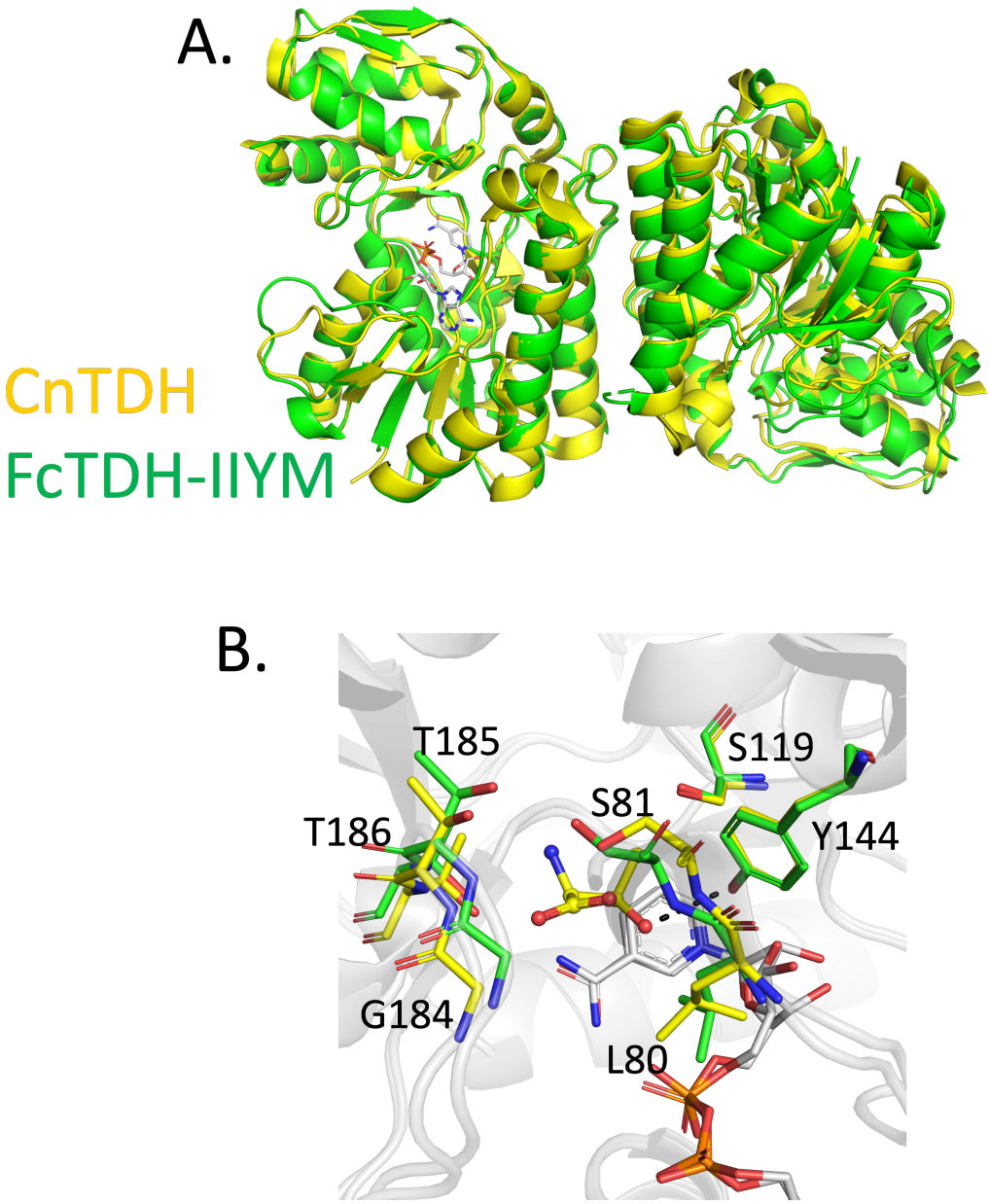


Figure S6. Structural comparison between CnTDH (PDB ID: 3WMX, yellow) and FcTDH-IIYM (green). Both overall structure (A) and active site structure (B) are highly conserved in each other. The root mean square deviation value for Cα atoms of these two structures was 0.49 Å.

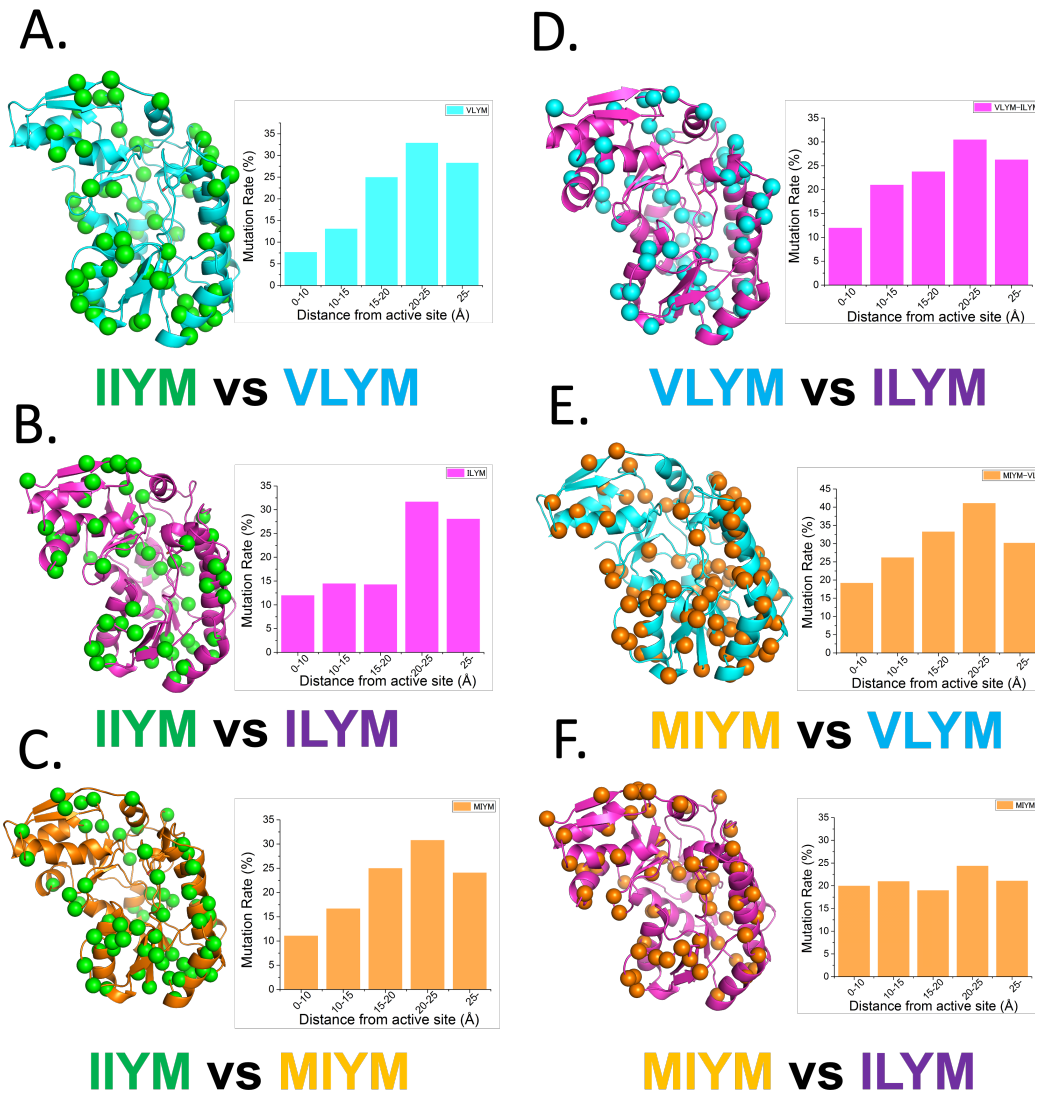


Figure S7. Location of mutation sites on homology models of each of the designed TDHs (A-F). The mutation sites which were confirmed by comparing sequence between FcTDH-IIYM (green sphere) and other three FcTDHs (FcTDH-VLYM, FcTDH-ILYM and FcTDH-MIYM were shown in Fig. S4A, S4B and S4C, respectively, utilizing the same color indicated in Fig.4.The comparison of mutation sites with other combinations was shown in Fig. S4D (FcTDH-VLYM and FcTDH-ILYM), S4E (FcTDH-MIYM and FcTDH-VLYM) and S4F (FcTDH-MIYM and FcTDH-ILYM), respectively.

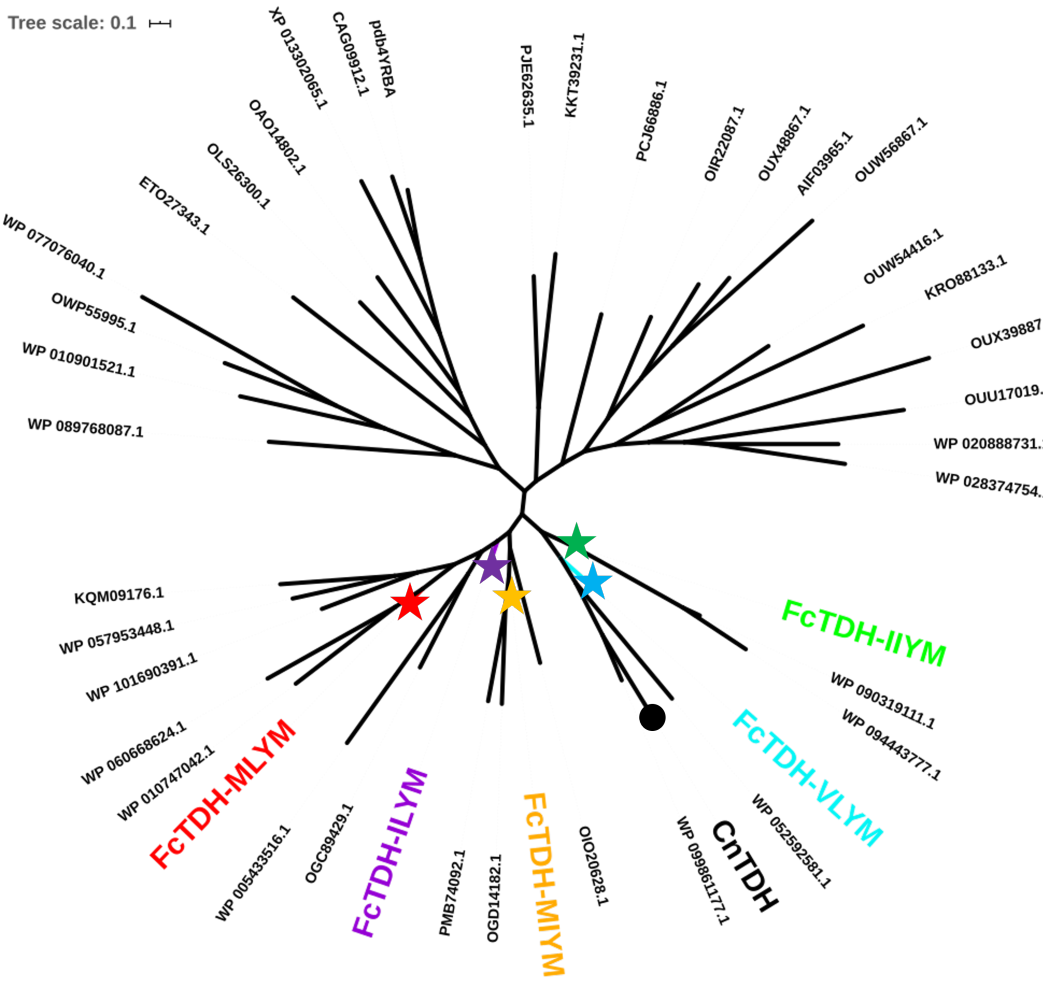


Figure S8. Phylogenetic analysis of the five designed TDHs and their homologous sequences. The phylogenetic tree was built by MEGA utilizing maximum likelihood method. Sequences in five of the curated libraries sharing >50% identity with other sequence were omitted by the Python script. The selected sequences are summarized in one data and used to build the tree.

Table S1. Protein sequences of five designed L-threonine 3-dehydrogenase (TDHs): FcTDH-VLYM, FcTDH-ILYM, FcTDH-IIYM, FcTDH-MIYM and FcTDH-MLYM.

>FcTDH-VLYM
MEAGKEKILIIGACGQIGTELTLALREIYGNENVIASDIREPNLELSGPFEILDVLDKEALEEVVEKY
KITQVYLLAALLSATGEKNPMFAWDLNMNSLLNVLELAKEGKIKKVFWPSSIAVFGPTTPKENTP
QYTVMEPSTVYGISKQAGERWCEYYFNKYGVDVRSIRYPGLISYKTLPGGGTTDYAVDIFHKAL
KGGKYTCFLSEDTALPMMYMDDAIRATIELMEAPAEKIKIRSSYNLAGMSFTPKEIAAEIKKHIPD
FEISYEPDFRQAIADSWPQSIDDSVARKDWGWKPEFDLEKMTKDMLKNLKKKLKK

>FcTDH-ILYM
MEAGKKKILVIGATGQIGSELTLALRKRYGNDNVVASDIREPELLLKGPFEILDVLDKEALAEVVK
KYNIDTIYHLAALLSATAEKNPQLAWKINMNGLLNVLEVAREEKGCKVFWPSSIAAFGPSTPKDN
TPQDTIQRPTTIYGVSKVAGELLCDYYFKKYGVDTRSLRFPGLISYKTLPGGGTTDYAVDIYYEA
VKGGKYTCFLKEGTYLDMMYMPDAIRATIELMEADPSKLKHRNSYNITAMSFTPEEIAAEIKKHIP
GFEISYEVDVRQAIADSWPNSIDDSAAREEWGWKPKYDLEKMTADMLENLKAKLLK

>FcTDH-IIYM
MEAGKKRILVIGANGQIGSELVPALRKRYGADNVIASDIRPPNLYEAGPFEYLDVLDKEALAELVK
KYKITQIYHLAALLSATGEKNPQKAWDLNMDGLLNVLEIARERGPLKVFWPSSIAAFGPNTPKDN
TPQDTIMRPTTIYGISKVAGELLCEYYHTKYGVDVRSVRYPGIISYKTPPGGGTTDYAVEIFYEAL
KGGKYECFLGPDTTLPMMYMPDAIRATIELMEAPAEKLKHRSSYNVAAMSFTPEELAAEIKKHIP
DFEITYKPDVRQAIADSWPKSLDDSAARADWGWKPRYDLEEMTKDMLENLDVKLGK

>FcTDH-MIYM
MEAGKKKILVTGATGQIGSELTPELRKRYGNDNVVAGDIKTKPLKESGPFEVLDVTDREALEEVV
KKYDIDTIYHLAAILSAVGEKNPQLAWDVNMNGLYNVLEVAREYGLCRVFTPSSIAAFGPETPRD
NTPQDTILRPTTMYGVTKVAGELLCDYYFKKYGVDVRGVRYPGIISNETLPGGGTTDYAVEIFYE
AIKGKKYTCFLREDTVLPMMYMPDALKATIDLMEADPSKLKHRNDFNVAAMSFSPEELAAEIKK
HIPDFEIEYKPDFRQAIADSWPNSIDDSAAREEWGWKPEYDLESMTKDMLEKLKEKLKK

>FcTDH-MLYM
MEAGKKKILVTGALGQIGSELVMKLREIYGADNVIATDIRKPEVVESGPFEILDVTDAKRMAELAK
KYKVDTIIHLAALLSATAEAKPLLAWNLNMGGLVNALEVARELNNCQFFTPSSIGAFGPSTPKDN
TPQDTIQRPTTMYGVNKVAGELLCDYYFHKFGVDTRGVRFPGLISYVTPPGGGTTDYAVEIYYE
AIKNGKYTSYIAKGTYMDMMYMPDALNAIIDLMEADPSKLKHRNAFNVTAMSFEPEEIAAEIKKHI
PDFEMDYDVDVRQAIADSWPNSIDDSAAREEWGFKPEYDLDKMTKDMLEKLKEKLNK

Table S2, Sequence identity between the designed TDHs

Sequence identity (%)	FcTDH-VLYM	FcTDH-ILYM	FcTDH-IIYM	FcTDH-MIYM	FcTDH-MLYM
FcTDH-VLYM	100	75	76	69	61
FcTDH-ILYM	75	100	79	79	76
FcTDH-IIYM	76	79	100	76	69
FcTDH-MIYM	69	79	76	100	74
FcTDH-MLYM	61	76	69	74	100

Table S3. Specific activity of designed TDHs toward the 20 L-amino acids and other compounds.[a]

Substrate	Relative activity (%)				
	CnTDH(*2*)	FcTDH-VLYM	FcTDH-ILYM	FcTDH-IIYM	FcTDH-MIYM
L-Thr	100	100	100	100	100
DL-Thr	64	38.7	15.3	106	19.8
D-Thr	0	0	0	0	0
(*R*)-3-HB	n.m.	3.2	0	20.9	1.0
(*rac*)-3-HB	n.m.	0	0	7.8	0
(*S*)-3-HB	n.m.	0	0	0	0
L-Ser	0	0	0	0	0
L-Amino acid					
L-Ala	0	0	0	0	0
L-Cys	0	0	0	0	0
L-Asp	0	0	0	0	0
L-Glu	0	0	0	0	0
L-Phe	0	0	0	0	0
Gly	0	0	0	0	0
L-His	0	0	0	0	0
L-Ile	0	0	0	0	0
L-Lys	0	0	0	0	0
L-Leu	0	0	0	0	0
L-Met	0	0	0	0	0
L-Asn	0	0	0	0	0
L-Pro	0	0	0	0	0
L-Gln	0	0	0	0	0
L-Arg	0	0	0	0	0
L-Val	0	0	0	0	0
L-Trp	0	0	0	0	0
L-Tyr	0	0	0	0	0

[a] There is no activity toward 20 of D-amino acids.

Table S4. Taxonomy in the curated libraries to design the four full consensus TDHs.

	FcTDH-VLYM (403seqs)	FcTDH-ILYM (115seqs)	FcTDH-IIYM (103seqs)	FcTDH-MIYM (129seqs)
1	*Flavobacterium*: 8.9% (1.9%)[b]	*Prevotella*: 13.9% (1.4%)	*Massilia*: 14.7% (0.8%)	*Candidatus*: 18.6% (7.3%)
2	*Mucilaginibacter*: 5.0% (1.1%)	*Chitinophaga*: 7.0% (0.8%)	*Candidatus*: 14.7% (7.3%)	*Bacteroidetes*: 16.3% (3.1%)
3	*Pedobacter*: 4.5% (1.2%)	*Flavobacteriales*: 6.1% (0.9%)	*Sphingobacterium*: 9.8% (0.5%)	*Bacteroides*: 9.3% (1.8%)
4	*Flavobacteriaceae*: 3.7%	*Candidatus*: 5.2%	*Alistipes*: 4.9%	*Thermoplasmatales*: 4.7%
5	*Cryseobacterium*: 3.5%	*Bacteroides*: 4.3%	*Prevotella*: 4.9%	*Ignavibacteria*: 3.9%
6	*Cupriavidus*[a]: 3.0%	*Flexibacter*: 2.6%	*Janthinobacterium*: 2.9%	*Marinilabiliales*: 3.1%
7	*Algoriphagus*: 3.0%	*Dethiosulfovibrio*: 2.6%	*Noviherbaspirillum*: 2.9%	*Janthinobacterium*: 2.3%
8	*Hymenobacter*: 3.0%	Others: 58.3%	Others: 45.2%	Others: 41.8%
9	*Sphingobacteriales*: 1.7%			
	Others: 63.7%			

[a]Genus of *Cupriavidus* includes *Ralstonia*.

[b]Rate of genes which are expressed in the corresponding genus in the non-curated library was calculated for top 3 of the genus in the curated libraries.

References

1. Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) WebLogo: a sequence logo generator, *Genome Res 14*, 1188-1190.
2. Ueatrongchit, T., and Asano, Y. (2011) Highly selective L-threonine 3-dehydrogenase from Cupriavidus necator and its use in determination of L-threonine, *Anal. Biochem. 410*, 44-56.