Supplementary Materials

1 *Supplementary methods*

The AEMD was written by perl5 and python 2.7 and the webserver is mainly based on Django. AEMD-Web provides users with an intuitive interface, enabling users to conveniently run protein engineering for improving enzyme stability, selectivity and activity.

1.1 Stability analysis process

The analysis pipeline about designing mutation sites for stability was showed in Fig. 1A. From the input target sequence, the tool executes the evolution- and energy-based approaches in parallel. For the evolution-based analysis, the homologous sequences of target protein were first detected through blast in local UniRef90 database. Then, we used hmmbuild to build a profile HMM (Eddy, 1998) which was used to search UniRef90 database again. After filtering the high similar proteins by cd-hit (Li and Godzik, 2006), we made a multiple sequence alignment for the remaining homologous sequences by Muscle (Edgar, 2004). We then obtained the intensity of coevolution and conservation for each residue, as well as the frequency of amino acids in each position. For the energy-based analysis, we first detected all homologous PDB structures by blast in local PDB database. We then sorted the homologous structures to obtain the best template PDB structure (Template PDB) by taking identity, coverage and resolution information into account. If the best PDB structure has 100% identity with the target enzyme, we use it directly in the next analysis; If the best PDB structure has identity less than 100% but more than 30%, we generated the PDB model of the target enzyme (Target PDB) by RosettaCM (Song, et al., 2013). After that, the ΔΔG was estimated for all point mutations (Length of enzyme * 19) by the FoldX (Guerois, et al., 2002) and Rosetta-ddG (Kellogg, et al., 2011), respectively.

Through the evolution- and energy-based analysis, we obtained four stability-associated properties for all point mutations, including intensity of coevolution (the number of residues coevolving with target residue), feasibility (frequency difference between original and other residues), $\Delta\Delta G^{foldx}$ and $\Delta\Delta G^{Rosetta}$. Then a computational prediction for the selection of point mutations was implemented based on the integration of these properties, and machine learning methods (SVM, support vector machines), and a training set from ProTherm database (Gromiha, et al., 2004). Finally, parts of point mutations which had the highest predictive score were selected and emailed to the users for further experimental verification. The reliability and applicability of this analysis had been demonstrated in the FRESCO (Wijma, et al., 2014) and FireProt (Bednar, et al., 2015). In the further, we hope to collect more precise stability-associated mutations and properties for improving the accuracy of the computational model.

1.2 Selectivity analysis process

The analysis about selecting mutations for specificity design was showed in Fig. 1B. The inputs need two files: one is the target sequence or target PDB file; the second is a substrate file with SDF format. If one of the input is protein sequence, we obtain the target PDB file in a similar way to that does in the stability design module. Based on the protein PDB and substrate SDF files, we first determined the interaction between ligand and protein backbones in two ways: 1). If the input substrate is the native substrate of the target enzyme, we directly used the native substrate for design; 2). If the input substrate is similar with native substrate of the target enzyme, we first make a flexible ligand alignment between the input substrate and native substrate using the “flex_align” function of Schrodinger software (QikProp, 2015), then the native substrate was replaced by the input substrate. Subsequently, the residues within 5Å distance from substrate were selected as the resfile input for the Rosetta “coupled moves” design method (Ollikainen, et al., 2015). This method will redesign (with 20 amino acids) and repack these residues. After multicycle optimizations for these candidate positions, an optimal residue assembly was offered for next round of experimental validation. To make a straightforward way to visualize the result, the optimal residues were shown with sequence logos using weblogo (Crooks, et al., 2004). It had been proved that the analysis can significantly increase the accuracy in both predicting ligand specificity altering mutations and binding site sequences (Ollikainen, et al., 2015).

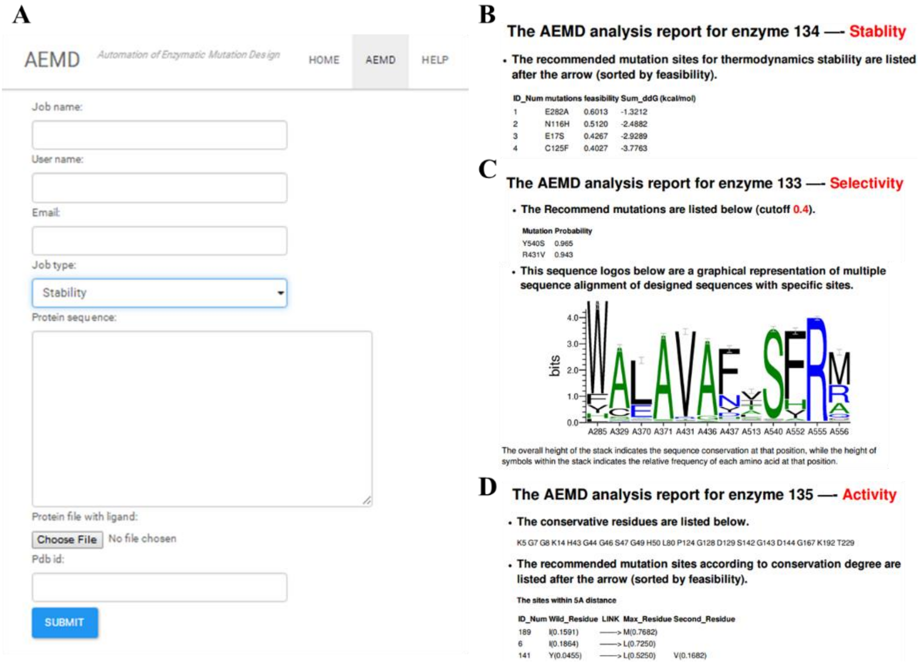
1.3 Activity analysis process

Because of the complexity of enzyme catalysis, it ' s difficult to predict point mutation improving protein activity accurately. We recently described a method which is able to identify desired mutations by analyzing the coevolution information of protein sequences (Liu, et al., 2016). In the AEMD-web, some point mutations are suggested by this method. Besides, our analysis generated some residues close to active center and transport tunnels which are recommended to saturated mutation to improve activity (Fig. 1C). For the input of target protein sequence, we first obtain the PDB file using RosettaCM (Song, et al., 2013). Next, the substrate of template PDB was mapped into target PDB using the " struct_align " funciton of Schrodinger software (QikProp, 2015). The spatial location of substrate in target PDB can help to determine the ligand-binding pocket of target enzyme. If all potential template PDB had no substrate in the PDB file, we predicted the ligand-binding pocket by a Rosetta script (gen_apo_grids.linuxgccrelease) (Zanghellini, et al., 2006). After the determination of ligand-binding pocket, we generated the possible catalytic sites by search local Catalytic Site Atlas (Furnham, et al., 2014); the residues within 5Å distance from ligands by calculating the minimum distance between residue and substrate; and the residues located within 3 Å distance from transport tunnels by CAVER (Chovancova, et al., 2012).

2 Supplementary Results

The AEMD-Web interface and interactive reports in the form of PDF are shown in figure S1. The analysis report for improving stability was showed in Fig. S1B, which showed the conservative residues in target enzyme, and the recommended mutation sites for thermodynamics stability. For example, " 1 M252L 0.7575 -2.2988 " represented that, we suggested to mutate the 252th methionine to leucine, the frequency difference between the 252th leucine and methionine is 0.7575 in all homologous enzymes, and the mutation ΔΔG is -2.2988 kcal/mol. The analysis report for improving specificity was showed in Fig. S1C. Firstly, the report lists the recommended mutations whose probability are greater than a cutoff (0.4) for selectivity engineering. For example, “Y540S 0.965” represented that, we advised to mutate the 540th tyrosine to serine at a probability of 0.965. Then, the relative amino acid bias of all designed positions is shown with a Sequence logos, and the height of each symbol within the stack indicates the relative frequency of each amino acid at that position. The analysis report for improving activity was showed in Fig. S1D, which showed the conservative residues in target enzyme, the residues located within 5 Å distance from substrate and cofactors and the residues located within 3 Å distance from transport tunnels. The residues close to active center and transport tunnel are recommended to saturated mutation to improve activity. We also showed some site-directed mutations based on the evolutional analysis. For example, “350 S(0.1349) -------> H(0.7084)” represented that, the frequency of the 350th native serine and the recommended histidine is 0.1349 and 0.7084, and we suggested to mutate the serine to histidine.

Fig. S1 The AEMD-Web interface and analysis reports.


A
AEMD
Automation of Enzymatic Mutation Design
HOME
AEMD
HELP
Choose File
SUBMIT
B
C
D
bits
4.0
3.0
2.0
1.0
0.0
A285 A329 A370 A371 A431 A436 A437 A513 A540 A552 A555 A556

Job name:

User name:

Email:

Job type: Stability

Protein sequence:

Protein file with ligand: No file chosen

Pdb id:

The AEMD analysis report for enzyme 134 —- Stablity

- The recommended mutation sites for thermodynamics stability are listed after the arrow (sorted by feasibility).

ID_Num	mutations	feasibility	Sum_ddG (kcal/mol)
1	E282A	0.6013	-1.3212
2	N116H	0.5120	-2.4882
3	E17S	0.4267	-2.9289
4	C125F	0.4027	-3.7763

The AEMD analysis report for enzyme 133 —- Selectivity

- The Recommend mutations are listed below (cutoff 0.4).

Mutation	Probability
Y540S	0.965
R431V	0.943

- This sequence logos below are a graphical representation of multiple sequence alignment of designed sequences with specific sites.

The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino acid at that position.

The AEMD analysis report for enzyme 135 —- Activity

- The conservative residues are listed below.

K5 G7 G8 K14 H43 G44 G46 S47 G49 H50 L80 P124 G128 D129 S142 G143 D144 G167 K192 T229

- The recommended mutation sites according to conservation degree are listed after the arrow (sorted by feasibility).

The sites within 5A distance

ID_Num	Wild_Residue	LINK	Max_Residue	Second_Residue
189	I(0.1591)	——>	M(0.7682)	
6	I(0.1864)	——>	L(0.7250)	
141	Y(0.0455)	——>	L(0.5250)	V(0.1682)

Note: The interface of the AEMD-web pipeline (A). In this page, we could select one of the three engineering types to use. And for the detailed inputs information for different types, please refer to the “HELP”. B, C and D represent the part of analysis reports for stability, selectivity and activity design, respectively.

Table S1. Overview the computational tools for enzyme engineering

Resources	Functions	Types	References
MSPocket	Detecting ligand-binding pocket	Software	(Zhu and Pisabarro, 2011)
TRITON	Detecting ligand-binding pocket	Software	(Prokop, et al., 2008)
CAVER	Analysis and visualization of tunnels and channels	Software	(Chovancova, et al., 2012)
Foldx	Engineering the stability of proteins and protein complexes	Software	(Van, et al., 2011)
ELASPIC	Predicting stability changes upon mutation	Web Services	(Witvliet, et al., 2016)
I-Mutant2.0	Predicting stability changes upon mutation	Web Services	(Capriotti, et al., 2005)
INPS	Predicting stability changes upon mutation	Web Services	(Fariselli, et al., 2015)
DUET	Predicting stability changes upon mutation	Web Services	(Pires, et al., 2014)
MAESTRO	Predicting stability changes upon mutation	Software + Web Services	(Laimer, et al., 2016)
PoPMuSiC	Predicting stability changes upon mutation	Web Services	(Dehouck, et al., 2011)
SABER	Selection of Active/Binding sites for Enzyme Redesign	Computational strategy	(Nosrati and Houk, 2012)
Janus	Prediction of Mutations Required for Functional Interconversion of Enzymes	Software	(Addington, et al., 2014)
ROSETTA	Enzyme design, structure modeling, ddG calculation and so on	Software	(Leaver-Fay, et al., 2011)
FRESCO	Computationally designed libraries for rapid enzyme stabilization	Computational strategy	(Wijma, et al., 2014)
FireProt	Computational Design of Thermostable Multiple-Point mutations	Computational strategy	(Bednar, et al., 2015)
ProSAR	Directed evolution approach	Computational strategy	(Fox, et al., 2007)

Table S2. Overview the computational tools using in AEMD

Resources	Functions	Types	References
ROSETTA	Enzyme design, structure modeling, ddG calculation and so on	Software	(Leaver-Fay, et al., 2011)
FoldX	Engineering the stability of proteins and protein complexes	Software	(Van Durme, et al., 2011)
CAVER	Analysis and visualization of tunnels and channels	Software	(Chovancova, et al., 2012)
HMMER	Protein sequence similarity searches	Software + Web Services	(Finn, et al., 2011)
MUSCLE	Multiple sequence alignment	Software	(Edgar, 2004)
ClustalW	Multiple sequence alignment	Software	(Thompson, et al., 2002)
trimAl	Automated alignment trimming	Software	(Capella-Gutirrez, et al., 2009)
SCA	Statistical coevolution analysis	Matlab based algorithm	(Süel, et al., 2003)
UCSF Chimera	structure preparation and refinement	Software	(Pettersen, et al., 2004)
Openbabel	structure format identification and conversion	Software	(O'Boyle, et al., 2011)
WebLogo	sequence logo graph construction	Software + Web Services	(Crooks, et al., 2004)
cd-hit	clustering and comparing large sets of protein sequences	Software + Web Services	(Li and Godzik, 2006)
Circos	An information aesthetic for comparative genomics	Software	(Krzywinski, et al., 2009)

Table S3. The running time of all examples

Type	Testing set	Sequence length	Running time
Activity	3LKK	245	7.2h
Selectivity	2FZN+HYP	602	1.5h
	1FCB +173	511	45min
	2O7B +TCA	523	1.8h
	1A80+NAD	277	20min
	1PK7+TAL	237	15min
	1K70+FPY	426	25min
	2H6F +GER	382	15min
	3HG5+A2G	398	33min
Stability	1BN6	294	6.5h
	1BNI	110	4h
	1BVC	153	3h
	1CSP	67	2h
	1LZ1	130	4h
	1RN1	104	3h
	1VQB	87	1.5h
	2CI2	83	1.5h
	2LZM	164	2.5h
	2RN2	155	2.5h
	4LYZ	129	5h

Note. The time consumption statistic of three different type of engineering were show in table S3, the difference of time consumption mainly dependent on the sequence length and Job type. The calculations were implemented in CentOS 6.6. Jobs were executed using machines running 64 bit, 12-core, two 2.2GHz processors with 24 GB of memory.

Table S4. Comparing the stability analysis pipeline in AEMD with FireProt

FireProt			**AEMD**			
mutations	**ddG_FoldX**	**ddG_Rosetta**	**mutations**	**feasibility**	**ddG_FoldX**	**ddG_Rosetta**
E20Q	−1.09	−2.13	E20Q	0.012	−1.41	−2.38
C128F	−2.21	−8.45	C128F	0.428	−1.26	−2.51
C128M	−3.48	−2.96	C128M	−0.006	−3.32	−1.63
T148W	−1.09	−2.65	T148W	−0.061	−0.93	−0.07
T148L	−1.96	−2.00	T148L	0.188	−2.06	−2.59
C176F	−2.22	−7.07	C176F	0.006	−2.78	−4.75
C176L	−2.01	−5.28	C176L	0.004	−2.97	−3.63
C176H	−1.08	−4.82	C176H	0.006	−2.11	−3.95
C176M	−2.51	−4.24	C176M	0.002	−2.91	−3.14
D187W	−1.37	−2.58	D187W	−0.261	−0.88	−3.39
D198W	−1.36	−4.55	D198W	−0.218	−0.66	−3.41
D198F	−1.98	−2.95	D198F	−0.224	−1.77	−0.24
D198Y	−1.85	−2.75	D198Y	−0.210	−1.78	−0.65
D198L	−1.92	−2.53	D198L	−0.220	−1.28	−1.12
N217Y	−2.38	−2.38	N217Y	−0.018	−2.56	0.98
V219W	−1.77	−3.04	V219W	−0.392	−1.54	−4.49
C262L	−1.64	−4.93	C262L	0.234	−2.01	−0.93
C262M	−1.42	−2.94	C262M	−0.065	−2.42	1.56
D266Y	−2.43	−2.90	D266Y	0.008	−1.22	−1.91
D266F	−2.31	−2.41	D266F	−0.038	−1.56	−1.87

Note: The proposed mutations by FireProt (the left three columns) were obtained from Table S4 in Bednar, et al., 2015. The ddG represented the change of Gibbs free energy (ΔΔG) after the mutation, and the unit of ddG is kcal/mol. The feasibility represented the frequency difference between the native and the recommended residues in all homologous enzymes. 20 out of 22 mutations were proposed by the stability analytic pipeline. The rest two candidates may be due to a low resolution protocol and a different weight file (“soft_rep_design”) were used in Rosetta ddg-monomer module in our pipeline for improving the efficiency.

Table S5. The AEMD selectivity analysis pipeline results

PDB	Ligand	Mutation	Catalytic center	Rank
2FZN	HYP	Y540S	√	1
1FCB	173	-	√	-
2O7B	TCA	H89F	√	37
1A80	NAD	K232G	√	-
1PK7	TAL	M64V	√	46
1K70	FPY	D314A	√	14
2H6F	GER	-	√	-
3HG5	A2G	E203S	√	3

Note: The AEMD selectivity pipeline analysis result of eight experimentally validated specificity engineering mutations, low ranking results my due to the limits of current selectivity engineering strategies.

Table S6. The AEMD activity analysis pipeline results

Mutation	Catalytic center	Tunnels	Surface	Recommend
G45A	√	-	-	√
V73T	√	-	-	√
V73I	√	-	-	-
V130A	√	-	-	-
I140V	√	-	-	-
Y141L	√	√	-	√
Y141V	√	-	-	-
K204A	√	-	√	-
K204G	√	-	√	-

Note: The AEMD activity pipeline analysis result for isopentenyl phosphate kinase (IPK) mutants, all of the six positions (45, 73, 130, 140, 141 and 204) reported were accurately predicted in the substrate binding pocket or the substrate channels, and 3 out of 9 point mutations (G45A, V73T and Y141L) were listed in the recommendation part of final report.

References

Addington, T.A.*, et al.* (2014) JANUS: Prediction and Ranking of Mutations Required for Functional Interconversion of Enzymes, *Journal of molecular biology*, **425**, 1378-1389.

Bednar, D.*, et al.* (2015) FireProt: Energy- and Evolution-Based Computational Design of Thermostable Multiple-Point Mutants, *Plos Computational Biology*, **11**, e1004556.

Capella-Gutirrez, S., Silla-Martínez, J.M. and Gabaldn, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses, *Bioinformatics*, **25**, 1972-1973.

Capriotti, E., Fariselli, P. and Casadio, R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure, *Nucleic acids research*, **33**, 306-310.

Chovancova, E.*, et al.* (2012) CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures, *Plos Computational Biology*, **8**, e1002708.

Crooks, G.E.*, et al.* (2004) WebLogo: a sequence logo generator, *Genome research*, **14**, 1188-1190.

Dehouck, Y.*, et al.* (2011) PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality, *BMC Bioinformatics*, **12**, 151.

Eddy, S.R. (1998) Profile hidden Markov models, *Bioinformatics*, **14**, 755-763.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic acids research*, **32**, 1792-1797.

Fariselli, P.*, et al.* (2015) INPS: predicting the impact of non-synonymous variations on protein stability from sequence, *Bioinformatics*, **31**, 2816-2821.

Finn, R.D., Clements, J. and Eddy, S.R. (2011) HMMER web server: interactive sequence similarity searching, *Nucleic acids research*, **39**, W29-W37.

Fox, R.J.*, et al.* (2007) Improving catalytic function by ProSAR-driven enzyme evolution, *Nature Biotechnology*, **25**, 338.

Furnham, N.*, et al.* (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes, *Nucleic acids research*, **42**, 485-489.

Gromiha, M.M.*, et al.* (2004) ProTherm, version 4.0: Thermodynamic Database for Proteins and Mutants, *Nucleic acids research*, **32**, D120-121.

Guerois, R., Nielsen, J.E. and Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations, *Journal of molecular biology*, **320**, 369-387.

Kellogg, E.H., Leaverfay, A. and Baker, D. (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability, *Proteins Structure Function & Bioinformatics*, **79**, 830-838.

Krzywinski, M.*, et al.* (2009) Circos: An information aesthetic for comparative genomics, *Genome research*, **19**, 1639-1645.

Laimer, J.*, et al.* (2016) MAESTROweb: a web server for structure-based protein stability prediction, *Bioinformatics*, **32**.

Leaver-Fay, A.*, et al.* (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules, *Methods in enzymology*, **487**, 545.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, **22**, 1658-1659.

Liu, Y.*, et al.* (2016) Improving the catalytic activity of isopentenyl phosphate kinase through protein coevolution analysis, *Sci Rep*, **6**, 24117.

Nosrati, G.R. and Houk, K.N. (2012) SABER: a computational method for identifying active sites for new reactions, *Protein Science*, **21**, 697-706.

O'Boyle, N.M.*, et al.* (2011) Open Babel: An open chemical toolbox, *Journal of cheminformatics*, **3**, 33.

Ollikainen, N., de Jong, R.M. and Kortemme, T. (2015) Coupling Protein Side-Chain and Backbone Flexibility Improves the Re-design of Protein-Ligand Specificity, *Plos Computational Biology*, **11**, e1004335.

Pettersen, E.F.*, et al.* (2004) UCSF Chimera⍰a visualization system for exploratory research and analysis, *Journal of computational chemistry*, **25**, 1605-1612.

Pires, D.E., Ascher, D.B. and Blundell, T.L. (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach, *Nucleic acids research*, **42**, 314-319.

Prokop, M.*, et al.* (2008) TRITON: a graphical tool for ligand-binding protein engineering, *Bioinformatics*, **24**, 1955-1956.

QikProp (2015) Schrodinger, LLC, New York.

Song, Y.*, et al.* (2013) High-resolution comparative modeling with RosettaCM, *Structure*, **21**, 1735-1742.

Süel, G.M.*, et al.* (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins, *Nature structural & molecular biology*, **10**, 59-69.

Thompson, J.D., Gibson, T. and Higgins, D.G. (2002) Multiple sequence alignment using ClustalW and ClustalX, *Current protocols in bioinformatics*, 2.3. 1-2.3. 22.

Van, D.J.*, et al.* (2011) A graphical interface for the FoldX forcefield, *Bioinformatics*, **27**, 1711-1712.

Van Durme, J.*, et al.* (2011) A graphical interface for the FoldX forcefield, *Bioinformatics*, **27**, 1711-1712.

Wijma, H.J.*, et al.* (2014) Computationally designed libraries for rapid enzyme stabilization, *Protein Engineering Design & Selection Peds*, **27**, 49-58.

Witvliet, D.K.*, et al.* (2016) ELASPIC web-server: proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity, *Bioinformatics*, **32**.

Zanghellini, A.*, et al.* (2006) New algorithms and an in silico benchmark for computational enzyme design, *Protein Science*, **15**, 2785-2794.

Zhu, H. and Pisabarro, M.T. (2011) MSPocket: an orientation-independent algorithm for the detection of ligand binding pockets, *Bioinformatics*, **27**, 351.