从新冠 Spike 蛋白抗原抗体复合物探究表位与抗体的空间识别关系

王彩翠[1]，毛甜甜[1]，裘天颐[2]，汪源[1]，郑根徽[1]，曹志伟[1,3*](1.同济大学生命科学与技术学院，上海 200092；2.复旦大学附属中山医院，上海 200032；3.复旦大学生命科学学院，上海 200433)

摘要：目的 基于新冠 Spike 蛋白抗原抗体复合物结构数据，探究表位与抗体的空间识别关系。**方法** 基于 718 对新冠 Spike 蛋白抗原抗体复合物结构数据，首先分析抗原表位分布的热点区域及抗体基因片段使用偏好性；其次对抗原抗体结合界面提取抗原空间表位和 CDR 结构，分别构建抗原空间表位的相似性聚类树以及对应抗体重链 CDR 结构的相似性聚类树，通过比较对应聚类树之间的相似性评估抗原空间表位与抗体 CDR 结构的整体识别关系。**结果** 表位位点中 94.02%分布于 RBD，4.44%分布于 NTD，且前 10 位热点位置均分布于 RBM；抗体 VJ 基因的偏好性片段为 IGHV3-30/IGHJ4 和 IGHV1-58/IGHJ3。空间表位聚类树与对应抗体 CDR 结构聚类树之间的相似性显著高于随机聚类树，提示表位与 CDR 结构存有内在关联。进一步分析数据发现，免疫原性相似的抗原表位，会被结构相似的 CDR 结构识别，且 CDR3 结构的贡献最大。**结论** 从目前数据来看，表位位点分布以 RBD 区域为主，抗体 IGHV3-30/IGHJ4 和 IGHV1-58/IGHJ3 片段突出，相似的抗原空间表位会被相似的 CDR 结构识别，这些发现为后续的新冠病毒抗体设计及优化提供一定的理论支撑。

关键词：抗原表位；抗体；CDR；层次聚类；SARS-CoV-2
中图分类号：R914.2　　**文献标志码：**B　　**文章编号：**1007-7693(2022)21-2850-06
DOI: 10.13748/j.cnki.issn1007-7693.2022.21.022
引用本文：王彩翠，毛甜甜，裘天颐，等. 从新冠 Spike 蛋白抗原抗体复合物探究表位与抗体的空间识别关系[J]. 中国现代应用药学, 2022, 39(21): 2850-2855.

Exploring the Spatial Recognition Relationship Between Epitopes and Antibodies from SARS-CoV-2 Spike Protein-antibody Complexes

WANG Caicui[1], MAO Tiantian[1], QIU Tianyi[2], WANG Yuan[1], ZHENG Genhui[1], CAO Zhiwei[1,3*](*1.School of Life Sciences and Technology, Tongji University, Shanghai 200092, China; 2.Zhongshan Hospital, Fudan University, Shanghai 200032, China; 3.School of Life Sciences, Fudan University, Shanghai 200433, China*)

ABSTRACT: OBJECTIVE To explore the spatial recognition relationship between epitopes and antibodies based on the structure data of SARS-CoV-2 Spike protein antigen-antibody complex. **METHODS** Seven hundred and eighteen available SARS-CoV-2 antigen-antibody structural complexes were analyzed in multiple ways. Firstly, the epitope hotspots and the usage preference of antibody gene fragments were analyzed. Secondly, the spatial epitopes and CDR structures were extracted from the antigen-antibody binding interfaces, and clustering trees of the spatial epitopes and the antibody CDR structures were respectively constructed. The recognition relationship between the epitopes and the CDR structures was evaluated by comparing the similarity between the corresponding clustering trees. **RESULTS** The 94.02% epitope sites were mapped on RBD, and 4.44% on NTD, with the top10 hotspots being all located in RBM. The most common antibody VJ genes were identified as IGHV3-30/IGHJ4 and IGHV1-58/IGHJ3. The similarity between the clustering trees of spatial epitopes and those of the corresponding antibody CDR structures was significantly higher than that expected from random clustering trees, suggesting an intrinsic epitope-CDR matching. Further analysis of the data revealed that the epitopes with similar immunogenicity would be recognized by similar CDR structures, with CDR3 domain making the greatest contribution. **CONCLUSION** The current data identifies that the epitopes are concentrated in its RBD region while the IGHV3-30/IGHJ4 and IGHV1-58/IGHJ3 are the preferred VJ gene combination utilized for the production of Spike protein-targeting antibodies. It also suggests that similar epitopes are likely to be recognized by similar CDR structures. Collectively, these findings add a new theoretical basis for the SARS-CoV-2 antibody design and optimization.

KEYWORDS: epitope; antibody; CDR; hierarchical clustering; SARS-CoV-2

基金项目：国家自然科学基金项目(32070657)
作者简介：王彩翠，女，硕士生　E-mail: 2131444@tongji.edu.cn　***通信作者：**曹志伟，女，博士，教授　E-mail: zwcao@fudan.edu.cn

新型冠状病毒肺炎(COVID-19)是由新型冠状病毒(SARS-CoV-2)导致的急性呼吸道传染病，自2019年12月出现后，COVID-19引发全球大流行[1]。至2022年8月，全球累计确诊病例超过五亿八千万，死亡病例超过六千四百万，对人类健康造成极大影响[2]。抗体是适应性免疫的重要组成部分，抗体通过 CDR 与抗原表位(抗原决定簇)特异性结合以阻断病原体入侵，在 COVID-19 的治疗中发挥不可或缺的作用[3]。新冠病毒出现后，产生并积累了许多新冠 Spike 蛋白抗原抗体复合物结构数据，这些数据中的抗原抗体相互识别与作用是否存在一定规律对于新冠病毒中和抗体的研究有重要意义，但目前尚未有针对新冠病毒抗原表位与抗体空间识别关系的系统性研究。因此，本研究旨在探究新冠病毒 S 蛋白抗原表位与对应抗体 CDR 结构的空间识别关系，从而为新冠病毒中和抗体的设计与优化提供一定参考。

1 材料

1.1 数据

截至2022年4月，从 SAbDab 数据库[4]共获得718对新冠 Spike 蛋白抗原抗体复合物结构。

1.2 实验网站及软件

SAbDab(http://opig.stats.ox.ac.uk/webapps/newsabdab/sabdab/)；ANARCI (http://opig.stats.ox.ac.uk/webapps/ newsabdab/sabpred/anarci/)；CE-BLAST (https://www.biosino.org/ce_blast/)；iTOL (https://itol.embl.de/)；WebLogo3 (https://weblogo.threeplusone.com/)；EMBOSS-6.6.0-Needleall 软件；Pymol 软件；MEGA11 软件。

2 方法

2.1 数据收集及预处理

截至2022年4月，从 SAbDab 数据库[4]下载抗原抗体复合物共4 577对，按照以下步骤过滤：①抗原名称为 SARS-CoV-2 Spike protein；②抗体序列同时包含轻重链；③按抗体序列相似性为100%去冗余。共获得809对新冠 Spike 蛋白抗原抗体复合物，其中人源750对。对750对人源新冠 Spike 蛋白抗原抗体复合物进一步筛选，以4Å为阈值，提取抗原抗体相互作用界面中的表位和对位，并去除表位<5个氨基酸的复合物，最终得到718对。

2.2 抗原表位分布及抗体 VJ 基因偏好性

统计表位位点在 Spike 蛋白不同区域的占比，NTD：13-305；RBD：319-541；S1/S2：542-787；S2：788-1273[5]。使用 ANARCI[6]对抗体进行注释并统计重链 VJ 基因的使用频率。

2.3 抗原表位及抗体 CDR 相似性计算

使用 CE-BLAST[7]计算两两表位的相似性，得到表位相似性矩阵。对抗体使用 IMGT 编号，并根据 CDR1：27-38；CDR2：56-65；CDR3：105-117分别提取 CDR1、CDR2、CDR3[8]，使用 CE-BLAST 计算两两 CDR 的相似性，得到 CDR 相似性矩阵。

2.4 层次聚类及聚类树相似性比较

首先，使用 R 中的 hclust(method="average")函数分别对表位及 CDR 相似性矩阵进行层次聚类；其次，使用 cutree(k=40)将表位聚类树划分为8大类[9]；最后，使用 iTOL 在线网站[10]对聚类树进行注释。

使用 R 包 TreeDist 中的 NyeSimilarity[11]函数进行聚类树相似性比较。将 NyeSimilarity 值的随机分布作为聚类树相似性的参考，即随机产生1 000个718×718的随机矩阵，根据该矩阵聚类得到1 000棵树，并从中进行1 000次随机的两两配对，计算得到1 000个 NyeSimilarity，获得其分布情况。

2.5 不同相似性阈值的表位及其对应抗体的聚类情况

相似性阈值分别设置为0.6，0.7和0.8，提取包含相似性表位最多的簇，将对应的表位及抗体用红色标注，并分别观察其在空间表位聚类树及抗体 CDR 结构聚类树中的聚类情况。

3 结果

3.1 新冠病毒抗原表位位点分布

来自 SAbDab 数据库[4]的 750 对人源新冠 Spike 蛋白抗原抗体复合物结构数据，预处理后得到 718 对。718 对复合物的抗原表位位点中，94.02%分布在 RBD 区域(319~541[5])，4.44%分布在NTD区域(13~305)，均位于S1蛋白上(1~541)(图1)。S1 蛋白上的表位位点主要分布于 350~500 附近区域，前10位热点位置为486，489，493，455，487，456，484，475，485 和 449，均位于RBM(438~508)区域，即 ACE2 受体与 S 蛋白结合的核心区域[12]，见表1。

3.2 新冠病毒抗体偏好性 VJ 基因片段

VJ 等基因片段重排，使得抗体多样性增加[13]，对新冠病毒抗体 VJ 基因的偏好性分析，有利于了

表 1　SARS-CoV-2 前 10 位表位热点

Tab. 1　SARS-CoV-2 top10 epitope hotspots

排序	位点	计数	占比/%
1	486	343	2.86
2	489	320	2.67
3	493	316	2.64
4	455	275	2.29
5	487	260	2.17
6	456	254	2.12
7	484	224	1.87
8	475	218	1.82
9	485	207	1.73
10	449	201	1.68

解目前已有新冠中和抗体的特征。使用 ANARCI[6] 对 718 个新冠病毒抗体进行注释，统计后发现重链 VJ 基因使用频率较高的是 IGHV3-30/IGHJ4 和 IGHV1-58/IGHJ3，见图 2。

3.3　抗原空间表位及对应抗体 CDR 结构相似性分析

3.3.1　空间表位聚类树与 CDR 结构聚类树相似性分析　为了研究新冠病毒 S 蛋白抗原表位与对应中和抗体 CDR 的空间识别关系，本研究首先基于大规模数据集对抗原空间表位进行了抗原性距离聚类。利用抗原表位比对算法 CE-BLAST[7]，对上述数据集中的 718 个抗原空间表位进行抗原性聚类，共得到 8 个大类，分别标注为 Group1 至 Group8。同时，考虑到重链 CDR 结构在抗原抗体结合过程中的重要作用，抗体结构的相似性聚类基于重链 CDR 进行。同样利用 CE-BLAST 算法，对上述数据集中的 718 个抗体的重链 CDR 结构进行相似性聚类，并使用空间表位聚类树中的 8 个大类对其进行标注(见图 3A 和 3B)。最后，利用两树相似性评分函数 NyeSimilarity[11]对空间表位聚类树和 CDR 结构聚类树进行相似性评估，为增加结果的可信度，将 1 000 个两两随机聚类树间的 NyeSimilarity 作为对照。结果显示，CDR 结构聚类树与空间表位聚类树间的 NyeSimilarity 为 0.556，显著高于 1 000 次随机 NyeSimilarity 取值分布(0.219~0.230，见图 3C)。上述结果表明，新冠病毒 S 蛋白抗原空间表位的相似性与对应中和抗体 CDR 结构相似性存在一定对应关联性，而非随机匹配。

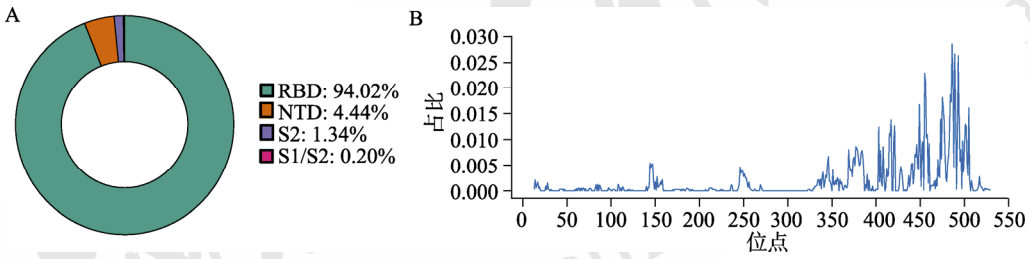

图 1　SARS-CoV-2 表位位点分布

A–表位位点在各区域分布占比；B–S1 蛋白(0~541)表位位点分布图。占比为该位点在总表位位点中的比例。

Fig. 1　Epitope site distribution of SARS-CoV-2

A–distribution proportion of epitope sites in each region; B–distribution map of S1 protein epitope(0–541) sites. The proportion was the proportion of this locus in the total epitope site.

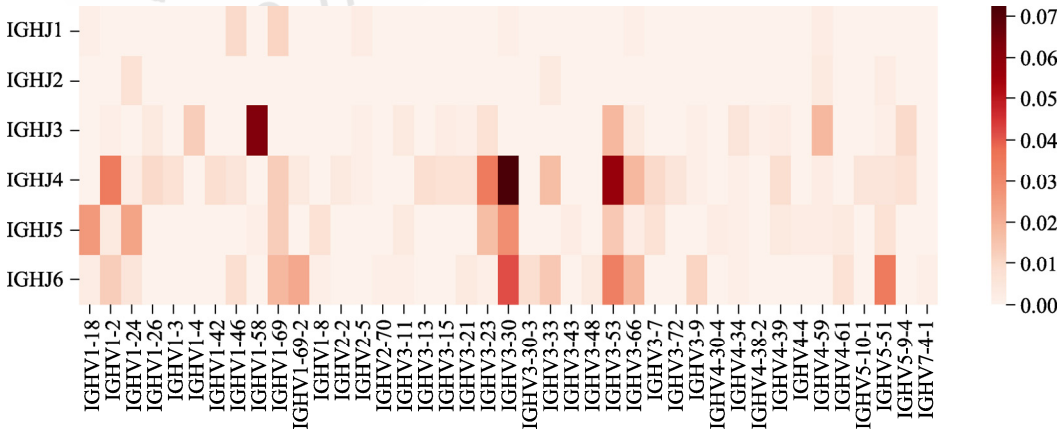

图 2　抗体重链 VJ 基因使用频率分布图

颜色越深则表示该 VJ 基因出现频率越高。

Fig. 2　Frequency distribution of antibody heavy chain VJ gene usage

The deeper the red color indicated a higher frequency of this VJ gene.

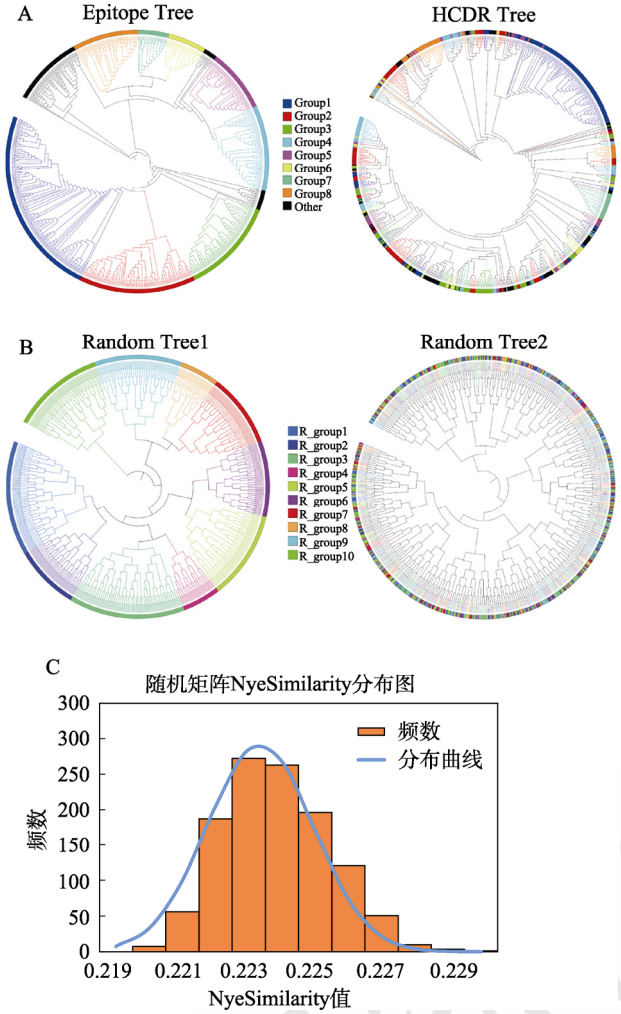

图 3 空间表位及抗体 CDR 结构聚类树

A–空间表位聚类树和抗体重链 CDR 结构聚类树，HCDR Tree 中的分布为 Epitope Tree 中不同簇在其上的映射；B–随机矩阵聚类树，Random Tree 2 中的分布为 Random Tree 1 中不同簇在其上的映射；C–1 000 次随机矩阵 NyeSimilarity 值分布图，均值=0.223，最小值=0.219，最大值=0.230。

Fig. 3 Clustering tree of spatial epitopes and antibody CDR structures

A–spatial epitopes clustering Tree and antibody heavy chain CDR structures clustering Tree, the groups of HCDR Tree was the same as Epitope Tree; B–random matrix clustering tree, the groups of Random Tree 2 was the same as Random Tree 1; C–distribution of 1 000 times random matrix NyeSimilarity, mean=0.223, minimum=0.219, maximum=0.230.

3.3.2 空间表位相似性对抗体 CDR 结构聚类效果的影响　在空间表位聚类树与抗体 CDR 结构聚类树的相似性分析中，笔者所在课题组发现表位与抗体的空间识别关系存在一定关联性。为进一步揭示潜在规律，本研究对 718 个空间表位免疫原性设置了不同的相似性阈值，分别为 0.6，0.7 和 0.8，提取不同阈值下包含相似性空间表位最多的类，并观察其在空间表位聚类树和抗体 CDR 结构聚类树中的聚类情况。分析结果显示，当表位相似性阈值为 0.6 时，共有 69 个相似性表位(见图 4A)，其对应抗体 CDR 结构聚类效果不佳，有 2 个小簇发生了明显的偏离，见图 4B；当表位相似性阈值为 0.7 时，共有 23 个相似性表位，见图 4C，其对应抗体 CDR 结构聚类效果相较阈值为 0.6 时有一定提升，但也有一小簇发生偏离，见图 4D；当表位相似性阈值为 0.8 时，共有 11 个相似性表位，见图 4E，其对应抗体 CDR 结构聚类效果良好，且相较阈值为 0.6 和 0.7 时有明显提升，见图 4F。综上，当抗原空间表位相似性逐渐提升时，对应识别的抗体 CDR 结构的相似性聚类效果也更好，说明相似的抗原空间表位会被相似的抗体 CDR 结构识别。

3.4 CDR3 在抗原抗体结合中的作用

互补决定区 CDR 由 CDR1、CDR2 和 CDR3 组成，其中 CDR3 变异最大，在抗体识别抗原过程中发挥关键作用[14]。因此，将抗体重链 CDR 结构拆分为 CDR1、CDR2 和 CDR3，并构建聚类树，分别计算与空间表位聚类树间的相似性。结果表明，基于 CDR1 和 CDR2 结构构建的聚类树与空间表位聚类树间的相似性得分为 0.480，已显著大于随机 NyeSimilarity 取值分布(0.219~0.230)；而基于 CDR3 结构的 NyeSimilarity 得分更高达 0.538，见表 2。上述结果表明，相对于 CDR1 和 CDR2，CDR3 结构聚类树与抗原空间表位聚类树的相似性更高，这表明在抗原抗体识别过程中，CDR3 贡献最大，这与以往的认知相吻合[14]。同时，值得注意的是，单独的 CDR1 和 CDR2，或 CDR3 结构的得分比 CDR 整体结构的得分(0.556)更低，见表 2，这表明虽然 CDR3 作用更大，但是 CDR1、CDR2 和 CDR3 还是作为一个整体参与抗原抗体的识别。

4 讨论

本研究基于大规模新冠 Spike 蛋白抗原抗体复合物数据集对表位与抗体的空间识别关系进行了探究。统计分析结果表明，新冠病毒抗体中，IGHV3- 30/IGHJ4 和 IGHV1-58/IGHJ3 片段使用频率较高；抗原表位中，94.02%的表位位点分布在 RBD 区域，且前 10 位的表位位点均位于 ACE2 结合的核心区域 RBM。但由于 RBM 是新冠病毒中的研究热点，相较其他区域，大量 RBM 相关结晶结构可能会对高频表位位点造成统计误差。

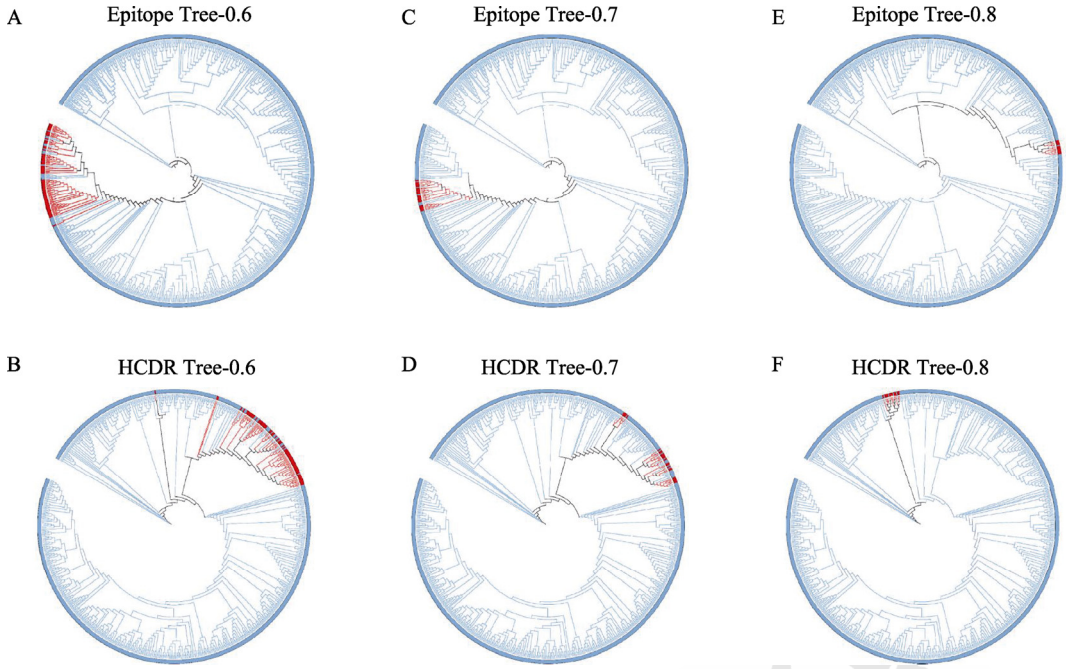

图 4　不同相似性阈值下的空间表位与抗体 CDR 结构聚类情况

A–表位相似性阈值为 0.6 时的空间表位聚类情况，共 69 个相似性表位；B–表位相似性阈值为 0.6 时的抗体 CDR 结构聚类情况，共 69 个 CDR；C–表位相似性阈值为 0.7 时的空间表位聚类情况，共 23 个相似性表位；D–表位相似性阈值为 0.7 时的抗体 CDR 结构聚类情况，共 23 个 CDR；E–表位相似性阈值为 0.8 时的空间表位聚类情况，共 11 个相似性表位；F–表位相似性阈值为 0.8 时的抗体 CDR 结构聚类情况，共 11 个 CDR。

Fig. 4　Clustering of spatial epitopes and antibody CDR structures under different similarity thresholds

A–spatial epitopes clustering when the epitope similarity threshold was 0.6, with a total of 69 similar epitopes; B–antibody CDR structures clustering when the epitope similarity threshold was 0.6, with a total of 69 CDR structures; C–spatial epitopes clustering when the epitope similarity threshold was 0.7, with a total of 23 similar epitopes; D–clustering of antibody CDR structures at the epitope similarity threshold of 0.7, with a total of 23 CDR structures; E–spatial epitopes clustering when the epitope similarity threshold was 0.8, with a total of 11 similar epitopes; F–antibody CDR structures clustering at an epitope similarity threshold of 0.8, with a total of 11 CDR structures.

表 2　聚类树间的 NyeSimilarity

Tab. 2　NyeSimilarity of clustering trees

聚类树 1	聚类树 2	NyeSimilarity
Epitope Tree	HCDR Tree	0.556
Epitope Tree	HCDR12 Tree	0.480
Epitope Tree	HCDR3 Tree	0.538
Random Tree1	Random Tree2	0.219~0.230

注：随机聚类树间的 NyeSimilarity 为 1 000 次随机得到的分布。

Note: NyeSimilarity between random clustering trees was the distribution obtained randomly for 1 000 times.

本研究发现，抗原空间表位相似性越高，则对应的抗体 CDR 结构相似性也越高。且 CDR3 是抗原抗体结合过程中最重要的区域[14]，抗原聚类树与抗体聚类树间的相似性主要来源于 CDR3 结构与空间表位间的相似性。虽基于 CDR1 和 CDR2 结构的聚类树与空间表位聚类树间的相似性也显著高于随机 NyeSimilarity 的取值分布，但小于 CDR3 结构聚类树与空间表位聚类树间的相似性。另外，分析结果中也存在一些不理想之处，例如在抗体聚类树中，有的簇虽呈一定的“热节点”分布，但整体聚簇并不明显，如 Group2、Group3 等。这可能是由于 SARS-CoV-2 抗原抗体复合物结晶过程中存在缺失，结构不完整所导致的；也有可能是因为 CE-BLAST 是抗原空间表位相似性计算工具，对抗原表位较敏感，对抗体 CDR 结构可能效果不佳。

总之，系统性的统计学分析表明新冠病毒 S 蛋白抗原表位与对应抗体 CDR 的空间识别关系具有一定规律，即抗原空间表位相似性越高，对应识别的抗体 CDR 结构相似性也越高，且 CDR3 在抗原抗体结合过程发挥关键作用。该规律探索为新冠病毒中和抗体的结构特征研究提供了一定参考，从而有助于设计及优化得到特异性高、疗效好、安全性强的新冠病毒中和抗体。

REFERENCES

[1] ZHU N, ZHANG D Y, WANG W L, et al. A novel coronavirus from patients with pneumonia in China, 2019[J]. N Engl J Med, 2020, 382(8): 727-733.

[2] World Health Organization. WHO Coronavirus(COVID-19) Dashboard[J/OL]. (2022-08-09). https://covid19.who.int/.

[3] LAGUNAS-RANGEL F A, CHÁVEZ-VALENCIA V. What

do we know about the antibody responses to SARS-CoV-2?[J]. Immunobiology, 2021, 226(2): 152054.
[4] DUNBAR J, KRAWCZYK K, LEEM J, et al. SAbDab: the structural antibody database[J]. Nucleic Acids Res, 2014, 42(Database issue): D1140-D1146.
[5] SONG S H, MA L N, ZOU D, et al. The global landscape of SARS-CoV-2 genomes, variants, and haplotypes in 2019nCoVR[J]. Genomics Proteomics Bioinformatics, 2020, 18(6): 749-759.
[6] DUNBAR J, DEANE C M. ANARCI: antigen receptor numbering and receptor classification[J]. Bioinformatics, 2016, 32(2): 298-300.
[7] QIU T Y, YANG Y Y, QIU J X, et al. CE-BLAST makes it possible to compute antigenic similarity for newly emerging pathogens[J]. Nat Commun, 2018, 9(1): 1772.
[8] LEFRANC M P, POMMIÉ C, RUIZ M, et al. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains[J]. Dev Comp Immunol, 2003, 27(1): 55-77.
[9] PARADIS E, SCHLIEP K. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R[J]. Bioinformatics, 2019, 35(3): 526-528.
[10] LETUNIC I, BORK P. Interactive Tree of Life(iTOL) v5: An online tool for phylogenetic tree display and annotation[J]. Nucleic Acids Res, 2021, 49(W1): W293-W296.
[11] NYE T M W, LIÒ P, GILKS W R. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees[J]. Bioinformatics, 2006, 22(1): 117-119.
[12] WANG M Y, ZHAO R, GAO L J, et al. SARS-CoV-2: Structure, biology, and structure-based therapeutics development[J]. Front Cell Infect Microbiol, 2020(10): 587269.
[13] HOZUMI N, TONEGAWA S. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions[J]. PNAS, 1976, 73(10): 3628-3632.
[14] 李宁丽, 张冬青, 周光炎. B 淋巴细胞抗原互补决定区序列分析的应用[J]. 细胞与分子免疫学杂志, 2001, 17(2): 104-105.

收稿日期：2022-08-23
(本文责编：沈倩)