Protein Design-Introduction and Datasets

介绍-中文 2

1. 目标、规则、提交内容 2

2. 数据介绍 3

Introduction-English 4

1.Objective, Guidelines, and Submission Content 4

2. Data Introduction 5

参考 /Reference 6

1.数据来源 /Data sources 7

2.相关研究 /Some papers in the field of protein engineering 7

介绍-中文

1. 目标、规则、提交内容

目标：您需要开发一个 GFP 的荧光强度预测模型，该模型能够预测哪些氨基酸**突变序列**可能具有最高的荧光强度。

1.1 提交内容

- 您需要对原始 **avGFP** 氨基酸序列进行改造，您可自由选最多 **6 个**不同位置上的氨基酸进行突变。
- 改造后**序列**：您需提交 **10 条**突变后的 avGFP 氨基酸序列，请确保提交的突变序列**未在**“GFP data.xlsx”数据集中出现，并以 Excel 格式提供（如下图）

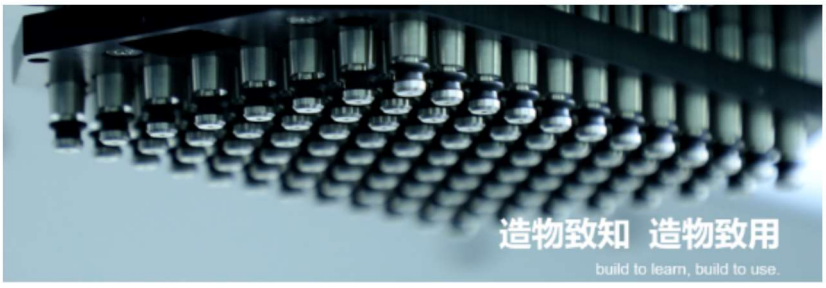
	A	B
1	Submitted sequences	
2	MSKGEELFTGVVPILVELDGDVNGHKFSSSGEGEGDATY(	
3	MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATY(	
4	MSKGEELFTGPPPPLVELDGDVNGHKFSVSGEGEGDATY	
5	MSKGLLLFTGVVPILVELDGDVNGHKFSVSGEGEGDATY(	
6	MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATY(	
7	MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATY(	
8	.....	

> >avGFP 氨基酸序列
> MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLSYGVQCFSRY
> PDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSH
> NVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLE
> FVTAAGITHGMDELYK

- **代码**：您设计的**模型代码**的压缩包（包括：ipynb 格式的代码及其补充文件）
- **思路**：您的代码或者队伍设计筛选的**思路**简介（word）

1.2 实验流程

- 合成序列：我们将借助自动化大设施合成您提交的 10 条 GFP 序列。



- 检测结果：使用 **400~420 nm** 波长激发荧光，在 **510~530 nm** 波长检测 GFP 发出的最亮荧光，并记录结果，**荧光强度**作为得分依据。
- 统计反馈：结合得分规则和提交的内容，给出最后的队伍排名

1.3 得分规则

在所有的参赛序列中，排名前0.5%的序列每个序列得 5 分，排名前0.6%-2%的序列每个序列得 1 分，排名前2.1-10%的序列每个序列得 0.1 分，其他序列不得分。

1.4 奖项

奖项 1：对参赛队伍的所有 10 条序列进行评分并加和，总分第 1 者获胜

奖项 2：在前0.5%的序列中占比最高的队伍获胜

2. 数据介绍

2.1 GFP 数据（GFP data.xlsx）

GFP data.xlsx
5.2 MB

查看

- 为协助您更有效地构建预测模型，我们提供了四种 GFP 蛋白（avGFP、amacGFP、cgreGFP、pluGFP2）的数据集，包含突变信息与荧光强度关系，以便模型学习不同 GFP 间的相似性并提高预测精度。
- 数据包括：aaMutations（氨基酸突变位点），GFP type（GFP 种类），Brightness（荧光亮度）。

aaMutations	GFP type	Brightness
WT	avGFP	3.719212132
A109D	avGFP	1.301030004
A109D:N145D:I187V:M232T:L235P	avGFP	1.30103124
A109D:Y142N:H147L:E221G	avGFP	1.301189044
A109G	avGFP	3.708478076
A109G:K139M:R167C:L235P	avGFP	3.58276407
A109G:K155E:F164S:L193Q:L194P	avGFP	1.49957272
A109G:K157R	avGFP	3.659012971
A109G:K157R:I160V:I187V:T224S	avGFP	3.573854919

aaMutations：突变位点和氨基酸变化的描述，例如 G101A 表示第 101 位的 G>A 替换，WT 表示野生型序列。

GFP type：表示 GFP 的种类，总共四种，来源于不同物种/途径。不同种类的 GFP 的氨基酸序列会有区别。

Brightness：具有相同核苷酸基因型的荧光对数亮度值的中位数。

2.2 蛋白质的氨基酸序列（AAseqs of 4 GFP proteins.txt）

AAseqs of 4 GFP proteins.txt
978.0 B

查看

- 包含 4 种 GFP 蛋白质（avGFP、amacGFP、cgreGFP、pluGFP2）的完整**氨基酸序列**。

2.3 蛋白质结构 PDB（GFP Protein structures.zip）

- 包括 4 种 GFP 蛋白质的三维结构，为 PDB 格式

Introduction-English

1. Objective, Guidelines, and Submission Content

Objective: Develop a predictive model for the fluorescent intensity of GFP that is capable of forecasting which amino acid mutant sequences are likely to exhibit the **highest** fluorescence intensities.

1.1 Submission Content

- You are required to engineer the original **avGFP** amino acid sequence by selecting up to **six** different positions for amino acid mutations freely.
- Modified **Sequences**: **Ten** mutated avGFP amino acid sequences must be submitted, ensuring that these variant sequences have **not** been previously reported in the "GFP data.xlsx" dataset. Submissions should be provided in Excel format (as illustrated below).

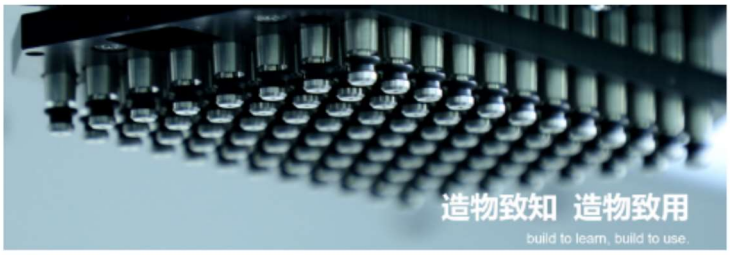
	A	B
1	Submitted sequences	
2	MSKGEELFTGVVPILVELDGDVNGHKFSSSGEGEGDATYG	
3	MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYG	
4	MSKGEELFTGPPPPLVELDGDVNGHKFSVSGEGEGDATY	
5	MSKGLLLFTGVVPILVELDGDVNGHKFSVSGEGEGDATYG	
6	MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYG	
7	MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYG	
8	.....	

> >avGFP amino acid sequence
> MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLSYGVQCFSRY
> PDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSH
> NVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLE
> FVTAAGITHGMDELYK

- **Code**: code zip package for the model your design
  (including: ipynb format code and its supplementary files)
- **Idea**: code or team design method **idea** brief (word)

1.2 Experimental Process

- Synthesize Sequences: We will use automated large-scale facilities to synthesize the 10 GFP sequences you submit.



- Detection Results: Fluorescence is excited at wavelengths of **400~420** nm and detected at wavelengths of **510~530** nm. The brightest fluorescence emitted by GFP is recorded, and fluorescence intensity serves as the basis for scoring.
- Statistical Feedback: Combining the scoring rules and submitted content, the final team rankings will be announced.

1.3 Scoring Rules In all participating sequences, those ranked in the top 0.5% will receive 5 points per sequence, those ranked from 0.6% to 2% will receive 1 point per sequence, and those ranked from 2.1% to 10% will receive 0.1 point per sequence. Other sequences will not receive points.

1.4 Awards

1. Award 1: Score all 10 sequences submitted by each team and sum them up. The team with the highest total score wins.
2. Award 2: The team with the highest proportion of sequences in the top 0.5%.

2. Data Introduction

GFP Data (GFP data.xlsx)

- To assist you in constructing prediction models more efficiently, we provide datasets of four GFP proteins (avGFP, amacGFP, cgreGFP, pluGFP2) containing mutation information in relation to the fluorescence intensity in order for the models to learn the similarities between different GFPs and improve prediction accuracy.
- Data include: aaMutations (amino acid mutation sites), GFP type (type of GFP), Brightness (log median brightness value of fluorescence for the same nucleotide genotype).
    - *aaMutations:*describes mutation sites and amino acid changes, for example, G101A indicates a G>A replacement at position 101, WT indicates the wild-type sequence.

- *GFP type:*represents the type of GFP, totaling four kinds, originating from different species/routes. The amino acid sequences of different GFPs will vary.
- *Brightness:*is the log median brightness value of fluorescence for identical nucleotide genotypes.

aaMutations	GFP type	Brightness
WT	avGFP	3.719212132
A109D	avGFP	1.301030004
A109D:N145D:I187V:M232T:L235P	avGFP	1.30103124
A109D:Y142N:H147L:E221G	avGFP	1.301189044
A109G	avGFP	3.708478076
A109G:K139M:R167C:L235P	avGFP	3.58276407
A109G:K155E:F164S:L193Q:L194P	avGFP	1.49957272
A109G:K157R	avGFP	3.659012971
A109G:K157R:I160V:I187V:T224S	avGFP	3.573854919

Amino Acid Sequences of 4 GFP Proteins (AAseqs of 4 GFP proteins.txt)

- Contains the complete amino acid sequences of 4 GFP proteins (avGFP, amacGFP, cgreGFP, pluGFP2).

Protein Structure PDB (GFP Protein structures.zip)

- Includes the three-dimensional structures of 4 types of GFP proteins, in PDB format.

参考 /Reference

1. 数据来源 /Data sources

Local fitness landscape of the green fluorescent protein

https://www.nature.com/articles/nature17995

Heterogeneity of the GFP fitness landscape and data-driven protein design

https://elifesciences.org/articles/75842

Wiki-GFP information

Green fluorescent protein - Wikipedia

2.相关研究 /Some papers in the field of protein engineering

Accurate proteome-wide missense variant effect prediction with AlphaMissense

https://www.science.org/doi/10.1126/science.adg7492

Learning protein fitness landscapes with deep mutational scanning data from multiple sources

https://www.sciencedirect.com/science/article/abs/pii/S2405471223002107

NBT (2024) Review: Machine learning for functional protein design

https://www.nature.com/articles/s41587-024-02127-0

Nat Method (2019) Machine-learning-guided directed evolution for protein engineering

https://www.nature.com/articles/s41592-019-0496-6