

# Tutorial para predição de estrutura de proteínas

## VIII Escola de Modelagem Molecular em Sistemas Biológicos

Priscila V. Z. Capriles Goliatt - [priscila.capriles@ufjf.edu.br](mailto:priscila.capriles@ufjf.edu.br)

Fábio Lima Custódio - [flc@Incc.br](mailto:flc@Incc.br), Gregório Kappaun Rocha - [gregorio@Incc.br](mailto:gregorio@Incc.br), Karina Baptista dos Santos - [karinabs@Incc.br](mailto:karinabs@Incc.br)

### Comandos Básicos do Linux:

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

**Atenção:** veja também o tutorial básico de linux em <http://www.emmsb.Incc.br/index.php?pagina=26>

## Parte 1: Predição via modelagem comparativa usando MHOLline e Modeller

### Introdução

Este tutorial está orientado para usuários do sistema operacional Linux e tem por objetivo exemplificar a aplicação do portal MHOLline v2.0 (<http://www.mholline2.Incc.br>) e do software Modeller (<http://www.salilab.org/modeller/>), para predição de estruturas de proteínas (PSP) via modelagem comparativa (MC). O MHOLline é um *workflow* que combina uma série de programas usados para análise de função de proteínas e PSP via Modelagem Comparativa, em grande escala. Neste tutorial ele será usado na obtenção de modelos tridimensionais (3D) de proteínas, que servirão como ponto de partida para refinamentos das mesmas via Modeller.

### Exemplo 1: Modelagem com alta taxa de identidade

#### Passo 1: Identificando a sequência via NCBI

- 1 Acessar o site do NCBI: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>
- 2 Selecionar a ferramenta “protein blast”.
- 3 No campo “Enter Query Sequence”, colar a sequência que se deseja modelar (no formato FASTA)<sup>1</sup>:

<sup>1</sup> O arquivo FASTA contém a sequência de resíduos de aminoácidos, e pode ser obtido de diversos bancos de sequências, ou mesmo editado pelo usuário. É composto por uma primeira linha sempre iniciada por “>” seguida do nome (identificação) da sequência e por uma segunda linha contendo a sequência propriamente dita.

>TESTE

MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQORLIFAGKQLEDGRTLSDYNIQKESTLHLVLRRLGG

**Atenção:** Neste curso o arquivo FASTA já está salvo em:

***./modelagem/basico/sequences/teste\_emmsb.fasta***

4 Buscar pela sequência usando os seguintes parâmetros:

4.1 “Database”: Protein Data Bank proteins (pdb)

4.2 “Algorithm”: blastp

5 Clicar em “Algorithm parameters

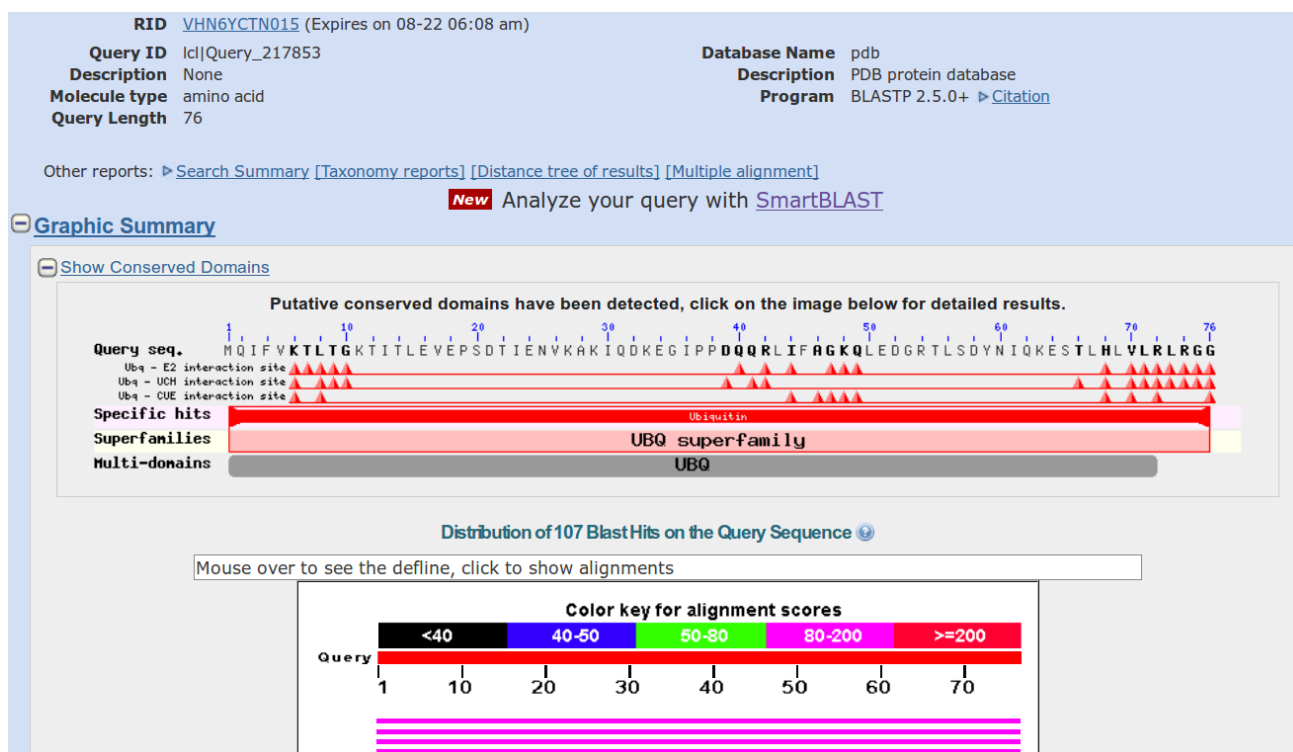
5.1 “Expected Treshold”: 0.0001

5.2 “Matrix”: BLOSUM62

5.3 “Filter”: Low complexity regions

The screenshot displays the NCBI BLAST search interface. The 'Database' is set to 'Protein Data Bank proteins(pdb)'. The 'Algorithm' is set to 'blastp (protein-protein BLAST)'. The 'Expected threshold' is set to '0.0001'. The 'Matrix' is set to 'BLOSUM62'. The 'Filter' is set to 'Low complexity regions'. The 'Max target sequences' is set to '100'. The 'Word size' is set to '3'. The 'Max matches in a query range' is set to '0'. The 'Gap Costs' are set to 'Existence: 11 Extension: 1'. The 'Compositional adjustments' are set to 'Conditional compositional score matrix adjustment'. The 'Mask' is set to 'Mask for lookup table only'.

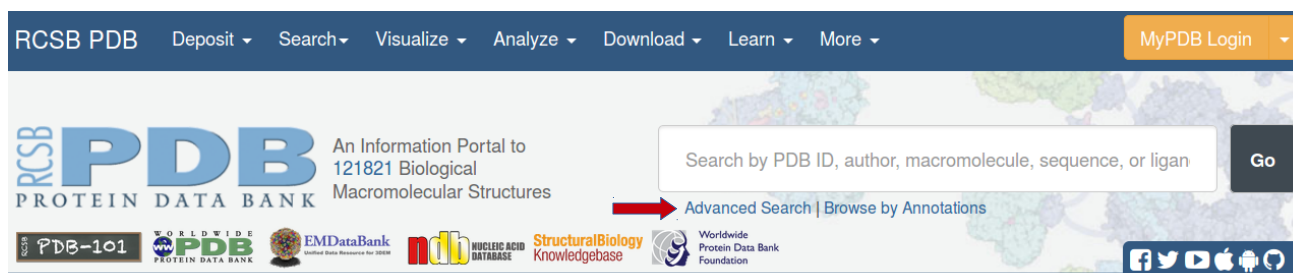
6 Analisar os resultados.



## Passo 2: Identificando a sequência via PDB

7 Acessar o site do PDB: <http://www.pdb.org/pdb/home/home.do>

8 Selecionar a ferramenta “Advanced”, para realizar uma busca avançada.



9 No campo “Choose a Query Type”, selecionar a opção “Sequence (BLAST/FASTA/PSI-BLAST)” e no campo “Sequence” colar a sequência que se deseja modelar (no formato FASTA)<sup>1</sup>.

10 Search Tool: BLAST

11 Mask Low Complexity: Yes

12 E-Value Cutoff: 0.0001

Advanced Search Interface

Sequence (BLAST/FASTA/PSI-BLAST)

Sequence search (BLAST or FASTA)

Structure Id

Chain Id

Sequence

MQIFVKLTGKTTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLLRG

Search Tool

BLAST

Mask Low Complexity

Yes

E-Value Cutoff

0.0001

Sequence Identity Cutoff (%)

0

Result Count

13 Compare o resultado obtido no PDB com o obtido pelo NCBI.

### Passo 3: Construindo o modelo via MHOLline

- 1 Acessar o site do MHOLline: <http://www.mholline2.lncc.br>
- 2 Clicar em “Submit Your Job Now” e em “Submit Now”.
- 3 Selecionar todos os programas
- 4 Fazer o upload do arquivo FASTA (**Atenção: não submeter o JOB!!!!**)

MODULE	Module Description
<input checked="" type="checkbox"/> HMMTOP	Prediction of protein transmembrane regions using hidden Markov model.
<input checked="" type="checkbox"/> TMHMM	Prediction of transmembrane helices regions.
<input checked="" type="checkbox"/> SIGNALP	Predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: <input checked="" type="radio"/> Euk <input type="radio"/> Gram+ <input type="radio"/> Gram-
<input checked="" type="checkbox"/> PSIPRED	A highly accurate method for protein secondary structure prediction.
<input checked="" type="checkbox"/> BLAST	A tool to find regions of similarity between biological sequences.
<input checked="" type="checkbox"/> BATS	A method to analyse BLAST results, defining the best template to be used in the comparative modelling.
<input checked="" type="checkbox"/> FILTERS	A tool to group sequences according to BATS score getting information about model quality.
<input checked="" type="checkbox"/> ECNGET	A tool that associates at least one Enzyme Commission (EC) number to each sequence that can be modelled. The EC number is assigned according to the PDB template.
<input checked="" type="checkbox"/> MODELLER	Production of 3D homology models of proteins structures.
<input checked="" type="checkbox"/> PROCHECK	Protein structure validation program.
<input checked="" type="checkbox"/> MOLPROBITY	A tool that offers quality validation for three-dimensional structures of macromolecules.

Upload your FASTA file:  Nenhum arquivo selecionado.

5 Acesse o resultado na caixa lateral “My Results → All” ou “My Results → Finished”

**Atenção:** Os resultados do MHOLline já estão salvos no diretório:  
./modelagem/basico/JOBteste\_emmsb

### Passo 4: Analisar o modelo construído via MHOLline

- 1 Visualizar o resultado de cada módulo no respectivo ícone:
- 2 Visualiza o resumo do resultado de modelagem clicando no botão “View”

ICON LEGENDS

Other Files	Independent				Interdependent		
<b>O</b> Original	<b>H</b> Hmmtop	<b>S</b> Signalp	<b>B</b> Blast	<b>B</b> Bats	<b>F</b> Filters	<b>E</b> ECNGet	
<b>S</b> Summary	<b>T</b> Tmhmm	<b>P</b> Psipred	<b>M</b> Modeller	<b>P</b> Procheck	<b>M</b> Molprobrity	<b>G2</b> G2	

Module Status Legend

**M** Executed

**M** Not Executed

**M** Not Selected

'EMMSB' Results - FINISHED JOBS

Showing 1 - 1 of 1 Results

Results: 10 Page: 1 of 1

JOBteste_emmsb					
Date	Status	Original	Result Files	Summary	Actions
2016-08-20 16:40:49	Finished	<b>O</b>	<div>OUTPUT</div> <div><b>H T S P B B F E M</b></div> <div><b>P M</b></div> <div>RESUME</div> <div><b>B F E M</b></div> <div><b>S</b></div> <div>INPUT</div> <div><b>H T S P B B F E M</b></div>		<div>VIEW</div> <div>DOWNLOAD</div> <div>CANCEL</div> <div>DELETE</div>

3 Clique no botão “Refine” e analise o alinhamento global e conseração de estrutura secundária

Filtering the results

Please, select at least one Quality.

Group

☐ G0
 ☐ G1
 ☒ G2
 ☐ G3

Quality [ [Select all](#) | [Unselect all](#) ]


☒ Very High
 ☒ High
 ☒ Good
 ☒ Medium to Good
 ☒ Medium to Low
 ☒ Low
 ☒ Very Low

RETURN

SEARCH

Showing 1 - 1 of 1 Results

Results: 10 Page: 1 of 1

PROTEIN	Sequence ID	Length	HETATM	Quality
	<div>Sequence name: teste</div> <div>Seq. #: 1</div> <div>PDB ID: 1YX6</div> <div>Length: 76</div> <div>HETATM:</div> <div>Organism: HOMO SAPIENS</div>	<div>Length: 76</div> <div>PDB Chain: B</div> <div>Resolution: 99</div>	<div>Group: G2</div> <div>Method: SOLUTION NMR</div> <div>EC #: Not Available</div>	<div>Quality: Very High</div>

Alignment Refinement

Previous refinements

Add (up to three) / Change Template

Restrictions

Labels:

DSSP (in the sequence):

Alpha structure

Beta structure

Coil

TMHMM (over the sequence):

Inside

Transmembrane Helix

Outside

Alignment

1\_teste

MQIFVK

TLTGK

TITLE

VEPSD

TENVKAK

IQDK

GIPPD

QRLIF

AGKQ

LEDGR

TSDYN

4XOF\_A

MQIFVK

TLTGK

TITLE

VEPSD

TENVKAK

IQDK

GIPPD

QRLIF

AGKQ

LEDGR

TSDYN

1\_teste

IQKES

TLHLVL

RLRGG

4XOF\_A

IQKES

TLHLVL

R---

RETURN

REFINE

4 Visualizar a estrutura do molde (*template*) encontrado pelo MHOLline (acesse o site: <http://www.pdb.org/pdb/home/home.do>)

4.1 Em “Search” digite o nome da estrutura 4XOF.

RCSB PDB

Deposit

Search

Visualize

Analyze

Download

Learn

More

MyPDB Login

PDB

PROTEIN DATA BANK

An Information Portal to

121821 Biological

Macromolecular Structures

4xof

Go

PDB ID

• 4XOF

4.2 Analisar os dados da estrutura (e.g. Espécie, Método Experimental, Resolução, Ligantes).

4.3 Clicar em “Sequence”: Analisar a estrutura secundária e verificar a existência ou não de ligações particulares (e.g. ligações dissulfídicas).

5 Visualizar a estrutura do modelo gerado via MHOLline: abrir com um visualizador 3D (vmd) o arquivo ./modelagem/basico/JOBteste\_emmsb/Modeller/1\_teste/1\_teste.B99990001.pdb:

5.1 Digite na linha de comando de um terminal:

```
vmd -m 1_teste.B99990001.pdb 4XOF_A.atm
```

**Atenção:** veja o tutorial básico de VMD em <http://www.emmsb.incc.br/index.php?pagina=26>

- 5.2 Para mudar a visualização no vmd: Clique em “Graphics → Representations”, em “Coloring Method” use preferencialmente a opção “Secondary Structure” e em “Drawing Method” use preferencialmente a opção “New Cartoon”.
  - 5.3 Alinhar as estruturas 3D:
    - 5.3.1 Clique em: “Extensions” → “Analysis” → “MultiSeq”.
    - 5.3.2 Selecionar: “VMD Protein Structures”.
    - 5.3.3 Clicar em: “Tools” → “Stamp Structural Alignment” → “Marked Structures” → “OK”.
  - 5.4 Use também outros visualizadores como rasmol ou pymol.
- 6 Visualizar a estrutura do modelo gerado via MHOLline: abrir com um visualizador 3D (pymol) o arquivo ./modelagem/basico/JOBteste\_emmsb/Modeller/1\_teste/1\_teste.B99990001.pdb:
- 6.1 Digite na linha de comando de um terminal:

```
pymol 1_teste.B9999000*.pdb 4XOF_A.atm
```
  - 6.2 Para mudar a visualização no pymol: No quadrante superior direito use o “Show (S) → as → cartoon” e em “Coloring (C)” use preferencialmente diferentes cores para cada estrutura.
  - 6.3 Alinhar as estruturas 3D:
    - 6.3.1 Na linha do template 4XOF\_A Clique em: “Action (A)” → “align” → “all to this”.
- 7 Visualizar a qualidade da estrutura avaliada com o programa Procheck: abrir a figura ./modelagem/basico/JOBteste\_emmsb/Procheck/1\_teste.B99990001\_01.ps
- 7.1 Para abrir o gráfico de Ramachandran gerado, digite na linha de comando de um terminal: `okular 1_teste.B99990001_01.ps&`
- 8 Visualizar a qualidade da estrutura avaliada com o programa Molprobity: abrir a figura ./modelagem/basico/JOBteste\_emmsb/Molprobity/1\_teste.B99990001.pdf
- 8.1 Para abrir o gráfico de Ramachandran gerado, digite na linha de comando de um terminal: `okular alvo.B99990001.pdf&`

*Observação: Para analisar a qualidade do modelo gerado, experimente também os seguintes programas online:*

**Qmean Server:** <http://swissmodel.expasy.org/qmean/cgi/index.cgi?>

**ProSA-Web:** <https://prosa.services.came.sbg.ac.at/prosa.php>

**Verify3D:** [http://nihserver.mbi.ucla.edu/Verify\\_3D/](http://nihserver.mbi.ucla.edu/Verify_3D/)

---

## **Passo 5: Construindo o modelo via Modeller**

### **5.1. Modelagem baseada em um único molde (template):**

O modeller trabalha basicamente com arquivos de entrada no formato PIR e *scripts* em linguagem python.

```
>P1;4XOF
```

```
MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQORLIFAGKQLEDGRTLSDYNIQKESTLHLVLRRLRG*
```

Exemplo de arquivo no formato PIR.

```
from modeller import *

log.verbose()

env = environ()

env.io.hetatm = env.io.water = True

code = './sequences/4xof'

mdl = model(env, file=code)

aln = alignment(env)

aln.append_model(mdl, align_codes=code)

aln.write(file=code+'.seg')
```

Exemplo de *script* em linguagem python.

### Preparando os arquivos de entrada:

- 1 Em um terminal, entre no diretório ./modelagem/basico/modeller
- 2 Digite na linha de comando do terminal: mod9.17 readseq.py

**Atenção:** irá aparecer a mensagem: 'import site' failed; use -v for traceback.

**Não se preocupe!**

- 3 O *script* readseq.py lê a estrutura PDB e gera arquivo.seg no formato PIR.
- 4 Abra com um editor de texto o arquivo ./sequences/4xof.seg e observe a formatação.
- 5 Copie o arquivo ./sequences/teste\_emmsb.fasta para ./sequences/alvo.seg.
- 6 Editar o arquivo ./sequences/alvo.seg para ficar no formato de entrada do Modeller.

```
>P1;./sequences/alvo

sequence:teste_emmsb:::::::::

MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQORLIFAGKQLEDGRTLSDYNIQKESTLHLVLRRLRG*
```

Exemplo de arquivo no formato.seg do Modeller.

### Gerando o alinhamento entre 4xof.seg e alvo.seg com o Modeller:

- 1 Digite na linha de comando do terminal: mod9.17 align2d.py

**Atenção:** irá aparecer a mensagem: 'import site' failed; use -v for traceback.

**Não se preocupe!**



- 2 O *script* align2d.py lê os arquivos 4xof.pdb, 4xof.seg e alvo.seg e gera os arquivos 4xof\_alvo.pap e 4xof\_alvo.ali. O primeiro contém o alinhamento entre as duas sequências no formato ALN e o segundo contém o alinhamento no formato de entrada do Modeller. Observe que nesse *script* nós definimos que a cadeia a ser usada na modelagem será a cadeia A (como selecionada pelo MHOLline).
- 3 Abra esses dois arquivos com um editor de texto e observe as diferenças.

#### Gerando o modelo da proteína alvo com o Modeller:

- 4 Digite na linha de comando do terminal: mod9.17 model-single.py

**Atenção:** irá aparecer a mensagem: 'import site' failed; use -v for traceback.

**Não se preocupe!**

- 5 O *script* model-single.py lê o arquivo 4xof\_alvo.ali, gera três modelos para a proteína alvo 4xof através da função automodel(), lista dentro do arquivo model-single.log a pontuação dos modelos gerados de acordo com a função de energia molpdf e com os métodos DOPE (*Discrete Optimized Protein Energy*) e GA341, e retorna o nome da melhor estrutura gerada de acordo com a pontuação do DOPE. Através da função cluster(cluster\_cut=1.00), clusteriza os modelos gerados e gera uma estrutura representativa já otimizada via Gradiente Conjugado (esta função é interessante para quando se deseja gerar uma grande quantidade de modelos). O resumo dos resultados é gerado dentro do arquivo model-single.out.

**Atenção:** Os modelos gerados (alvo.B99990001.pdb, alvo.B99990002.pdb e alvo.B99990003.pdb) já encontram-se no diretório ./sequences/singleinit/

**Atenção:** A estrutura representativa inicial (cluster.ini) e otimizada (cluster.opt) encontram-se no diretório ./sequences/singleinit/cluster/

#### Avaliando os modelos 3D gerados:

- 6 Abra com um editor de texto o arquivo ./modelagem/basico/modeller/model-single.out
- 7 Observe os resultados do DOPE, DOPE normalizado e GA341 para os modelos e para a estrutura representativa (cluster.opt).
- 8 Via terminal, entre no diretório ./sequences/singleinit/
- 9 Digite na linha de comando do terminal:  

```
vmd -m alvo.B9999000*.pdb cluster/cluster.opt ../4xof.pdb
```
- 10 Alinhe todas as estruturas.

#### Gerando os modelos otimizados da proteína alvo com o Modeller

(Usando o método *Variable Target Function Method - VTFM*):

- 11 Digite na linha de comando do terminal: mod9.17 model-single-opt.py

**Atenção:** irá aparecer a mensagem: 'import site' failed; use -v for traceback.

### **Não se preocupe!**

- 12 O *script* model-single-opt.py lê o arquivo 4xof\_alvo.ali e gera três modelos otimizados para a proteína alvo 4xof através da função automodel(). O resumo dos resultados é gerado dentro do arquivo model-single-opt.out.

**Atenção:** Os modelos gerados (alvo.B99990001.pdb, alvo.B99990002.pdb e alvo.B99990003.pdb) já encontram-se no diretório ./sequences/singleopt/

### **Avaliando os novos modelos 3D gerados:**

- 13 Abra com um editor de texto o arquivo ./modelagem/basico/modeller/model-single-opt.out
- 14 Observe os resultados do DOPE, DOPE normalizado e GA341 para os modelos e para a estrutura representativa (cluster.opt).
- 15 Via terminal, entre no diretório ./sequences/singleopt/
- 16 Digite na linha de comando do terminal:
- ```
vmd -m alvo.B9999000*.pdb cluster/cluster.opt ../4xof.pdb
```
- 17 Alinhe todas as estruturas.

### **Validando os modelos gerados via Procheck:**

- Avaliar os modelos gerados no diretório ./sequences/singleinit/ e no diretório ./sequences/singleopt/
- Entre no diretório ./sequences/
- Digite na linha de comando do terminal:

```
okular ./singleinit/procheckresults/*.pdf ./singleopt/procheckresults/*.pdf &
```

### **Passo 6: Analisando todos os modelos gerados**

- 1 Via terminal, entre no diretório './modelagem/basico/'
- 2 Digite na linha de comando do terminal:

```
vmd -m modeller/sequences/4xof.pdb modeller/sequences/singleinit/alvo.B9999000*.pdb  
modeller/sequences/singleopt/alvo.B9999000*.pdb JOBteste_emmsb/Modeller/1_teste/1_teste.B9999000*.pdb
```

ou

```
pymol modeller/sequences/4xof.pdb modeller/sequences/singleinit/alvo.B9999000*.pdb  
modeller/sequences/singleopt/alvo.B9999000*.pdb JOBteste_emmsb/Modeller/1_teste/1_teste.B9999000*.pdb
```

- 3 Alinhe todas as estruturas.