Bioinformática I – Setembro/2023

Tutorial

Prof. Célio Dias Santos Júnior

Neste tutorial, você utilizará ferramentas de cada aula que cursamos e que poderão analisar online. Para este tutorial, utilizaremos as seguintes sequências:

>protein_a
MAAQKSALILLPPEDAEEIEVIVTGDVLVRGGLQVLYAGSSTEPVKCAKGARIVPDVALKDVKNKTFDIIIIPGGPGCSKLAE
CPVIGELLKTQVKSGGLIGAICAGPTVLLAHGIVAERVTCHYTVKDKMTEGGYKYLDDNVVISDRVITSKGPGTAFEFALKI
VETLEGPEKTNSLLKPLCLAK
>protein_b
MAQKSALIILAAEGAEEMEVIITGDVLARGEIRVVYAGLDGAEPVKCARGAHIVPDVKLEDVETEKFDIVILPGGQPGSNTL
AESLLVRDVLKSQVESGGLIGAICAAPIALLSHGVKAELVTSHPSVKEKLEKGGYKYSEDRVVVSGKIITSRGPGTAFEFALK
IVELLEGKDKATSLIAPMLLKL
>protein_c
MLSLLAVVSLAAATLAAPAASDAPGWRFDLKPNLSGIVALEAIVVNSSLVVIFDRATGDQPLKINGESTWGALWDLDTSTV
RPLSVLTDSFCASGALLSNGTMVSMGGTPGGTGGDVAAPPGNQAIRIFEPCASPSGDGCTLFEDPATVHLLEERWYPSSVRIF
DGSLMIIGGSHVLTPFYNVDPANSFEFFPSKEQTPRPSAFLERSLPANLFPRAFALPDGTVFIVANNQSIIYDIEKNTETILPDIP
NGVRVTNPIDGSAILLPLSPPDFIPEVLVCGGSTADTSLPSTSLSSQHPATSQCSRIKLTPEGIKAGWQVEHMLEARMMPELVH
VPNGQILITNGAGTGFAALSAVADPVGNSNADHPVLTPSLYTPDAPLGKRISNAGMPTTTIPRMYHSTVTLTQQGNFFIGGN
NPNMNFTPPGTPGIKFPSELRIETLDPPFMFRSRPALLTMPEKLKFGQKVTVPITIPSDLKASKVQVALMDLGFSSHAFHSSAR
LVFMESSISADRKSLTFTAPPNGRVFPPGPAVVFLTIDDVTSPGERVMMGSGNPPPTLE

Predição de características físico-químicas

Vamos primeiro realizar uma análise das sequências quanto à potenciais características de interesse, assim, primeiro precisamos saber algumas informações, que o ProtParam é capaz de nos informar. Atente-se que o programa aceita apenas uma sequência por vez, tendo como exemplo o resultado da proteína A:

- *Molecular weight: 19555.98* (19,6kDa)
- *Theoretical pI: 6.15* (Proteína ácida)

- *Amino acid composition*
  Rica em G - presença de domínios com pouca conservação de sequências
  Rica em L – se presente em *repeats*, revela domínio estrutural de ferradura com conformações α/β
  Rica em V - termoestável

Total number of negatively charged residues (Asp + Glu): 22
Total number of positively charged residues (Arg + Lys): 21

Cargas relativamente iguais 21/22 ~ 1 → Neutralidade mostra solubilidade.

Podemos determinar características ópticas para calcular a concentração da proteína em solução, no entanto, veja o que o programa nos informa sobre essas predições:

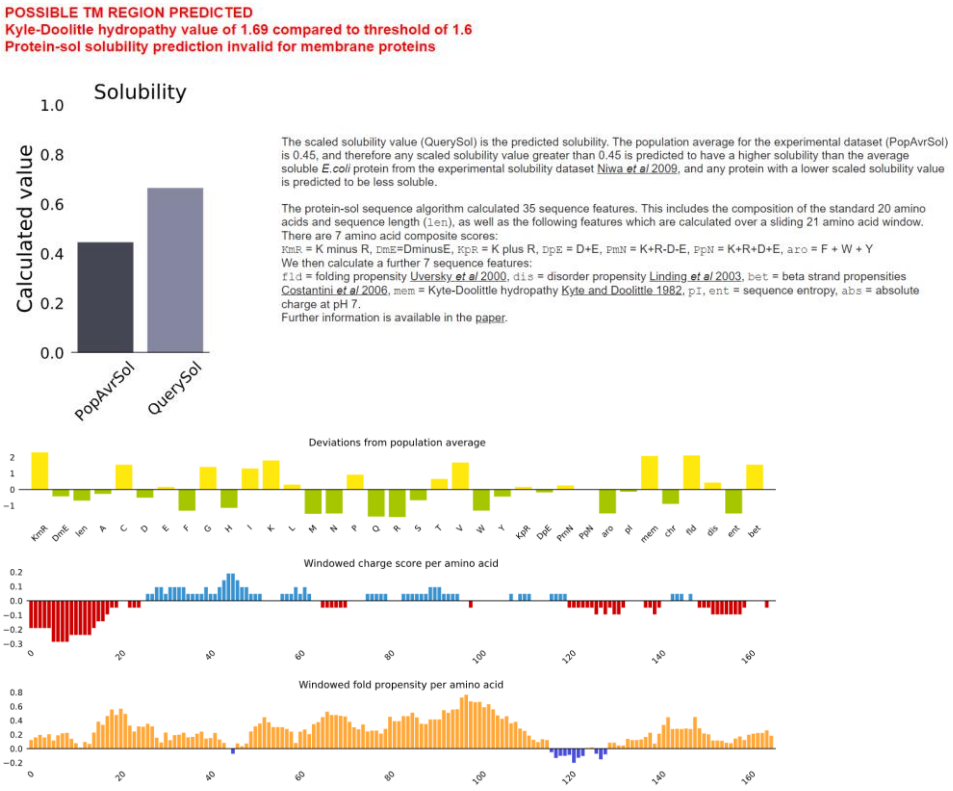
Extinction coefficients:
This protein does not contain any Trp residues. Experience shows that
this could result in more than 10% error in the computed extinction coefficient.
Extinction coefficients are in units of M-1 cm-1, at 280 nm measured in water.
Ext. coefficient 6335
Abs 0.1% (=1 g/l) 0.324, assuming all pairs of Cys residues form cystines
Ext. coefficient 5960
Abs 0.1% (=1 g/l) 0.305, assuming all Cys residues are reduced

Outros resultados incluem:

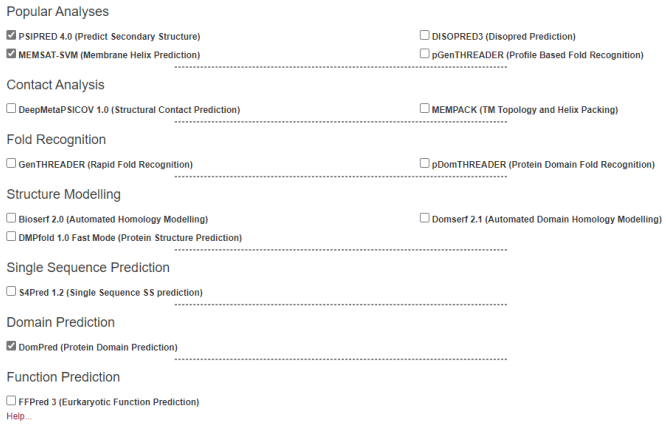
- *The instability index (II) is computed to be 31.51* (Proteína estável)
- *Aliphatic index: 111.61* (Termoestabilidade)

- *Grand average of hydropathicity (GRAVY): 0.244* (Tende a ser solúvel)

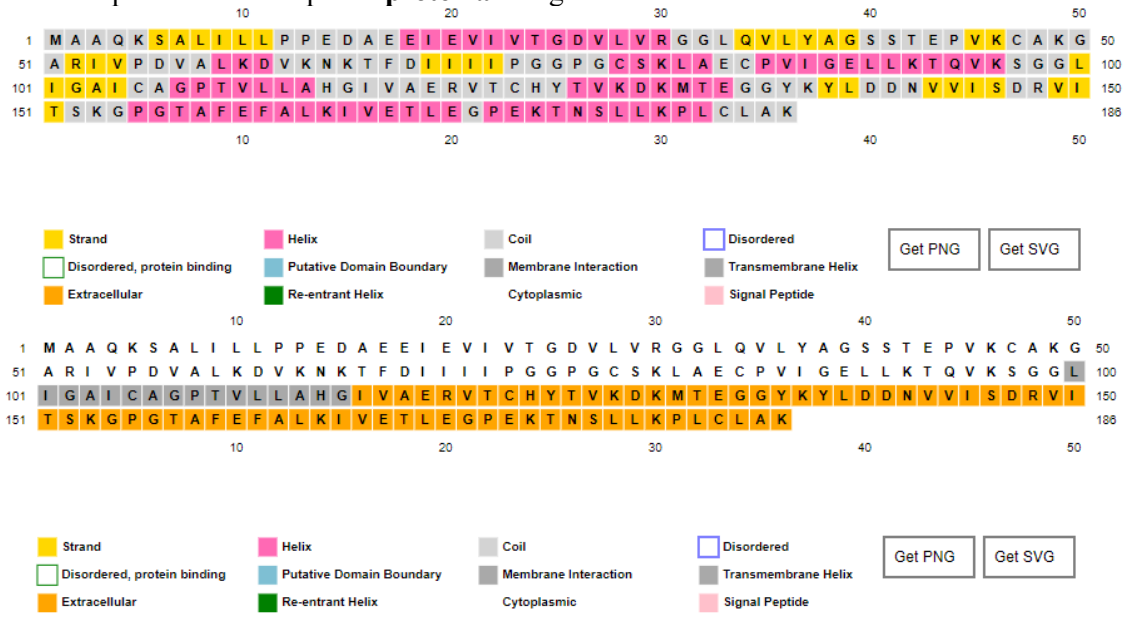
Podemos ainda buscar por informações com base numa predição mais robusta de solubilidade utilizando o Protein-Sol. Ele nos informa, por exemplo, a potencial existência de uma região transmembranar (TM) com base no perfil de hidrofobicidade utilizando a escala de Kyte-Doolitle. O gráfico ainda mostra que a nossa proteína de interesse (A), ainda tem um score muito maior que a média das populações de proteínas média, o que indica elevada solubilidade.



O programa ainda revela que há uma condição para enovelamento, uma alternância de regiões carregadas com cargas opostas, o que indica interação de segmentos subsequentes com os precedentes. Com base na sequência primária ainda podemos verificar a existência de sequência secundária, e como esses motifs se organizam. Para isso podemos utilizar o software PsiPred. Ele nos revela informações úteis, por exemplo, onde começam e terminam as folhas-beta pregueadas ou mesmo as alfa-hélices. O PsiPred ainda pode fornecer muitas outras informações, como hélices transmembranares (MEMSAT), disordem localizada (DISOPRED), análise de contato de resíduos/estruturas (MEMPACK), reconhecimento de folding (GenTHREADER) e ainda simulação de enovelamentos (DMPfold, Domserf). Também podemos predizer domínios e muitas outras informações.



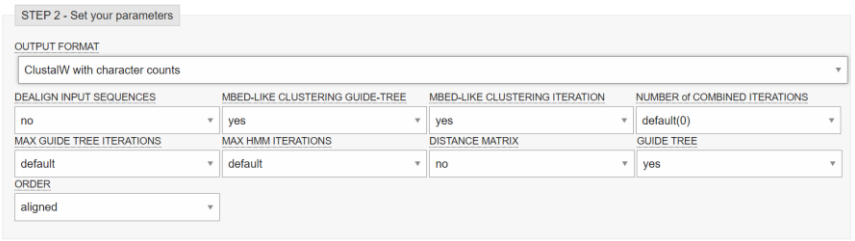
Um exemplo de resultado para a **proteína A** segue abaixo:



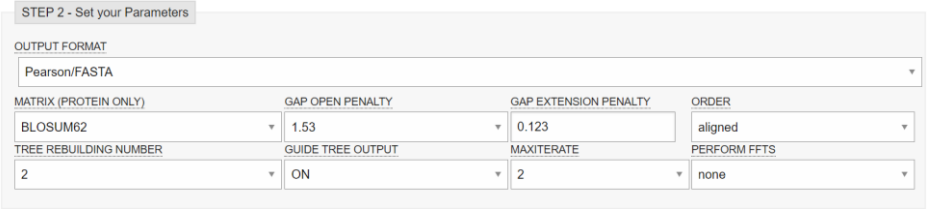
Testando a relação entre as sequências

Uma vez tendo informações de cunho físico-químico dessas proteínas, podemos também tentar alinhá-las para visualizar se há uma conservação entre elas. A primeira etapa é selecionar um bom programa para isso, sabemos que para sequências relacionadas um bom alinhamento depende mais da matriz a ser utilizada e o algoritmo dita a velocidade e exaustividade de busca entre o melhor alinhamento. Para testar 2 *approaches* diferentes, vamos até o ClustalOmega e o MAFFT.

Existem diferentes opções para esses programas funcionarem, por exemplo, o ClustalO tem parâmetros não tão óbvios:



Isto porque este programa realiza um alinhamento com base probabilística com cadeias ocultas de Markov (HMM). Nesse alinhamento, os parâmetros dependem de árvores filogenéticas geradas a partir de blocos que assim gradualmente compõe o score de pontuação. Já no MAFFT, os parâmetros são mais simples, no entanto, o algoritmo de alinhamento múltiplo depende de uma matemática mais aplicada, com base em Transformada Rápida de Fourier (FFT). Assim, os parâmetros do MAFFT tendem a ser mais compreensíveis:



Realize o alinhamento das três sequências com os dois programas em configuração Default, e depois compare os alinhamentos em termos de contiguidade.

Busca por homólogos

Utilizando blast, podemos ter uma ideia de quais organismos e funções as sequências têm. Vá até o site blastp. Uma vez lá, entre com as três sequências na janela de *query* e utilize diferentes parâmetros. Para testar cada um deles, selecione o parâmetro desejado e aperte o botão **BLAST**, que está no canto inferior esquerdo da página. Você deve testar os seguintes parâmetros:

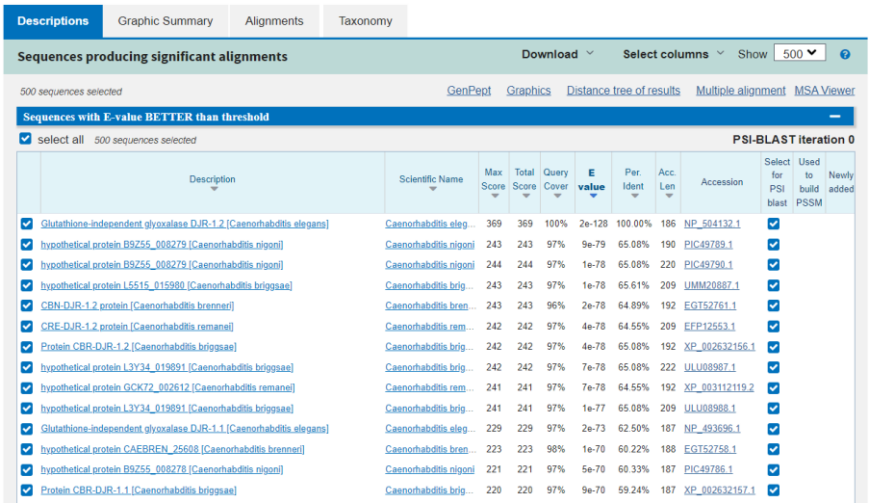
1. Database (RefSeq, Uniprot, NR)
2. Selecione um organismo: *C. elegans*
3. Algoritmo: QuickBlast, Blastp, PSI-Blast

Além de parâmetros básicos, você deve testar também parâmetros avançados do algoritmo. Neles você vai encontrar parâmetros já falados em aula, como:

- Expected threshold - teste com valores menores, como 1e-5
- Word size - Maiores e menores
- Matrix - Blosum62, Blosum90, PAM30, PAM250
- Gap costs - Teste pelo menos 2 tipos de custos para extensão e existência.

Ao utilizar esses parâmetros, vemos que é possivel regular a sensibilidade do algoritmo, bem como sua resposta ao buscar sequências no banco de dados selecionado, simplesmente alterando o campo **Max. Target Sequences**.

Testando o BlastP com as configurações Default, temos para proteína A:



Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI blast	Used to build PSSM	Newly added
Glutathione-independent glyoxalase DJR-1.2 [Caenorhabditis elegans]	Caenorhabditis eleg...	369	369	100%	2e-128	100.00%	186	NP_504132.1	☑		
hypothetical protein B9Z55_008279 [Caenorhabditis nigoni]	Caenorhabditis nigoni	243	243	97%	9e-79	65.08%	190	PIC49789.1	☑		
hypothetical protein B9Z55_008279 [Caenorhabditis nigoni]	Caenorhabditis nigoni	244	244	97%	1e-78	65.08%	220	PIC49790.1	☑		
hypothetical protein L5515_015980 [Caenorhabditis briggsae]	Caenorhabditis brig...	243	243	97%	1e-78	65.61%	209	UMM20887.1	☑		
CBN-DJR-1.2 protein [Caenorhabditis brenneri]	Caenorhabditis bren...	243	243	96%	2e-78	64.89%	192	EGT52761.1	☑		
CRE-DJR-1.2 protein [Caenorhabditis remanei]	Caenorhabditis rem...	242	242	97%	4e-78	64.55%	209	EFP12553.1	☑		
Protein CBR-DJR-1.2 [Caenorhabditis briggsae]	Caenorhabditis brig...	242	242	97%	4e-78	65.08%	192	XP_002632156.1	☑		
hypothetical protein L3Y34_019891 [Caenorhabditis briggsae]	Caenorhabditis brig...	242	242	97%	7e-78	65.08%	222	ULU08987.1	☑		
hypothetical protein GCK72_002612 [Caenorhabditis remanei]	Caenorhabditis rem...	241	241	97%	7e-78	64.55%	192	XP_003112119.2	☑		
hypothetical protein L3Y34_019891 [Caenorhabditis briggsae]	Caenorhabditis brig...	241	241	97%	1e-77	65.08%	209	ULU08988.1	☑		
Glutathione-independent glyoxalase DJR-1.1 [Caenorhabditis elegans]	Caenorhabditis eleg...	229	229	97%	2e-73	62.50%	187	NP_493696.1	☑		
hypothetical protein CAEBREN_25608 [Caenorhabditis brenneri]	Caenorhabditis bren...	223	223	98%	1e-70	60.22%	188	EGT52758.1	☑		
hypothetical protein B9Z55_008278 [Caenorhabditis nigoni]	Caenorhabditis nigoni	221	221	97%	5e-70	60.33%	187	PIC49786.1	☑		
Protein CBR-DJR-1.1 [Caenorhabditis briggsae]	Caenorhabditis brig...	220	220	97%	9e-70	59.24%	187	XP_002632157.1	☑		

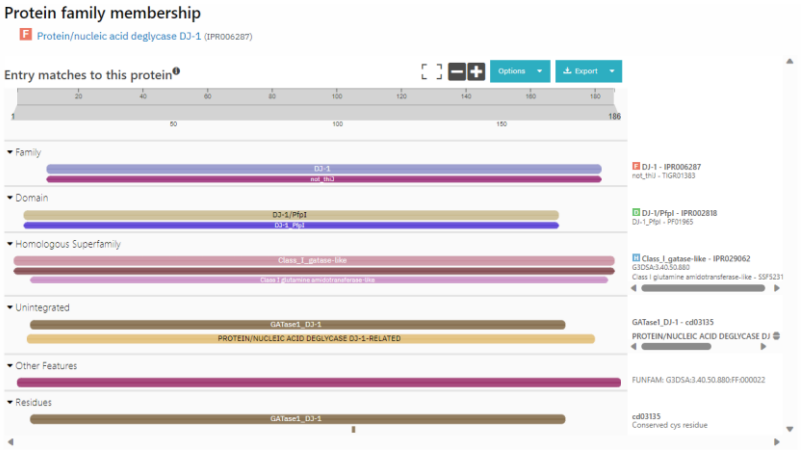
Alternativamente, podemos buscar por homólogos em bancos como o Uniprot que também revela diversas proteínas similares com níveis de informação e anotação mais específicos que aqueles presentes no NCBI.

Busca por domínios conservados

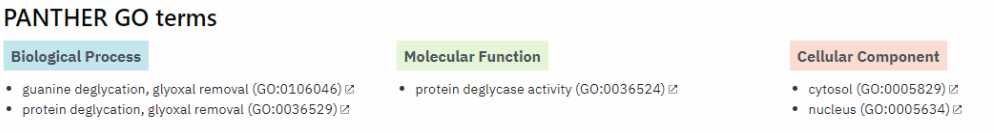
Domínios são importantes para estrutura e função das proteínas, geralmente acontecem em regiões que se conservam e que podem ser preditas por homologia. Existem vários meios de se buscar por domínios em bancos públicos, o mais utilizado é por meio de homologia à famílias, nesse âmbito, o InterproScan é um banco extremamente rico e atualizado, com ferramentas muito úteis.

Vá até o InterproScan e entre com as sequências que vamos analisar. – ATENTE-SE que a ferramenta aceita apenas uma sequência por vez. Nas opções avançadas, tente identificar outros bancos de dados dos quais falamos em aula, como o CDD do NCBI, o ANTIFAM, o PFAM e imagine uma melhor combinação deles para que se busque a função/estrutura de suas proteínas. Os resultados para essa análise serão discutidos durante a prática.

Detectamos família e conservação da sequência



E também *insights* sobre seu potencial funcional e localização:



Modelando a proteína

Para modelarmos uma proteína candidata, podemos utilizar vários métodos, neste tutorial, iremos tentar 2 deles:

1. SWISSMODEL
2. ALPHAFOLD2 via ColabFOLD

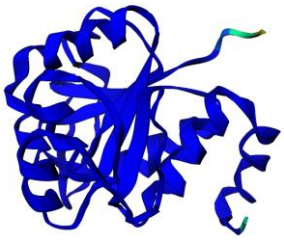
Ao se inserir as sequências nesses programas, poucos parâmetros podem e devem ser alterados. No SWISSMODEL, o processo é praticamente todo automatizado. Enquanto no ColabFOLD, podemos testar parâmetros como:

- template mode: None (para modelagem *ab initio*) ou PDB100 (para iniciar com homólogos estruturais)

- num relax: parâmetro que testa diversos modelos a partir de um relaxamento da estrutura com base em campos de energia (o número se refere à iterações)

Obtemos a estrutura seguinte com SWISSMODEL, onde azul indica confiança estatística e vermelho incerteza (note que apenas a posição terminal da proteína tem incerteza):

E a estrutura seguinte com o AlphaFold2:



Explore os critérios de qualidade da estrutura. Com a estrutura gerada no ALPHAFOLD2, teste sua qualidade utilizando a ferramenta STRUCTURE ASSESSMENT do SWISSMODEL webserver e compare os resultados com a estrutura modelada por homologia. Agora compararemos as estruturas obtidas pelos 2 métodos, por meio da ferramenta RCSB PDB - Structure Pairwise Alignment Tool. Nela carregue os 2 modelos e selecione a opção 'upload file' para ambas as entradas e depois a opção de algoritmo (neste caso, utilizaremos um modelo de estrutura rígida). Assim, vemos que as estruturas apresentam a seguinte sobreposição:

Podemos ver que ambas as estruturas estão praticamente modeladas à perfeição e se sobrepõe mesmo sendo geradas por métodos diferentes. O que indica que a sequência modelada é uma sequência conservada e semelhante à outras que tem modelos validados experimentalmente. Além disso, os motifs estão organizados corretamente em relação aos dados do sítio ativo da glyoxal reductase utilizada como proteína exemplo.