生物信息学之软件分析平台

报告人：苏晓峰

制作:一班E队集体

（孟志刚、苗猛猛、孙豹、邹良平、徐明、张怡、苏晓峰、张健飞、王金辉）

一、生物信息学

生物信息学（Bioinformatics）：是在生命科学的研究中，以计算机为工具对生物信息进行储存、检索和分析的科学。它是当今生命科学和自然科学的重大前沿领域之一，同时也将是21世纪自然科学的核心领域之一。

生物信息学网站 http://abc.cbi.pku.edu.cn/内含有丰富的资源，我们这里着重对里面的生物学软件分析平台进行讲解，进入ABC主页后，可在其右侧打开Tools，里面有好多的软件包，在此我们以 EMBOSS explore为例进行演示：

EMBOSS explore的应用

1. 对输入信息的加工分析
2. 对基因的分析
3. 对蛋白质的氨基酸序列性质的分析
4. 对蛋白质的氨基酸序列或核酸的核苷酸序列的相似性分析
5. 对蛋白质一级结构的分析
6. 对蛋白质二级结构的分析
7. 对蛋白质三级结构的分析
8. 对蛋白质进行酶学分析
9. 对多个序列之间进化关系上的分析
10. 其他软件功能

1. 对输入信息的加工分析

coderet 可以把输入的信息进行整合加工，再以更直观的形式表现出来。输入的时候要把其基因的说明信息等都要输入，而不能只输入核苷酸或氨基酸序列，否则只输出序列的个数，没有意义。

以NCBI中的NM_000517为例进行操作：

OUTPUT FILE outfile

CDS	mRNA	non-c	Trans	Total	Sequence
1	0	4	1	6	NM_000517

OUTPUT FILE cdsoutseq

>nm_000517_cds_1
atggtgctgtctcctgccgacaagaccaacgtcaaggccgcctggggtaaggtcggcgcg
cacgctggcgagtatggtgcggaggccctggagaggatgttcctgtccttccccaccacc
aagacctacttcccgcacttcgacctgagccacggctctgcccaggttaagggccacggc
aagaaggtggccgacgcgctgaccaacgccgtggcgcacgtggacgacatgcccaacgcg
ctgtccgccctgagcgacctgcacgcgcacaagcttcgggtggacccggtcaacttcaag
ctcctaagccactgcctgctggtgaccctggccgcccacctccccgccgagttcaccgct
gcggtgcacgcctccctggacaagttcctggcttctgtgagcaccgtgctgacctccaaa
taccgttaa

OUTPUT FILE translationoutseq

>nm_000517_pro_1
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR

OUTPUT FILE restoutseq

>nm_000517_noncoding_606
aaaaaaaaaaaaaaaaa

- *Seqretsplit*其可以把一起输入的多个核酸或氨基酸序列进行拆分，便于我们的操作，这样可以节省时间。

- >Human - HBA_HUMAN Hemoglobin alpha - Homo sapiens (Human).
- MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQV KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHL PAEFTPAVHASLDKFLASVSTVLTSKYR
- >Mouse - HBA_MOUSE Hemoglobin alpha - Mus musculus (Mouse).
- MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVK GHGKKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHP ADFTPAVHASLDKFLASVSTVLTSKYR
- >Dolphin - HBA_TURTR Hemoglobin alpha - Tursiops truncatus (Atlantic bottle-nosed dolphin).
- MVLSPADKTNVKGTWSKIGNHSAEYGAEALERMFINFPSTKTYFSHFDLGHGSAQIKG HGKKVADALTKAVGHIDNLPDALSELSDLHAHKLRVDPVNFKLLSHCLLVTLALHLPAD FTPSVHASLDKFLASVSTVLTSKYR
- >Chicken - HBA_CHICK Hemoglobin alpha-A - Gallus gallus (Chicken).
- MVLSAADKNNVKGIFTKIAGHAEEYGAETLERMFTTYPPTKTYFPHFDLSHGSAQIKG HGKKVVAALIEAANHIDDIAGTLSKLSDLHAHKLRVDPVNFKLLGQCFLVVVAIHHPAAL TPEVHASLDKFLCAVGTVLTAKYR
- >Snake - HBA_DRYCE Hemoglobin alpha-A - Drymarchon corais erebennus (Texas indigo snake).
- MVLTEEDKSRVRAAWGPVSKNAELYGAETLTRLFTAYPATKTYFHHFDLSPGSSNLKT HGKKVIDAITEAVNNLDDVAGALSKLSDLHAQKLRVDPVNFKLLGHCLEVTIAAHNGGP LKPEVILSLDKFLCLVAKTLVSRYR
- >Frog - HBA1_XENLA Hemoglobin subunit alpha-1 - Xenopus laevis (African clawed frog).
- MLLSADDKKHIKAIMPAIAAHGDKFGGEALYRMFIVNPKTKTYFPSFDFHHNSKQISAH GKKVVDALNEASNHLDNIAGSMSKLSDLHAYDLRVDPGNFPLLAHNILVVVAMNFPKQ FDPATHKALDKFLATVSTVLTSKYR
- >Goldfish - HBA_CARAU Hemoglobin alpha - Carassius auratus (Goldfish).
- MSLSDKDKAVVKALWAKIGSRADEIGAEALGRMLTVYPQTKTYFSHWSDLSPGSGPV KKHGKTIMGAVGDAVSKIDDLVGALSALSELHAFKLRIDPANFKILAHNVIVVIGMLFPG DFTPEVHMSVDKFFQNLALALSEKYR

>Human - HBA_HUMAN Hemoglobin alpha - Homo sapiens (Human).
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR

OUTPUT FILE mouse fasta

>Mouse - HBA_MOUSE Hemoglobin alpha - Mus musculus (Mouse).
MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG
KKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPADFTP
AVHASLDKFLASVSTVLTSKYR

OUTPUT FILE dolphin fasta

>Dolphin - HBA_TURTR Hemoglobin alpha - Tursiops truncatus (Atlantic bottle-nosed dolphin).
MVLSPADKTNVKGTWSKIGSHSAEYGAEALERMFINFPSTKTYFSHFDLGHGSAQIKGHG
KKVADALTKAVGHIDNLPDALSELSDLHAHKLRVDPVNFKLLSHCLLVTLALHLPADFTP
SVHASLDKFLASVSTVLTSKYR

OUTPUT FILE chicken fasta

>Chicken - HBA_CHICK Hemoglobin alpha-A - Gallus gallus (Chicken).
MVLSAADKSNVKGIFTKIAGHAEEYGAETLERMFTTYPPTKTYFPHFDLSHGSAQIKGHG
KKVVAALIEAANHIDDIAGTLSKLSDLHAHKLRVDPVNFKLLGQCFLVVVAIHHPAALTP
EVHASLDKFLCAVGTVLTAKYR

OUTPUT FILE snake fasta

>Snake - HBA_DRYCE Hemoglobin alpha-A - Drymarchon corais erebennus (Texas indigo snake).
MVLTEEDKSNVAAAWGPVSKNAELYGAETLTRLFTAYPATKTYFHHFDLSPGSSQLKTHG
KKVIDAITEAVNHLDDVAGALSKLSDLHAQKLRVDPVNFKLLGHCLEVTIAAHNCGPLKP
EVILSLDKFLCLVAKTLVSMYR

OUTPUT FILE frog fasta

>Frog - HBA1_XENLA Hemoglobin subunit alpha-1 - Xenopus laevis (African clawed frog).
MLLSADDKKHIKAIWPAIAAHGDEFGCEALYRMFIVNPKTKTYFPSFDFHHNSEQISAHG
KKVVDALNEAVNHLDNIAGSMSKLSDLHAYDLRVDPGNFPLLAHNILVVVAAHFPKQFDP
ATHKALDKFLATVSTVLTSKYR

OUTPUT FILE goldfish fasta

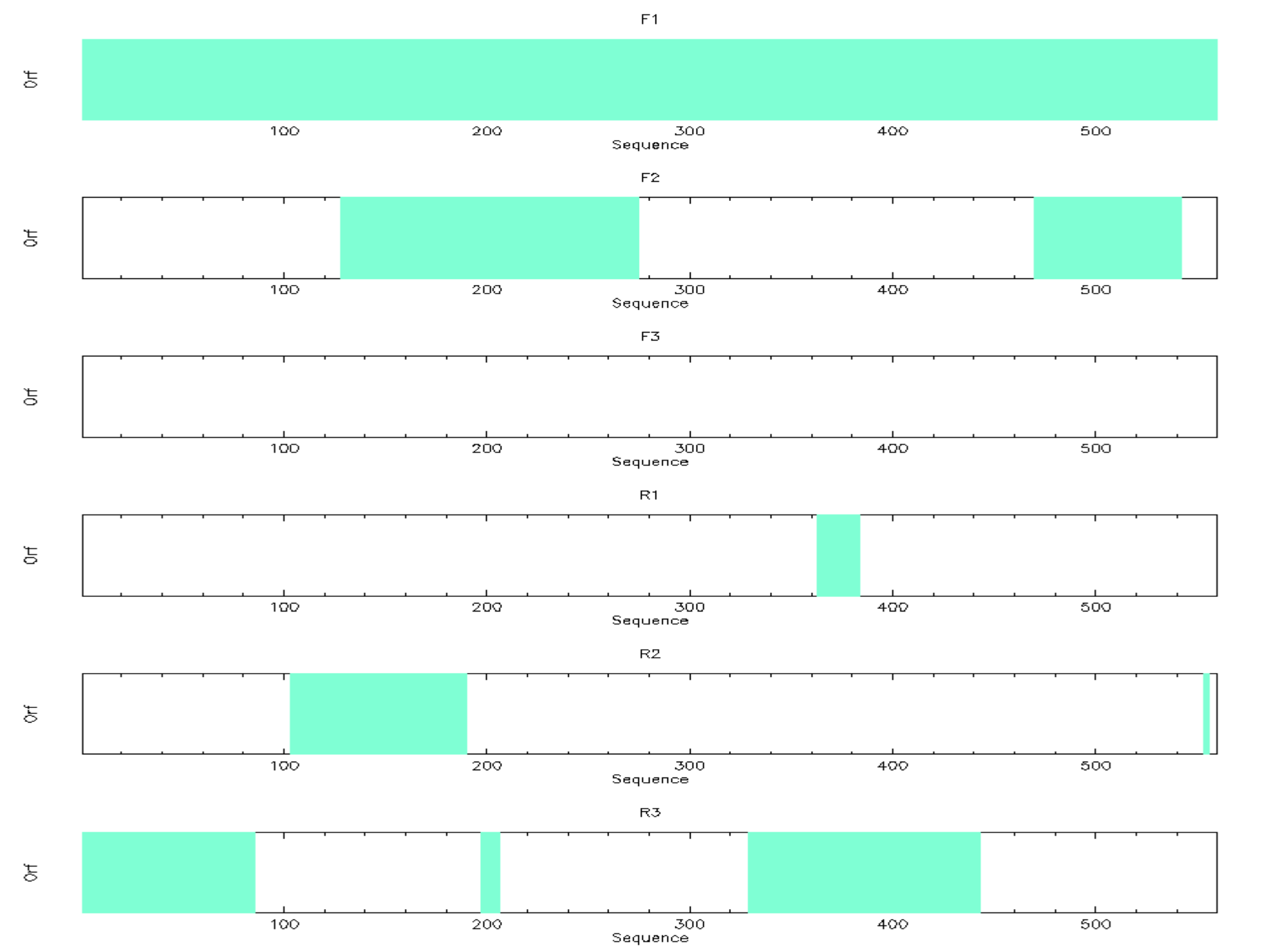
>Goldfish - HBA_CARAU Hemoglobin alpha - Carassius auratus (Goldfish).
MSLSDKDKAVVKALWAKISPKADEIGAEALGRMLTVYPQTKTYFSHWSDLSPGSGPVKKH
GKTIMGAVGDAVSKIDDLVGALSALSELHAFKLRVDPANFKILAHNVIVVIGMLFPGDFT
PEVHMSVDKFFQNLALALSEKYR

2.对基因的分析

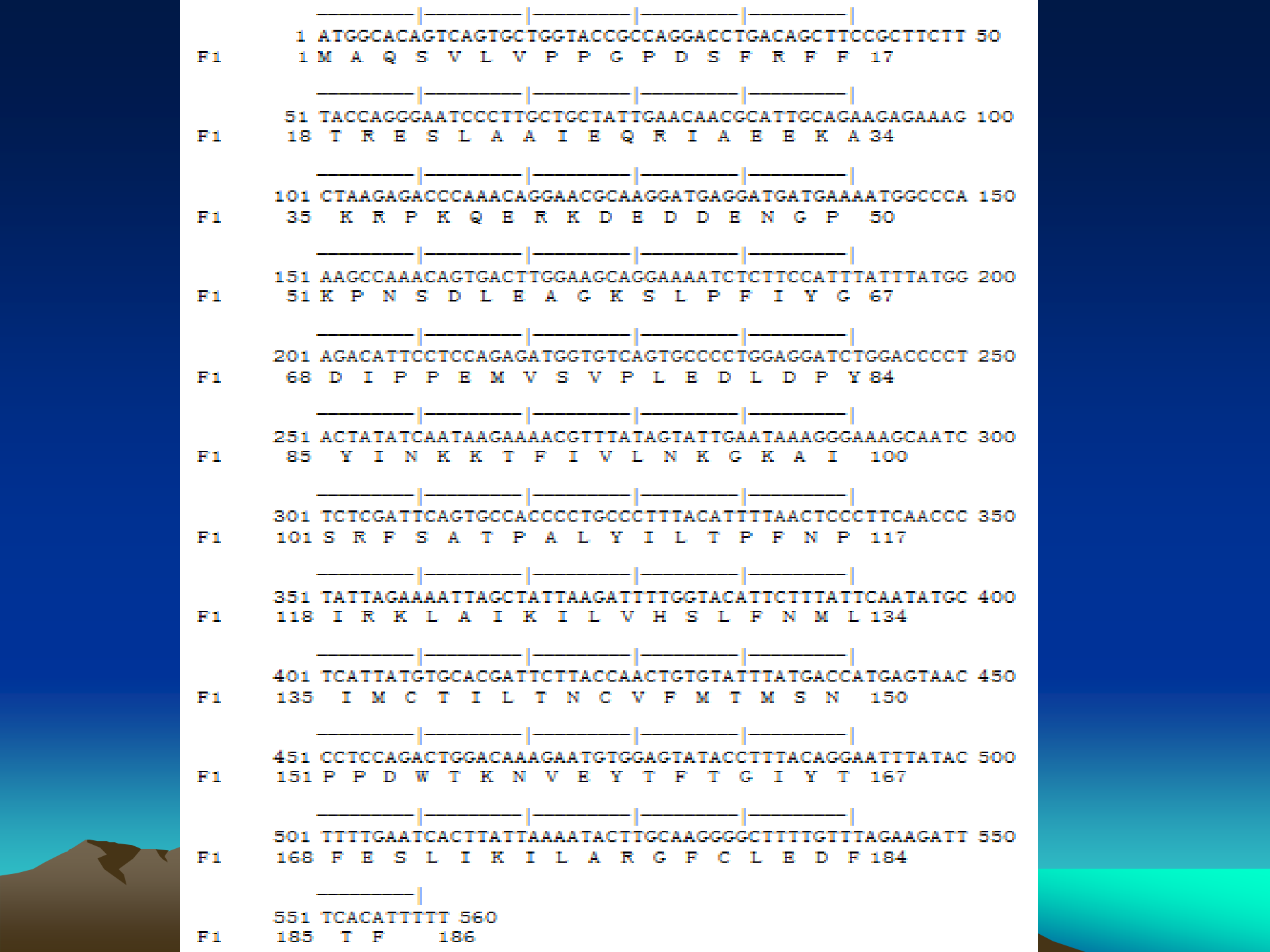
以下我们以此基因为例进行一系列的操作：

- ATGGCACAGTCAGTGCTGGTACCGCCAGGACCTGACAGCTTCC
  GCTTCTTTACCAGGGAATCCCTTGCTGCTATTGAACAACGCATT
  GCAGAAGAGAAAGCTAAGAGACCCAAACAGGAACGCAAGGATG
  AGGATGATGAAAATGGCCCAAAGCCAAACAGTGACTTGGAAGC
  AGGAAAATCTCTTCCATTTATTTATGGAGACATTCCTCCAGAGAT
  GGTGTCAGTGCCCCTGGAGGATCTGGACCCCTACTATATCAAT
  AAGAAAACGTTTATAGTATTGAATAAAGGGAAAGCAATCTCTCG
  ATTCAGTGCCACCCCTGCCCTTTACATTTTAACTCCCTTCAACC
  CTATTAGAAAATTAGCTATTAAGATTTTGGTACATTCTTTATTCAA
  TATGCTCATTATGTGCACGATTCTTACCAACTGTGTATTTATGAC
  CATGAGTAACCCTCCAGACTGGACAAAGAATGTGGAGTATACCT
  TTACAGGAATTTATACTTTTGAATCACTTATTAAAATACTTGCAAG
  GGGCTTTTGTTTAGAAGATTTCACATTTTT

- *plotorf* 用图形的形式来预测它的开放阅读框。可对你输入的序列进行分析，由于其仅仅是对其预测，所以把其可能的形式都以图形的形式表现出来，以防有所疏漏。注：输入的形式为核酸序列，不必加入其它信息。


F1
Orf
100
200
300
400
500
Sequence
F2
Orf
100
200
300
400
500
Sequence
F3
Orf
100
200
300
400
500
Sequence
R1
Orf
100
200
300
400
500
Sequence
R2
Orf
100
200
300
400
500
Sequence
R3
Orf
100
200
300
400
500
Sequence

- *Showorf*是把我们输入的核酸序列翻译成蛋白质的氨基酸序列。其有6种方式可以选择R1、R2、R3、F1、F2和F3等6种方式可对它所翻译出来的序列方式进行预测。注：R为reverse，F为forward。其为从正向或反向第几个核苷酸序列进行翻译。


1 ATGGCACAGTCAGTGCTGGTACCGCCAGGACCTGACAGCTTCCGCTTCTT 50
F1 1 M A Q S V L V P P G P D S F R F F 17
51 TACCAGGGAATCCCTTGCTGCTATTGAACAACGCATTGCAGAAGAGAAAG 100
F1 18 T R E S L A A I E Q R I A E E K A 34
101 CTAAGAGACCCAAACAGGAACGCAAGGATGAGGATGATGAAAATGGCCCA 150
F1 35 K R P K Q E R K D E D D E N G P 50
151 AAGCCAAACAGTGACTTGGAAGCAGGAAAATCTCTTCCATTTATTTATGG 200
F1 51 K P N S D L E A G K S L P F I Y G 67
201 AGACATTCCTCCAGAGATGGTGTCAGTGCCCCTGGAGGATCTGGACCCCT 250
F1 68 D I P P E M V S V P L E D L D P Y 84
251 ACTATATCAATAAGAAAACGTTTATAGTATTGAATAAAGGGAAAGCAATC 300
F1 85 Y I N K K T F I V L N K G K A I 100
301 TCTCGATTCAGTGCCACCCCTGCCCTTTACATTTTAACTCCCTTCAACCC 350
F1 101 S R F S A T P A L Y I L T P F N P 117
351 TATTAGAAAATTAGCTATTAAGATTTTGGTACATTCTTTATTCAATATGC 400
F1 118 I R K L A I K I L V H S L F N M L 134
401 TCATTATGTGCACGATTCTTACCAACTGTGTATTTATGACCATGAGTAAC 450
F1 135 I M C T I L T N C V F M T M S N 150
451 CCTCCAGACTGGACAAAGAATGTGGAGTATACCTTTACAGGAATTTATAC 500
F1 151 P P D W T K N V E Y T F T G I Y T 167
501 TTTTGAATCACTTATTAAAATACTTGCAAGGGGCTTTTGTTTAGAAGATT 550
F1 168 F E S L I K I L A R G F C L E D F 184
551 TCACATTTTT 560
F1 185 T F 186

*chips*依据某个特定的基因序列计算密码子偏爱性，计算结果为一个Nc值，该值越低，则密码子偏爱性越高，反之则越低。此序列的Nc值为：

OUTPUT FILE outfile

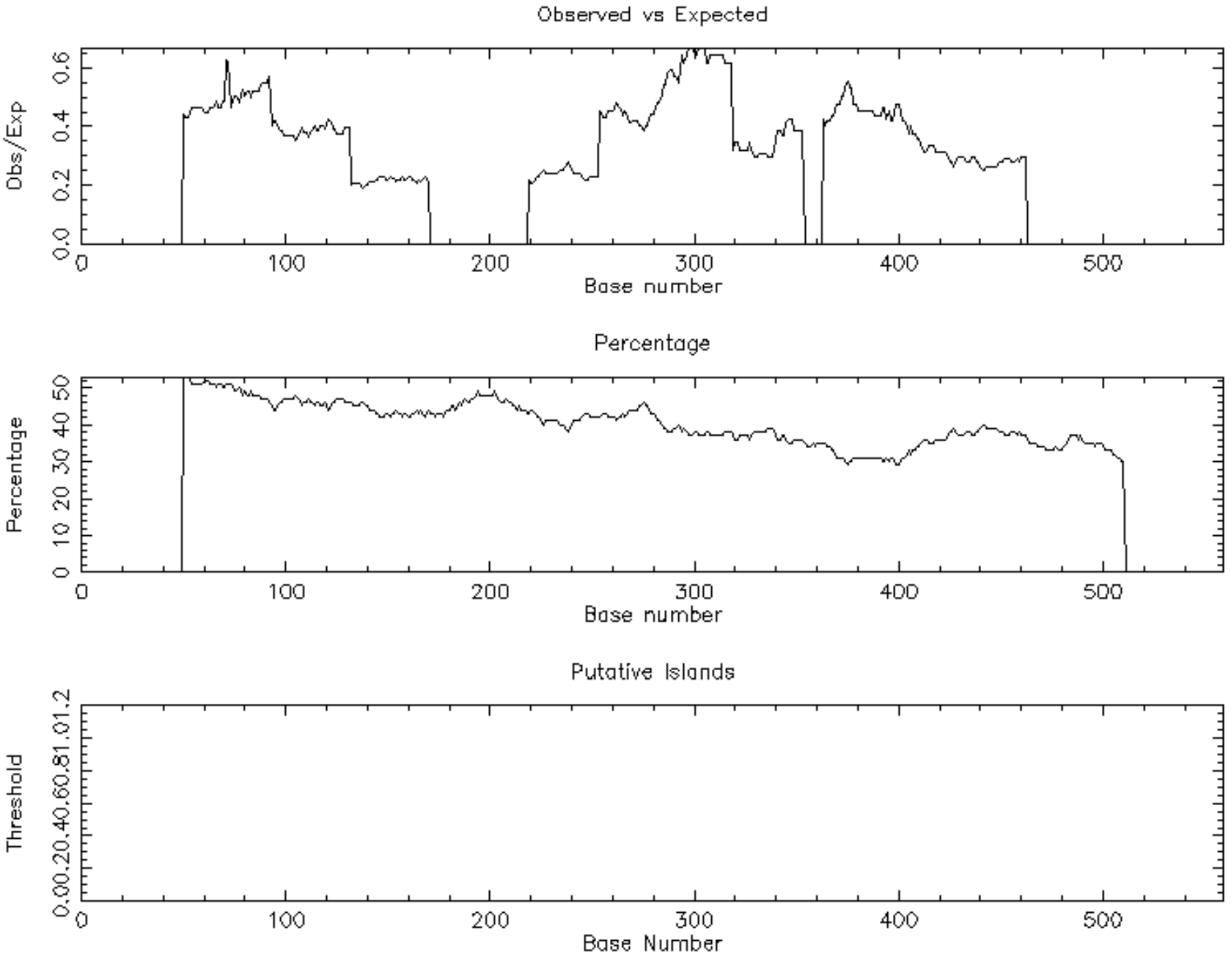
```
# CHIPS codon usage statistics

Nc = 57.312
```

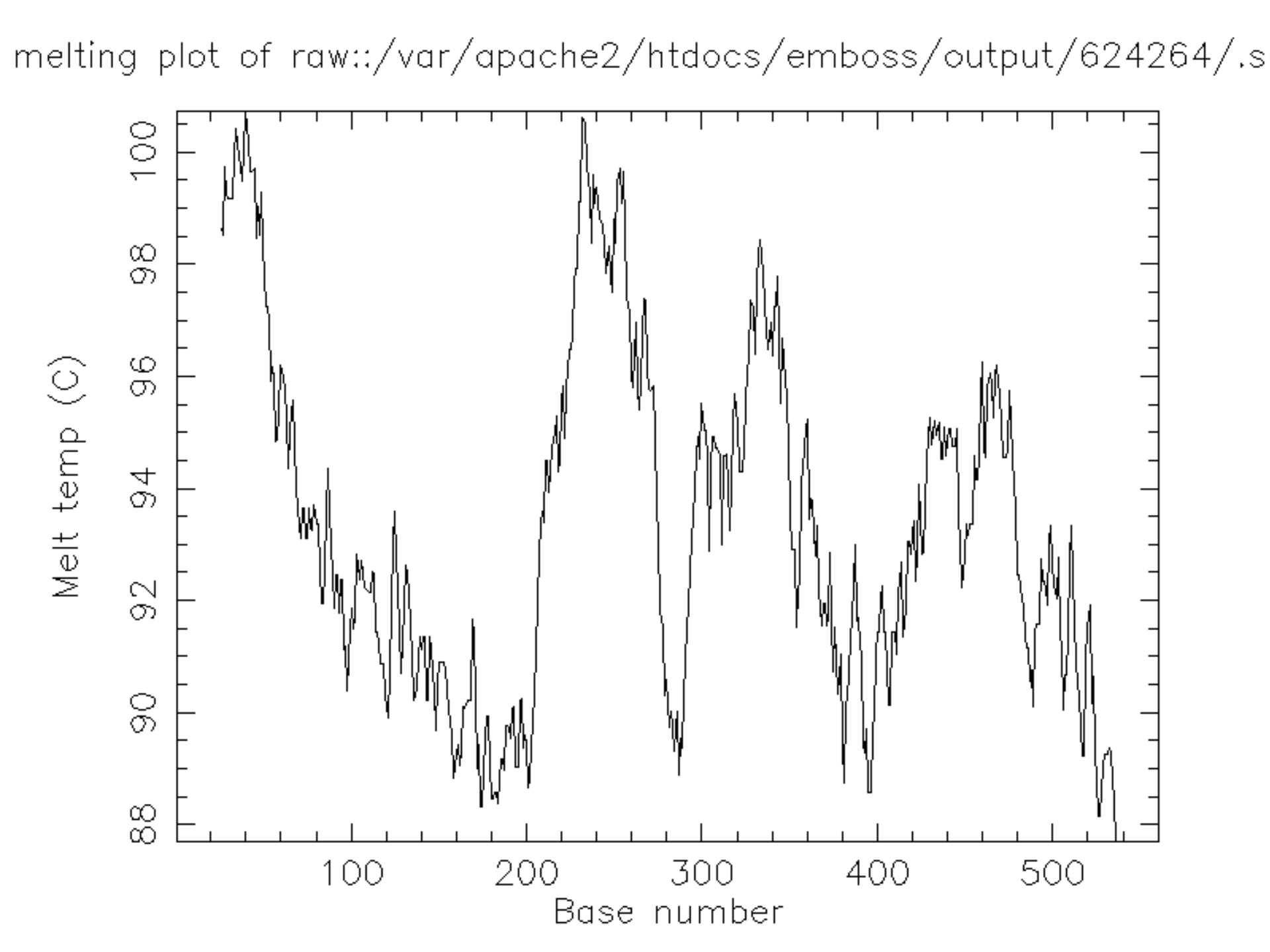
- *cpgplot*以图形文件和表格文件的形式表示核酸序列中CpG分布特征。由于CpG是基因组中高表达区域的特征，因此可以用来预测某个基因在基因组中的表达水平。

```
CPGPLOT islands of unusual CG composition
 from 1 to 560

      Observed/Expected ratio > 0.60
      Percent C + Percent G > 50.00
      Length > 200
```


Observed vs Expected
Obs/Exp
0.0
0.2
0.4
0.6
0
100
200
300
400
500
Base number
Percentage
Percentage
0
10
20
30
40
50
0
100
200
300
400
500
Base number
Putative Islands
Threshold
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
100
200
300
400
500
Base Number

*dan*计算DNA、RNA序列的熔点温度。该软件可用于southern blot、northern blot探针以及PCR引物的TM值的计算。

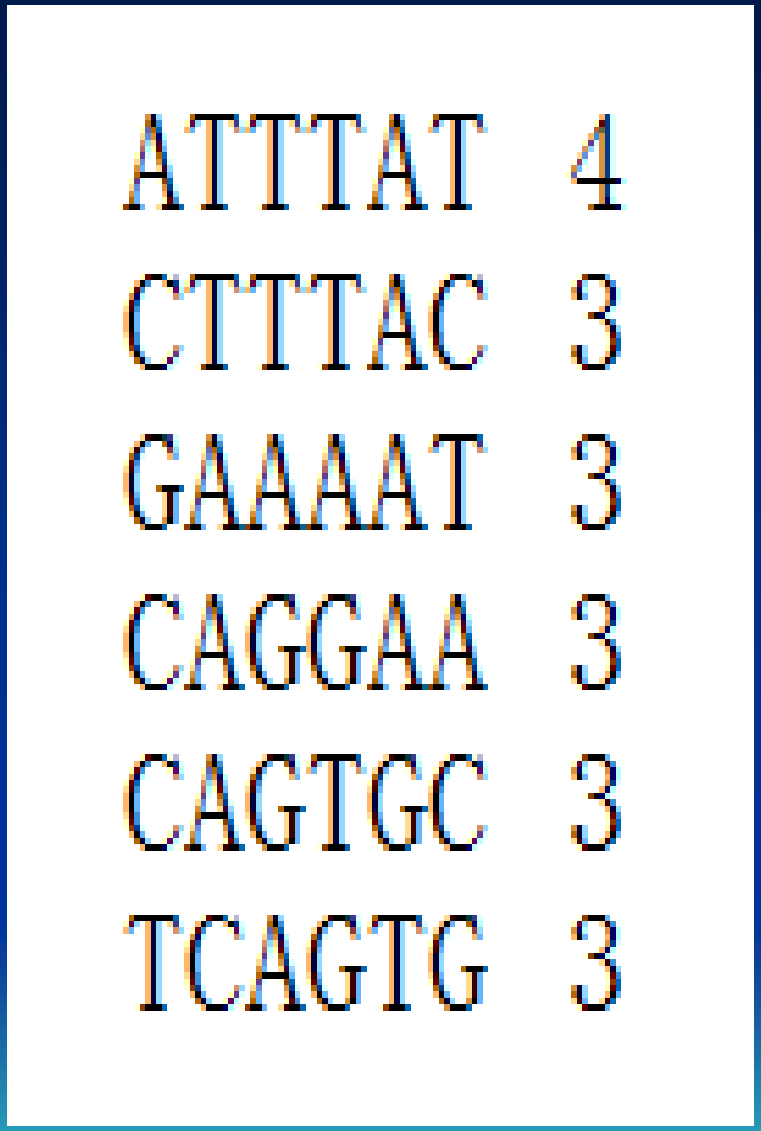

melting plot of raw::/var/apache2/htdocs/emboss/output/624264/.s
Melt temp (C)
88
90
92
94
96
98
100
100
200
300
400
500
Base number

geecee计算核苷酸的GC含量。输入所要计算的核苷酸序列，程序运行后可以得到G+C的百分含量。输入此序列为：

OUTPUT FILE outfile

#Sequence GC content
0.41

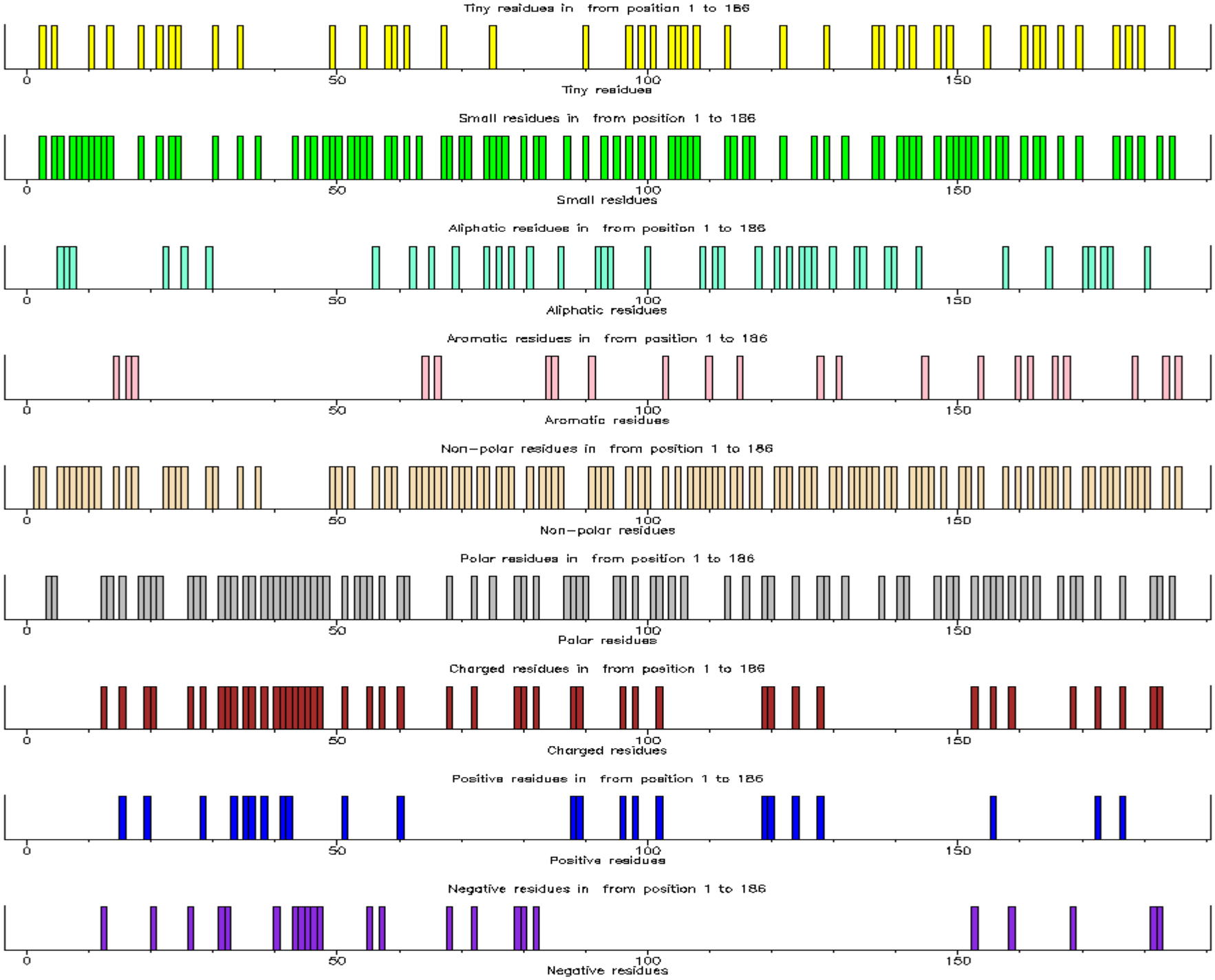
*Wordcount*在DNA序列中计算一定长度的连续序列在DNA序列中出现个数。可以选择相同序列的核苷酸个数，也可以选择>×的显示。

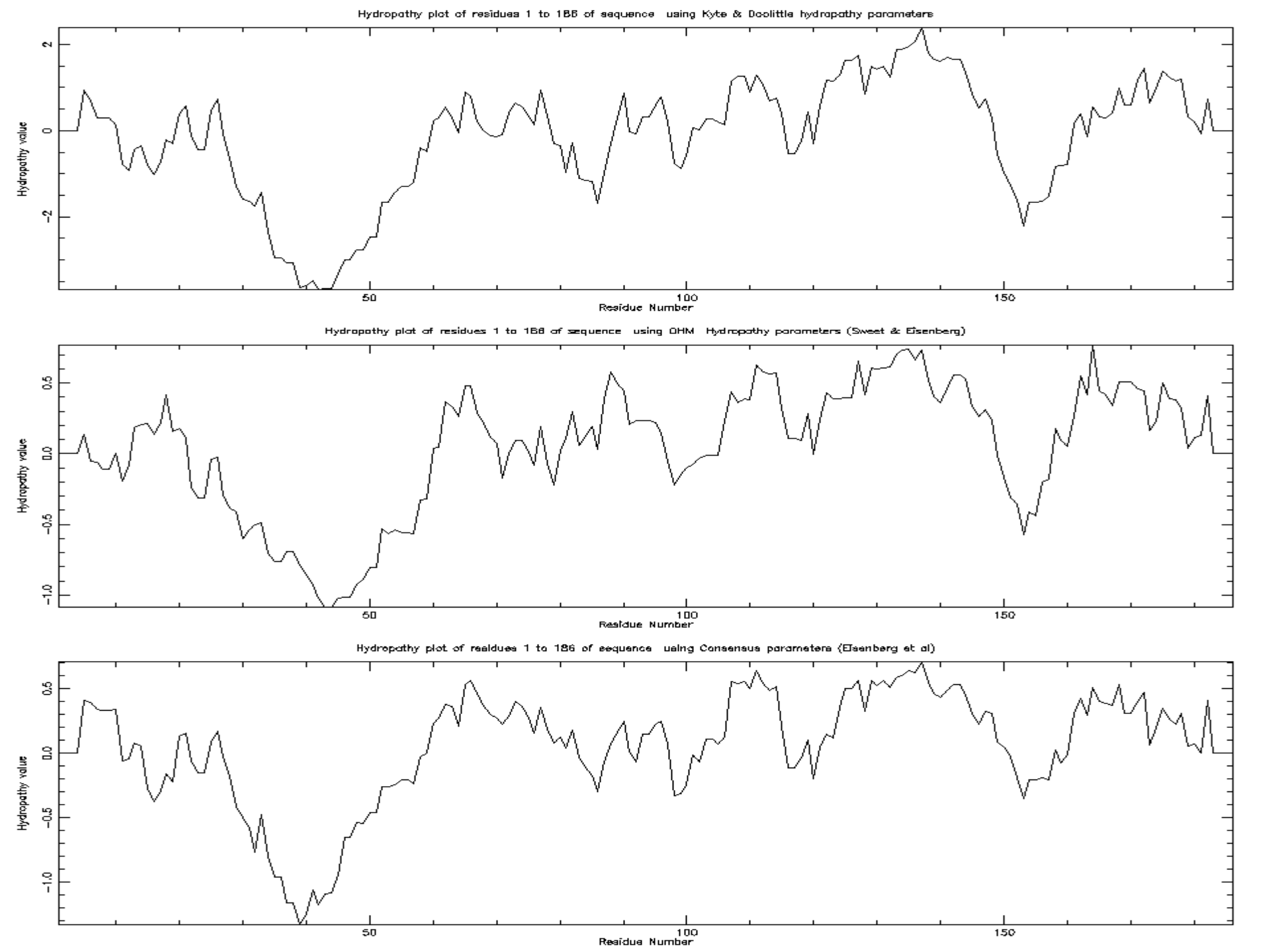


3. 对蛋白质的氨基酸序列性质的分析

- 以下以此氨基酸序列为例进行一系列的操作：
- MAQSVLVPPGPDSFRFFTRESLAAIE QRIAEEKAKRPKQERKDEDDENGPK PNSDLEAGKSLPFIYGDIPPEMVSVPL EDLDPYYINKKTFIVLNKGKAISRFSAT PALYILTPFNPIRKLAIKILVHSLFNMLI MCTILTNCVFMTMSNPPDWTKNVEY TFTGIYTFESLIKILARGFCLEDFTF

- pepinfo 能以图形方式显示蛋白质序列中各种不同性质的氨基酸残基的含量(较小R残基的氨基酸、小R残基的氨基酸、脂肪族的氨基酸、芳香族的氨基酸、带电荷的氨基酸、不带电荷的氨基酸、氨基酸对水的亲和程度等)，能够输出两张不同的图。


Tiny residues in from position 1 to 186
0
50
100
150
Tiny residues
Small residues in from position 1 to 186
0
50
100
150
Small residues
Aliphatic residues in from position 1 to 186
0
50
100
150
Aliphatic residues
Aromatic residues in from position 1 to 186
0
50
100
150
Aromatic residues
Non-polar residues in from position 1 to 186
0
50
100
150
Non-polar residues
Polar residues in from position 1 to 186
0
50
100
150
Polar residues
Charged residues in from position 1 to 186
0
50
100
150
Charged residues
Positive residues in from position 1 to 186
0
50
100
150
Positive residues
Negative residues in from position 1 to 186
0
50
100
150
Negative residues


Hydropathy plot of residues 1 to 186 of sequence using Kyte & Doolittle hydropathy parameters
Hydropathy value
2
0
-2
50
100
150
Residue Number
Hydropathy plot of residues 1 to 186 of sequence using OHM Hydropathy parameters (Sweet & Eisenberg)
Hydropathy value
0.5
0.0
-0.5
-1.0
50
100
150
Residue Number
Hydropathy plot of residues 1 to 186 of sequence using Consensus parameters (Eisenberg et al)
Hydropathy value
0.5
0.0
-0.5
-1.0
50
100
150
Residue Number

- *iep*计算蛋白的等电点。输入的是蛋白序列，以EBI数据库中laci_ecoli为例，得出pH、Bound、Charge等结果。

IEP of from 1 to 186
Isoelectric Point = 5.9915

pH	Bound	Charge
1.00	56.97	23.97
1.50	56.92	23.92
2.00	56.75	23.75
2.50	56.22	23.22
3.00	54.72	21.72
3.50	51.10	18.10
4.00	44.95	11.95
4.50	38.81	5.81
5.00	35.20	2.20
5.50	33.66	0.66
6.00	32.99	-0.01
6.50	32.54	-0.46
7.00	32.14	-0.86
7.50	31.73	-1.27
8.00	31.04	-1.96
8.50	29.85	-3.15
9.00	28.35	-4.65
9.50	26.50	-6.50
10.00	23.53	-9.47
10.50	19.00	-14.00
11.00	13.86	-19.14
11.50	9.84	-23.16
12.00	6.98	-26.02
12.50	4.30	-28.70
13.00	2.02	-30.98
13.50	0.76	-32.24
14.00	0.25	-32.75

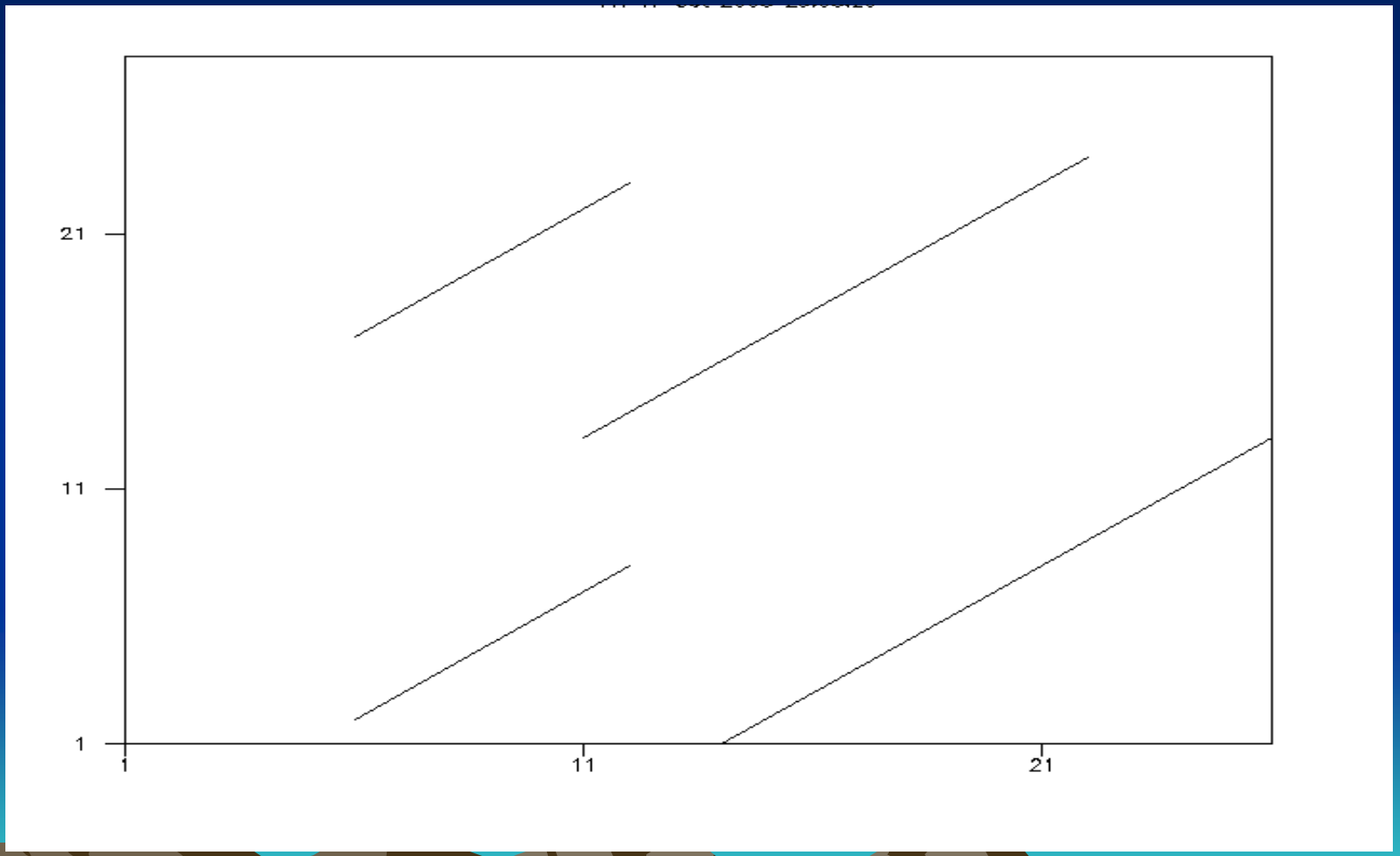
4. 对核酸的核苷酸序列或蛋白质的氨基酸序列的相似性分析

dottup 是两条序列精确匹配的作图方法，这个程序的执行方式是在给定序列长度(word size) 下逐一比对，即在水平轴和竖直轴上的两个序列，将每个序列的每个残基同另一个序列的全部残基比较，有相同的残基就在图表中用“点”作为标记，否则就空白。当两个序列有相同的区域出现的时候，很多点相连接就形成斜线，显示出序列比对。

>F1 AREALFRIENDISAFRIENDINNEED

>F2 AFRIENDINNEEDISAFRIENDINDEED

其中X轴为F1序列，Y轴为F2序列。（Word size 5）



从这张分析图中我们可以知道：

>F1 AREALFRIENDISAFRIENDINNEED

>F2 AFRIENDINNEEDISAFRIENDINDEED

和

>F1 AREALFRIENDISAFRIENDINNEED

>F2 AFRIENDINNEEDISAFRIENDINDEED

和

>F1 AREALFRIENDISAFRIENDINNEED

>F2 AFRIENDINNEEDISAFRIENDINDEED

和

>F1 AREALFRIENDISAFRIENDINNEED

>F2 AFRIENDINNEEDISAFRIENDINDEED

- *Water* DNA或蛋白质的局部比对软件，在比对后给出两序列的相同性，相似性，gap以及分数。

- 我们以这两条氨基酸序列为例进行操作：
- MVLSGEDKSNIKAAWGKIGGHGAEYGAE
  ALERMFASFPTTKTYFPHFDVSHGSAQV
  KGHGKKVADALASAAGHLDDLPGALSAL
  SDLHAHKLRVDPVNFKLLSHCLLVTLASH
  HPADFTPAVHASLDKFLASVSTVLTSKYR
- MVLSPADKTNVKAAWGKVGAHAGEYGA
  EALERMFLSFPTTKTYFPHFDLSHGSAQV
  KGHGKKVADALTNAVAHVDDMPNALSAL
  SDLHAHKLRVDPVNFKLLSHCLLVTLAHL
  PAEFTPAVHASLDKFLASVSTVLTSKYR


1 MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS 50
1 MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVS 50
51 HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFK 100
51 HGSAQVKGHGKKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFK 100
101 LLSHCLLVTLA-HLPAEFTPAVHASLDKFLASVSTVLTSKYR 141
101 LLSHCLLVTLASHHPADFTPAVHASLDKFLASVSTVLTSKYR 142

5.对蛋白质一级结构的分析

- *Pepstats* 蛋白质的统计，可以在该程序中得到一条蛋白质的各个残基的统计量。

MVLSPADKTNVKAAWGKVGAHAGEYGAE
ALERMFLSFPTTKTYFPHFDLSHGSAQVK
GHGKKVADALTNAVAHVDDMPNALSALSD
LHAHKLRVDPVNFKLLSHCLLVTLAAHLPA
EFTPAVHASLDKFLASVSTVLTSKYR

PEPSTATS of from 1 to 142

Molecular weight = 15257.50 Residues = 142
Average Residue Weight = 107.447 Charge = 7.0
Isoelectric Point = 9.0879
A280 Molar Extinction Coefficient = 9530
A280 Extinction Coefficient 1mg/ml = 0.62
Improbability of expression in inclusion bodies = 0.724

Residue	Number	Mole%	DayhoffStat
A = Ala	21	14.789	1.720
B = Asx	0	0.000	0.000
C = Cys	1	0.704	0.243
D = Asp	8	5.634	1.024
E = Glu	4	2.817	0.469
F = Phe	7	4.930	1.369
G = Gly	7	4.930	0.587
H = His	10	7.042	3.521
I = Ile	0	0.000	0.000
J = ---	0	0.000	0.000
K = Lys	11	7.746	1.174
L = Leu	18	12.676	1.713
M = Met	3	2.113	1.243
N = Asn	4	2.817	0.655
O = ---	0	0.000	0.000
P = Pro	7	4.930	0.948
Q = Gln	1	0.704	0.181
R = Arg	3	2.113	0.431
S = Ser	11	7.746	1.107
T = Thr	9	6.338	1.039
U = ---	0	0.000	0.000
V = Val	13	9.155	1.387
W = Trp	1	0.704	0.542
X = Xaa	0	0.000	0.000
Y = Tyr	3	2.113	0.621
Z = Glx	0	0.000	0.000

Property	Residues	Number	Mole%
Tiny	(A+C+G+S+T)	49	34.507
Small	(A+B+C+D+G+N+P+S+T+V)	81	57.042
Aliphatic	(A+I+L+V)	52	36.620
Aromatic	(F+H+W+Y)	21	14.789
Non-polar	(A+C+F+G+I+L+M+P+V+W+Y)	81	57.042
Polar	(D+E+H+K+N+Q+R+S+T+Z)	61	42.958
Charged	(B+D+E+H+K+R+Z)	36	25.352
Basic	(H+K+R)	24	16.901
Acidic	(B+D+E+Z)	12	8.451

6.对蛋白质二级结构的分析

- 我们以这条氨基酸序列为例进行操作：
- MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR

- *garnier* 预测蛋白的二级结构。输入蛋白序列从而进行预测。以EBI数据库中的amic_pseae序列为例，程序运行得出了helix、sheet、turns、coil等序列位点。

```
          .    10     .    20     .    30     .    40     .    50
      MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS
helix     HHHHHHHHHHHHH       HHHHHHHHHHHH
sheet EEEE                                      EEE E
turns                                        TTTTT   T T
 coil                   CCCCCC            CC           CC
          .    60     .    70     .    80     .    90     .   100
      HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFK
helix    HHHH       HHHHHHHHH  HHHHHHHHHHHHHHHHH    HHHHH
sheet              EE          EE                  E EE
turns         TT                                    T
 coil CCC       CC
          .   110     .   120     .   130     .   140
      LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
helix HHHHHHHHHHHH HHHHHHHHHHHHHHHHHHH
sheet                                    EEEE
turns                                       TTTT
 coil               C
```

- *sigcleave* 在真核生物中,信号肽介导跨膜蛋白定位到质膜上;而在原核生物中信号肽介导跨膜蛋白定位贯穿到内外质膜.这个软件可以预测信号肽和成熟蛋白之间的切点.在原核和真核生物中预测的准确率在75-80%之间。

(1) Score 5.015 length 13 at residues 99->111

```
 Sequence: FKLLSHCLLVTLA
           |           |
          99          111
 mature_peptide: AHLPAEFTPAVHASLDKFLASVSTVLTSKYR
```

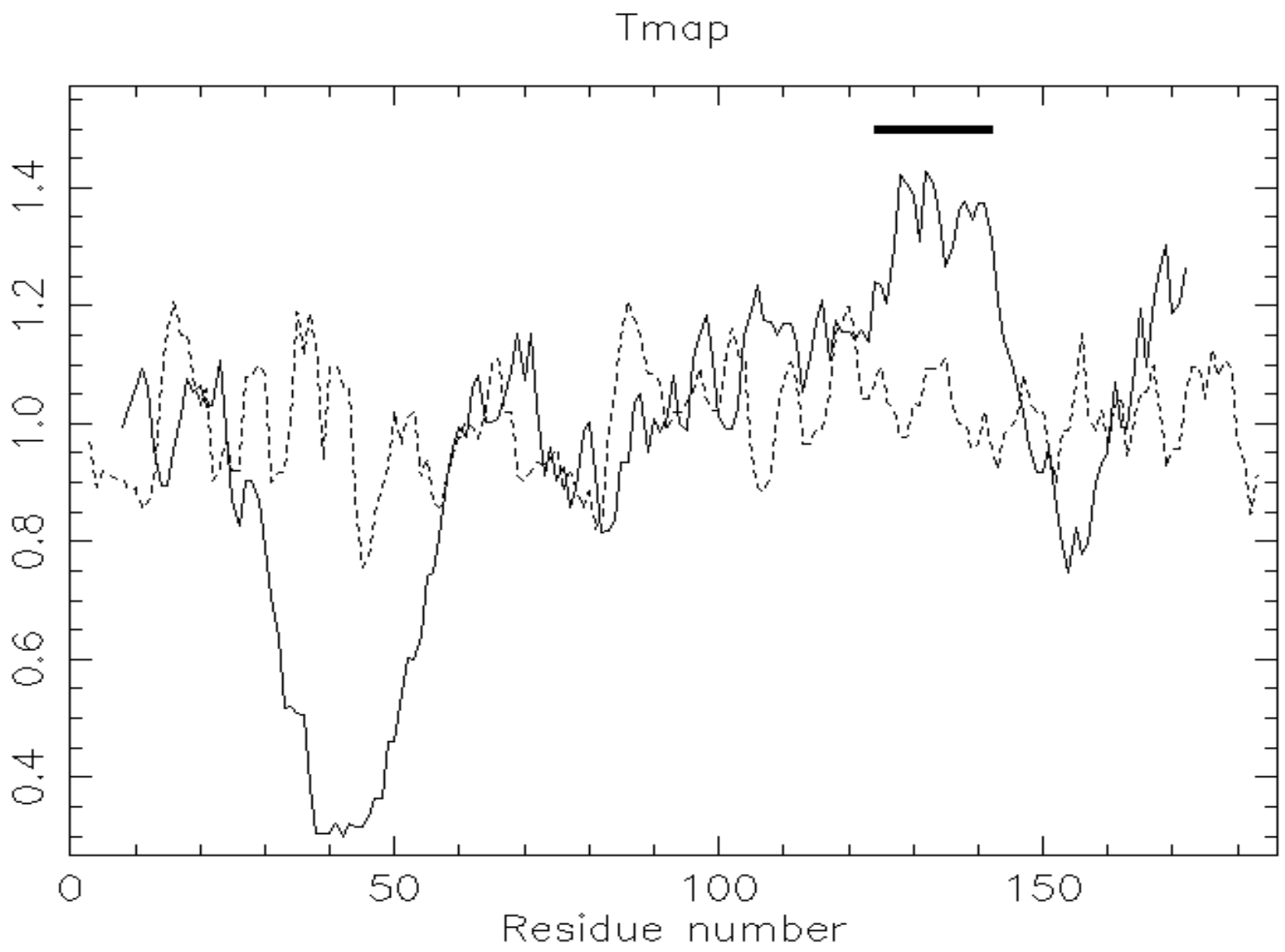
(2) Score 3.539 length 13 at residues 104->116

```
 Sequence: HCLLVTLAAHLPA
           |           |
         104          116
 mature_peptide: EFTPAVHASLDKFLASVSTVLTSKYR
```

7.对蛋白质三级结构的分析

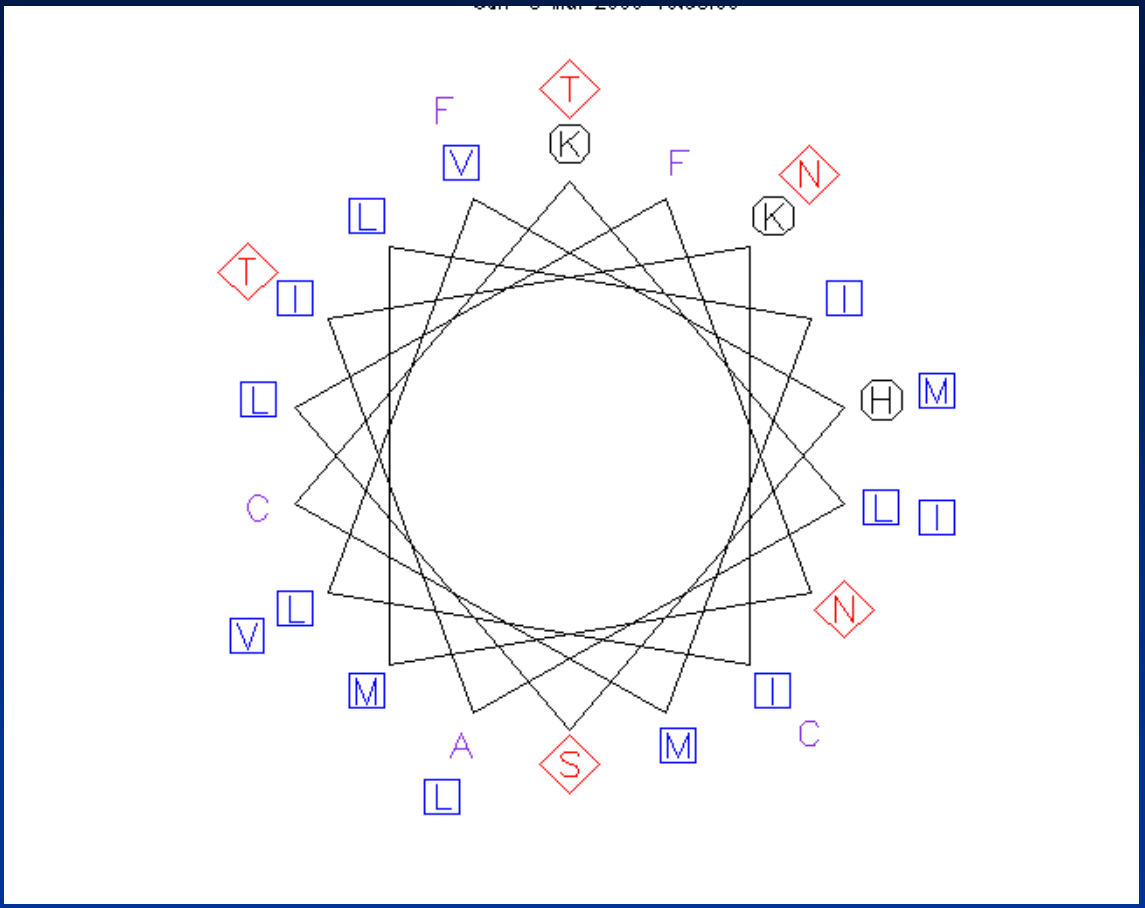
- 以下我们以此氨基酸序列为例进行一系列的操作：
- MAQSVLVPPGPDSFRFFTRESLAAIEQRI AEEKAKRPKQERKDEDDENGPKPNSDL EAGKSLPFIYGDIPPEMVSVPLEDLDPYY INKKTFIVLNKGKAISRFSATPALYILTPFN PIRKLAIKILVHSLFNMLIMCTILTNCVFMT MSNPPDWTKNVEYTFTGIYTFESLIKILA RGFCLEDFTF

- tmap用来查看蛋白质是否具有跨膜区。输入其蛋白质的氨基酸序列。
- 一般跨膜区是一段高度疏水的氨基酸序列，其二级结构为螺旋。但是由此预测出来其不一定为跨膜区，因为其是否跨膜还与其在细胞中的位置等许多因素有关。



120—146 KLAIKILVHSLFNMLIMCTILTNCVFM

- pepwheel 程序可以对一段链作螺旋状轮图，从而识别疏水和亲水的区域。它所显示的是检测序列螺旋的横切面，从而更加直观的反应序列的各种性质。我们输入可以把我们从上图中找到一个疏水区域。但只可以对一个疏水的序列进行，不可输入整个的氨基酸序列，太多了我们反而无法分析，这是我们大家应该注意的地方。通过这个分析我们也可以验证Tmap预测结果的正确性。



轮状图的周围是螺旋上的氨基酸，其中黑色带圆圈的是极性带电荷的氨基酸，亲水性比较强；红色带方框的是极性不带电荷的氨基酸，也是亲水性的；紫色字母不带框的是非极性氨基酸，具有疏水性；蓝色带方框的是非极性氨基酸，疏水性比紫色的强。

8.对蛋白质进行酶学分析

以下我们以此基因为例进行一系列的操作：

- ATGGCACAGTCAGTGCTGGTACCGCCAGGACCTGACAGCTTCC
  GCTTCTTTACCAGGGAATCCCTTGCTGCTATTGAACAACGCATT
  GCAGAAGAGAAAGCTAAGAGACCCAAACAGGAACGCAAGGATG
  AGGATGATGAAAATGGCCCAAAGCCAAACAGTGACTTGGAAGC
  AGGAAAATCTCTTCCATTTATTTATGGAGACATTCCTCCAGAGAT
  GGTGTCAGTGCCCCTGGAGGATCTGGACCCCTACTATATCAAT
  AAGAAAACGTTTATAGTATTGAATAAAGGGAAAGCAATCTCTCG
  ATTCAGTGCCACCCCTGCCCTTTACATTTTAACTCCCTTCAACC
  CTATTAGAAAATTAGCTATTAAGATTTTGGTACATTCTTTATTCAA
  TATGCTCATTATGTGCACGATTCTTACCAACTGTGTATTTATGAC
  CATGAGTAACCCTCCAGACTGGACAAAGAATGTGGAGTATACCT
  TTACAGGAATTTATACTTTTGAATCACTTATTAAAATACTTGCAAG
  GGGCTTTTGTTTAGAAGATTTCACATTTTT

- *restrict*利用的是限制性酶的数据库来进行搜索分析，输入的方式为核苷酸序列，可以把它的限制性内切酶的位点与酶的名称都能罗列出来。运行的时候，我们可以进行一些选择，如序列是否为环形的、酶切的大小、末端等。

Start	End	Enzyme_name	Restriction_site	5prime	3prime	5primerev	3primerev
15	25	MwoI	GCNNNNNNGC	21	18	.	.
26	34	AlwNI	CAGNNNCTG	31	28	.	.
27	33	DraII	RGGNCCY	28	31	.	.
27	33	PpuMI	RGGWCCY	28	31	.	.
224	214	Hin4I	GAYNNNNNVTC	237	232	205	200
313	323	MwoI	GCNNNNNNGC	319	316	.	.
444	454	BplI	GAGNNNNCTC	435	430	467	462
454	444	BplI	GAGNNNNCTC	467	462	435	430
464	475	CspCI	CAANNNNGTGG	452	450	487	485

- *redata*利用的是限制性酶的数据库来进行搜索分析，输入的方式为某一种酶的名字，然后把它的基本性质与相关的文献都列出来，可以使我们对这个酶有一个更加全面的认识。

例如我们输入BamHI，运行后得到：

BamHI

Recognition site is GGATCC leaving sticky ends
  Cut positions 5':1 3':5
Organism: Bacillus amyloliquefaciens H
Methylated: 5(4)
Source: ATCC 49763

Suppliers:
GE Healthcare
Invitrogen Corporation
Fermentas International Inc.
Qbiogene
American Allied Biochemical, Inc.
SibEnzyme Ltd.
Nippon Gene Co., Ltd.
Takara Bio Inc.
Roche Applied Science
New England Biolabs
Toyobo Biochemicals
Molecular Biology Resources - CHIMERx
Promega Corporation
Sigma Chemical Corporation
Bangalore Genei
Vivantis Technologies
MP Biomedicals
EURx Ltd.
CinnaGen Inc.

References:
Brooks, J.E., Nathan, P.D., Landry, D., Sznyter, L.A., Waite-Rees, P., Ives, C.L., Moran, L.S., Slatko, B.E., Benner, J.S., (1991) Nucleic Acids Res., vol. 19, pp. 841-850.
Endo, M., Majima, T., Japanese Patent Office, 2003.
Hattman, S., Keister, T., Gottehrer, A., (1978) J. Mol. Biol., vol. 124, pp. 701-711.
Majima, T., Endo, M., Japanese Patent Office, 2006.
Roberts, R.J., Wilson, G.A., Young, F.E., (1977) Nature, vol. 265, pp. 82-84.
Usami, S., Kurimura, H., Kino, K., Kamigaki, K., Kirimura, K., Japanese Patent Office, 2003.
Wilson, G.A., Young, F.E., (1975) J. Mol. Biol., vol. 97, pp. 123-125.

9.多个序列之间进化关系上的分析

- 我们以下面为例进行操作：
- >Human - HBA_HUMAN Hemoglobin alpha - Homo sapiens (Human).
- MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTN AVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKY R
- >Mouse - HBA_MOUSE Hemoglobin alpha - Mus musculus (Mouse).
- MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHGKKVADALASA AGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPADFTPAVHASLDKFLASVSTVLTSKYR
- >Dolphin - HBA_TURTR Hemoglobin alpha - Tursiops truncatus (Atlantic bottle-nosed dolphin).
- MVLSPADKTNVKGTWSKIGNHSAEYGAEALERMFINFPSTKTYFSHFDLGHGSAQIKGHGKKVADALTKAV GHIDNLPDALSELSDLHAHKLRVDPVNFKLLSHCLLVTLALHLPADFTPSVHASLDKFLASVSTVLTSKYR
- >Chicken - HBA_CHICK Hemoglobin alpha-A - Gallus gallus (Chicken).
- MVLSAADKNNVKGIFTKIAGHAEEYGAETLERMFTTYPPTKTYFPHFDLSHGSAQIKGHGKKVVAALIEAAN HIDDIAGTLSKLSDLHAHKLRVDPVNFKLLGQCFLVVVAIHHPAALTPEVHASLDKFLCAVGTVLTAKYR
- >Snake - HBA_DRYCE Hemoglobin alpha-A - Drymarchon corais erebennus (Texas indigo snake).
- MVLTEEDKSRVRAAWGPVSKNAELYGAETLTRLFTAYPATKTYFHHFDLSPGSSNLKTHGKKVIDAITEAVN NLDDVAGALSKLSDLHAQKLRVDPVNFKLLGHCLEVTIAAHNGGPLKPEVILSLDKFLCLVAKTLVSRYR
- >Frog - HBA1_XENLA Hemoglobin subunit alpha-1 - Xenopus laevis (African clawed frog).
- MLLSADDKKHIKAIMPAIAAHGDKFGGEALYRMFIVNPKTKTYFPSFDFHHNSKQISAHGKKVVDALNEASN HLDNIAGSMSKLSDLHAYDLRVDPGNFPLLAHNILVVVAMNFPKQFDPATHKALDKFLATVSTVLTSKYR

*Distamt*该软件可以计算出多重比对中的每对序列之间的进化距离。经过多重比对的序列结果才可以实用该软件，多重比对的质量对本软件计算的准确性有很大的影响。可以有jukes、kimura、tamura、tajima-nei、jin-nei gamma五种方法可以用。

Distance Matrix

Using the Jukes-Cantor correction method
Gap weighting is 0.100000

1	2	3	4	5	6	
0.00	15.24	17.76	35.44	49.76	58.47	Human 1
	0.00	20.34	34.42	53.39	53.39	Mouse 2
		0.00	35.44	59.78	58.47	Dolphin 3
			0.00	50.95	53.39	Chicken 4
				0.00	82.01	Snake 5
					0.00	Frog 6

10.其他软件功能

- *seealso*输入出一个软件名字，其可以找出与之功能相似的软件。可以扩大你所知软件的种类，对你要的结果进行反复验证或是查找到更多的信息。例如你输入water，则

SEE ALSO

matcher	Finds the best local alignments between two sequences
seqmatchall	All-against-all comparison of a set of sequences
supermatcher	Match large sequences against one or more other sequences
wordfinder	Match large sequences against one or more other sequences
wordmatch	Finds all exact matches of a given size between 2 sequences

*backtranambig*将蛋白序列按照通用密码子表翻译成由简并密码子组成的核酸序列。其也有很多的标准可以选，如大肠杆菌、酵母菌等。

我们以此为例进行说明:

MVLSPADKTNVKAAWGKVGAHAGEYGAE
ALERMFLSFPTTKTYFPHFDLSHGSAQVK
GHGKKVADALTNAVAHVDDMPNALSALSD
LHAHKLRVDPVNFKLLSHCLLVTLAAHLPA
EFTPAVHASLDKFLASVSTVLTSKYR

>EMBOSS_001

ATGGTNYTNWSNCCNGCNGAYAARACNAAYGTNAARGCNGCNTGGGGNAARGTNGGNGCN
CAYGCNGGNGARTAYGGNGCNGARGCNYTNGARMGNATGTTYYTNWSNTTYCCNACNACN
AARACNTAYTTYCCNCAYTTYGAYYTNWSNCAYGGNWSNGCNCARGTNAARGGNCAYGGN
AARAARGTNGCNGAYGCNYTNACNAAYGCNGTNGCNCAYGTNGAYGAYATGCCNAAYGCN
YTNWSNGCNYTNWSNGAYYTNCAYGCNCAYAARYTNMGNGTNGAYCCNGTNAAYTTYAAR
YTNYTNWSNCAYTGYYTNYTNGTNACNYTNGCNGCNCAYYTNCCNGCNGARTTYACNCCN
GCNGTNCAYGCNWSNYTNGAYAARTTYYTNGCNWSNGTNWSNACNGTNYTNACNWSNAAR
TAYMGN

tfm 这个工具是帮助找到EMBOSS软件用法的软件。在搜索栏中输入你要了解的工具，便可轻松找到相关的信息、说明、应用举例等，让你可以学习到更多的东西。

*Biosed*是一个核酸、蛋白序列编辑软件，可以进行序列删除、替换，产生一个新的序列。例如可以将mRNA序列中的U用T替换，获得cDNA序列。但是有一个缺点，就是不能只替换一个，一旦替换所有序列中的X都由Y所替换。

现在我们以此核苷酸序列为例进行说明：

ATGGCACAGTCAGTGCTGGTACCGCCAGGACCTGACA
GCTTCCGCTTCTTTACCAGGGAATCCCTTGCTGCTATT
GAACAACGCATTGCAGAAGAGAAAGCTAAGAGACCCA
AACAGGAACGCAAGGATGAGGATGATGAAAATGGCCC
AAAGCCAAACAGTGACTTGGAAGCAGGAAAATCTCTTC
CATTTATTTATGGAGACATTCCTCCAGAGATGGTGTCA
GTGCCCCTGGAGGATCTGGACCCCTACTATATCAATAA
GAAAACGTTTATAGTATTGAATAAAGGGAAAGCAATCT
CTCGATTCAGTGCCACCCCTGCCCTTTACATTTTAACT
CCCTTCAACCCTATTAGAAAATTAGCTATTAAGATTTTG
GTACATTCTTTATTCAATATGC

如果我们用U全部替换其中的T，则

>EMBOSS_001
AUGGCACAGUCAGUGCUGGUACCGCCAGGACCUGACAGCUUCCGCUUCUUUACCAGGGAA
UCCCUUGCUGCUAUUGAACAACGCAUUGCAGAAGAGAAAGCUAAGAGACCCAAACAGGAA
CGCAAGGAUGAGGAUGAUGAAAAUGGCCCAAAGCCAAACAGUGACUUGGAAGCAGGAAAA
UCUCUUCCAUUUAUUUAUGGAGACAUUCCUCCAGAGAUGGUGUCAGUGCCCCUGGAGGAU
CUGGACCCCUACUAUAUCAAUAAGAAAACGUUUAUAGUAUUGAAUAAAGGGAAAGCAAUC
UCUCGAUUCAGUGCCACCCCUGCCCUUUACAUUUUAACUCCCUUCAACCCUAUUAGAAAA
UUAGCUAUUAAGAUUUUGGUACAUUCUUUAUUCAAUAUGCUCAUUAUGUGCACGAUUCUU
ACCAACUGUGUAUUUAUGACCAUGAGUAACCCUCCAGACUGGACAAAGAAUGUGGAGUAU
ACCUUUACAGGAAUUUAUACUUUUGAAUCACUUAUUAAAAUACUUGCAAGGGGCUUUUGU
UUAGAAGAUUUCACAUUUUU

tfscan TRANSFAC 数据库是一个商业数据库。从酵母到人类所有的真核生物顺式作用原件及转录因子都可以在这个数据库中找到。通过输入一段核苷酸序列并选择其所属的物种，便可搜寻与TRANSFAC数据库中相匹配的数据。

```
TFSCAN of  from 1 to 400

Y$HIS4_13          R00656   7     12    CAGTCA
                   T00321; GCN4;Quality: 2; Species: yeast, Saccharomyces cerevisiae.
Y$GAL1_08          R00493   235   239   GAGGA
Y$GAL1_08          R00493   130   134   GAGGA
                   T00303; GAL80;Quality: 4; Species: yeast, Saccharomyces cerevisiae.
Y$GAL1_07          R00492   235   239   GAGGA
Y$GAL1_07          R00492   130   134   GAGGA
```

感謝