“BUSCA DE SEQUÊNCIAS CONSERVADAS UTILIZANDO EXPRESSÕES REGULARES”

DAVI SALLES

CONTEXTUALIZANDO…

No decorrer da evolução dos organismos, muitas estruturas permanecem extremamente conservadas por conta da importância funcional que elas exercem. Essa conservação pode até estar microscopicamente presente em biomoléculas, como ácidos nucleicos (DNA e RNA) e proteínas. Com o avanço das tecnologias de sequenciamento, podemos acessar as sequências de aminoácidos de alguns organismos e verificar a sua existência conservada em proteínas de outras espécies. A presença dessas sequências pode ser descrita em bancos de dados biológicos com a utilização de nomenclaturas padronizadas:

- x : refere-se a presença de qualquer aminoácido
- [ ] : refere-se a presença de um dos aminoácidos dentro dos colchetes. Por exemplo, “[GAL]” pode ser G ou A ou L.
- ( ) : refere-se ao número de vezes que o aminoácido anterior aparece. Por exemplo, “IV(3)” seria IVVV. Ou mesmo “V[PF](2)” poderia ser VPF, VPP ou VFF.
- < : refere-se a presença da sequência na porção N-terminal da proteína (no “início”)
- > : refere-se a presença da sequência na porção C-terminal da proteína (no “final”)

Exemplos:

Proteínas que são transportadas para o núcleo celular frequentemente apresentam um sinal de endereçamento composto pela sequência “P(2)-L(3)-A-L-V” que pode ser compreendido por “PPLLLALV”. Já as proteínas que permanecem na membrana do retículo endoplasmático possuem um sinal de endereçamento de permanência na porção C-terminal que frequentemente apresenta a sequência “A(2)-x(2)->” que é equivalente a “AAXXCOO^{-}”, sendo o “X” uma representação de qualquer aminoácido e o “COO^{-}” a porção carboxila terminal da proteína.

PROBLEMÁTICA...

No curso, você viu que podemos fazer buscas textuais simples com o comando `grep`. Poderíamos, por exemplo, saber a quantidade de proteínas que apresentam o sinal de endereçamento nuclear apresentado anteriormente, na espécie humana com o comando:

```
grep -c 'PPLLLALV' Homo_sapiens.faa
```

Explorando a sintaxe do `grep`, sabemos que "-c" é um parâmetro para retornar somente a quantidade de vezes que ocorrência textual "PPLLLALV" está presente no arquivo "Homo_sapiens.faa".

Mas como poderíamos fazer uma busca textual mais complexa utilizando o exemplo do sinal de permanência no retículo endoplasmático apresentado anteriormente? Como representar a dupla ocorrência de qualquer aminoácido? Como informar ao comando que queremos encontrar essa ocorrência somente no final da sequência proteica?

O comando `grep` não entende o padrão utilizado nos bancos de dados biológicos e instruções como `grep -c 'A(2)-x(2)->' Homo_sapiens.fna` não funcionam para busca da sequência.

Alternativamente, temos uma forma de representar ocorrências textuais mais complexas na programação com o uso de **expressões regulares** (regular expression, REGEX), que são basicamente um consenso de caracteres semelhantes às nomenclaturas apresentadas anteriormente. Você pode compreender um pouco mais sobre expressões regulares nesse material do Aurélio Marinho Jargas, onde ele introduz o conceito, mostra exemplos e relaciona sua utilização de busca textual com o comando `grep`.

Depois de ler o material, você já será capaz de reformular o exemplo anterior utilizando expressões regulares e terá um comando próximo disso:

```
grep -c 'AA..$' Homo_sapiens.faa
```

MÃO NA MASSA!

Tá na hora de usarmos o arquivo “Trypanosoma_cruzi.faa” que te enviamos por email: faça o download para o seu computador, abra a linha de comando e navegue pelos diretórios até conseguir acessá-lo… Esse arquivo possui todas as sequências proteicas do parasito *Trypanosoma cruzi*, causador da doença de Chagas.

Você vai utilizar o `grep` no arquivo para buscar as sequências conservadas descritas abaixo, evidenciado o comando que utilizou e a quantidade de ocorrências encontradas.

- **RNAm da beta-tubulina:** <-M-R-[DE]-[IL]
- **Sinal de endereçamento de proteínas secretadas pela via clássica:** [KRHQSA]-[DENQ]-E-L->
- **Região A (P-loop) de ligação ao ATP/GTP:** [AG]-x(4)-G-K-[ST]
- **Assinatura molecular da proteína ribossomal L30e:** [STA]-x(5)-G-x-[QKRN]-x(2)-[LIVMQ]-[KRQT]-x(2)-[KR]-x-[GS]-x(2)-[KQ]-x-[LIVM](3)

Por fim, caso você queira descobrir informações sobre as proteínas do *Trypanosoma cruzi* que possuem a assinatura molecular da proteína ribossomal L30e, qual estratégia você utilizaria?

DICA: Ao explorar o arquivo fasta com o comando `less` você pode observar que ele possui a estrutura de “cabeçalho informativo” (header) em uma linha e a respectiva sequência proteica na linha logo abaixo, sem apresentar quebra. Explore o manual do comando `grep` (ou pesquise em fóruns!) e encontre o parâmetro que também retorna a linha anterior da ocorrência que você está procurando, assim você consegue ter acesso ao header e consequentemente as informações da proteína!