**Использование расширенных регулярных выражений**

**при работе с egrep (grep –E)**

**1. Пробелы с разных сторон**

Чтобы обозначить, что искомый фрагмент текста отгорожен от всего остального в строке пробелами/табуляцией (иначе говоря, искать фрагмент текста как слово), можно использовать \< слева от фрагмента и \> справа. Чтобы обозначить, что фрагмент является началом слова, то следует поместить только \< слева от него, для того, чтобы искать фрагмент как конец слова – поместить \> справа.

egrep ‘\<rrs\>’ ~/term1/block2/Zlobin\_NC\_012563\_genes.txt

(здесь и далее ~ значит /home/students/y13/alexander.zlobin/)

Эта команда ищет в хромосомной таблице только записи о генах, с которых синтезируется 16S рибосомальная РНК.

**2. Использование диапазонов**

Для обозначения любого символа из введенного набора или из диапазона используются квадратные скобки. Для поиска всего, кроме заданного набора, перед ним ставится ^: [^набор].

egrep ‘^CLM\_000[0-9]’ ~/term1/block2/Zlobin\_NC\_012563\_genes.txt

Эта команда ищет в хромосомной таблице записи, начинающиеся с CLM\_0000 по CLM\_0009 включительно (в случае их наличия).

egrep ‘\<rr[slf]\>’ ~/term1/block2/Zlobin\_NC\_012563\_genes.txt

Эта команда ищет в хромосомной таблице записи о генах, с которых синтезируются 5S, 16S, 23S рибосомальные РНК.

**3. Использование альтернатив**

Для поиска строчек, содержащих либо один фрагмент, либо другой, используется разделитель фрагментов |.

egrep ‘\<rrs\>|\<rrl\>|\<rrf\>’ ~/term1/block2/Zlobin\_NC\_012563\_genes.txt

Эта команда делает то же самое, что и использующая ‘\<rr[slf]\>’.

**4. Использование повторов**

Некоторые символы обозначают, сколько раз встречается символ, стоящий за ними, в фрагменте.

\* – символ либо не встречается, либо встречается любое количество раз.

? – символ либо не встречается, либо встречается один раз.

+ – символ встречается минимум один раз.

{n,m} – символ встречается от n до m раз включительно.

{n},{n,},{,m} – символ встречается ровно n раз, минимум n раз, максимум m раз соответственно.

egrep ‘^CLM\_00[0-9]{2}’ ~/term1/block2/Zlobin\_NC\_012563\_genes.txt

Эта команда ищет в хромосомной таблице записи, начинающиеся с CLM\_0000 по CLM\_0099 включительно (в случае их наличия).

egrep ‘^ATOM {4,}[0-9]{1,2}\>’ ~/term2/block2/pr3/3QIX.pdb

Эта команда ищет в моем pdb-файле строчки с информацией об атомах с номерами от 1 до 99 включительно.

**5. Сочетание условий**

Чтобы найти строчки, содержащие оба фрагмента, можно использовать операции с потоками.

egrep ‘^CLM\_00[0-9]{2}’ ~/term1/block2/Zlobin\_NC\_012563\_genes.txt | egrep ‘\<rr[slf]\>’

Эта команда ищет в хромосомной таблице записи, начинающиеся с CLM\_0000 по CLM\_0009 включительно (в случае их наличия) и содержащие информацию о генах, с которых считываются рибосомальные РНК.