Identify hotspots for the SARS-CoV-2

Content

1.Abstract

After Covid-19 pandemic, understanding the evolutionary principles of SARS-Cov-2 is crucial to further research. Based on 8 million SARS-Cov-2 sequence data, this article will introduce four algorithms for finding mutation hotspots. Weighted average proximity (WAP) score, which assess the significant of mutation based on standardized distance and Virus Number (VN) is the key for the article Comprehensive assessment of cancer missense mutation clustering in protein structure (Atanas K, et al., 2015) and will continually be used for all methods here. We first reproduce the method on this protein. We then improve the method to consider the density of the residual and high frequency points in each cluster, which is not considered before. The greedy method could choose the radius for each sphere, which will adjust the size of each cluster. The fourth method, using ML algorithm including RF and SVM is demonstrated It can decrease the calculation and identify more 'representative' area of hotspots.

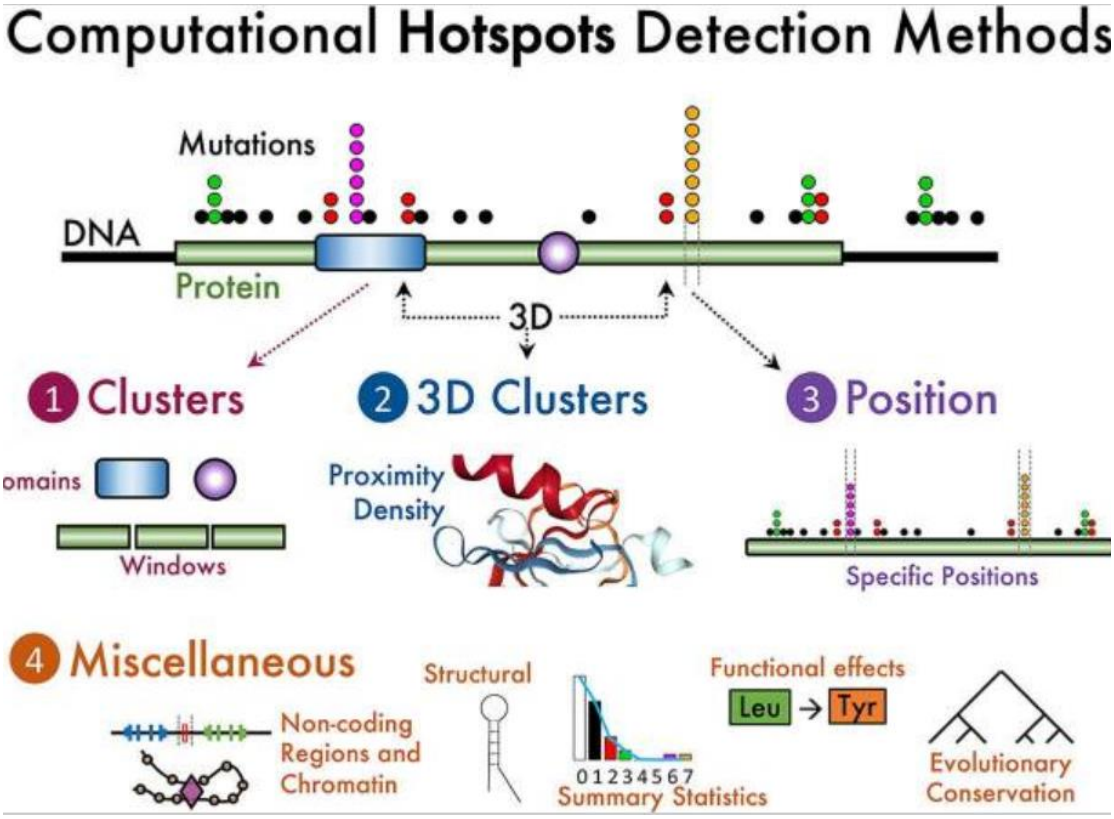
Statistical analysis is performed to compare and evaluate the results of the four methods. The greedy method is estimated to bring the most adaptive result. The first two methods, which prioritize overall mutation occurrence within an area, consistently identify a specific region. The fourth method focuses more on individual residues with high VN. To

provide a comprehensive assessment, we also compare these methods with a k-means clustering approach. Additionally, we conduct further analysis to explore long-term correlations within the protein's 3D structure. Combined with biological analysis, our findings suggest that considering the 3D structure systematically can aid in understanding the functional implications of these mutations. Together with biological analysis, our results indicate that systematic consideration of 3D structure can assist in understanding of the functional role of their mutations.

2.Introduction:

A hotspot can be defined as an area that has higher concentration of events compared to the expected number given a random distribution of events. Hotspot detection evolved from the study of point distributions or spatial arrangements of points in a space (Chakravorty, 1995). When examining point patterns, the density of points within a defined area is compared against a complete spatial randomness model, which describes a process in which point events occur completely at random (i.e., homogeneous spatial Poisson process). Beyond assessing the density of points in each area, hotspot techniques also measure the extent of point event interaction to understand spatial patterns (Baddeley, 2010).

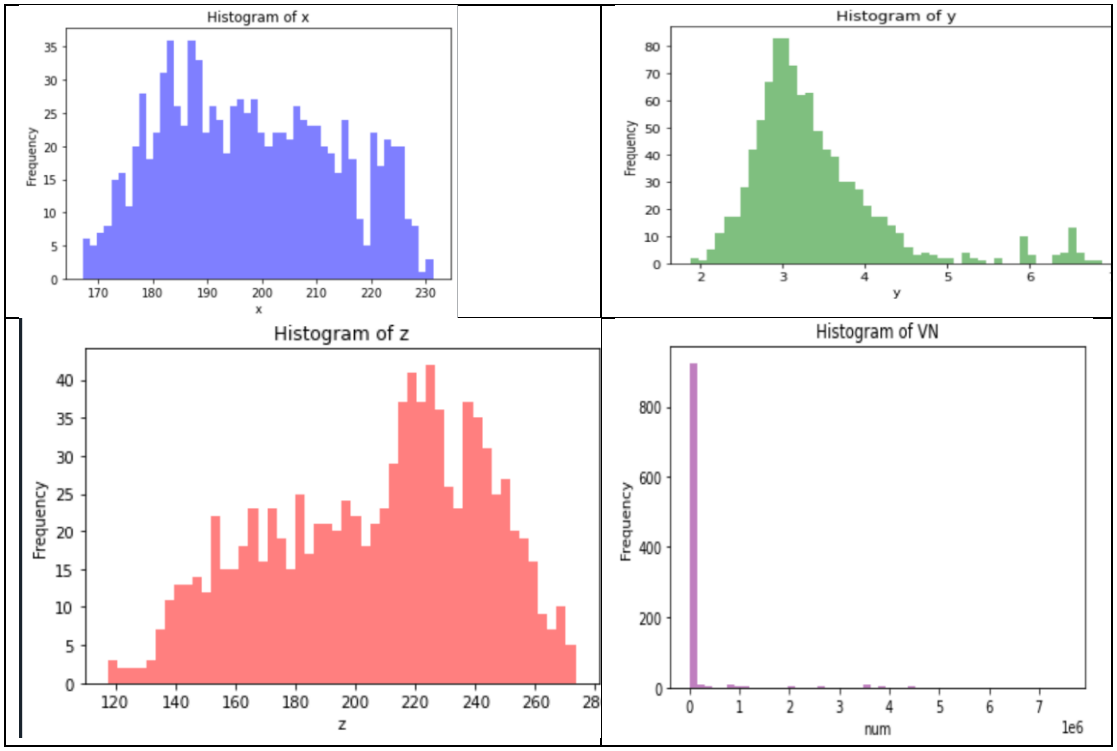
Definition of hotspots is crucial. Usual definition of a hotspot refers to a region or area where there is an unusually high concentration or intensity of a particular phenomenon or event. Here in a protein, we define a hotspot as a place where mutation is easily to occur.



3.Data exploration

We get the coordinates of the residues, ('x','y','z'), indicating their position in 3D space and number of mutation for each residues. (VariusNumber and VariusPercentage, all use VN below).

As the three types of SARS-Cov-2 protein(6vxx,6vyb,7dzw) are spatially identical. We focus first on 6vxx protein. CA atom is the only consideration, which has 972 residues.

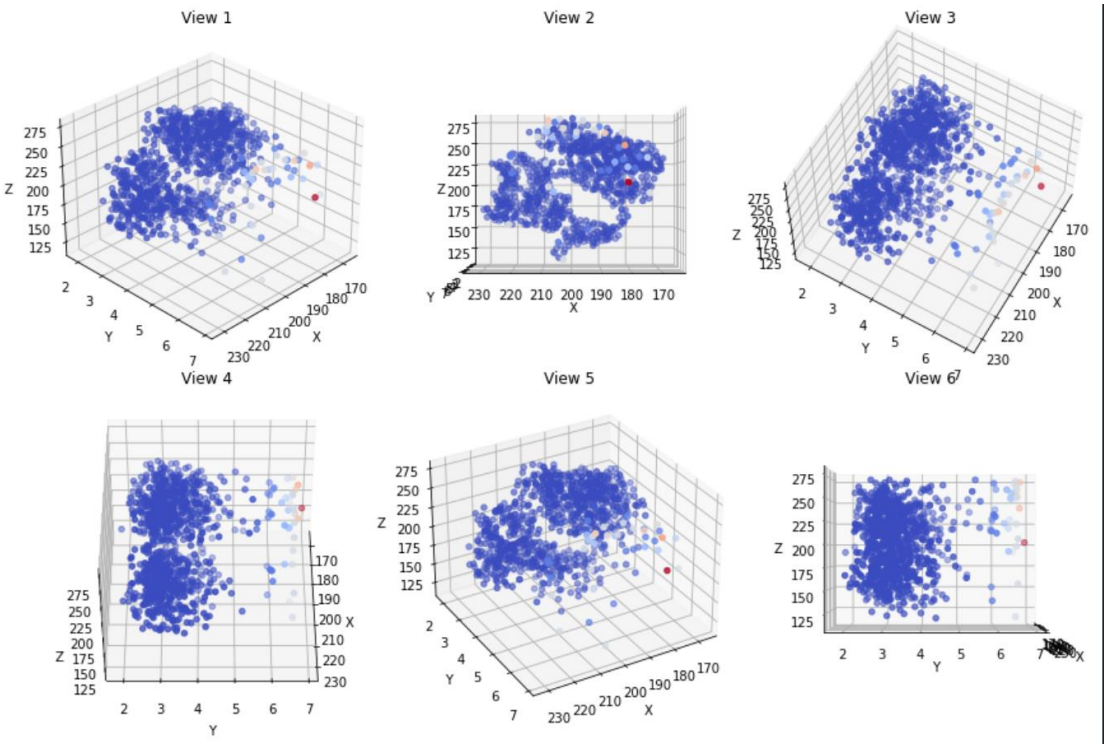


The histograms for the four variables are demonstrated above. It could be seen that although wide range, the majority of the resides have a relatively low VN. Value for y has little right skew.

It should be noticed that we do not need to perform any transformation here. As WAP method will consider the scale itself.

We have not got the time of for any mutations as the data is collected from every test for the virus. At this stage, we choose to research for a

static model, which do not consider time or dynamic motion.



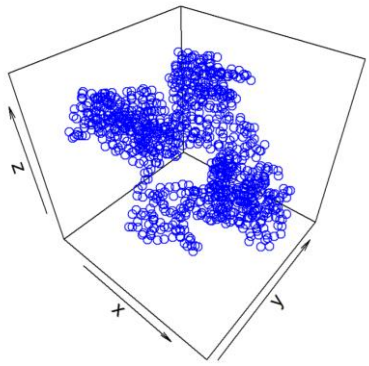
In the above figures, we use color to represent the magnitude of VN. Redder places indicate higher VN while the bluer places indicate smaller VN. Although high VN points is relative scarce, it seems they are somehow clustered.

4.Method

4.1 Basic WAP-Sphere

Each cluster here is defined by a 5A radius sphere with each point as its center. Sphere itself has a lot of benefits in dealing with the problem.

As 5A is typically defined as a distance which two residues could interact with each other, we could say that the center residue interacts with another one if it is in the sphere. The shape of sphere itself is good for interpretation.



Radius of the sphere is crucial. Therefore, before selecting clusters, we make a statistical analysis of the cluster situation in different radius. The average number of residuals in sphere is 3.5 for radius of 5A and 18.1 for the radius of 10A. The distribution of number of residues is quite different: For radius of 5A, most number is 3 and 4, but increase radius to 10A, we have a much wider range from 4 to 31. Generally, we think 5A is a reasonable choice.

CLUMPS Methodology. To identify significant clustering of mutations in proteins with available structural data, we first defined a WAP score summarizing all pairwise distances between mutated residues in a given 3D protein structure (*SI Appendix*, Fig. S1) as

$$\mathrm{WAP} = \sum_{q,r} n_q n_r e^{-\frac{d_{q,r}^2}{2t^2}}, \quad [1]$$

where q and r ($q \neq r$) are protein residues; $d_{q,r}$ is the Euclidean distance (in Å) between the centroids of those residues; and n_q (or n_r) is the number of samples where q (or r) is found mutated, normalized to the range [0,1] using the sigmoidal Hill function

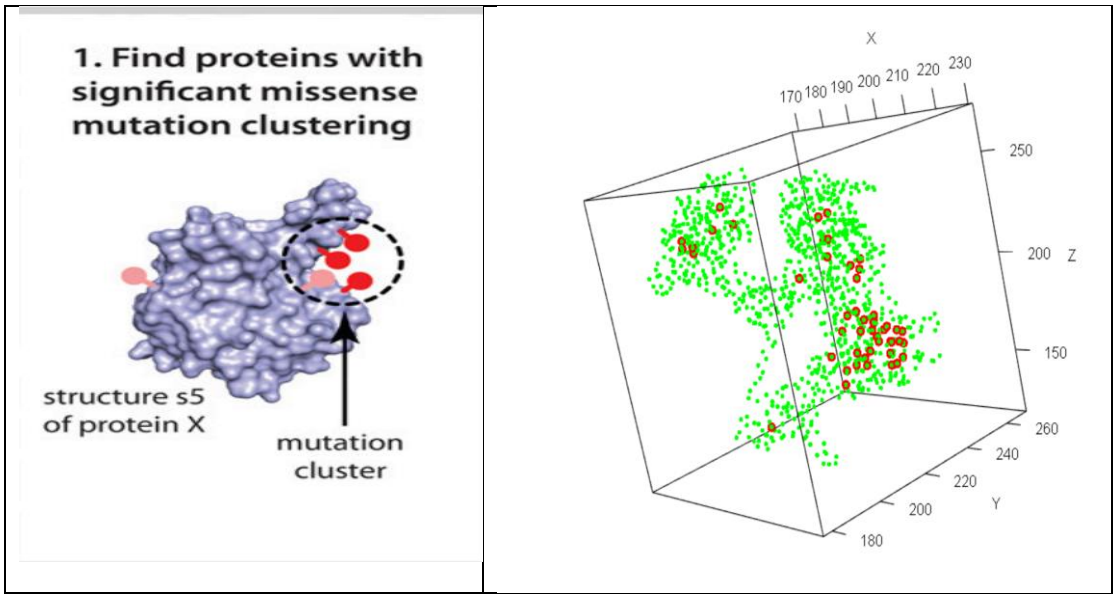
$$n_q = \frac{N_q^m}{\theta^m + N_q^m}. \quad [2]$$

Here, N_q is the number of samples with a missense mutation impacting residue q of the protein; and $\theta = 2$ and $m = 3$ are parameters of the Hill function controlling the critical point (center) and steepness of the sigmoid function, respectively. The exponential function in Eq. **1** transforms the absolute spatial distance $d_{q,r}$ between residues to the interval [0,1] with shorter distances (relative to the parameter t that can be interpreted as a "soft" distance threshold and was set to $t = 6$ Å) mapping to a value close to 1 and longer distances

Weighted average proximity (WAP) scoring function is then used. It summarizes the pairwise Euclidean distances of all mutated residues in the structure, weighted by the normalized number of samples in which they are mutated is a good criterion for evaluating each cluster.

We assess the significance of a given WAP score by comparing it to the null distribution obtained by randomly permuting the positions of the mutations across all residues in the structure (preserving the distribution of the number of samples mutated at a given residue) to obtain an empirical P value. After [10^5] times of random permutation, If the p-value of the original WAP score less than 5%, then we have an area which has an unusually high VN.

Results are follows:



Indices (on the 972 set):

The indices are the center of the spheres selected.

8, 10, 124, 176, 305, 315, 407, 427, 582, 578, 593, 625, 629, 633, 656, 670, 696, 698, 699, 702, 704, 705, 707, 854, 859, 874, 875, 877, 879, 881

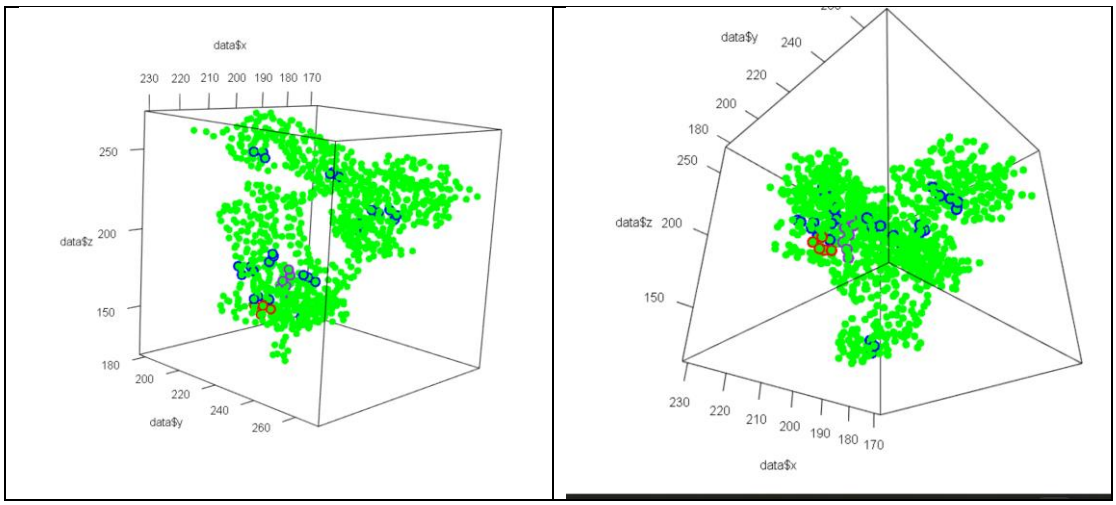
we find the chosen indices of the residues is somewhat cluster in the 3D structure.

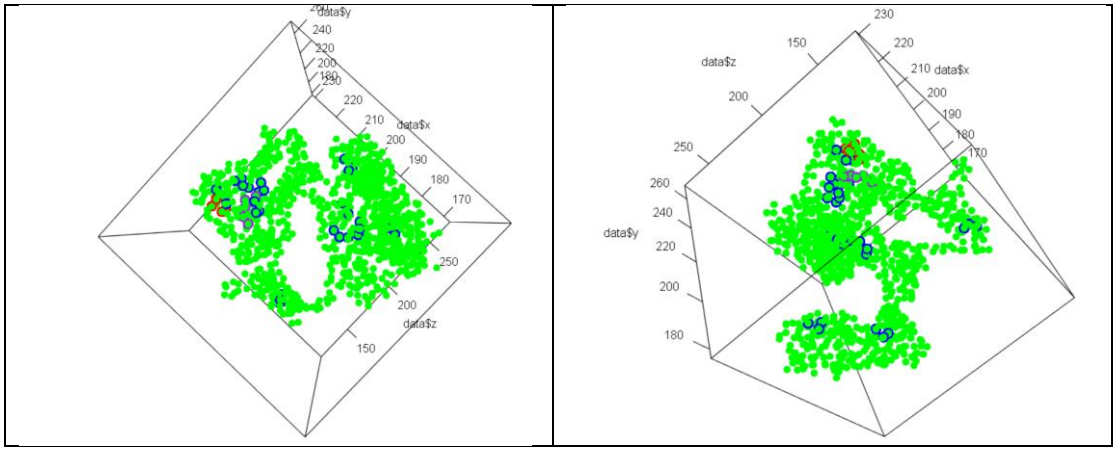
4.2 WAP-Sphere+Intersect

We should notice that the last method has drawbacks. Firstly, it does not consider the intersection of spheres. This should be considered as consecutive indices are common in the last result, indicating possible intersection. Not analyze this will lead to inaccuracy of the results. Secondly, we cannot learn exactly how many high VN points in those spheres. If we contain an extremely low(nearly zero points) in the sphere, we could loss that cluster regardless of an extreme high points.

Thus, we check the residues which appears not in a single cluster in the last result. It turns out quite a lot of clusters, or spheres we found earlier are intersect with each other. This is not surprising as in this case high frequency points have tendency to cluster.

We consider those spheres which intersect with each other as one category. 36 categories have been found. A category can either be a single sphere, or two or more intersected sphere. We defined those single sphere as ‘single area’ and those intersected sphere as ‘combined area’. To identify high VN points in those categories, we now search for the residues with VN higher then 8000, or the top 17.5% of the points. Among the 36 categories, we have one cluster have three those points, three with two high VN points. Most of the cluster have at least one distinguished point. We call the category with three high VN points the area3, and other area2 and area1 respectively. The graph shows the area3,2,1 in red, purple, and blue respectively.





Result: Indices (on the 972 set):

Please mind that those are the indices for all the points in those area:

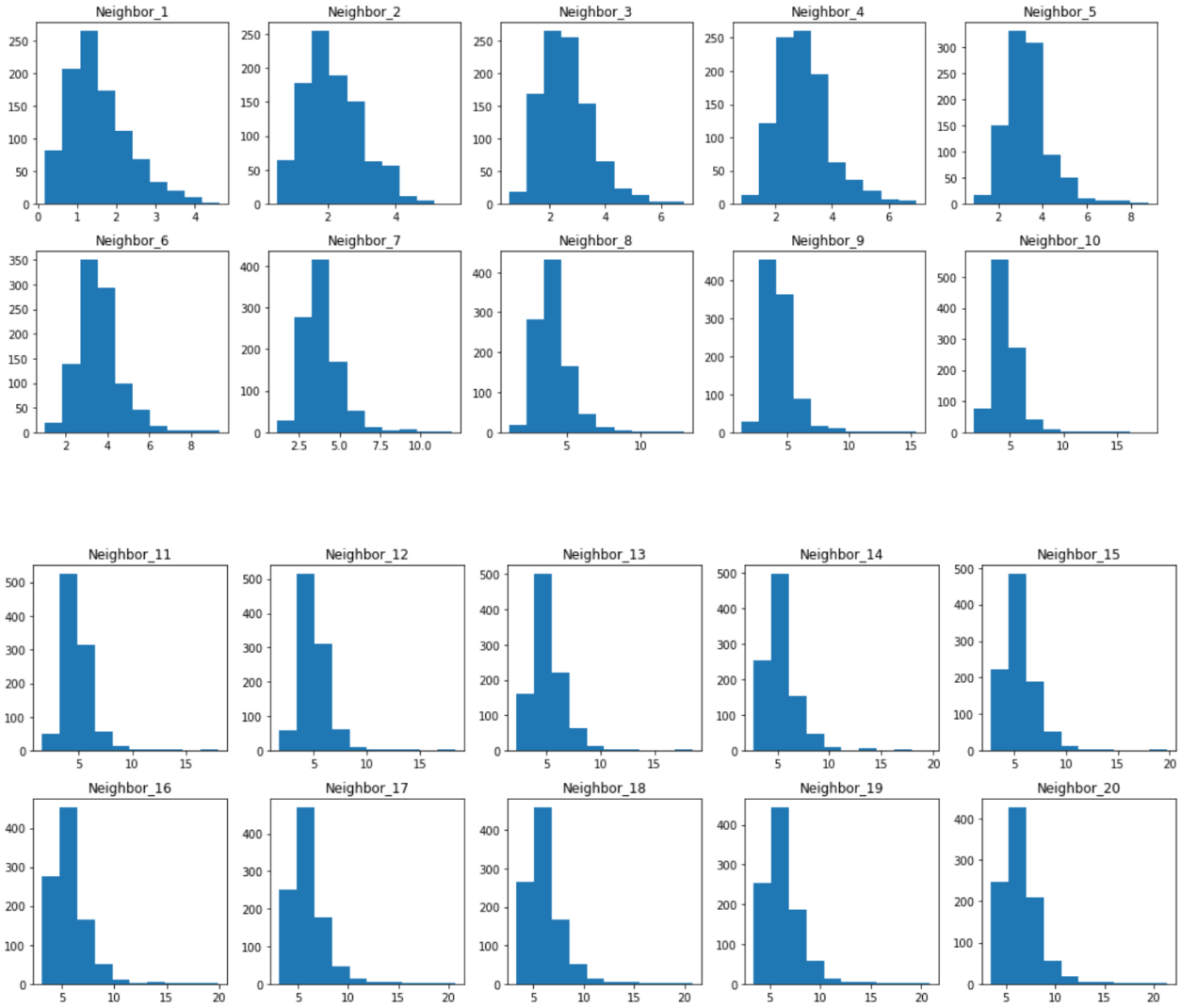
Area3: 704 705 706 707 708

Area2: 878 879 880 876 877 881 882

Area1: 701 702 703 9 10 11 7 8 628 629 630 627 581 582 583 426 427 428 211 212 213 905 906 907 672 673 674 322 323 324

4.3 Greedy WAP-Sphere+Intersect

We try to revise the last method by designing a greedy algorithm to give a variable radius for each cluster. The histograms below show the distances to the neighbor range from the nearest one to the 20th. As 5 is a typical number for the distances of 7th to 15th neighbor, fixing radius to 5 might import to a lot of human bias in it and loss some information.



We add a greedy search step when constructing the sphere. Starting from radius of three, we continuously search for the next point and update the value of WAP for the sphere. We aim to search for a radius with the largest WAP and in the range of 3 to 10, which is the optimum radius for a sphere.

Keep the following step the same to the second method, the result is:

Area3: 701 702 703 704 705 706 707 708

Area2: 878 879 880 876 877 881 882

Area1: 9 10 11 7 8 628 629 630 627 581 582 583 426 427 428 211 212 213 905 906 907 672 673 674 322 323 324

4.4 High center+WAP for RF

All improved CLUMPS requires a lot of computations and cannot make sure the center of the sphere is high mutated. We develop another greedy algorithm, which start from high frequency points as the center, and using Random Forest to decide which low frequency points should be included in hotspots based on a revised WAP value.

In this procedure, we first pick top alpha%(called high area) of the residues in VN. Those residues will be considered as the center of the clusters. Secondly, we will consider other residues (called low area, not in the top alpha%) and decide which will be included in the hotspots. This is done by comparing WAP score if one residue is considered. In the cluster. we will finally use random forest to find how to divide the space near the hotspots (in the cluster or out of cluster) can we find the optimum solution.

Result: Indices(on the 972 set):

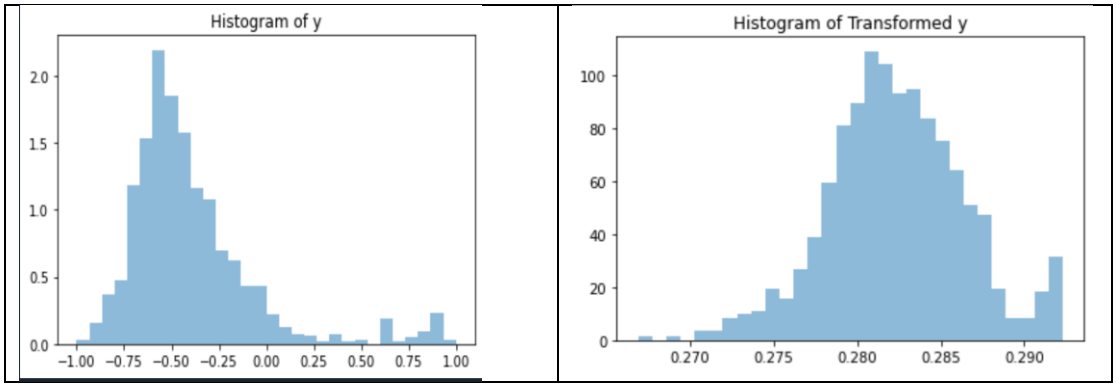
41 42 59 106 107 143 252 259 284 286 288 289 318 321 330 353 363

377 382 385 388 497 518 615 647 681 775 779 794 971

4.5 k-means

k-means method could also be used. First, we need to normalize these data to bring them to the same scale as we are dealing with data from spatial coordinates and mutation frequency. Histograms above have shown that although x, y, z, and VN are all in the range [0, 1] after scaling, the distributions of these variables are quite different.

Num is highly right-skewed. This could potentially affect the performance of the clustering algorithm, as it may assign a lot of emphasis to VN and group those high-frequency points in one cluster, regardless of their coordinates. To address this issue, we choose to use the Box-Cox method to transform the four variables and make them closer to a normal distribution.



We then apply the K-means method to perform clustering in this 4D coordinate space. Here, we have designed an algorithm to determine

the optimal number of clusters for the K-means method. Detailed methods and results could be found on the summary of my teammate.

5Result and analysis

We will Compare the results for hotspots based on the indices and position in 3D structure.

1. The relative position of hotspots on 1D sequence

Basic WAP-Sphere:

WAP-Sphere+Intersect:(red for area3, purple for area2, blue for area1)

Greedy WAP-Sphere+Intersect

High center+WAP for RF

K-means(red for cluster 61, orange for cluster 33 and 65)

2. The indices (in 972) of the five methods

Indices (on the 972 set):

The indices are the center of the spheres selected.

Basic WAP-Sphere:

8, 10, 124, 176, 305, 315, 407, 427, 582, 578, 593, 625, 629, 633, 656, 670, 696, 698, 699, 702, 704, 705, 707, 854, 859, 874, 875, 877, 879, 881

WAP-Sphere+Intersect:

Area3: 704 705 706 707 708

Area2: 878 879 880 876 877 881 882

Area1: 701 702 703 9 10 11 7 8 628 629 630 627 581 582 583 426 427 428 211 212 213 905 906 907 672 673 674 322 323 324

Greedy WAP-Sphere+Intersect:

Area3: 701 702 703 704 705 706 707 708

Area2: 878 879 880 876 877 881 882

Area1: 9 10 11 7 8 628 629 630 627 581 582 583 426 427 428 211 212 213 905 906 907 672 673 674 322 323 324

High center+WAP for RF:

41 42 59 106 107 143 252 259 284 286 288 289 318 321 330 353 363 377 382 385 388 497 518 615 647 681 775 779 794 971

K-means:

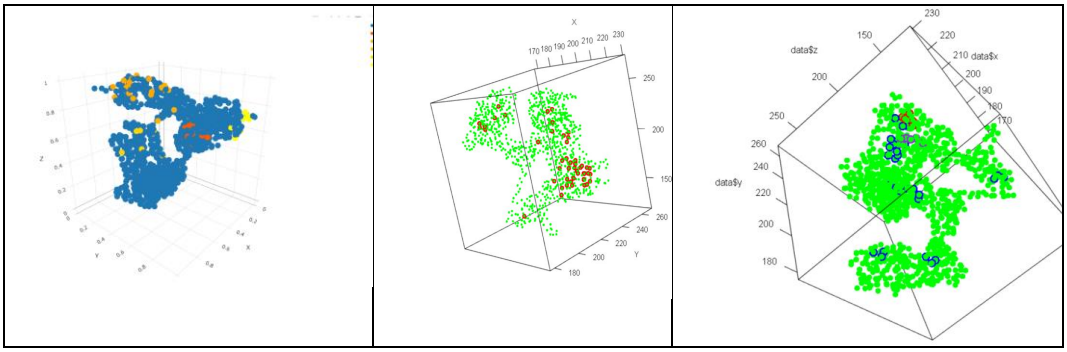
cluster33: 36 52 53 54 55 221 222 272 276 289 290 296 299 302 306

cluster61: 375 376 405 408 410 417 440 444 452 493 496 498 501 505

cluster65: 339 346 368 371 373 547 570

3. The position of hotspots on 3D structure

k-means | Basic WAP-Sphere | WAP-Sphere+Intersect



4. Analyze from the results

It seems the first two methods give a quite similar answer: statistical shows that the same indices from two methods are {8, 9, 10, 426, 427, 428, 701, 702, 703, 704, 705, 707, 581, 582, 876, 877, 879, 881, 627, 628, 629}, giving a similarity of 0.568. If we further consider gap between two indices less than two can be considered as 'the same', the similarity even increases. The similar of the results for the first two answers are not surprise as the second method is derived from the first one.

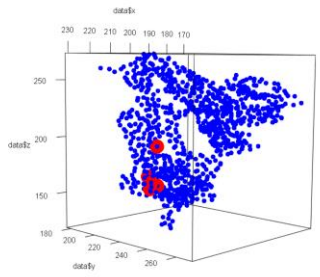
6.Discussion

1. Long-distance interaction analysis:

We first do the long-distance interaction analysis here. As we make our analysis based on sphere and distance, we could easily find those pair of points with close in Euclidean distance but far in 1D sequence. Those pair might show interactions under biology perspective. Intuitively speaking, this may refer to a place where overlapping happens. We search for pair of points which distance are less than 10 but index differences are bigger than 100 and both have virus number bigger than 5000. The result is shown below.

Indices of points that satisfy the criteria:

583 702 704 707 708 (all appear in the second methods)



2. Significance of hotspots based on permutation test.

The basic idea behind the permutation test is to randomly shuffle the labels of the data points and calculate a test statistic under the null hypothesis. By repeating this process multiple times and comparing

the observed test statistic with the null distribution of test statistics obtained from the random permutations, we can determine the p-value or the statistical significance of the observed hotspots.

3. How to compare the result of hotspots from different methods remained a difficult problem. Maybe we should combine empirical experiments with statistical criteria like p-value from permutation test to manage the problem.