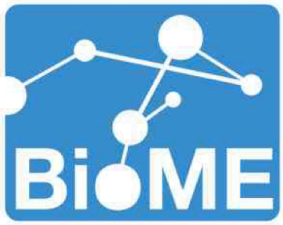


Bioinformatics Multidisciplinary Environment | Centro Multiusuário de Bioinformática

1a Avaliação - Exercícios

Parte 1 - Padrões em Sequências de Proteínas

Considere o seguinte alinhamento de um grupo de sequências de proteínas. Neste grupo, apenas sequências com motivos verdadeiros são encontradas. No geral, elas possuem cerca de 16,5% de identidade.

```
P49918      .......MSD ASLRSTSTME RLVARGTFPV LVRTSACRSL FG...PVDHE
Q96TE0      .......MSN .VRVSNGSPS LERMDARQAD HPKPSACRNL FG...PVDHE
Q91603      .....MAAFH IALQEEMIVA SPAALPRLSL GTGRGACRNL FG...PIDHD
Q4FK34      .......... .........M SNL.GDVRPV PHRSKVCRCL FG...PVDSE
Q9U6R5      .......... .......... ...MAATTAG DGKRKAARCL FGK..PDPEE
Q179M8      MSARVCNPVA LSEIAKLRSP AVVRKPMTNS ISLARVKRDL FG...PVDKQ
Q61CE7      .......... .......... .......... ...MSARRCL FGRPTPEQRA


P49918      ELSRELQARL AELNAEDQNR WDYDFQQDMP LR....GPG. RLQWTEVDSD
Q96TE0      ELTRDLEKHC RDMEEASQRK WNFDFQNHKP L......EG. KYEWQEVEKG
Q91603      ELRSELKRQL KEIQASDCQR WNFDFESGTP L......KG. TFCWEPVETK
Q4FK34      QLRRDCDALM AGCLQEARER WNFDFVTETP L......EG. NFVWERVRSL
Q9U6R5      QVSRQLNSSL EEMYKKDSRK FNFDFSGGVP IVG...SRG. DYEFESISAS
Q179M8      ESKNFIDRQL AAQNDALSKK WGFDFTAGEP L.....QNHE QYQWERVPPT
Q61CE7      RTREWLDNAC KRIREEESKK WGFDFELGMP LPSLMISTEV DYKYEILPEC


P49918      SVPAFYRETV QVGRCRLLLA PRP.VAVAVA VSPPLEPAAE SLDGLEEAPE
Q96TE0      SLPEFYYRPP RPPKGACKVP AQESQDGSGS RPAAPLIGAP A...NSEDTH
Q91603      DVPSFYS... PS.RSLAANT TPQSRQQQPL L...VSRQPE P.....REEA
Q4FK34      GLPKVYLSPG SRS.RDDLGG DKRPSTSSAL LQG..PAPED HVALS.LSCT
Q9U6R5      EVPSFYREKI VRPRKIIARR NSTPVSDTVE MPSESPPVVE SNETPLLIAS
```

```
Q179M8      SAPACFTGMV TLTRGAHRVP QSSTISEDLL DQRAERENAS ....LYRHPS
Q61CE7      SVPEFYRTKV ISVNTSHSTH TDLNLSSTTL TPLSSPSTSE K...EEPSLM


P49918      QLPSVPVPAP ASTPPPVPVL APAPAPAPAP VAAPVAAPVA VAVLAPAPAP
Q96TE0      LVDPKTDPSD SQTGLAEQCA GIRKRPATDD SS..TQNKRA NRTEENVSDG
Q91603      PVDTVRNVPN PPCAKENAEK IIKRCQGVKG PTKASANTST QRRKREITTP
Q4FK34      LVS...ERPE DSPGGPGTSQ GRKRRQTSLT DFYHSKRRLV FCKRKP....
Q9U6R5      TSTEVTVYEK PVTRSSAAKQ SIEQQETYNL KQTKLTNYMP VRKRRSETCL
Q179M8      SISSPASVSG SDSESDCSFE TVRTHPLVLR SETIVSINTA STTTITSSST
Q61CE7      DHNSSFEDEE EPKKWLFREP PTPRKSPQKR QQKVTDFYTI TRKKNSMSP.


P49918      APAPAPAPAP VAAPAPAPAP APA......P APAPAPAPDA APQESAEQGA
Q96TE0      SPNAGSVEQT P....KKPGL RRRQT..... .......... ..........
Q91603      IT........ .......... .......... .......... ..........
Q4FK34      .......... .......... .......... .......... ..........
Q9U6R5      VTAAVSMSRS VSIDSSMESC KEKRGSKIVH NNKGAPKRPL RFVASNVPKS
Q179M8      PSFPATVNRA KRQQRITDYL KERKRLSTGA PKSTAAKKAR QMLMTSASPS
Q61CE7      KMSPKNVIYT P..KSRRPTV STR...SPY. .......... ..........


P49918      NQGQRGQEPL ADQLHSGISG RPAAGTAAAS AN...GAAIK KLSGPLISDF
Q96TE0      .......... .......... .......... .......... ..........
Q91603      .......... .......... .......... .......... ........DY
Q4FK34      .......... .......... .......... .......... ..........
Q9U6R5      AQSSTSDTVL VSSPRSPPAK KMTTSTRRSR RPIEAGDF.. ..........
Q179M8      AASSISSSSS ANATAQQDH. .......... .......... ..........
Q61CE7      .......... .......... .......... .......... ..........


P49918      FAKRKRSAPE .KSSGDVPAP CPSPSAAPGV GSVEQTPRKR LR
Q96TE0      .......... .......... .......... .......... ..
Q91603      FPKRKKILSA KPDAT..... ....KGVHLL CPLEQTPRKK IR
Q4FK34      .......... .......... .......... .......... ..
Q9U6R5      .......... .......... .......... .......... ..
Q179M8      .......... .......... .......... .......... ..
Q61CE7      .......... .......... .......... .......... ..
```

Clique aqui para baixar o arquivo alinhado.

Clique aqui para baixar o arquivo não alinhado.

> *O arquivo não alinhado pode utilizado para você treinar suas habilidades de alinhamento múltiplo de sequências. Para isso use o muscle ou MAFFT.*

Parte 1 (para entrega):

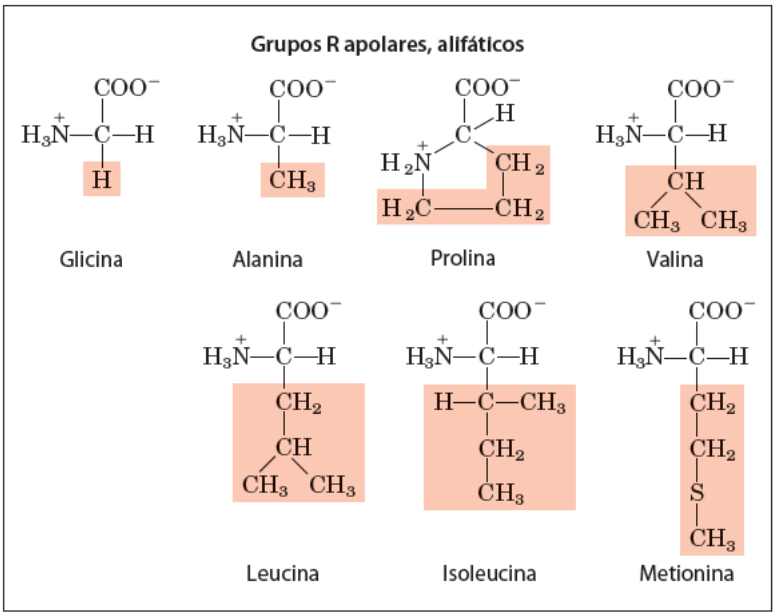
Identifique as proteínas (função e qual organismo) do arquivo acima. Descreva apenas os cabeçalhos de cada uma delas.

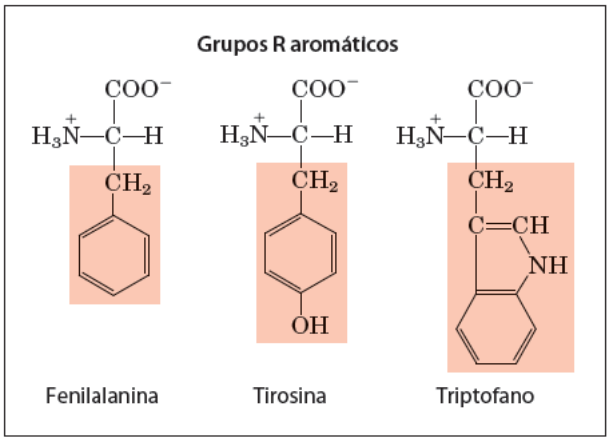
O seu objetivo é construir uma assinatura do tipo PROSITE, completamente funcional, a partir do alinhamento acima.

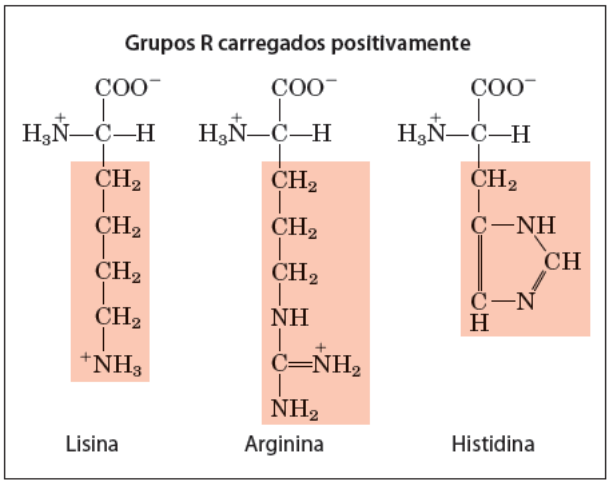
Exemplo de uma assinatura PROSITE:

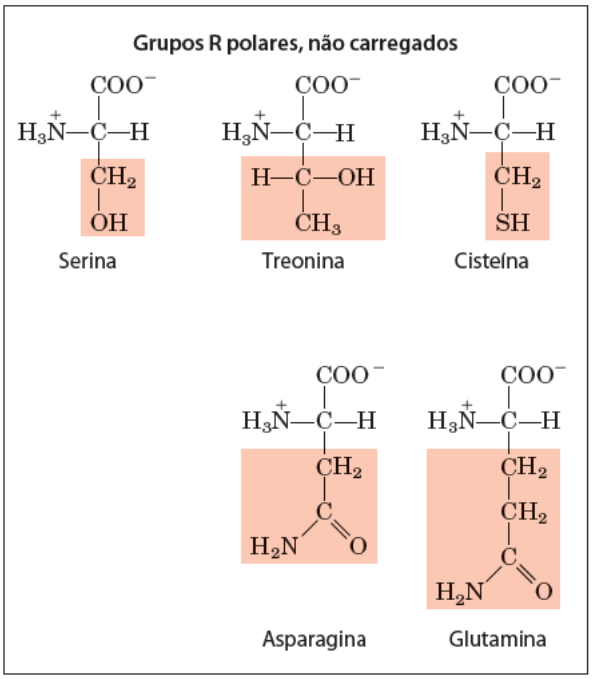
```
C-x(5)-PVCC-x(1,4)-G-x(1,6)-T-x(2)-N-x(1)-C-x(7,14)-G-x(1)-C-x(1,5)-[HN]-x(4)-P
```

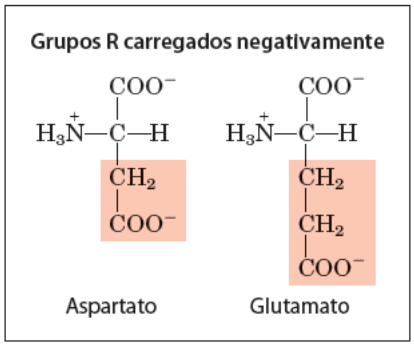
Para isso, abra o arquivo alinhado identifique os sítios conservados com ajuda de uma visualização em cores do alinhamento e vá fazendo testes para a sua assinatura. Use a tabela de classificação dos aminoácidos abaixo para incluir variações possíveis na assinatura. Adicionalmente você também pode utilizar matrizes do tipo BLOSUM ou PAM para guiar sua decisão em relação as variações possíveis (neste caso, não esqueça que as sequências acima possuem 16,5% de identidade, portanto, use a matriz adequada a este valor).











Lembre-se que a assinatura deve ser funcional para todas as sequências do *dataset* acima e para isso você pode testá-la realizando uma busca PHI-BLAST (em caso de dúvidas, veja o tutorial de Busca de Similaridades, inserido no SIGAA).

Parte 2 - Matrizes de frequência e PSSMs

Considere as sequências abaixo (formato multifasta).

```
>Q07108
MSSENCFVAENSSLHPESGQENDATSPHFSTRHEGSFQVPVLCAVMNVVFITILIIALIA
LSVGQYNCPGQYTFSMPSDSHVSSCSEDWVGYQRKCYFISTVKRSWTSAQNACSEHGATL
AVIDSEKDMNFLKRYAGREEHWVGLKKEPGHPWKWSNGKEFNNWFNVTGSDKCVFLKNTE
VSSMECEKNLYWICNKPYK
>Q95MQ1
MNSEDFSATETSSLHLKREQQSHATGTYSATYHEGSIQVPIPCAVVNVVFITTLIIALVA
LSVGQYNCPGQYASSAPPNTHVFPCSDDWIGHKGKYYLISKKTKNWTLAQNFCSKHGATL
AVIDSKEDMNFLKQHVGRAEHWIGLKNEAGQTWKWSNGQEFNNWFNLTGSENCAVLNSAE
ISSTECDKNLHWICSKPSK
>Q8SPX1
MGSENCSTTETNSLHPNRGQPSNATGPHFATHHEGSLQVPIPCAVVNVVFITVLIIALIA
LSVGQYNCPGQYVPSVPSNMHVSSCPDDWIGYQTKCYFISKKTKNWTLAQSFCSKHHGAT
LALLESKEDMVFLKQHVGRAEHWIGLKNEDGQTWKWSNGKEFNNWFKLTGSKNCPFLNST
EVGSMECEKNLHWICSKSSI
>Q5M851
MNSEECSITENSSSHLERGQRDHGTSVHFEKHREGSIQVPIPCAVLVVVLITSLIIALFA
LSVGKYNCPGFYENLESFDHHAASCKNEWFSYNGKCYFFSTTTKTWALAQKSCSEDDATL
AVIDSEKDMAFLKRYAGGLKHWIGLRNEASQTWKWANGKEFNSWFNVTGSKKCVSLNHTD
VASVDCEANLHWICSKASL
>Q3U6A8
MDSENCSITENSSSHLERGQKDHGTSIHFEKHHEGSIQVSIPWAVLIVVLITSLIIALIA
LNVGKYNCPGLYEKLESSDHHVATCKNEWISYKRTCYFFSTTTKSWALAQRSCSEDAATL
AVIDSEKDMTFLKRYSGELEHWIGLKNEANQTWKWANGKEFNSWFNLTGSGRCVSVNHKN
VTAVDCEANFHWVCSKPSR
```

Este grupo de sequências possui cerca de 63,1% de identidade e apenas um *gap* em seu alinhamento. O arquivo multifasta pode ser obtido AQUI.

O objetivo aqui é construir uma matriz de frequência de aminoácidos sítio-específica, que é o primeiro passo para a construção de uma PSSM. Segundo o NCBI:

> *Uma PSSM, ou Matriz de Pontuação Específica de Posição, é um tipo de matriz de pontuação usada em buscas BLAST nas quais as pontuações de substituição de aminoácidos são dadas separadamente para cada posição em um alinhamento de múltiplas seqüências de proteína. Assim, uma substituição de Tyr-Trp na posição A de um alinhamento pode receber uma pontuação muito diferente da mesma substituição na posição B. Isto está em contraste com as matrizes independentes de posição, como as matrizes PAM e BLOSUM, nas quais o Tyr-Trp a substituição recebe a mesma pontuação, independentemente da posição em que ocorre.*
>
> *As pontuações em uma PSSM são geralmente mostradas como inteiros positivos ou negativos. Escores positivos indicam que a substituição de aminoácidos dada ocorre com mais freqüência no alinhamento do que o esperado por acaso, enquanto os escores negativos indicam que a substituição ocorre com menos frequência do que o esperado. Grandes pontuações positivas indicam frequentemente resíduos funcionais críticos, que podem ser resíduos do local ativo ou resíduos*

> *necessários para outras interações intermoleculares.* (FONTE:NCBI).

Veja como calcular uma PSSM ou PWM (*Protein weight matrix*) é fácil. Até na própria Wikipedia mostra como calcular. Leia este site com atenção.

Identifique as proteínas (função e qual organismo) do arquivo acima. Descreva apenas os cabeçalhos de cada uma delas.

Faça o alinhamento das sequências acima utilizando os programas Muscle ou MAFFT e calcule as matrizes PFM (*Position Frequency Matrix*) e PPM (*Position Probability Matrix*).

Compare sua matriz com a matriz destas proteínas no CDD. A matriz mostrada será apenas para o sítio de ligação, que é bem conservado.

Depois proponha uma PSSM (pode ser com ou sem pseudocontagens) para estas sequências utilizando os passos descritos em *Wikipedia*.

> *Leiam com bastante atenção o exemplo dado na Wikipedia para sequências de DNA. A lógica é a mesma para proteínas, só aumenta para 20 o número possível de caracteres. Aminoácidos que não aparecem no sítio você poderá representar na matriz PSSM como -∞ ("menos infinito"). Qualquer dúvida enviem um email ou procurem o professor.*

Antes de **ampliar para o alinhamento inteiro acima**, você pode fazer o teste **apenas para o espaço entre os sítios 85 a 115 do alinhamento**, descrito abaixo:

```
CSEDWVGYQRKCYFISTVKRSWTSAQNACSE
CSDDWIGHKGKYYLISKKTKNWTLAQNFCSK
CPDDWIGYQTKCYFISKKTKNWTLAQSFCSK
CKNEWFSYNGKCYFFSTTTKTWALAQKSCSE
CKNEWISYKRTCYFFSTTTKSWALAQRSCSE
```

Parte 3 - HMMs

A partir do alinhamento proteico abaixo, proponha uma HMMs com as probabilidades de transição e emissão descritas em uma matriz.

```
Seq1    SYK--HFTYL
Seq2    NYG--PFTFL
Seq3    SYG--PL.FL
Seq4    TYE--QLSFL
Seq5    KFAG-NVDFL
Seq6    QY-GSQVTFA
Seq7    QYKG-DLSLV
```

O esquema da HMM pode ser feito em qualquer programa de desenho ou usando uma ferramenta específica (veja o item abaixo *One more thing*). A matriz pode ser entregue em formato de planilha ou tabela.

Desafio da Unidade

Escrever um programa/script, em qualquer linguagem, que resolva de forma automática, pelo menos uma das partes (1, 2 ou 3) deste exercício. O desafio pode ser efetuado individualmente ou em duplas. O aluno/dupla irá apresentar rapidamente (5-10 minutos) a estratégia utilizada em sala de aula.

One more thing

Para visualizar o alinhamento em cores, seguem algumas sugestões (existem inúmeras):

- Jalview. *Suite para trabalho com sequências.*
- ClustaX. *Programa de alinhamento múltiplo de sequências, com algumas opções de visualização.*
- Seaview. *Apenas visualização de alinhamentos no Linux. Para instalar, digite no terminal* `sudo apt-get install seaview` .
- UGene. *Suite um pouco mais completa para trabalho com sequências em bioinformática.*
- Para montar a HMM você pode utilizar a ferramenta HMMEditor. Basta selecionar na caixa de seleção no final da página o programa HMMVE_1.2.tar.gz. O programa é em Java.