北京大学
PEKING UNIVERSITY

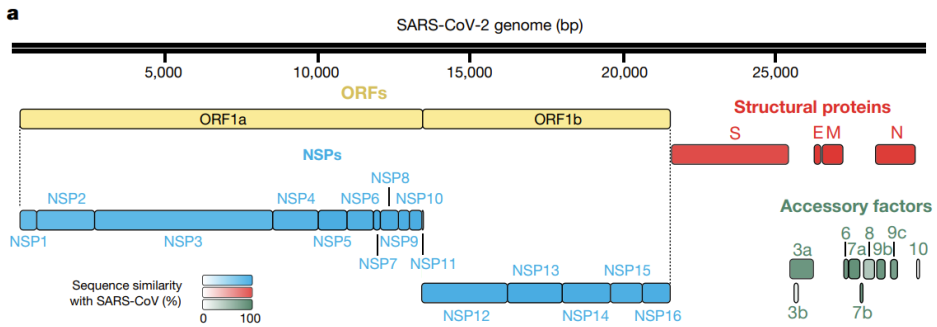
基于序列的新型冠状病毒的进化分析及蛋白质结构和相互作用网络预测

Sequence-based evolution analysis of SARS-CoV-2 and the prediction of both protein structure and interaction network

- 小组成员：邓启东 刘鋆 刘文清 林芮昀
- 报告人：刘文清

新冠病毒蛋白背景知识

Entry	Protein names	Gene names
P0DTC2	Spike glycoprotein	S 2
P0DTD1	Replicase polyprotein 1ab	rep 1a-1b
P0DTC1	Replicase polyprotein 1a (pp1a)	
P0DTC7	ORF7a protein	7a
P0DTC9	Nucleoprotein (N)	N
P0DTC3	ORF3a protein (ORF3a)	3a
P0DTC5	Membrane protein (M)	M
P0DTC4	Envelope small membrane protein (E)	E 4
P0DTC8	Non-structural protein 8	8
P0DTC6	Non-structural protein 6	6
P0DTD2	ORF9b protein (ORF9b) (ORF-9b)	9b
P0DTD8	ORF7b protein (ORF7b)	7b
P0DTF1	ORF3b protein (ORF3b)	
P0DTD3	ORF9c protein (ORF9c) (ORF14)	9c
P0DTG0	ORF3d protein	
P0DTG1	ORF3c protein (ORF3c) (ORF3h protein) (ORF3h)	



30kb的基因组
14个开放阅读框（ORF）
ORF1a和ORF1ab→编码多聚蛋白水解加工成16种非结构蛋白（NSP1–NSP16）
在病毒基因组的3'端，从9个预测的亚基因组RNA中表达多达13个开放阅读框
棘突蛋白（S）、包膜蛋白（E）、膜蛋白（M）和核衣壳蛋白（N）及九种辅助因子

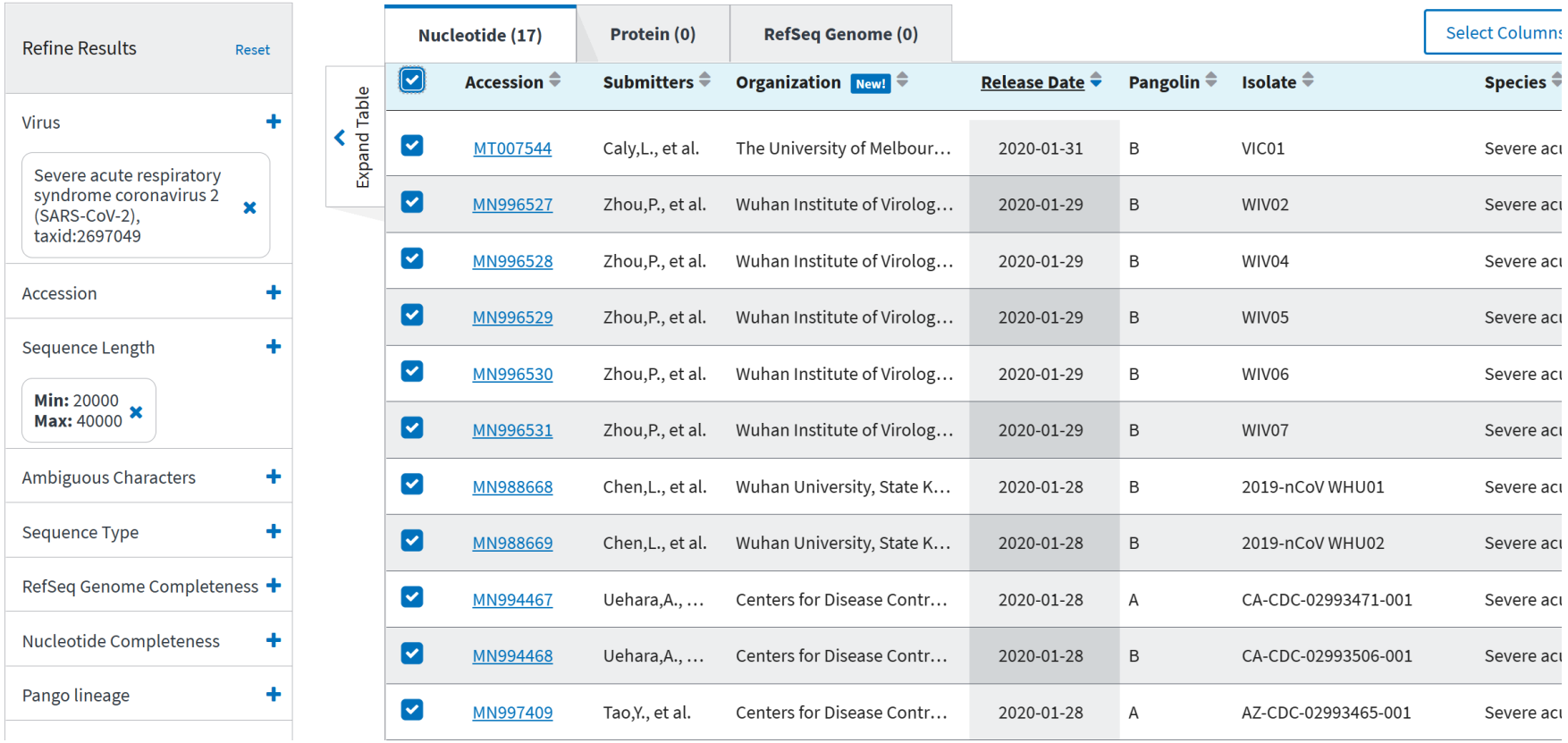
[1] Gordon D E , Jang G M , Bouhaddou M , et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing[J]. Nature, 2020, 583(7816):1-13.

01

Peking University

新冠病毒是否起源于武汉?

病毒基因组数据的收集--NCBI Virus

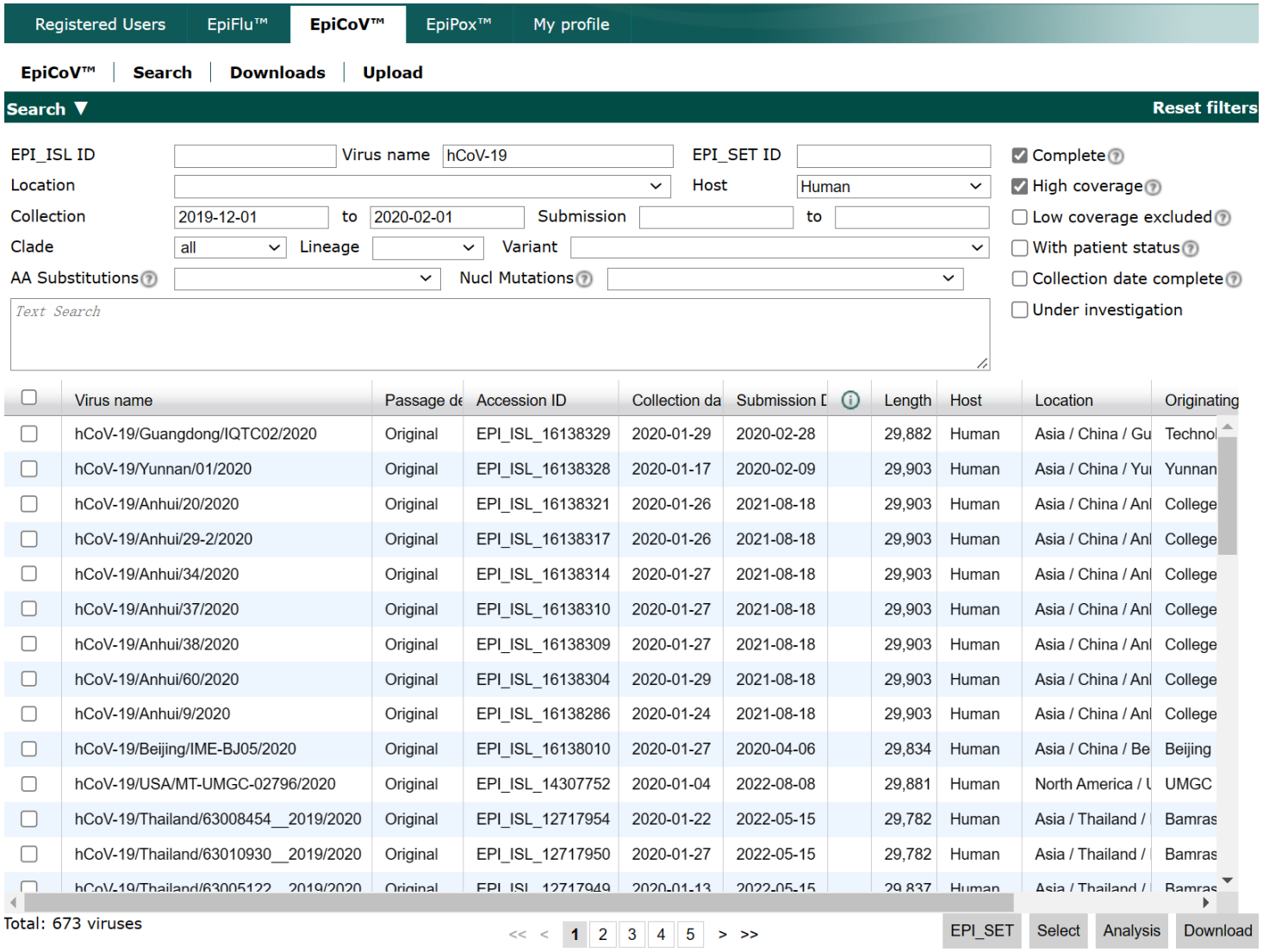


Accession	Submitters	Organization	Release Date	Pangolin	Isolate	Species
MT007544	Caly,L., et al.	The University of Melbour…	2020-01-31	B	VIC01	Severe acu
MN996527	Zhou,P., et al.	Wuhan Institute of Virolog…	2020-01-29	B	WIV02	Severe acu
MN996528	Zhou,P., et al.	Wuhan Institute of Virolog…	2020-01-29	B	WIV04	Severe acu
MN996529	Zhou,P., et al.	Wuhan Institute of Virolog…	2020-01-29	B	WIV05	Severe acu
MN996530	Zhou,P., et al.	Wuhan Institute of Virolog…	2020-01-29	B	WIV06	Severe acu
MN996531	Zhou,P., et al.	Wuhan Institute of Virolog…	2020-01-29	B	WIV07	Severe acu
MN988668	Chen,L., et al.	Wuhan University, State K…	2020-01-28	B	2019-nCoV WHU01	Severe acu
MN988669	Chen,L., et al.	Wuhan University, State K…	2020-01-28	B	2019-nCoV WHU02	Severe acu
MN994467	Uehara,A., …	Centers for Disease Contr…	2020-01-28	A	CA-CDC-02993471-001	Severe acu
MN994468	Uehara,A., …	Centers for Disease Contr…	2020-01-28	B	CA-CDC-02993506-001	Severe acu
MN997409	Tao,Y., et al.	Centers for Disease Contr…	2020-01-28	A	AZ-CDC-02993465-001	Severe acu

Virus: SARS-CoV-2 (taxid:2697049)

Sequence Length: 20000-40000 Release Date: 2019.8.31-2020.2.1

病毒基因组数据的收集--GISAID



多序列比对与单倍型的划分

多序列比对与单倍型的划分

Output of: LeiCZ_17_renameHT_MT.fas

DNA Polymorphism

```
Input Data File: D:\...\LeiCZ_17_renameHT_MT.fas
Number of sequences: 18    Number of sequences used: 18
Selected region: 1-29825   Number of sites: 29825
Total number of sites (excluding sites with gaps / missing data): 29795

Number of polymorphic (segregating) sites, S: 1152
Total number of mutations, Eta: 1154

Number of Haplotypes, h: 13
Haplotype (gene) di                    .902
   Variance of Haplotype diversity: 0.00441
      Standard Deviation of Haplotype diversity: 0.066

Nucleotide diversity, Pi: 0.00435
Theta (per site) from Eta: 0.01126

Theta (per site) from S, Theta-W: 0.01124
   Variance of theta (no recombination): 0.0000150
      Standard deviation of theta (no recombination): 0.00388
   Variance of theta (free recombination): 0.0000001
      Standard deviation of theta (free recombination): 0.00033

Finite Sites Model
   Theta (per site) from Pi: 0.00438
   Theta (per site) from S: 0.01151
   Theta (per site) from Eta: 0.01140

Average number of nucleotide differences, k: 129.693
   Stochastic variance of k (no recombination), Vst(k): 3023.605
   Sampling variance of k (no recombination), Vs(k): 384.938
      Total variance of k (no recombination), V(k): 3408.542
   Stochastic variance of k (free recombination), Vst(k): 43.231
   Sampling variance of k (free recombination), Vs(k): 5.086
      Total variance of k (free recombination), V(k): 48.317

Theta (per sequence) from S, Theta-W: 334.927
   Variance of theta (no recombination): 13359.806
   Variance of theta (free recombination): 97.375
```

```
[Hap#  Freq. Sequences]
[Hap_1:  6   NC_045512.2_|Severe_acute_respiratory_syndrome_coronavirus_2
[Hap_2:  1   MT007544.1_|Severe_acute_respiratory_syndrome_coronavirus_2_
[Hap_3:  1   MN997409.1_|Severe_acute_respiratory_syndrome_coronavirus_2
[Hap_4:  1   MN996532.2_Bat_coronavirus_RaTG13_complete_genome]
[Hap_5:  1   MN996531.1_|Severe_acute_respiratory_syndrome_coronavirus_2
[Hap_6:  1   MN996529.1_|Severe_acute_respiratory_syndrome_coronavirus_2
[Hap_7:  1   MN996527.1_|Severe_acute_respiratory_syndrome_coronavirus_2
[Hap_8:  1   MN994468.1_|Severe_acute_respiratory_syndrome_coronavirus_2
[Hap_9:  1   MN994467.1_|Severe_acute_respiratory_syndrome_coronavirus_2
[Hap_10:  1   MN988713.1_|Severe_acute_respiratory_syndrome_coronavirus_
[Hap_11:  1   MN985325.1_|Severe_acute_respiratory_syndrome_coronavirus_
[Hap_12:  1   MN975262.1_|Severe_acute_respiratory_syndrome_coronavirus_
[Hap_13:  1   MN938384.1_|Severe_acute_respiratory_syndrome_coronavirus_
```

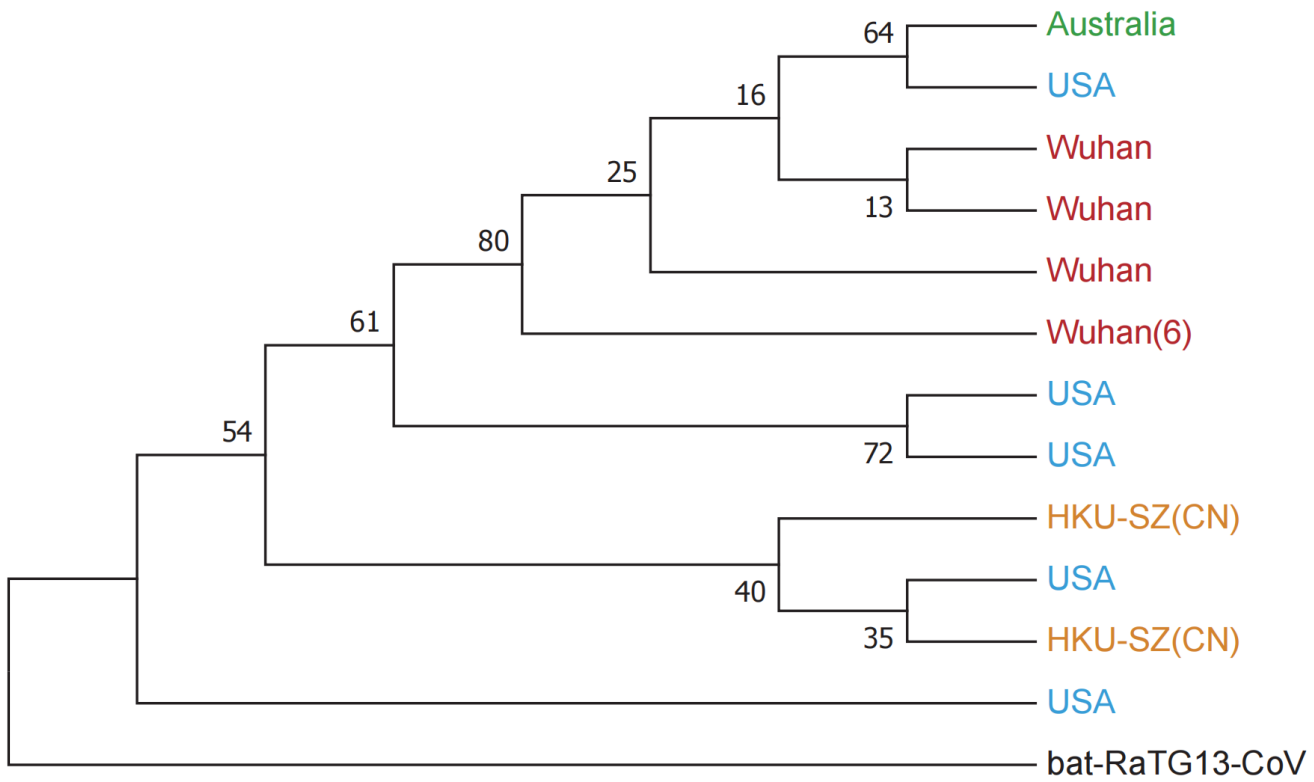
进化树的构建--MEGA

M7: Analysis Preferences

Options Summary

Option	Selection
Analysis	Phylogeny Reconstruction
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replications	1000
Substitution Model	
Substitutions Type	Nucleotide
Model/Method	Tamura-Nei model
Rates and Patterns	
Rates among Sites	Uniform rates
No of Discrete Gamma Categories	*Not Applicable*
Data Subset to Use	
Gaps/Missing Data Treatment	Complete deletion
Site Coverage Cutoff (%)	*Not Applicable*
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically (Default - NJ/BioNJ)
Initial Tree File	*Not Applicable*
Branch Swap Filter	None
System Resource Usage	
Number of Threads	12

Help Compute Cancel

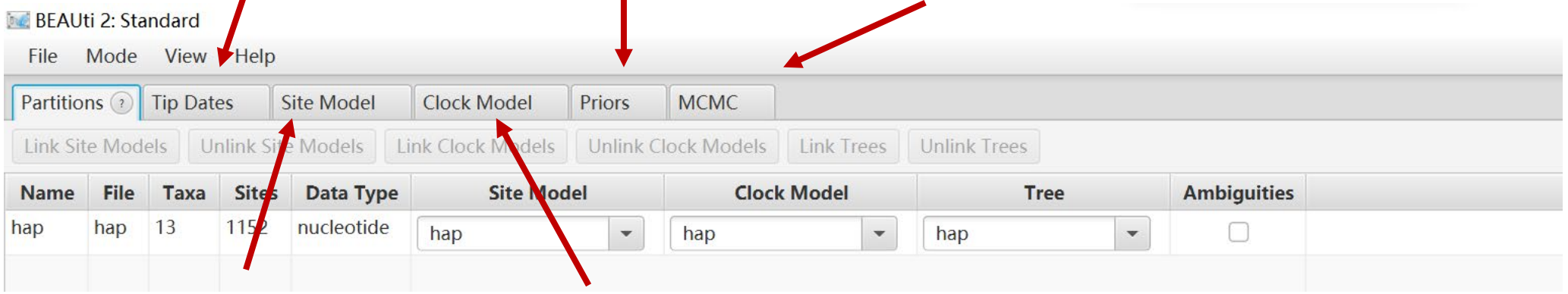


进化树的构建--BEAST

设置样本时间节点

分歧时间

3000*序列条数平方(20000000)
打印10000个
链长与步长



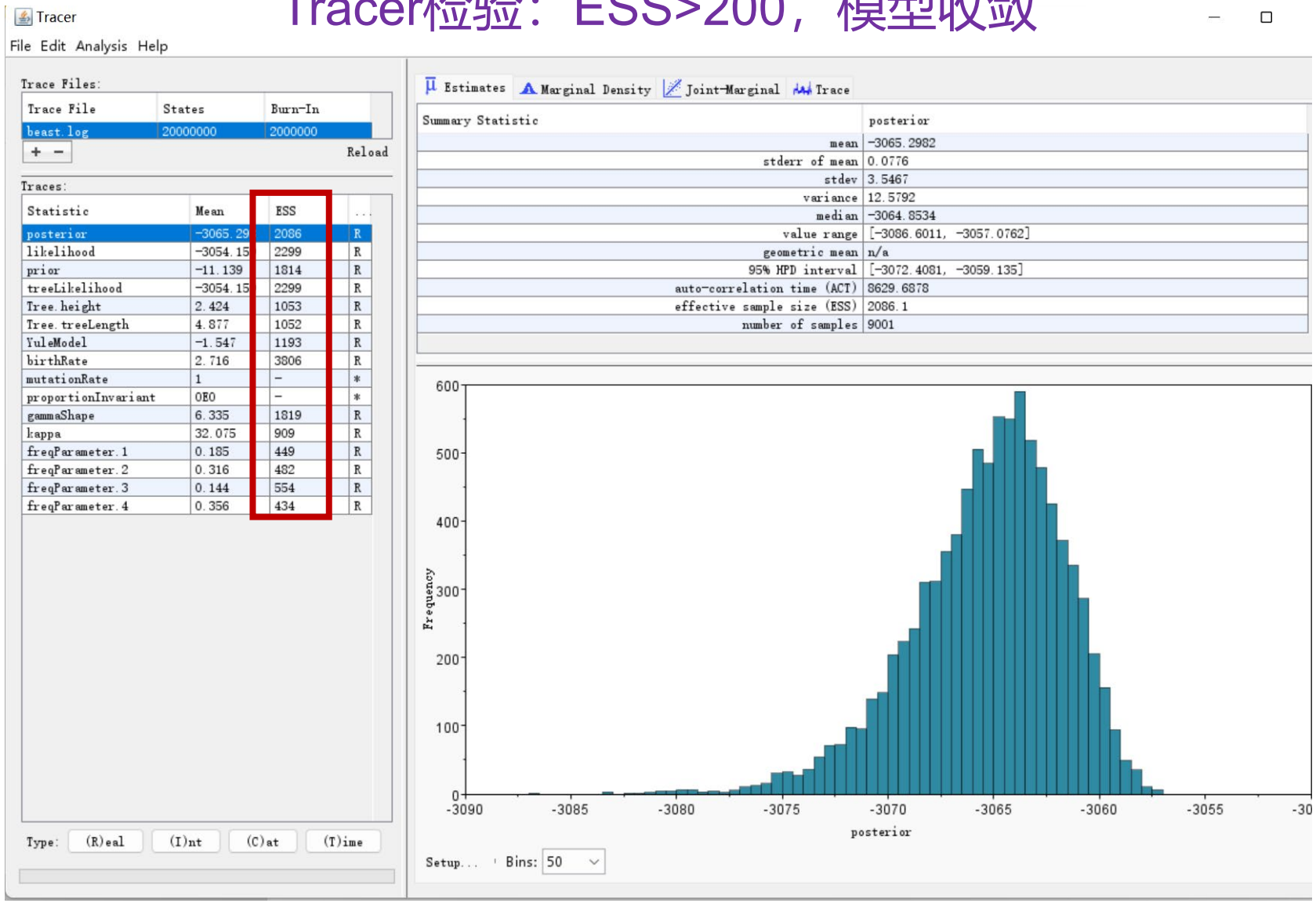
模型选择
替换速率等选择“estimate”
HKY核苷酸取代模型
(考虑转换>颠换)

设置分子钟
strict clock
(不同分支进化速率相同)

导入非中文命名的NEXUS格式文件，在BEAUti中完成参数设置

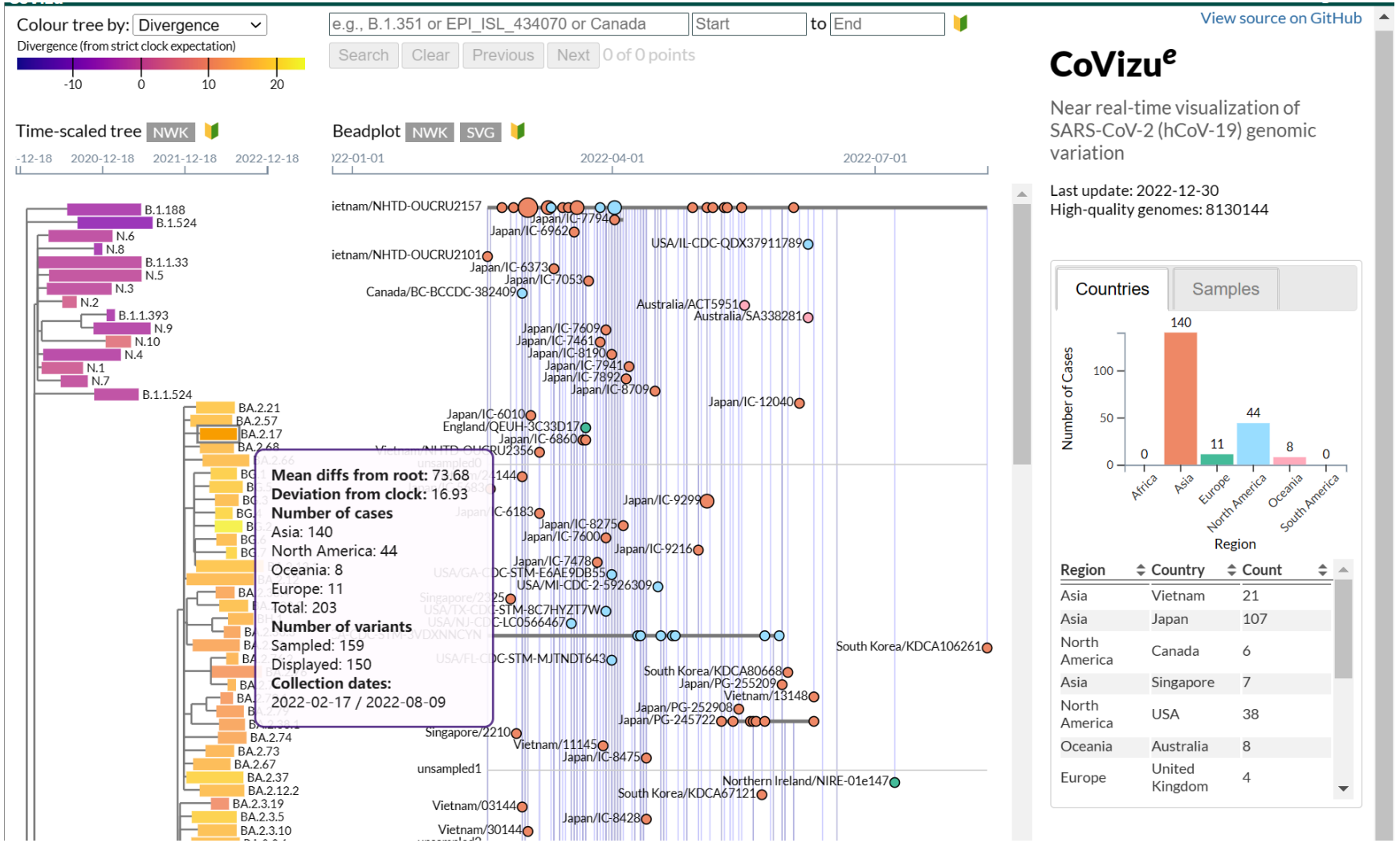
进化树的构建--BEAST

Tracer检验：ESS>200，模型收敛



FigTree：截取部分结果

新冠病毒进化树的在线构建及可视化--GISAID



02

Peking University

新冠病毒刺突糖蛋白与受体结合区域（RBD）的蛋白质结构预测

使用的新冠病毒刺突糖蛋白信息

突变株	序列标识
Wuhan-Hu-1 (wild-type)	NCBI ID: P0DTC2
德尔塔变体 （B.1.617.2）	NCBI ID: QWK65230.1
奥密克戎（XBB.1）	NCBI ID: OQ380107.1

二级结构预测结果

S蛋白结构域的细胞受体结合区（**Receptor binding domain，RBD**）直接参与了宿主受体的识别，

预测工具： GOR IV

	Wuhan-Hu-1 (wild-type)	德尔塔变体 （B.1.617.2）	奥密克戎（XBB.1）
刺突糖蛋白整体序列	GOR4 : Alpha helix (Hh) : 274 is 21.52% 3_{10} helix (Gg) : 0 is 0.00% Pi helix (Ii) : 0 is 0.00% Beta bridge (Bb) : 0 is 0.00% Extended strand (Ee) : 281 is 22.07% Beta turn (Tt) : 0 is 0.00% Bend region (Ss) : 0 is 0.00% Random coil (Cc) : 718 is 56.40% Ambiguous states (?) : 0 is 0.00% Other states : 0 is 0.00%	GOR4 : Alpha helix (Hh) : 280 is 22.03% 3_{10} helix (Gg) : 0 is 0.00% Pi helix (Ii) : 0 is 0.00% Beta bridge (Bb) : 0 is 0.00% Extended strand (Ee) : 279 is 21.95% Beta turn (Tt) : 0 is 0.00% Bend region (Ss) : 0 is 0.00% Random coil (Cc) : 712 is 56.02% Ambiguous states (?) : 0 is 0.00% Other states : 0 is 0.00%	GOR4 : Alpha helix (Hh) : 281 is 22.14% 3_{10} helix (Gg) : 0 is 0.00% Pi helix (Ii) : 0 is 0.00% Beta bridge (Bb) : 0 is 0.00% Extended strand (Ee) : 273 is 21.51% Beta turn (Tt) : 0 is 0.00% Bend region (Ss) : 0 is 0.00% Random coil (Cc) : 715 is 56.34% Ambiguous states (?) : 0 is 0.00% Other states : 0 is 0.00%
刺突糖蛋白RBD区域	GOR4 : Alpha helix (Hh) : 15 is 6.73% 3_{10} helix (Gg) : 0 is 0.00% Pi helix (Ii) : 0 is 0.00% Beta bridge (Bb) : 0 is 0.00% Extended strand (Ee) : 53 is 23.77% Beta turn (Tt) : 0 is 0.00% Bend region (Ss) : 0 is 0.00% Random coil (Cc) : 155 is 69.51% Ambiguous states (?) : 0 is 0.00% Other states : 0 is 0.00%	GOR4 : Alpha helix (Hh) : 13 is 5.83% 3_{10} helix (Gg) : 0 is 0.00% Pi helix (Ii) : 0 is 0.00% Beta bridge (Bb) : 0 is 0.00% Extended strand (Ee) : 55 is 24.66% Beta turn (Tt) : 0 is 0.00% Bend region (Ss) : 0 is 0.00% Random coil (Cc) : 155 is 69.51% Ambiguous states (?) : 0 is 0.00% Other states : 0 is 0.00%	GOR4 : Alpha helix (Hh) : 11 is 4.93% 3_{10} helix (Gg) : 0 is 0.00% Pi helix (Ii) : 0 is 0.00% Beta bridge (Bb) : 0 is 0.00% Extended strand (Ee) : 54 is 24.22% Beta turn (Tt) : 0 is 0.00% Bend region (Ss) : 0 is 0.00% Random coil (Cc) : 158 is 70.85% Ambiguous states (?) : 0 is 0.00% Other states : 0 is 0.00%

内在无序区域的结构预测

预测工具： PONDR® VLXT在线工具

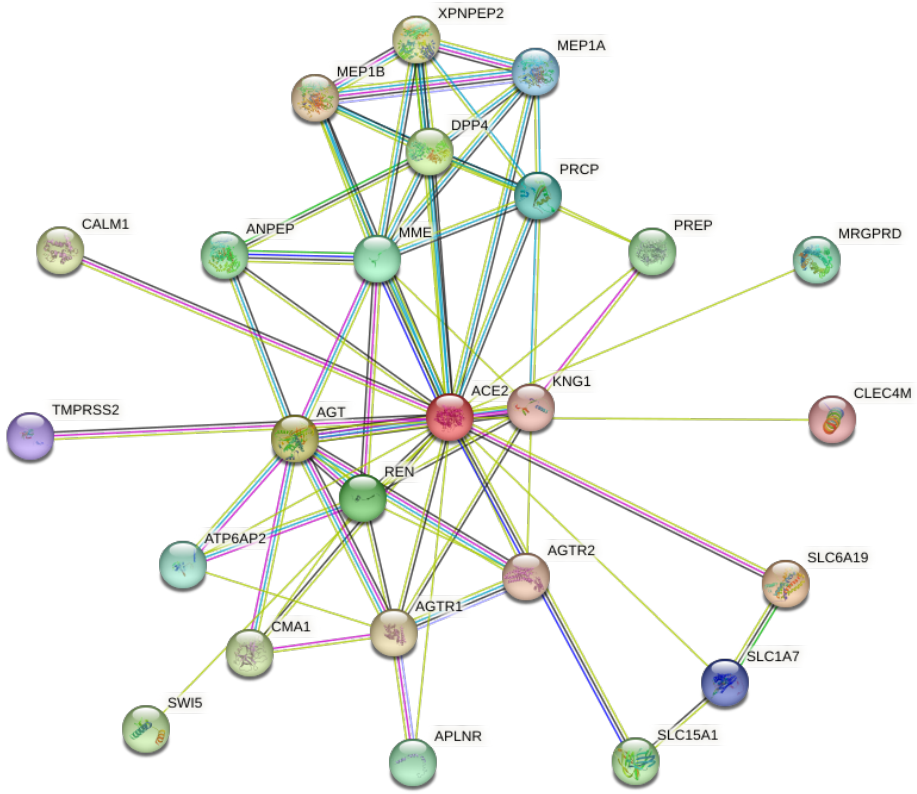
	无序残基的数量	整体无序的百分比	预测到的无序片段	无序的区域
Wuhan-Hu-1 (S)	98	7.7%	17 - 20; 468-475; 601-608; 672-709; etc	11
Wuhan-Hu-1 (RBD)	14	6.28%	1-6; 150-157	2
Delta (S)	101	7.95%	469-471; 599-608; 673-707; 867-869; etc	10
Delta (RBD)	6	2.69%	1-3; 151-153	2
Omicron (S)	85	6.7%	212-218; 405-406; 597-606; 674-705; etc	10
Omicron (RBD)	3	1.35%	87-88; 223-223	2

03

Peking University

新冠的S蛋白和人ACE2受体结合机制的生信分析

ACE2蛋白质互作预测



关联通路和代谢途径

term	description	FDR
GO:0001991	regulation of systemic arterial blood pressure by circulatory renin-angiotensin	2.59e-10
GO:0008015	Blood circulation	3.61e-08
GO:0010817	Regulation of hormone levels	6.80e-06
GO:0002034	Regulation of blood vessel diameter by renin-angiotensin	5.26e-05
hsa04924	Renin secretion	0.0076
hsa04614	Renin-angiotensin system	2.16e-25
hsa04974	Protein digestion and absorption	1.16e-12

STRING数据库参数：
minimum required interaction score: 0.700
max number of interactors to show: 1st shell 0.700, 2nd shell
其余为默认参数

基于机器学习的方法获取新冠病毒与人类蛋白相互作用

1.数据收集

481对新冠病毒-宿主的实验PPI [1]
包含12个新冠蛋白和304个人类蛋白。**①去重**得到304对PPI。

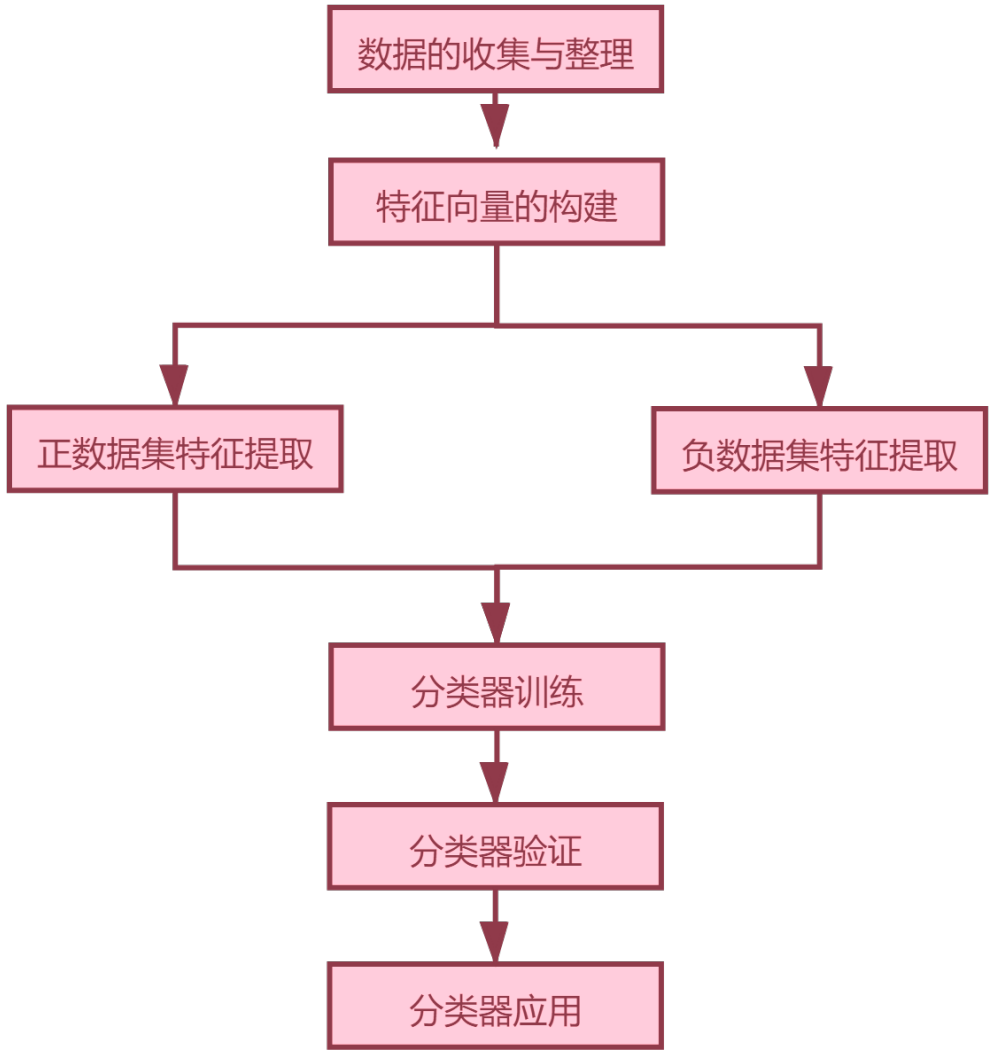
②去掉序列比对重复性比较高的蛋白

种内序列进行比对，相似性超过百分之30的蛋白进行删除操作
P0DTD1和P0DTC1的Aligned. Score达到 99.9092
删除后者，保留11个蛋白质
人类蛋白重要性相似，如重复则同时删除

	新冠	人类
比对前	12	304
比对后	11	270

PPI数量变化

实验PPI	去重	序列比对去除相似
418	304	269



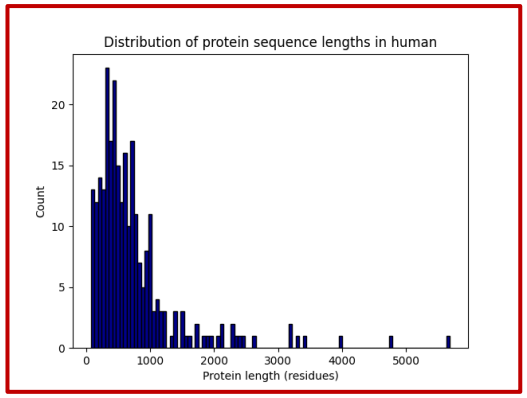
流程图

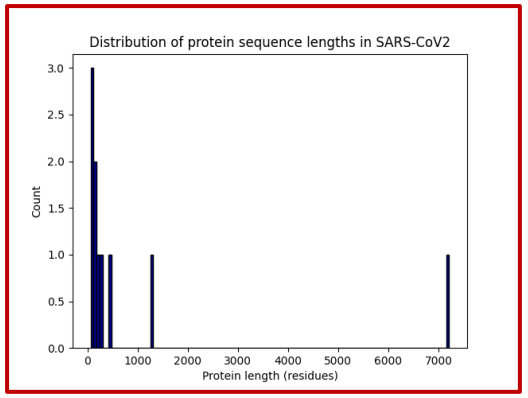
D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, and N. J. Krogan. A sars-cov-2 protein interaction map reveals targets for drug repurposing. *Nature*, 583(7816):1–13, 2020.

基于机器学习的方法获取新冠病毒与人类蛋白相互作用

2.转换为序列

Uniprot转换序列，得到病毒及人类蛋白长度分布直方图





(a) 人类蛋白序列长度分布
μ = 736 ± 720 残基

(b) SARS-CoV-2 蛋白序列长度分布
μ = 993 ± 2103 残基

3.正负样本构建

	正样本	负样本
PPI数目	269	2701

随机生成负样本
11*270-269=2701

2.特征抽提及特征构造思路

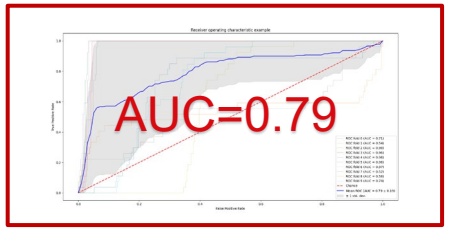
使用BioSeq-Analysis2.0 蛋白质分析平台[2]

kmer，参数k优化，SVM，chi-suqre检验

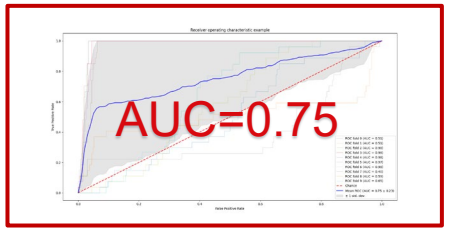
$$\{\varphi_A^+\} = \{\mathbf{y}_k\}^1 \oplus \{\mathbf{y}_k\}^2 \oplus \cdots \oplus \{\mathbf{y}_k\}^M$$

$$\{\varphi_{AB}^+\} = \{\varphi_A^+\} \oplus \{\varphi_B^+\}$$

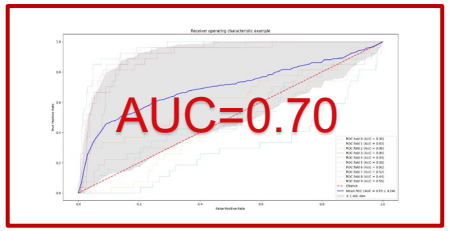
噪声引入



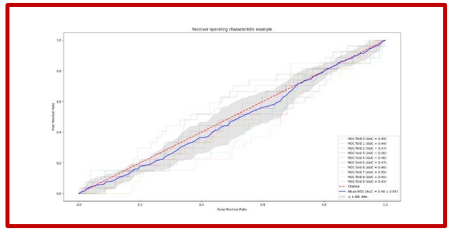
(a) 不加入噪声



(b) 2 倍特征噪声



(c) 5 倍特征噪声



(d) 20 倍特征噪声

预测得到311 个PPI

B. Liu, X. Gao, and H. Zhang. Bioseq-analysis2.0: an updated platform for analyzing dna, rna and protein sequences at sequence level and residue level based on machine learning approaches. *Nuclc Acids Research*, (20):20, 2019.

BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches

基于机器学习的方法获取新冠病毒与人类蛋白相互作用

预测出的蛋白互作网络可视化及分析

P0DTC2（S蛋白）

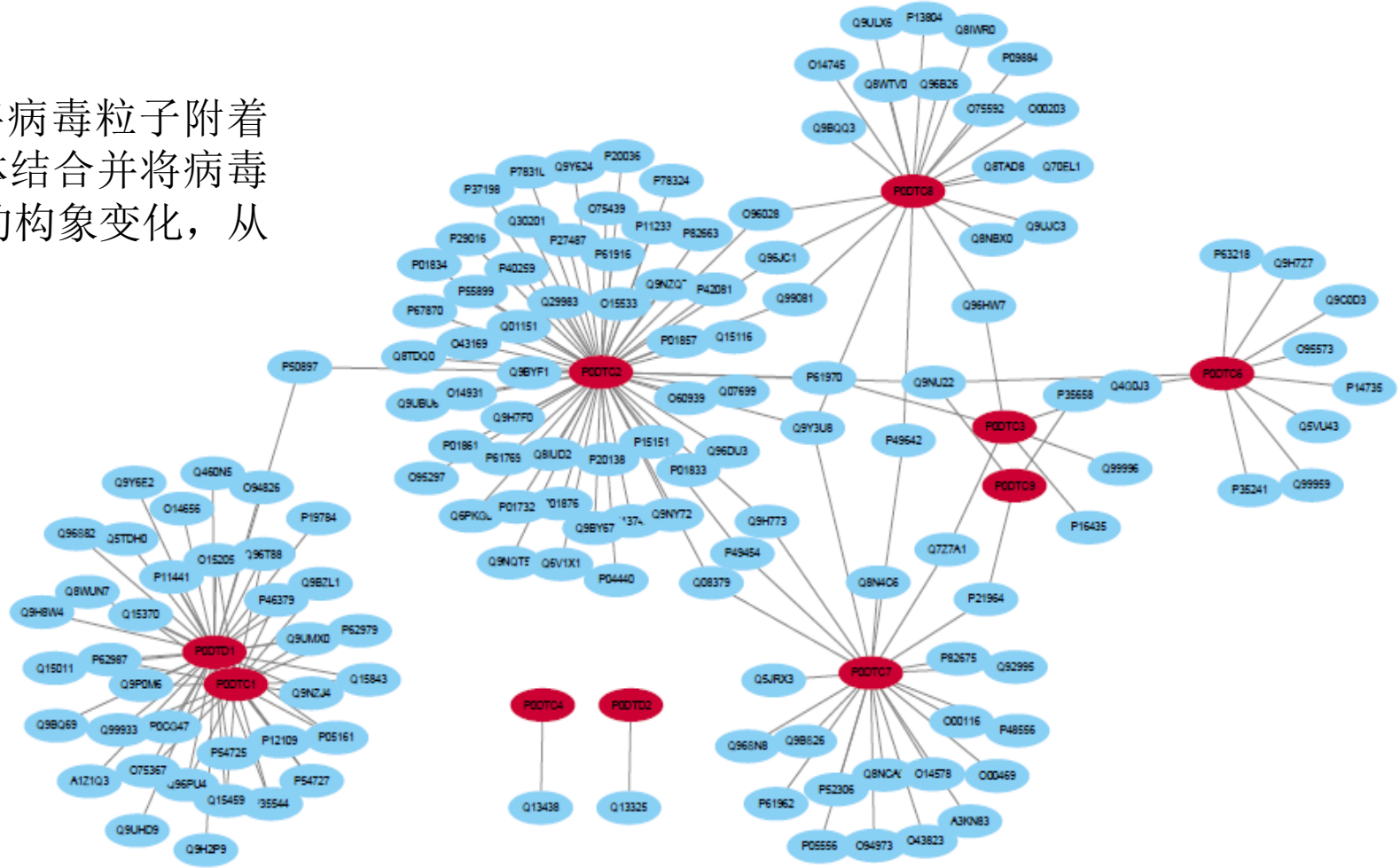
刺突蛋白,通过与宿主受体相互作用，将病毒粒子附着在细胞膜上，引发感染。与人ACE2受体结合并将病毒内化到宿主细胞的内体中会诱导糖蛋白的构象变化，从而增强病毒粒子进入宿主细胞的能力。

P0DTC8

的功能则是在调节宿主免疫反应中发挥作用，与IL17RA受体结合，导致IL17通路激活和促炎因子分泌增加，在感染期间导致细胞因子风暴。

激素治疗过程中炎症因子风暴是新冠重症死亡的一大原因。

新冠病毒侵染人类细胞的过程中，P0DTC2与P0DTC8起到了重要作用。



北京大学
PEKING UNIVERSITY

Thank you !

- 小组成员：邓启东 刘鋆 刘文清 林芮昀
- 报告人：刘文清