同源结构搜索工具 Foldseek简介

汇报人：邱嘉禾　G3成员：徐慧东、黄沁

2024-01-14

1.1 背景-作者

Martin Steinegger，首尔国立大学生物系助理教授，AF2（唯一非Deepmind作者）。

研究课题：开发数据密集型计算方法。

代表性方法：

①MMseqs2：搜索和聚类大量的蛋白质和核苷酸序列。

比BLAST快10000倍/比PSI-BLAST快400倍。

②ColabFold：方便的蛋白质结构预测。

将AlphaFold2和RoseTTAFold与MMseqs2结合起来，与AlphaFold系统相比，MSA生成速度提高了16倍。

③Foldseek：搜索和聚类蛋白质结构，与目前同源结构搜索工具相比快 10^4 - 10^5。

Steinegger, M. et al. *Nat Biotechnol* 35, 1026–1028 (2017).
Mirdita, M. et al. *Nat Methods* 19, 679–682 (2022).
van Kempen. et al. *Nat Biotechnol* (2023).

1.2 背景-同源搜索

同源搜索：将查询与数据库进行比较。搜索报告具有同源性的目标结果。应用：功能注释、结构预测、揭示进化关系等。

同源序列搜索：将蛋白质序列与序列数据库进行比较并计算统计显着性。

方法：BLAST、PSI-BLAST、HMMER、HHBlits、MMSeqs2。

限制：要求较高的高序列相似性。

同源结构搜索：将蛋白质结构与结构数据库进行比较并计算统计显着性。

方法：Dali、TM-align、FAST、Mammoth。

好处：与序列相比，蛋白质结构在进化过程中的保存时间更长，更保守。

限制：由于无预过滤算法，搜索速度慢。

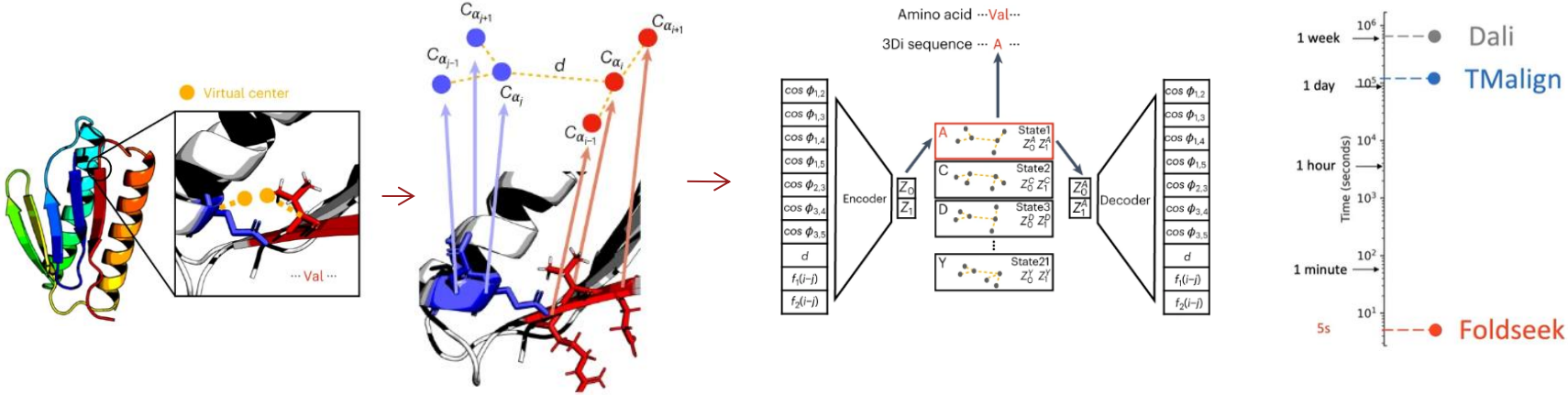
据估计：

数据库	方法	CPU	目的	时间消耗
1亿序列	MMseqs2	1K CPU core	All-versus-all 序列同源搜索	1周
1亿结构	TM-align	1 CPU core	单结构同源搜索	1月
1亿结构	TM-align	1K CPU core	All-versus-all 结构同源搜索	1000 年

2.1 Foldseek模型简介

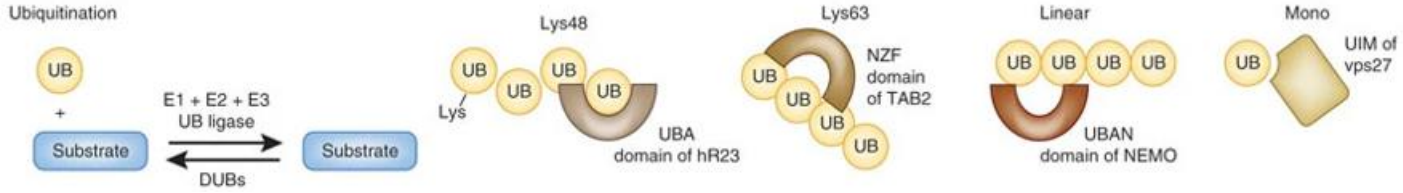
Highlight：开发了一种描述三级相互作用（3Di）的结构字母表，将结构转换为3Di序列，并使用序列比对（MMSeqs2）来比较结构。相比目前的结构比对方法将氨基酸主链描述为序列，能获得结构核心的高密度信息+状态频率更均匀。

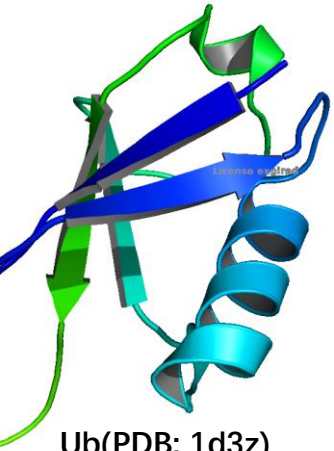
①寻找近邻残基　②提取3Di距离+角度+序列特征　③VQ-VAE训练保守3Di序列　对800k结构搜索SARS-Cov2的RdRp



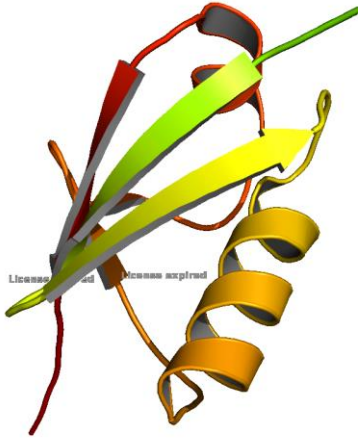
2.3 Foldseek使用例子1-Ub 和 SUMO

翻译后修饰系统中包含通过共价连接短修饰蛋白的修饰，通过泛素样蛋白 (UBL)如泛素(Ub)、小泛素样修饰物 (SUMO)介导 。泛素化通过E1E2E3三步完成修饰。Ub和SUMO通常具有显着序列同源性。





Ub(PDB: 1d3z)



SUMO(PDB: 1u4a)

需解决问题：部分Ub与SUMO具有高度结构相似性，但是在序列相似性较低。

Ub 和 SUMO EMBOSS Needle 结果：

Identity: 12/76 (15.8%)

Similarity: 33/76 (43.4%)

Deribe, Y., Pawson, T. & Dikic, I. *Nat Struct Mol Biol* 17, 666–672 (2010).

2.3 Foldseek使用例子1-Ub 和 SUMO Foldseek结果

搜索步骤：①输入PDB结构（也可用ESMFold/ColabFold对序列进行结构预测）。

②选择搜索数据库、搜索方式，分类限制等。

Foldseek

```
1d3z (3Di)  1 DKEWEAEPVGDIDIDDDDQADFPLNVLVVVCVVPVAPSVFKWKDFPRDTGDRVGGNVVVVADPHGYIYIDGDDPPD 76
              ||||||| | | | ||  || | |  ||||||      ||| || || ||    | |    |||||  ||  |
1u4a (3Di)  4 AWEWEAEPVRDTDTDTDDLQDFLLVVVCVVCVVPPHDPPWKWKAFPNDTDDRRDRVVSNVDRHDGYIYIYTDDDDD 79
```

Foldseek E-value = 3e-5

Sequence identity = 54% (41/76)

PDB100 1000 hits

GRAPHICAL | NUMERIC

Target	Description	Scientific Name	Prob.	Seq. Id.	E-Value	Position in query	Alignment
1u4a_MODEL_7_A	Solution structure of h...	Homo sapiens	1.00	16	4.37e-5	2 — 76	

```
Q  2 QIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG
     ++ V    G  + ++++    +  +     +++G+   Q R+ F G+ + +  T +   ++ E T+ +  +  GG
T  5 NLKVAGQDGSVVQFKIKRHTPLSKLMKAYSERQGLSMRQIRFRFDGQPINETDTPAQLEMEDEDTIDVFQQQTGG
```

Select target residues to highlight their structure

PDB | PNG

2.3 Foldseek使用例子2-人唾液酸酶与病毒神经氨酸酶

用于禽流感的药物奥司他韦的使用在部分亚洲地区与神经疾病和皮肤反应有关。

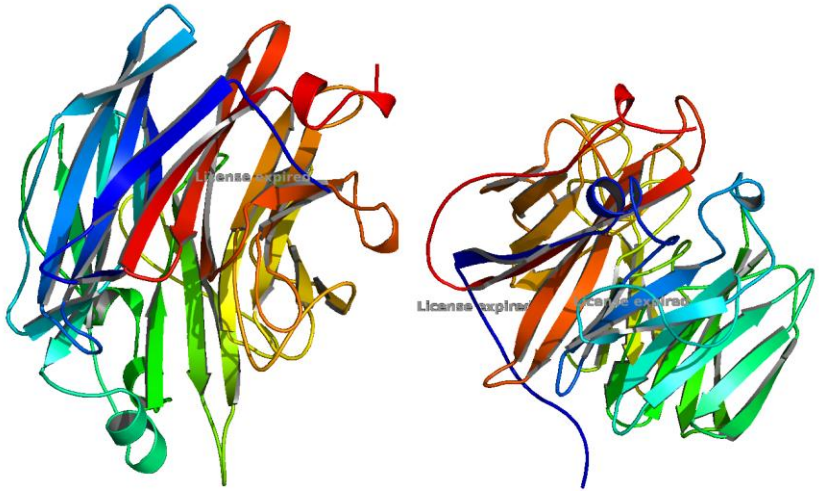
研究人员得到人唾液酸酶HsNEU2是奥司他韦靶点的流感病毒神经氨酸酶N2的同源物，并通过人胞质唾液酸酶非同义SNP R41Q（单核苷酸多态性，出现在9.29%的亚洲人口中）解释上述现象。

但是人唾液酸酶HsNEU2与流感病毒神经氨酸酶N2序列相似性很低，当时通过活性位点信息搜索得到。

Ub 和 SUMO EMBOSS Needle 结果

Identity: 21/751 (2.8%)

Similarity: 43/751 (5.7%)

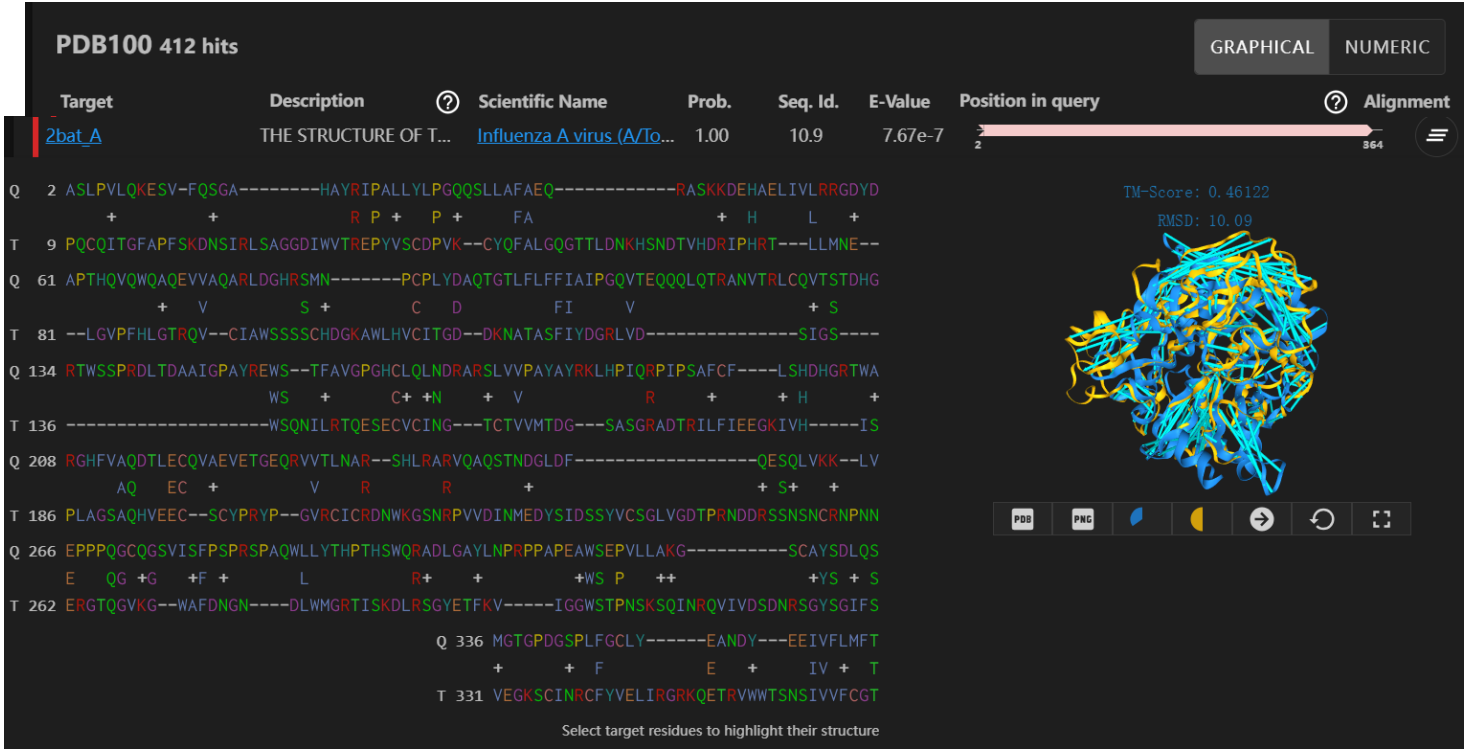


HsNEU2(PDB: 1VCU) N2(PDB: 2BAT)

Li, CY., Yu, Q., Ye, ZQ. et al. Cell Res 17, 357–362 (2007).

2.3 Foldseek使用例子2 Foldseek结果

搜索人唾液酸酶HsNEU2(PDB: 1VCU)结构：结果中包含流感病毒神经氨酸酶N2(PDB: 2BAT)。



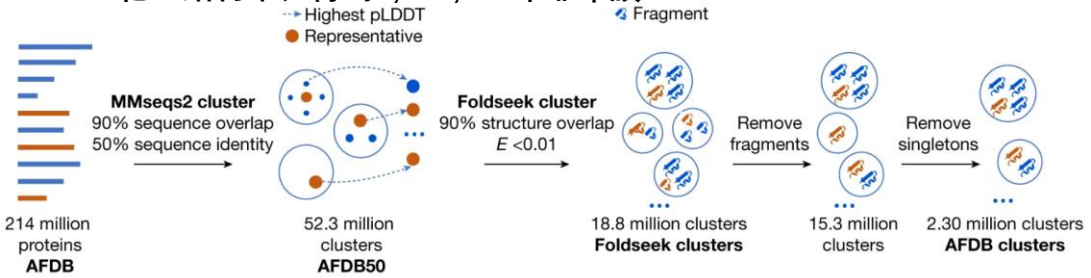
2.3 Foldseek使用例子2 Foldseek后续搜索

文章讨论部分建议对人类四种唾液酸酶都进行研究。

搜索流感病毒神经氨酸酶N2(PDB: 2BAT)结构，同时得到了人唾液酸酶HsNEU2、HsNEU3，可能酶HsNEU3可进行后续研究。

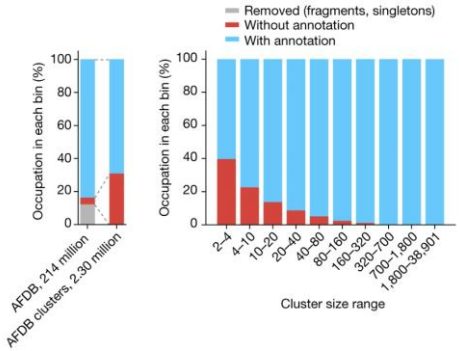
3.1 Foldseek Cluster-对目前蛋白质宇宙的聚类

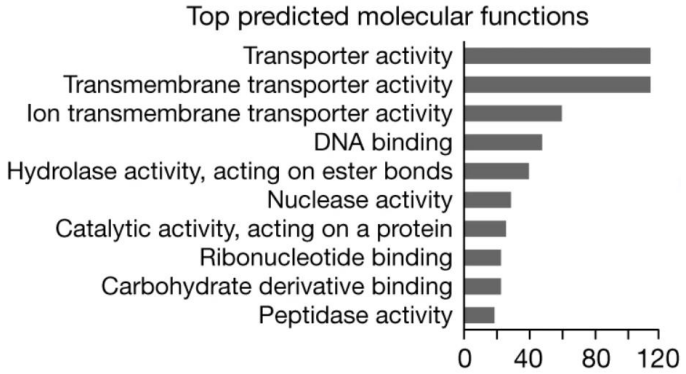
Foldseek Cluster工作流程：MMseqs2聚类（得AFDB50）+Foldseek聚类+去除片段和单簇。最终从AlphaFold UniProt v.3数据库的214,684,311（2.14亿）结构中，得到2,302,908个非单簇。



无注释的蛋白质（5%）占30%的簇（定义为暗簇）。

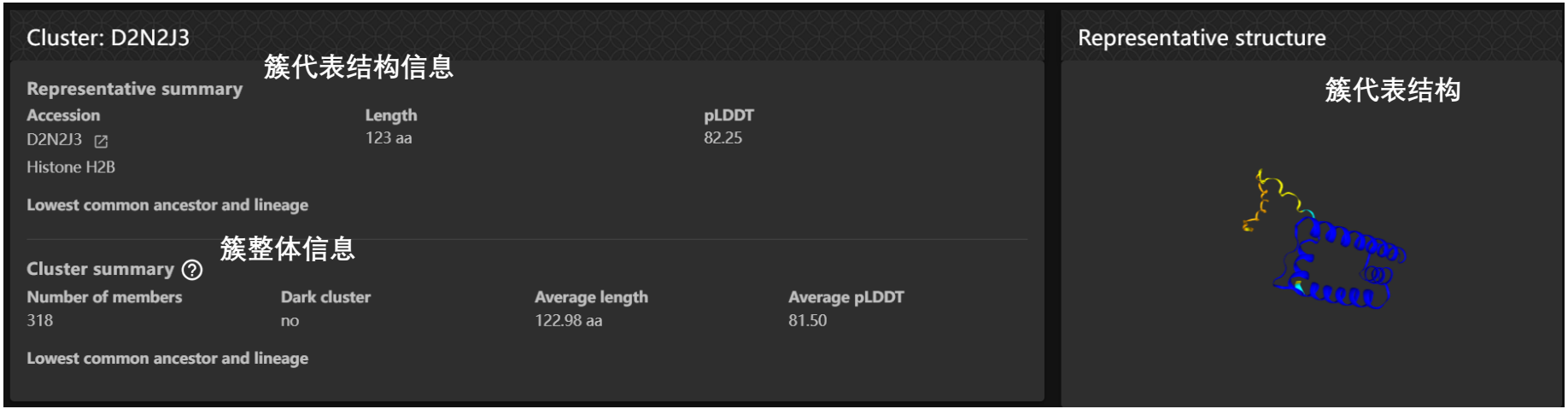
暗簇功能预测中最多的分子功能：跨膜转运蛋白活性。



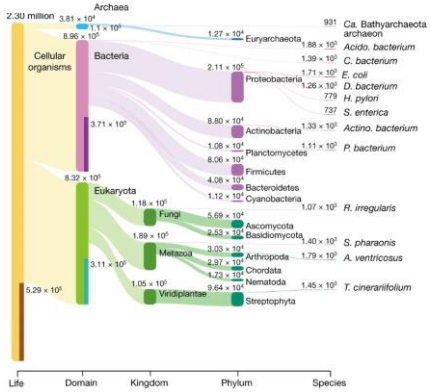


3.2.1 Foldseek Cluster 结果1-簇简要信息

以A0A1G5ASE0，一种细菌组蛋白为例

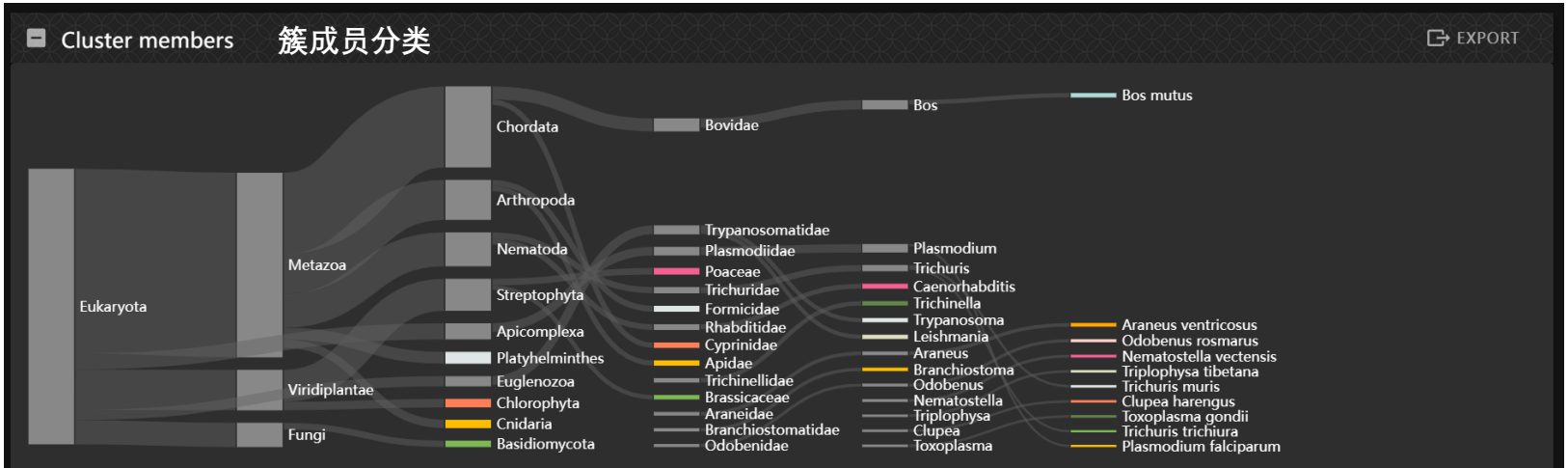


3.2.2 Foldseek Cluster 结果2-簇分类

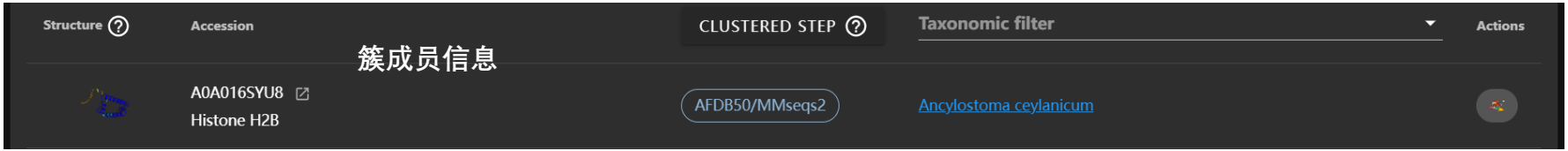


对非单簇分类并确定簇的LCA（最近共同祖先）。

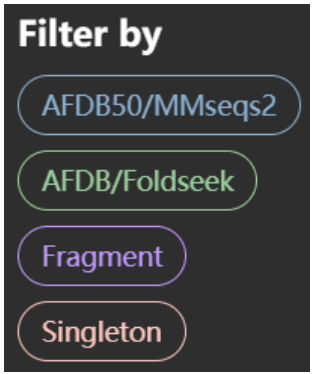
大多数簇的起源非常古老，4%物种特异性簇可能是基因从头诞生的例子。



3.2.3 结果3-簇成员（具同源结构）



簇成员信息



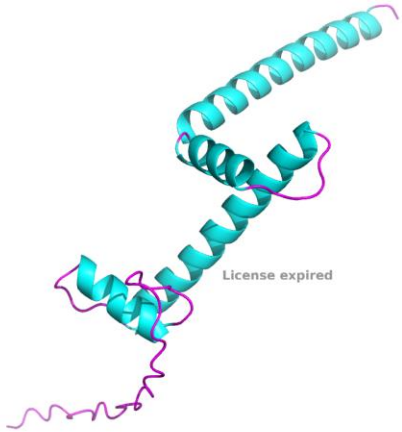
可通过聚类步骤进行筛选

①序列同一性50%聚类

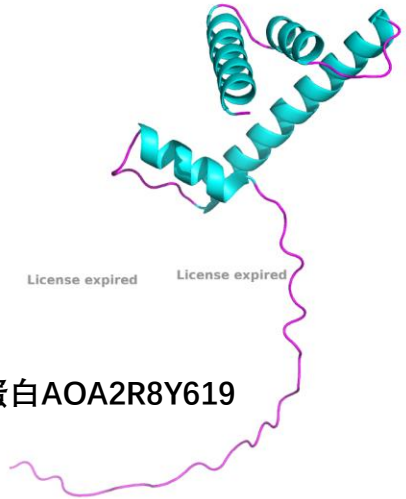
②结构相似性聚类

③删除的片段

④删除的单例

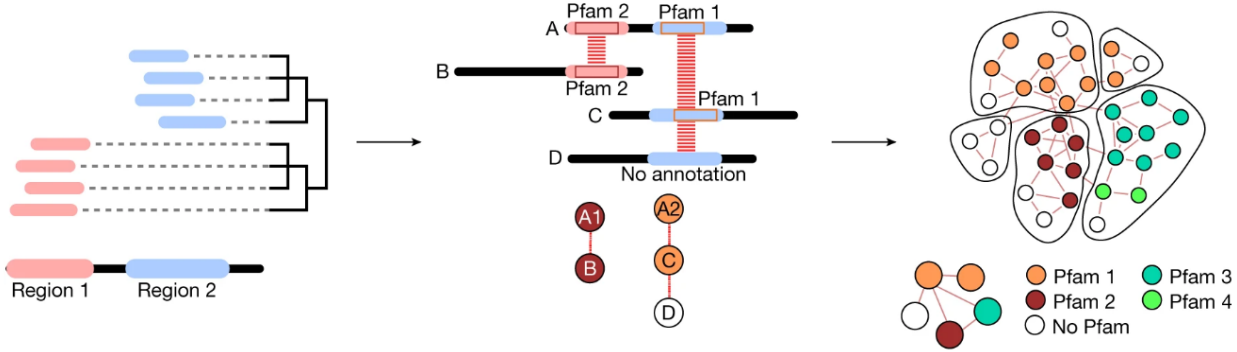


细菌组蛋白A0A1G5ASE0



人组蛋白AOA2R8Y619

3.2.4结果-相似簇（具相似Domain）



①搜索每个簇的代表性结构。

②对Hit的开始和结束位置聚类并定义可能的Domain边界。

③将具有结构相似性的Domain连接起来，并将其聚类成假定的Domain家族

Similar clusters 相似簇的代表结构 EXPORT

Structure	Accession	Average length	Average pLDDT	Number of members	Taxonomic filter	Dark cluster	Rep pLDDT	Rep length	E-value	Actions
	A0A0R3Q197 Histone H4	340.59	71.97	22	cellular organisms	0	74.50	357	7.108E-16	

4.总结

1.Foldseek(search.foldseek.com)

①使用描述三级相互作用（3Di）的结构字母表将结构转换为3Di序列，并使用序列比对（MMSeqs2）来搜索同源结构。

②与目前同源结构搜索工具相比快 10^4 - 10^5，并保持相似灵敏度。

③结果：查询结构对应的具同源结构蛋白。

2.Foldseek cluster(cluster.foldseek.com)

①对目前蛋白质宇宙的聚类，2.14亿个结构聚类得到230万个较高结构质量非单簇。

②结果：查询结构对应的目标簇的①簇简要信息②簇分类③簇成员（具同源结构）④相似簇（具相似Domain）。

3.课程感悟

①生物信息学，针对同一问题往往存在多种工具选择，须明辨工具特点和适用范围。

②ABC课程为我们提供了对这些工具的全面阐释，通过实践练习让我们理解如何将它们有效地融入到各自的科研工作中。

③Half day on the Web, saves you half month in the lab!

感谢罗老师的ABC课程！

感谢大家的聆听！