

Tutorial para Modelagem de Proteínas por Reconhecimento de Enovelamentos

~XI Escola de Modelagem Molecular em Sistemas Biológicos LNCC/MCTI

Predição de Estruturas de Proteínas (PSP)

Matheus José Novais Landim - mjni.landim76@gmail.com

Maria Luiza Pereira Baltazar - malubaltazar19@gmail.com

Priscila V. Z. Capriles Goliatt – priscila.capriles@uff.edu.br

Comandos Básicos do Linux:

cd diretório/	Entra na pasta diretório/
ls diretório/	Lista as subpastas e arquivos existentes em diretório/
cd ..	Retorna para a pasta anterior
mkdir diretório/	Cria a pasta diretório/
rm arquivo.txt	Remove arquivo.txt (.txt ou outro formato de arquivo)
rmdir diretório/	Remove a pasta diretório/
cp diretório1/arquivo.txt diretório2/	Copia arquivo.txt no diretório1/ para diretório2/

Comandos Básicos do Windows:

cd diretório\	Entra na pasta diretório\
dir diretório\	Lista as subpastas e arquivos existentes em diretório\
cd ..	Retorna para a pasta anterior
mkdir diretório\	Cria a pasta diretório\
del arquivo.txt	Remove arquivo.txt (.txt ou outro formato de arquivo)
rd diretório\	Remove a pasta diretório\
copy diretório1\arquivo.txt diretório2\	Copia arquivo.txt no diretório1\ para diretório2\

Sumário

Introdução:	4
Dia 1: Modelagem com alta taxa de identidade - Monômero	5
Compreendendo a biologia da proteína	5
Predição Estrutural Usando Modeller:	8
Passo 1: Identificando estruturas de referências usando alinhamento local	8
Passo 2: Seleção do(s) Molde(s):	14
Passo 3: Alinhamento Global entre Sequências: Alvo X molde(s):	17
Passo 4: Construção do Modelo Tridimensional	22
Gerando os modelos otimizados da proteína alvo com o Modeller:	25
Passo 5: Análise inicial dos modelos gerados	26
Alinhamento estrutural	26
Obtenção do RMSD	29
Dia 2: Predições estruturais	33
Predições Estruturais Importantes:	33
Passo 1: Predição de Peptídeo Sinal via SignalP 3.0	33
Passo 2: Predição de estrutura secundária	35
NetSurfP - 3.0	36
PSIPRED	38
JPRED	39
PSSpred	40
Passo 3: Predição de regiões transmembrana	41
TMHMM	42
UniTmp	42
Passo 4: Predições de contato	43
MapPred	43
Avaliação dos resultados obtidos	47
Dia 3: Modelagem de um monômero com restrições estruturais	49
Predição Estrutural Usando Modeller	49
Passo 1: Revisão dos resultados	49
Passo 2: Construção do modelo tridimensional	50
Passo 3: Avaliação de viabilidade dos modelos	52
Validação dos modelos	54
Gráfico de Ramachandran via SAVES-Procheck e Molprobity	54
Passo 1: SAVES-Procheck	55
Passo 2: Molprobity	58
Avaliação de energia	62
Molpdf	62
DOPE score	62

GA341 score	63
Normalized DOPE (z-DOPE)	63
RMSD	63
Comparação com os modelos sem otimização	64
Avaliação final	64
Dia 4: Utilização de técnicas alternativas para a predição estrutural de proteínas	66
Método não baseado em modelagem comparativa:	66
Utilização de Inteligência Artificial	69
Comparação da estrutura da COMT obtida com AlphaFold, Robetta e Modeller	71
Comparação de outras estruturas	73
Beta-Lactoglobulina	73
Malato Desidrogenase (MDH)	74
Hemoglobina:	75
Transportador GLUT1:	76
Modelagem comparativa x Técnicas Alternativas	76

Introdução:

O curso de predição estrutural de proteínas está dividido em duas partes. A primeira irá abordar sobre a técnica de modelagem comparativa ou baseada em molde, isto é, utiliza moléculas com estrutura tridimensional (3D) já resolvida por algum e que compartilha certo grau de homologia com a proteína alvo. A segunda parte do curso dará enfoque à técnicas independentes de referências, as quais utilizam informações físico-químicas das interações entre os aminoácidos a fim de encontrar um modelo 3D.

O presente tutorial está orientado para usuários do sistema operacional Linux e tem por objetivo exemplificar a utilização do software Modeller (<http://www.salilab.org/modeller/>), para predição de estruturas de proteínas (PSP) via modelagem baseada em molde (*template based*), e dos softwares AlphaFold (<https://alphafold.ebi.ac.uk/>) e Robetta (<https://robbetta.bakerlab.org/>), para a predição de estrutura de proteínas por meio de técnicas estatísticas, matemáticas e por inteligência artificial.

Principais etapas da técnica de Modelagem Comparativa:

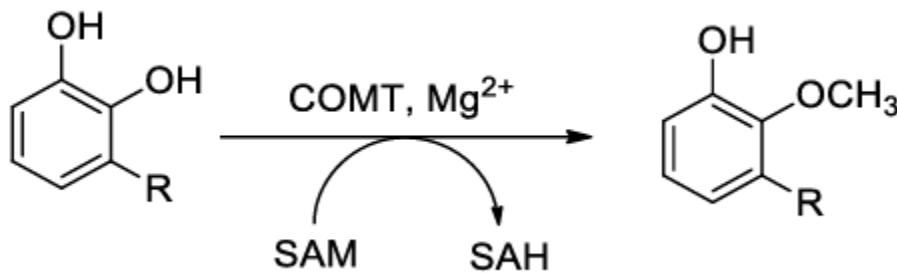
- Identificação de estruturas de referências;
- Seleção do molde (*template*)
- Alinhamento entre sequências alvo e molde(s);
- Predições estruturais;
- Construção do modelo;
- Validação do modelo.

Dia 1: Modelagem com alta taxa de identidade - Monômero

Compreendendo a biologia da proteína

Antes de realizar a predição de estrutura de proteínas é necessário compreender não só os processos computacionais envolvidos, mas também a biologia molecular e estrutural do alvo que estamos trabalhando. Esse processo irá permitir que os resultados obtidos na modelagem sejam os mais próximos possíveis da estrutura *in vivo*, possibilitando previsões mais precisas e confiáveis. Neste momento iremos apresentar um breve resumo sobre a proteína, mas solicitamos que leiam os artigos disponíveis em [~Minicursos/Predicao de estrutura 3D de proteinas por modelagem comparativa/dia1/Artigos](#) para uma maior compreensão.

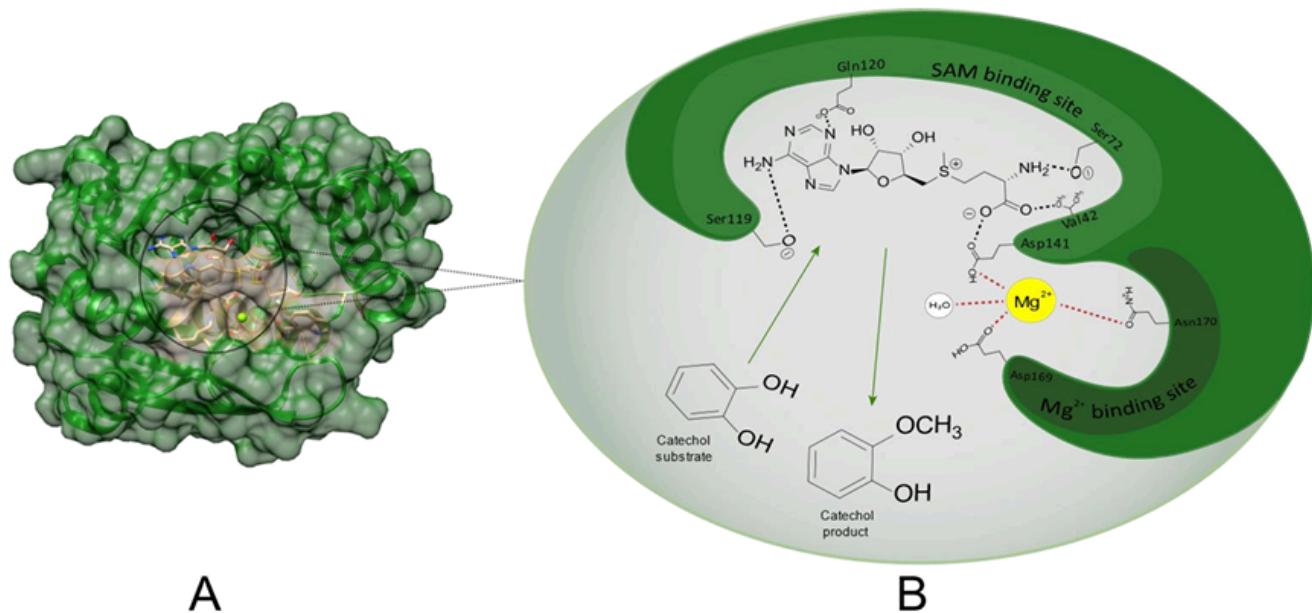
A Catecol-O-metiltransferase (COMT, E.C. 2.1.1.6) é uma enzima encontrada em diversos tecidos humanos, como o fígado, rins, trato gastrointestinal e cérebro. Ela é responsável por realizar a metilação do oxigênio no grupamento catecol de diversas moléculas, principalmente as catecolaminas, como a dopamina, epinefrina e norepinefrina. Para que a metilação ocorra, a enzima necessita de 2 cofatores, sendo eles um íon de magnésio (Mg^{2+}) e uma molécula de S-adenosil-L-metionina (SAM), produzindo o produto metilado e uma molécula de S-adenosil-L-homocisteína (SAH).



Nos humanos, a COMT pode ser encontrada em **duas isoformas distintas**, a forma **solúvel** (SCOMT) e a forma **ancorada em membranas** (MBCOMT). No SNC, a expressão da isoforma ancorada em membrana é maior, enquanto nos outros tecidos há maior presença da isoforma solúvel. A localização celular das duas isoformas também se difere entre ambas, enquanto a SCOMT pode ser encontrada no citoplasma, com um peso molecular de 24.7 kDa e 221 resíduos de aminoácidos, a MBCOMT é encontrada ancorada na membrana de retículo endoplasmático rugoso, orientada em direção ao citoplasma, com um peso molecular de 30 kDa e 271 resíduos de

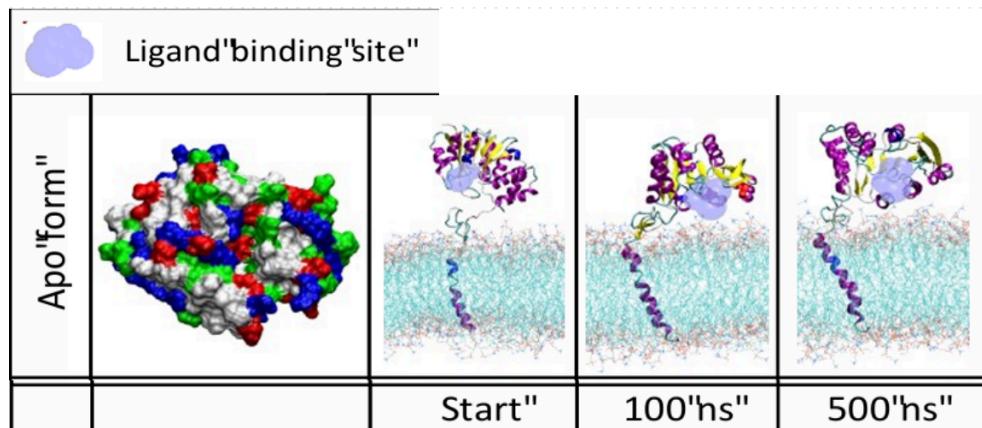
aminoácidos. Os resíduos adicionais dessa isoforma compreendem a sequência de peptídeo responsáveis pelo ancoramento na membrana e um *linker*, os quais são clivados para a formação da SCOMT.

Estruturalmente, a COMT humana é composta de uma estrutura proteica mista α/β , consistindo em um núcleo de folhas β com sete fitas, das quais as fitas $\beta_1\text{-}\beta_6$ são paralelas e a fita β_7 é antiparalela, inserida entre duas camadas de α -hélices. O sítio catalítico é moldado para acomodar o substrato catecol e o sítio de ligação ao Mg^{2+} em um sulco raso na superfície externa da enzima, enquanto o sítio de ligação ao SAM está localizado dentro da estrutura da enzima, em uma fenda mais interna. Nesta reação de transferência SN_2 , a sequência de ligação precisa seguir uma ordem crucial para metilar o substrato catecol com sucesso. Assim, o SAM deve ser o primeiro a se ligar, seguido pelo Mg^{2+} e, por fim, pelo substrato. As principais interações moleculares dos cofatores com o sítio ativo da proteína estão ilustradas na figura abaixo (extraída de Vicente, 2020).

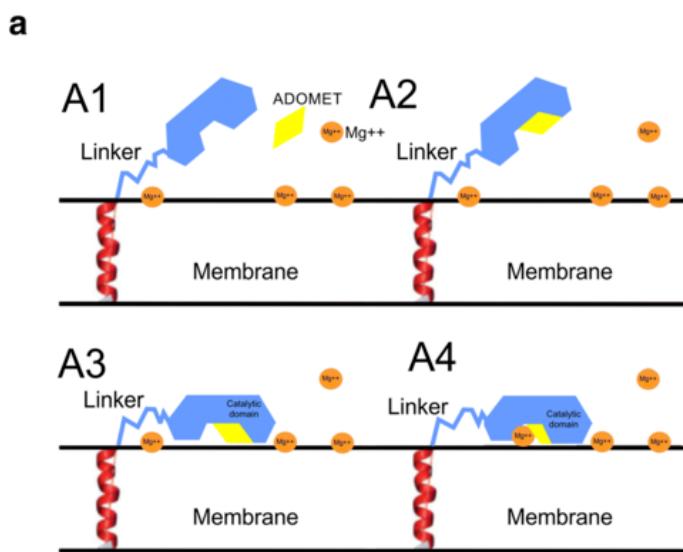


Enzimas bitópicas, como a MB-COMT, possuem um domínio catalítico que é conectado à membrana lipídica por meio de um único segmento de ligação e uma hélice transmembranar. A única estrutura completa disponível de uma enzima bitópica individual sugere que a hélice transmembrana e o segmento de ligação ajudam a定位 o domínio catalítico em relação à

membrana. Isso indica que a membrana lipídica pode ter um papel direto no processo catalítico da enzima e na escolha dos substratos. No geral, as evidências mostram que tanto os substratos potenciais quanto o domínio catalítico interagem com a membrana lipídica, influenciando a atividade catalítica da MB-COMT. Uma representação do ancoramento da COMT na membrana pode ser observado nas imagens a seguir:



Avaliação das trajetórias por dinâmica molecular da conformação apo da MBCOMT (Magakar et. al., 2018).



Representação esquemática do mecanismo de ancoramento dos cofatores na MBCOMT. **(A1)**: O domínio catalítico da MB-COMT interage fracamente com a membrana; **(A2)**: O cofator S-adenosil-l-metionina (ADOMET) se liga ao sítio catalítico da MB-COMT; **(A3)**: Quando a MB-COMT se complexa com o ADOMET, o sítio catalítico se abre em direção à membrana, permitindo que a

enzima se ligue à superfície da membrana; (A4): Finalmente, a MB-COMT se liga a um íon de Mg²⁺ que já está presente na superfície da membrana (Postila e Rög, 2019).

Predição Estrutural Usando Modeller:

Passo 1: Identificando estruturas de referências usando alinhamento local

Neste primeiro passo, iremos realizar um alinhamento local da sequência de aminoácidos da molécula de interesse aos principais bancos de dados de estruturas 3D de proteínas. Para isto será utilizado o algoritmo Blast.

1. Identificando referências via UniProt

Antes de apresentarmos como realizar o BLAST pelo UniProt, iremos apresentar algumas funções importantes desse banco de dados. Para acessá-lo, utilize o link <https://www.uniprot.org/> e, na barra de pesquisa, pesquise pela proteína “COMT” ou “Catechol O-methyltransferase” (o site é capaz de reconhecer tanto o nome completo, quanto a sigla).

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share					
Entry	Entry Name	Protein Names	Gene Names	Organism	Length
P21964	COMT_HUMAN	Catechol O-methyltransferase[...]	COMT	Homo sapiens (Human)	271 AA
O88587	COMT_MOUSE	Catechol O-methyltransferase[...]	Comt, Comt1	Mus musculus (Mouse)	265 AA
A7MBI7	COMT_BOVIN	Catechol O-methyltransferase[...]	COMT	Bos taurus (Bovine)	272 AA
P22734	COMT_RAT	Catechol O-methyltransferase[...]	Comt	Rattus norvegicus (Rat)	264 AA
Q5H879	COMT_HORSE	Catechol O-methyltransferase[...]	COMT	Equus caballus (Horse)	269 AA
Q6T1F5	COMT1_AMMMJ	Caffeic acid 3-O-methyltransferase[...]	COMT	Ammi majus (Bishop's weed)	365 AA
Q43609	COMT1_PRUDU	Caffeic acid 3-O-methyltransferase[...]	COMT1	Prunus dulcis (Almond) (Amygdalus dulcis)	365 AA
Q8GU25	COMT1_ROSCH	Caffeic acid 3-O-methyltransferase[...]	COMT1	Rosa chinensis (China rose)	365 AA

Note que nesse momento já podemos identificar informações importantes sobre a proteína que estamos buscando, como o código dela no banco de dados, o organismo, e o tamanho da sequência de aminoácidos. Note que após o código temos um símbolo amarelo. Esse símbolo significa que essa proteína já teve seus dados validados manualmente pela curadoria do UniProt. O ideal é trabalharmos com proteínas que já foram validadas, mas nem todas já passaram por esse

processo, e às vezes precisamos trabalhar com as informações disponíveis. Como iremos trabalhar com a COMT de *Homo sapiens*, clique no código dela.

P21964 · COMT_HUMAN

Protein ⁱ	Catechol O-methyltransferase	Amino acids	271 (go to sequence)
Gene ⁱ	COMT	Protein existence ⁱ	Evidence at protein level
Status ⁱ	UniProtKB reviewed (Swiss-Prot)	Annotation score ⁱ	5/5
Organism ⁱ	Homo sapiens (Human)		

[Entry](#) [Variant viewer](#) 330 [Feature viewer](#) [Genomic coordinates](#) [Publications](#) [External links](#) [History](#)

Neste cabeçalho podemos ver novamente o status da proteína, indicando que ela já foi validada pelo banco de dados. À direita, há a validação da existência da proteína (**Protein existence**), isto indica o tipo de evidência que suporta a existência da proteína. Note-se que a prova de "existência de proteína" não dá informações sobre a exatidão ou correção da sequência apresentada. Embora forneça informações sobre a existência de uma proteína, pode acontecer que a sequência seja ligeiramente diferente das sequências genômicas, especialmente no caso de sequências derivadas de previsões de modelos de genes. O valor "**Experimental evidence at protein level**" indica que existem provas experimentais claras da existência da proteína. Os critérios incluem sequenciação Edman parcial ou completa, identificação clara por espectrometria de massa, estrutura de raios X ou RMN, interação proteína-proteína de boa qualidade ou deteção da proteína por anticorpos.

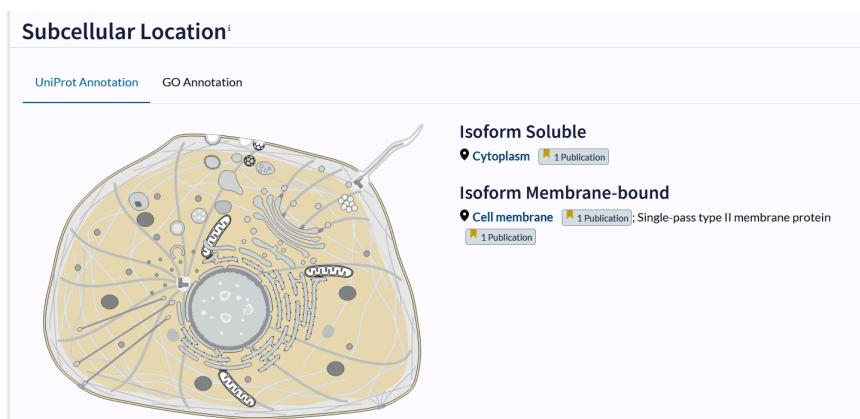
O “**Annotation score**” é uma medida heurística que avalia o conteúdo das anotações de uma entrada ou proteoma, mas não sua precisão. Esse score é calculado com base na presença e quantidade de diferentes tipos de anotações, atribuindo maior valor às que possuem evidências experimentais. A soma dessas pontuações é traduzida em uma escala de 5 pontos, onde 5 indica entradas com anotações mais completas e 1 representa anotações básicas. Esse score é utilizado para avaliar rapidamente o nível de anotação em resultados de busca, escolher membros representativos de clusters UniRef e selecionar proteomas de referência. No entanto, ele não deve ser confundido com uma medida da precisão das anotações.

Na barra à esquerda podemos observar um índice com os títulos das informações disponíveis no UniProt, como a função, sua localização celular, doenças associadas, expressão, estrutura, domínios, etc.

- | [Function](#)
- [Names & Taxonomy](#)
- [Subcellular Location](#)
- [Disease & Variants](#)
- [PTM/Processing](#)
- [Expression](#)
- [Interaction](#)
- [Structure](#)
- [Family & Domains](#)
- [Sequence & Isoform](#)
- [Similar Proteins](#)

Em “**Function**”, várias informações podem ser extraídas, como a sua função propriamente dita, seus substratos e cofatores, seus parâmetros cinéticos, sua ontologia genética (que avalia a função molecular, seus processos biológicos e seus componentes celulares), bases de dados de enzimas e vias de funcionamento, dentre outras informações.

Em “**Subcellular Location**”, há a apresentação de todos os locais celulares em que essa proteína pode ser encontrada. Note que o banco de dados informou que a COMT apresenta as isoformas solúveis e ancoradas em membrana, o que está de acordo com a literatura disponível.



Em “Structure” é possível visualizar as estruturas disponíveis no PDB e no AlphaFold para essa proteína. Além disso, na lista fornecida, é possível observar o identificador da proteína no banco de dados que a estrutura foi retirada, o método de obtenção da estrutura, sua resolução, as cadeias presentes, e os resíduos que a estrutura está cobrindo. Iremos entender a importância de cada um desses parâmetros no decorrer deste tutorial.

Structure¹

The screenshot shows a protein structure viewer with a ribbon model of a protein. Below it is a table of structures:

SOURCE	IDENTIFIER	METHOD	RESOLUTION	CHAIN	POSITIONS	LINKS
PDB	4PYI	X-ray	1.35 Å	A	51-271	PDBe · RCSB-PDB · PDBj · PDBeSum Foldseek
PDB	4PYJ	X-ray	1.90 Å	A	51-271	PDBe · RCSB-PDB · PDBj · PDBeSum Foldseek
PDB	4PYK	X-ray	2.22 Å	A	51-271	PDBe · RCSB-PDB · PDBj · PDBeSum Foldseek
PDB	4XUC	X-ray	1.80 Å	A	48-265	PDBe · RCSB-PDB · PDBj · PDBeSum Foldseek
PDB	4XUD	X-ray	2.40 Å	A	48-265	PDBe · RCSB-PDB · PDBj · PDBeSum Foldseek
PDB	4XUE	X-ray	2.30 Å	A/B	52-265	PDBe · RCSB-PDB · PDBj · PDBeSum Foldseek
PDB	5LSA	X-ray	1.50 Å	A	51-271	PDBe · RCSB-PDB · PDBj · PDBeSum Foldseek
PDB	6I3C	X-ray	1.34 Å	A	52-271	PDBe · RCSB-PDB · PDBj · PDBeSum Foldseek
PDB	6I3D	X-ray	1.45 Å	A/B	52-271	PDBe · RCSB-PDB · PDBj · PDBeSum Foldseek
AlphaFold	AF-P21964-F1	Predicted			1-271	AlphaFold Foldseek

Recomendamos que entre em um outro momento no UniProt e explore cada uma de suas funções, ele poderá ser de grande utilidade futuramente :)

Agora vamos colocar a mão na massa!

1.1. Acesse o site do UniProt: <https://www.uniprot.org/>

The screenshot shows the UniProt homepage with a dark blue header. In the top left, there's a navigation bar with links: UniProt, BLAST, Align, Peptide search, ID mapping, and SPARQL. The 'BLAST' link is highlighted with a red box and an arrow pointing to it. The main title 'Find your protein' is centered below the header. Below the title is a search bar with the placeholder 'UniProtKB' and a search button. A message below the search bar says 'Examples: Insulin, APP, Human, P05067, organism_id:9606'. On the right side of the header, there are links for 'Release 2024_03 | Statistics', 'Help', 'Feedback', and 'Cite UniProt'. At the bottom of the page, a footer states 'UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. Cite UniProt'.

1.2. Selecionar a ferramenta “BLAST”;

1.3. No campo “Enter one or more sequences (5 max). ”, deve-se colar a sequência alvo que se deseja modelar (no formato FASTA):

```
>ALVO
MPEAPPLLLAAVLLGLVLLVVLLLLLRLHWGWLCLIGWNEFILQPIHNLLMGDTKEQRIL
NHVLQHAEPGNAQSVLLEAIDTYCEQKEWAMNVGDKKGKIVDAVIQEHQPSVLLELGAYCG
YSAVRMARLLSPGARLITIEINPDCAAITQRMVDFAGVKDKVTLVVGASQDIIPQLKKY
DVDTLDMVFLDHWKDRYLPDTLLLEECGLLRKGTVLLADNVICPGAPDFAHVRGSSCFE
CTHYQSFLEYREVVDGLEKAIYKGPGSEAGP
```

The screenshot shows the UniProt BLAST search interface. At the top, there are two dropdown menus: 'Target database' set to 'UniProtKB with 3D structure (PDB)' and 'Restrict by taxonomy' with a search input field. Below these, a 'Name your BLAST job' input field contains the text 'my job title'. Under the heading 'Advanced parameters', there are several dropdown menus: 'Sequence type' (Protein), 'Program' (blastp), 'E-Threshold' (0.0001), 'Matrix' (Auto - BLOSUM62), 'Filter' (Filter low complexity regions), 'Gapped' (yes), 'Hits' (250), and 'HSPs per hit' (All). Red boxes and arrows highlight the 'Target database', 'E-Threshold', 'Matrix', and 'Filter' fields.

1.4. Buscar pela sequência usando os seguintes parâmetros:

a) “Target database”: UniProtKB with 3D structure (PDB).

1.5. Clicar em “Advanced parameters” ao final da página e alterar os valores para:

a) “Sequence type”: Protein;

b) “Program”: blastp;

c) “E-Threshold”: 0.0001;

d) “Matrix”: Auto-BLOSUM62;

e) “Filter”: Low complexity regions;

1.6. Analisar os resultados retornados:

- a) Os resultados marcados com “Reviewed (Swiss-Prot)” devem ser a escolha preferencial;
- b) Maior valor de identidade;
- c) Maior valor de score;
- d) Menor valor de E-value.

2. Identificando referência via PDB

2.1. Acessar o site PDB: <https://www.rcsb.org/>

2.2. Selecionar a opção “Advanced Search”, para realizar uma busca avançada no banco de dados;

The screenshot shows the RCSB PDB homepage. At the top, there is a dark blue header with various navigation links: Deposit, Search, Visualize, Analyze, Download, Learn, About, Documentation, Careers, and COVID-19. On the right side of the header are MyPDB, Contact us, and a user icon. Below the header is the main search area. It features the RCSB PDB logo and two statistics: 222,415 Structures from the PDB and 1,068,577 Computed Structure Models (CSMs). To the right of these stats is a search bar with the placeholder "Enter search term(s), Entry ID(s), or sequence". Next to the search bar are buttons for "Include CSM" (with a toggle switch) and a magnifying glass icon. Below the search bar are two buttons: "Advanced Search" (which has a red arrow pointing to it and is enclosed in a red box) and "Browse Annotations". At the bottom of the page, there is a footer with links to PDB-101, wwPDB, EMDDataResource, NAKB, wwPDB Foundation, and PDB-Dev. There is also a social media section with icons for Facebook, Twitter, YouTube, and LinkedIn. A banner at the very bottom encourages users to "Access Computed Structure Models (CSMs) of available model organisms" with a "Learn more" link.

2.3. Selecione “Sequence similarity”. No campo, cole a sequência alvo no formato fasta sem cabeçalho;

2.4. Defina o valor de cutoff desejado (em geral utilizamos 0.0001);

2.5. Em “Return” selecione “polymer entities” para que os resultados mostrem os valores de **identidade, E-value e gaps** do alinhamento;

2.6. Então clique em “Search”.

Advanced Search Query Builder Help

Full Text ?

Structure Attributes ?

Chemical Attributes ?

Sequence Similarity ? ←

MPEAPPLLAVALGLVLLVLLLRLHWGWLGLCLIGWNEFILQPIHNLUMGDTKEQRILNHLQHAEPGNAQSYLEAIDTYCEQKEWAMNVGDKGKIVDAVQEHQPSVLLEGAYCGYSAVRMARLLSPGARLITIEINPD
CAATQRMVFAGVKDKVTLVVGASQDIPQLKKYDVDTLDMVFLDHWKDRYLPDTLLEECGLLRKGTVLLADNVICPGAPDFLAHYRGSSCFETHYQSFLEYREWVDGLEKANKGPGSEAGP █

Entry ID Sequence Type ? E-Value Cutoff ? Identity Cutoff % (Integer only) ? Count Clear

Sequence Motif ?

Structure Similarity ?

Structure Motif ?

Chemical Similarity ?

Return ? grouped by ? ↑

Include Computed Structure Models (CSM) Count Clear Search

2.7. Análise dos resultados.

Pergunta: Houve diferença nos resultados obtidos via PDB ou Uniprot???

Passo 2: Seleção do(s) Molde(s):

A estrutura de referência deve ser selecionada a partir do alinhamento realizado nos passos anteriores usando o banco de dados PDB (<https://www.rcsb.org/>). Os principais pontos devem ser considerados para escolha de um molde para modelagem comparativa:

- I. Mesma função que a proteína alvo;
- II. Pelo menos 25-30% de identidade com a sequência alvo;
- III. Menor quantidade *gaps*;

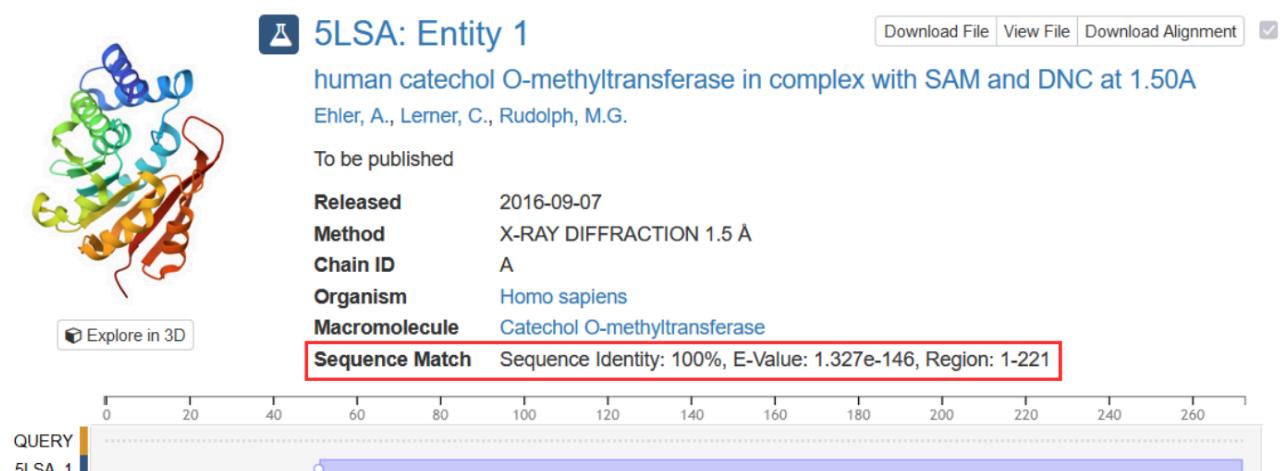
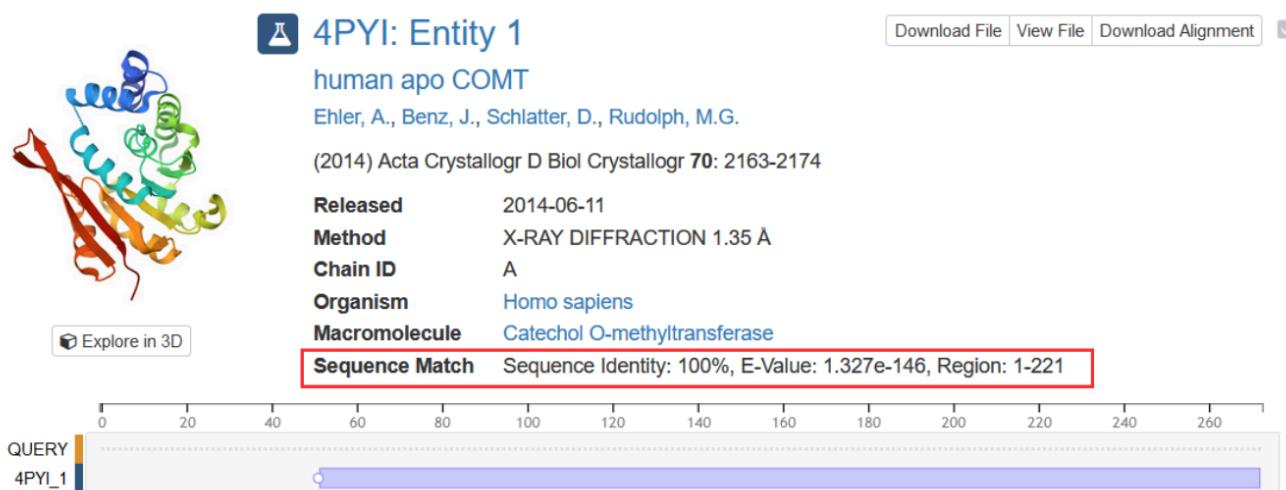
IV. E-value o mais próximo de zero;

V. Observar a presença de ligantes e cofatores;

VI. Analisar se a sequência possui mutações;

VII. Melhor valor de resolução obtido pela técnica experimental (quanto menor, melhor).

VII.I. Geralmente optamos por modelos obtidos pela técnica de difração em raios-X
(X-RAY DIFFRACTION)



Nota: devido às atualizações do banco de dados PDB com o depósito de novas estruturas, talvez as estruturas mostradas acima não estejam entre as primeiras. Porém para fins didáticos iremos mantê-las. Para selecioná-la basta procurá-la na lista gerada pelo site.

Após a seleção das moléculas de referência no PDB, usando os pontos destacados acima, deve-se realizar o download da estrutura 3D da mesma:

1. clique sobre o nome ou imagem da estrutura escolhida;
2. clique em “Download Files” e selecione “PDB Format”;
3. salve o arquivo “.pdb” na pasta “~Minicursos/Predicao de estrutura 3D de proteinas por modelagem comparativa/dia1/modeller/sequences” que se encontra no diretório do presente tutorial.

Structure Summary Structure Annotations Experiment Sequence Genome Ligands Versions

Biological Assembly 1

5LSA
human catechol O-methyltransferase in complex with S-adenosyl methionine

PDB DOI: <https://doi.org/10.2210/pdb5LSA/pdb>

Classification: TRANSFERASE
Organism(s): Homo sapiens
Expression System: Escherichia coli BL21(DE3)
Mutation(s): No

Deposited: 2016-08-24 Released: 2016-09-07
Deposition Author(s): Ehler, A., Lerner, C., Rudolph, M.G.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 1.50 Å
R-Value Free: 0.210
R-Value Work: 0.167
R-Value Observed: 0.169

WWPDB Validation

Validation Metric	Value
Clarity	0.21%
Ramachandran	5.0%
Sidechain	0.0%
RSR2	2.2%
Biological Assembly 1 (CIF - gz)	0.5%
Biological Assembly 1 (PDB - gz)	0.5%

Ligand Str

Display Files Download Files Data API

FASTA Sequence

PDBx/mmCIF Format
PDBx/mmCIF Format (gz)
BinaryCIF Format (gz)

PDB Format
PDB Format (gz)

PDBML/XML Format (gz)

Structure Factors (CIF)
Structure Factors (CIF - gz)

Report Full Report

Value
0.21% 5.0% 0.0% 2.2% 0.5% Better

Explore in 3D: Structure | Sequence Annotations | Electron Density | Validation Report | Ligand Interaction (SAM)
Global Symmetry: Asymmetric - C1
Global Stoichiometry: Monomer - A1
Find Similar Assemblies
Biological assembly 1 assigned by authors and

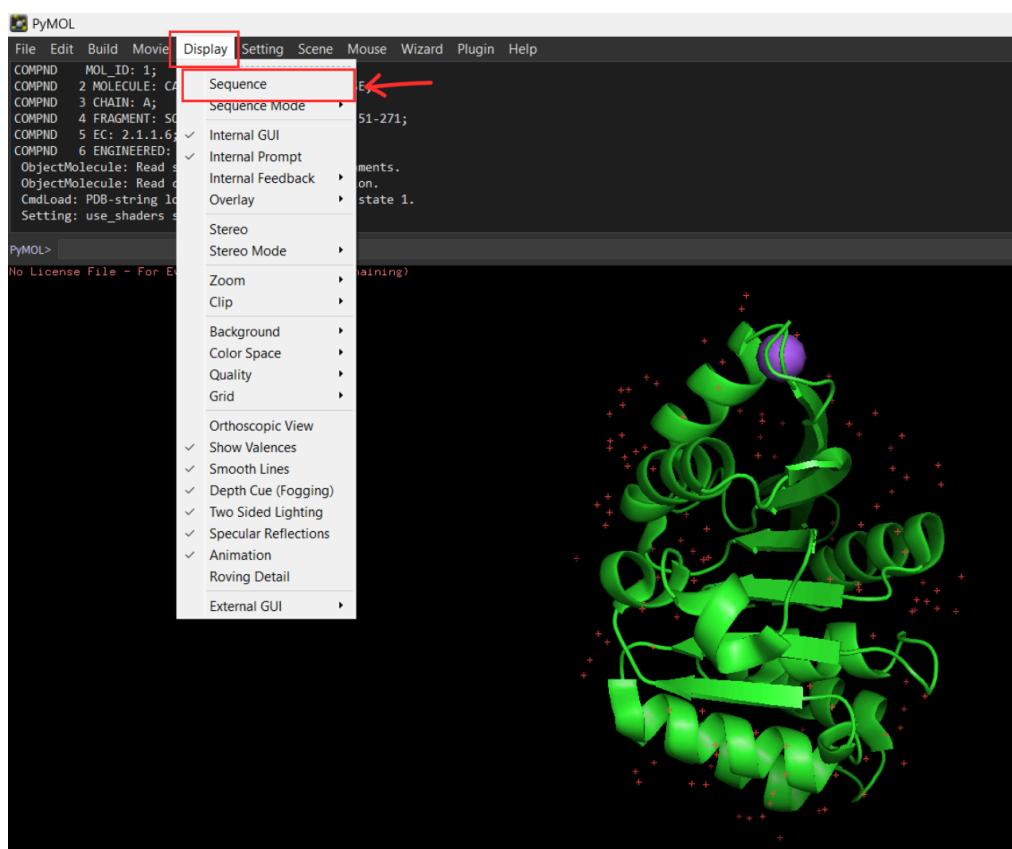
Passo 3: Alinhamento Global entre Sequências: Alvo X molde(s):

Para realizar o alinhamento global entre duas sequências iremos usar o programa Clustal Omega.

1. Note que ambos os templates selecionados possuem água e ligantes em sua estrutura. Visando manter esses elementos presentes no modelo “**5lسا.pdb**”, devemos remover heteroátomos presentes no modelo “**4pyi.pdb**”

1.1. Na pasta “**~Minicursos/Predicao de estrutura 3D de proteinas por modelagem comparativa/dia1/modeller/sequences**”, abra o modelo “**4pyi.pdb**” utilizando o software **Pymol**;

1.2. Na menu “**Display**”, clique em “**Sequence**”, para aparecer a sequência de todos os resíduos presentes na estrutura;



1.3. Após aparecer a sequência, arraste a barra até o final. Você notará a presença de um “**NA**” e uma sequência de **0 (zeros)**;

266 401 406 411 416 421 426 431 436 441 446 451 456 461 466 471 476 481 486 491 496 501 506 511 516 521 526 531

- 1.4. Selecione (clicando com o cursor e arrastando) o “NA” e todos os zeros;

a) O NA é a codificação para o íon de sódio presente na estrutura, enquanto os zeros são a codificação para moléculas de água.

1.5. Irá surgir um novo elemento no índice, recebendo o nome de “**sele**”;



- 1.6. Deve-se clicar com o botão esquerdo em **A** (action), e selecionar a opção “**remove atoms**”;

Dica: também é possível remover todas as águas clicando em “action” (A) na estrutura (4pyi). Ao surgir o menu flutuante, basta clicar em “remove waters”.

- 1.7. Salve a estrutura em **File -> Export molecule**. Nesse momento irá abrir uma nova janela. Não precisa alterar nada. Clique em “**Save...**”;

- 1.8. Altere o formato do arquivo de “.cif” para “.pdb” e salve o arquivo na mesma pasta de antes (`~/Minicursos/Predicao de estrutura 3D de proteinas por modelagem comparativa/dia1/modeller/sequences`), utilizando o mesmo nome da estrutura (4pyi). Ao surgir a janela perguntando se deseja substituir o arquivo, selecione que **sim**.

2. Antes de realizar o alinhamento global da molécula alvo com o molde selecionado, deve-se obter o arquivo .seq. Para isto, vamos rodar o script em Python “readseq.py” e executá-lo usando o programa Modeller. No código do script iremos alterar as seguintes linhas:

- 2.1. entre na pasta “**~Minicursos/Predicao de estrutura 3D de proteinas por modelagem comparativa/dia1/modeller/sequences**”;
- 2.2. abra o arquivo “**readseq.py**” com o editor de texto de sua preferência;
- 2.3. na linha “**code = ”** deve-se informar o caminho, juntamente com o nome do arquivo “**.pdb**” referente ao molde a ser usado durante a modelagem comparativa. Em nosso exemplo iremos usar o molde “**5lسا**” e “**4pyi**”;

```
# -*- coding: utf-8 -*-
# File: readseq.py
# Reading the structure file and passing the sequence to a seg file.

from modeller import *

log.verbose()
env = environ()

#Considering heteroatoms and waters molecules
env.io.hetatm = env.io.water = True
# Directories with input atom files:
env.io.atom_files_directory = './.../atom_files'

#Reading file.pdb and writing file.seq
codes = ['./5lسا', './4pyi']
for code in codes:
    mdl = model(env, file=code)
    aln = alignment(env)
    aln.append_model(mdl, align_codes=code)
    aln.write(file=code+'.seq')
```

2.4. Entre na pasta sequences usando o terminal do Linux ou o prompt de comando no Windows;

2.5. Dentro da pasta, use o comando “mod10.5 readseq.py”

```
You can find many useful example scripts in the
examples\automodel directory.
It is recommended that you use Python to run
Modeller scripts. However, if you don't have Python installed,
you can type 'mod10.5' to run them instead.

C:\Program Files\Modeller10.5>cd C:\Users\mjnla\Desktop\aulas-PSP-2024\dia1\modeller\sequences
C:\Users\mjnla\Desktop\aulas-PSP-2024\dia1\modeller\sequences>mod10.5 readseq.py|
```

2.6. Acesse novamente a pasta “~Minicursos/Predicao de estrutura 3D de proteinas por modelagem comparativa/dia1/modeller/sequences” e veja que um novo arquivo “.seq” foi gerado.

3. Abra o arquivo “.seq” gerado com o editor de texto de sua preferência e copie a primeira sequência de aminoácidos. Cole-a em um novo arquivo texto vazio;
4. Copie a sequência de aminoácidos da molécula alvo e cole-a abaixo da sequência molde dentro do novo arquivo, gerando um arquivo multi-fasta, veja o exemplo abaixo:

```
>alvo
MPEAPPLLLAAVLLGLVLLVVLLLLRHWGWGLCLIGWNEFILQPIHNLLMGDTKEQRIL
NHVLQHAEPGNAQSVDLEAIDTYCEQKEWAMNVGDKKKGKIVDAVIQEHQPSVLLELGAYCG
YSAVRMARLLSPGARLITIEINPDCAAITQRMVDFAGVKDKVTLVVGASQDIIPQLKKY
DVDTLDMVFLDHWKDRYLPDTLLEECGLLRKGTVLLADNVICPGAPDFLAHVRGSSCF
CTHYQSFLEYREVDGLEKAIYKGPGSEAGP

>5lsa
DTKEQRILNHVLQHAEPGNAQSVDLEAIDTYCEQKEWAMNVGDKKKGKIVDAVIQEHQPSVLLELGAYCGYSAVRMA
RLLSPGARLITIEINPDCAAITQRMVDFAGVKDKVTLVVGASQDIIPQLKKYDVDTLDMVFLDHWKDRYLPDTL
LLEECGLLRKGTVLLADNVICPGAPDFLAHVRGSSCFETHYQSFLEYREVDGLEKAIYKGPG

>4pyi
TKEQRILNHVLQHAEPGNAQSVDLEAIDTYCEQKGDKKGKIVDAVIQEHQPSVLLELGAYCGYSAVRMARLLSPGA
RLITIEINPDCAAITQRMVDFAGVKDKVTLVVGASQDIIPQLKKYDVDTLDMVFLDHWKDRYLPDTLLEECGL
LRKGTVLLADNVICPGAPDFLAHVRGSSCFETHYQSFLEYREVDGLEKAIYKGPG
```

5. Salve o atual arquivo com o nome “**multi-fasta.fasta**” dentro da pasta “**~Minicursos/Predicao de estrutura 3D de proteinas por modelagem comparativa/dia1/modeller/sequences**”.

Nota: caso esteja utilizando o Windows, é importante alterar o tipo do arquivo para “Todos os arquivos”, caso contrário, irá salvar no formato “.txt”.

6. Após isso, acesse o site do programa Clustal Omega: <https://www.ebi.ac.uk/jdispatcher/msa/clustalo>, para o alinhamento global.
7. Vá em “upload a file” e clique em “Browse...” e selecione o arquivo multi-fasta gerado no passo anterior.
8. Em “Output Format” selecione a saída no formato Pearson/FASTA;
9. Clique em “Submit”;
10. Após a execução, em “Tool output” clique em “Download” e salve o alinhamento na pasta “**~Minicursos/Predicao de estrutura 3D de proteinas por modelagem comparativa/dia1/modeller/alignment/clustalo**” dentro do diretório do tutorial com o nome “**alvo.ali**”;

```

>alvo
MPEAPLLLAAVLLGLVLLVVLLLLRHWGWLCLIGWNEFILQPIHNLLMGDTKEQRIL
NHVLQHAEPGNAQSVDLEAIDTYCEQKEWAMNVGDKKKGKIVDAVIQEHQPSVLLELGAYCG
YSAVRMARLLSPGARLITIEINPDCAAITQRMVDFAGVKDKVTLVVGASQDIIPQLKKY
DVDTLDLDMVFLDHWDKDRYLPDTLLLEECGLLRKGTVLLADNVICPGAPDFLAHVRGSSCFE
CTHYQSFLEYREVVDGLEKAIYKGPGSEAGP

>5lsa
-----DTKEQRIL
NHVLQHAEPGNAQSVDLEAIDTYCEQKEWAMNVGDKKKGKIVDAVIQEHQPSVLLELGAYCG
YSAVRMARLLSPGARLITIEINPDCAAITQRMVDFAGVKDKVTLVVGASQDIIPQLKKY
DVDTLDLDMVFLDHWDKDRYLPDTLLLEECGLLRKGTVLLADNVICPGAPDFLAHVRGSSCFE
CTHYQSFLEYREVVDGLEKAIYKGPG-----
```



```

>4pyi
-----TKEQRIL
NHVLQHAEPGNAQSVDLEAIDTYCEQK-----GDKKGKIVDAVIQEHQPSVLLELGAYCG
YSAVRMARLLSPGARLITIEINPDCAAITQRMVDFAGVKDKVTLVVGASQDIIPQLKKY
DVDTLDLDMVFLDHWDKDRYLPDTLLLEECGLLRKGTVLLADNVICPGAPDFLAHVRGSSCFE
CTHYQSFLEYREVVDGLEKAIYKGPG-----
```

Perceba que há a presença de **gaps** que não foram resolvidos por nenhum dos dois templates escolhidos.

- Spoiler: trabalharemos isso nos próximos dias ;)

Passo 4: Construção do Modelo Tridimensional

O Modeller trabalha basicamente com arquivos de entrada no formato **PIR** e scripts em linguagem **python**. No link <https://salilab.org/modeller/manual/node104.html> pode-se obter uma lista de extensões com as quais o Modeller trabalha. Em geral as extensões não são obrigatórias, apenas muito úteis para informar o tipo de dado armazenado no arquivo.

Exemplo de arquivo .PIR

```
>P1;EXEMPLO
sequence:EXEMPLO::::::::::
MPEAPLLLAAVLLGLVLLVVLLLLRHWGWLCLIGWNEFILQPIHNLLMGDTKEQRIL
NHVLQHAEPGNAQSVDLEAIDTYCEQKEWAMNVGDKKKGKIVDAVIQEHQPSVLLELGAYCG
YSAVRMARLLSPGARLITIEINPDCAAITQRMVDFAGVKDKVTLVVGASQDIIPQLKKY
DVDTLDLDMVFLDHWDKDRYLPDTLLLEECGLLRKGTVLLADNVICPGAPDFLAHVRGSSCFE
CTHYQSFLEYREVVDGLEKAIYKGPGSEAGP*
```

- Quando está “**sequence**”, é referente ao ALVO que desejamos modelar;
- Quando está “**structure**”, é referente ao TEMPLATE obtido no PDB.

Exemplo de arquivo .py usado no modeller

```
from modeller import *
log.verbose()
env = environ()
env.io.het atm = env.io.water = True
code = 'EXEMPLO'
mdl = model(env, file=code)
aln = alignment(env)
aln.append_model(mdl, align_codes=code)
aln.write(file=code+'.seq')
```

Os arquivos necessários para gerar os modelos usando modeller são:

- a) alinhamento no formato “**.PIR**”;
- b) arquivo “**.pdb**” do(s) molde(s);
- c) arquivo “**.fasta**” referente à estrutura alvo.

1. O primeiro passo para gerar os modelos no modeller será editar o arquivo de alinhamento global gerado pelo clustal omega nos passos anteriores. Para isto deve-se seguir as seguintes etapas:

1.1. abra os arquivos “**.ali**” gerado pelo clustalW e os arquivos “**.seq**” gerado anteriormente;

1.2. No arquivo de alinhamento, deve-se substituir o cabeçalho referente a sequência molde pelo cabeçalho presente no arquivo “**.seq**”;

1.3. Ainda dentro do arquivo “**.ali**”, substitua o cabeçalho da sequência alvo por:

```
>P1;ALVO
sequence:ALVO: FIRST :A:LAST :A:::0.00:0.00
```

Nota: O nome “ALVO” deve ser substituído pelo nome do arquivo “.fasta” da molécula alvo !

1.4. Transfira todos os heteroátomos presentes no template “5lسا.seq” para o arquivo “.ali”. Os heteroátomos deverão ficar depois da sequência de resíduos de aminoácidos da própria 5lسا e também após a sequência alvo. Irá ficar da seguinte forma:

Nota: não se esqueçam de conferir se todas as estruturas presentes estão no formato PIR

Após as alterações salve o alinhamento.

2. Copie os arquivos “.ali”, “.pdb” do molde e o “.fasta” do alvo para dentro da pasta “~/Minicursos/Predicao de estrutura 3D de proteinas por modelagem comparativa/dia1/modeller/single-init”;
 3. Dentro da pasta “single-init” abra o script “**model-single.py**” com o editor de texto (ex. kate);
 4. Na função “**automodel**” deve-se mudar os seguintes parâmetros:
 - a) **alnfile**: deve ser indicado o endereço e nome do arquivo de alinhamento **.ali** entre aspas;
 - b) **knowns**: refere-se ao endereço e nome das estruturas usadas como molde (arquivo **.pdb**) para a modelagem. Assim como em “alnfile”, o nome dos arquivos também necessita estar entre aspas;

c) **sequence:** é o nome do “.fasta” da sequência alvo. Também deve estar entre aspas.

```
# MODEL CONSTRUCTION
# Modelling 'sequence' with file.ali
a = automodel(env, alnfile='alvo.ali',
              knowns = ('./51sa', './4pyi'),
              sequence='alvo',
              assess_methods=(assess.DOPE, assess.normalized_dope, assess.GA341)
              )

# Generating 5 models
a.starting_model = 1
a.ending_model = 5
a.make()

# Get clusters
a.cluster(cluster_cut=1.00)
# END OF MODEL CONSTRUCTION
```

5. Para gerar os modelos deve-se entrar dentro da pasta “**~Minicursos/Predicao de estrutura 3D de proteínas por modelagem comparativa/dia1/modeller/single-init**” via **terminal ou prompt de comando**. Após isso digite: “**mod10.5 model-single.py**”. O script irá gerar **5 modelos** para a proteína alvo através da função **automodel()**, um arquivo “**.log**” com informações de energia dos modelos gerados (molpdf, DOPE e GA341).

Gerando os modelos otimizados da proteína alvo com o Modeller:

Nesta etapa iremos gerar modelos variando os parâmetros das funções de otimização do modeller.

1. Para isto, dentro da pasta “**~Minicursos/Predicao de estrutura 3D de proteínas por modelagem comparativa/dia1/modeller/single-opt**” abra o script “**model-single-opt.py**”;
2. As novas linhas que foram inseridas referem-se aos diferentes graus de otimização:
 - a) “**autosched.slow()**” = refere-se a otimização lenta executada pelo método VTFM (*Variable Target Function Method*);
 - b) “**max_var_iterations**” = número máximos de iterações do algoritmo de gradiente conjugado;
 - c) “**refine.slow()**” = refere-se a forma de otimização realizada usando dinâmica molecular

(MD) com *simulated annealing*;

- Aceita os valores de refine.**very_fast**, refine.**fast**, refine.**slow**, refine.**very_slow** ou refine.**slow_large**
- **Quanto mais lento for, maior será o nível de otimização dos modelos. Como estamos trabalhando com um modelo didático, podemos deixar esse parâmetro em “refine.slow”**

- d) “**repeat_optimization**” = número de repetições do passo de otimização do modelo;
- e) “**max_molpdf**” = valor máximo da função de densidade e probabilidade calculado a partir das restrições espaciais e distâncias dos átomos;

3. De forma similar ao passo sem otimização (single-init), acesse a pasta “**single-opt**” e rode o comando: mod10.5 “**model-single-opt.py**” para gerar os modelos com otimização.

Passo 5: Análise inicial dos modelos gerados

Neste momento, iremos utilizar o software **Pymol** para realizar uma inspeção inicial acerca dos modelos obtidos. Para isso, utilizaremos 2 parâmetros importantes: o **alinhamento estrutural** e o **RMSD**.

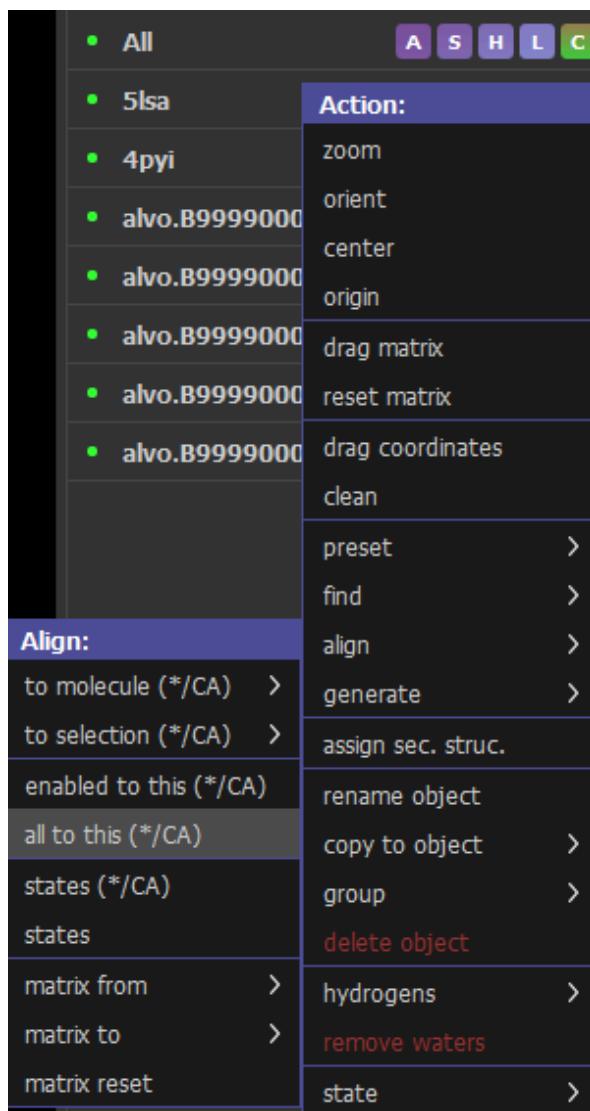
Alinhamento estrutural

Para realizarmos o alinhamento estrutural, iremos abrir todos os modelos gerados utilizando o Pymol.

1. Abra inicialmente o template 5lسا.pdb utilizando o Pymol;
2. Em File -> Open..., selecione o template 4pyi.pdb, e todos os 5 modelos otimizados gerados;
3. O menu irá ficar da seguinte forma:

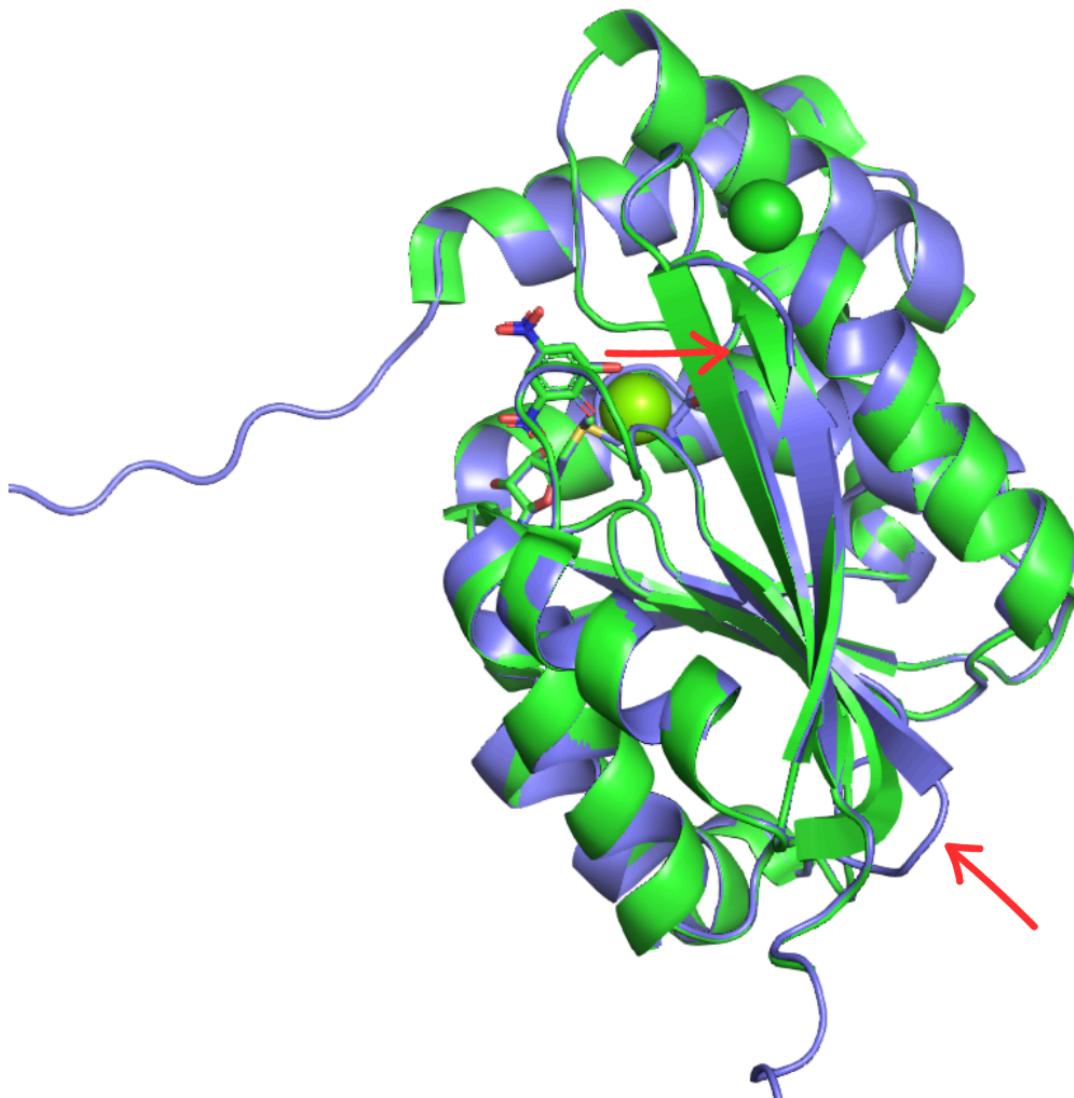
• All	A S H L C
• 5lسا	A S H L C
• 4pyi	A S H L C
• alvo.B99990001	A S H L C
• alvo.B99990002	A S H L C
• alvo.B99990003	A S H L C
• alvo.B99990004	A S H L C
• alvo.B99990005	A S H L C

4. Na estrutura 5lسا, vá em **action (A)**, e selecione “align -> align all to this (*/CA)”;



5. Para facilitar a visualização da estrutura, você pode remover as águas;
6. Realize a inspeção visual dos modelos, tentando perceber diferenças estruturais entre os modelos e os templates;

Dica: você pode clicar no nome de cada uma das estruturas no menu para ocultá-las. Ao clicar novamente elas reaparecem. Teste dessa forma para conseguir analisar uma à uma.



Note que pode ocorrer diferenças estruturais entre o template (verde) e o modelo gerado (azul). Um bom modelo é aquele que está mais semelhante aos templates.*

* Isso não é uma regra para TODOS os casos de modelagem de proteínas, já que às vezes estamos realizando a modelagem justamente para suprir alguma conformação que não está disponível nos modelos experimentais presentes no PDB.

Pergunta: Qual foi o melhor modelo otimizado gerado???

7. Agora repita o processo para os modelos gerados sem otimização, presentes em “~/Minicursos/Predicao de estrutura 3D de proteinas por modelagem comparativa/dia1/modeller/single-init”.

Pergunta: Qual processo gerou modelos de maior qualidade? O com ou sem otimização???

Obtenção do RMSD

O RMSD (**Root Mean Square Deviation**) é um parâmetro utilizado para medir a diferença média entre as posições de átomos correspondentes em duas estruturas proteicas.

* Raiz quadrada da média dos quadrados das diferenças (desvios) entre as posições dos átomos equivalentes de duas estruturas. Ele fornece uma medida quantitativa de quão similar ou diferente duas estruturas são.

O RMSD é amplamente utilizado para comparar estruturas de proteínas, seja para avaliar a qualidade de um modelo comparado a uma estrutura de referência conhecida (o nosso template) ou para comparar duas conformações diferentes de uma proteína.

Quando você realiza o alinhamento de duas proteínas, o objetivo é sobrepor as duas estruturas da melhor forma possível para minimizar as diferenças entre elas. O RMSD é então calculado para quantificar a magnitude dessas diferenças. Um RMSD baixo indica que as duas estruturas são muito semelhantes, enquanto um RMSD alto sugere que há grandes diferenças entre as duas estruturas. O RMSD mede a distância média entre pares de átomos correspondentes (geralmente os átomos C-alfa ou os átomos pesados equivalentes) após o alinhamento estrutural. Ele é dado em unidades de distância, tipicamente angstroms (Å).

Matematicamente, o RMSD é definido como:

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N |r_i^A - r_i^B|^2}$$

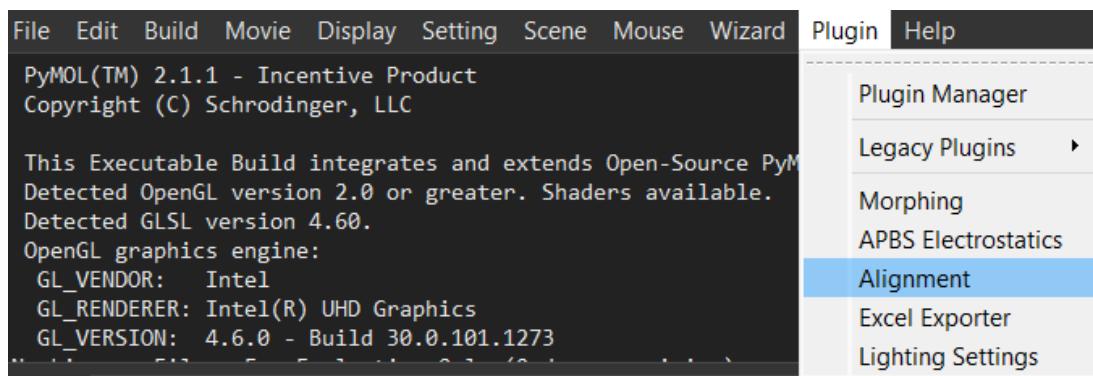
Onde:

- N é o número de pares de átomos correspondentes nas duas estruturas.
- r_i^A e r_i^B são os vetores de posição (por exemplo, (x_i^A, y_i^A, z_i^A) e (x_i^B, y_i^B, z_i^B) dos átomos correspondentes iii nas duas estruturas).

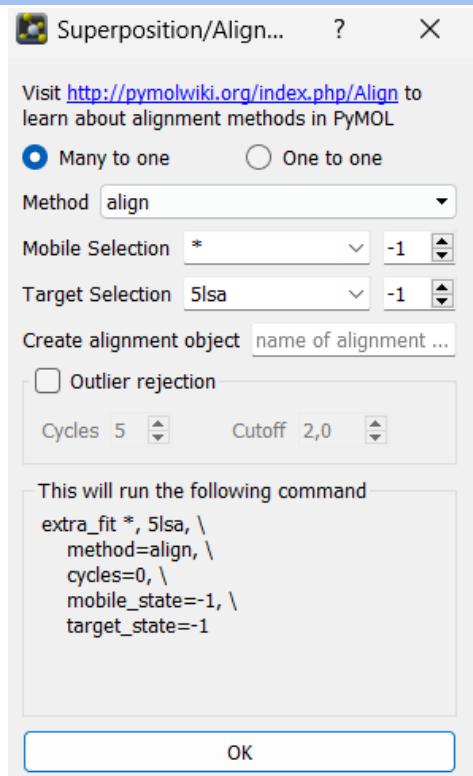
- $|r_i^A - r_i^B|$ representa a distância euclidiana entre os átomos correspondentes i nas duas estruturas.

A fórmula calcula a soma dos quadrados das distâncias entre os átomos correspondentes, divide pelo número de átomos N , e então toma a raiz quadrada desse valor.

1. Para obter o RMSD pelo Pymol, faça o upload de todas as estruturas geradas (com e sem otimização) no software;
2. Abra o template **5lسا.pdb**;
3. Na barra superior, à direita, clique em **Plugin -> Alignment** (algumas versões do PyMol podem estar como “Superposition/Alignment”);



4. Na nova janela que se abriu, deixe as configurações da seguinte forma:
 - a. Many to one;
 - b. Method: align;
 - c. Mobile section: * e -1;
 - d. Target section: 5lسا e -1;
 - e. Desmarque a opção “Outlier rejection”;



5. Aperte OK;
6. Retorne ao PyMol. O RMSD irá se encontrar no menu de comandos;

7. Uma vez obtido o RMSD de todos os modelos gerados com o template **5lسا**, repita o processo utilizando o template **4pyi.pdb**;

8. Complete a tabela abaixo:

Modelos otimizados	RMSD com 5LSA	RMSD com 4PYI	Modelos não otimizados	RMSD com 5LSA	RMSD com 4PYI
1			1		
2			2		
3			3		
4			4		
5			5		

Pergunta: Qual modelo gerou o menor RMSD? Ele foi com ou sem otimização??



VOCÊ CONCLUIU O PRIMEIRO DIA!



Dia 2: Predições estruturais

Predições Estruturais Importantes:

Antes de modelar uma proteína, é fundamental realizar predições estruturais para se obter informações importantes sobre sua estrutura e função. Essas predições incluem peptídeo sinal, estrutura secundária, regiões transmembrana e regiões de contato entre aminoácidos.

Essas análises preliminares ajudam a guiar a modelagem, fornecendo uma compreensão melhor da localização e papel de diferentes segmentos da proteína. Elas permitem prever domínios funcionais, interações com a membrana e áreas de contato importantes, resultando em modelos mais precisos e informativos.

Passo 1: Predição de Peptídeo Sinal via SignalP 3.0

O servidor SignalP 3.0 prevê a presença e a localização dos locais de clivagem do peptídeo sinal em sequências de aminoácidos de diferentes organismos: procariotas Gram-positivos, procariotas Gram-negativos e eucariotas. O método incorpora uma previsão dos locais de clivagem e uma previsão do péptido sinal/peptídeo não sinal com base numa combinação de várias redes neurais artificiais e modelos de Markov ocultos.

1. No site do SignalP 3.0 (<https://services.healthtech.dtu.dk/services/SignalP-3.0/>) em “Paste a single sequence or several sequences in FASTA format into the field below” você deverá inserir sua sequência alvo no **formato FASTA**.
2. Devemos alterar os parâmetros
 - a. Organism group: Eukaryotes;
 - b. Method: both;
 - c. “Graphics”: “GIF (inline) and EPS (as links)”;
 - d. Output format: standard;
 - e. Truncation: 0 (zero).

Submission

Sequence submission: paste the sequence(s) and/or upload a local file

Paste a single sequence or several sequences in [FASTA](#) format into the field below:

```
>alvo
MPEAPPPLLAVALLGLVLLVVLRLRHGWGLCLIGWNEFILOPIHNLLMGDTKEQRL
NHVLOHAEPGNAOSVLFAIDTYCEOKEWAMNIVDKKGKIVDAVTOEHOPSVLLELGAYCG
```

Submit a file in [FASTA](#) format directly from your local disk:

Procurar... Nenhum arquivo selecionado.

Organism group

- Eukaryotes
- Gram-negative bacteria
- Gram-positive bacteria

Method

- Neural networks
- Hidden Markov models
- Both

Graphics

- No graphics
- GIF** (inline)
- GIF** (inline) and **EPS** (as links)

Output format

- Standard
- Full
- Short (no graphics!)

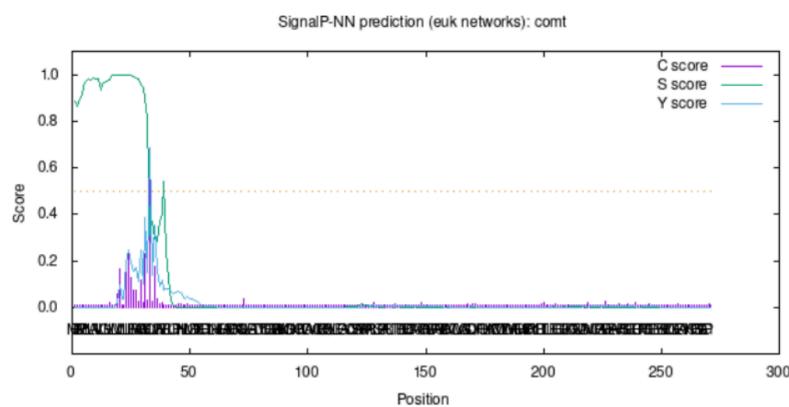
Truncation

Truncate each sequence to max. residues.

We recommend that only the N-terminal part of each protein sequence is submitted. Enter 0 (zero) to disable truncation.

[Submit](#) [Clear fields](#)

O resultado será apresentado dessa forma:



```
# data
# plot
in EPS format
```

```
>comt                               length = 271
# Measure  Position  Value  Cutoff  signal peptide?
max. C      7    0.625  0.32   YES
max. Y      2    0.687  0.33   YES
max. S      2    0.998  0.87   YES
mean S     1    0.961  0.48   YES
D      1-2    0.824  0.43   YES
# Most likely cleavage site between pos. 3 : GWG-LC
```

3. Salve os resultados na pasta “~/Minicursos/Predicao de estrutura 3D de proteinas por modelagem comparativa/dia2/SignalP”. Iremos utilizá-los em outro momento.

Pergunta: A proteína apresenta clivagem de peptídeo sinal? Onde essa clivagem ocorre?

Passo 2: Predição de estrutura secundária

A predição de estrutura secundária é um método utilizado para identificar as regiões de uma proteína que formarão alfa hélices, folhas beta ou coils/loops. Essa análise ajuda a entender a organização básica da proteína e fornece informações essenciais para construir modelos tridimensionais mais precisos. Além disso, a predição de estrutura secundária facilita a identificação de domínios funcionais e possíveis sites de interação, contribuindo para a compreensão da função proteica.

Para isso, iremos construir uma tabela no Excel de forma que facilite nossa visualização.

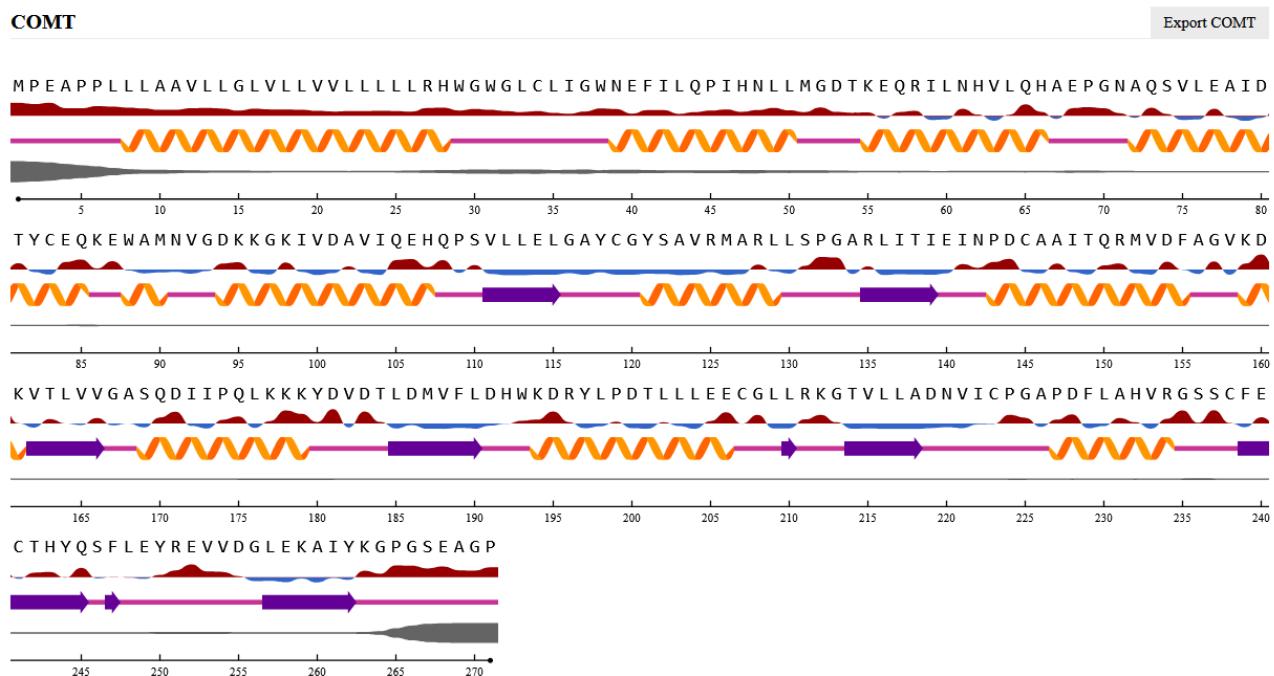
- Iremos trabalhar na tabela presente em:

“~/Minicursos/Predicao de estrutura 3D de proteinas por modelagem comparativa/dia2/predEstSec/Pred_Est_Sec.xlsx”;

Os resultados dos preditores sugerem qual conformação uma região assume, que pode ser: Hélice, Coil e Fita. Sugestão de cor:											
Hélice (H)											
Fita (E, S)											
Coil (C, R) - sem preenchimento	Resíduo	1	2	3	4	5	6	7	8	9	10
	Resíduo	M	P	E	A	P	P	L	L	L	A
http://www.compbio.dundee.ac.uk/jpred4/index.html	Jpred	Conformação									
http://bioinf.cs.ucl.ac.uk/psipred/	PSIPRED										
https://services.healthtech.dtu.dk/services/NetSurfP-3.0/	NetSurfP										
https://zhanggroup.org/PSSpred/	PSSpred										

- Na linha 5 está presente a numeração de cada um dos resíduos de nossa estrutura;
 - Na linha 6 a sequência de aminoácidos da nossa estrutura;
 - As linhas seguinte são os preditores de estrutura secundária que iremos utilizar;
 - É importante utilizar mais de um preditor de estrutura secundária para que possamos avaliar o consenso entre cada um deles. Neste exemplo iremos utilizar apenas cinco, mas geralmente utilizamos mais.
- Abra cada um dos preditores disponíveis em:
 - Jpred: <http://www.compbio.dundee.ac.uk/jpred4/index.html>
 - PSIPRED: <http://bioinf.cs.ucl.ac.uk/psipred/>
 - NetSurfP - 3.0: <https://services.healthtech.dtu.dk/services/NetSurfP-3.0/>
 - PSSpred: <https://zhanggroup.org/PSSpred/>

1. Cole sua sequência alvo no formato FASTA;
2. Clique em Submit;
3. O resultado irá aparecer dessa forma:



4. Os resíduos que estão com a alfa hélice (em amarelo) sobre ele, significa que eles possuem maior probabilidade de formar alfas hélices; Os resíduos com um coil (em rosa), formará coils; Os resíduos com uma fita beta (seta roxa), formará folhas beta;
5. Em “Export”, você pode utilizar o “JSON format”, que irá facilitar sua utilização no Excel, pois irá mostrar o resultado na forma de C (para coils), E (para folhas beta) e H (para alfa hélices)
6. Em “JSON”, oculte a linha “q8prob”
7. Irá aparecer a linha “q3”. Nela está presente seus resultados referentes a predição de estrutura secundária;

JSON	Dados brutos	Cabeçalhos
Salvar	Copiar	Recolher tudo
Expandir tudo	Filtrar JSON	
<pre> ▼ q8: "CCCCCCCCHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCHHHHHHHHHHHHHHCCCCHHHHHHHHHHHHCCCCHHHHHHHHCCCCHHHHHHHSCGGGGCCH ► q8_prob: [...] ▼ q3: "CCCCCCCCHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCHHHHHHHHHHHHHHCCCCHHHHHHHHHHHHCCCCHHHHHHHHCCCCHHHHHHHCCCTCHHHHHHHHHHHHHHHSCGGGGCCH ▼ q3_prob: </pre>		

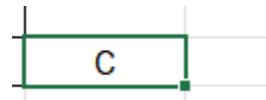
8. Copie a sequência e cole em uma célula qualquer do Excel, desde que seja fora da tabela de resultados;
9. Para este exemplo, eu coloquei o resultado na célula B25:

	A	B
23		
24		CCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCC CCHHHHHHHHHHHHHHHCCCCHHHHHHHHHHHHCCCC HHHHHHHHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHH CCCEEEEEECCCCCHHHHHHHHHCCCCCEECCCHHHH HHHHHHHHHHCCCCHHHEEEECCHHHHHHHHHHHHHCCCC CEEEEEECCHHHHHHHHHCCCCCEECCCEEEEEECCCC CCCCHHHHHHHHCCCCCEECCCECCCCCCCCCCCC CCCCCCCC
25		

10. Na célula referente aos resultados do TMHMM, utilize a seguinte fórmula

=EXT.TEXTO(\$B\$25;COL(A1);1)

- Na primeira variável deve-se colocar a célula em que estão presentes os dados. É importante colocar “\$” antes da linha e da coluna para travá-la. Como meus dados estão presentes na célula B25, a minha fórmula ficou dessa maneira, mas você pode adaptá-la para se adequar ao seu caso;
- Na segunda variável, quando você usa a função COL(A1) dentro de uma fórmula, ela retorna o número da coluna em que a célula A1 está. Neste caso, COL(A1) retorna 1, porque A é a primeira coluna;
- Na terceira variável utilizamos o número 1 (**um**), para informar que o texto presente na célula B25 será dividido em 1 caractere por célula;
- Depois disso basta clicar no quadradinho presente no canto inferior direito, e arrastar a fórmula por todas as células.



11. O resultado ficará presente dessa forma:

Resíduo	1	2	3	4	5	6	7	8	9	10	11	12
Resíduo	M	P	E	A	P	P	L	L	L	A	A	V
Jpred	Conformação											
PSIPRED												
NetSurfP		C	C	C	C	C	C	H	H	H	H	H
PSSpred												

12. Neste momento, basta utilizar a ferramenta “Cor de preenchimento” para colorir as células com as cores recomendadas, sendo amarelo para fitas, rosa para hélices e sem

preenchimento para coils.

PSIPRED

No PSIPRED existem diversas opções de predições as quais o site consegue nos fornecer.

Neste momento iremos nos atentar apenas a predição de estrutura secundária (PSIPRED 4.0)

1. Selecione a opção: **PSIPRED 4.0 (Predict Secondary Structure);**

Choose prediction methods (hover for short description)

Popular Analyses

PSIPRED 4.0 (Predict Secondary Structure) DISOPRED3 (Disored Prediction)
 MEMSAT-SVM (Membrane Helix Prediction) pGenTHREADER (Profile Based Fold Recognition)

.....

2. No final da página, informe sua sequência no formato **FASTA**;
3. Clique em **Submit**;
4. O resultado será exibido dessa forma:

The screenshot shows the PSIPRED 4.0 results page. On the left, there is a "Sequence Plot" showing a protein sequence from position 1 to 271. The plot uses color-coded boxes to represent different secondary structure predictions: yellow for Strands, pink for Helices, grey for Coils, light blue for Putative Domain Boundaries, dark grey for Membrane Interaction, green for Re-entrant Helices, and red for Signal Peptides. Below the plot is a legend with the following categories and their corresponding colors:

- Strand (Yellow)
- Helix (Pink)
- Coil (Grey)
- Putative Domain Boundary (Light Blue)
- Membrane Interaction (Dark Grey)
- Re-entrant Helix (Green)
- Disordered (Light Blue Box)
- Transmembrane Helix (Dark Grey Box)
- Cytoplasmic (Light Blue Box)
- Signal Peptide (Red Box)
- Get PNG (Button)
- Get SVG (Button)

On the right side of the page, there is a sidebar titled "Downloads" which includes links to "RESULTS ZIP FILE" and "JOB CONFIGURATION". Below that is a section titled "PSIPRED DOWNLOADS" with links to "Horiz Format Output" and "SS2 Format Output". At the bottom of the sidebar is a section titled "Segment Resubmission".

5. À direita, em “Downloads”, baixe o “Horizon Format Output”;

```
# PSIPRED HFORMAT (PSIPRED V4.0)

Conf: 99992799999999999999999999999999999979974799998614995189999
Pred: CCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCHHHHH
AA: MPEAPPLLAVALGLVLLVVLLLRLHWGWLCLIGWNEFILQPIHNLLMGDTKEQRIL
      10     20     30     40     50     60

Conf: 999993699999999999999535977446899999999999979989999799989
Pred: HHHHHHHCCCCCHHHHHHHHHHHHHCCCCCCCCCHHHHHHHHHHHHHCCCEEEECCCC
AA: NHVLQHAEPGNAQSVLLEAIDTYCBEQKEWAMNVGDKGKIVDAVIQEHQPSVLLLGAYCG
      70     80     90    100    110    120

Conf: 9999996289998999869999999990987658999588899996732
Pred: HHHHHHHCCCCCEEEEECCHHHHHHHHHHHHHCCCCCEEEBECCHHHHHHHHHHC
AA: YSAVRMARLLSPGARLITIEINPDCAAITQRMVDFAGVKDKVTLVVGASQDIIPQLKKY
      130    140    150    160    170    180

Conf: 988669999589134379999982998688699907889998899997199948
Pred: CCCCEEEEEEECCHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCHHHHHHHHHCCCCCEE
AA: DVDTLDMVFLDHWDYLPDTLLLEECGLLRKGTVLLADNVICPGAPDFLAHVRGSSCFE
      190    200    210    220    230    240

Conf: 8998415663402784689997499988999
Pred: EEEEBCEEEECCCCCEEEBCCCCCCCC
AA: CTHYQSFLEYREVVDGLEKAIYKGPGEAGP
      250    260    270
```

6. A linha “Conf” diz respeito a confiabilidade do resultado, a linha “Pred” é a predição de estrutura secundária e a linha “AA” é a sequência de resíduos de aminoácidos. Copie as linhas referentes à predição de estrutura secundária, transfira-a para a planilha e faça a coloração correspondente.

JPRED

1. Em “Input sequence”, coloque a sequência de aminoácidos da proteína no formato FASTA;
2. Clique em “Make prediction”;
3. O resultado será fornecido dessa forma:

- View results summary in SVG - displayed below (details on acronyms used):



- View full results in HTML
 - View simple results in HTML
 - View results in PDF
 - View results in Jalview (Link to a separate page with the Jalview Java Desktop application)
 - View everything in a results directory (details on data each file contains are available through README file)
 - Get all (but PS) files in TAR.GZ archive

4. Clique em “View simple results in HTML”;
 5. Na nova janela que se abre, a primeira linha é referente a sequência de resíduos de aminoácidos, enquanto a segunda é a predição de conformação de estrutura secundária. Copie e cole a segunda linha no Excel e faça as colorações necessárias.

Nota: neste preditor, os coils são representados por traços (-). Não é necessário alterar.

PSSpred

1. Utilize sua sequência alvo no formato FASTA;
 2. Insira seu e-mail no campo seguinte (é obrigatório para o funcionamento do servidor, mas os resultados serão mostrados no próprio navegador);
 3. Clique em Run PSSpred;
 4. Os resultados obtidos serão representados dessa forma:

```

seq: 1 MPEAPPPLLAALVLLGLVLLVVLLLLRHWGWLCLIGWNEFILQPIHNLLMGDTKEQRIL 60
SS: CCCCCCCHHHHHHHHHHHHHHHHHHHCCCCCEEHHHHHHHHHHHHHHCCCCCHHHH
conf: 99864127899999999999999962354200012211233445443124665079999

seq: 61 NHVLQHAEPGNAQSYLEAIDTYCEQKEWAMNVGDKKGKIVDAVIQEHQPSVLLELGAYCG 120
SS: HHHHHHCCCCCHHHHHHHHHHHHHHHCCCCCCCCHHHHHHHHHHHHCCEEEECCCCCH
conf: 9999855899989999999999727977667879999999999979987999806740

seq: 121 YSAVRMARLLSPGARLITIEINPDCAAITQRMVDFAGVKDKVTLVGASQDIIPQLKKY 180
SS: HHHHHHHHCCCCCEECCCCCHHHHHHHHHHHHHHHCCCCCEECCCCCHHHHHHHHHHC
conf: 8999999589997899997493799999999983966627999847588888877536

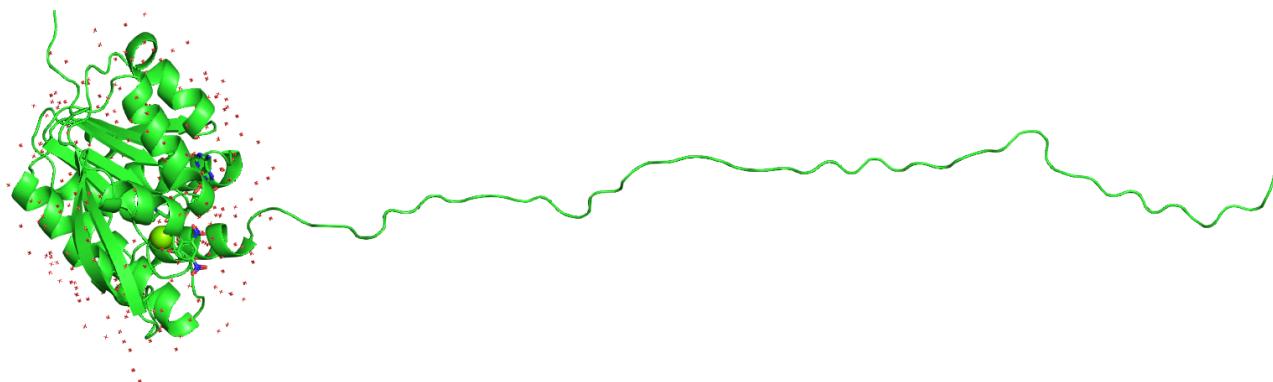
seq: 181 DVDTLDMVFLDHWKDRYLPDTLLEECGLLRKGTVLLADNVICPGAPDFLAHVRGSSCFE 240
SS: CCCCEECCCCCCCCCHHHHHHHCCCCCEECCCCCCCCCHHHHHCCCCCEE
conf: 776336999817831056767999996768998799992667899951578863277502

seq: 241 CTHYQSFLEYREVVDGLEKAIYKGPGEAGP 271
SS: EEEHHHHHHHCCCCCEECCCCCCCC
conf: 1014555541555551688887168988999

```

5. Assim como no PSIPRED, copie apenas a linha “SS”. Cole-a no Excel e faça as colorações necessárias.

Pergunta: A partir das previsões de estrutura secundária, a proteína continuará igual aos modelos gerados no dia anterior, ou será necessário realizar modificações em sua estrutura?



Passo 3: Predição de regiões transmembrana

A predição de regiões transmembrana é o processo de identificar os segmentos de uma proteína que atravessam a membrana celular. Essas regiões geralmente consistem em **alfa-hélices** hidrofóbicas que se inserem na bicamada lipídica da membrana. A identificação dessas regiões é crucial, pois elas desempenham um papel fundamental na função das proteínas de membrana, como receptores, canais iônicos e transportadores.

Na modelagem computacional de proteínas, a predição de regiões transmembrana é importante para construir modelos mais precisos. Saber se a proteína possui, e onde essas regiões estão localizadas permite simular adequadamente a interação da proteína com a membrana, influenciando a estabilidade e a funcionalidade do modelo. Além disso, as características hidrofóbicas e a orientação das hélices transmembrana são essenciais para entender como a proteína se insere e se comporta na membrana, afetando diretamente a interpretação de experimentos e o desenvolvimento de fármacos.

TMHMM

1. Acesse o site do TMHMM (<https://services.healthtech.dtu.dk/services/TMHMM-2.0/>)
2. Insira sua sequência FASTA no campo disponível;
3. Output format: Extensive, with graphics;
4. Submeta para o servidor e analise os resultados gerados.

UniTmp

Outra ferramenta importante que pode ser utilizada é o Human Transmembrane Proteome (<https://htp.unitmp.org/>). A partir dele, é possível utilizar a função UNIfied database of TransMembrane Proteins (<https://www.unitmp.org/>), que fornece dados já existentes acerca de proteínas de *Homo sapiens*.

1. Abra o site do UNIfied database of TransMembrane Proteins;
2. No campo disponível, digite o nome da sua proteína (COMT);
3. Dos resultados obtidos, clique em See all of them, na database “HTP”;
4. Clique em “Details” do primeiro resultado:

The screenshot shows a search interface for the UniTmp database. At the top, there is a search bar with the text "Search results for comt". Below the search bar, there is a small icon of a yellow stick figure. The main results area displays information for the protein "COMT_HUMAN". The protein name is listed as "COMT_HUMAN" followed by its full name, "Catechol O-methyltransferase". It indicates "#TM: 1" and "Evidence level: 3D". To the right of this information is a blue rectangular button with the text "Details" and a small arrow icon. On the left side of the results area, there is a large, empty white box with a thin black border.

5. Na aba “1D” terá disponível a sequência de aminoácidos e sua localização (intracelular, transmembrana ou extracelular);
6. Analise os resultados.

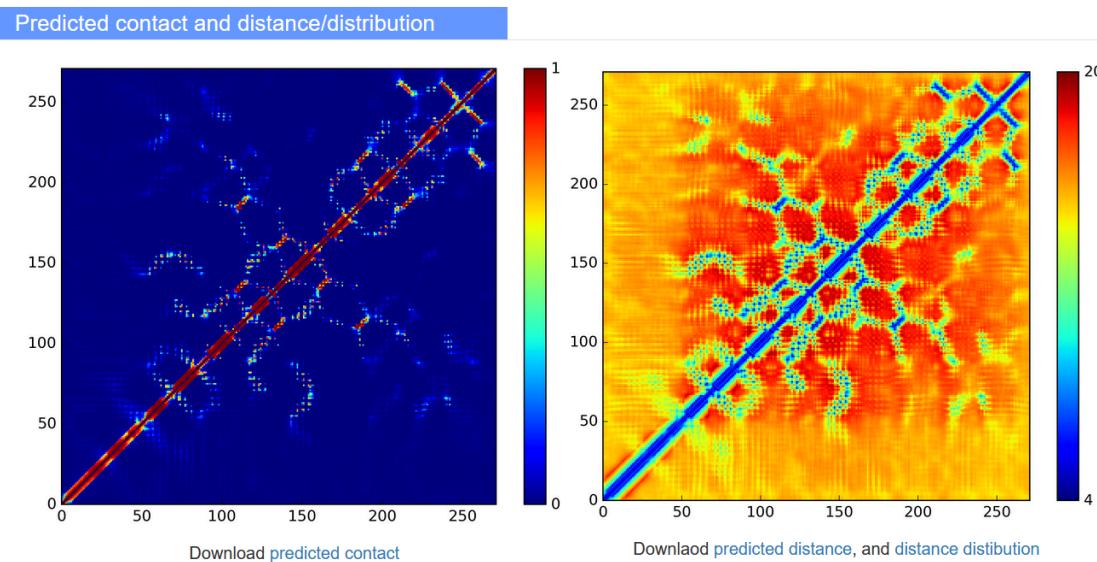
Pergunta: Houve distinção nos dois resultados obtidos? A proteína apresenta regiões transmembrana?

Passo 4: Predições de contato

Agora que identificamos a presença de uma alfa-hélice além da região transmembrana, é necessário avaliar e predizer sua posição tridimensional. Para isso, podemos utilizar preditores de contato, que permitirão identificar se os resíduos da alfa-hélice interagem com a porção globular da proteína.

MapPred

1. Acesse o site do MapPred (<https://yanglab.qd.sdu.edu.cn/MapPred/>);
2. Insira sua sequência no formato FASTA;
3. Caso preferir, pode fornecer seu email e um nome de identificação para sua predição;
4. Clique em “Submit”
5. O resultado será apresentado dessa forma:



6. No gráfico à esquerda temos a predição de contato, e no gráfico a direita uma predição de distância entre os resíduos;
 7. Note que o resultado na parte superior também inclui predição de estrutura secundária.
Insira esses resultados na tabela de estrutura secundária.
 8. Clique em “Download predicted contact” e “predicted distance”;
 9. Copie o resultado e cole em uma planilha do Excel;
 10. No Excel, selecione a primeira coluna;
 11. Na aba “Dados”, utilize a ferramenta “Texto para colunas” para criar uma tabela com os resultados;

A1

	A	B
1	1 2 0 8 0 9	0.986532
2	1 3 0 8 0 7	0.711588
3	1 4 0 8 0 2	0.212281
4	1 5 0 8 0 1	0.184883
5	1 6 0 8 0 1	0.109699
6	1 7 0 8 0 0	0.71953
7	1 8 0 8 0 0	0.49533
8	1 9 0 8 0 0	0.35058
9	1 10 0 8 0 0	0.029754
10	1 11 0 8 0 0	0.023229
11	1 12 0 8 0 0	0.014322
12	1 13 0 8 0 0	0.011804
13	1 14 0 8 0 0	0.009890
14	1 15 0 8 0 0	0.0063027

12. Na janela que se abrir, selecione “Delimitado”, para definir quais elementos ficarão em cada coluna;

Tipo de dados originais

Escolha o tipo de campo que melhor descreva seus dados:

Delimitado - Caracteres como vírgulas ou tabulações separam cada campo.

Largura fixa - Campos são alinhados em colunas com espaços entre cada campo.

13. Clique em “Avançar”;

14. Em “Delimitadores”, selecione a opção “Espaço” e “Considerar delimitadores consecutivos como um só”;

Delimitadores

Tabulação
 Ponto e vírgula
 Vírgula
 Espaço
 Outros:

Considerar delimitadores consecutivos como um só

Qualificador de texto: "

Visualização dos dados

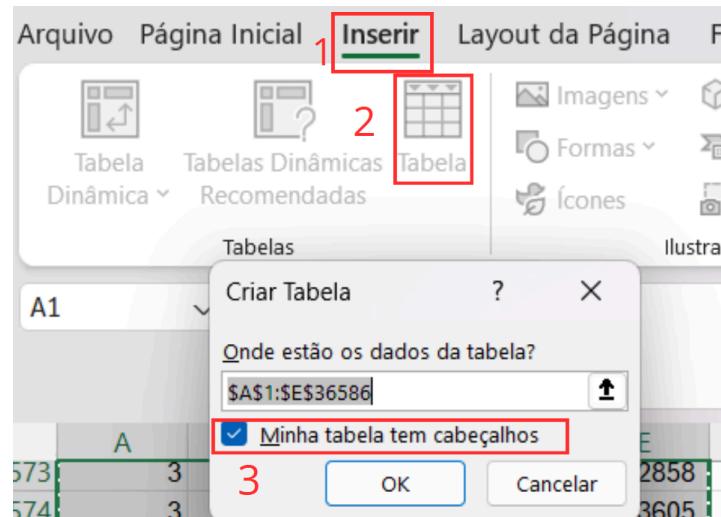
#REMARK	MapPred	1.0		
1	2	0	8	0.986532
1	3	0	8	0.711588
1	4	0	8	0.212281
1	5	0	8	0.184883
1	6	0	8	0.109699

15. Clique em “Avançar”, e em seguida “Concluir”;

16. Sua tabela ficará desta forma:

	A	B	C	D	E
1					
2	1	2	0	8	0.986532
3	1	3	0	8	0.711588
4	1	4	0	8	0.212281
5	1	5	0	8	0.184883
6	1	6	0	8	0.109699
7	1	7	0	8	0.071953
8	1	8	0	8	0.049533
9	1	9	0	8	0.035058
10	1	10	0	8	0.029754
11	1	11	0	8	0.023229
12	1	12	0	8	0.014322
13	1	13	0	8	0.011804
14	1	14	0	8	0.009890
15	1	15	0	8	0.006927

17. A primeira coluna (A) é referente ao índice do primeiro resíduo de aminoácido a ser avaliado; a segunda coluna (B) representa o índice do resíduo que está sendo avaliado a possível interação com o primeiro; a terceira e quarta coluna (C e D) indicam os limites de distância que definem um contato; a última coluna (E) indica a probabilidade de contato entre esses resíduos.
18. Coloque um cabeçalho em cada uma das colunas. Podemos utilizar, respectivamente, a nomenclatura: r1, r2, d1, d2 e p;
19. Selecione todas as células;
20. Na aba “Inserir”, clique em “Tabela”. Na janela que se abrir marque a opção “Minha tabela tem cabeçalhos”;



21. Em uma nova aba do Excel, faça o mesmo processo para os resultados de “Predicted

“distance”. As colunas 1 e 2 continuam sendo r_1 e r_2 , a terceira coluna é a probabilidade de contato (p), a quarta a coluna é a média estimada da distância C-beta entre os dois resíduos, em angstroms (d) e a última coluna o desvio padrão da distância, refletindo a variação esperada dessa média (dpm).

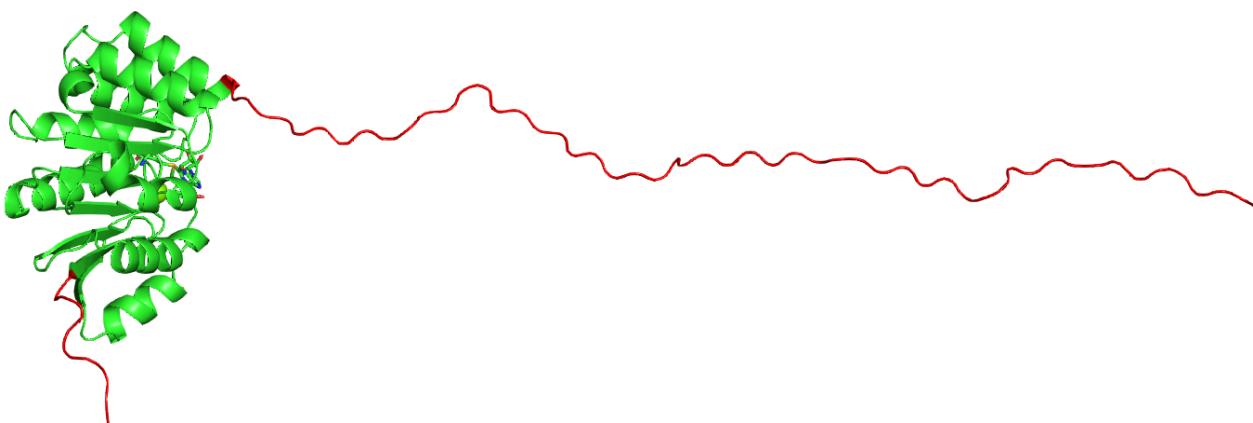
22. A partir disso é possível avaliar se os resíduos iniciais da proteína (os que estavam na conformação de coil) apresentam interações com outros resíduos da proteína. Isso irá nos auxiliar a posicionar essa cadeia de forma adequada no nosso modelo final.

Dica 1: Para facilitar a visualização, filtre os resultados indicando na coluna "r1" os resíduos que você deseja avaliar.

Dica 2: Os valores de “ p ” estão em formato decimal. Portanto, um valor de 0.986532 significa uma probabilidade de 98,6532%. Você pode multiplicar esses valores por 100 para convertê-los em porcentagem.

Avaliação dos resultados obtidos

Nesse momento, iremos avaliar os resultados obtidos e como podemos interpretá-los para utilizá-los na modelagem da nossa proteína.



Note que a proteína gerada apresentou um grande coil na porção inicial, e um menor na porção final (regiões coloridas em vermelho). Essas regiões correspondem, respectivamente, aos resíduos M1-T54 e Y262-P271.

- a) Com base no consenso dos resultados obtidos pela predição de estrutura secundária, essa região permanecerá na forma de coil ou deverá assumir outra conformação (alfa-hélice ou folha-beta)?
- b) A proteína apresenta peptídeo sinal?
- c) A proteína apresenta regiões transmembrana?
- d) Os resíduos em vermelho fazem interação com algum outro resíduo da proteína?

Para responder essas perguntas e facilitar a sua visualização, complete a tabela abaixo:

Preditores	Resíduos
Conformação alfa-hélice	
Conformação folhas-beta	
Peptídeo sinal	
Região transmembrana	
Contato entre resíduos	

Com base nos resultados fornecidos pelos preditores, juntamente com as informações da biologia da proteína que vimos ontem, quais alterações teremos que realizar na estrutura para que ela fique adequada?



Dia 3: Modelagem de um monômero com restrições estruturais

Predição Estrutural Usando Modeller

Neste momento iremos aplicar os dados de predições estruturais obtidas no dia anterior para obtermos modelos mais adequados da nossa estrutura.

Passo 1: Revisão dos resultados

Preditores	Resíduos
Conformação alfa-hélice	7-29 e 30-49
Conformação folhas-beta	N/A
Peptídeo sinal	50
Região transmembrana	7-29
Contato entre resíduos	Nenhum resultado suficientemente adequado

A partir desses resultados nós sabemos o que irá ocorrer com a porção inicial da nossa proteína, que até então estava na conformação de Coil.

É importante salientar que os dados obtidos nas predições são apenas uma estimativa e, caso tenha essas informações disponíveis em artigos baseados em experimentos, devemos sempre considerar o que está sendo informado na literatura. Um exemplo claro disso é o peptídeo sinal, no qual foi estimado que ocorreria uma clivagem entre os resíduos 32 e 33, mas a literatura fornece que a clivagem ocorre no resíduo 50. Iremos trabalhar em cima disso para a obtenção da forma **solúvel** da COMT. Apesar disso, os preditores fornecem informações suficientes para que possamos trabalhar com a proteína quando não temos resultados experimentais adequados, portanto não devemos descartá-los.

Sabemos que os resíduos 7-29 correspondem à porção transmembrana da proteína, o que está de acordo com os dados presentes na literatura. Apesar dos preditores de estrutura secundária não terem um consenso sobre a conformação dos resíduos de 30 a 40, é extremamente raro que uma proteína apresente uma alfa-hélice transmembrana tão longa (caso ela fosse de 7 a 49). Neste caso, iremos considerar que os resíduos 7-29 são uma alfa-hélice, e do 31-49 outra, com um coil ligando as duas. Esse coil será importante para fornecer maior

flexibilidade à proteína. Lembre-se que a MBCOMT necessita ter uma flexibilidade adequada para se dobrar e adquirir o íon Mg²⁺ na membrana, essencial para que ela possa atuar de forma adequada. Caso ela mantivesse uma alfa-hélice tão longa, isso não seria possível.

Passo 2: Construção do modelo tridimensional

1. Acesse o diretório **~/Minicursos/Predicao de estrutura 3D de proteinas por modelagem comparativa/dia3/modeller**. Nele está presente os arquivos necessários para a modelagem (o script, os templates no formato .pdb, o arquivo de alinhamento .ali e a sequência fasta do alvo - todos obtidos no dia 1);
2. Abra o script “**model-single-opt-ss.py**” utilizando um editor de texto de sua preferência

```
# Override the 'special_restraints' and 'user_after_single_model' methods:
class MyModel(automodel):
    def special_restraints(self, aln):

        rsr = self.restraints
        at = self.atoms

        #rsr.add(forms.gaussian(group=physical.xy_distance,
        #                      feature=features.distance(at['CB:44:A'],at['CB:58:A']),
        #                      mean=11.0, stdev=0.5))

        #rsr.add(secondary_structure.strand(self.residue_range('X:A', 'Y:A')))

        # An anti-parallel sheet composed of the two strands:
        # rsr.add(secondary_structure.sheet(at['N:1:A'], at['O:2:A'],
        #                                   sheet_h_bonds=-11))
        #rsr.add(secondary_structure.sheet(at['N:3:A'], at['O:4:A'],
        #                                   sheet_h_bonds=-11))

        rsr.add(secondary_structure.alpha(self.residue_range('7:A', '29:A')))
        rsr.add(secondary_structure.alpha(self.residue_range('31:A', '49:A')))
```

3. As linhas

```
"#rsr.add(forms.gaussian(group=physical.xy_distance,
                      feature=features.distance(at['CB:44:A'],at['CB:58:A']),
                      mean=11.0, stdev=0.5))"
```

Correspondem à restrição de distância entre resíduos. Utilizariamos esse comando caso o resultado de predição de contato fosse positivo para algum dos resíduos. Nesse exemplo,

estamos informando de que o carbono beta do resíduo 44 da cadeia A, está a uma distância de 11 Å do carbono beta do resíduo 58 da cadeia A, com um desvio padrão de 0.5 Å. Como não iremos utilizar essa linha, colocamos uma #;

4. A linha “#rsr.add(secondary_structure.strand(self.residue_range('X:A', 'Y:A')))” corresponde a restrição de estrutura secundária, informando que do resíduo X ao Y da cadeia A é uma fita Beta. Como não iremos utilizar essa restrição, manteremos com #;

5. As linhas

“#An anti-parallel sheet composed of the two strands:
rsr.add(secondary_structure.sheet(at['N:1:A'], at['O:2:A'],
#sheet_h_bonds=-11))
#rsr.add(secondary_structure.sheet(at['N:3:A'], at['O:4:A'],
#sheet_h_bonds=-11))”

Correspondem à restrição de folha beta. Neste exemplo estamos informando ao Modeller que o nitrogênio do resíduo 1 da cadeia A faz interação de hidrogênio com o oxigênio do resíduo 2 da cadeia A, a uma distância de 11 Å. Esse seria o início da folha beta. O final da folha beta corresponde ao nitrogênio do resíduo 3 fazendo interações de hidrogênio com o oxigênio do resíduo 4, a uma distância de 11 Å. Note que a distância no script está negativa. Utilizamos essa forma quando se trata de uma folha beta anti-paralela. Caso queira trabalhar com folhas-beta paralelas, basta tirar o sinal negativo.

Como também não iremos necessitar dessa restrição, manteremos com o #;

6. As linhas

```
rsr.add(secondary_structure.alpha(self.residue_range('7:A', '29:A')))  
rsr.add(secondary_structure.alpha(self.residue_range('31:A', '49:A')))
```

Correspondem a uma restrição de alfa hélice. Estamos informando ao Modeller que do resíduo 7 ao 29 da cadeia A temos uma alfa-hélice. O mesmo ocorre com os resíduos 31-49. Como necessitamos dessa restrição, deverá ser mantida sem o #;

7. O restante do script segue da mesma forma que já trabalhamos anteriormente no dia 1;

```

a = MyModel(env,
            alnfile='alvo.ali',
            knowns =('~/5lsa', './4pyi'),
            sequence='alvo',
            assess_methods=(assess.DOPE, assess.normalized_dope, assess.GA341)

        )

a.starting_model= 1                      # index of the first model
a.ending_model   = 10                     # index of the last model
                                         # (determines how many models to calculate)

# Very thorough Variable Target Function Method (VTFM) optimization:
a.library_schedule = autosched.slow
a.max_var_iterations = 300

# Thorough MD optimization:
a.md_level = refine.slow

# Repeat the whole cycle 2 times and do not stop unless obj.func. > 1E6
a.repeat_optimization = 5
a.max_molpdf = 1e6
#####
a.make()                                # do homology modeling

# Get clusters
a.cluster(cluster_cut=1.00)
# END OF MODEL CONSTRUCTION

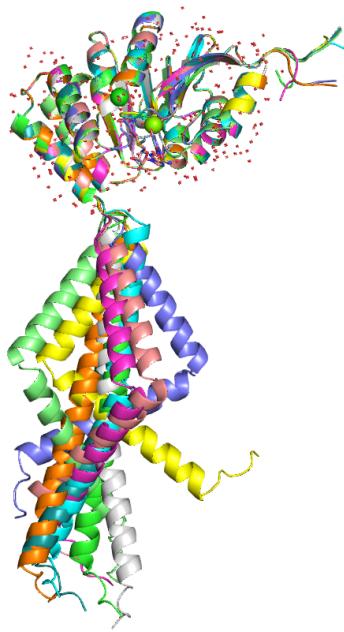
```

8. Iremos obter 10 modelos, com um refinamento lento, sendo realizado 5 ciclos de otimização. O valor máximo de molpdf que adotamos foi de 1e6 (1.000.000);
9. Salve o script e rode-o no Modeller da mesma maneira que realizamos no dia 1;

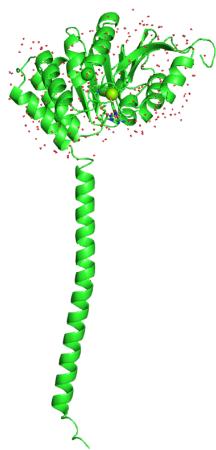
Passo 3: Avaliação de viabilidade dos modelos

Nesse momento iremos avaliar qual dos modelos obtidos está mais adequado de acordo com a literatura. Lembre-se que a proteína deve estar ancorada na membrana e com flexibilidade o suficiente para obter o íon de Mg²⁺ da membrana plasmática.

1. Abra todos os modelos ao mesmo tempo no Pymol;
2. No modelo 1, clique em **action (A) > Align > all to this (*/CA)**
3. Irá ficar dessa forma:



4. Note que onde está presente o íon de magnésio e o SAM é o sítio catalítico da enzima. Ele quem deve estar voltado em direção da membrana para o acoplamento do magnésio;
5. Analise as estruturas uma a uma e tente observar qual se enquadra mais no modelo esperado;



Note que a estrutura 1, por exemplo, não formou um coil no meio da alfa-hélice inicial. Portanto esse modelo não irá apresentar a flexibilidade adequada para que a enzima funcione da maneira correta.

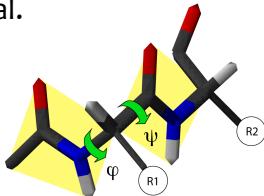
Pergunta: Qual modelo apresentou a estrutura tridimensional adequada para o funcionamento da proteína? Quais modelos foram descartados?

Validação dos modelos

Nesse momento iremos analisar a qualidade estrutural, a estereoquímica e a energia dos modelos gerados. Para isso, iremos avaliar o gráfico de Ramachandran, RMSD e os parâmetros de molpdf, DOPE score, GA341 score e Normalized DOPE.

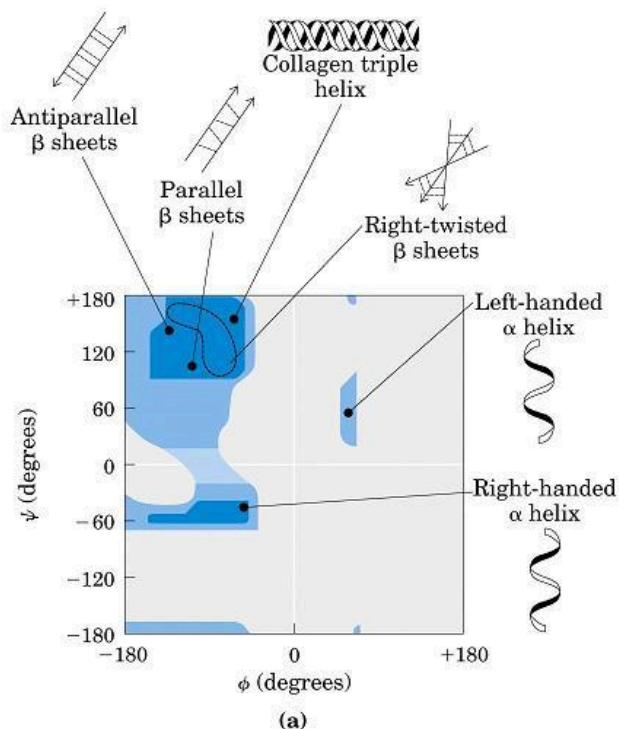
Gráfico de Ramachandran via SAVES-Procheck e Molprobity

O gráfico de Ramachandran é uma ferramenta fundamental na bioquímica, especialmente útil na modelagem de proteínas, que mapeia as possíveis conformações de um peptídeo com base nos ângulos de torção φ (phi) e ψ (psi) das ligações entre os átomos do esqueleto da proteína. Esses ângulos determinam como as cadeias laterais dos aminoácidos se orientam, influenciando diretamente a forma tridimensional da proteína. No gráfico, certas regiões correspondem a combinações de ângulos que são energeticamente favoráveis e que aparecem frequentemente em estruturas secundárias comuns, como hélices α e folhas β . Ao analisar os pontos correspondentes aos ângulos φ e ψ de cada resíduo de uma proteína no gráfico de Ramachandran, os cientistas podem prever e avaliar a conformação da proteína e identificar possíveis erros na modelagem estrutural.



O gráfico é uma representação visual das combinações possíveis desses ângulos, com base na rotação em torno das ligações C α -N (φ) e C α -C (ψ) para cada resíduo da proteína. Em termos simples, ele mostra quais conformações do esqueleto da proteína são mais prováveis ou estáveis. Como cada ângulo define a rotação ao redor de uma ligação específica dentro da proteína, o gráfico de Ramachandran permite visualizar as regiões onde a conformação é mais estável, ajudando a prever a estrutura secundária da proteína e a entender melhor sua conformação final. Mesmo que a matemática envolvida seja complexa, o uso do gráfico de Ramachandran oferece uma maneira prática de verificar se uma estrutura modelada de proteína faz sentido do ponto de vista biológico e químico.

O gráfico deve ser interpretado dessa forma:



Passo 1: SAVES-Procheck

1. Abra o site <https://saves.mbi.ucla.edu/>
2. Em “Browse” selecione sua estrutura gerada pelo Modeller

UCLA-DOE LAB — SAVES v6.1



To run any or all programs:
upload your structure, in PDB format only

The server is slower, please be patient. Send any questions or complaints to holton at mbi.ucla.edu

No file selected.

3. Clique em Run programs

Job 16978 has been created

New Job**job #16978: alvo.B99990001.pdb [job link] [3D Viewer]**

ERRAT Analyzes the statistics of non-bonded interactions between different atom types and plots the value of the error function versus position of a 9-residue sliding window, calculated by a comparison with statistics from highly refined structures. Start	Verify3D Determines the compatibility of an atomic model (3D) with its own amino acid sequence (1D) by assigning a structural class based on its location and environment (alpha, beta, loop, polar, nonpolar etc) and comparing the results to good structures. Start	PROVE Temporarily down at the moment
WHATCHECK Derived from a subset of protein verification tools from the WHATIF program (Vriend, 1990), this does extensive checking of many stereochemical parameters of the residues in the model. Start	PROCHECK Checks the stereochemical quality of a protein structure by analyzing residue-by-residue geometry and overall structure geometry. Start	OPEN We are open to suggestions for a 6th program to operate in this window. If you know of a program that we could run locally on our server that would be most useful, please let us know: email holton at mbi dot ucla dot edu with your suggestion

4. Na opção PROCHECK, clique em Start;

Start	to good structures Start	
WHATCHECK Derived from a subset of protein verification tools from the WHATIF program (Vriend, 1990), this does extensive checking of many stereochemical parameters of the residues in the model. Start	PROCHECK Complete Out of 9 evaluations • Errors: 4 • Warning: 2 • Pass: 3 Results	OPEN We are open to suggestions for a 6th program to operate in this window. If you know of a program that we could run locally on our server that would be most useful, please let us know: email holton at mbi dot ucla dot edu with your suggestion

5. Ao finalizar, clique em “Results”;

Summary	
Ramachandran plot	Warning
All Ramachandrans	Error
Chi1-chi2 plots	Pass
Main-chain params	
Side-chain params	Error
Residue properties	Pass
Bond len/angle	Pass
M/c bond lengths	
M/c bond angles	
Planar groups	Pass
Program output	

```
+-----<<< P R O C H E C K   S U M M A R Y >>>-----+
| /var/www/SAVES/Jobs/16978/saves.pdb  1.5          275 residues
+| Ramachandran plot:  94.0% core    5.1% allow    0.4% gener   0.4% disall
+| All Ramachandrans: 3 labelled residues (out of 269)
+| Chi1-chi2 plots:   7 labelled residues (out of 166)
+| Side-chain params: 5 better      0 inside     0 worse
+| Residue properties: Max.deviation: 18.6          Bad contacts: 53
+|                               Bond len/angle: 5.5          Morris et al class: 1 1 2
+|                               G-factors           Dihedrals: 0.05        Covalent: -0.23      Overall: -0.05
+| Planar groups: 100.0% within limits 0.0% highlighted
+-----+
+ May be worth investigating further. * Worth investigating further.
```

Summary file

6. Na aba à esquerda, clique em “Ramachandran plot” (segunda opção);
7. Em **Main Ramachandran plot**, escolha a opção “PDF”;
8. Nele estará presente o gráfico propriamente dito e, na parte inferior, terá um índice com as informações importantes, tais como:
 - **Residues in most favoured regions** - Resíduos nas regiões mais favorecidas. Esses resíduos possuem ângulos φ e ψ que estão dentro das regiões mais estáveis e frequentemente observadas em proteínas naturais. Nessas regiões, a conformação é energeticamente favorável e geralmente corresponde a estruturas secundárias bem definidas, como hélices α e folhas β .
 - **Residues in additional allowed regions** - Resíduos em regiões adicionalmente permitidas. Resíduos nessas regiões possuem ângulos φ e ψ que ainda são permitidos, mas são menos comuns em proteínas naturais do que aqueles na região mais favorecida. Essas conformações ainda são consideradas aceitáveis, mas são menos ideais em termos de estabilidade energética.
 - **Residues in generously allowed regions** - Resíduos em regiões generosamente permitidas. Esses resíduos possuem ângulos φ e ψ que são raros, mas ainda podem ocorrer em certas proteínas, especialmente em situações específicas como voltas e alças. Essas regiões são consideradas "generosamente permitidas" porque, embora sejam incomuns, não são completamente desfavoráveis ou impossíveis.
 - **Residues in disallowed regions** - Resíduos em regiões não permitidas. Resíduos nesta categoria têm ângulos φ e ψ que caem em regiões do gráfico onde as conformações são energeticamente desfavoráveis ou impossíveis. Resíduos nestas regiões geralmente indicam problemas na modelagem da proteína, como tensões internas ou erros de construção, que podem necessitar de revisão.
 - **Number of non-glycine and non-proline residues** - Número de resíduos não-glicina e não-prolina. A glicina e a prolina apresentam condições especiais no gráfico de Ramachandran. Todos os outros resíduos devem estar em regiões favoráveis;
 - **Number of end-residues (excl. Gly and Pro)** - Número de resíduos terminais, excluindo Glicina e Prolina. Resíduos terminais (aqueles localizados no início ou no final de uma cadeia peptídica) têm conformações menos restritas devido à ausência de ligações peptídicas adicionais. Excluindo glicina e prolina, esses resíduos são

analisados separadamente porque suas conformações podem não seguir os padrões típicos dos resíduos internos, e são considerados em análises específicas;

- **Number of glycine residues (shown as triangles)** - Número de resíduos de glicina, mostrados como triângulos. A glicina é um aminoácido especial porque não possui um grupo lateral volumoso, o que lhe confere uma grande flexibilidade conformacional. Ela pode ocupar regiões do gráfico de Ramachandran que são inacessíveis para outros aminoácidos. Por isso, a glicina é destacada separadamente (geralmente com triângulos) para não confundir sua conformação mais flexível com possíveis erros em outros resíduos;
 - **Number of proline residues** - Número de resíduos de prolina. A prolina também é única, mas por razões opostas às da glicina. Sua estrutura rígida, devido à ligação do grupo amina ao ciclo pirrolidínico, restringe severamente seus ângulos φ , o que limita as conformações possíveis. Prolinas frequentemente aparecem em regiões específicas e são tratadas de forma diferenciada porque suas restrições podem influenciar o ajuste e a avaliação da estrutura.
9. Anote os resultados de Residues in most favoured region, Residues in additional allowed regions, Residues in generously allowed regions e Residues in disallowed regions na tabela presente em **~/Minicursos/Predicao de estrutura 3D de proteinas por modelagem comparativa/dia3/validacao**
 10. Salve todos os gráficos na pasta **~/Minicursos/Predicao de estrutura 3D de proteinas por modelagem comparativa/dia3/validacao/Procheck**

Passo 2: Molprobity

1. Abra o site <http://molprobity.biochem.duke.edu/>

Main page

Dear MolProbity users, MolProbity Server Problems? Please contact us at molprobity.bugreports AT gmail.com and provide a date, time, time zone, and description. Thank you.

User at University of Iowa: Please contact us through our bug report email (see link in this site) or github molprobity project page regarding recent usage incident and other past warnings. We may be able to help you develop alternatives to using this public server for your project.

Looking at deposited SARS-CoV-2 related structures? Check PDB for updated versions as well as new structures. (Our Fetch > always returns the latest version.) Solving or improving them? Look at MolProbity's CaBLAM outliers, and at sparse H-bonds.

FILE UPLOAD/RETRIEVAL (MORE OPTIONS)

PDB/NDB code: type: PDB coords

No file selected. type: PDB coords

Usage Guidelines:
These web services are provided for analysis of individual structures.
For batch runs, please download and install your own copy of MolProbity.

2. Em “Browse” selecione sua estrutura modelada;
3. Na nova página que carregar, clique em “Continue”;

Your file from local disk was uploaded as alvo.B99990001.pdb.

- This structure was solved by THEORETICAL MODEL, MODELLER 10.5 2024/08/10 19:26:38.
- 1 chain(s) is/are present [1 unique chain(s)]
- A total of 271 residues are present.
- Protein mainchain and sidechains are present.
- No explicit hydrogen atoms are included.
- 254 hetero group(s) is/are present.
- 0 PDBv2.3 atoms were found. Proceeding assuming PDBv3 formatted file.



4. Na nova página que carregar, clique em “Analyze geometry without all-atom contacts”

Due to the parameter adjustments to hydrogen bondlengths and van der Waals radii, the current default behavior for MolProbity is to remove hydrogens, if they are present, before analysis. Please re-add hydrogens using the "Add hydrogens" option below, where you will have the option to choose either the default electron-cloud position hydrogens (i.e. for crystal structures) or nuclear-position hydrogens (i.e. for neutron-diffraction structures or for NMR structures).

Currently working on: alvo.B99990001.pdb

Add hydrogens

Edit PDB file

Downgrade file to PDBv2.3 format (for download only)

Make simple kinematics

Fill gaps in protein backbone with JiffiLoop (beta test)

Analyze geometry without all-atom contacts

5. Agora selecione os seguintes parâmetros:

3-D kinemage graphics

Universal

- Clashes
- Hydrogen bonds
- van der Waals contacts
- Geometry evaluation

Protein

- Ramachandran plots
- Rotamer evaluation
- C β deviations
- Cis-Peptide evaluation
- CaBLAM backbone markup

RNA

- RNA sugar pucker analysis
- RNA backbone conformations

Other options

- Make views of trouble spots even if it takes longer
- Alternate conformations
- Model colored by B-factors
- Model colored by occupancy
- Ribbons

Charts, plots, and tables

Universal

- Clashes & clashscore
- Geometry evaluation

Protein

- Ramachandran plots
- Rotamer evaluation
- C β deviations
- Cis-Peptide evaluation
- Show cis-nonPro and twisted peptide statistics even if the model has none
- CaBLAM backbone evaluation

Other options

- Horizontal chart with real-space correlation data
- Chart for use with Coot (may take a long time, but should take less than 1 hour)
- Suggest / report on automatic structure fix-ups
- Create html version of multi-chart
 - List all residues in multi-chart, not just outliers
 - Remove residue rows with '' altloc when other alternate(s) present

[Run programs to perform these analyses >](#)

Analyze all-atom contacts and geometry

6. Clique em “Run programs to perform these analyses”;

Summary statistics

		Poor rotamers	15	6.55%	Goal: <0.3%
		Favored rotamers	185	80.79%	Goal: >98%
		Ramachandran outliers	1	0.37%	Goal: <0.05%
		Ramachandran favored	261	97.03%	Goal: >98%
		Rama distribution Z-score	2.45 ± 0.51		Goal: abs(Z score) < 2
		C β deviations >0.25Å	11	4.38%	Goal: 0
		Bad bonds:	1 / 2198	0.05%	Goal: 0%
		Bad angles:	42 / 2993	1.40%	Goal: <0.1%
		Cis Prolines:	1 / 14	7.14%	Expected: ≤1 per chain, or ≤5%
		Low-resolution Criteria	CaBLAM outliers	4	1.5%
			CA Geometry outliers	1	0.37%
		Additional validations	Tetrahedral geometry outliers	1	

In the two column results, the left column gives the raw count, right column gives the percentage.

Key to table colors and cutoffs here:

7. Anote o parâmetro “Ramachandran outliers” na mesma tabela que você anotou os dados do procheck;
8. Desça a página até em **Single-criterion visualizations**, e selecione **Ramachandran plot PDF: View**

9. O gráfico de Ramachandran irá abrir em uma nova aba. Anote os valores de **residues were in favored regions** e **residues were in allowed regions**. Salve o gráfico na pasta **~/Minicursos/Predicao de estrutura 3D de proteinas por modelagem comparativa/dia3/validacao/Molprobity**
10. Além desses parâmetros, o Molprobity fornece:
- **Rama distribution Z-score** - Um Z-score que avalia a distribuição dos resíduos no gráfico de Ramachandran em comparação com uma distribuição ideal. Um Z-score negativo extremo pode indicar problemas na qualidade da estrutura.
 - **Z-score** é uma medida estatística que indica quantos desvios padrão um valor está afastado da média de um conjunto de dados. **Z = 0**: O valor é exatamente igual à média; **Z positivo**: O valor está acima da média; **Z negativo**: O valor está abaixo da média.
 - Em modelagem, o Z-score é usado para comparar características estruturais (como ângulos de ligação, distâncias atômicas, etc.) da proteína modelada com valores de referência obtidos a partir de estruturas de proteínas de alta qualidade.
 - **C_B deviations** - Avalia desvios na posição do átomo de carbono beta (C_B) em relação ao que é esperado. Desvios significativos podem sugerir problemas como erros na atribuição de coordenadas atômicas.
 - **Bad bonds** - Indica a presença de ligações covalentes com comprimentos anômalos, o que pode ser um sinal de erros na modelagem da estrutura.
 - **Bad angles** - Refere-se a ângulos de ligação que desviam significativamente dos valores esperados. Ângulos ruins podem sugerir problemas na conformação local da estrutura.
 - **Peptide Omegas - Cis Prolines** - A maioria das ligações peptídicas adota uma conformação trans, mas em alguns casos (especialmente com prolina), podem adotar uma conformação cis. Esse parâmetro verifica se as prolina cis estão presentes e se estão em conformações apropriadas.
 - **CaBLAM outliers** - (Ca-based Ramachandran-like Analysis Method) identifica segmentos de estrutura secundária que podem estar incorretamente modelados. Outliers indicam possíveis erros nas conformações locais.

- **CA Geometry outliers** - Avalia desvios na geometria dos átomos de carbono alfa (Ca), fundamentais para a estrutura do esqueleto da proteína. Outliers podem sugerir problemas estruturais.
- **Additional validations - Tetrahedral geometry outliers** - Verifica a geometria dos átomos tetraédricos, que são comuns em ligações de carbono. Outliers indicam desvios significativos da geometria esperada, o que pode sugerir problemas na estrutura.

Avaliação de energia

Os valores de energia estão disponíveis nos resultados da modelagem, no arquivo `model-single-opt.OUT`. Inclua todos os resultados gerados na tabela utilizada para o Procheck e o Molprobity.

Molpdf

O **molpdf** é uma pontuação global que o Modeller usa para avaliar o quanto bem o modelo gerado atende às restrições geométricas e energéticas impostas durante o processo de modelagem. Esse valor é composto por diferentes componentes, incluindo a energia interna do modelo e as penalidades associadas a desvios das restrições que foram aplicadas, como distâncias entre átomos, ângulos e outras características geométricas.

Um valor de molpdf menor indica que o modelo respeita melhor as restrições impostas e é energeticamente mais estável. É importante ressaltar, no entanto, que o molpdf não deve ser o único critério de avaliação, pois ele é específico para o conjunto de restrições e métodos utilizados pelo Modeller. Assim, ele é mais útil para comparar diferentes modelos gerados dentro da mesma sessão de modelagem.

DOPE score

O **DOPE score** é uma pontuação baseada em um potencial de energia que avalia a probabilidade de uma estrutura proteica ser realista com base em um modelo estatístico. O DOPE calcula a energia do modelo comparando-o com um banco de dados de estruturas de proteínas conhecidas e deposita essa energia em diferentes componentes, como interações entre átomos, distâncias entre pares de resíduos, e conformações gerais.

Um valor mais baixo de DOPE score sugere que a estrutura é mais estável e, portanto, mais provável de estar correta. Isso significa que as interações entre os átomos na estrutura modelada

são consistentes com aquelas observadas em estruturas de proteínas conhecidas. O DOPE score é frequentemente utilizado para comparar diferentes modelos de uma mesma proteína, ajudando a identificar qual modelo tem a conformação mais realista.

GA341 score

O **GA341 score** é um escore de confiabilidade do modelo gerado pelo MODELLER. Ele combina diferentes critérios, calculando um score que vai de 0 a 1. Valores próximos de 1 indicam que o modelo é de alta qualidade e confiável, ou seja, tem uma alta probabilidade de representar com precisão a estrutura natural da proteína. Essa pontuação é particularmente útil para uma avaliação global da qualidade do modelo e pode ser usada para selecionar os melhores modelos entre vários gerados. Dentre os critérios, podemos citar:

- **Similaridade Estrutural:** Compara a estrutura modelada com estruturas reais conhecidas.
- **Compatibilidade dos Pares de Resíduos:** Avalia se as interações entre pares de resíduos no modelo são consistentes com as interações observadas em proteínas reais.
- **Estimativa de Confiabilidade:** Oferece uma previsão da confiabilidade do modelo, levando em consideração as informações disponíveis sobre a sequência e a estrutura.

Normalized DOPE (z-DOPE)

O **Normalized DOPE** é uma versão ajustada do DOPE score original. Ele é normalizado de forma que os valores possam ser comparados não apenas entre diferentes modelos da mesma proteína, mas também entre diferentes proteínas ou diferentes regiões dentro de uma mesma proteína.

Assim como o DOPE score, valores mais baixos de Normalized DOPE indicam uma estrutura mais estável e provável de ser correta. A normalização permite uma análise mais refinada, especialmente quando se está examinando regiões específicas de uma proteína, como domínios ou loops, que podem ter diferentes conformações e energias associadas.

RMSD

Assim como realizamos o alinhamento e a obtenção do RMSD no dia 1, utilize o plugin Alignment do Pymol para obter os valores de RMSD de todos os modelos, comparando-os com ambos os templates.

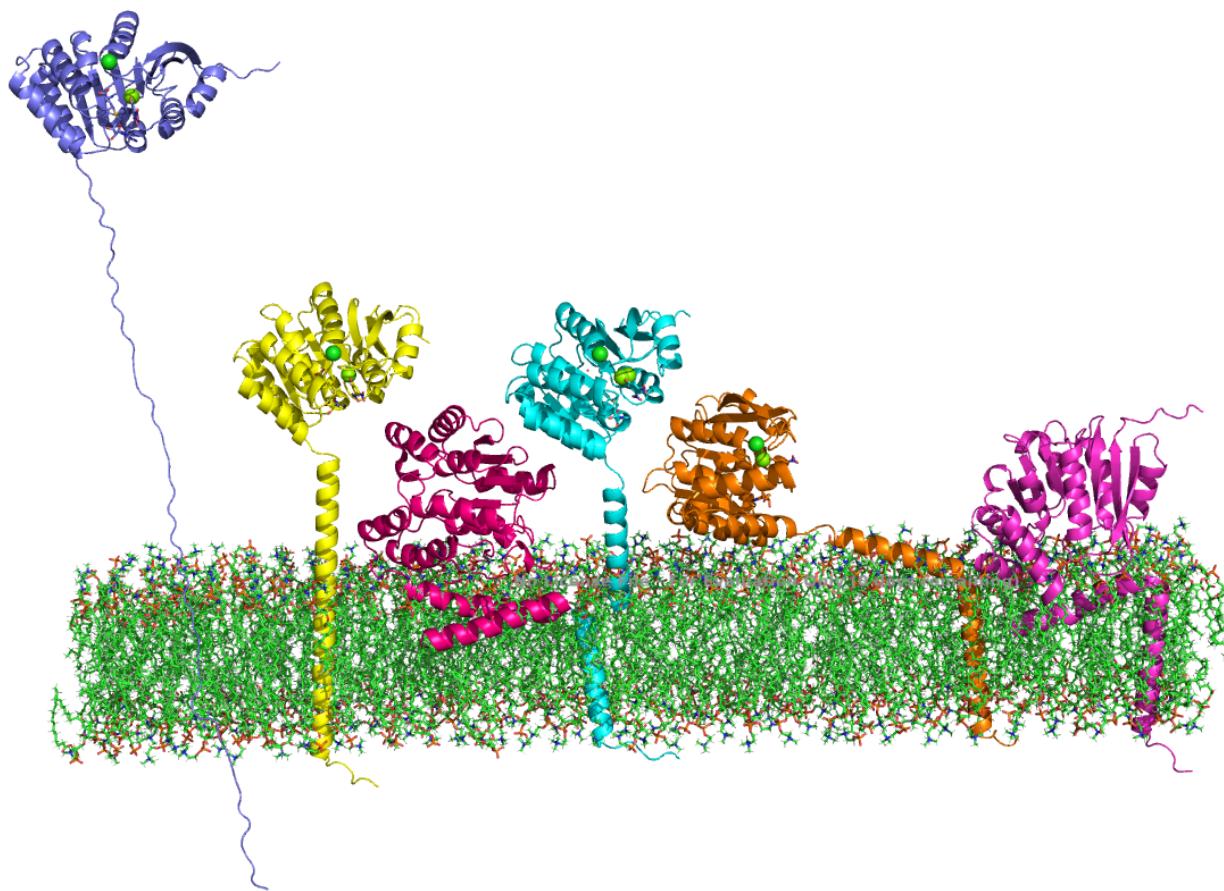
Assim será possível completar a tabela de validação de modelo.

Comparação com os modelos sem otimização

Após realizar as análises de validação dos modelos gerados sem otimização no primeiro dia, será possível avaliar a importância das restrições aplicadas.

Modelo	Molprobit			Grafico de Ramachandran (Procheck)				MOLPDF	DOPE score	GA341 score	Normalized DOPE	RMSD com 4PYI	RMSD com 5LSA
	Residues in most favoured region	Residues in additional allowed regions	Residues in disallowed regions	Residues in most favoured region	Residues in additional allowed regions	Residues in generously allowed regions	Residues in disallowed regions						
1	97.0%	99.3%	0.74%	94.5%	4.7%	0.4%	0.4%	7805,89697	-26701,11523	1,0000	0.53736	2.396	0.906
2	96.7%	99.3%	0.74%	94.0%	5.5%	0.0%	0.4%	7868,48047	-26137,26953	1,0000	0.68472	2.418	1.061
3	96.3%	99.3%	0.74%	94.0%	5.5%	0.0%	0.4%	7964,48242	-26649,41602	1,0000	0.55088	2.350	0.950
4	96.3%	99.3%	0.74%	93.6%	6.0%	0.0%	0.4%	8423,00977	-26130,03906	1,0000	0.68661	2.400	0.931
5	97.8%	99.6%	0.37%	94.5%	4.7%	0.4%	0.4%	8310,16699	-26105,37305	1,0000	0.69305	2.404	1.052

Compare esses resultados iniciais com os obtidos a partir dos modelos otimizados e com restrições gerados hoje. Vale destacar que a avaliação visual da proteína também é crucial, assegurando que a modelagem atenda às especificações funcionais previstas.



- **Azul** - modelo sem otimização e sem restrições (dia 1);
- **Amarelo** - modelo sem *coil* no meio da estrutura transmembrana;
- **Rosa** - modelo gerado pelo Robetta (falaremos melhor sobre esse modelo amanhã);
- **Ciano** - modelo com *coil* mas sem as torções e flexibilidade adequada;
- **Laranja** - modelo com *coil*, torções e flexibilidade adequada;
- **Roxo** - modelo do AlphaFold (falaremos melhor sobre esse modelo amanhã).

Agora com todos os dados obtidos, a avaliação estrutural, juntamente com as informações obtidas da literatura, responda:

Qual modelo gerado você utilizaria para seguir para os testes de docking e dinâmica molecular?

VOCÊ CONCLUIU O TERCEIRO DIA!



Dia 4: Utilização de técnicas alternativas para a predição estrutural de proteínas

Método não baseado em modelagem comparativa:

Tal método tem como princípio as Leis da Física, uma vez que são utilizados quando há a necessidade de se modelar uma proteína cuja taxa de identidade com estruturas já conhecidas não é suficiente. Assim, nesse método, o conhecimento acerca da estrutura como ângulos de torção e inserção de átomos, é feito por meio de modelos matemáticos, estatísticos e Inteligência Artificial. Para exemplificar a técnica, utilizaremos o servidor ROSETTA.

Rosetta:

Compreende-se como um conjunto de softwares que incluem algoritmos computacionais para modelagem e análise de estruturas de proteínas. Dentro desse programa será utilizado o servidor Robetta para realizar a predição de estrutura de proteínas.

Tutorial:

1. Acessar o servidor pelo link : <http://robetta.bakerlab.org>
2. Criar uma conta gratuita para acesso aos trabalhos submetidos e seus status

Create account

Required

User Name	<input type="text"/>
First Name	<input type="text"/>
Last Name	<input type="text"/>
Email	<input type="text"/>
Institution	<input type="text"/>
Password	<input type="text"/>
Must be at least 4 characters	
Confirm password	<input type="text"/>
3 + 2 = <input type="text"/>	

Optional

State/Province/Administrative Division		<input type="text"/>
Postal or ZIP Code		<input type="text"/>
Country	<input type="button" value="None"/>	
Add me to your email list <input checked="" type="radio"/> Yes <input type="radio"/> No		

Buttons:

-
- [Privacy Policy](#)

* conferir se já possui um login

3. Submeter a sequência FASTA da proteína no local indicado

The screenshot shows the Robetta web interface for protein structure prediction. At the top, there are navigation links: 'Robetta', 'Project ▾', 'Structure Prediction ▾', and a menu icon. Below this, a banner reads 'Robetta is a protein service' with a subtext about features and a learning base. A note at the bottom of the banner says 'Please do not submit jobs under different user accounts. Such jobs will be removed.' The main form area has two sections: 'Required' and 'Optional'. In the 'Required' section, there are fields for 'Target Name' (with a red '1') and 'Protein sequence' (with a red '2'). Below these is a note 'or upload FASTA' with a 'Escolher arquivo' (Select file) button. In the 'Optional' section, there are checkboxes for 'RoseTTAFold', 'CM' (with a red arrow pointing to it), 'AB' (with a red arrow pointing to it), and 'Predict domains'. There are also fields for 'Upload PDB template' and 'or enter PDB + chain IDs'. At the bottom of the form are buttons for 'Open constraints panel', 'Open fragments panel', 'Submit' (in red), and a CAPTCHA field '3 + 2 = []'. The footer includes links to 'Baker Lab', 'Rosetta@home', 'Contact', 'Terms of Service', and '©2024 University of Washington'.

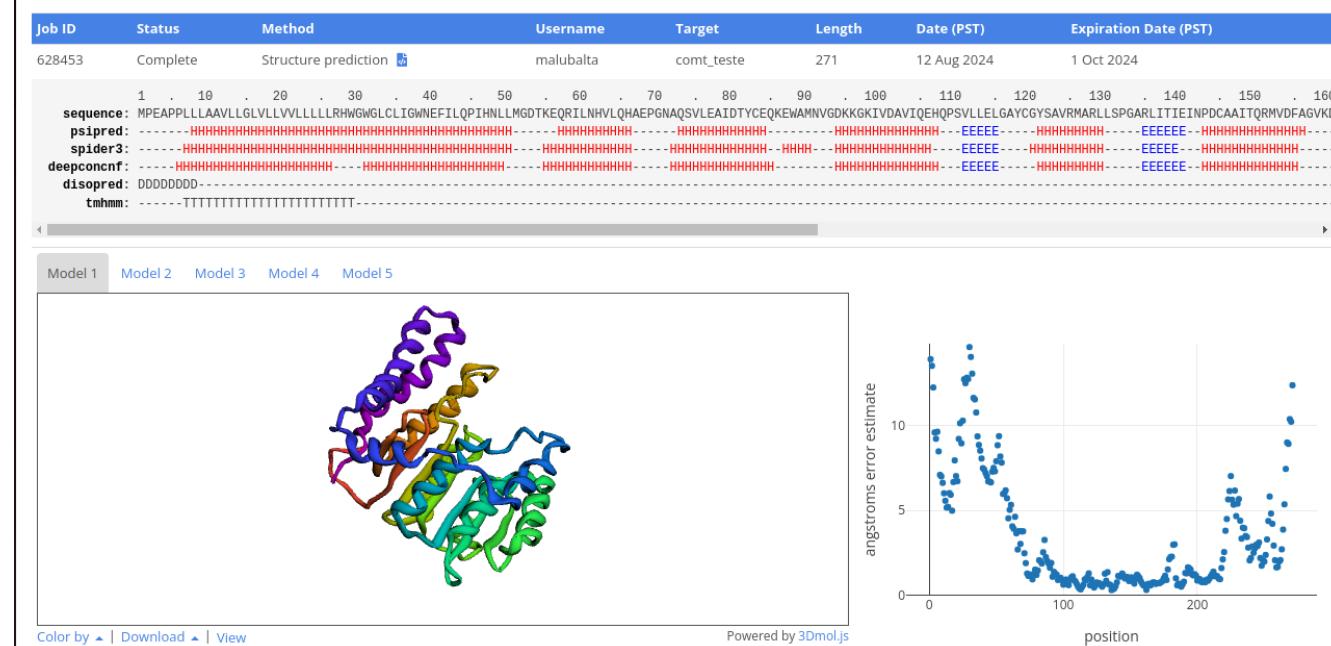
1 - Identificar o nome da proteína que será submetida

2- Colar a sequência no espaço apropriado ou fazer o upload do arquivo .FASTA

```
>ALVO
MPEAPPLLLAAVLLGLVLLVVLLLLLRRHWGWGLCLIGWNEFILQPIHNLLMGDTKEQRIL
NHVLQHAEPGNAQSVLLEAIDTYCEQKEWAMNVGDKKGKIVDAVIQEHQPSVLLEGAYCG
YSAVRMARLLSPGARLITIEINPDCAAITQRMVDFAGVKDKVTLVVGASQDIIPQLKKKY
DVDTLDMVFLDHWKDRYLPDTLLLLEECGLLRKGTVLLADNVICPGAPDFAHVRGSSCFE
CTHYQSFLEYREVVDGLEKAIYKPGPSEAGP
```

→ Selecionar a caixa “AB” (Ab Initio) e em seguida selecionar “submit”

4. Quando o resultado estiver pronto será enviada uma notificação no e-mail cadastrado
5. O resultado será apresentado dessa forma:



6. Na parte superior é possível observar os diferentes tipos de preditores que o Robetta utiliza, como o **psipred**, **spider3** e **deepconcnf**, que realizam predição de estrutura secundária; **disopred** que realiza predição de desordem; e o **tmhmm**, que realiza predição de região transmembrana.
 - a. **Predição de desordem** - identificação de regiões em uma proteína que não adotam uma estrutura tridimensional fixa ou estável. As regiões desordenadas são desafiadoras para a modelagem computacional porque não têm uma conformação única a ser prevista. Durante o processo de modelagem, essas regiões podem ser identificadas para que se tome cuidado ao interpretar as estruturas geradas, reconhecendo que essas partes da proteína podem ser flexíveis ou adaptáveis.
7. Apesar de ter realizado a predição de região transmembrana, o Robetta acoplou essa região ao restante da proteína, tornando inviável sua atividade;
8. Ao lado direito temos um gráfico de erro. A partir dele é possível avaliar o nível de confiabilidade da conformação de cada resíduo da proteína gerada. Note que a região que interage com a membrana apresentou uma alta estimativa de erro;
9. O Robetta gera 5 modelos distintos. É possível avaliar cada um deles separadamente;
10. Na porção inferior (logo abaixo da estrutura), temos a opção “**color by**”, sendo possível alterar a visualização das cores pela sequência da proteína (visualização atual), ou por estimativa de erro. Essa coloração irá mostrar as regiões com maior probabilidade de

estarem errada na cor vermelha;

11. Ao lado, em “**download**” é possível baixar a estrutura para utilizá-la para realizar seus testes.

Utilização de Inteligência Artificial

AlphaFold:

O AlphaFold é uma inteligência artificial desenvolvida pela subsidiária DeepMind da Google no qual, por meio da sequência de aminoácidos, essa IA consegue predizer a estrutura 3D da proteína. Para isso, utiliza técnicas híbridas sendo a modelagem comparativa uma delas. A predição começa com a submissão da sequência de aminoácidos que se deseja modelar. Em seguida, essa sequência é analisada por uma rede neural profunda chamada Invariant Point Attention (IPA), que foi treinada com dados de repositórios como o PDB e o Uniprot. Durante o processo, a rede neural utiliza essas informações para alinhar as sequências. A partir da análise, a rede neural prediz duas informações principais: (i) a distância entre os aminoácidos e (ii) os ângulos das ligações químicas entre eles. Essas previsões são então otimizadas por um processo chamado gradiente descendente, o que resulta na obtenção da estrutura tridimensional final da proteína (mais informações no artigo publicado pela Nature: Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). <https://doi.org/10.1038/s41586-021-03819-2>).

Competências do AlphaFold na PSP		
Consegue predizer	Prediz com dificuldade	Não consegue predizer
Cadeias de proteínas simples	Múltiplas conformações para a mesma sequência	Complexos proteína-DNA e proteína-RNA
Multímeros de proteínas	Efeitos de mutações pontuais	Estrutura de ácidos nucleicos
Complexos proteína-proteína multi subunidade	Interações antígeno-anticorpo	Interação de ligantes e íons
		Plano da membrana para domínios transmembranares

OBS: Apesar da versão AlphaFold2 conseguir predizer estruturas proteicas multiméricas, essa função é recente e ainda não se tem avaliada a porcentagem de acurácia que essa função produz.

Tutorial:

1. Acessar o servidor pelo link : <https://alphafold.ebi.ac.uk/>
2. Copiar a mesma sequência utilizada anteriormente, colar na barra de pesquisa e clicar em “search”.

The screenshot shows the AlphaFold Protein Structure Database homepage. At the top, there's a navigation bar with links to EMBL-EBI home, Services, Research, Training, About us, and EMBL-EBI logo. Below the navigation bar, the page title 'AlphaFold Protein Structure Database' is displayed, along with the subtitle 'Developed by Google DeepMind and EMBL-EBI'. A large search bar at the bottom of the main content area contains the placeholder text 'Search for protein, gene, UniProt accession or organism or sequence search'. To the right of the search bar is a 'BETA' button, and further right is a blue 'Search' button. Two red arrows are overlaid on the image: one pointing down to the search bar and another pointing to the right of the 'Search' button.

3. Buscar pelo melhor resultado (símbolo de revisão e do proteoma de referência)

The screenshot shows the search results for the protein Catechol O-methyltransferase (P21964). On the left, there's a sidebar with a list of other organisms and their counts: Rattus norvegicus (3), Bos taurus (3), Danio rerio (5), Other organisms (30), Adineta steineri (30), Rotaria sordida (18), Rotaria sp. Silwood1 (16), Sphaeramia orbicularis (orbiculate cardinalfish) (13), Pan troglodytes (11), Phytophthora parasitica (11), Sus scrofa (11), Macaca mulatta (10), Carassius auratus (9), Seriola lalandi dorsalis (9), Xenopus laevis (9), and Mycolicibacterium peregrinum (8). The main content area displays the protein details for P21964. It includes the protein name 'Catechol O-methyltransferase', UniProt ID 'P21964 (COMT_HUMAN)', and a detailed table of its characteristics. The table shows the protein is COMT, from Homo sapiens, with a UniProt ID of P21964. It also lists experimental structures (12 PDB structures for P21964) and provides a link to PDBe-KB. Below this, a sequence alignment table compares the user's query sequence (MPEAPLLAAVLLGLVLLVLLLRLHWGWLCLIGWNEFILQPIHNLLMGDTKEQRIL) with the reference sequence (P21964). The alignment shows an HSP score of 1426, an E-value of 0, and 100% identity. A 'Show full alignment' link is provided at the bottom of the alignment table.

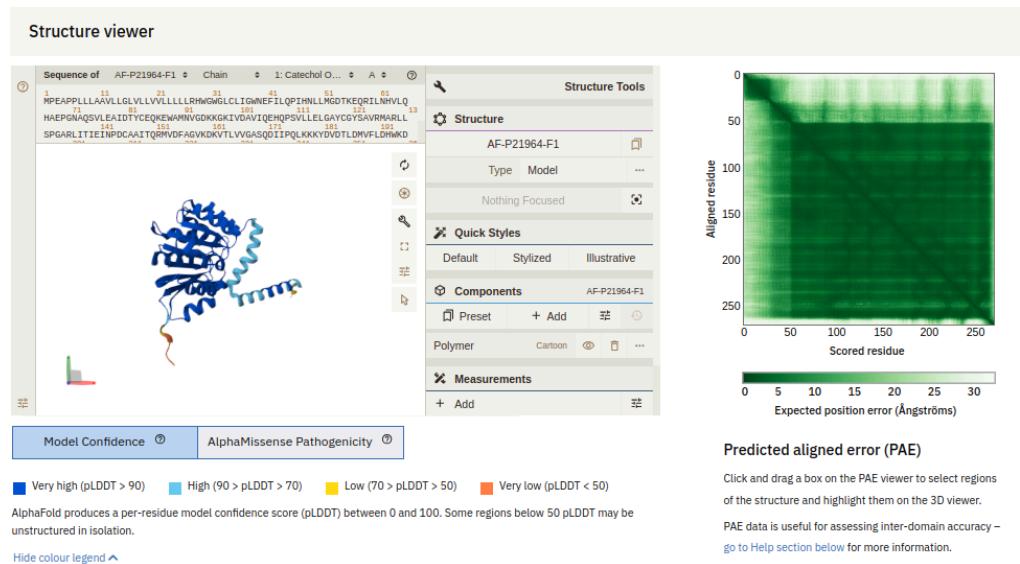
Observação: O código-fonte do AlphaFold e o Google Colab podem ser utilizados para predizer as estruturas de proteínas que não estão no banco de dados do AlphaFold.

Ambos os recursos são capazes de predizer estruturas proteicas monoméricas e multiméricas.

O código-fonte está disponível em: <https://github.com/google-deepmind/alphafold>

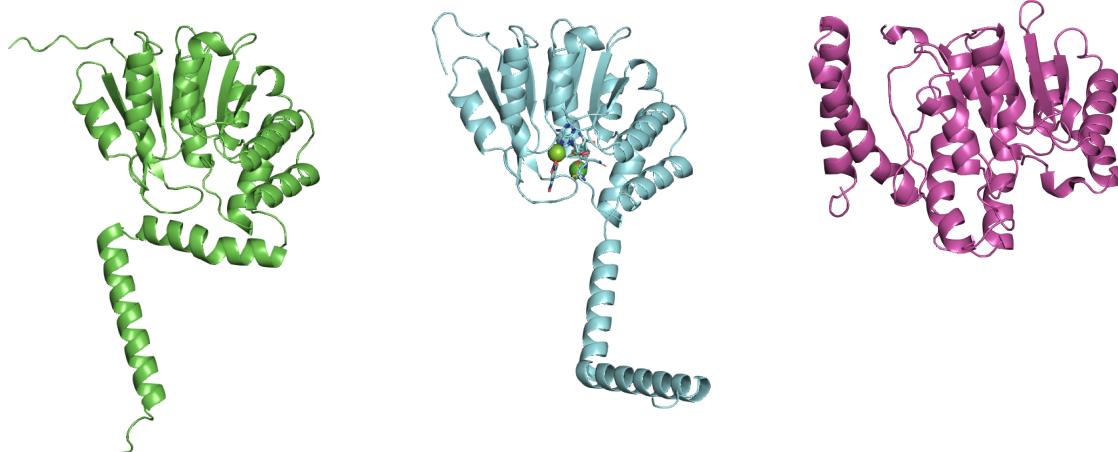
GoogleColab: <https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/>

4. Visualização do resultado

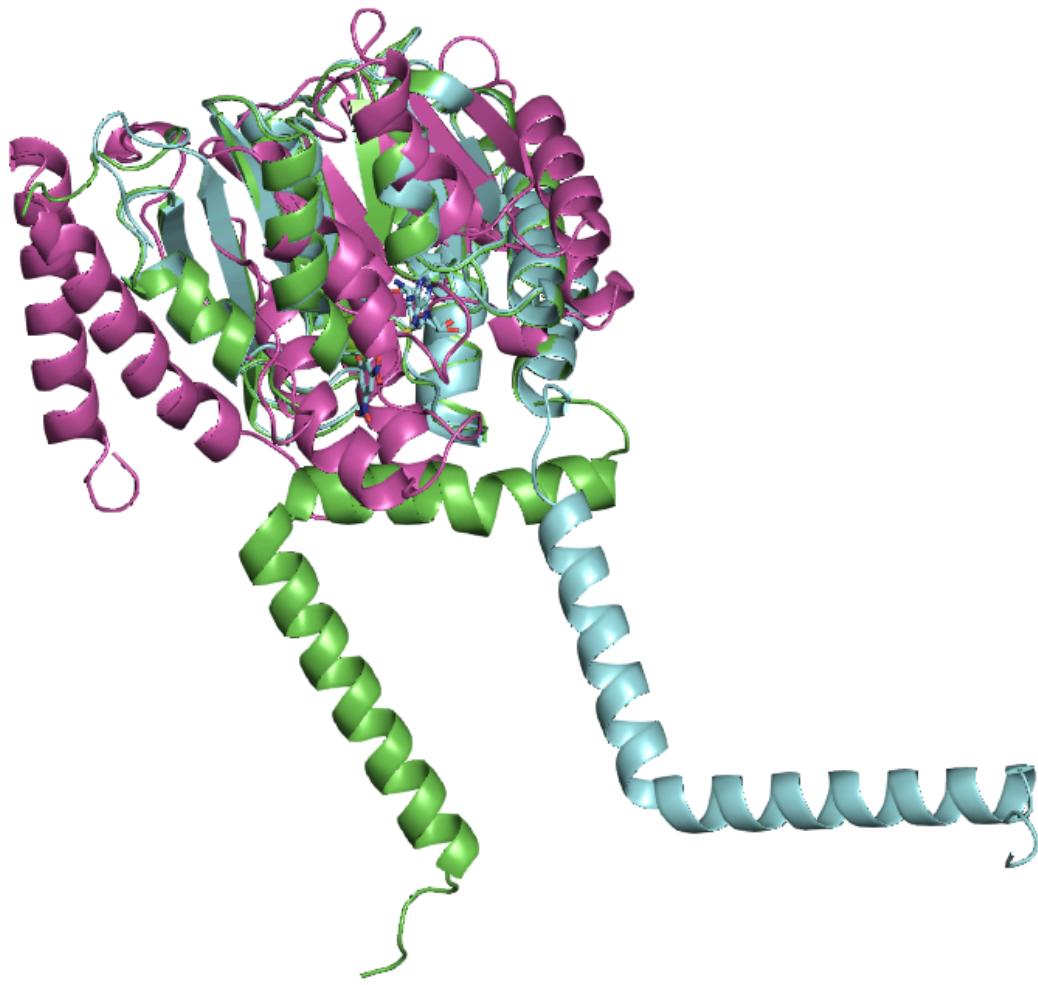


Comparação da estrutura da COMT obtida com AlphaFold, Robetta e Modeller

Neste momento iremos avaliar as diferenças entre a COMT disponível no AlphaFold e a obtida por nós durante o curso, gerada por modelagem comparativa.



Em verde é a estrutura gerada pelo AlphaFold, e em ciano a estrutura modelada no dia anterior e em roxo a estrutura gerada pelo Robetta. O alinhamento foi realizado considerando o sítio catalítico da enzima, causando maior distinção na porção transmembrana

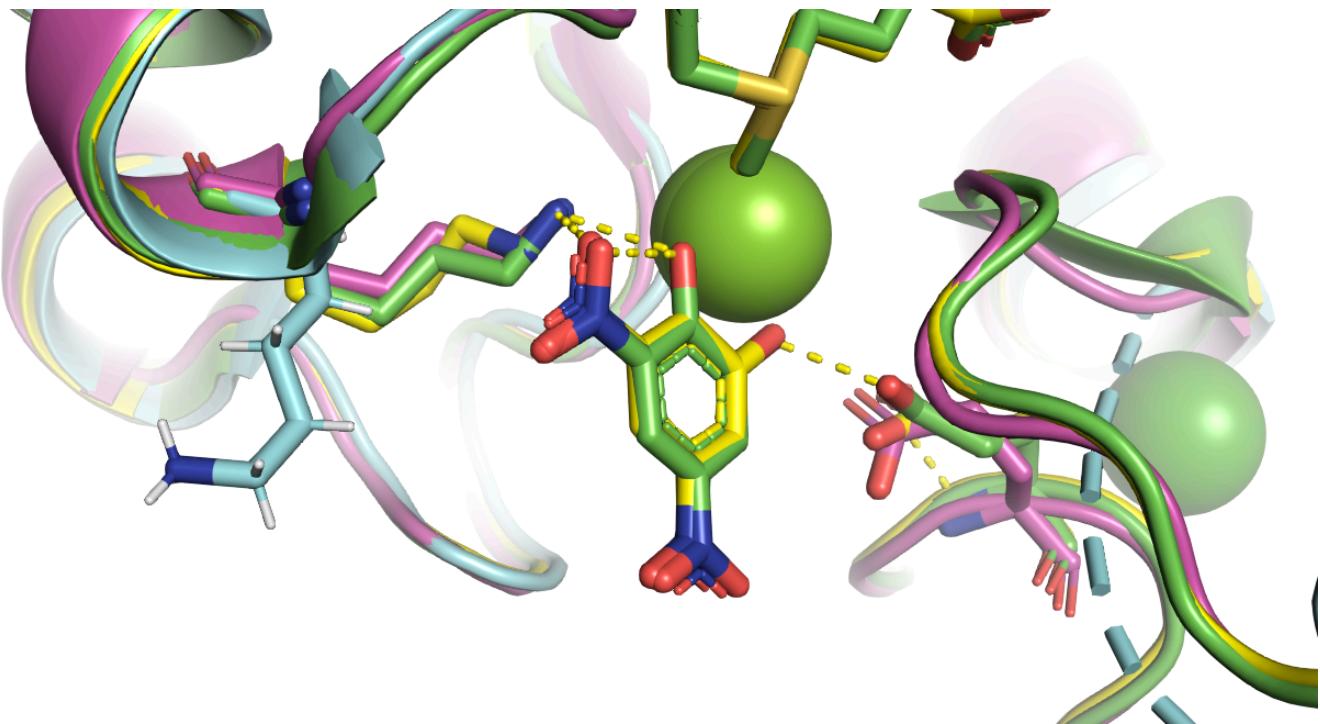


Alinhamento estrutural entre os modelos gerados pelo AlphaFold (verde), Modeller (azul) e Robetta (roxo).

Como discutimos no dia anterior, a conformação adotada pelo modelo gerado pelo AlphaFold não correspondeu ao esperado para a região transmembrana. Nesse modelo, o sítio ativo da enzima estava completamente enterrado na membrana, o que impediria a ligação com seus co-fatores e substratos, inviabilizando também a realização correta de testes *in silico*, como dinâmica molecular e docking molecular. O mesmo pode ser dito para o modelo gerado pelo Robetta, que apresentou uma incongruência ainda maior quanto a essa porção. O software não foi capaz de predizer que a porção N-terminal era composta por duas alfa-hélices, alocando-a próxima ao sítio ativo da enzima. Essa conformação não comprehende a morfologia tridimensional da proteína e a tornaria inválida para qualquer teste *in silico* futuro. Além disso, é possível notar que o alinhamento desse modelo difere muito dos demais, causando uma grande variação na

conformação e na disposição dos resíduos no sítio ativo. Sendo assim, os alinhamentos a seguir não levarão o resultado do Robetta em consideração.

Compreender a biologia da proteína em estudo é o aspecto mais crítico na modelagem de proteínas, pois os resultados podem parecer corretos, mas podem não atender às especificações funcionais. Além disso, o AlphaFold ainda não é capaz de fornecer estruturas com ligantes, o que pode levar a resultados incorretos, já que a conformação e a atividade da proteína muitas vezes dependem da presença de ligantes em sua estrutura. No caso da COMT, o AlphaFold foi capaz de gerar uma estrutura com a proteína na forma Holo (ligada ao substrato/ligante), conforme apontado na figura a seguir.



Resíduo a esquerda - Lys144; Resíduo à direita - Glu199;

Verde - Forma Holo (PDB5LSA); **Ciano** - Forma Apo (PDB2PYI); **Amarelo** - Estrutura gerada por nós no dia anterior; **Rosa** - modelo obtido no AlphaFold.

Comparação de outras estruturas

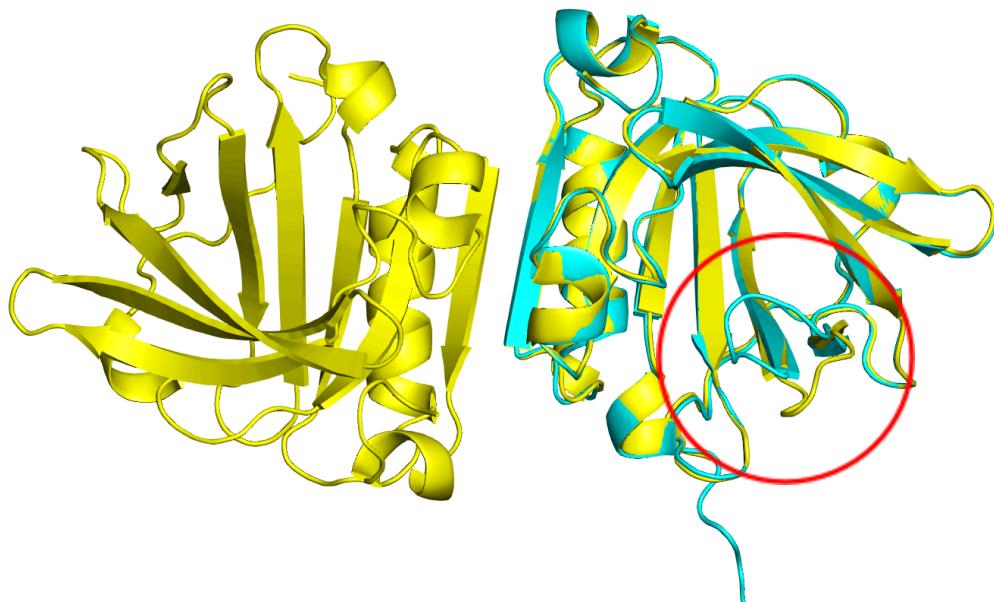
Agora, apresentaremos alguns exemplos em que a predição do AlphaFold não foi adequada devido à presença de modificações estruturais que ocorrem na fisiologia da proteína e que o AlphaFold não consegue prever.

Beta-Lactoglobulina

A beta-lactoglobulina é uma proteína globular encontrada no leite de ruminantes, desempenhando um papel crucial no transporte de moléculas hidrofóbicas, como ácidos graxos e

vitaminas lipossolúveis. Sua estrutura é composta principalmente por folhas beta que formam um barril beta, com algumas hélices alfa para estabilidade. A proteína existe tanto na forma monomérica quanto na **dimérica**, com a forma dimérica **predominando** em soluções mais concentradas e **em condições fisiológicas**. Essa dimerização pode influenciar suas interações com ligantes e outras moléculas.

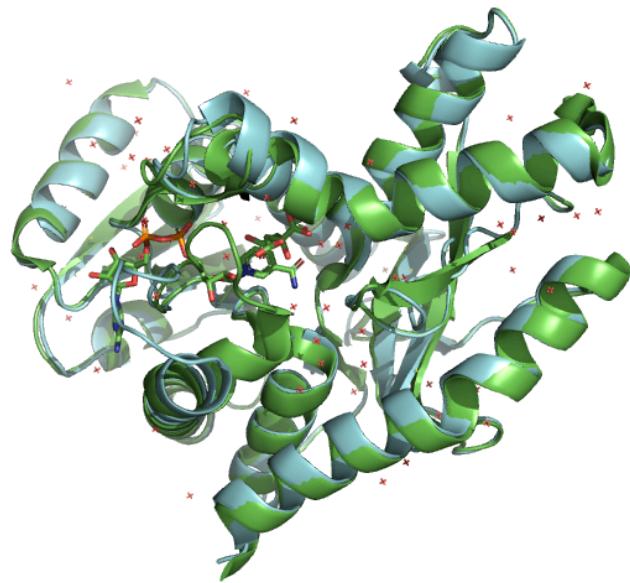
A proteína pode existir em duas formas: **apo**, sem ligantes, onde uma alça conhecida como **alça EF adota uma conformação aberta**, expondo o sítio de ligação; e **holo**, com um ligante, cuja essa **alça se fecha**, encapsulando o ligante dentro do barril beta. A alternância da alça EF entre as conformações aberta e fechada é essencial para a função de transporte da beta-lactoglobulina, permitindo a captura e liberação de ligantes. Na imagem a seguir, apresentamos um modelo gerado pelo Modeller que foi capaz de atender a dimerização e a conformação aberta da proteína, em contraste com o modelo gerado pelo AlphaFold, que resultou em uma conformação fechada.



Em amarelo temos a proteína obtida pelo Modeller, e em ciano a proteína obtida pelo AlphaFold. No círculo vermelho temos a diferença no coil para a conformação aberta e fechada. A proteína fechada esconde o seu sítio de ligação com ligantes, o que impossibilitaria testes de docking, por exemplo.

Malato Desidrogenase (MDH)

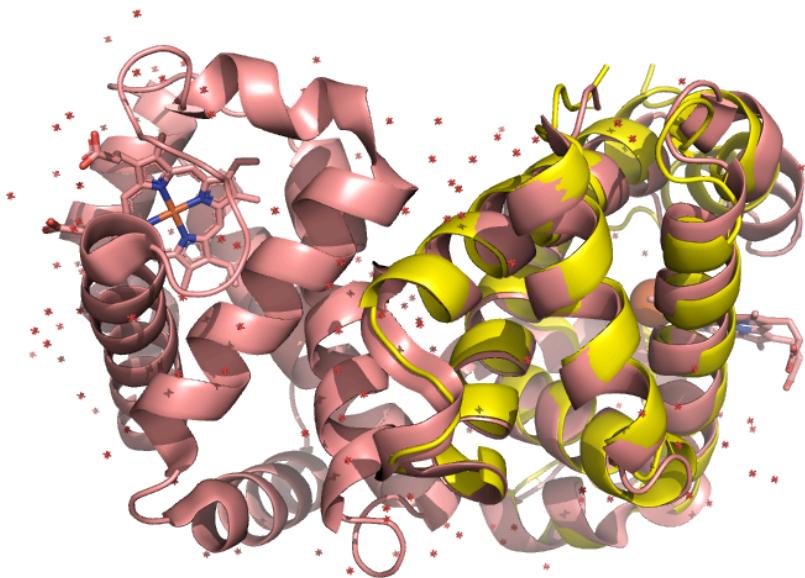
A Malato Desidrogenase (MDH), é uma enzima essencial no ciclo do ácido cítrico e utiliza como cofator a molécula NAD⁺. Quando o NAD⁺ se liga à MDH, ocorre uma mudança conformatacional que leva a uma estrutura mais fechada. Essa conformação é crítica porque aproxima os resíduos catalíticos e estabiliza a estrutura ativa da enzima, permitindo que a reação de conversão de malato a oxaloacetato ocorra de maneira eficiente. Apesar de ser uma estrutura monomérica, o AlphaFold não é capaz de identificar a presença do cofator e a conformação da proteína. Além disso, mesmo que ele tenha conseguido predizer as porções α-hélice de forma muito semelhante, em algumas porções β-pregueada não foi possível construir a estrutura tridimensional.



Legenda: Em verde, tem-se a estrutura resolvida por difração de RAIO-X no PDB (código: 1EMD) e em azul a estrutura predita pelo AlphaFold). Código Uniprot: P61889

Hemoglobina:

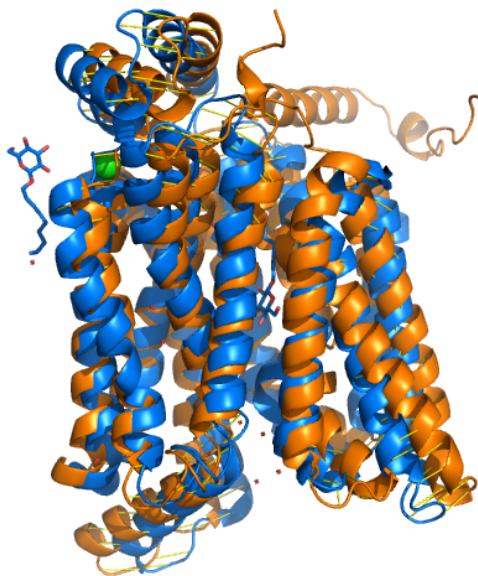
A Hemoglobina é um heterotetrâmero, sendo assim possui estrutura multimérica, também chamada de estrutura quaternária. Multímeros apresentam-se como um conjunto de estruturas terciárias formando um complexo que se liga por meio de ligações não covalentes tal qual ligações de hidrogênio, interações hidrofóbicas e pontes salinas. No alinhamento abaixo foi evidenciado apenas a cadeia B da subunidade β já que essa proteína não possui resolução do complexo protéico na sua forma multimérica. Porém, mesmo evidenciando apenas uma cadeia de uma das quatro subunidades da hemoglobina, é possível notar que a predição do AlphaFold ficou incompleta além de não mostrar o grupamento prostético heme, cuja presença é essencial para o transporte de oxigênio.



Legenda: Em rosa claro, tem-se a estrutura resolvida por difração de RAIO-X no PDB (código: 6NQ5) e em amarelo a estrutura predita pelo AlphaFold). Código Uniprot: P68871

Transportador GLUT1:

O GLUT1 é uma proteína transportadora de glicose que facilita a entrada de glicose nas células através da membrana plasmática. É composto por 12 hélices alfa transmembranares que se organizam para formar um canal através da membrana celular. Essas hélices são fundamentais para a formação da estrutura do canal e para o transporte da glicose. O alinhamento é referente à cadeia A de uma das α -hélices em questão. Apesar do AlphaFold ter reconhecido as α -hélices, nota-se que em algumas partes ele adicionou uma porção que não estava contida na estrutura resolvida do PDB além de algumas partes não se sobreponem e adquirirem conformação distinta, o que certamente influencia na atividade dessa proteína.



Legenda: Em azul escuro tem-se a estrutura resolvida por difração de RAI-O-X no PDB (código: 6THA) e em azul a estrutura predita pelo AlphaFold). Código Uniprot: P11166

Modelagem comparativa x Técnicas Alternativas

A princípio, técnicas experimentais como cristalografia de Raio-X, Ressonância Magnética Nuclear (RMN) e Crio Microscopia Eletrônica (cryo-EM) conseguem obter as coordenadas atômicas a um nível de acurácia e confiabilidade muito grande, porém são técnicas onerosas que requerem treinamento especializado. Além disso, a taxa de novas sequências de proteínas que vêm sendo descobertas atualmente é alta e o número de resoluções tridimensionais delas em bancos de dados como PDB e Uniprot não é compatível. Assim os métodos *in silico* se apresentam como uma boa alternativa para preencher essa lacuna. Nesse sentido, a partir das técnicas de PSP mencionadas durante o curso, podemos fazer um comparativo acerca das propriedades e como elas podem ser utilizadas a fim de predizer a melhor estrutura. Para isso, é necessário entender como a estrutura 3D de uma proteína é importante. A relação do arranjo tridimensional impacta diretamente na

função da proteína pois interfere na posição dos resíduos catalíticos no sítio ativo além de interferir na interação da proteína com outras moléculas e com seus próprios aminoácidos.

A modelagem comparativa é classificada como “*template based*”, cuja abordagem é baseada na teoria de que a estrutura tridimensional de uma proteína se mantém conservada ao longo da evolução. Dessa forma, os métodos dessa categoria consideram que sequências de aminoácidos semelhantes se enovelam de forma parecida. Dessa forma, pode-se elencar como ponto positivo desse método a assertividade da modelagem que irá se aproximar ao máximo da forma biologicamente ativa do alvo que se deseja modelar. Já a modelagem “*template free*” não parte do mesmo princípio. Isso porque, devido à lacuna entre o número de estruturas primárias e o número de estruturas tridimensionais resolvidas, muitas sequências não compartilham similaridade com as proteínas já resolvidas, isso requer outros métodos os quais sejam independentes de “*template*” como o método *Ab Initio* e o método *De novo*. Na modelagem *Ab initio* os programas baseiam-se em princípios termodinâmicos seguindo modelos matemáticos e estatísticos para determinar os ângulos de torção e inserção dos átomos. Assim, pode-se dizer que tal modelo é mais limitado, uma vez que é computacionalmente mais exigente e fica restrito a sequências de aminoácidos menores. Já no método *De novo*, são usadas informações provenientes de bancos de estruturas determinadas empiricamente, em forma de fragmentos estruturais sem identidade com a sequência alvo, para orientar o estado enovelado do modelo.

Por fim tem-se a utilização inovadora da Inteligência Artificial na predição de estruturas proteicas tridimensionais sendo a Google empresa pioneira com o lançamento do AlphaFold. Apesar desse sistema utilizar ambas as técnicas de modelagem mencionadas para modelar uma sequência alvo e emitir o resultado com rapidez, ele ainda apresenta algumas restrições, como a predição de estados conformacionais dinâmicos, previsão de modificações pós-tradicionais (como fosforilação, glicosilação e ligação com cofatores) e interpretação de resultados, já que a estrutura prevista por ele não possui tanto refinamento, sendo ainda necessária a análise detalhada e aplicação de técnicas de refinamento já mencionadas. No entanto, o mais importante é ter conhecimento dessas estratégias e saber avaliá-las a fim de obter um resultado promissor.



PARABÉNS, VOCÊ CONCLUIU O CURSO!



