08.04.22

АДБМ

Практическая биоинформатика

Домашнее задание 5

Pfam

Прокопчук С. Р

1.1 Бактерия

Для анализа была выбрана бактерия *Bacillus subtilis*.

1.2 Белки бактерии

Файл со всеми белками приложен в архиве. Название файла: GCF_000009045.1_ASM904v1_protein.faa

1.3 Выравнивания-затравки

Все выравнивания приложены в архиве.

1.4 HMMER

Был установлен на Linux. Версия 3.3.2

1.5 Анализ с HMMER

Ниже приведены скрины работы. Полный набор команд и вывода приложены в файле hw5 pfam terminal в архиве.

Пример команды для построения модели:

```
~/hmmer-3.3.2/src/hmmbuild ~/profiles/HD_GYP.hmm
          ~/domains_Pfam/HD_GYP.fasta
```

Пример для поиска совпадений:

```
~/hmmer-3.3.2/src/hmmsearch ~/profiles/HD_GYP.hmm
 ~/my_bac/GCF_000009045.1_ASM904v1_protein.faa >
              ~/rez/rez_HD_GYP.txt
```

```
# Copyright (C) 2020 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# input alignment file:             /home/sofia/domains_Pfam/STYK.fasta
# output HMM file:                  /home/sofia/profiles/STYK.hmm
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

# idx name                  nseq  alen  mlen eff_nseq re/pos description
#---- -------------------- ----- ----- ----- -------- ------ -----------
1     STYK                   38   419   263     2.63  0.590

# CPU time: 0.09u 0.00s 00:00:00.09 Elapsed: 00:00:00.09
(base) sofia@asrock:~$ ~/hmmer-3.3.2/src/hmmsearch ~/profiles/hweHK.hmm ~/my_ba
c/GCF_000009045.1_ASM904v1_protein.fasta > rez_hweHK.txt

Error: Failed to open sequence file /home/sofia/my_bac/GCF_000009045.1_ASM904v1
_protein.fasta for reading

(base) sofia@asrock:~$ ~/hmmer-3.3.2/src/hmmsearch ~/profiles/hweHK.hmm ~/my_ba
c/GCF_000009045.1_ASM904v1_protein.faa > rez_hweHK.txt
(base) sofia@asrock:~$ cd /home/sofia/domains_Pfam
(base) sofia@asrock:~/domains_Pfam$ ls
AC1.fasta  EAL.fasta        HD_GYP.fasta  hisKa.fasta  STYK.fasta
AC2.fasta  GGDEF.fasta      hisKa2.fasta  hweHK.fasta
AC3.fasta  hatpase_c.fasta  hisKa3.fasta  MCP.fasta
(base) sofia@asrock:~/domains_Pfam$ cd /home/sofia
(base) sofia@asrock:~$ ~/hmmer-3.3.2/src/hmmsearch ~/profiles/MCP.hmm ~/my_bac/
GCF_000009045.1_ASM904v1_protein.faa > ~/rez/rez_MCP.txt
(base) sofia@asrock:~$
```

```
(base) sofia@asrock:~/domains_Pfam$ cd /home/sofia
(base) sofia@asrock:~$ ~/hmmer-3.3.2/src/hmmsearch ~/profiles/MCP.hmm ~/my_bac/
GCF_000009045.1_ASM904v1_protein.faa > ~/rez/rez_MCP.txt
(base) sofia@asrock:~$ ~/hmmer-3.3.2/src/hmmsearch ~/profiles/STYK.hmm ~/my_bac
/GCF_000009045.1_ASM904v1_protein.faa > ~/rez/rez_STYK.txt
(base) sofia@asrock:~$ ~/hmmer-3.3.2/src/hmmsearch ~/profiles/EAL.hmm ~/my_bac/
GCF_000009045.1_ASM904v1_protein.faa > ~/rez/rez_EAL.txt
(base) sofia@asrock:~$ ~/hmmer-3.3.2/src/hmmsearch ~/profiles/GGDEF.hmm ~/my_ba
c/GCF_000009045.1_ASM904v1_protein.faa > ~/rez/rez_GGDEF.txt
(base) sofia@asrock:~$ ~/hmmer-3.3.2/src/hmmsearch ~/profiles/hatpase_c.hmm ~/m
y_bac/GCF_000009045.1_ASM904v1_protein.faa > ~/rez/rez_hatpase_c.txt
(base) sofia@asrock:~$ ~/hmmer-3.3.2/src/hmmsearch ~/profiles/HD_GYP.hmm ~/my_b
ac/GCF_000009045.1_ASM904v1_protein.faa > ~/rez/rez_HD_GYP.txt
(base) sofia@asrock:~$ ~/hmmer-3.3.2/src/hmmsearch ~/profiles/hisKa.hmm ~/my_ba
c/GCF_000009045.1_ASM904v1_protein.faa > ~/rez/rez_hisKa.txt
(base) sofia@asrock:~$ ~/hmmer-3.3.2/src/hmmsearch ~/profiles/hisKa2.hmm ~/my_b
ac/GCF_000009045.1_ASM904v1_protein.faa > ~/rez/rez_hisKa2.txt
(base) sofia@asrock:~$ ~/hmmer-3.3.2/src/hmmsearch ~/profiles/hisKa3.hmm ~/my_b
ac/GCF_000009045.1_ASM904v1_protein.faa > ~/rez/rez_hisKa3.txt
(base) sofia@asrock:~$ ~/hmmer-3.3.2/src/hmmsearch ~/profiles/AC1.hmm ~/my_bac/
GCF_000009045.1_ASM904v1_protein.faa > ~/rez/rez_AC1.txt
(base) sofia@asrock:~$ ~/hmmer-3.3.2/src/hmmsearch ~/profiles/AC2.hmm ~/my_bac/
GCF_000009045.1_ASM904v1_protein.faa > ~/rez/rez_AC2.txt
(base) sofia@asrock:~$ ~/hmmer-3.3.2/src/hmmsearch ~/profiles/AC3.hmm ~/my_bac/
GCF_000009045.1_ASM904v1_protein.faa > ~/rez/rez_AC3.txt
(base) sofia@asrock:~$
```

1.6 Анализ находок

В таблице рассмотрено количество находок выше inclusion threshold.

Таблица 1

Тип ферментов	Домен 1	Домен 2	Количество находок
Histidine kinases	phosphoacceptor domain: HisKA [Pfam:PF00512]	ATPase domain: HATPase_c [Pfam:PF02518]	43(ATPase)/ 17(hiska)
	HisKA_2 [Pfam:PF07568]	-	нет
	HisKA_3 [Pfam:PF07730]	-	9
	HWE_HK [Pfam:PF07536]	-	нет
Methyl-accepting chemotaxis proteins	Methyl-accepting protein (MCP) domain: [Pfam:PF00015]	-	11
Ser/Thr/Tyr kinases	Ser/Thr/Tyr kinase (STYK) domain: [Pfam:PF00069]	-	6
Diguanylate cyclases	GGDEF domains: [Pfam:PF00990]	-	4
Adenylate cyclases	AC1 domains: [Pfam:PF01295]	-	нет
	AC2 domains: [Pfam:PF01928]	-	1
	or AC3 domains: [Pfam:PF00211]	-	1
Predicted c-di-GMP phosphodiesterases	EAL domains :[Pfam:PF00563]	-	4
	HD-GYP domain: [Pfam:PF01966]	-	9

Всего белков в бактерии: 4237

Далее проводился анализ на повторную встречаемость последовательностей. Для гистидин киназы сначала проверялись находки на уникальность между HisKa и HisKa3. Для этого названия последовательностьей ‘sequense’ переносились в таблицу в Excel, а затем анализировались в R. Повторяющихся доменов обнаружено не было. Далее искались пересечения между ними и АТФазным доменом, чтобы можно было найти удачные находки, содержащие оба домена. Ниже приведены имена белков из последовательностей в файле бактерии.

"NP_390788.1" "NP_390192.1" "NP_391920.1" "NP_389209.1" "NP_389249.1" "NP_388138.2" "NP_389282.1" "NP_389332.1"

"NP_391023.2" "NP_391201.2" "NP_391182.2" "NP_388083.1" "NP_390519.1" "NP_391351.1" "NP_391844.1" "NP_389800.1"

"NP_391287.2" "NP_388813.1" "NP_391769.1" "NP_391189.1" "NP_388710.1" "NP_391430.1" "NP_388422.1"

Всего 23.

Далее данные последовательности и по остальным доменам были объединены в 1 дата фрейм и осуществлялся поиск уникальных последовательностей. Из 36 названий 35 оказалось уникальными.

Следовательно, всего оказалось 35 последовательностей.

(Файл all.csv содержит находки по всем доменам кроме гистидина киназы, так как число хороших находок для нее определялось другим способом).

Анализ проводился при помощи R. Файл будет приложен в архиве, а также все используемые таблицы с названиями последовательностей. Кодировка UTF-8.Код для анализа:

```
#######

#hisKa unigue

hiska=read.csv(file='hiska.csv')

table(hiska) #все последовательности, которые были в обнаружены в обоих видах HisKa

length(unique(hiska$hiska))#  кол-во  уникальных  совпало  с изначальнм кол-вом

#######

#hatpase_c

hatpase_c=read.csv(file='hatpase_c.csv')

####intersect hatpase_c and hisKa

length(intersect(hiska$hiska, hatpase_c$hatpase_c))

hist_kin <- (intersect(hiska$hiska, hatpase_c$hatpase_c))

hist_kin <- as.data.frame(hist_kin)

###all uniqe

all <- read.csv(file='all.csv')
```

```
length(all$all)

length(unique(all$all))

all_uniq <- as.data.frame(unique(all$all))

intersect(all$all,hist_kin$hist_kin)    #на всякий случай проверили нет ли совпадений с гистидиновым доменом и остальными последовательностями
```

1.7 IQ

Рассчитаем IQ бактерии *Bacillus subtilis*.

Всего белков	4237
Находок гистидин киназы	23
Остальных находок по оставшимся доменам	35

35+23=58

QI: 58/4237=0,0136889308