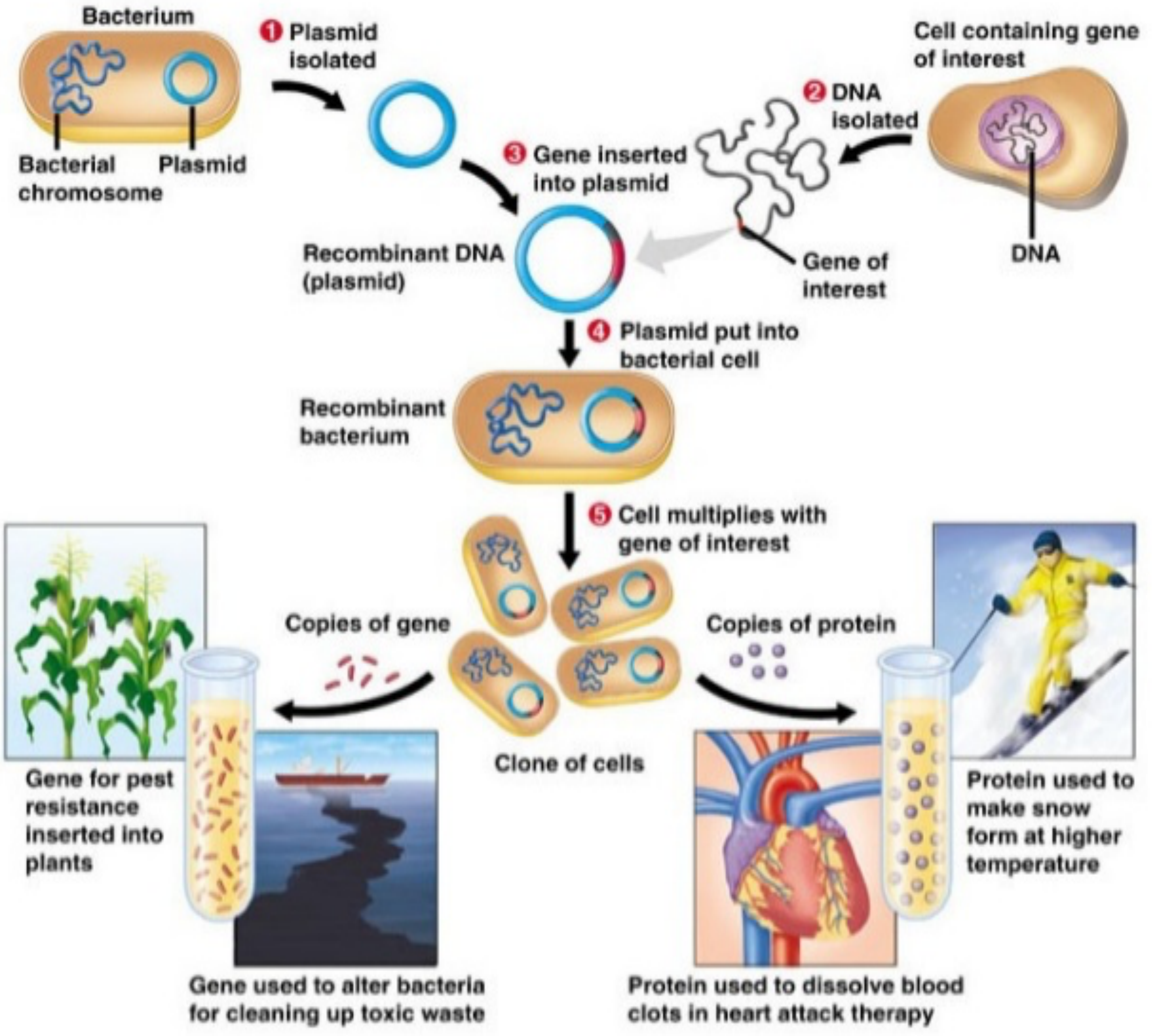
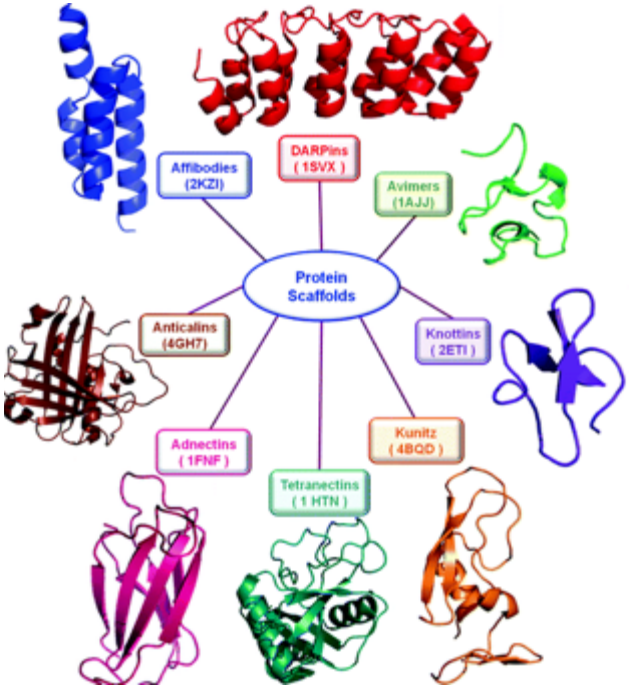
基因智能设计与蛋白智造

田 健

2020年2月17日

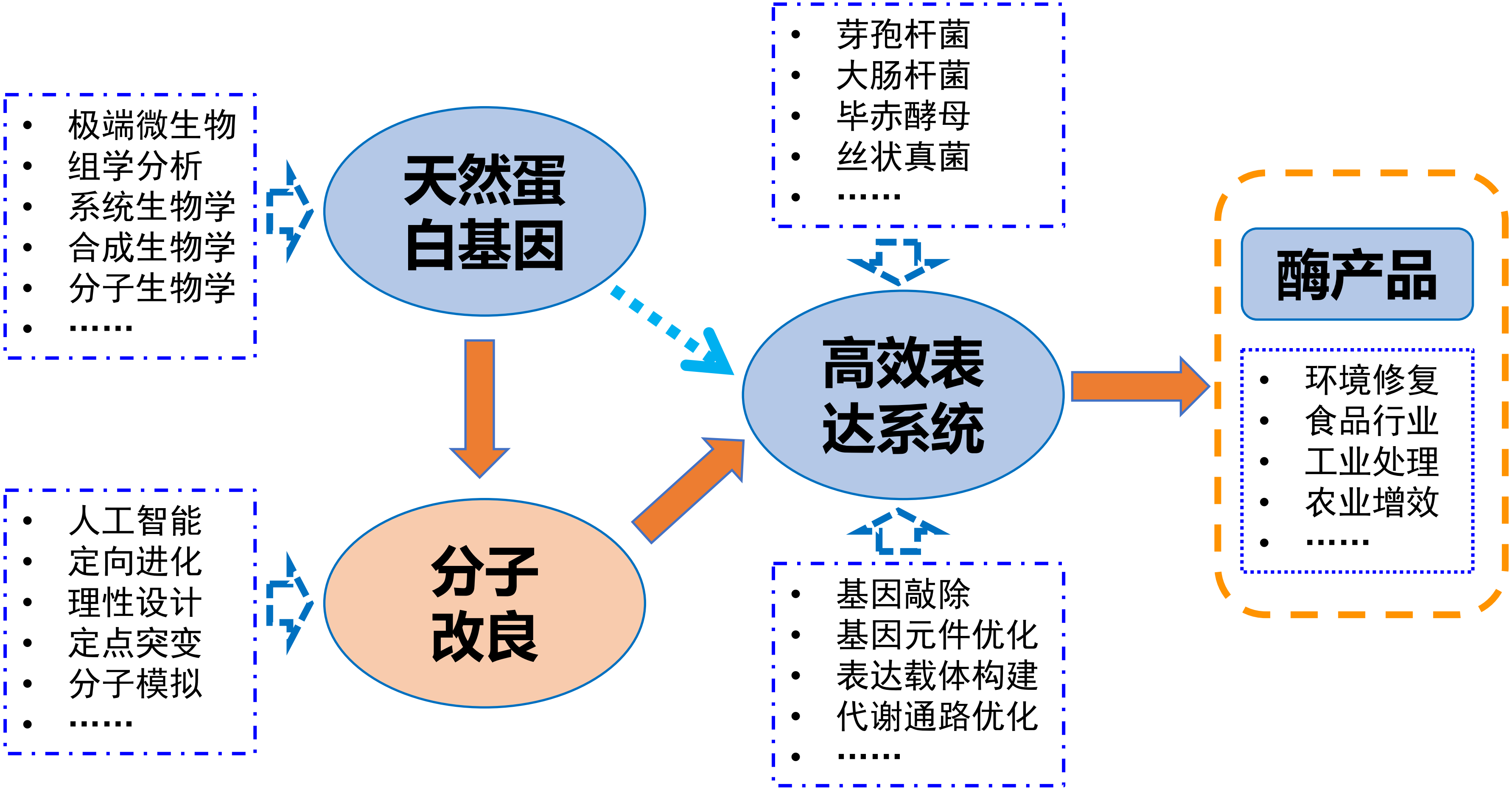
蛋白质工程


Bacterium
Bacterial chromosome
Plasmid
1 Plasmid isolated
Cell containing gene of interest
2 DNA isolated
DNA
3 Gene inserted into plasmid
Recombinant DNA (plasmid)
Gene of interest
4 Plasmid put into bacterial cell
Recombinant bacterium
5 Cell multiplies with gene of interest
Copies of gene
Copies of protein
Clone of cells
Gene for pest resistance inserted into plants
Gene used to alter bacteria for cleaning up toxic waste
Protein used to dissolve blood clots in heart attack therapy
Protein used to make snow form at higher temperature



蛋白质的结构

酶工程的研究流程



- 极端微生物
- 组学分析
- 系统生物学
- 合成生物学
- 分子生物学
- ……

天然蛋白基因

- 人工智能
- 定向进化
- 理性设计
- 定点突变
- 分子模拟
- ……

分子改良

- 芽孢杆菌
- 大肠杆菌
- 毕赤酵母
- 丝状真菌
- ……

高效表达系统

- 基因敲除
- 基因元件优化
- 表达载体构建
- 代谢通路优化
- ……

酶产品

- 环境修复
- 食品行业
- 工业处理
- 农业增效
- ……

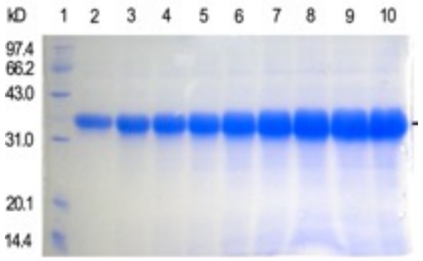
农业环保用酶的创制：有机磷降解酶

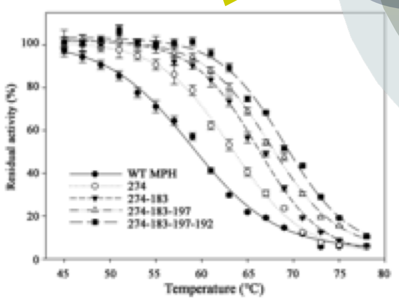
- 蛋白质耐热性的分子改良
- 蛋白质嗜酸性的分子改造
- 蛋白质外泌表达量的提升

基因克隆

- 高效降解有机磷农药菌株的筛选
- 有机磷污染土壤中宏基因组的研究
- 有机磷降解酶新基因的发现

蛋白质的分子改良





基因的高效表达

工业化生产

已克隆到高效降解多种有机磷农药的新基因，建立了一系列蛋白质分子改良方法

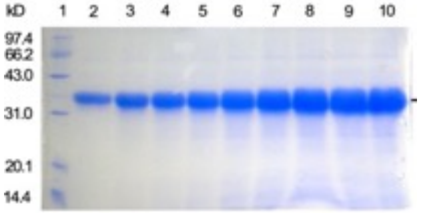
有机磷水解酶表达量提升140倍，现在该酶已商业化生产，可用于农药污染土壤的治理

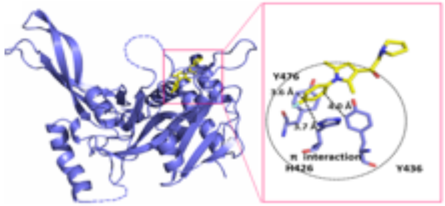
蛋白质分子改良的方向

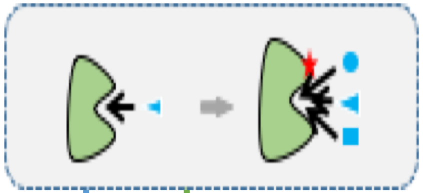
基因的高效表达

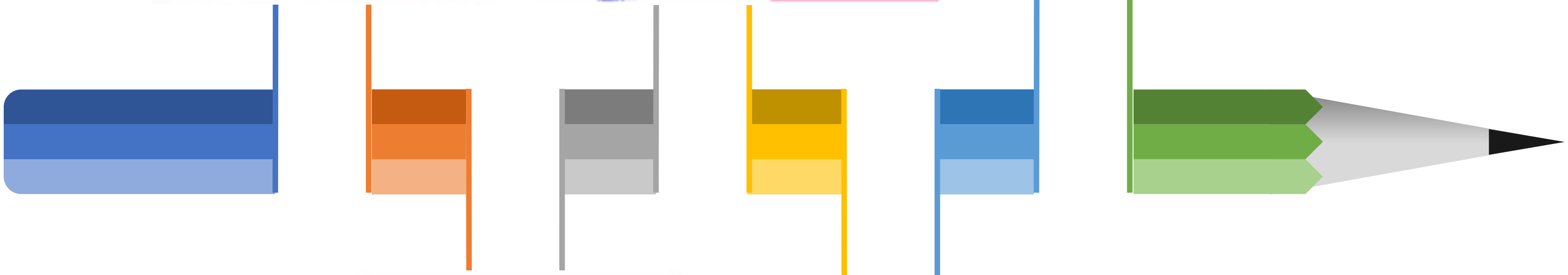
高催化活性

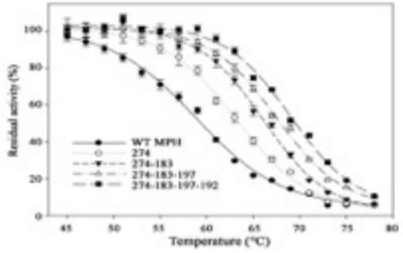
底物范围宽泛

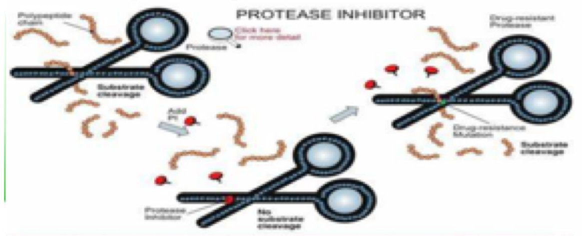












高稳定的蛋白质

蛋白酶抗性

提高基因的表达和蛋白质稳定性是使酶蛋白变成酶产品的关键技术，也是蛋白质研究中两个热点。

常用分子设计方法

- **定向进化（DNA随机突变）**

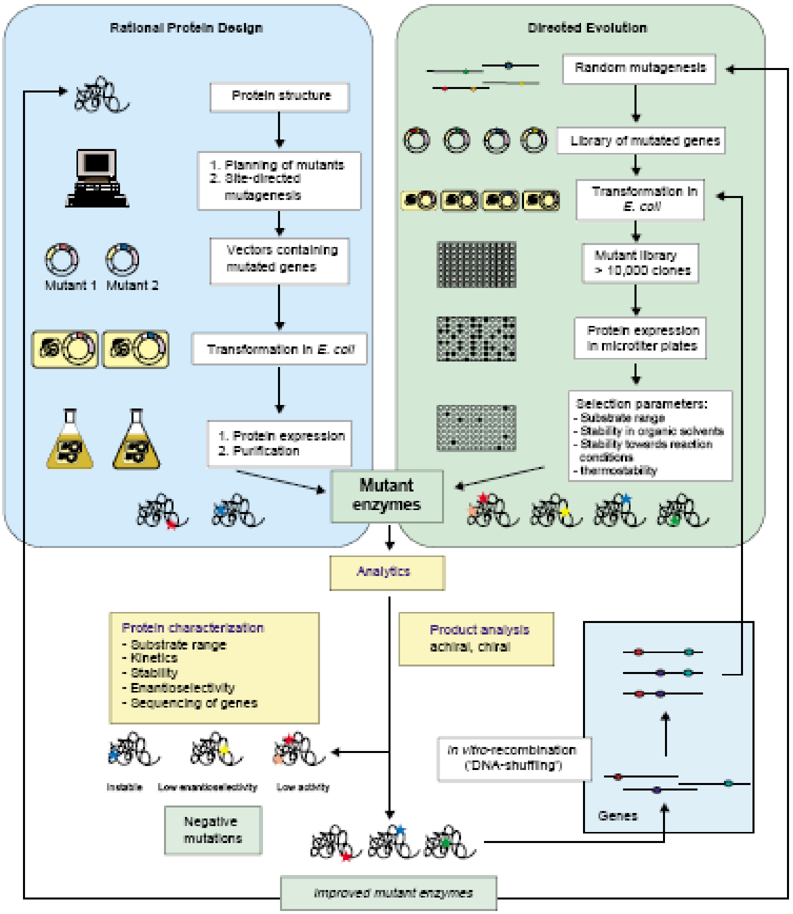
  工作量比较大，需要高通量、准确的筛选方法。

- **计算机辅助分子设计与智能设计**

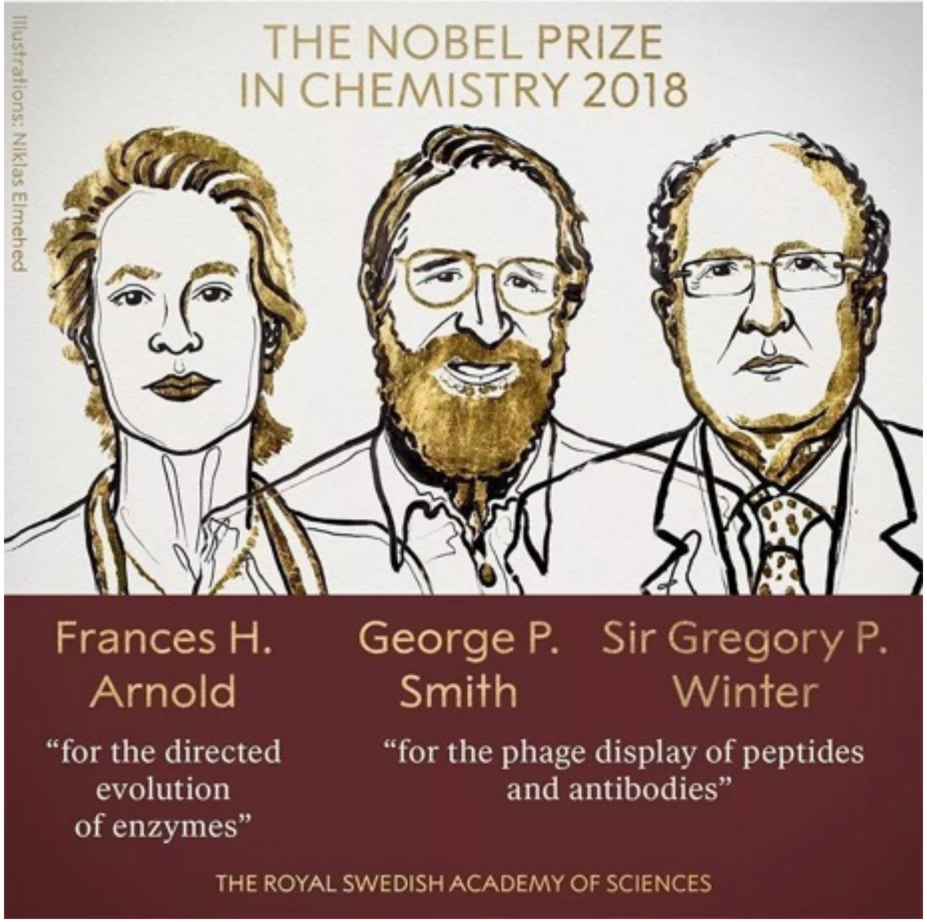
  需要对蛋白质的结构及影响蛋白质热性质的机制有着深入的了解。

- **序列比对**

  通过与热稳定性高的同源蛋白质进行序列比对确定突变位点，需要有一个热稳定性非常好的同源蛋白质。



2018年诺贝尔化学奖：驯服进化的力量


THE NOBEL PRIZE
IN CHEMISTRY 2018

Frances H. Arnold

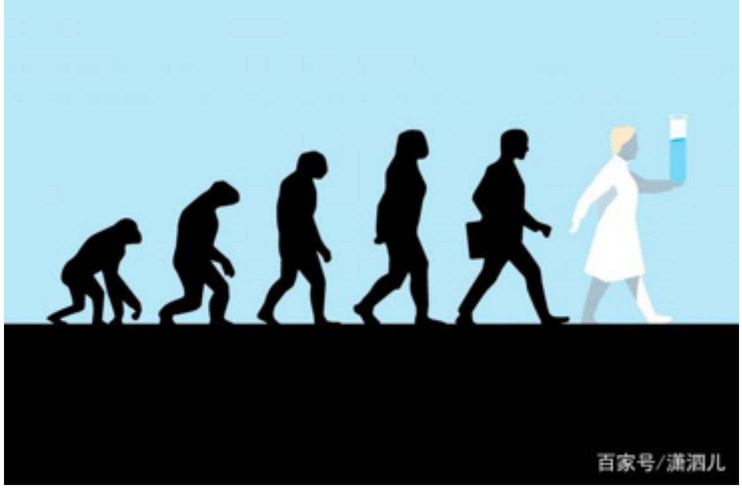
"for the directed evolution of enzymes"

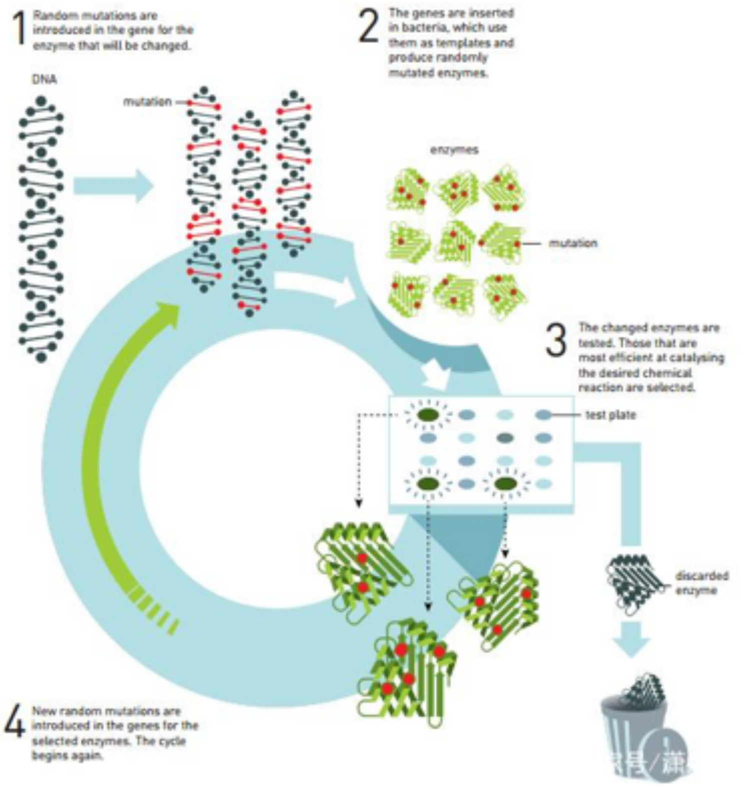
George P. Smith

Sir Gregory P. Winter

"for the phage display of peptides and antibodies"

THE ROYAL SWEDISH ACADEMY OF SCIENCES




1 Random mutations are introduced in the gene for the enzyme that will be changed.

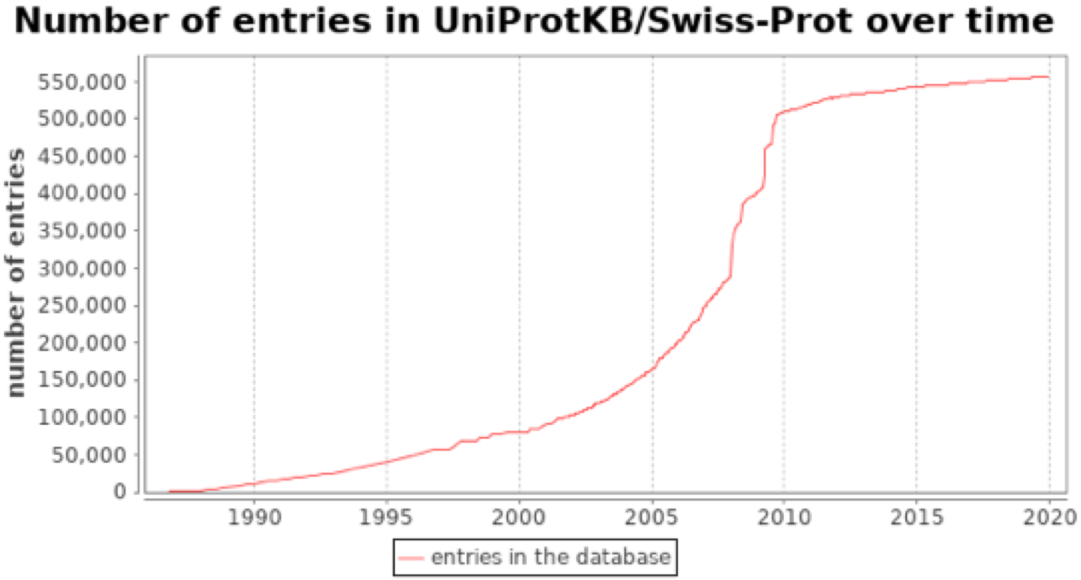
2 The genes are inserted in bacteria, which use them as templates and produce randomly mutated enzymes.

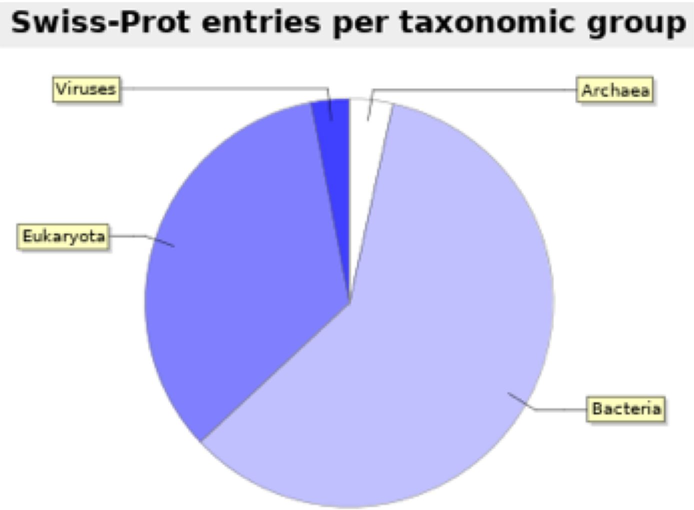
3 The changed enzymes are tested. Those that are most efficient at catalysing the desired chemical reaction are selected.

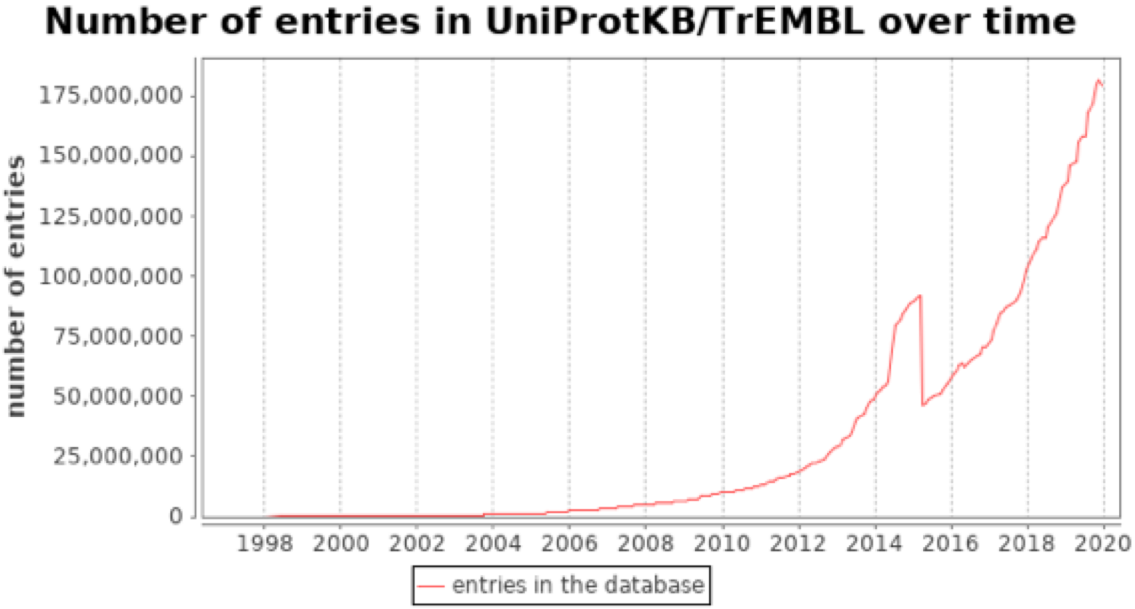
4 New random mutations are introduced in the genes for the selected enzymes. The cycle begins again.

通过定向进化制造的酶可用于生产各类产品，包括生物燃料、药品等等。

UniProt release statistics


Number of entries in UniProtKB/Swiss-Prot over time

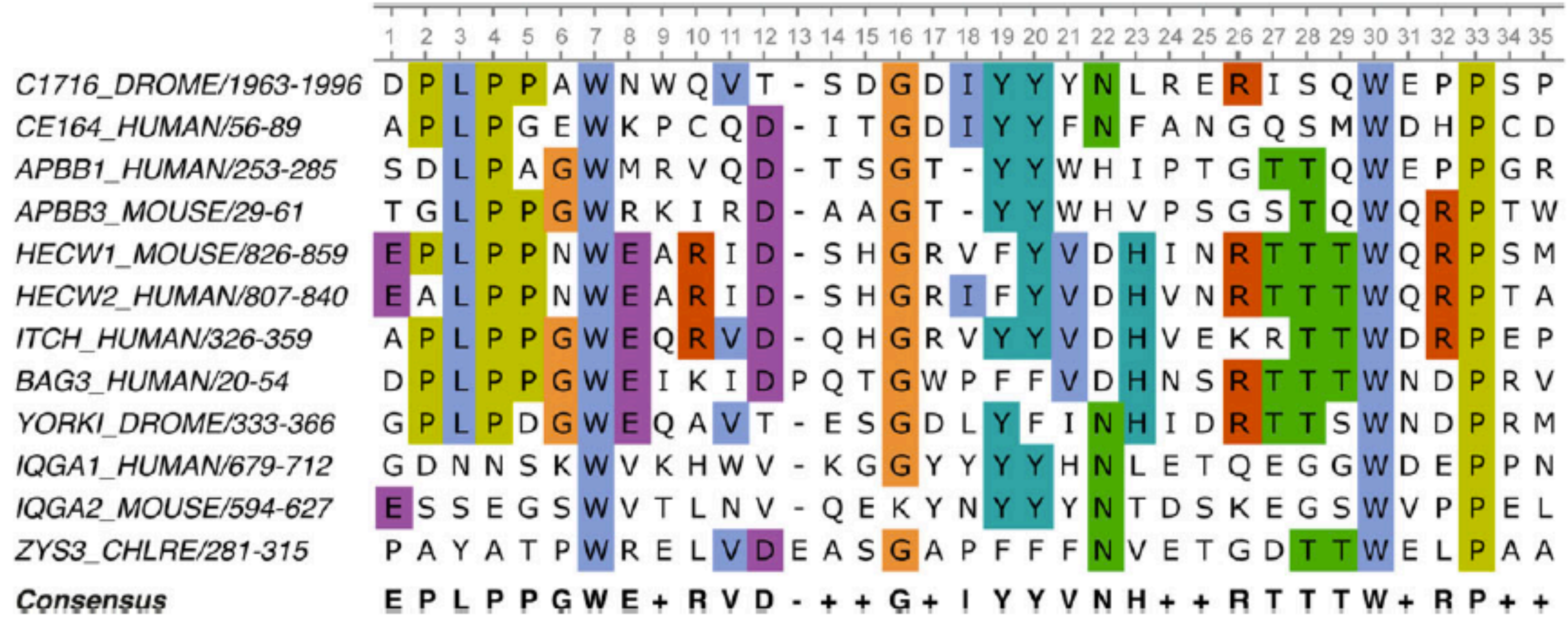

Swiss-Prot entries per taxonomic group


Number of entries in UniProtKB/TrEMBL over time

304,988,160 Entries in UniParc

December 11, 2019

Consensus protein design



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
C1716_DROME/1963-1996	D	P	L	P	P	A	W	N	W	Q	V	T	-	S	D	G	D	I	Y	Y	Y	N	L	R	E	R	I	S	Q	W	E	P	P	S	P
CE164_HUMAN/56-89	A	P	L	P	G	E	W	K	P	C	Q	D	-	I	T	G	D	I	Y	Y	F	N	F	A	N	G	Q	S	M	W	D	H	P	C	D
APBB1_HUMAN/253-285	S	D	L	P	A	G	W	M	R	V	Q	D	-	T	S	G	T	-	Y	Y	W	H	I	P	T	G	T	T	Q	W	E	P	P	G	R
APBB3_MOUSE/29-61	T	G	L	P	P	G	W	R	K	I	R	D	-	A	A	G	T	-	Y	Y	W	H	V	P	S	G	S	T	Q	W	Q	R	P	T	W
HECW1_MOUSE/826-859	E	P	L	P	P	N	W	E	A	R	I	D	-	S	H	G	R	V	F	Y	V	D	H	I	N	R	T	T	T	W	Q	R	P	S	M
HECW2_HUMAN/807-840	E	A	L	P	P	N	W	E	A	R	I	D	-	S	H	G	R	I	F	Y	V	D	H	V	N	R	T	T	T	W	Q	R	P	T	A
ITCH_HUMAN/326-359	A	P	L	P	P	G	W	E	Q	R	V	D	-	Q	H	G	R	V	Y	Y	V	D	H	V	E	K	R	T	T	W	D	R	P	E	P
BAG3_HUMAN/20-54	D	P	L	P	P	G	W	E	I	K	I	D	P	Q	T	G	W	P	F	F	V	D	H	N	S	R	T	T	T	W	N	D	P	R	V
YORKI_DROME/333-366	G	P	L	P	D	G	W	E	Q	A	V	T	-	E	S	G	D	L	Y	F	I	N	H	I	D	R	T	T	S	W	N	D	P	R	M
IQGA1_HUMAN/679-712	G	D	N	N	S	K	W	V	K	H	W	V	-	K	G	G	Y	Y	Y	Y	H	N	L	E	T	Q	E	G	G	W	D	E	P	P	N
IQGA2_MOUSE/594-627	E	S	S	E	G	S	W	V	T	L	N	V	-	Q	E	K	Y	N	Y	Y	Y	N	T	D	S	K	E	G	S	W	V	P	P	E	L
ZYS3_CHLRE/281-315	P	A	Y	A	T	P	W	R	E	L	V	D	E	A	S	G	A	P	F	F	F	N	V	E	T	G	D	T	T	W	E	L	P	A	A
Consensus	**E**	**P**	**L**	**P**	**P**	**G**	**W**	**E**	**+**	**R**	**V**	**D**	**-**	**+**	**+**	**G**	**+**	**I**	**Y**	**Y**	**V**	**N**	**H**	**+**	**+**	**R**	**T**	**T**	**T**	**W**	**+**	**R**	**P**	**+**	**+**

Fig. 1 Sequence alignment of 12 WW domains across several species and parent proteins. In the consensus, a '–' is a gap, whilst a '+' is an ambiguous position with no consensus. The most conserved residues are highlighted.

The consensus design approach has been widely successful in improving the stabilities of functional and non-functional proteins, for example increasing melting temperatures by 10–32°C.

Porebski, et al (2016) *Protein Eng Des Sel*, **29**, 245-251.

PSAP

position-specific amino acid probabilities

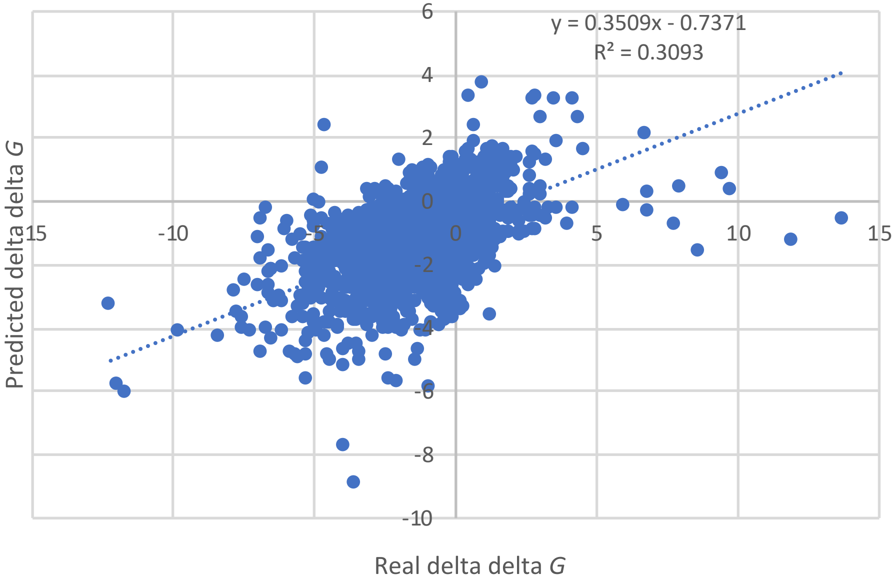
1. Retrieve sequences
2. Sequence alignment
3. Weight sequences
4. Calculate the PSAP based on Dirichlet mixture.
5. Calculate the position entropy

Tian et al. BMC Bioinformatics. 2007

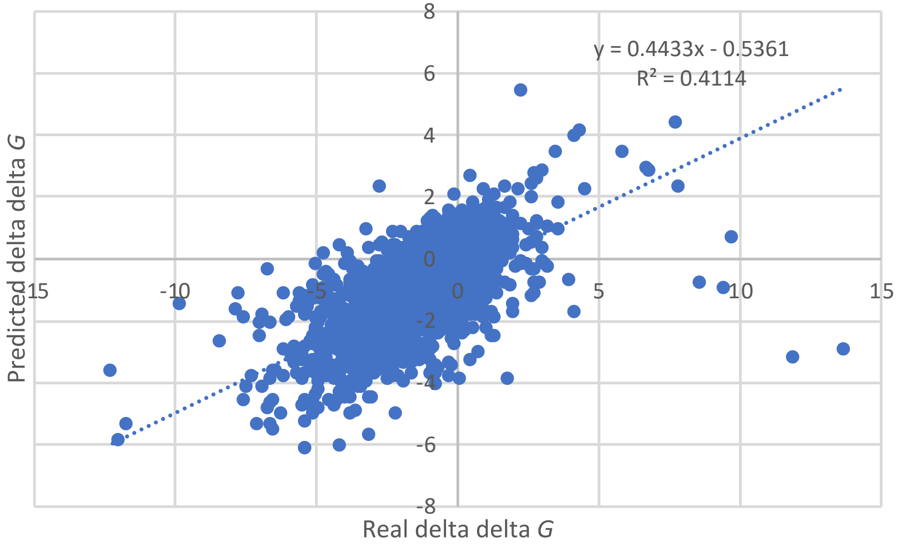
No	AA	Entropy	Wt	Max	Max_aa	Diff	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Num
194	L	0.009	0.996	0.996	L	0	0	0	1E-04	0	0	2E-04	0	0.003	0	0.996	0	0	1E-04	0	0	0	0	0	0	0	4368
1	M	0.928	0.122	0.156	H	0.033	0.052	0.01	0.065	0.039	0.022	0.039	0.156	0.036	0.041	0.064	0.122	0.026	0.034	0.029	0.063	0.044	0.069	0.065	0.008	0.019	4368
2	K	0.912	0.093	0.196	A	0.103	0.196	0.009	0.026	0.036	0.021	0.069	0.018	0.043	0.093	0.072	0.057	0.027	0.032	0.047	0.073	0.062	0.04	0.055	0.007	0.017	4368
3	F	0.717	0.087	0.37	M	0.283	0.031	0.03	0.004	0.004	0.087	0.019	0.006	0.026	0.016	0.208	0.37	0.007	0.014	0.003	0.044	0.026	0.022	0.029	0.043	0.012	4368
4	T	0.608	0.068	0.5	K	0.432	0.029	0.002	0.006	0.011	0.01	0.027	0.008	0.01	0.5	0.028	0.002	0.011	0.033	0.014	0.174	0.052	0.068	0.012	0.001	0.003	4368
5	T	0.738	0.072	0.262	F	0.19	0.094	0.01	0.008	0.004	0.262	0.014	0.009	0.032	0.007	0.106	0.011	0.023	0.013	0.003	0.041	0.027	0.072	0.248	0.003	0.012	4368
6	T	0.666	0.263	0.329	S	0.066	0.083	0.002	0.005	0.005	0.012	0.012	0.002	0.022	0.007	0.083	0.002	0.049	0.046	0.006	0.007	0.329	0.263	0.06	8E-04	0.005	4368
7	I	0.714	0.144	0.241	A	0.097	0.241	0.007	0.003	0.004	0.025	0.021	0.002	0.144	0.004	0.148	0.005	0.003	0.047	0.003	0.007	0.18	0.033	0.116	0.002	0.005	4368
8	T	0.694	0.239	0.261	S	0.022	0.171	0.002	0.004	0.005	0.081	0.021	0.002	0.012	0.004	0.064	0.008	0.005	0.011	0.03	0.014	0.261	0.239	0.063	9E-04	0.003	4368
9	G	0.734	0.046	0.238	V	0.192	0.16	0.004	0.004	0.004	0.044	0.046	0.003	0.099	0.004	0.14	0.158	0.004	0.008	0.006	0.004	0.027	0.036	0.238	0.003	0.009	4368
10	L	0.625	0.421	0.421	L	0	0.13	0.004	0.004	0.004	0.076	0.011	0.003	0.024	0.004	0.421	0.006	0.004	0.008	0.003	0.044	0.023	0.031	0.191	0.003	0.007	4368
11	L	0.649	0.433	0.433	L	0	0.079	0.004	0.004	0.004	0.027	0.011	0.003	0.051	0.004	0.433	0.031	0.004	0.008	0.003	0.044	0.165	0.028	0.089	0.003	0.007	4368
12	V	0.758	0.07	0.217	S	0.146	0.17	0.011	0.003	0.004	0.013	0.019	0.002	0.033	0.004	0.114	0.163	0.003	0.083	0.043	0.004	0.217	0.038	0.07	0.002	0.005	4368
13	G	0.665	0.093	0.371	A	0.278	0.371	0.01	0.029	0.004	0.003	0.093	0.005	0.026	0.004	0.203	0.003	0.003	0.049	0.003	0.003	0.086	0.036	0.067	9E-04	0.003	4368
14	T	0.585	0.073	0.538	A	0.465	0.538	0.007	0.003	0.004	0.01	0.041	0.002	0.012	0.004	0.09	0.046	0.003	0.021	0.006	0.003	0.03	0.073	0.097	0.006	0.005	4368
15	A	0.634	0.418	0.418	A	0	0.418	0.007	0.003	0.008	0.006	0.072	0.002	0.031	0.004	0.224	0.003	0.006	0.033	0.003	0.044	0.052	0.048	0.034	9E-04	0.003	4368
16	G	0.667	0.082	0.293	A	0.211	0.293	0.007	0.003	0.005	0.004	0.082	0.002	0.065	0.031	0.122	0.011	0.003	0.004	0.003	0.004	0.032	0.056	0.27	0.001	0.003	4368
17	V	0.629	0.3	0.325	L	0.025	0.141	0.014	0.004	0.004	0.037	0.013	0.003	0.023	0.004	0.325	0.006	0.004	0.006	0.003	0.004	0.024	0.071	0.3	0.007	0.007	4368
18	F	0.89	0.063	0.241	A	0.177	0.241	0.011	0.031	0.043	0.063	0.04	0.016	0.063	0.039	0.119	0.017	0.026	0.029	0.043	0.047	0.045	0.045	0.048	0.014	0.021	4368
19	A	0.829	0.349	0.349	A	0	0.349	0.011	0.031	0.042	0.024	0.042	0.016	0.041	0.039	0.08	0.017	0.026	0.029	0.028	0.036	0.051	0.056	0.054	0.009	0.021	4368
20	A	0.884	0.127	0.219	P	0.092	0.127	0.011	0.031	0.042	0.024	0.04	0.015	0.042	0.039	0.101	0.019	0.026	0.219	0.027	0.036	0.082	0.039	0.054	0.008	0.02	4368

模拟准确率

Predicted results based on the PSAPs



Predicted results of Prethermut



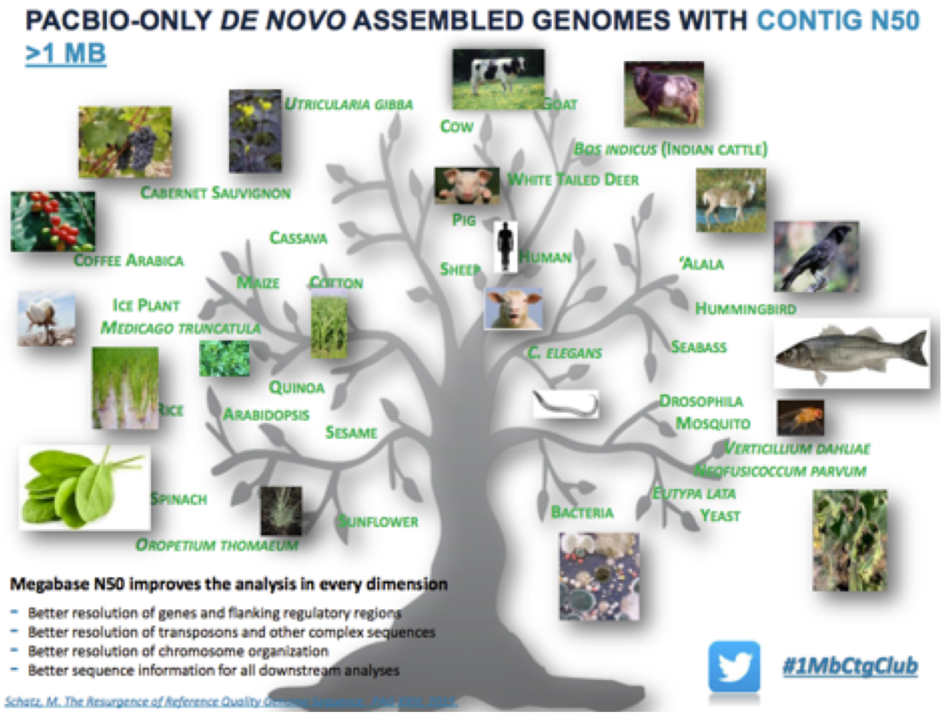
2633个单点突变体(突变是否会影响蛋白质的稳定性)，来自ProTherm

利用随机森林机器学习算法

10-folds 交叉验证

在基因的进化数据中包含量了众多的信息，可用来模拟基因的性质，并对蛋白质进行设计。

机遇与挑战



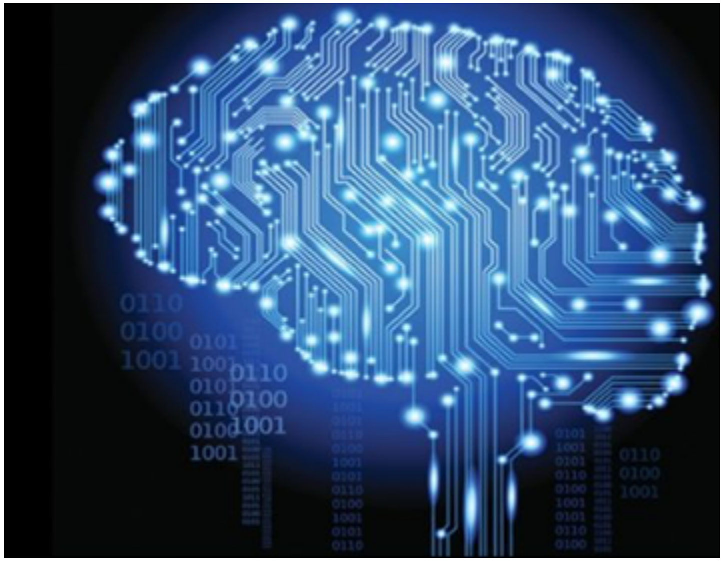
PACBIO-ONLY *DE NOVO* ASSEMBLED GENOMES WITH CONTIG N50 >1 MB

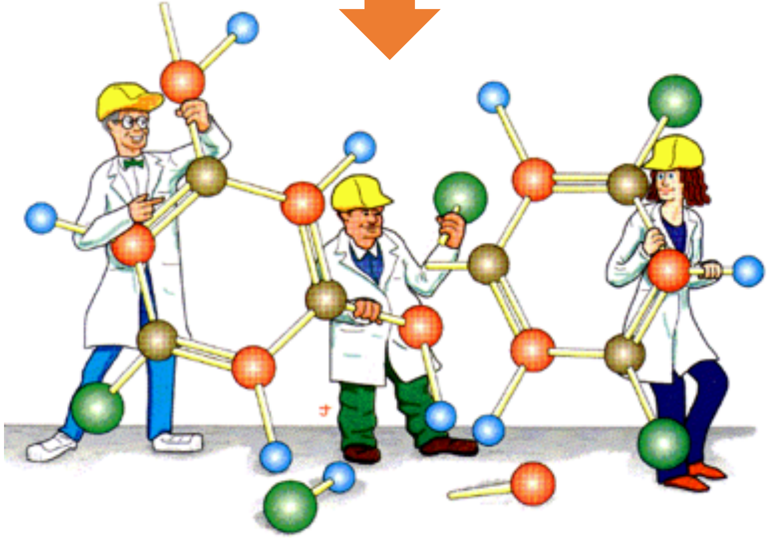
Utricularia gibba
Cow
Goat
Bos indicus (Indian cattle)
Cabernet Sauvignon
White Tailed Deer
Pig
Cassava
Coffee Arabica
Human
'Alala
Sheep
Maize
Cotton
Ice Plant
Medicago truncatula
Hummingbird
C. elegans
Seabass
Quinoa
Rice
Arabidopsis
Drosophila
Mosquito
Sesame
Verticillium dahliae
Neofusicoccum parvum
Spinach
Eutypa lata
Yeast
Bacteria
Sunflower
Oropetium thomaeum
#1MbCtgClub

Megabase N50 improves the analysis in every dimension

- Better resolution of genes and flanking regulatory regions
- Better resolution of transposons and other complex sequences
- Better resolution of chromosome organization
- Better sequence information for all downstream analyses

Schatz, M. The Resurgence of Reference Quality Genome Sequence. PAG XXIII, 2015.

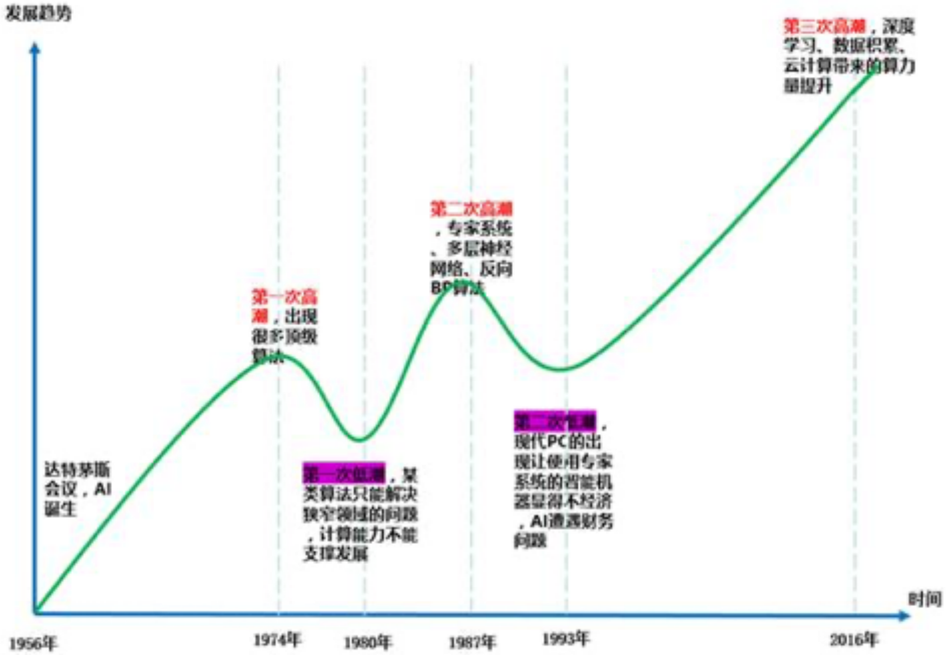


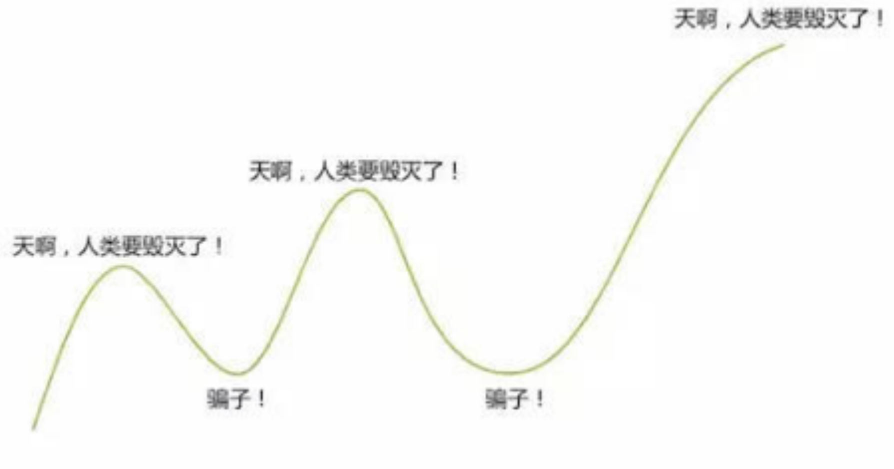


而近年来随着基因组学、合成生物学等学科的进步，给人类提供了海量的基因数据。而人工智能技术又给人类带来了高通量的模型构建的机遇。

人工智能发展的三起两落

2017年
AlphaGo



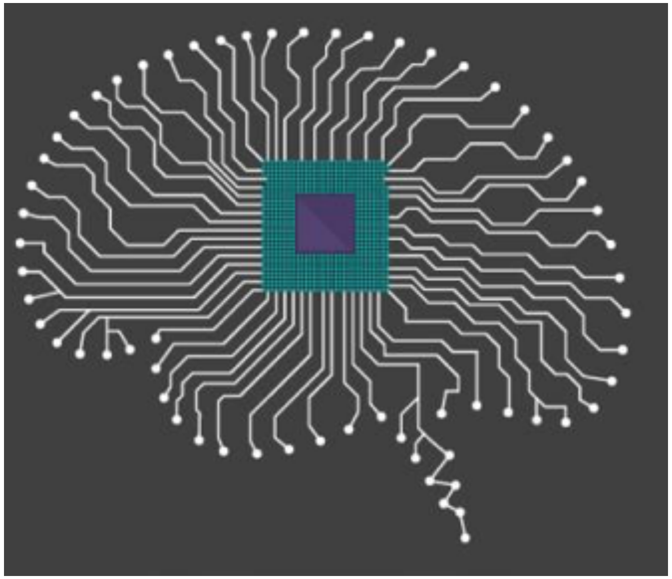


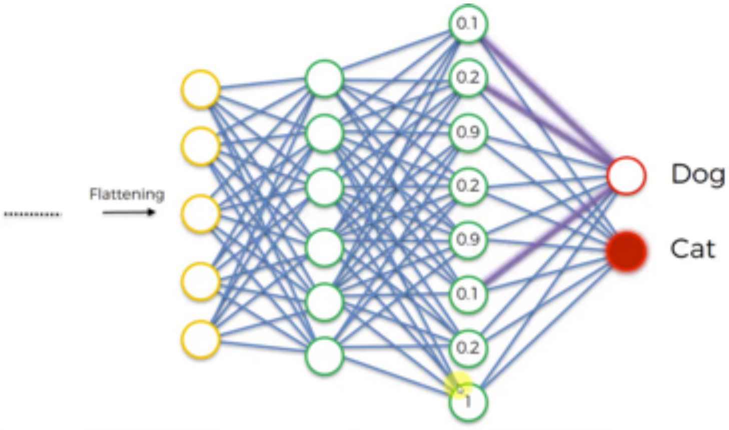
网上流传的漫画：人工智能发展成熟度曲线

人工智能的历史其实正好与计算机的历史差不多一样长，但两者的发展进度却大相径庭。一个很像一帆风顺的富二代，一个则起起落落很像白手起家的创业者。

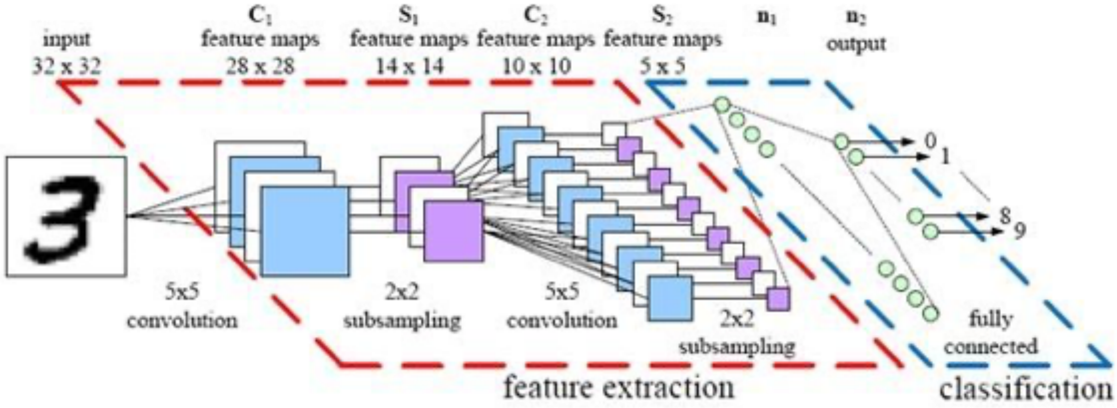
http://www.qianjia.com/html/2018-06/14_295433.html

神经网络





图像识别



数字识别

从“深蓝”到AlphaGo

1997年，美国IBM公司的“深蓝”超级计算机以2胜1负3平战胜了当时世界排名第一的国际象棋大师卡斯帕罗夫。

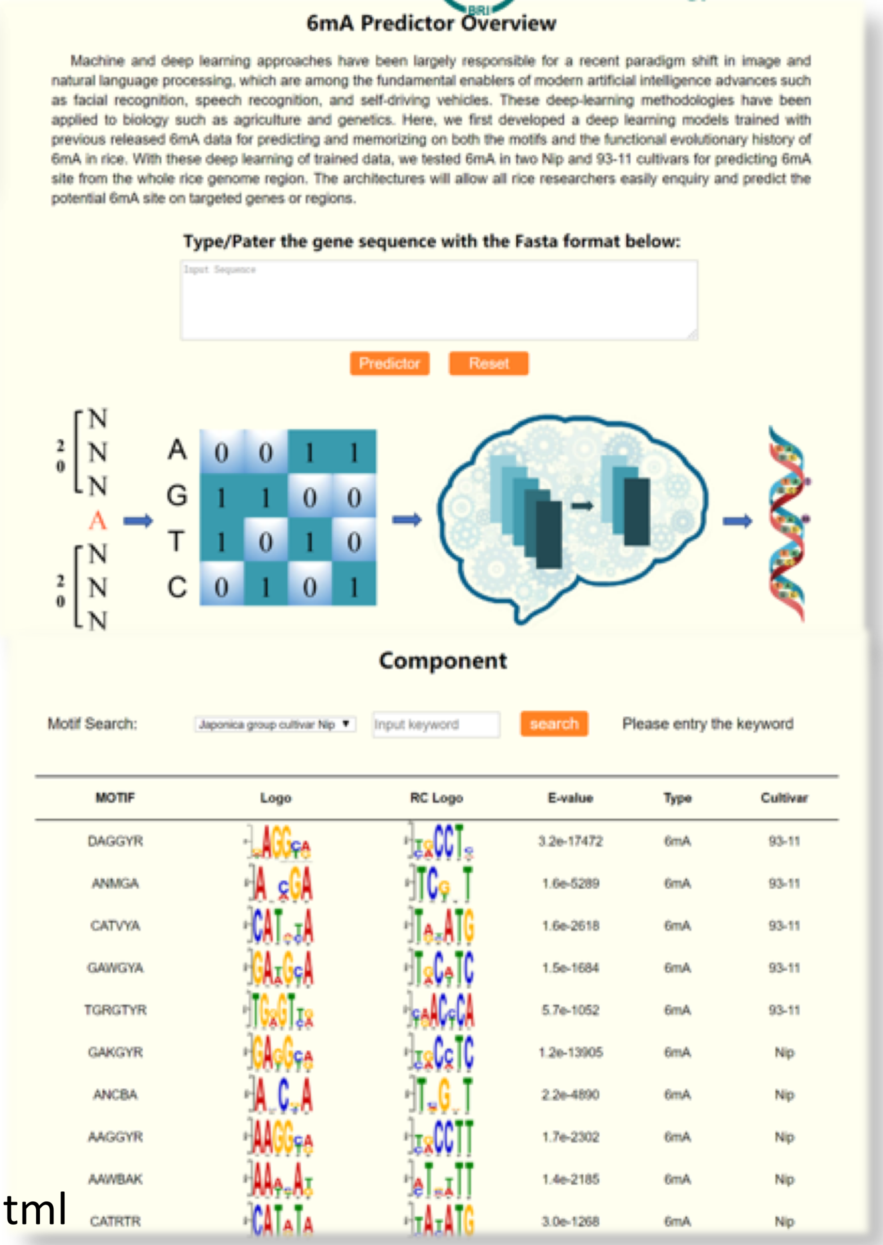
“深蓝” 每秒可运算2亿步，依靠强大的计算能力穷举所有路数来选择最佳策略。

“深蓝”靠硬算可以预判12步，卡斯帕罗夫可以预判10步，两者高下立现。

围棋的可能下法数量超越了可观测宇宙范围内的原子总数，显然“深蓝”式的硬算在围棋上行不通。

“阿尔法围棋” 通过大量数据分析学习了3000万步的职业棋手棋谱，通过策略网络和价值网络来决定棋路，不计算每一步的可能性，颇有人类棋手**“我感觉这样会赢”**的味道。

6mA AI predictor



6mA Predictor Overview

Machine and deep learning approaches have been largely responsible for a recent paradigm shift in image and natural language processing, which are among the fundamental enablers of modern artificial intelligence advances such as facial recognition, speech recognition, and self-driving vehicles. These deep-learning methodologies have been applied to biology such as agriculture and genetics. Here, we first developed a deep learning models trained with previous released 6mA data for predicting and memorizing on both the motifs and the functional evolutionary history of 6mA in rice. With these deep learning of trained data, we tested 6mA in two Nip and 93-11 cultivars for predicting 6mA site from the whole rice genome region. The architectures will allow all rice researchers easily enquiry and predict the potential 6mA site on targeted genes or regions.

Type/Pater the gene sequence with the Fasta format below:

Input Sequence

Predictor Reset

A	0	0	1	1
G	1	1	0	0
T	1	0	1	0
C	0	1	0	1

Component

Motif Search: Japonica group cultivar Nip ▼ Input keyword search Please entry the keyword

MOTIF	Logo	RC Logo	E-value	Type	Cultivar
DAGGYR			3.2e-17472	6mA	93-11
ANMGA			1.6e-5289	6mA	93-11
CATVYA			1.6e-2618	6mA	93-11
GAWGYA			1.5e-1684	6mA	93-11
TGRGTYR			5.7e-1052	6mA	93-11
GAKGYR			1.2e-13905	6mA	Nip
ANCBA			2.2e-4890	6mA	Nip
AAGGYR			1.7e-2302	6mA	Nip
AAWBAK			1.4e-2185	6mA	Nip
CATRTR			3.0e-1268	6mA	Nip

http://www.elabcaas.cn/rice/6mA_predictor.html

Gene Design – synonymous codon selection

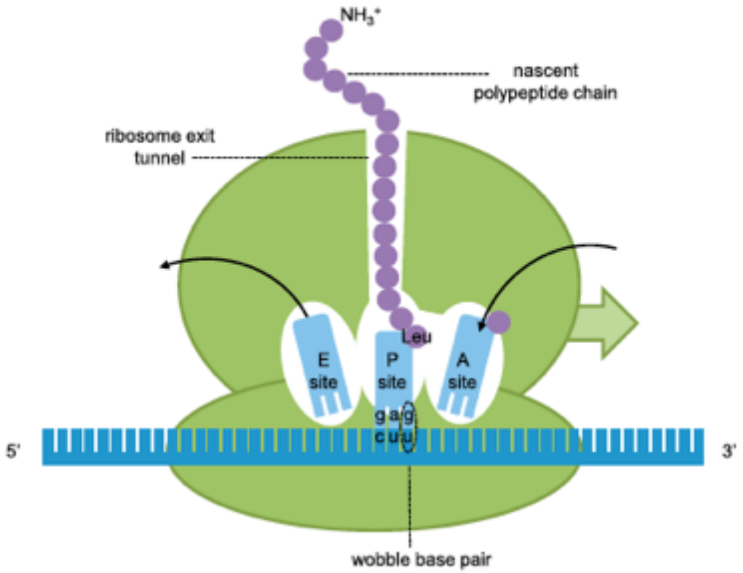
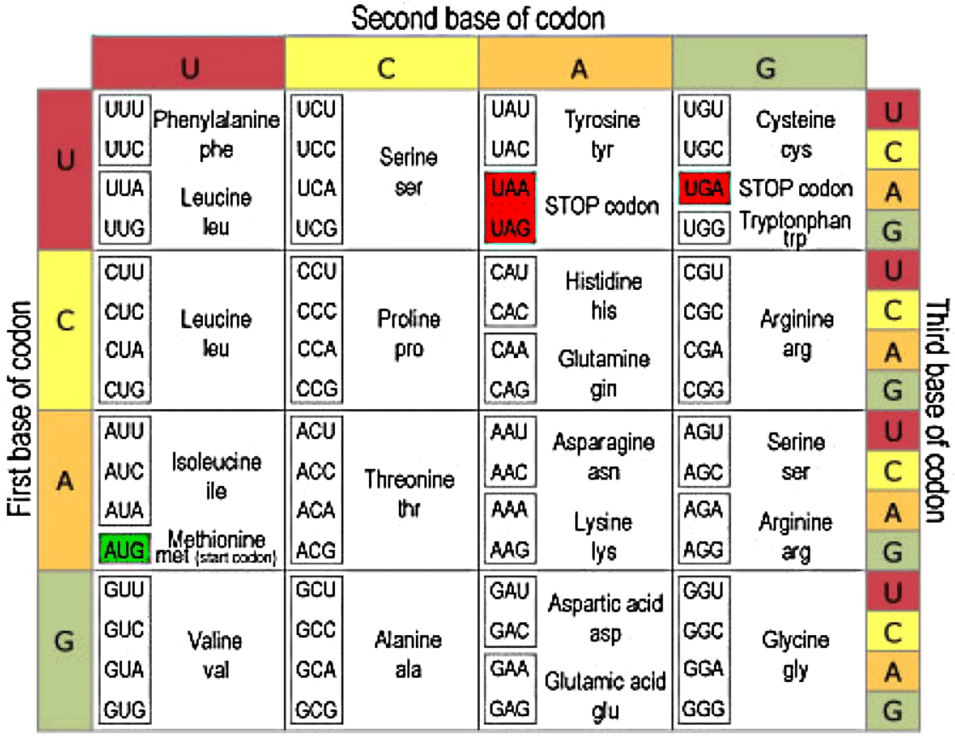
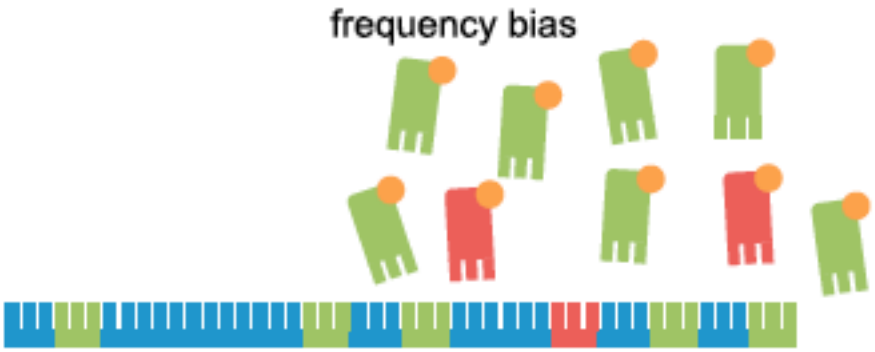


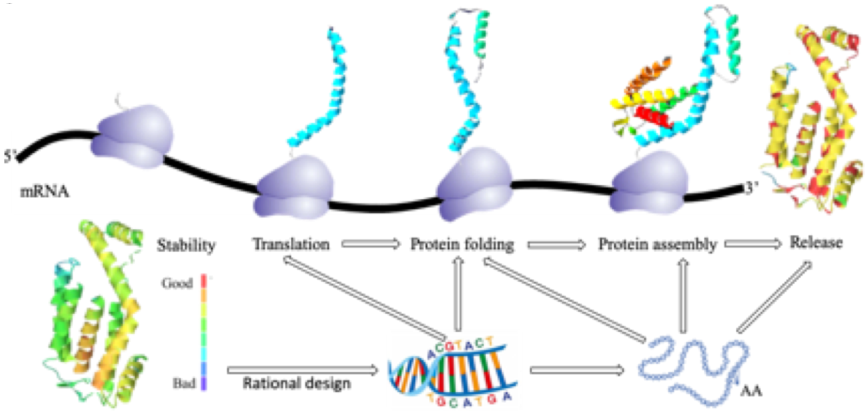
Figure 1. Translation in the Ribosome and tRNA Structure
Cartoon of the ribosome (green) during translation of a mRNA (blue) with a wobbling codon-anticodon base pair encoding a leucine amino acid. A site, aminoacyl-tRNA site; E site, exit site; P site, peptidyl-tRNA site.



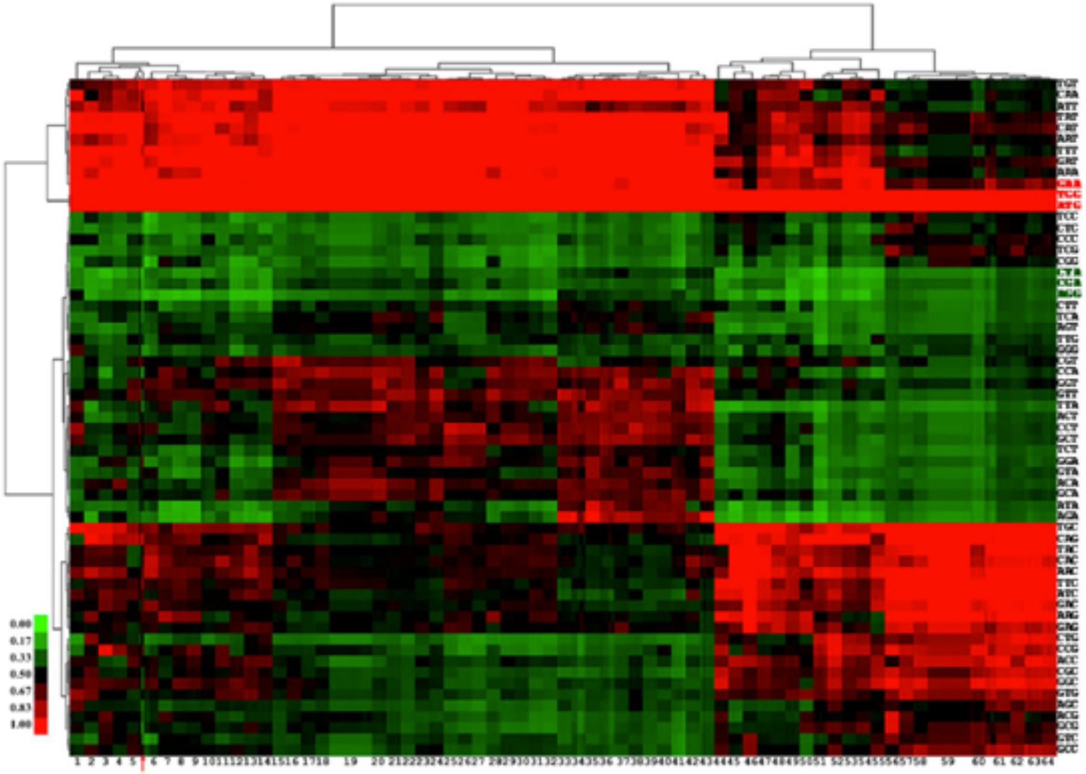
Second base of codon

First base of codon	U		C		A		G		Third base of codon
U	UUU	Phenylalanine phe	UCU	Serine ser	UAU	Tyrosine tyr	UGU	Cysteine cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leucine leu	UCA		UAA	STOP codon	UGA	STOP codon	A
	UUG		UCG		UAG		UGG	Tryptophan trp	G
C	CUU	Leucine leu	CCU	Proline pro	CAU	Histidine his	CGU	Arginine arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Glutamine gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Isoleucine ile	ACU	Threonine thr	AAU	Asparagine asn	AGU	Serine ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lysine lys	AGA	Arginine arg	A
	AUG	Methionine met (start codon)	ACG		AAG		AGG		G
G	GUU	Valine val	GCU	Alanine ala	GAU	Asparic acid asp	GGU	Glycine gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glutamic acid glu	GGA		A
	GUG		GCG		GAG		GGG		G





Gene optimizing methods

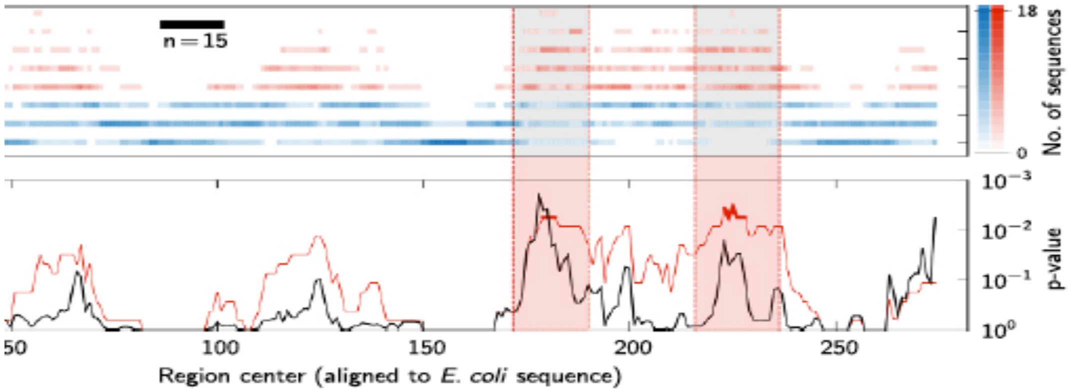


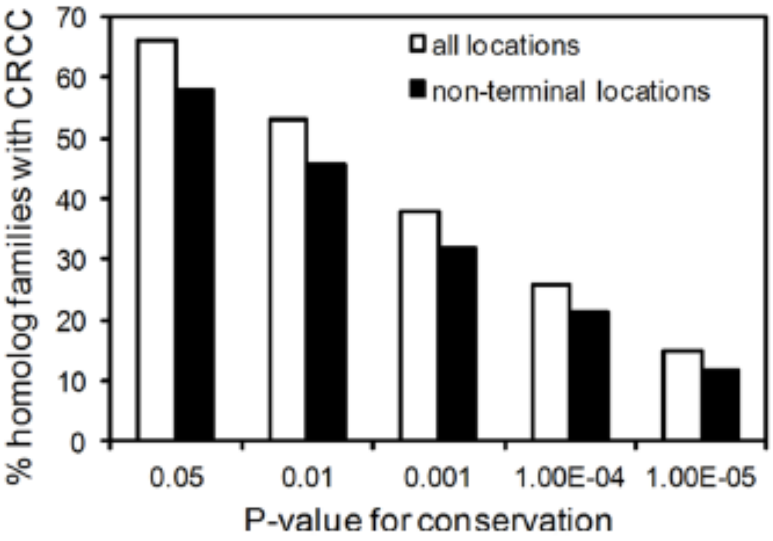
Clustering results of the codon usage pattern of different species.

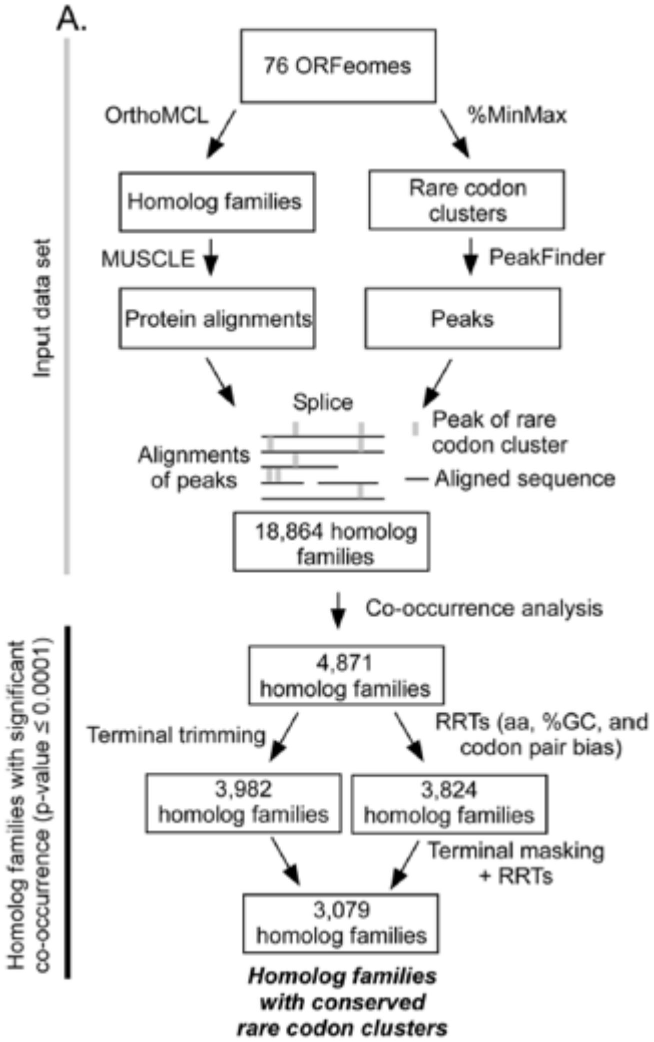
The row and column represent the codon usage pattern and the different bacterial subspecies. The species between species id 5 and 6 (red arrow) is Bacteroides fragilis NCTC 9343. The numbers from 1 to 64 refer to the bacterial genera.

Many methods, including COOL, Gene Designer, Gene composer, JCat, COStar and OPTIMIZER, have been proposed to design heterologous genes which are expected to be efficiently expressed in the host organism. Based on our knowledge, these methods are prone to select the high-frequency-usage codons.

Conservation of the rare codons in the evolution







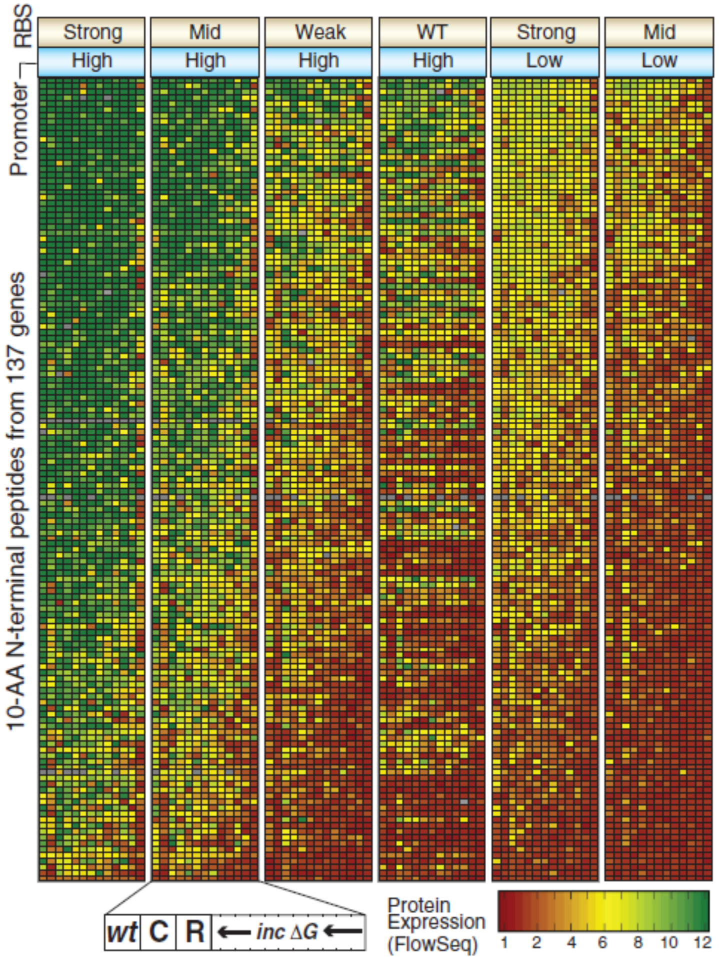
重要稀有密码子在进化上保守。

Jacobs, W.M. et al (2017). *Proc Natl Acad Sci U S A*, **114**, 11434-11439.
Chaney, J.L.et al . (2017). *PLoS Comput Biol*, **13**, e1005531.

Causes and Effects of N-Terminal Codon Bias in Bacterial Genes

Daniel B. Goodman *et al.*
Science **342**, 475 (2013);
DOI: 10.1126/science.1241934

Most amino acids are encoded by multiple codons, and codon choice has strong effects on protein expression. **Rare codons are enriched at the N terminus of genes in most organisms**, although the causes and effects of this bias are unclear. Here, we measure expression from >14,000 synthetic reporters in Escherichia coli and show that using **N-terminal rare codons instead of common ones increases expression by ~14-fold (median 4-fold).**



N端的稀有密码子提高了*gfp*基因的高表达。

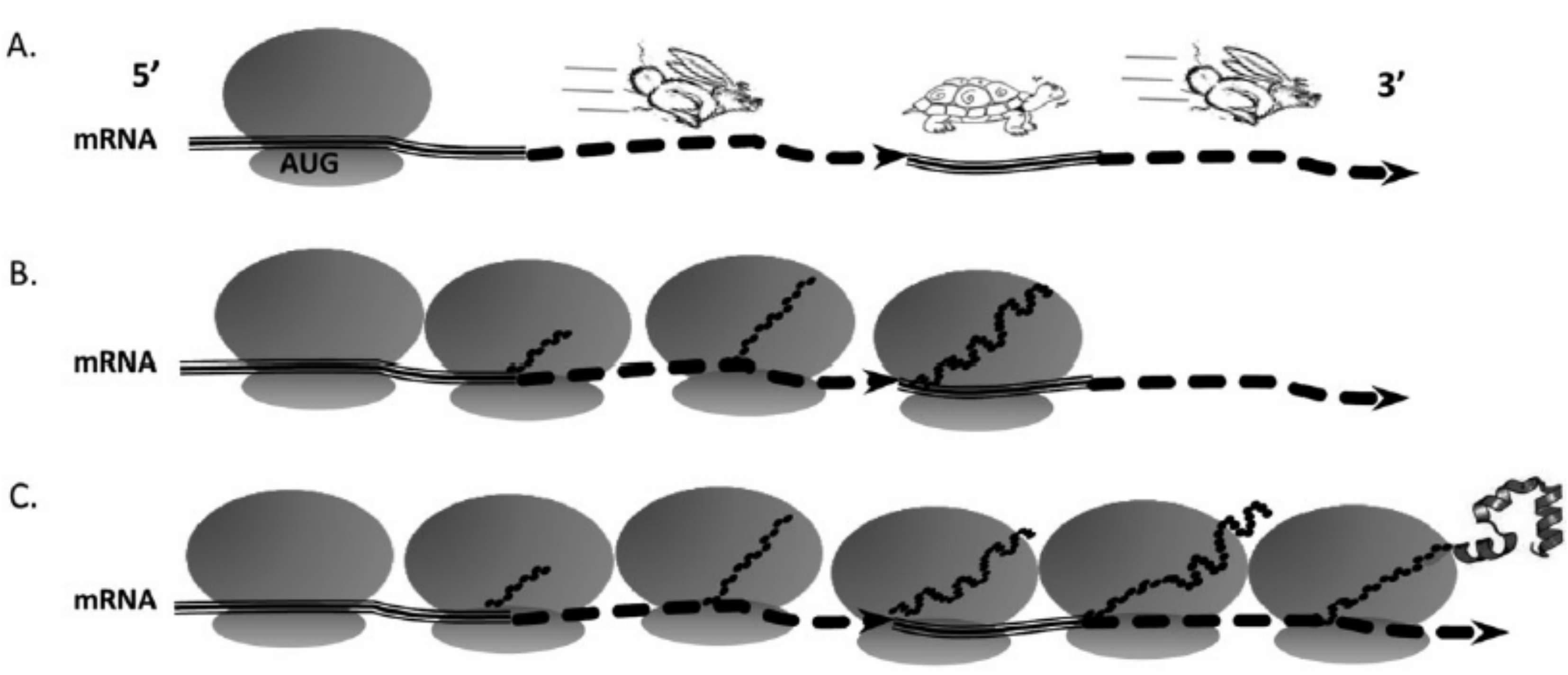
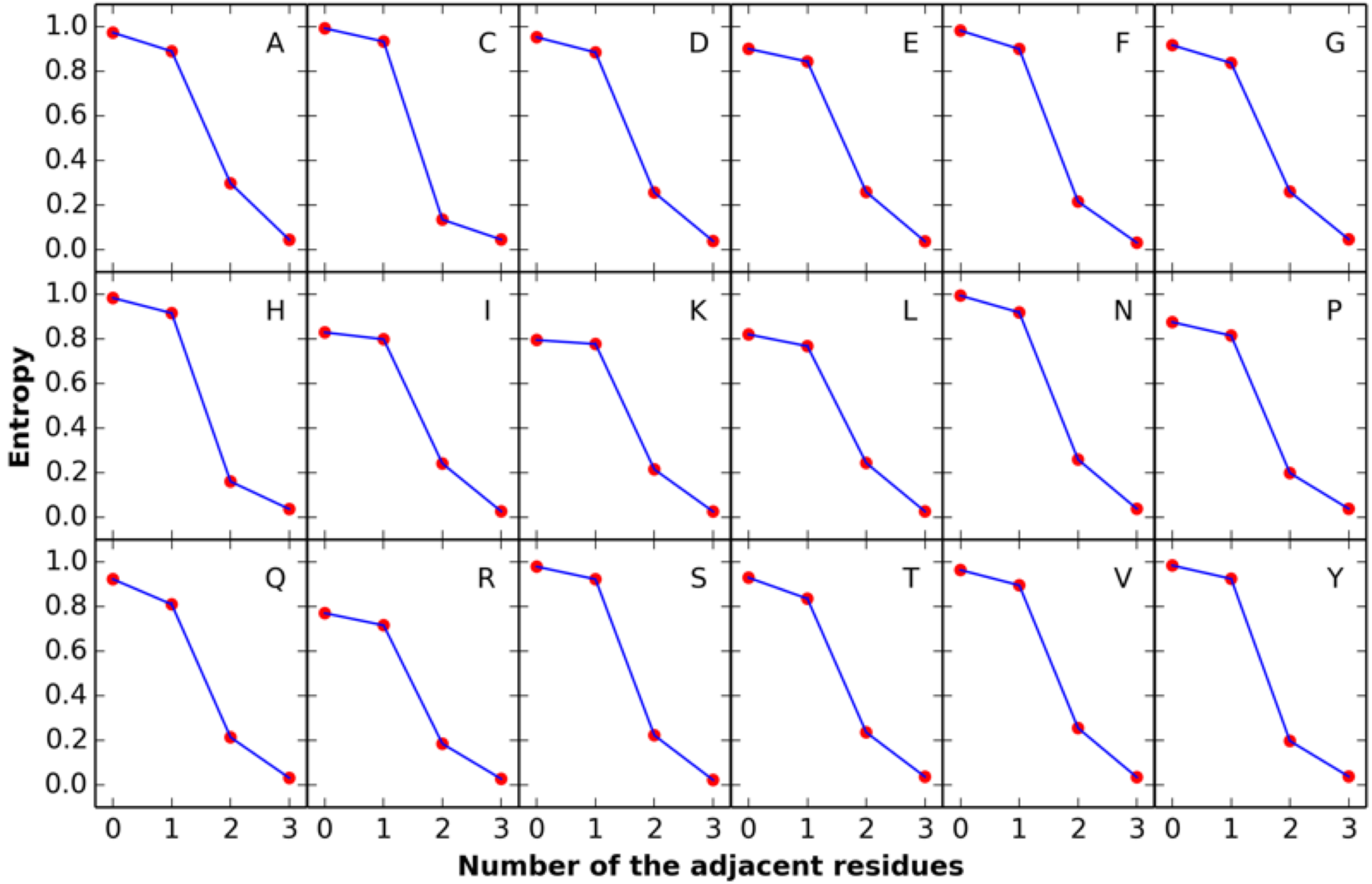


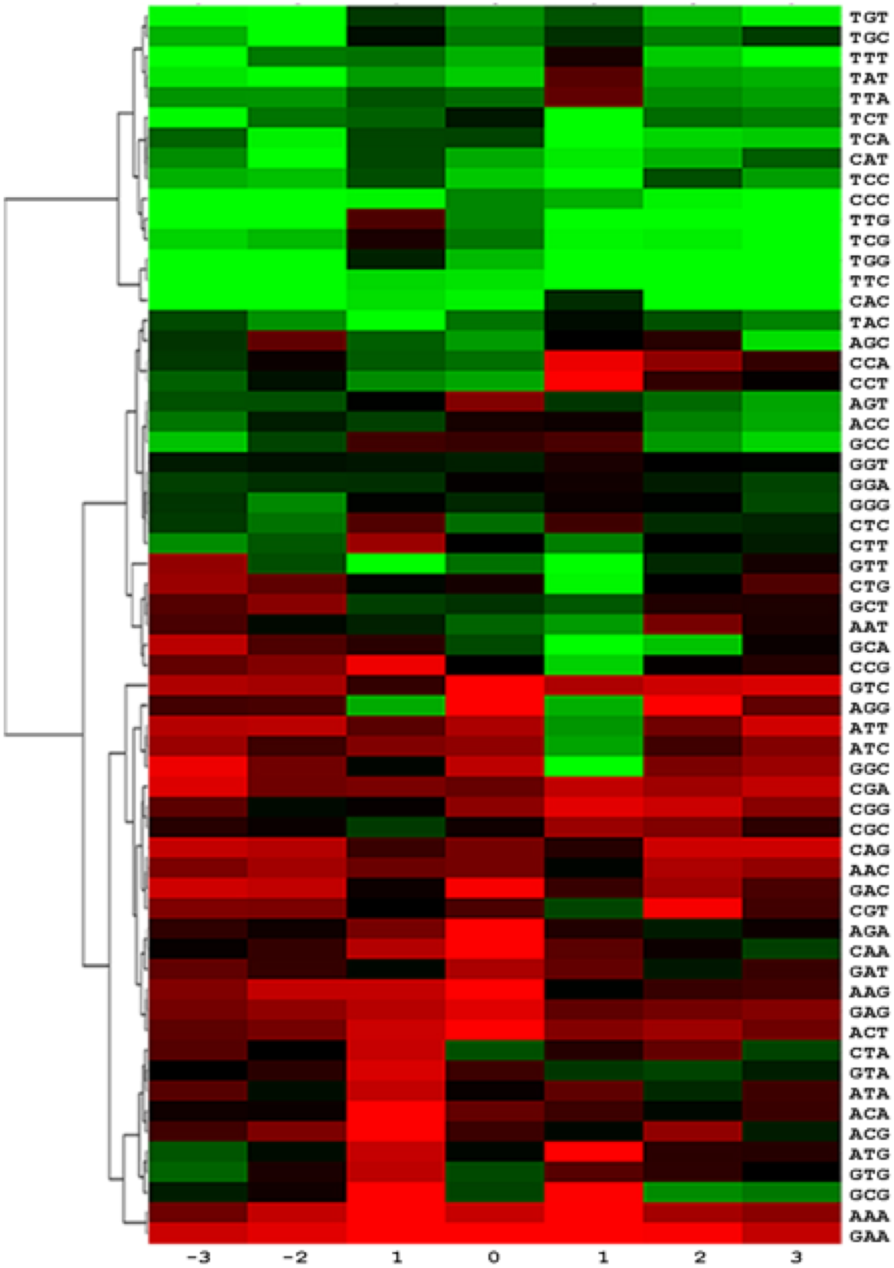
Figure 3. Schematic model of co-translational folding on mRNA by ribosomes. **(A)** Ribosomal complex centered on the translation initiation site, AUG (initiation codon). **(B)** Nascent polypeptide synthesis within the protective environment of the ribosomal tunnel. **(C)** Putative translational pause sites in conjunction with co-translational folding occur within the ribosomal tunnel. Differences in codon usage frequency are shown as thick dashed lines with arrowheads for areas representing high-frequency-usage codons, and therefore, translating rapidly (hare) and regions that are double lined represent segments of lower frequency usage codons (i.e., putative pause sites; tortoise) where translation proceeds more slowly to allow nascent polypeptide folding.

O'Brien, E. P., P. Ciryam, M. Vendruscolo, and C. M. Dobson. "Understanding the Influence of Codon Translation Rates on Cotranslational Protein Folding." ***Acc Chem Res*** 47, no. 5 (May 20 2014): 1536-44.

The codon usage entropy of the middle amino acid with different neighbors in *E.coli*.



基因中不同氨基酸密码子的选择相互关联、相互影响，形成了适应于自身的密码子模型。

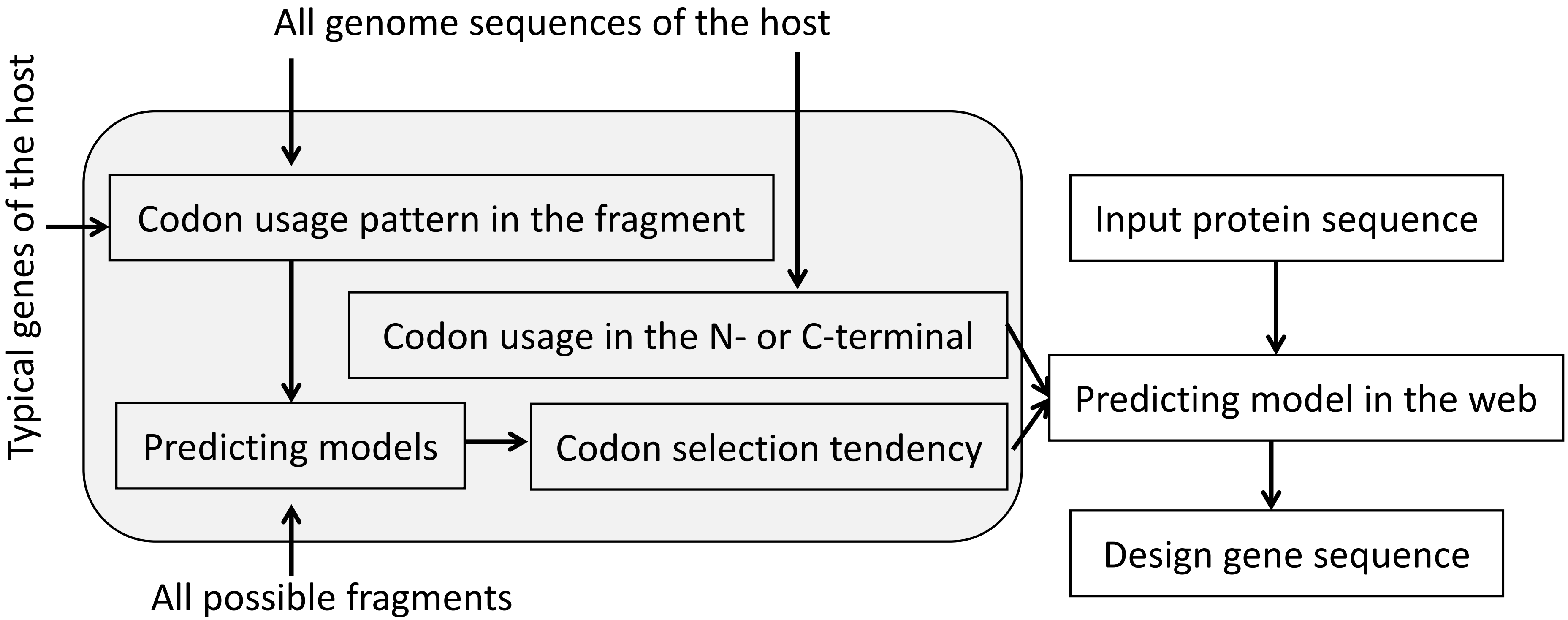


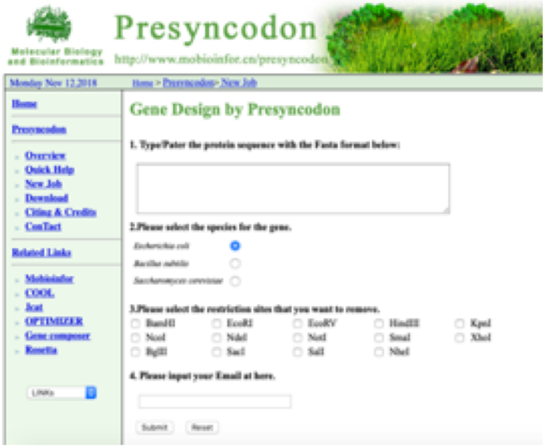
The effect of the neighbor codons on the target codon selection in *E. coli*.

All of the genes in E. coli were divided into seven-codon windows. All of the fragments with seven-codons were clustered into the 61 groups based on the middle codon, which were represented in the rows. In each group, every codon was represented as the codon usage value in E. coli.

物以类聚，人以群分。

密码子优化软件Presyncodon


All genome sequences of the host
Typical genes of the host
Codon usage pattern in the fragment
Codon usage in the N- or C-terminal
Predicting models
Codon selection tendency
All possible fragments
Input protein sequence
Predicting model in the web
Design gene sequence



Presyncodon

http://www.mobioinfor.cn/presyncodon

Monday Nov 12,2018

Home > Presyncodon > New Job

- Home
- Presyncodon
  - Overview
  - Quick Help
  - New Job
  - Download
  - Citing & Credits
  - ConTact
- Related Links
  - Mobioinfor
  - COOL
  - Jcat
  - OPTIMIZER
  - Gene composer
  - Rosetta

Gene Design by Presyncodon

1. Type/Pater the protein sequence with the Fasta format below:

2.Please select the species for the gene.

Escherichia coli ●
Bacillus subtilis ○
Saccharomyces cerevisiae ○

3.Please select the restriction sites that you want to remove.

☐ BamHI	☐ EcoRI	☐ EcoRV	☐ HindIII	☐ KpnI
☐ NcoI	☐ NdeI	☐ NotI	☐ SmaI	☐ XhoI
☐ BglII	☐ SacI	☐ SalI	☐ NheI	

4. Please input your Email at here.

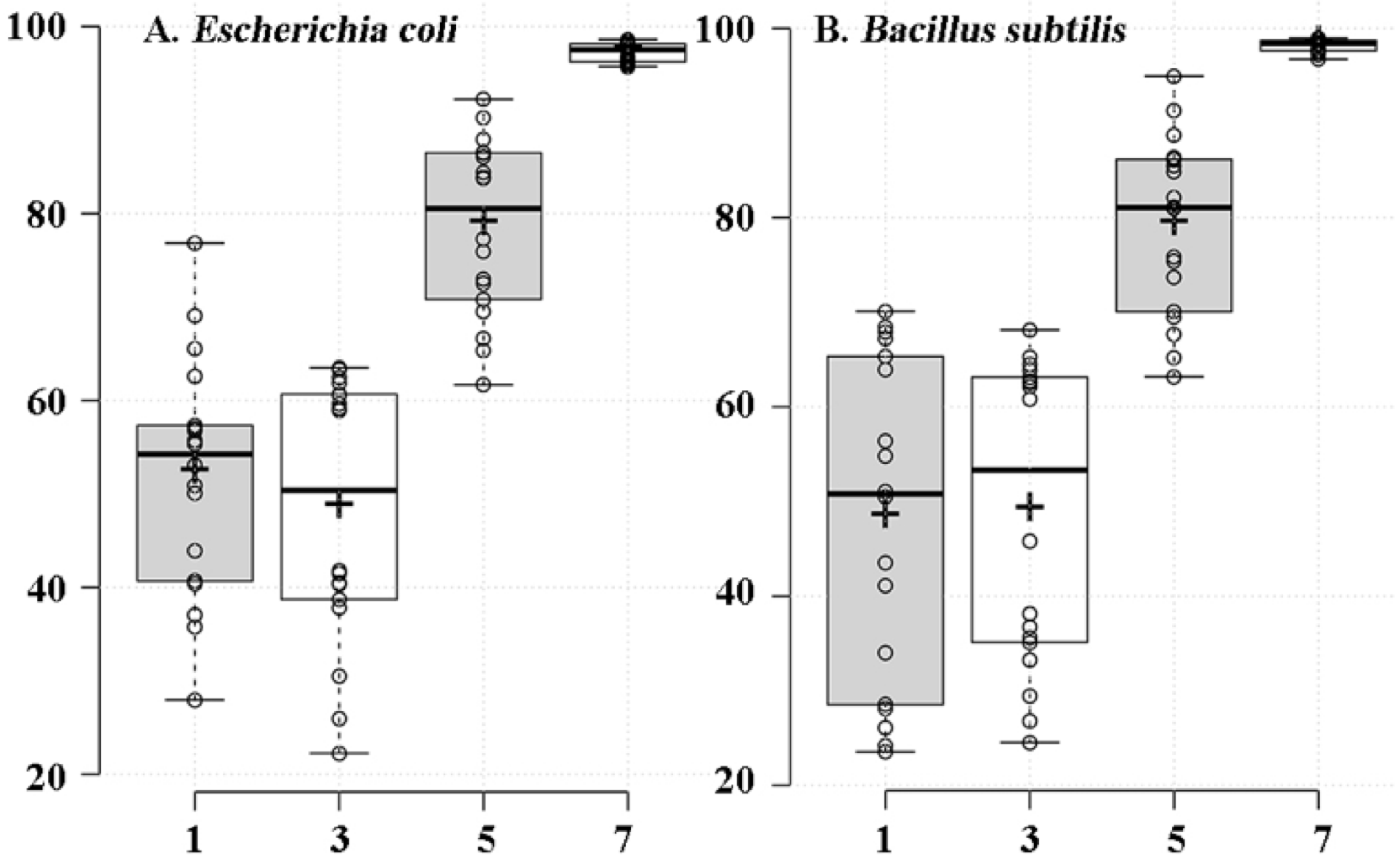
Submit Reset

目前可以做宿主*Escherichia coli*，*Bacillus subtilis*和*Saccharomyces cerevisiae*中的基因设计，截止到2020年1月12日，**已完成了449条基因的设计**。

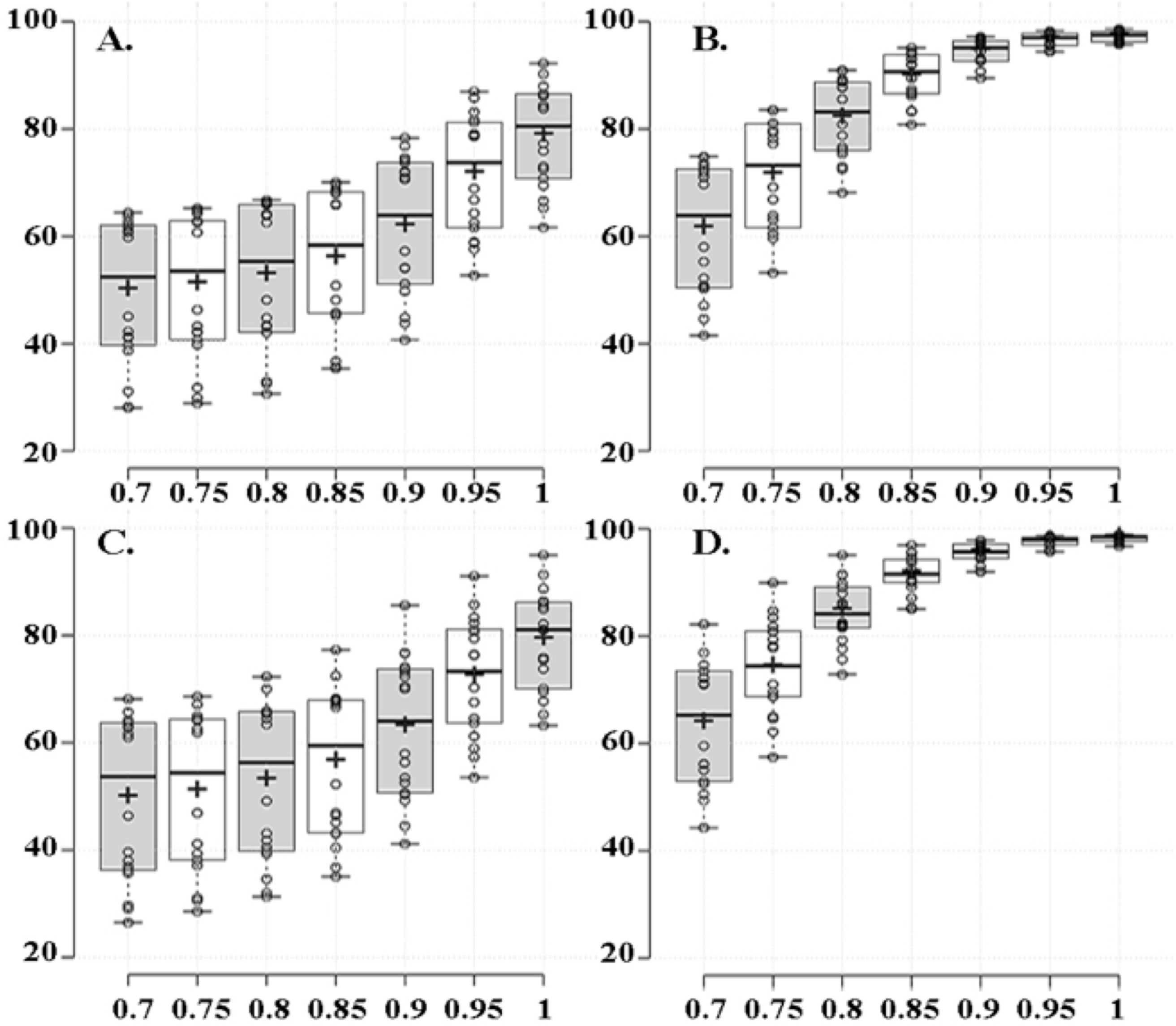
网址: http://www.elabcaas.cn/codon/newjob.html

软件著作权 No. 2016SR103338 Sci Rep. 2017, Int J Mol Sci, 2018

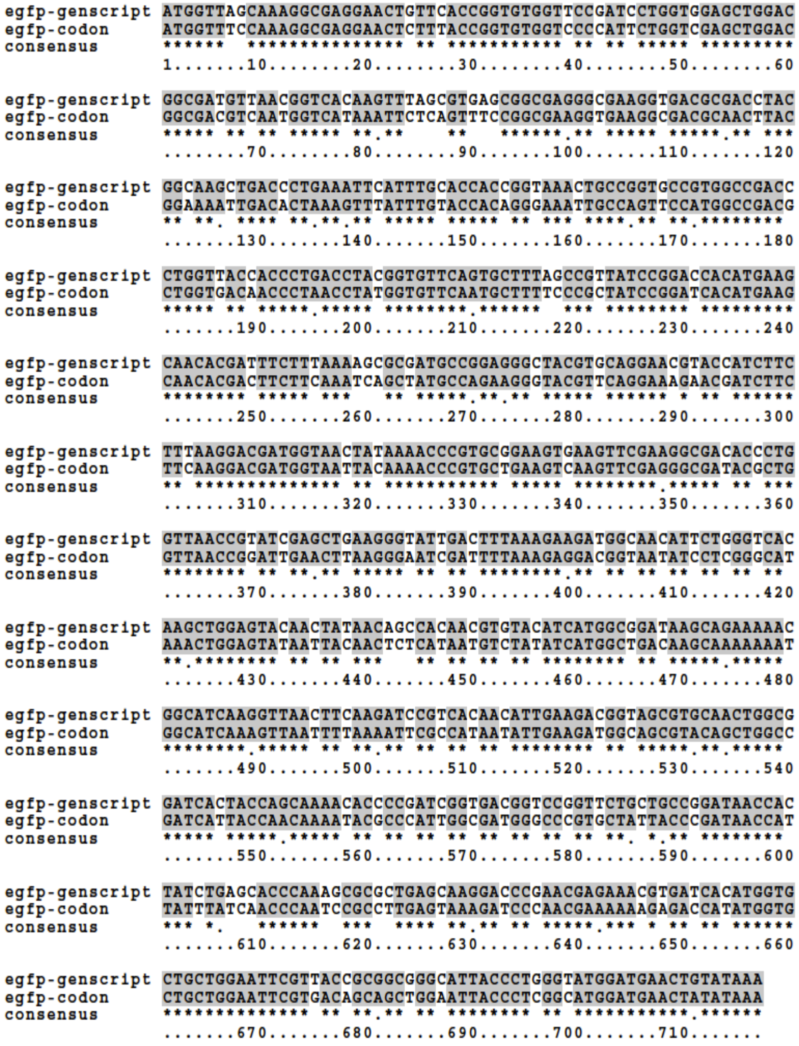
Prediction performance



The prediction accuracy with different adjacent residue number.

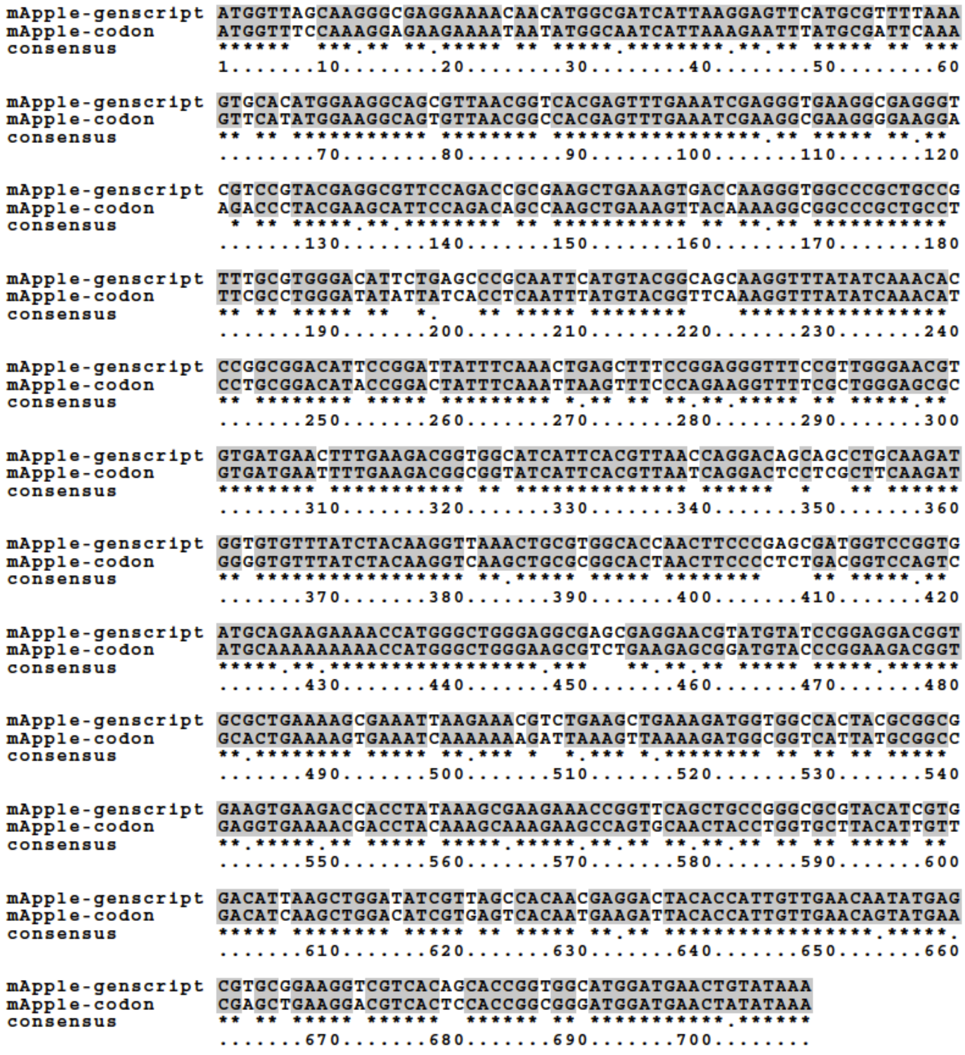


The prediction performance of the 18 classifiers for target codon selection with different matched cutoff in *E. coli* (A, B) and *B. subtilis* (C, D).



```
egfp-genscript ATGGTTAGCAAAGGCGAGGAACTGTTCACCGGTGTGGTTCCGATCCTGGTGGAGCTGGAC
egfp-codon     ATGGTTTCCAAAGGCGAGGAACTCTTTACCGGTGTGGTCCCCATTCTGGTCGAGCTGGAC
consensus      ******  *************** ** *********** ** ** ***** *********
               1.......10........20........30........40........50........60

egfp-genscript GGCGATGTTAACGGTCACAAGTTTAGCGTGAGCGGCGAGGGCGAAGGTGACGCGACCTAC
egfp-codon     GGCGACGTCAATGGTCATAAATTCTCAGTTTCCGGCGAAGGTGAAGGCGACGCAACTTAC
consensus      ***** ** ** ***** **.**    **   ******.** ***** *****.** ***
               ........70........80........90.......100.......110.......120

egfp-genscript GGCAAGCTGACCCTGAAATTCATTTGCACCACCGGTAAACTGCCGGTGCCGTGGCCGACC
egfp-codon     GGAAAATTGACACTAAAGTTTATTTGTACCACAGGGAAATTGCCAGTTCCATGGCCGACG
consensus      ** **. **** **.**.** ***** ***** ** *** ****.** **.********
               .......130.......140.......150.......160.......170.......180

egfp-genscript CTGGTTACCACCCTGACCTACGGTGTTCAGTGCTTTAGCCGTTATCCGGACCACATGAAG
egfp-codon     CTGGTGACAACCCTAACCTATGGTGTTCAATGCTTTTCCCGCTATCCGGATCACATGAAG
consensus      ***** ** *****.***** ********.******  *** ******** *********
               .......190.......200.......210.......220.......230.......240

egfp-genscript CAACACGATTTCTTTAAAAGCGCGATGCCGGAGGGCTACGTGCAGGAACGTACCATCTTC
egfp-codon     CAACACGACTTCTTCAAATCAGCTATGCCAGAAGGGTACGTTCAGGAAAGAACGATCTTC
consensus      ******** ***** ***   ** *****.**.** ***** ****** * ** ******
               .......250.......260.......270.......280.......290.......300

egfp-genscript TTTAAGGACGATGGTAACTATAAAACCCGTGCGGAAGTGAAGTTCGAAGGCGACACCCTG
egfp-codon     TTCAAGGACGATGGTAATTACAAAACCCGTGCTGAAGTCAAGTTCGAGGGCGATACGCTG
consensus      ** ************** ** *********** ***** ********.***** ** ***
               .......310.......320.......330.......340.......350.......360

egfp-genscript GTTAACCGTATCGAGCTGAAGGGTATTGACTTTAAAGAAGATGGCAACATTCTGGGTCAC
egfp-codon     GTTAACCGGATTGAACTTAAGGGAATCGATTTTAAAGAGGACGGTAATATCCTCGGGCAT
consensus      ******** ** **.** ***** ** ** ********.** ** ** ** ** ** **
               .......370.......380.......390.......400.......410.......420

egfp-genscript AAGCTGGAGTACAACTATAACAGCCACAACGTGTACATCATGGCGGATAAGCAGAAAAAC
egfp-codon     AAACTGGAGTATAATTACAACTCTCATAATGTCTATATCATGGCTGACAAGCAAAAAAAT
consensus      **.******** ** ** ***   ** ** ** ** ******** ** *****.*****
               .......430.......440.......450.......460.......470.......480

egfp-genscript GGCATCAAGGTTAACTTCAAGATCCGTCACAACATTGAAGACGGTAGCGTGCAACTGGCG
egfp-codon     GGCATCAAAGTTAATTTTAAAATTCGCCATAATATTGAAGATGGCAGCGTACAGCTGGCC
consensus      ********.***** ** **.** ** ** ** ******** ** *****.**.*****
               .......490.......500.......510.......520.......530.......540

egfp-genscript GATCACTACCAGCAAAACACCCCGATCGGTGACGGTCCGGTTCTGCTGCCGGATAACCAC
egfp-codon     GATCATTACCAACAAAATACGCCCATTGGCGATGGGCCCGTGCTATTACCCGATAACCAT
consensus      ***** *****.***** ** ** ** ** ** ** ** ** **. *.** ********
               .......550.......560.......570.......580.......590.......600

egfp-genscript TATCTGAGCACCCAAAGCGCGCTGAGCAAGGACCCGAACGAGAAACGTGATCACATGGTG
egfp-codon     TATTTATCAACCCAATCCGCCTTGAGTAAAGATCCCAACGAAAAAGAGACCATATGGTG
consensus      *** *.   ******  ***  **** **.** ** *****.*** * ** ** ******
               .......610.......620.......630.......640.......650.......660

egfp-genscript CTGCTGGAATTCGTTACCGCGGCGGGCATTACCCTGGGTATGGATGAACTGTATAAA
egfp-codon     CTGCTGGAATTCGTGACAGCAGCTGGAATTACCCTCGGCATGGATGAACTATATAAA
consensus      ************** ** **.** ** ******** ** ***********.******
               .......670.......680.......690.......700.......710.......
```

egfp

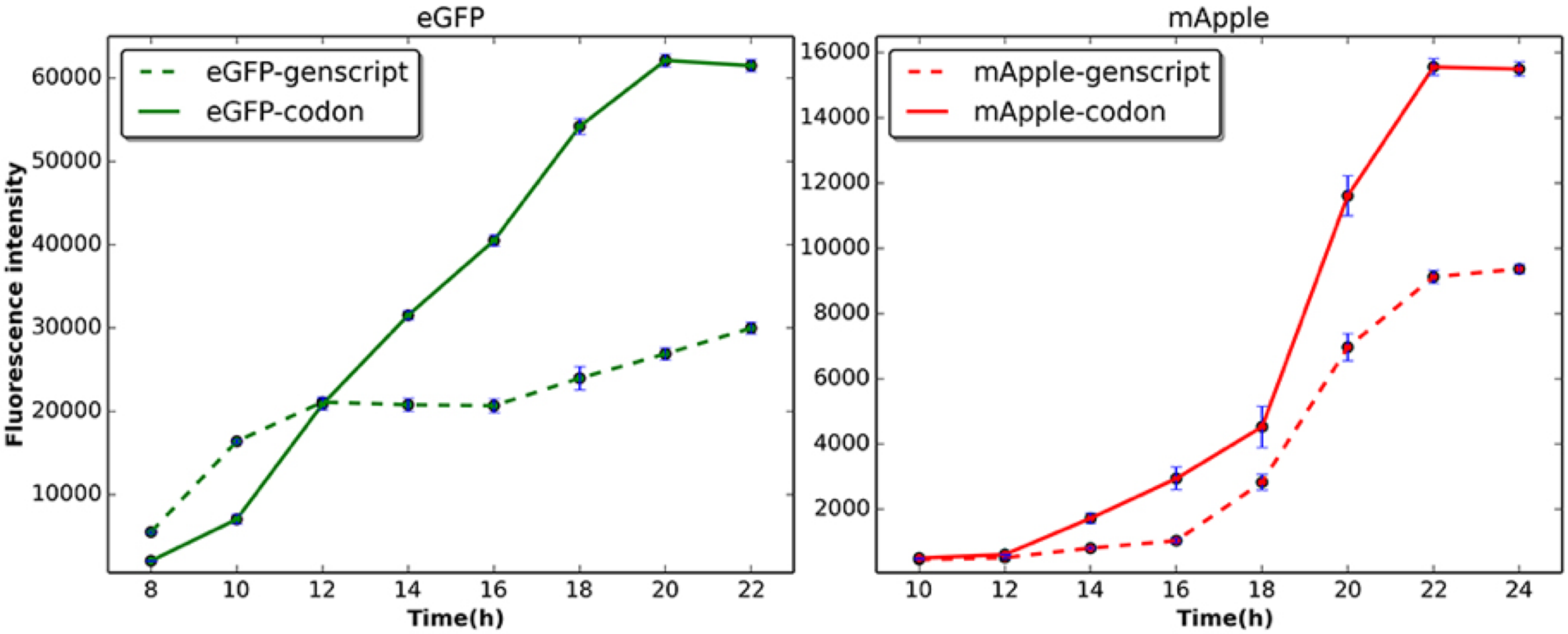


```
mApple-genscript ATGGTTAGCAAGGGCGAGGAAAACAACATGGCGATCATTAAGGAGTTCATGCGTTTTAAA
mApple-codon     ATGGTTTCCAAAGGAGAAGAAAATAATATGGCAATCATTAAAGAATTTATGCGATTCAAA
consensus        ******  ***.** **.***** ** *****.********.**.** ***** ** ***
                 1.......10........20........30........40........50........60

mApple-genscript GTGCACATGGAAGGCAGCGTTAACGGTCACGAGTTTGAAATCGAGGGTGAAGGCGAGGGT
mApple-codon     GTTCATATGGAAGGCAGTGTTAACGGCCACGAGTTTGAAATCGAAGGCGAAGGGGAAGGA
consensus        ** ** *********** ******** *****************.** ***** **.**
                 ........70........80........90.......100.......110.......120

mApple-genscript CGTCCGTACGAGGCGTTCCAGACCGCGAAGCTGAAAGTGACCAAGGGTGGCCCGCTGCCG
mApple-codon     AGACCCTACGAAGCATTCCAGACAGCCAAGCTGAAAGTTACAAAAGGCGGCCCGCTGCCT
consensus         * ** *****.**.******** ** *********** ** **.** ***********
                 .......130.......140.......150.......160.......170.......180

mApple-genscript TTTGCGTGGGACATTCTGAGCCCGCAATTCATGTACGGCAGCAAGGTTTATATCAAACAC
mApple-codon     TTCGCCTGGGATATATTATCACCTCAATTTATGTACGGTTCAAAGGTTTATATCAAACAT
consensus        ** ** ***** **  *.   ** ***** ********    *****************
                 .......190.......200.......210.......220.......230.......240

mApple-genscript CCGGCGGACATTCCGGATTATTTCAAACTGAGCTTTCCGGAGGGTTTCCGTTGGGAACGT
mApple-codon     CCTGCGGACATACCGGACTATTTCAAATTAAGTTTCCCAGAAGGTTTTCGCTGGGAGCGC
consensus        ** ******** ***** ********* *.** ** **.**.***** ** *****.**
                 .......250.......260.......270.......280.......290.......300

mApple-genscript GTGATGAACTTTGAAGACGGTGGCATCATTCACGTTAACCAGGACAGCAGCCTGCAAGAT
mApple-codon     GTGATGAATTTTGAAGACGGCGGTATCATTCACGTTAATCAGGACTCCTCGCTTCAAGAT
consensus        ******** *********** ** ************** ******  *   ** ******
                 .......310.......320.......330.......340.......350.......360

mApple-genscript GGTGTGTTTATCTACAAGGTTAAACTGCGTGGCACCAACTTCCCGAGCGATGGTCCGGTG
mApple-codon     GGGGTGTTTATCTACAAGGTCAAGCTGCGCGGCACTAACTTCCCCTCTGACGGTCCAGTC
consensus        ** ***************** **.***** ***** ********    ** *****.**
                 .......370.......380.......390.......400.......410.......420

mApple-genscript ATGCAGAAGAAAACCATGGGCTGGGAGGCGAGCGAGGAACGTATGTATCCGGAGGACGGT
mApple-codon     ATGCAAAAAAAAACCATGGGCTGGGAAGCGTCTGAAGAGCGGATGTACCCGGAAGACGGT
consensus        *****.**.*****************.***   **.**.** ***** *****.******
                 .......430.......440.......450.......460.......470.......480

mApple-genscript GCGCTGAAAAGCGAAATTAAGAAACGTCTGAAGCTGAAAGATGGTGGCCACTACGCGGCG
mApple-codon     GCACTGAAAAGTGAAATCAAAAAAAGATTAAAGTTAAAAGATGGCGGTCATTATGCGGCC
consensus        **.******** ***** **.*** *  *.*** *.******** ** ** ** *****
                 .......490.......500.......510.......520.......530.......540

mApple-genscript GAAGTGAAGACCACCTATAAAGCGAAGAAACCGGTTCAGCTGCCGGGCGCGTACATCGTG
mApple-codon     GAGGTGAAAACGACCTACAAAGCAAAGAAGCCAGTGCAACTACCTGGTGCTTACATTGTT
consensus        **.*****.** ***** *****.*****.**.** **.**.** ** ** ***** **
                 .......550.......560.......570.......580.......590.......600

mApple-genscript GACATTAAGCTGGATATCGTTAGCCACAACGAGGACTACACCATTGTTGAACAATATGAG
mApple-codon     GACATCAAGCTGGACATCGTGAGTCACAATGAAGATTACACCATTGTTGAACAGTATGAA
consensus        ***** ******** ***** ** ***** **.** *****************.*****.
                 .......610.......620.......630.......640.......650.......660

mApple-genscript CGTGCGGAAGGTCGTCACAGCACCGGTGGCATGGATGAACTGTATAAA
mApple-codon     CGAGCTGAAGGACGTCACTCCACCGGCGGGATGGATGAACTATATAAA
consensus        ** ** ***** ******  ****** ** ***********.******
                 .......670.......680.......690.......700........
```

mApple

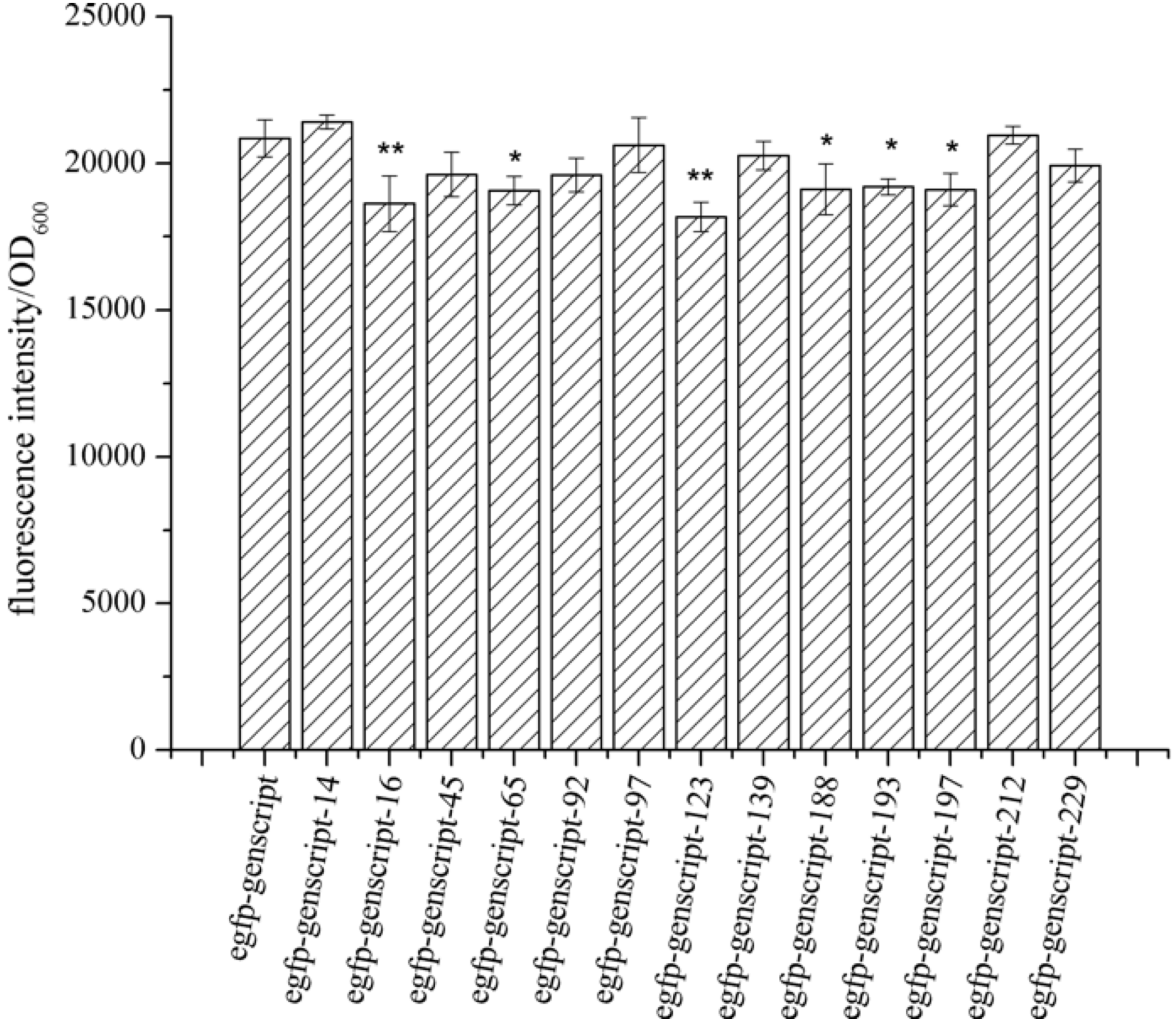
codons	AA	Codon usage	D4_GFP	D6_GFP	codons	AA	Codon usage	D4_GFP	D6_GFP
GCA	A	20.21	0	2	AAC	N	21.58	13	4
GCC	A	25.53	0	2	AAT	N	17.67	0	9
GCG	A	33.68	8	0	**CCA**	**P**	**8.46**	**0**	**3**
GCT	A	15.23	0	4	**CCC**	**P**	**5.47**	**0**	**5**
TGC	C	6.48	2	1	**CCG**	**P**	**23.25**	**10**	**2**
TGT	C	5.17	0	1	**CCT**	**P**	**6.99**	**0**	**0**
GAC	D	19.08	9	8	CAA	Q	15.36	4	6
GAT	D	32.1	9	10	CAG	Q	28.86	4	2
GAA	E	39.53	9	11	**AGA**	**R**	**2.07**	**0**	**2**
GAG	E	17.78	7	5	**AGG**	**R**	**1.2**	**0**	**0**
TTC	F	16.51	7	7	**CGA**	**R**	**3.56**	**0**	**0**
TTT	F	22.2	5	5	**CGC**	**R**	**22**	**0**	**2**
GGA	**G**	**7.92**	**0**	**3**	**CGG**	**R**	**5.44**	**0**	**1**
GGC	**G**	**29.63**	**10**	**9**	**CGT**	**R**	**20.9**	**6**	**1**
GGG	**G**	**11.05**	**0**	**4**	AGC	S	16.04	10	1
GGT	**G**	**24.73**	**12**	**6**	AGT	S	8.75	0	1
CAC	H	9.71	9	2	TCA	S	7.16	0	3
CAT	H	12.91	0	7	TCC	S	8.6	0	4
ATA	I	4.4	0	0	TCG	S	8.95	0	0
ATC	I	25.17	7	5	TCT	S	8.43	0	1
ATT	I	30.4	5	7	ACA	T	7.07	0	4
AAA	K	33.67	9	14	ACC	T	23.37	16	7
AAG	K	10.32	11	6	ACG	T	14.45	0	4
CTA	**L**	**3.92**	**0**	**4**	ACT	T	8.92	0	1
CTC	**L**	**11.1**	**0**	**3**	GTA	V	10.9	0	1
CTG	**L**	**52.8**	**21**	**8**	GTC	V	15.28	0	5
CTT	**L**	**11.03**	**0**	**1**	GTG	V	26.19	9	6
TTA	**L**	**13.92**	**0**	**2**	GTT	V	18.29	9	6
TTG	**L**	**13.69**	**0**	**3**	TAC	Y	12.21	6	6
					TAT	Y	16.12	5	5

Design of the genes (*gfp* and *mApple*)

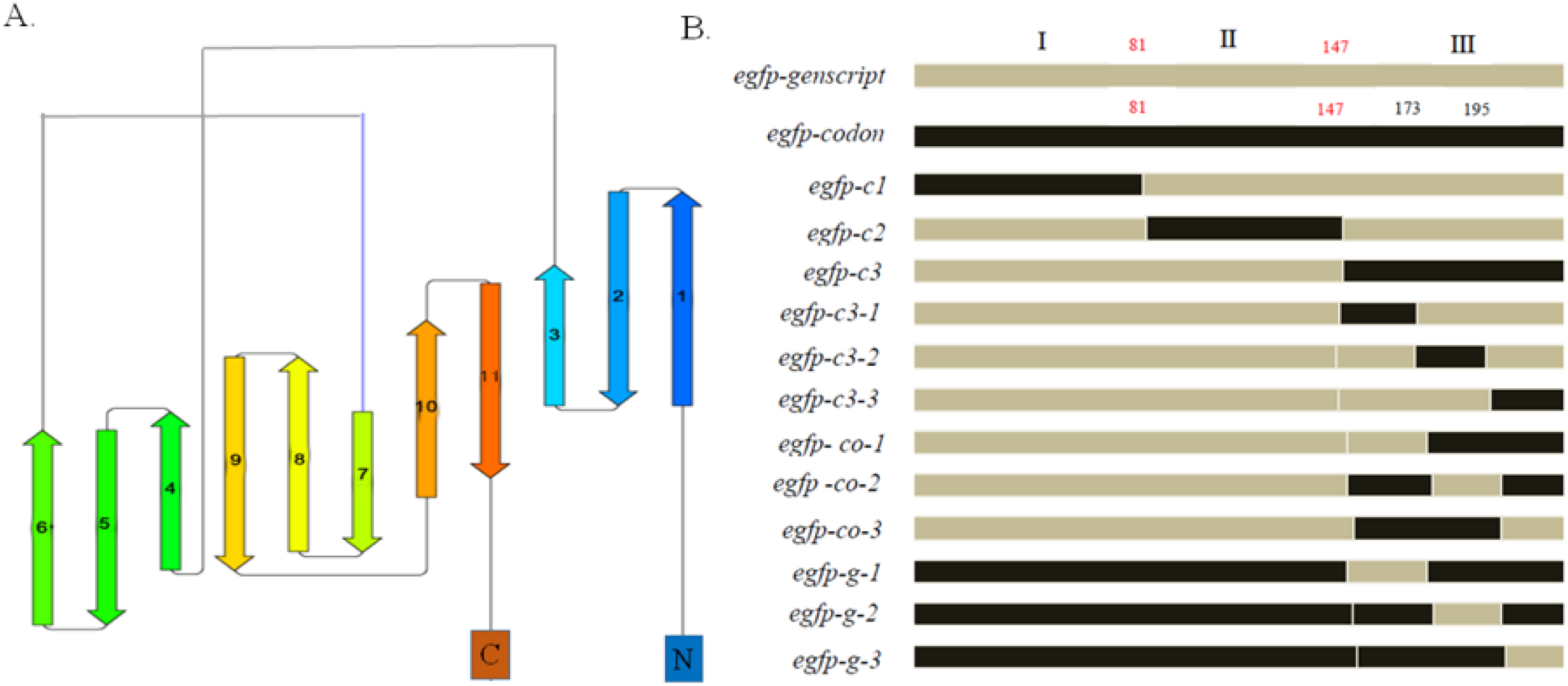


Fluorescence intensity of E. coli containing the reporter genes (egfp or mApple).
The reporter genes (egfp-codon and mApple-codon) were designed based on the model in this study

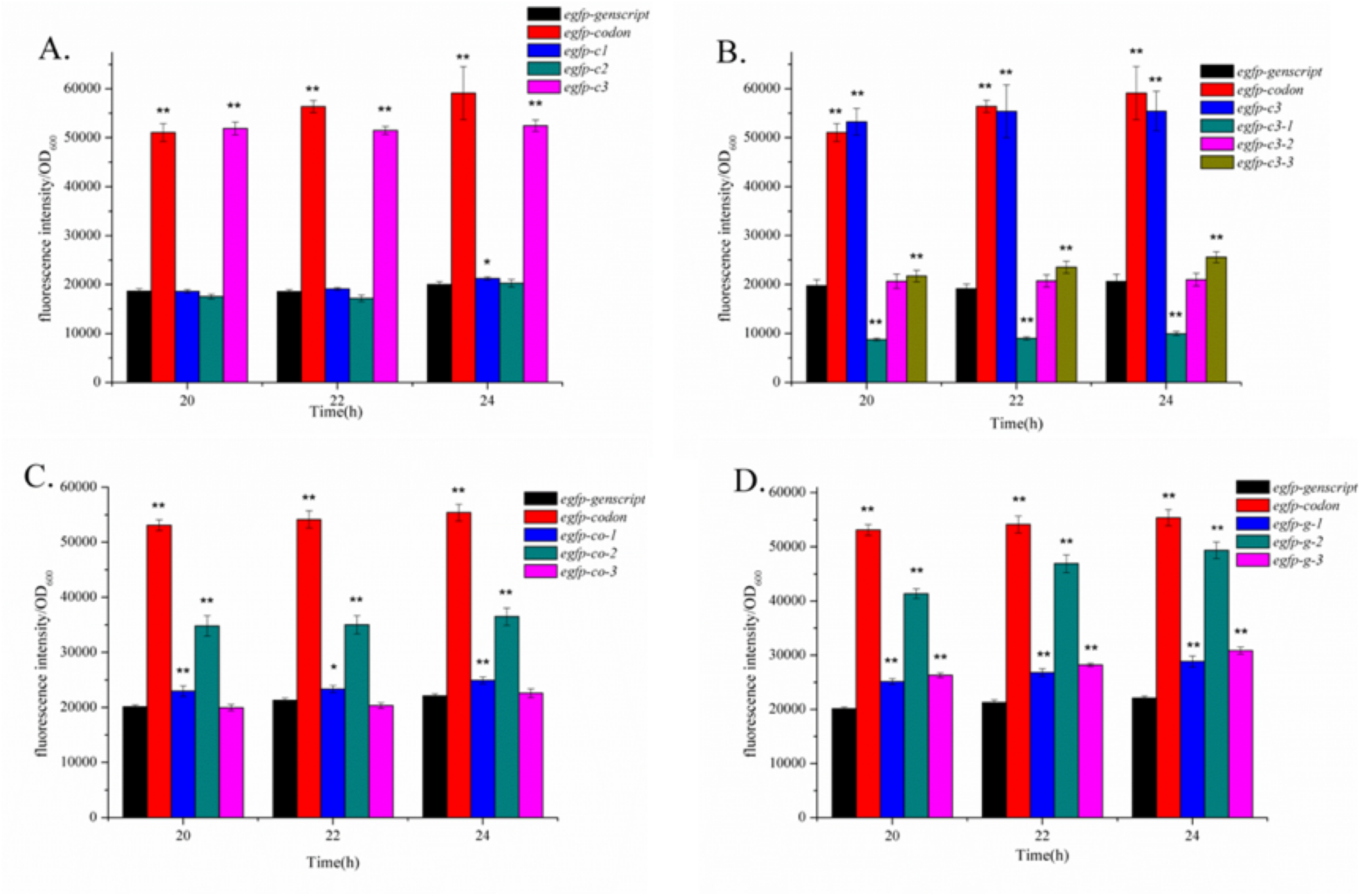
Single point mutations on eGFP



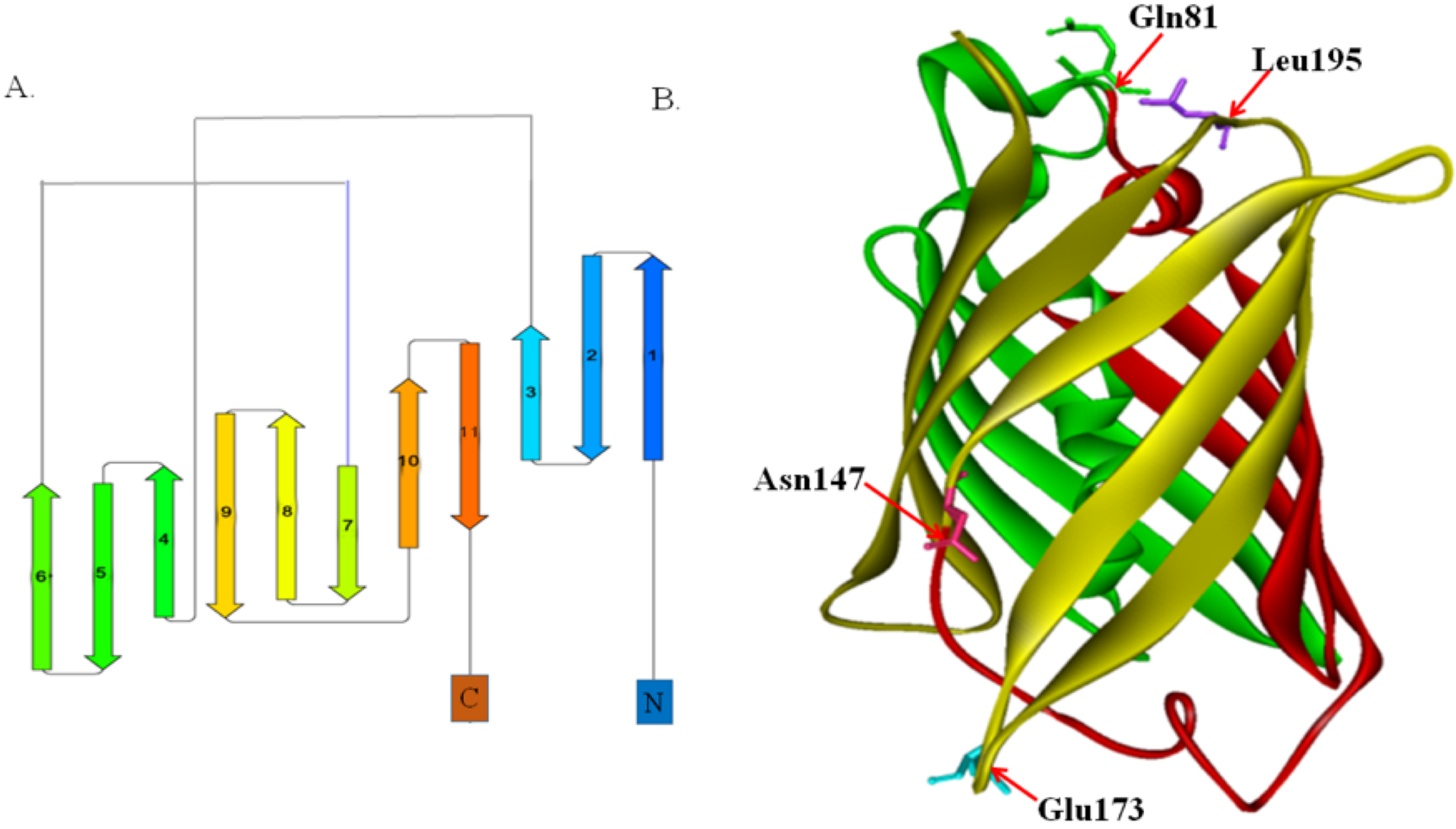
Structure-guided Domain recombination



C-terminal region of *egfp-codon* related to the production of eGFP

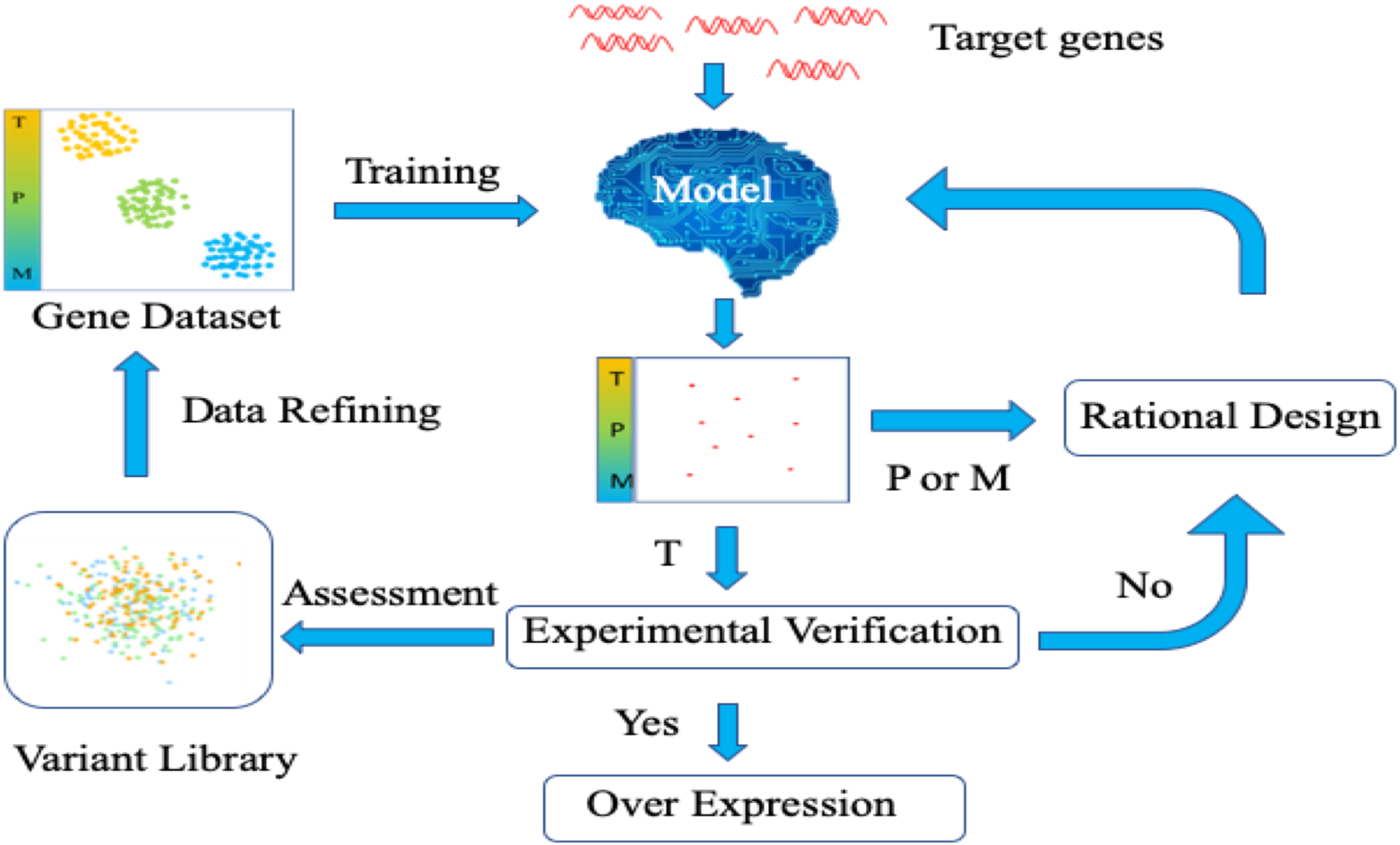


Structure of eGFP



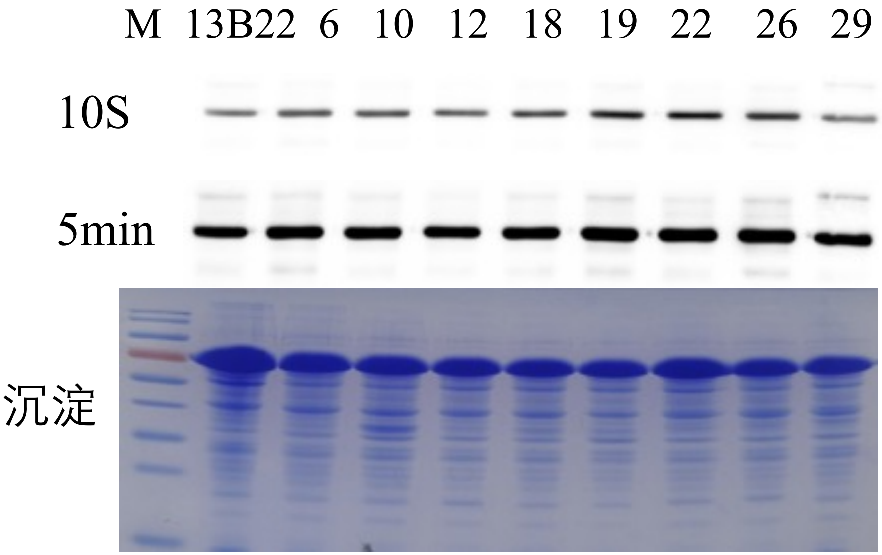
密码子的选择与其结构及表达密切相关。

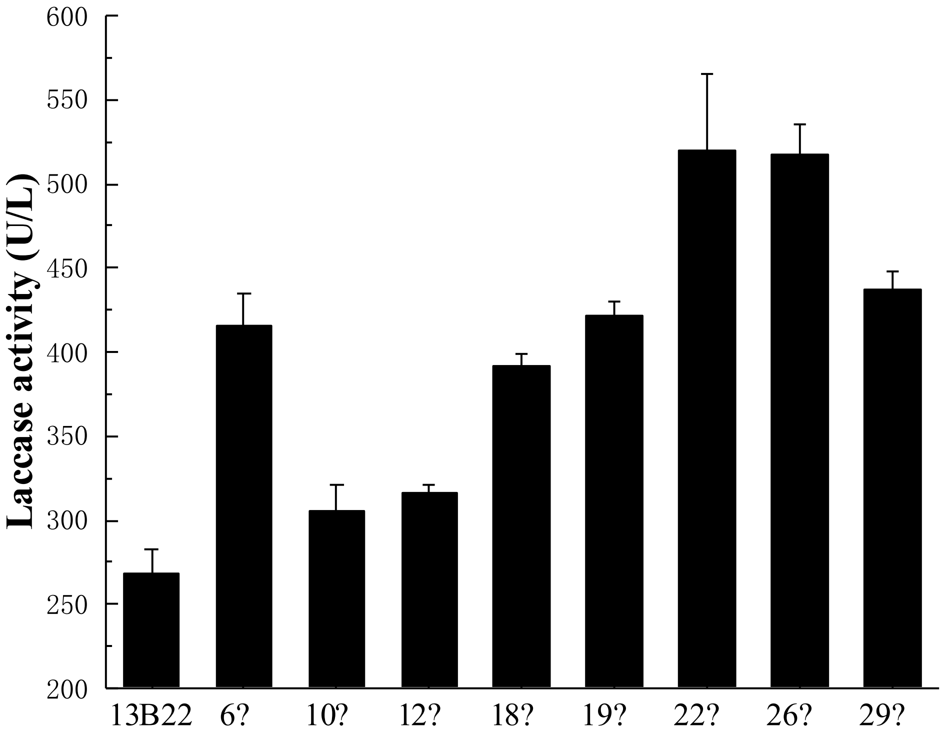
基于AI的高表达蛋白的学习算法



蛋白质表达数据来自于 PaxDb，包括56个不同的物种，基因数量220396条，预测准确率82%

漆酶13B22的验证实验





选择本实验室分离到的活力高，但难表达的漆酶13B22进行实验，发现在设计的30个突变体中有8个可以提高目标蛋白的表达量。

Unpublished Data, 2020

Advantages of thermostability in industrial processes

▶ Use of enzymes at high temperatures

▶ **Thermostability => stability against other denaturating agents (e.g., detergents, proteolytic attack, organic solvents)**

▶ **10 ^{o}C increase in temperature => $\underline{>}$2-fold increase in reaction rate**

▶ Endothermic reaction => equilibrium is shifted toward the final product at high temp.

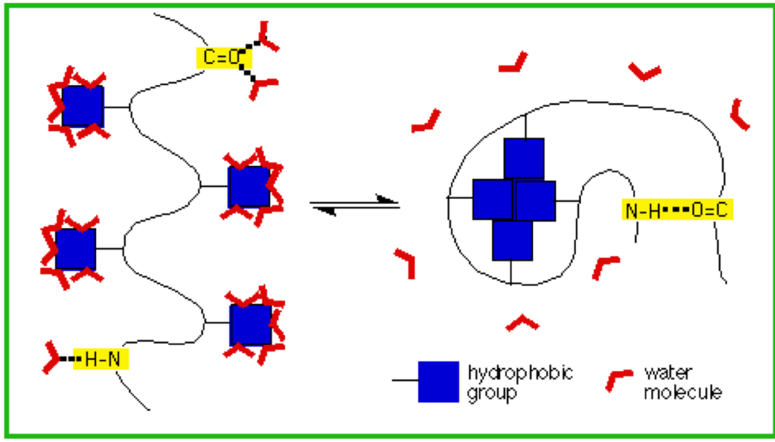
▶ Temperatures > 60 ^{o}C inhibit microbial growth and >70 ^{o}C kill almost all pathogenic bacteria.

Protein Stability

- Protein stability is an important protein character, related **to its application in biotechnology industry** and **organismal fitness**.
- **Lots of methods to be proposed to design the mutations.**

  Foldx, Eris, CC/PBSA, SDM, I-mutant, Prethermut ….
- There are lots of factor/forces affected the protein stability.
- Hydrophobic, H-bond, ionic bond, compactness, Quaternary structures, ……

http://www.mobioinfor.cn

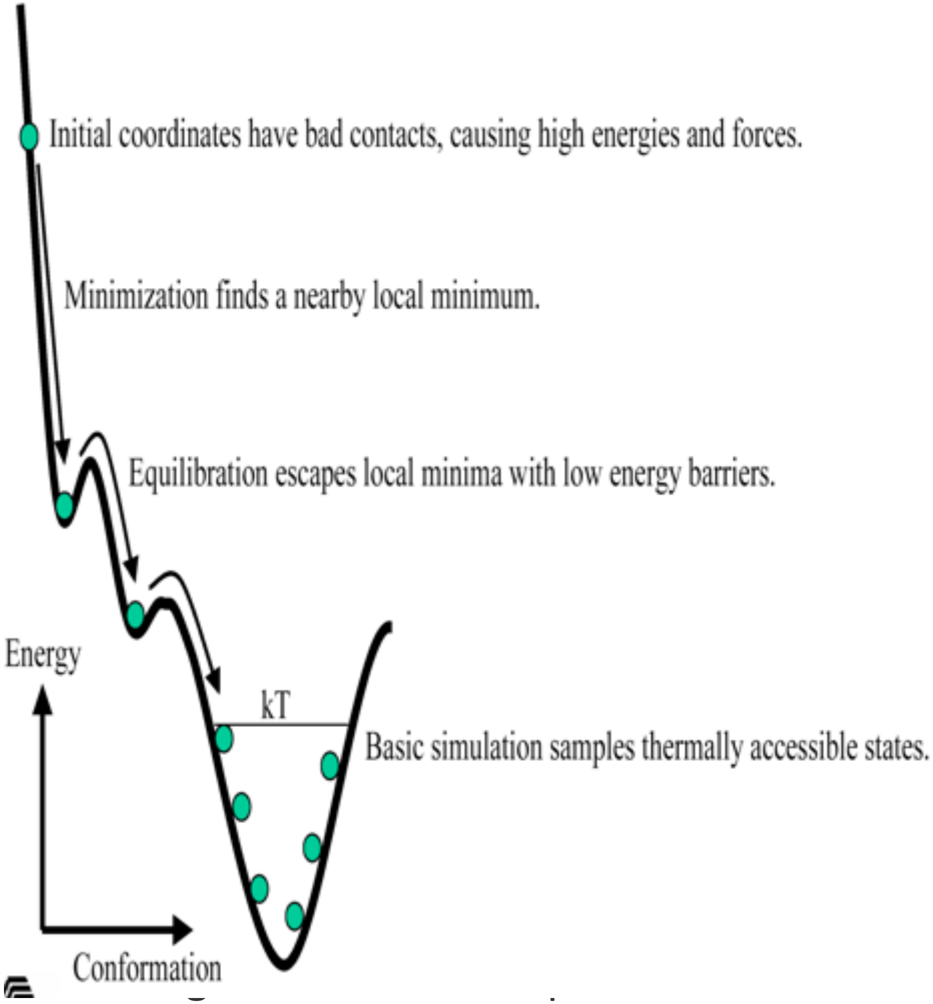


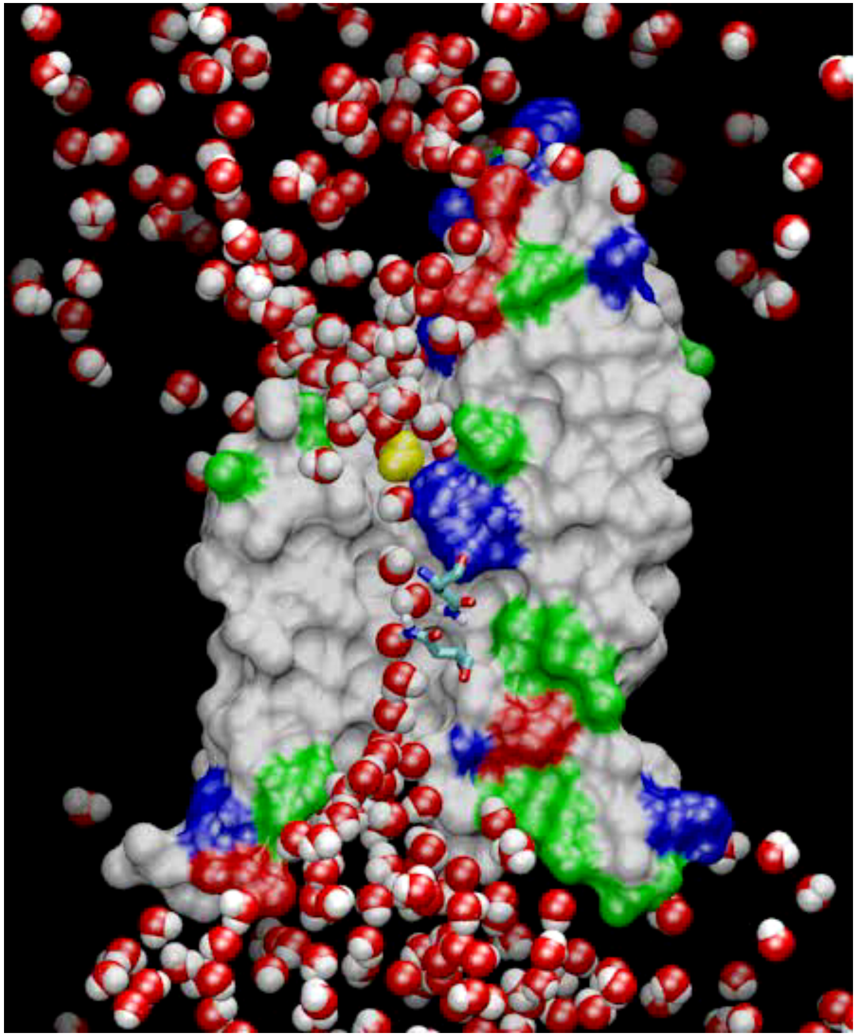
蛋白质的稳定性

蛋白质结构中最不稳定性的片段或氨基酸决定了蛋白质的稳定性。

木桶的
盛水量
取决于
桶壁上
最短的
木板
木桶理论

1.5 分子动力学模拟

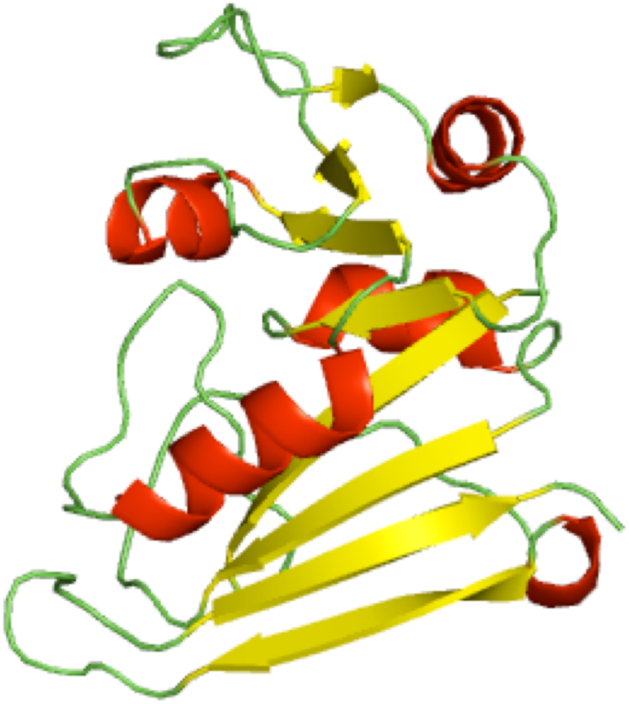

Initial coordinates have bad contacts, causing high energies and forces.
Minimization finds a nearby local minimum.
Equilibration escapes local minima with low energy barriers.
kT
Basic simulation samples thermally accessible states.
Energy
Conformation

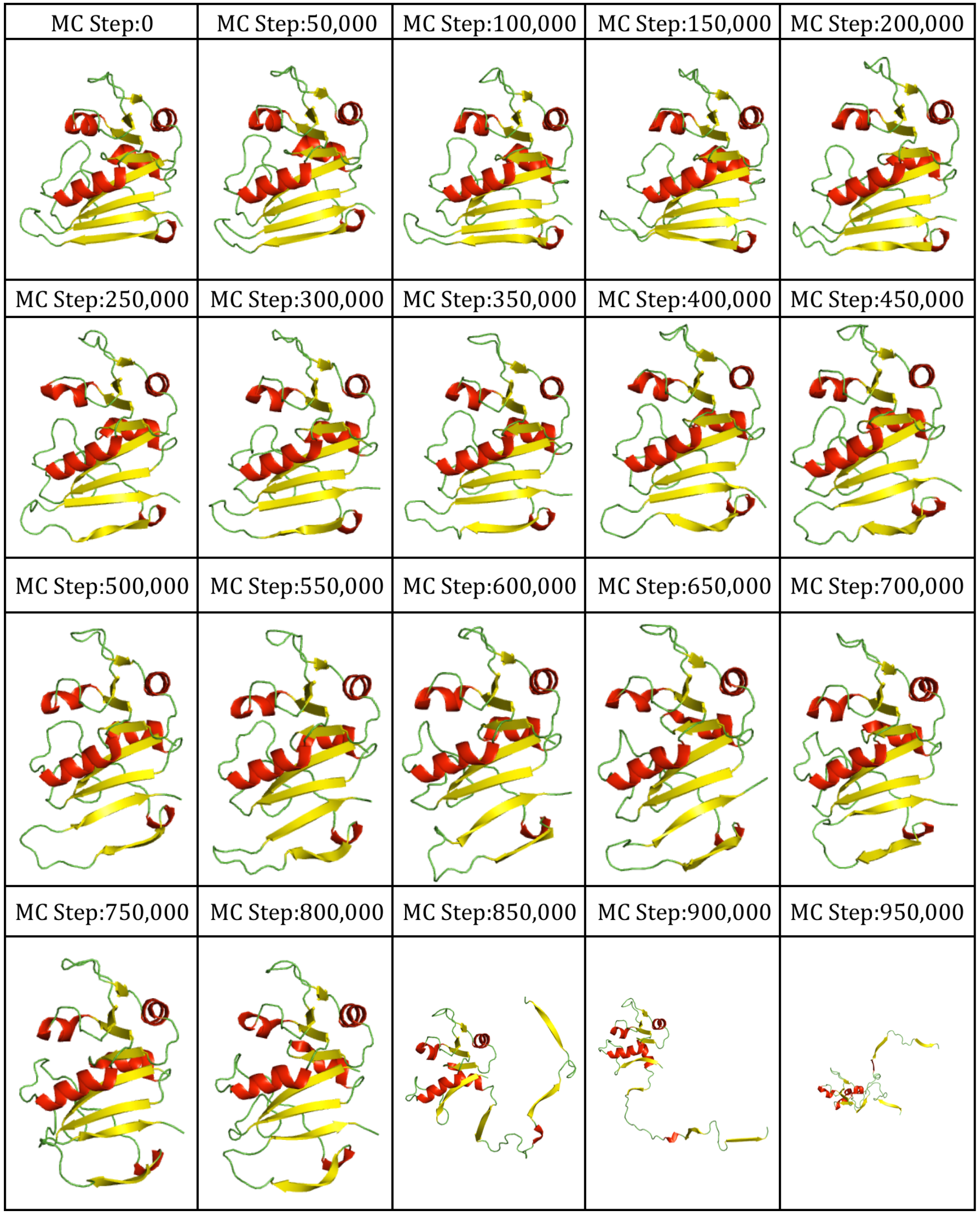


Water Channels in Cell Membranes

de Groot BL et al Science. 2001 Dec 14;294(5550):2353-7.

Unfolding DHFR with Monte Carlo method

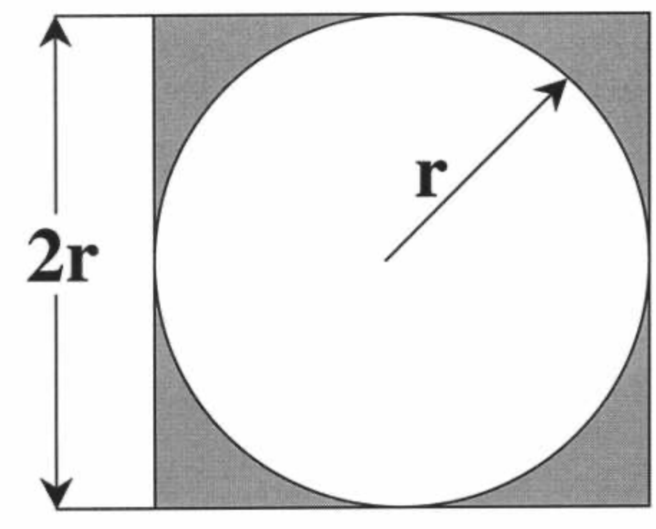


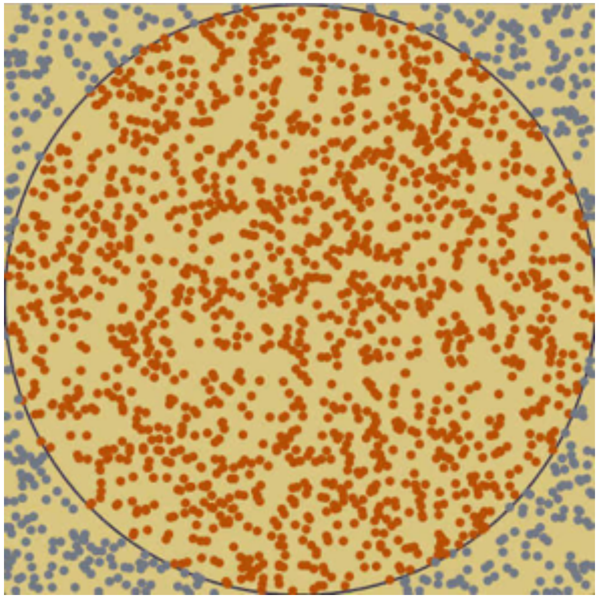


The unfolding figures of DHFR at the temperature (1.5) with different Monte Carlo Steps.

蒙特卡罗方法

计算机模拟经常采用随机模拟方法，不断产生随机数序列来模拟过程，这就是蒙特卡洛方法。





$$\frac{\textit{Area of Circle}}{\textit{Area of Square}} = \frac{\pi r^2}{(2r)^2} = \frac{\pi}{4}$$

```
#!/usr/bin/env python
from random import random
from math import sqrt
total = 10000
x = y = inn = out = 0.0
for i in range(total):
    x = random()
    y = random()
    if (i % (total/10) == 0):
        print(i)
    if (sqrt(x * x + y * y)<1.0):
        inn += 1.0
    else:
        out += 1.0
print(total, inn, out)
print(inn*4 / total)
```

Python程序16行

当total=10000时，计算的结果为3.15
当total=100000时，计算的结果为3.14528
当total=1000000时，计算的结果为3.141228
当total=10000000时，计算的结果为3.141518
当total=100000000时，计算的结果为3.14186572
当total=1000000000时，计算的结果为3.14166922

4.4.2 蒙特卡罗和原子弹

> 20 世纪 40 年代，美国在二战期间开始了制造原子弹的"曼哈顿计划"。由于这是人类历史上第一次操纵核裂变来制造原子弹，其危险性可想而知，尤其是其中中子的运动状态，非常复杂而且充满随机性，很难进行精确的计算。现代计算机之父、天才的数学家冯·诺依曼提出了随机模拟的方法，并使用摩纳哥的著名赌城蒙特卡罗来为这种方法命名，于是也称其为蒙特卡罗方法。

蒙特卡罗方法的本质是随机模拟，对于一些复杂的过程，如果通过数学来进行推理，通常会非常困难。但是如果考虑其中的随机性，利用随机数来进行模拟，模拟很多次之后，在大数据的情况下研究各种统计量，也能得到稳定的结果。原子弹对于我们太遥远，我们通过"第13页: 1.2.2 6 连号和 14 连号"中提到的生活中的小例子来进行介绍。

在这个例子里，武汉市 5141 名困难家庭市民参与一个经济适用房小区的公开摇号，结果中签的 124 名市民当中有 6 人的购房资格证明的编号是连号。经查，6 人申请材料系造假，购房资格被取消。

我们可以在 5141 位市民中采用简单随机抽样[1] 的方式抽取 124 人，然后将这 124 个编号排序，通过算法判断是否存在 6 连号。对于每一次抽样，只存在两个结果，有 6 连号或者没有 6 连号。我们可以把这种抽样模拟 1 亿次，统计其中出现 6 连号的次数，就可以算出这种情况下 6 连号的概率。

我们在普通的家用电脑上测试了一下这种方法，当时模拟 1 亿次花费了 2 小时，最后的结果是发生了 80 次 6 连号的情况，也就是说 6 连号的概率只有百万分之 0.8，这是一个非常小的数，所以我们有理由认为发生这种情况是非常不正常的。

后来在老河口又发生了 14 连号的事情，当时是在 1138 户具有购房资格的申请者中，抽中了 514 户购房者，其中有 14 户资格证编号相连。我们同样用这种方法模拟 1 亿次，结果得到了 829 546 次 14 连号的情况，这个概率是百分之 0.8，明显要大得多。

这个例子虽然简单，但是体现了蒙特卡罗方法的便捷之处。利用排列组合公式也可以计算这个概率，但是这个过程非常费脑[2]，即使写出了递推公式，计算起来也很费时。但我们可以用蒙特卡罗方法直接模拟抽签的过程，这里面几乎不用进行任何数学上的转换。计算抽签一次的概率和统计 1 亿次的频率，在大数定律的支持下，可以认为这两个数是很接近的，那么我们就把复杂的问题转化成了简单的统计问题。这种思路在业界非

[1] 在"第66页: 2.4.1 管中窥豹与一叶知秋"中我们介绍了抽样。

[2] 详情可以参见"第13页: 1.2.2 6 连号和 14 连号"。

Monte Carlo protein simulation approach

The all-atom energy function is as following:

$E = E_{con} + w_{trp} * E_{trp} + w_{hb} * E_{hb} + w_{sct} * E_{sct}$

E_{con} is the pairwise atom–atom contact potential.

E_{hb} is the hydrogen bonding potential.

E_{trp} is the sequence-dependent local torsional potential.

E_{sct} is the side-chain torsional angle potential.

Detailed information on the following publications:

Yang, J. S., S. Wallin, and E. I. Shakhnovich. *PNAS. 105, (2008): 895-900.*

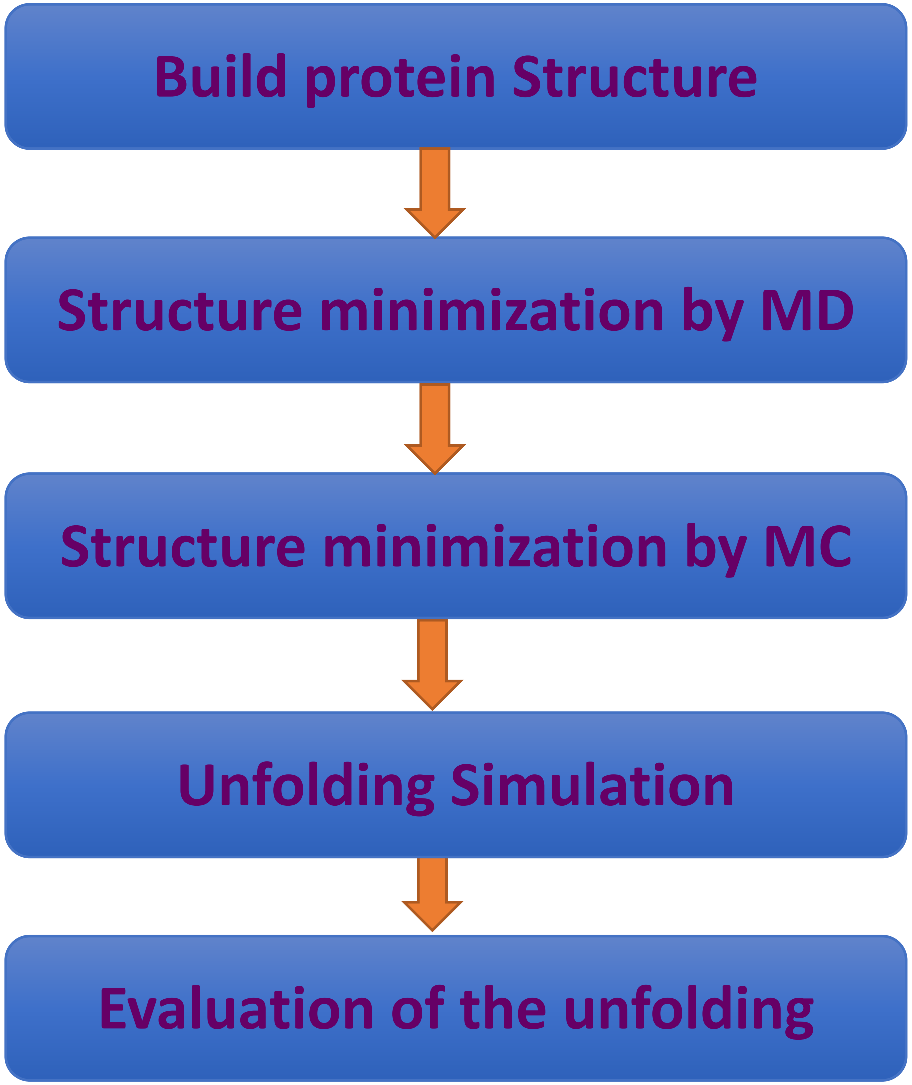
Yang, J. S., W. W. Chen, J. Skolnick, and E. I. Shakhnovich. *Structure 15, (2007): 53-63.*

Hubner, I. A., E. J. Deeds, and E. I. Shakhnovich. *PNAS*. 103, (2006): 17747-52.

Hubner, I. A., E. J. Deeds, and E. I. Shakhnovich. *PNAS. 102, (2005):* 18914-19.

Kussell, E., J. Shimada, and E. I. Shakhnovich. *PNAS. 99, (2002): 5343-48.*

Protein Unfolding procedure



Build protein Structure

Structure minimization by MD

Target structure was minimized with NAMD for 5000 steps.

Structure minimization by MC

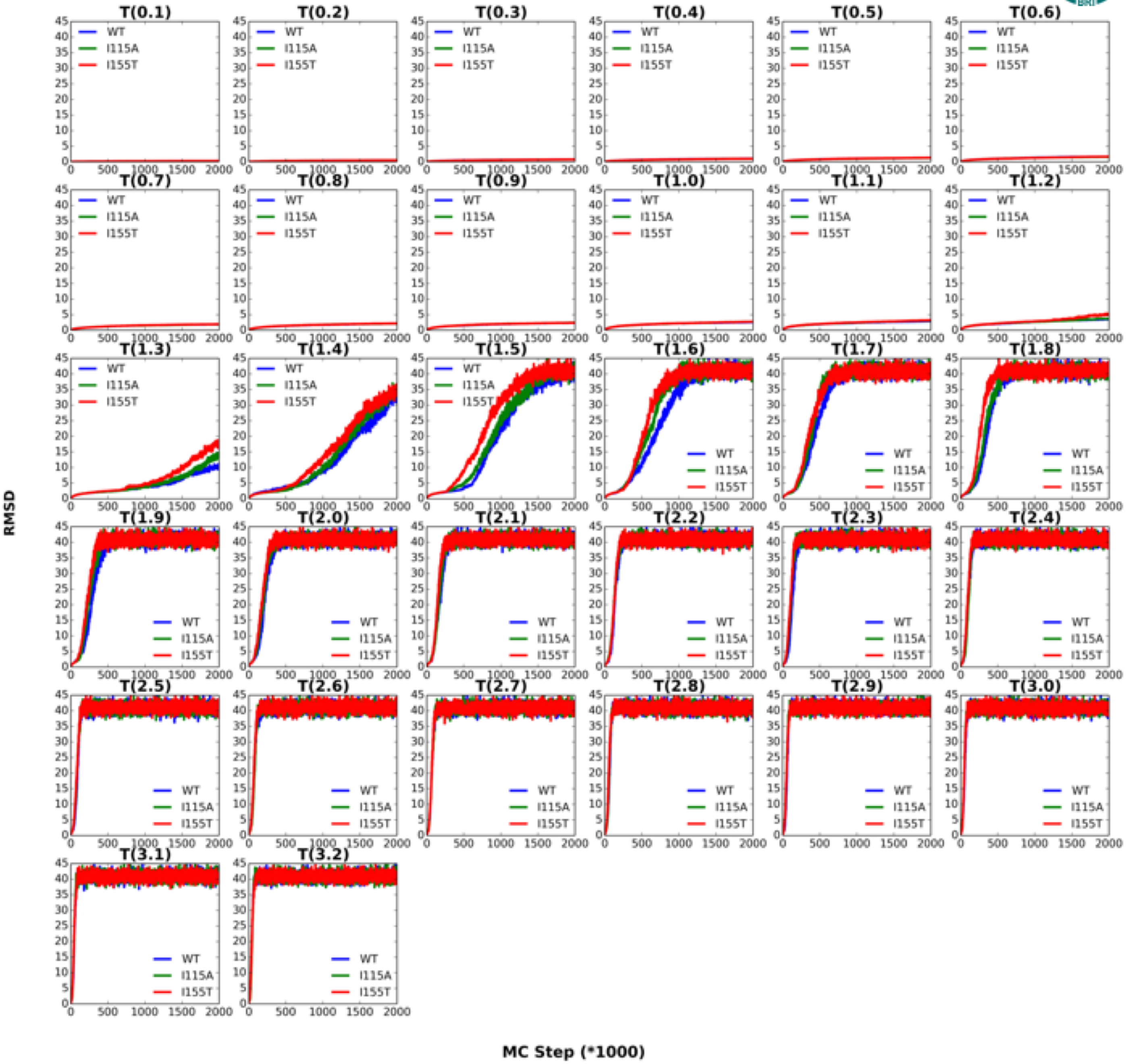
Temperature is 0.1, running of 1,000,000 steps.

Unfolding Simulation

Temperature ranges from 0.1 to 3.2, which the temperature step is 0.1.

Evaluation of the unfolding

Protein simulation results-RMSD



T_m of DHFRs:

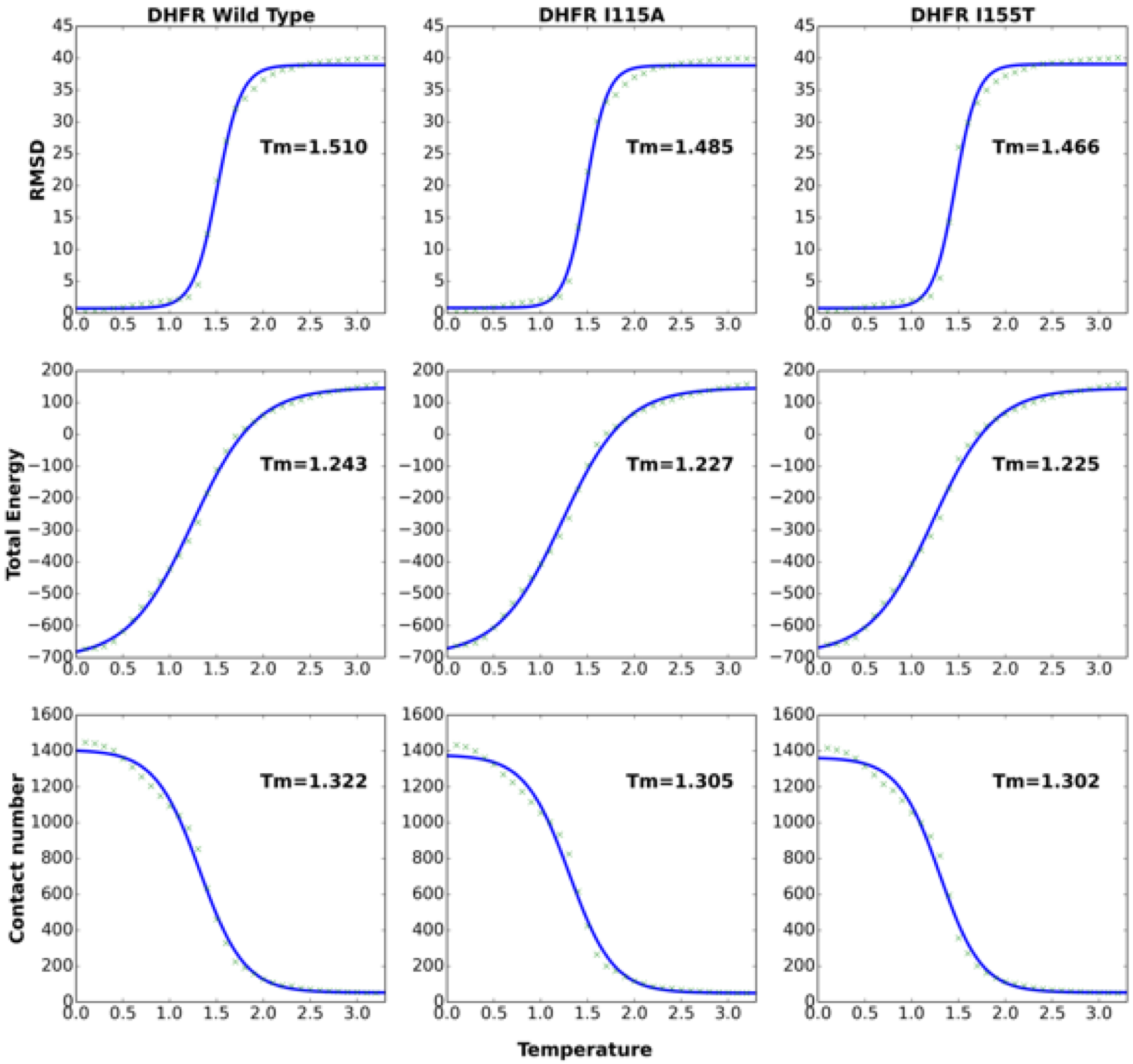
WT :51.7 °C

I115A:46.1 °C

I155T :43.4 °C

Each protein was simulated in 50 replications for 2,000,000 MC steps. The RMSD is the average RMSD in all of the 50 replications.

Unfolding Curves of DHFR WT



***T*$_m$ of DHFRs:**

WT :51.7 °C

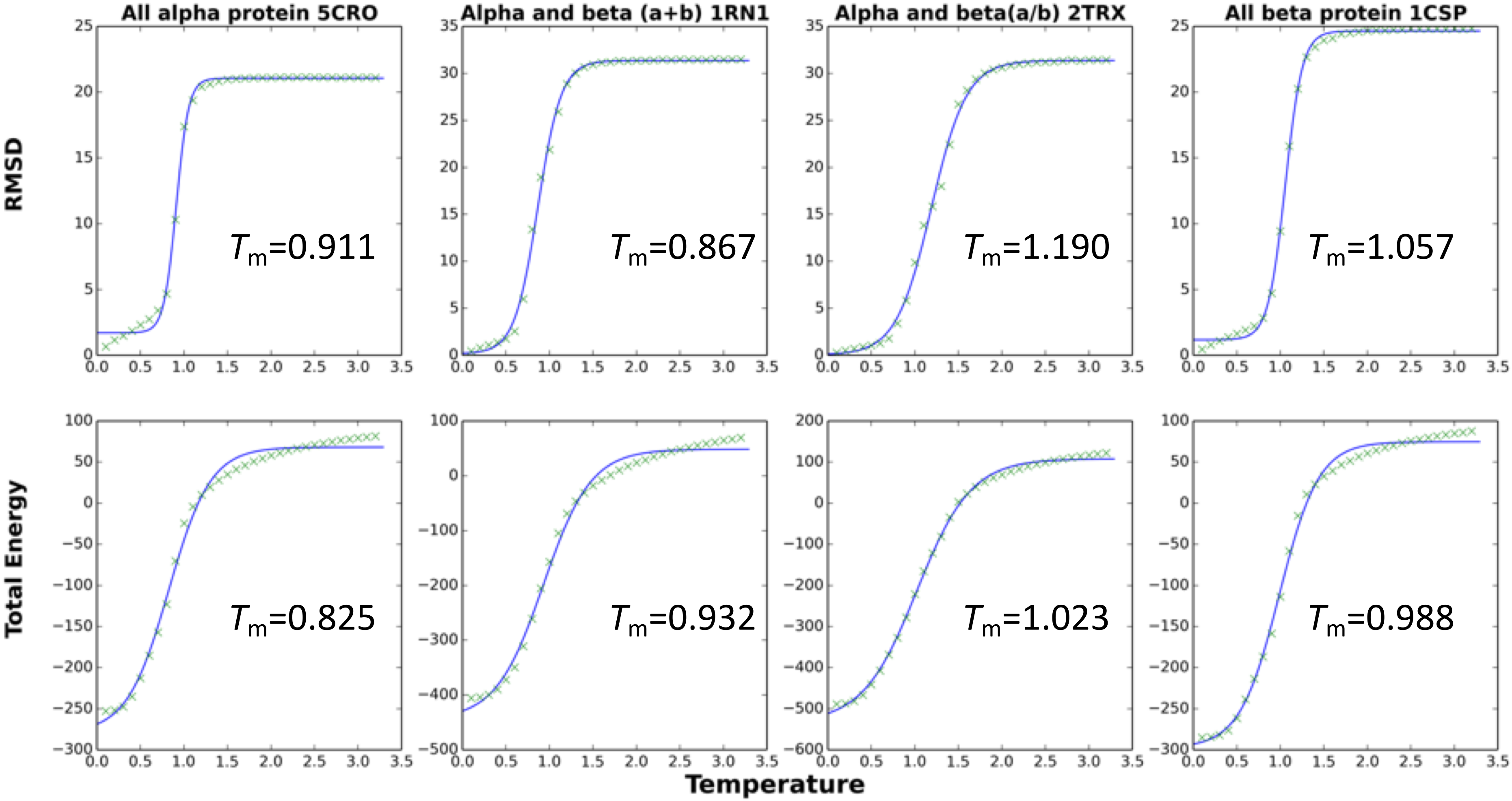
I115A:46.1 °C

I155T :43.4 °C

The protein was simulated in 50 replications for 2,000,000 MC steps.

The points were the average RMSD, energy or the contact number in the last 1,000,000 steps in the 50 replications. The Tm values were simulated the sigmoid function.

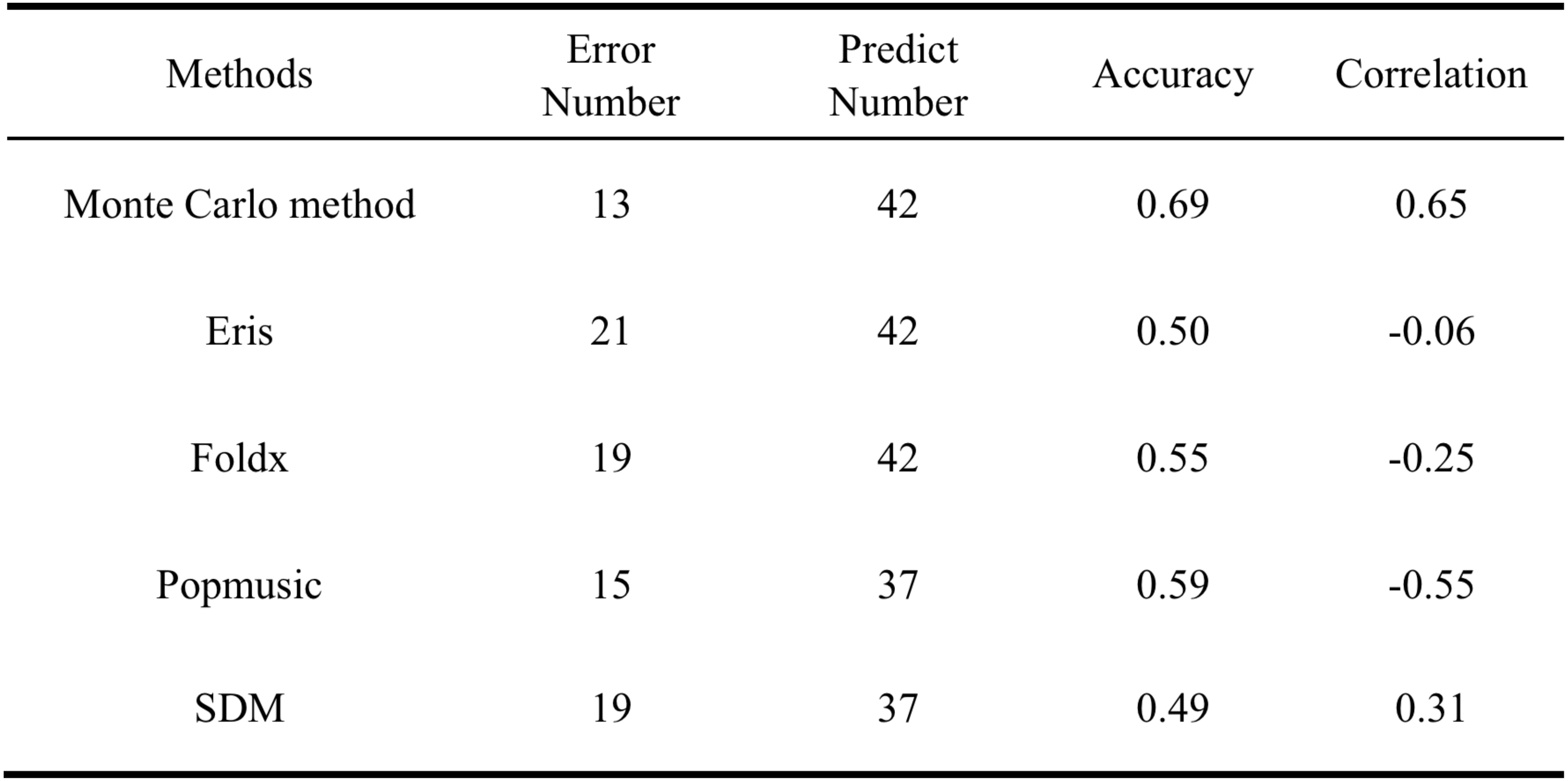
Unfolding Curves of other proteins



Each protein was simulated in 50 replications for 2,000,000 MC steps.

The points were the average RMSD or energy in the last 1,000,000 steps in the 50 replications. The Tm values were simulated the sigmoid function.

Comparison with other methods



Methods	Error Number	Predict Number	Accuracy	Correlation
Monte Carlo method	13	42	0.69	0.65
Eris	21	42	0.50	-0.06
Foldx	19	42	0.55	-0.25
Popmusic	15	37	0.59	-0.55
SDM	19	37	0.49	0.31

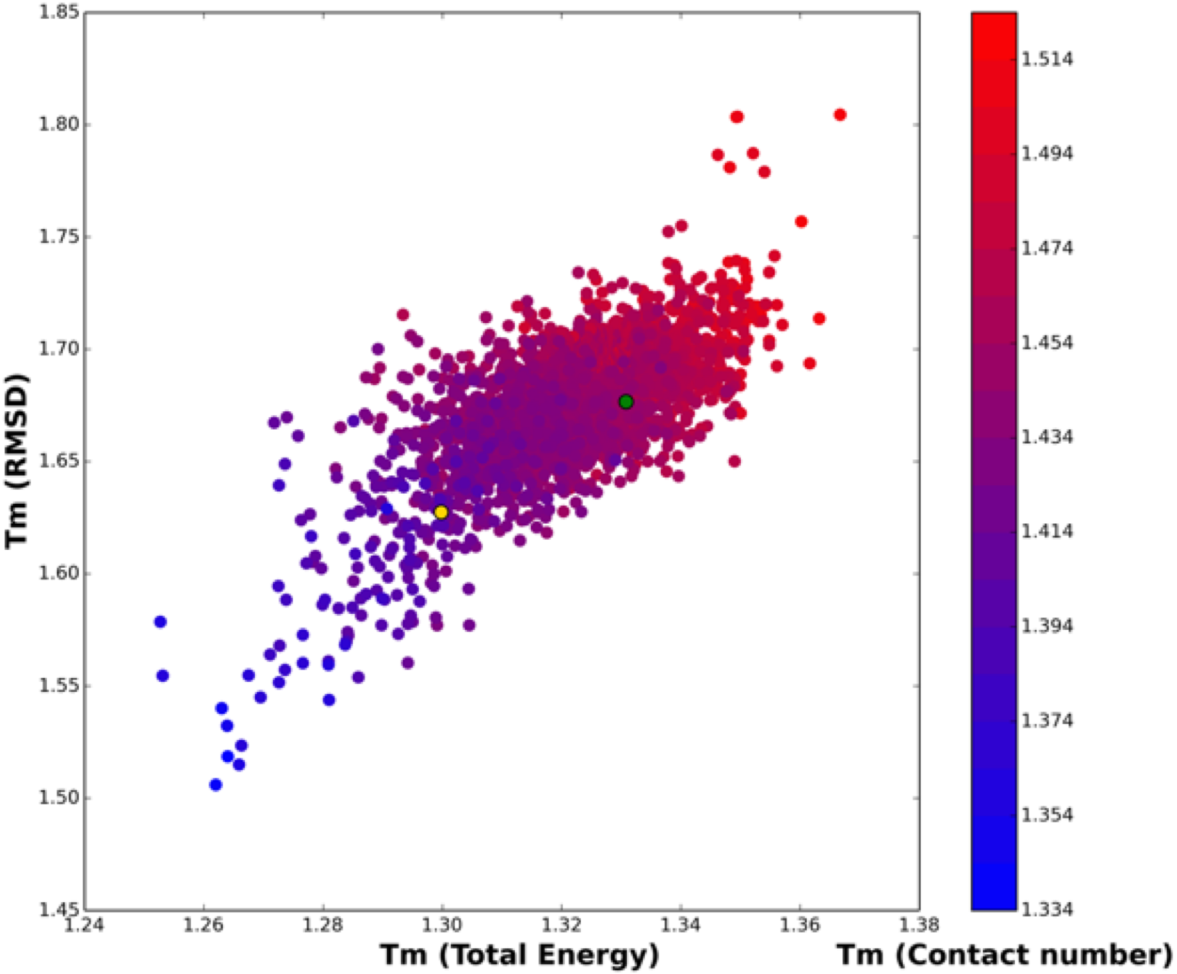
Foldx: Guerois R, Nielsen JE, Serrano L (2002). J Mol Biol 320: 369-387.

Popmusic: Dehouck Y, Grosfils A, Folch B, Gilis D, et al. (2009). Bioinformatics 25: 2537-2543.

SDM: Worth CL, Preissner R, Blundell TL (2011). Nucleic Acids Res 39: W215-222.

Eris: Yin S, Ding F, Dokholyan NV (2007) Nature Methods 4: 466-467.

Unfolding simulation of all DHFR mutations



Mutation number:

159*19=3,021

Positive mutation: 570

Positive

rate :570/3021=18.9%

Green: WT(T_m=51.7°C)

Yellow: I155A(T_m=38.5°C)

Every mutation was simulated

in 7 replications with

1,000,000 MC steps.

The last 500,000 MC step data

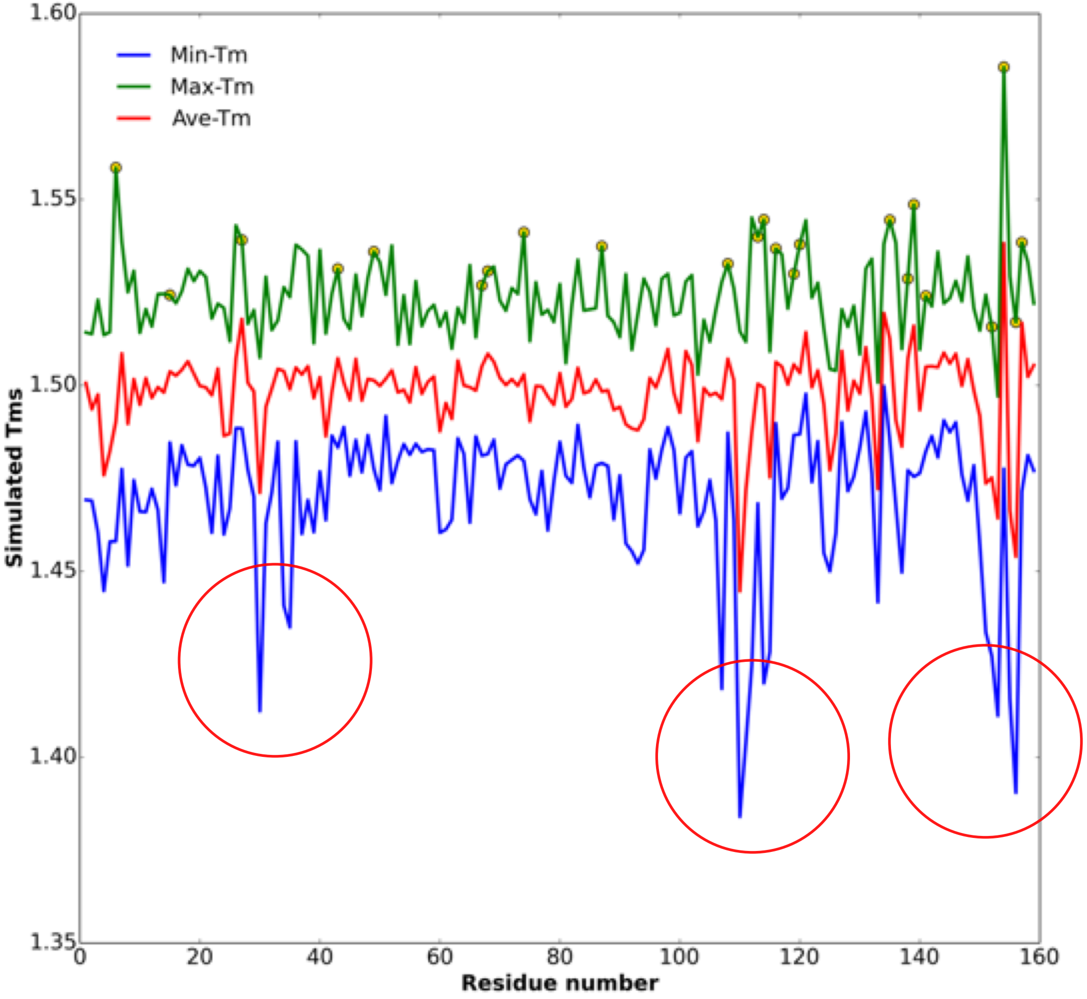
were used to simulate the Tm.

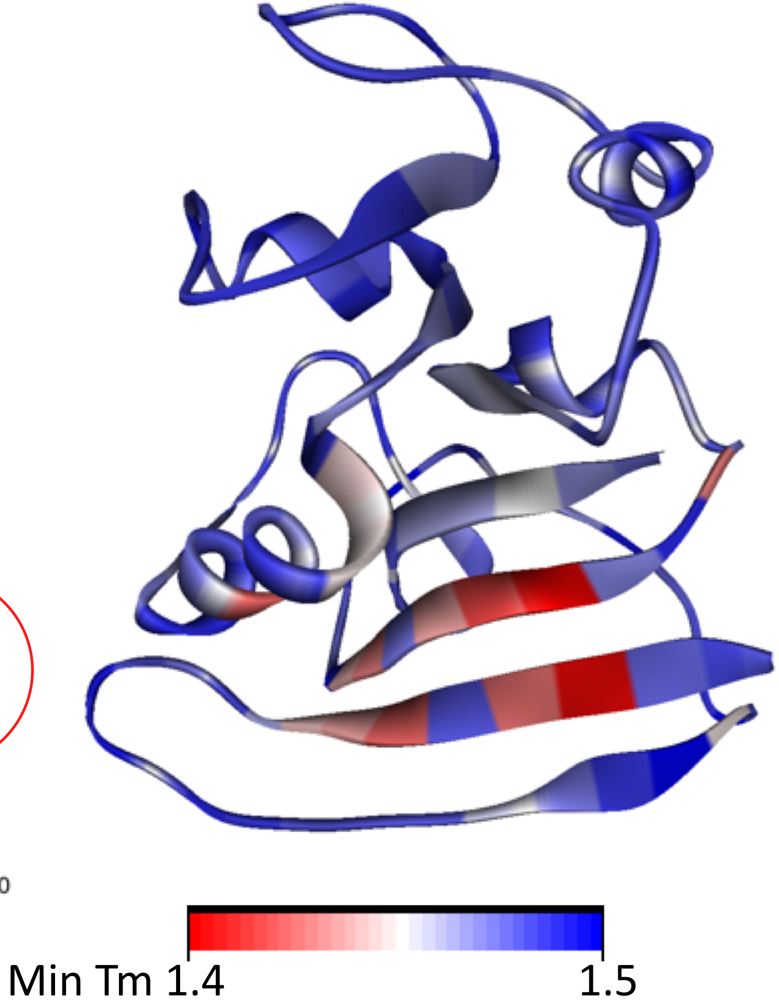
Experimental Results - single point mutations

Mutation	*Tm*(DSC)	Cm(CD)	kcat	kcat/Km	Mutation	Tm(DSC)	Cm(CD)	kcat	kcat/Km
wt	54.1	3.09	24.60	14.07	H114R	54.1	3.07	28.31	14.06
D27F	61.7	4.55	N.D.	N.D.	S49E	53.5	2.89	10.55	5.24
T113V	58.0	3.28	13.67	10.86	H141F	53.0	2.94	12.07	6.00
Q108D	55.7	3.18	24.60	10.35	E157F	52.4	3.07	29.07	14.45
S138Y	55.6	3.33	24.51	9.33	G15W	52.3	3.07	10.55	5.24
D116F	55.5	3.43	24.80	9.53	L156Y	51.3	2.62	6.02	3.00
T68N	55.5	3.26	29.36	13.32	E139V	51.3	2.73	24.80	12.31
E120P	55.3	3.25	30.02	13.91	D87P	51.0	2.88	25.73	13.18
V119F	54.9	3.12	28.50	12.57	G43P	51.0	2.83	10.07	5.02
S135I	54.8	3.33	33.35	16.66	W74F	50.5	2.96	3.44	1.71
C152I	54.2	3.15	22.99	11.44	G67H	48.1	2.65	17.20	8.57
					A6I	47.2	3.05	19.66	9.79

Units: T_m: °C , C_m: M, k_{cat}: s^{-1}, k_{cat}/K_M : s^{-1} μM^{-1}

T_m value of the residues in DHFR

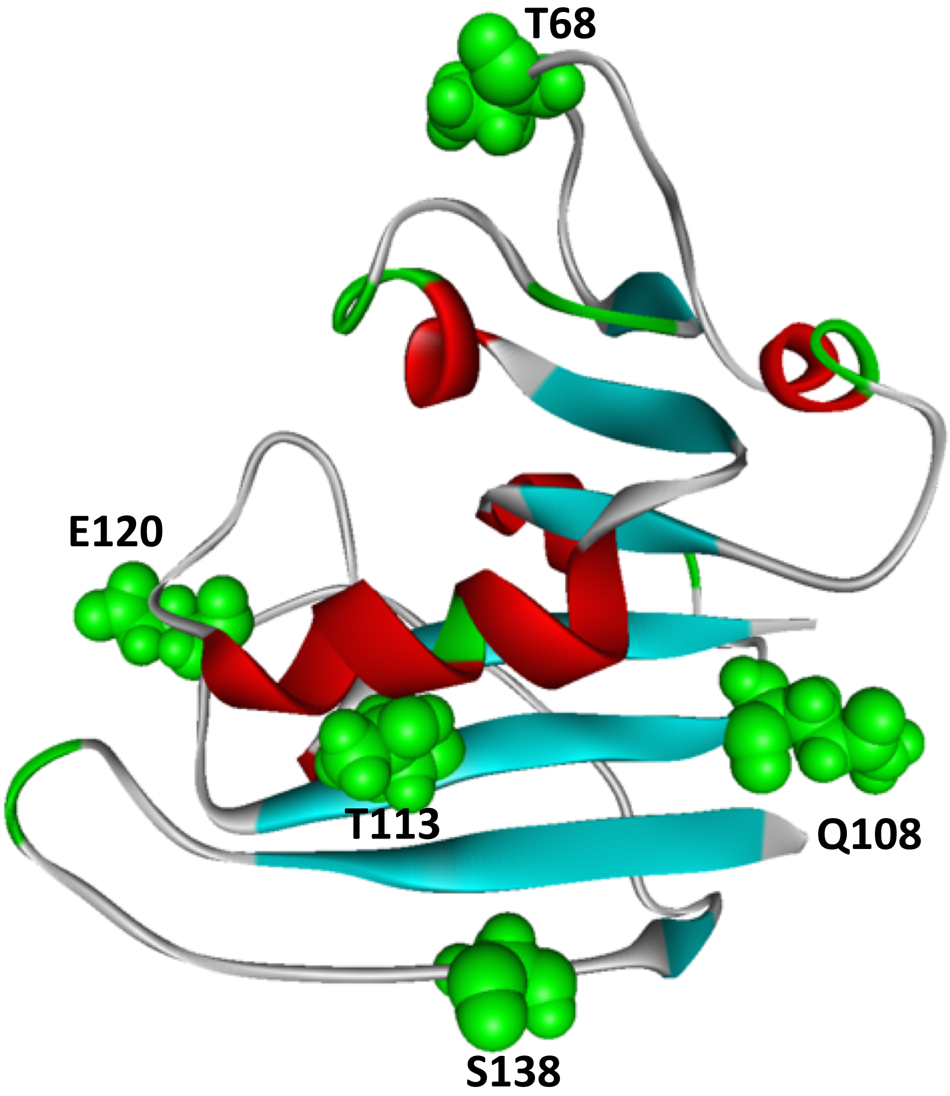




The residues with low min-Tm is very important for its structure stability.

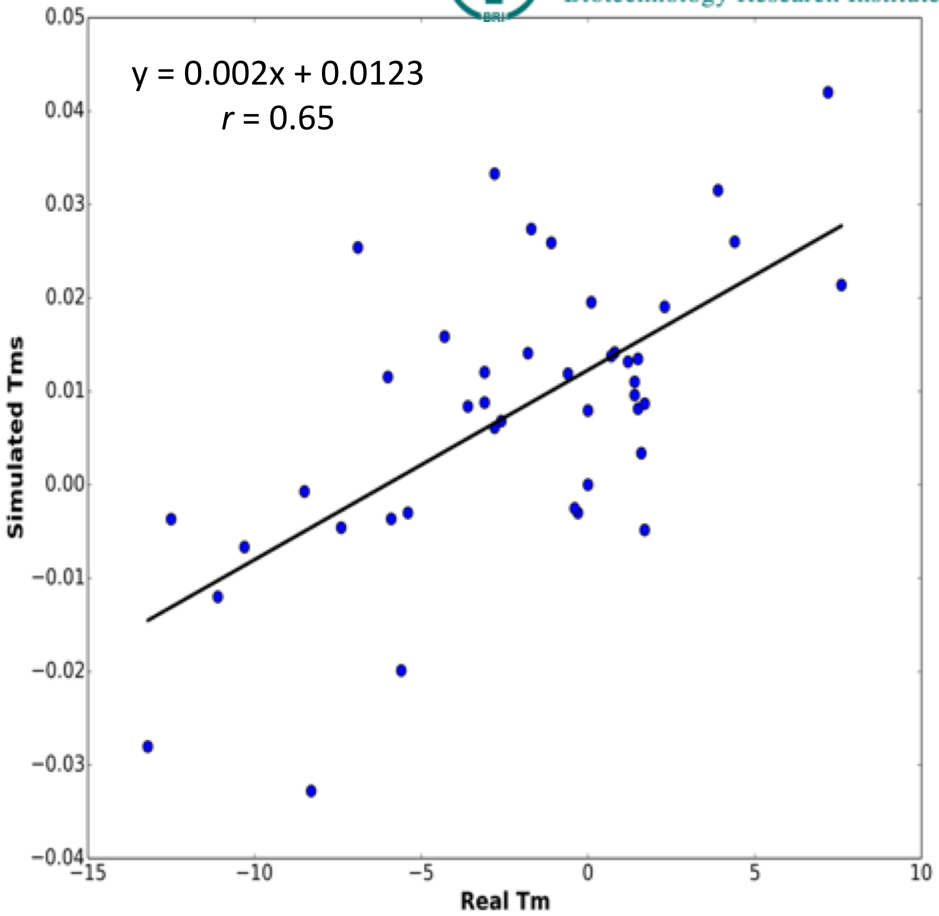
Experimental Results – Multiple point mutations

Mutation	Tm(DSC)	Cm(CD)	k_{cat}	k_{cat}/K_m
T68N, Q108D, T113V, E120P, S138Y	61.3	3.52	32.63	12.20
T113V, E120P, S138Y	58.5	3.49	31.13	13.10
T68N, Q108D, E120P, S138Y	56.4	3.47	22.80	10.94
T68N, Q108D	55.8	3.14	17.99	15.24
E120P, S138Y	55.6	3.29	16.01	10.81



Units: T_m: °C , C_m: M, kcat: (s^{-1}), k_{cat}/K_M : s^{-1} μM^{-1}

No	Mutation	Real-Tm	Simulated Tm
1	I155T	-8.3	1.343
2	V75H	-11.1	1.365
3	I155A	-13.2	1.343
4	I115A	-5.6	1.354
5	I91L	-10.3	1.367
6	I155L	-5.9	1.371
7	W133F	-5.4	1.371
8	V88I	-7.4	1.368
9	I61V	1.7	1.371
10	V40A	-8.5	1.375
11	I115V	-0.3	1.371
12	V75I	-12.5	1.372
13	A145T	-0.4	1.372
14	WT	0.0	1.375
15	I91V	-2.6	1.380
16	L112V	-4.3	1.395
17	D27F	7.6	1.393
18	T113V	3.9	1.405
19	Q108D	1.6	1.377
20	S138Y	1.5	1.382
21	D116F	1.4	1.386
22	T68N	1.4	1.383
23	E120P	1.2	1.388
24	V119F	0.8	1.386
25	S135I	0.7	1.387
26	C152I	0.1	1.393
27	H114R	0.0	1.378
28	S49E	-0.6	1.384
29	H141F	-1.1	1.400
30	E157F	-1.7	1.402
31	G15W	-1.8	1.387
32	L156Y	-2.8	1.379
33	E139V	-2.8	1.409
34	D87P	-3.1	1.381
35	G43P	-3.1	1.387
36	W74F	-3.6	1.382
37	G67H	-6.0	1.385
38	A6I	-6.9	1.401
39	T68N_Q108D_T113V_E120P_S138Y	7.2	1.420
40	T113V_E120P_S138Y	4.4	1.407
41	T68N_Q108D_E120P_S138Y	2.3	1.397
42	T68N_Q108D	1.7	1.380
43	E120P_S138Y	1.5	1.389
r			0.645



Each protein was simulated with 2,000,000 MC steps. The replication number is 50. The last 1,000,000 step data were used to simulated T_m.

小结

- **Monte Carlo protein unfolding method (MCPUM)**

  Simulate the protein unfolding.

  Evaluate the stability of mutations.

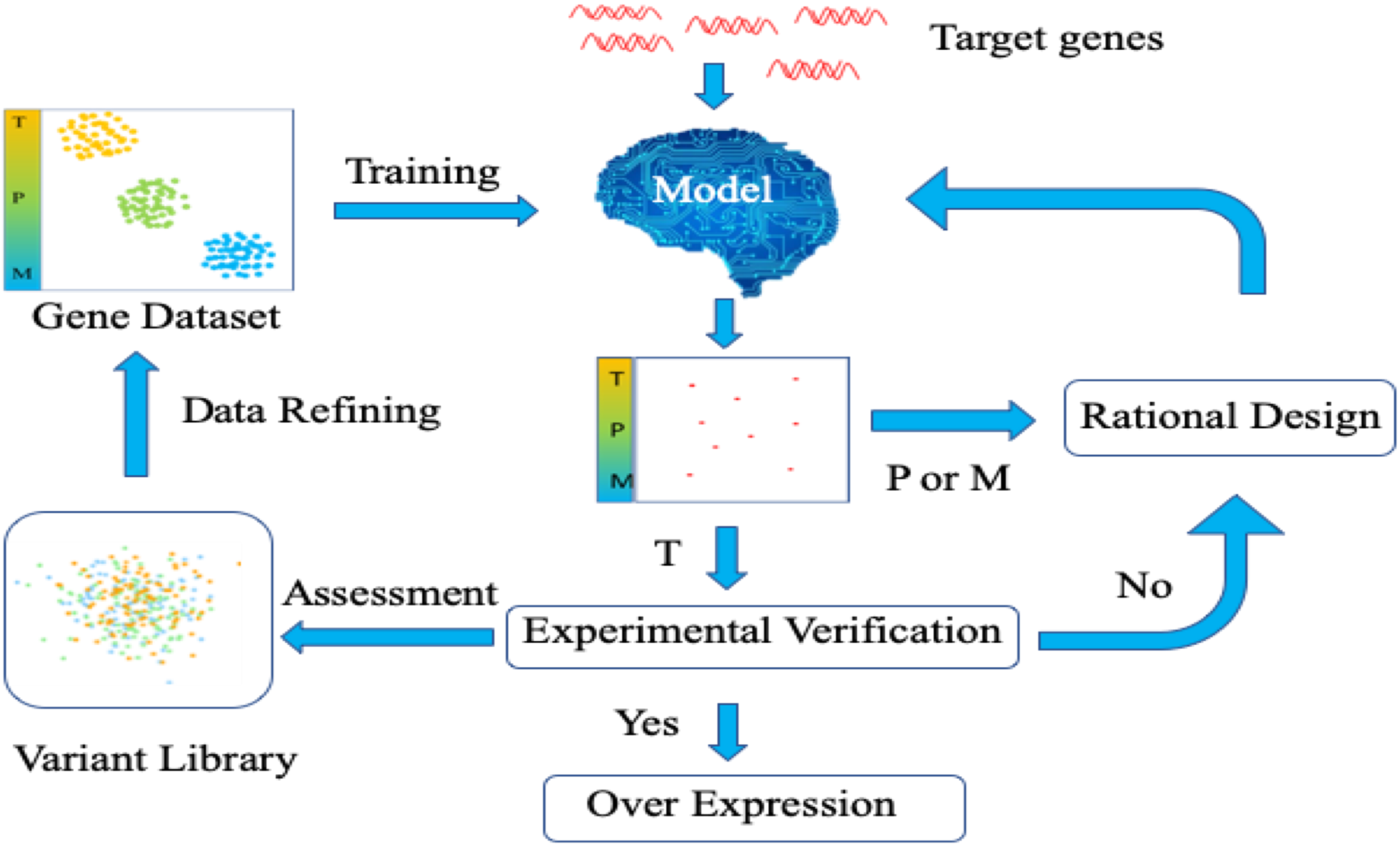
  Find the key contact of the protein structure.

  Calculate the protein unfolding free energy.

- **The simulation performance may be improved by optimizing the energy function, temperature function, and so on.**

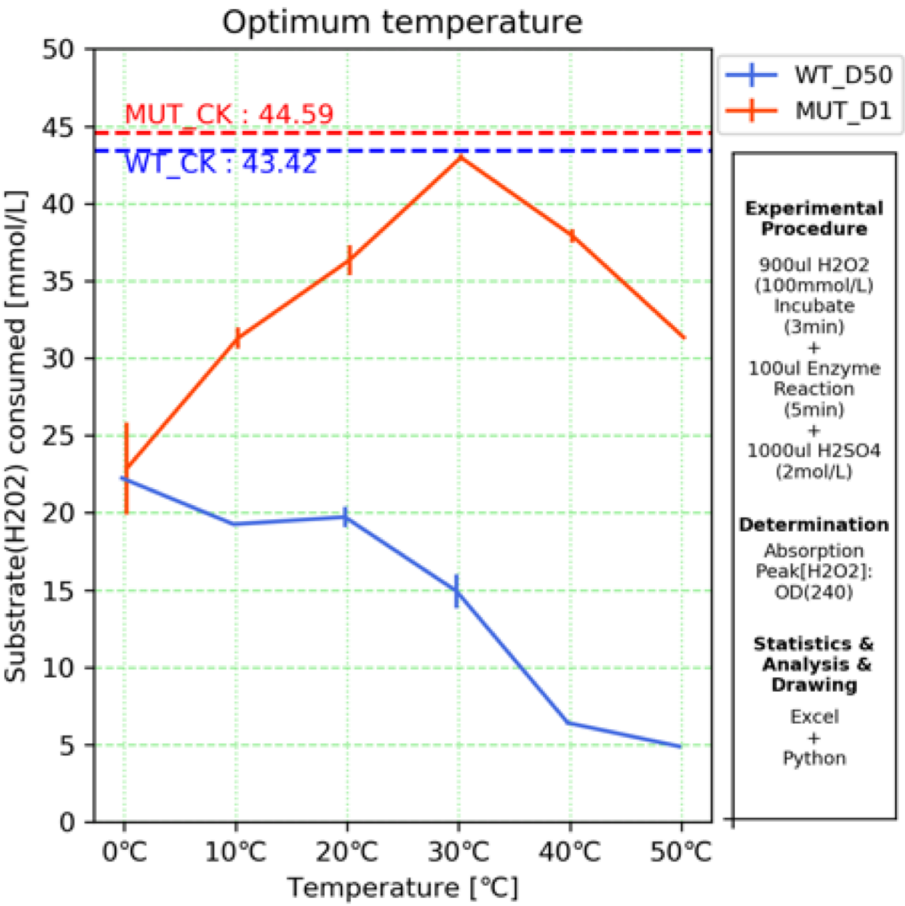
Tian, Jian et al, PLoS Computational Biology (2015).

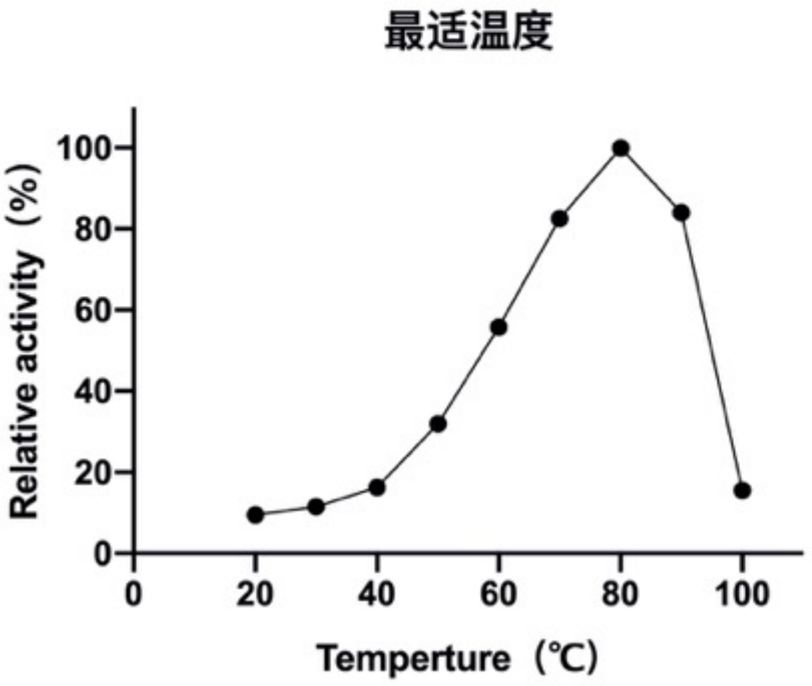
基于AI的蛋白稳定性的学习算法



蛋白质表达数据来自于 NCBI，包括456个不同的物种，基因数量1448174条，预测准确率76%

基于AI的蛋白稳定性算法的验证实验





Unpublished Data, 2020

基于该算法，成功从数据库中挖掘到低温过氧化氢酶（最适温度为0度），高温几丁质酶（最适温度80度），目前正在用该算法来设计定点突变体对蛋白稳定性影响的工作。

Deep Mind

Alphago

Alphafold

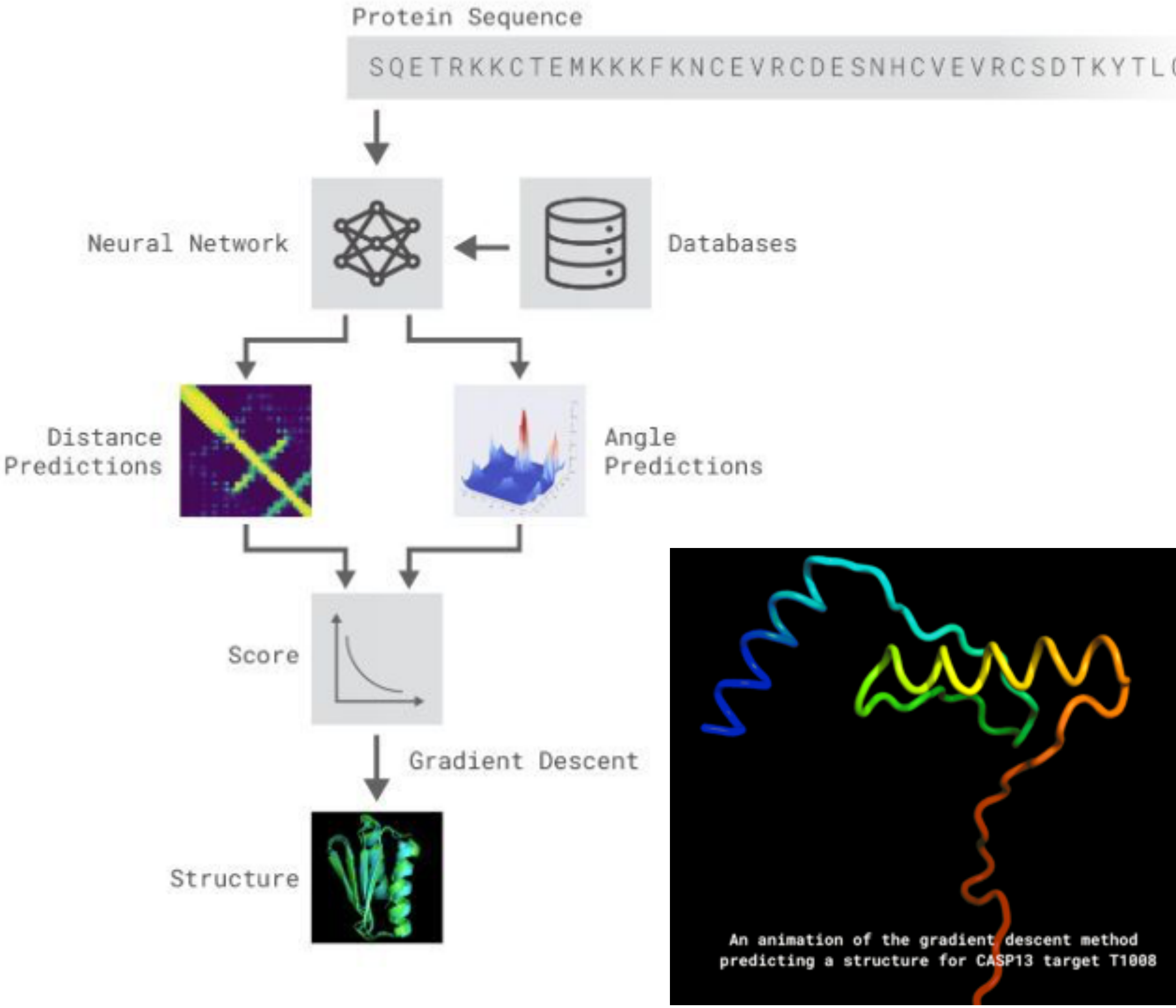
……

AlphaMutation

AlphaGene

AlphaProtein

……



An animation of the gradient descent method predicting a structure for CASP13 target T1008

致 谢

- 范云六院士
- 伍宁丰 研究员、张伟 研究员、姚斌 研究员、谷晓峰 研究员
- 初晓宇博士、刘晓青博士、关菲菲博士
- 哈佛大学 **Eugene Shakhnovich**教授、**Jaie Woodard**博士。
- 研究生：闫亚茹、秦伟彤、李庆宾、王平

NSFC
National Natural Science
Foundation of China

中华人民共和国
科学技术部

中国农业科学院
CAAS

谢谢

THANKS

田健

Email : tianjian@caas.cn

电话 : 010-82106354