Задание практикума: Поиск ДНК/РНК-фрагментов

1 Общее описание

Требуется написать одну из программ с контролем ошибок. В аргументах программы указывается 2 имени файла с последовательностями. В качестве первого имени файла укаказывается файл с «белковыми» (аминокислотными) последовательностями в FASTA формате[1], в качестве второго параметра указывается файл с последовтельностями РНК/ДНК. Далее могут следовать параметры специфичные для конкретного вида программы.

Нуклеотидная последовательность представлена символами алфавита {A,a, U,u, G,g, C,c, T,t}. Белковая последовательность представлена латинскими большими и маленькими символами составленными по первым буквам названия соответствующих аминокислот (см. 3-ий столбец таблицы 1).

Последовательность в файле может включать пробельные символы, переносы строк, дефисы: '-' и числа. Всё это носит вспомогательный характер для облегчения ориентирования в сосой последовательности. Предполагается, что нуклеотидные последовательности могут быть очень длинными, до миллионов «буковок». С учётом вспомогательных символов файлы могут быть довольно большими. Нуклеотидные последовательности можно скачать на сайте [2], а аминокислотные на сайте [3].

Дефис означает, что что в данной позиции может находится произвольный нуклеотид или аминокислота, но какой именно мы в текущий момент не знаем. В биоинформатике у дефиса может быть иной смысл, но в рамках данного задания практикума он рассматриваться не будет.

Как известно из школьного курса биологии белок получается в несколько стадий. Грубо эти стадии следующие:

1. **Транскрипция** — по последовательности ДНК строится цепочка матричной РНК. Место в ДНК откуда начинается построение матричной РНК будем называть геном. (Такое определение гена не точное, но для целей задания подойдет). Процесс построения матричной РНК от начала гена обрывается на некоторой позиции, сама эта позиция носит вероятностный характер и определяется так называемыми шпильками, которые возникают на матричной РНК. Матричная РНК также является нуклеотидной последовательностью, но все нуклеотиды 't' в ней заменены на нуклеотиды 'u'.

2. **Трансляция** — по последовательности матричной РНК, начиная со стартового кодона (определённой тройки нуклеотидов) генерируется аминокислотная цепочка. Генерация останавливается в момент встречи СТОП-кодона. В таблице 2 представлены правила соответствия РНК кодонов аминокислотным остаткам, тоесть по сути задают принцип преобразования РНК в ДНК.

2 Программы поиска

2.1 Поисковик кодирующей белок последовательности

По аминокислотной последовательности, которая предполагается фрагментом белка (например Гемоглобина) требуется найти все нуклеотидные последовательности, которые могли бы закодировать данный фрагмент белка. В качестве находки необходимо выдать каждый пункт, начиная с новой строчки:

1. имя последовательности закодированное в FASTA файле.

2. координаты начала и конца участка в нуклеотидной последовательности, имея в виду биологический порядок (без всяких дополнительных пробелов, переносов строк, и.т.п). Предполагается, что символы в последовательности пронумерованны начиная с единицы. В случае, если находка находится на комплементарной цепи ДНК (тоесть идёт в противоположную сторону, относительно прямой цепи) координаты нужно представлять отрицательными числами. Значение -1 означает первый с конца цепочки.

3. координаты начала и конца в файле в виде номера строчки в файле и номера столбца в файле. Для цепочек идущих в обратную сторону указывать координаты в файле начиная от начала файла, тоесть ранье будет конец реверсивной цепочки, а не не начало.

4. Сама последовательность находки. для нуклеотидов маленькими буквами, для белков большими буквами, в режиме 6 раз по 10 элементов в строчку, где каждая десятка отделяется пробелами. после последовательности два переноса строки.

Программа должна работать в одном из следующих режимов:

1. допускается только точное совпадение. В этом режиме дефисы просто игнорируются также как: числа, пробелы, и т.п.
2. в нуклеотидных последовательностях допускаются дефисы, в аминокислотных дефисы игнорируются. Ищутся находки в предположении, что на месте дефисов может находится любой нуклеотид, соотвественно в случае удачной подстановки, будет зафиксирована находка.
3. в аминокислотной последовательности допускаются дефисы, при этом в нуклеотидной дефисы игнорируются. В случае удачной подстановки фиксируется находка.
4. Дефисы игнорируются в обоих послеловательностях, но допускается удаление или вставка из последовательности не более чем некоторое число аминокислот задаваемое в параметре. Находка будет по неточному сопоставлению.

3 Примеры ввода и вывода

3.1 Ввод

```
>litle
MLII

>seq RNA
1 GCU-AAAAAAAAAA-CU
2 gcgcgcgcgcgcgcAUG
3 CUUAUCAUAUAGAAAAA
4 AC
```

3.2 вывод

```
seq RNA
30, 44
(2,15) - (3,12)
AUGCUUAUCA UAUAG
```

Список литературы

[1] Статья на википедии про FASTA формат. `https://ru.wikipedia.org/wiki/FASTA`.

[2] Сайт геномного браузера EnsEMBL. Ссылка на геном человека. `https://ftp.ensembl.org/pub/release-110/fasta/homo_sapiens/dna/`.

[3] Сайт белковых последовательностей. `https://www.uniprot.org/uniprotkb?query=reviewed:true`.

Таблица 1: Аминокислотные остатки

Аминокислота	сокращ.	1-буква
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Таблица 2: Аминокислотные остатки

Amino acid	Codons
START	AUG
Ala / A	GCU, GCC, GCA, GCG
Arg / R	CGU, CGC, CGA, CGG, AGA, AGG
Asn / N	AAU, AAC
Asp / D	GAU, GAC
Cys / C	UGU, UGC
Gln / Q	CAA, CAG
Glu / E	GAA, GAG
Gly / G	GGU, GGC, GGA, GGG
His / H	CAU, CAC
Ile / I	AUU, AUC, AUA
Leu / L	UUA, UUG, CUU, CUC, CUA, CUG
Lys / K	AAA, AAG
Met / M	AUG
Phe / F	UUU, UUC
Pro / P	CCU, CCC, CCA, CCG
Ser / S	UCU, UCC, UCA, UCG, AGU, AGC
Thr / T	ACU, ACC, ACA, ACG
Trp / W	UGG
Tyr / Y	UAU, UAC
Val / V	GUU, GUC, GUA, GUG
STOP	UAA, UGA, UAG