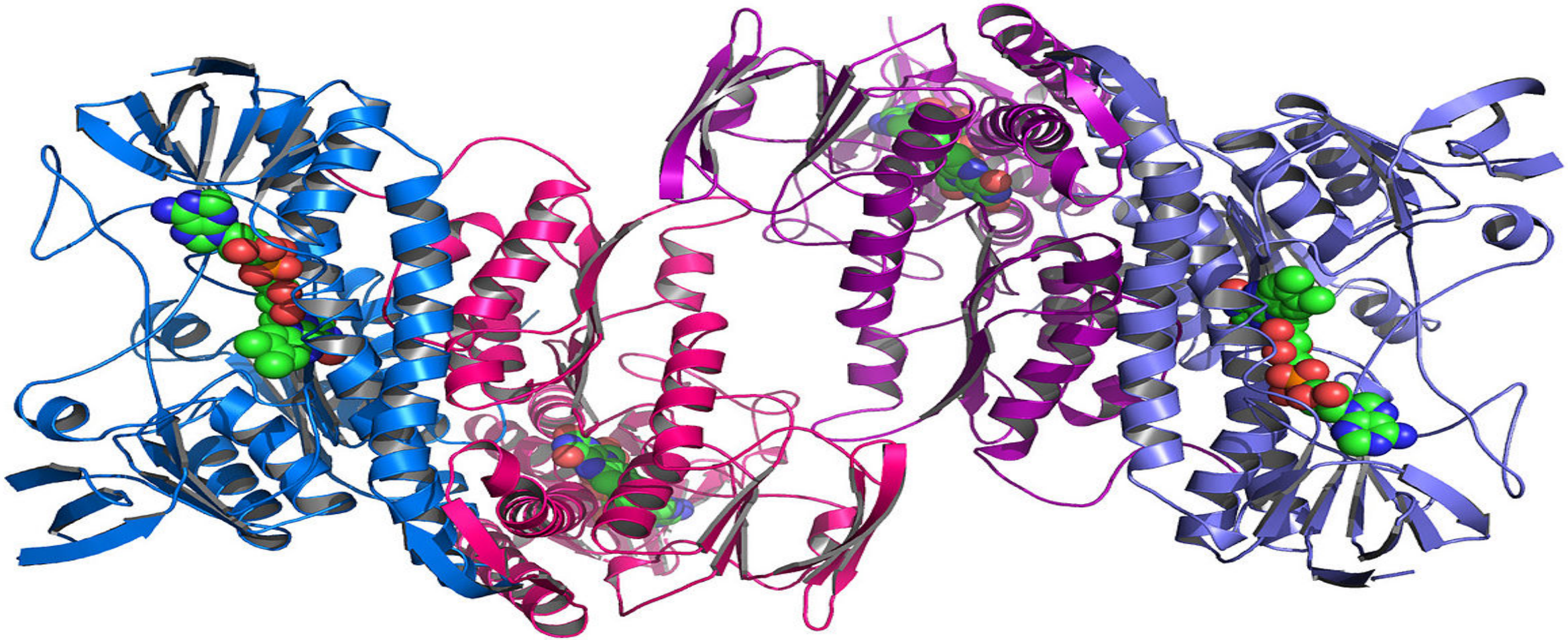
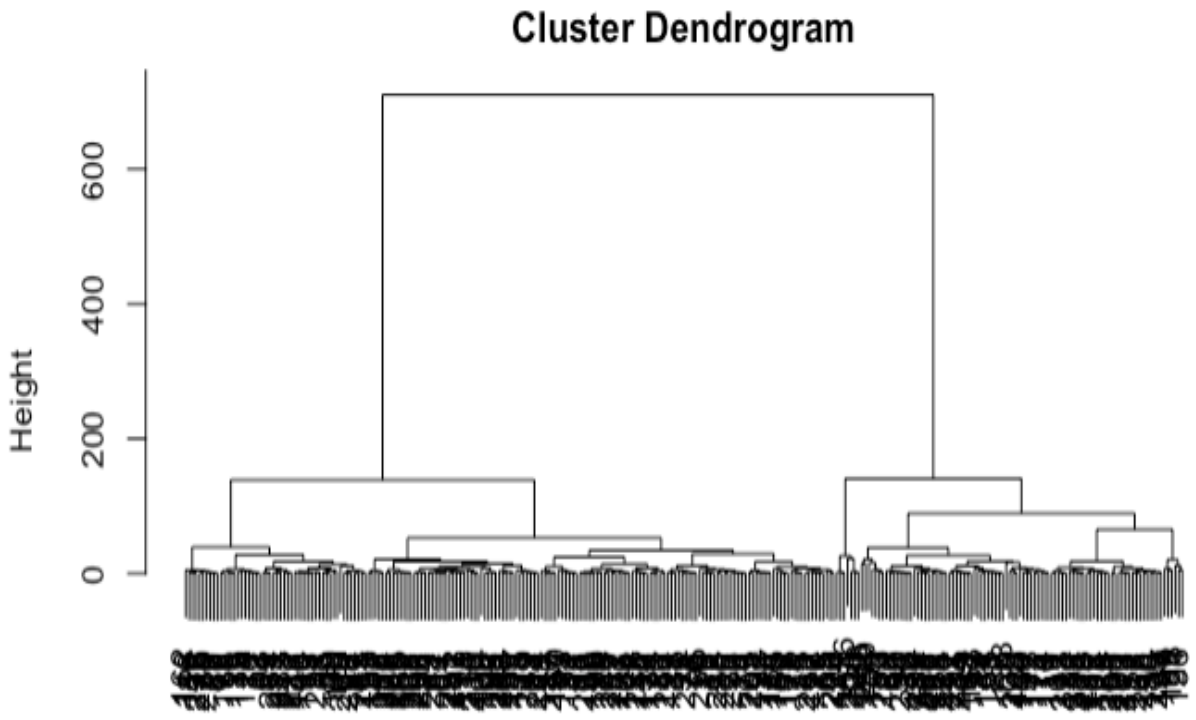
PROTEIN CLUSTERING FOR A. MARGINALE

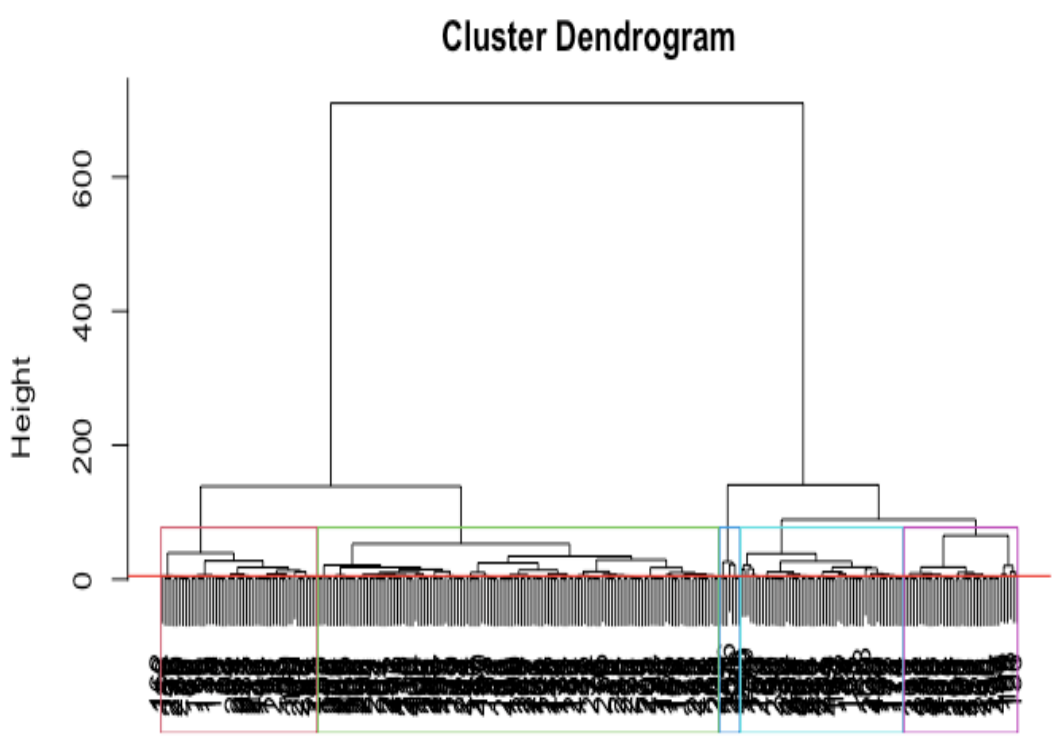
(ZHCHENG GU, YI CHOU)

THE FILE WE USE : AMARGREPEATS.FASTA



HIERARCHICAL CLUSTERING BASED ON THE ALIGNMENT SCORE





BECAUSE WE HAVE TO MANY DATA SO WE CANNOT CLEARLY SEE WHAT HAPPENED ON THE BOTTOM, SO WE WILL CHOOSE THE SAMPLE WHEN K=5 TO CALCULATE THE AVERAGE AND THIS IS THE CUT AVERAGE WE GET FOR THE FIRST CLUSTER DENDROGRAM.

When the k=5

```
[1] "k= 5 "
cut_avg
  1   2   3   4   5
 84 186   1   7   6
```

Put the result we get into the origin data set as a column

	nb	nam	seq	com	y1
1	284	A	ddsssasgqqqessvssqseastssqlg	NA	1
2	284	B	adsssaggqqqessvssqsdqastssqlg	NA	2
3	284	C	adsssaggqqqessvssqsgqastssqlg	NA	2
4	284	D	adsssasgqqqessvssqseastssqlgg	NA	1
5	284	E	adsssasgqqqessvssqseastssqlg	NA	1
6	284	F	tdsssasgqqqessvssqsgqastssqlg	NA	2

And the y1 is the classify result we get

K-MEAN CLUSTERING

- Use one-hot encoding sequence, turn it into the matrix with 0 and 1.

	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	f17	f18	f19	f20	f21	f22	f23	f24	f25	f26	f27	f28	f29	f30	f31
A	d	d	s	s	s	a	s	g	q	q	q	e	s	s	v	s	s	q	s	e	a	s	t	s	s	q	l	g	o	o	o
B	a	d	s	s	s	a	g	g	q	q	q	e	s	s	v	s	s	q	s	d	q	a	s	t	s	s	q	l	g	o	o
C	a	d	s	s	s	a	g	g	q	q	q	e	s	s	v	s	s	q	s	g	q	a	s	t	s	s	q	l	g	o	o
D	a	d	s	s	s	a	s	g	q	q	q	e	s	s	v	s	s	q	s	e	a	s	t	s	s	q	l	g	g	o	o
E	a	d	s	s	s	a	s	g	q	q	q	e	s	s	v	s	s	q	s	e	a	s	t	s	s	q	l	g	o	o	o
F	t	d	s	s	s	a	s	g	q	q	q	e	s	s	v	s	s	q	s	g	q	a	s	t	s	s	q	l	g	o	o

And then take the Euclidean distance, and use k-means clustering.
The y2 is the classify result we get.

	nb	nam	seq	com	y1	y2
1	284	A	ddsssasgqqqessvssqseastssqlg	NA	1	4
2	284	B	adsssaggqqqessvssqsdqastssqlg	NA	2	2
3	284	C	adsssaggqqqessvssqsgqastssqlg	NA	2	2
4	284	D	adsssasgqqqessvssqseastssqlgg	NA	1	4
5	284	E	adsssasgqqqessvssqseastssqlg	NA	1	4
6	284	F	tdsssasgqqqessvssqsgqastssqlg	NA	2	1

TEXT FEATURE EXTRACTION BASED ON TERM FREQUENCY

- Count different letters
- Count the number of occurrences of different letters in each sequence to get the document-term matrix.
- perform k-means clustering after standardization above the matrix with dtm.

	a	c	d	e	f	g	h	i	k	l	m	n	p	q	r	s	t	v	w	x	y
A	2	0	2	2	0	2	0	0	0	1	0	0	0	5	0	12	1	1	0	0	0
B	3	0	2	1	0	3	0	0	0	1	0	0	0	6	0	11	1	1	0	0	0
C	3	0	1	1	0	4	0	0	0	1	0	0	0	6	0	11	1	1	0	0	0
D	3	0	1	2	0	3	0	0	0	1	0	0	0	5	0	12	1	1	0	0	0
E	3	0	1	2	0	2	0	0	0	1	0	0	0	5	0	12	1	1	0	0	0
F	2	0	1	1	0	3	0	0	0	1	0	0	0	6	0	12	2	1	0	0	0
G;39	2	0	2	1	0	3	0	0	0	0	0	0	0	6	0	13	1	1	0	0	0
H	2	0	1	1	0	3	0	0	0	0	0	0	0	6	0	13	2	1	0	0	0
I	2	0	2	1	0	3	0	0	0	1	0	0	0	6	0	12	1	1	0	0	0
J	3	0	2	1	0	3	0	0	0	2	0	0	0	6	0	10	1	1	0	0	0
Australian	3	0	2	1	0	4	0	0	0	0	0	0	0	5	0	11	1	1	0	0	0
K;S	3	0	2	1	0	4	0	0	0	1	0	0	0	6	0	10	1	1	0	0	0
L	3	0	2	1	0	3	0	0	0	1	0	0	0	6	0	11	1	1	0	0	0
M;UP47	3	0	1	1	0	3	0	0	0	1	0	0	0	6	0	12	1	1	0	0	0
N	2	0	2	1	0	2	0	0	0	1	0	0	0	6	0	12	2	1	0	0	0
O	2	0	1	1	0	3	0	0	0	1	0	0	0	6	0	9	1	1	0	0	0
Q	3	0	2	1	0	2	0	0	0	1	0	0	0	6	0	12	1	1	0	0	0
R	3	0	2	1	0	2	1	0	0	1	0	0	0	5	0	11	1	1	1	0	0
T	3	0	1	1	0	4	0	0	0	1	0	0	0	6	0	11	1	1	0	0	0
U	2	0	3	1	0	2	0	0	0	1	0	0	0	6	0	12	1	1	0	0	0

The y3 is the classify result we get

	nb	nam	seq	com	y1	y2	y3
1	284	A	ddsssasgqqqessvssqseastssqlg	NA	1	4	5
2	284	B	adsssaggqqqessvssqsdqastssqlg	NA	2	2	5
3	284	C	adsssaggqqqessvssqsgqastssqlg	NA	2	2	5
4	284	D	adsssasgqqqessvssqseastssqlgg	NA	1	4	5
5	284	E	adsssasgqqqessvssqseastssqlg	NA	1	4	5
6	284	F	tdsssasgqqqessvssqsgqastssqlg	NA	2	1	1

Compare the three results we get from each clustering method we try

TABLE(AMARGREPEATS$Y1,AMARGREPEATS$Y2)
TABLE(AMARGREPEATS$Y1,AMARGREPEATS$Y3)
TABLE(AMARGREPEATS$Y2,AMARGREPEATS$Y3)

The final form we got:

Hc vs km1

	1	2	3	4	5
1	1	0	31	51	1
2	135	51	0	0	0
3	0	0	0	0	1
4	0	0	5	0	2
5	0	0	0	0	6

Hc vs. km2

	1	2	3	4	5
1	25	3	6	9	41
2	83	6	7	20	70
3	0	0	0	0	1
4	1	1	2	0	3
5	0	0	0	0	6

Km1 vs. km2

	1	2	3	4	5
1	69	6	4	10	47
2	15	0	3	10	23
3	11	0	4	8	13
4	14	3	4	1	29
5	0	1	0	0	9

OUR CONCLUSION:

- From the previous slides we can see, y2 an d y3 even we use same clustering method but the result we get is really different.
- Y2 is based on the Euclidean distance and y3 is based on the term frequency, It shows that the clustering model relies heavily on distance metrics, and different distance metrics produce different clustering results.
- Regardless of the clustering method, when there is a heterogeneous SSR in a given locus, it can help us find different repetitive patterns and then type genotypic strains based on the SSR. For this data set the hierarchical clustering is more suitable.