Rosalind problems

Студент: Антон Мордберг

Научный руководитель: Николай Вяххи

Цель моей работы

- Написать несколько задач для системы Rosalind, решая которые можно познакомиться с некоторыми из форматов данных, используемых биологами, и инструментами для их обработки.

Задача GBK

- Часто последовательности хранятся в формате gb (gbk) – GenBank file format. Он отличается от формата fasta в основном своим заголовком, который содержит некоторую ключевую информацию о последовательности (что это за организм, тип молекулы, авторы, ссылки в публикациях и т.д.)

Задача GBK

- В задача предлагается конвертировать файл из gbk формата в форматы fasta, nexus и phylip. Форматы nexus и phylip – форматы для хранения выравненных последовательностей.

```
>fish
ACATAGAGGGTACCTCTAAG
>frog
ACATAGAGGGTACCTCTAAG
>snake
ACATAGAGGGTACCTCTAAG
>mouse
ACATAGAGGGTACCTCTAAG
```

```
#NEXUS
begin data;
       dimensions ntax=4
nchar=20;
       format datatype=dna
missing=? gap=-;
matrix
fish  ACATAGAGGGTACCTCTAAG
frog  ACATAGAGGGTACCTCTAAG
snake ACATAGAGGGTACCTCTAAG
mouse ACATAGAGGGTACCTCTAAG
;
end;
```

```
 4 20
fish   ACATAGAGGG TACCTCTAAG
frog   ACATAGAGGG TACCTCTAAG
snake  ACATAGAGGG TACCTCTAAG
mouse  ACATAGAGGG TACCTCTAAG
```

Задача VCF

- VCF – Variant Call Format. В этом формате хранят информацию о мутациях в генотипе. Некоторые результаты проекта «1000 геномов» доступны в данном формате. Цель этого проекта создать наиболее полный каталог вариаций человеческого генотипа.
- В задаче предлагается прочитать файл в данном формате и вывести для каждой записи наиболее вероятную вариацию.

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T
20     1234567 microsat1 GTCT   G,GTACT 50   PASS   NS=3;DP=9;AA=G
```

Задача BAM

- В задаче предлагается перевести sam файл в bam файл, который является его сжатой, индексированной, бинарной версией.

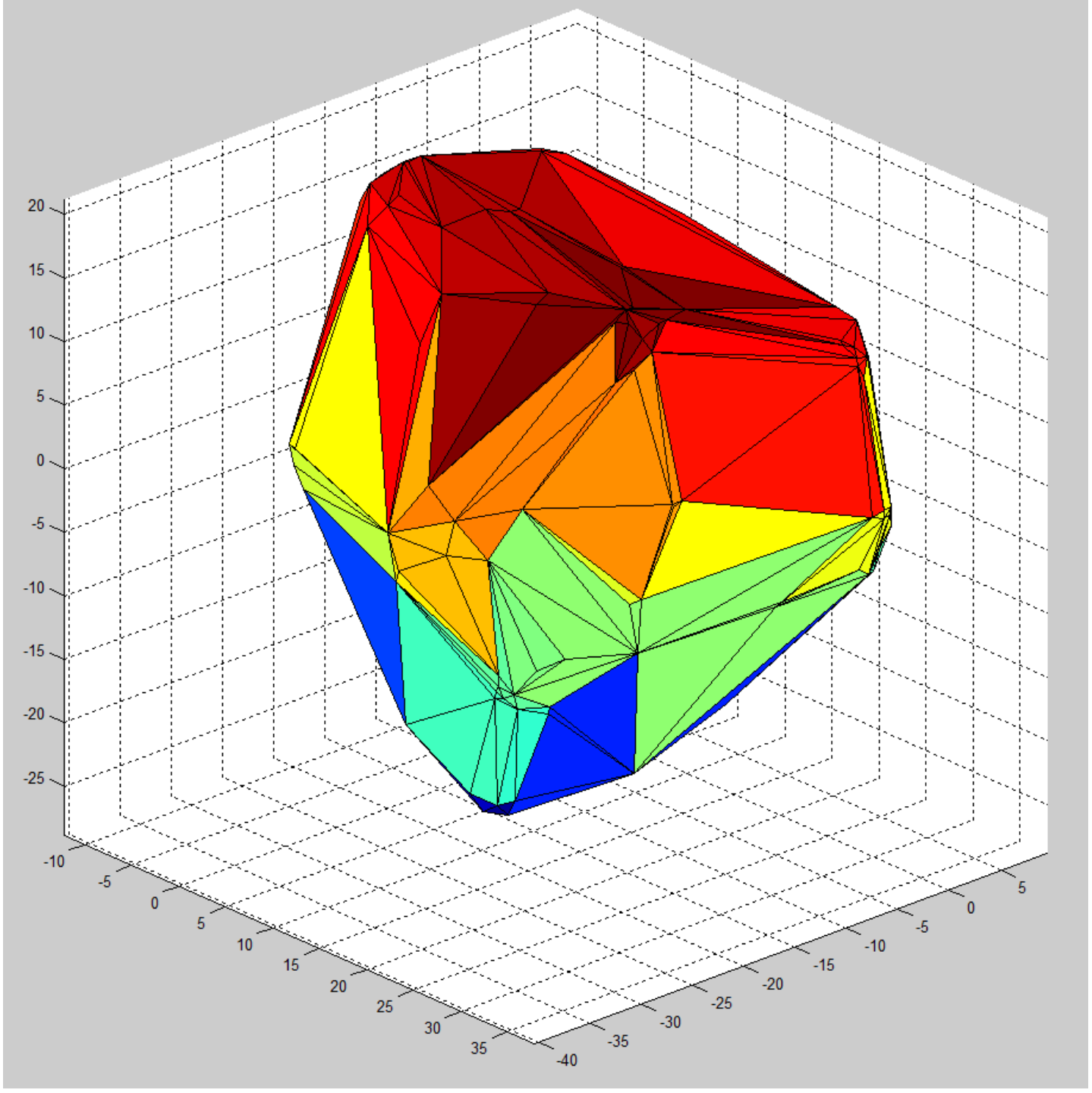
```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                           ttagctTAGGC
-r001/2                                      CAGCGCCAT
```

The corresponding SAM format is:

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA    *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC     *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

Задача PDB

- В данной задаче я постарался кроме чисто технической части сделать ещё небольшую алгоритмическую часть.
- Формат pdb (Protein Data Bank) хранит информацию о трёхмерной структуре белка – в частности в нём хранятся координаты всех атомов.
- Задача состоит в том, чтобы построить выпуклую оболочку по точкам, соответствующим атомам белка.


20
15
10
5
0
-5
-10
-15
-20
-25
-10
-5
0
5
10
15
20
25
30
35
-40
-35
-30
-25
-20
-15
-10
-5
0
5

Спасибо за внимание!