

Tutorial para predição de estruturas de proteínas

V Escola de Modelagem Molecular em Sistemas Biológicos

Priscila V. Z. Capriles Goliatt – capriles@lncc.br

Fábio Lima Custódio- flc@lncc.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/

* Neste curso os arquivos já estão salvos em pastas exemplificadas no Tutorial.

Parte 2: Predição *ab initio* usando o Rosetta

Introdução

Este tutorial está orientado para usuários com experiência em Linux e tem por objetivo exemplificar a aplicação do pacote de software para predição de estrutura de proteínas (PSP) Rosetta. Trata-se de uma atividade de nível introdutório para a realização da sua primeira predição.

O Rosetta compreende um grande pacote de programas complexos e que apresentam uma grande variedade de opções, de modo a serem aplicados em muitas situações diferentes. Durante o tutorial, serão utilizadas as principais opções que permitem que um usuário sem experiência prévia realize as suas primeiras predições.

Não é objetivo deste tutorial a instalação do pacote em si, uma vez que tais instruções são altamente dependentes do computador do usuário (e.g. distribuição do Linux) e podem ser encontradas na internet.

Funcionamento do Rosetta

A estratégia de PSP utilizada pelo Rosetta é baseada em fragmentos de proteínas conhecidas. Sendo assim, para uma dada sequência de aminoácidos sendo estudada (sequência alvo), os modelos construídos são compostos de partes (fragmentos) de proteínas de estrutura conhecida. O Rosetta utiliza fragmentos de dois tamanhos: três e nove aminoácidos. São utilizados fragmento, para todas as posições possíveis na sequência alvo. Cada fragmento de uma dada posição apresenta uma sequência semelhante a aquela da sequência alvo.

O primeiro passo para a utilização do Rosetta é então a geração dos fragmentos para a sua sequência alvo.

Em seguida esses fragmentos são utilizados para a construção dos modelos por meio de um processo de otimização baseado no método Monte Carlo (simulated annealing) utilizando um campo de forças contendo termos estatísticos (e.g. interação entre pares de resíduos) e também termos com significado físico (e.g. potencial de Lennard-Jones).

Algoritmo Rosetta

A montagem dos fragmentos segue um protocolo baseado em Monte Carlo, e a função potencial que guia as simulações é baseada em conhecimento, como está descrito em [1] e [2]. Os comprimentos e ângulos de ligação são mantidos fixos e as cadeias laterais são aproximadas, isto é, as interações são calculadas apenas pelo seu centróide.

Dessa forma, a conformação da proteína é determinada exclusivamente pelos ângulos de torção do esqueleto peptídico (phi, psi e omega). A função de energia nessa etapa da predição ab initio é denominada função de energia de baixa resolução.

De forma geral, o algoritmo é da seguinte forma:

1. A conformação inicial é uma cadeia completamente estendida.
2. O esqueleto peptídico é alterado por inserção de fragmentos:
 - a. Uma posição na cadeia é selecionada aleatoriamente.
 - b. Um fragmento é selecionado aleatoriamente da biblioteca, para essa posição.
 - c. Assim, alterando os valores dos ângulos phi, psi e omega da proteína, para os valores do fragmento daquela posição
3. Durante a fase inicial, fragmentos de nove resíduos são utilizados. Durante a segunda fase fragmentos de três resíduos são utilizados para refinar a estrutura.

Após um conjunto de modelos de estruturas, de baixa resolução, ser construído, vários são selecionados para refinamento com todos os átomos explícitos (full-atom). Essa etapa é necessária para que os detalhes das estruturas sejam descritos, por exemplo, a configuração das cadeias laterais. Para que possamos ser capazes de distinguir entre estruturas nativas e não nativas é necessário que não só a posição do esqueleto peptídeo esteja acurada, mas também a posição das cadeias laterais. Mais detalhes do processo podem ser encontrados em [3]. É comum realizar o refinamento full-atom apenas nas 5% melhores estruturas de baixa resolução geradas.

Organização do tutorial

O tutorial está organizado em exemplos de aplicação e é importante que seja seguida a ordem para melhor compreensão. Os textos que estiverem escritos com a seguinte formatação:

\$ echo "Hello World!"

São comandos para serem digitados em uma janela de terminal.

Antes de realizar cada exemplo sempre iremos copiar os arquivos de entrada para um diretório de trabalho.

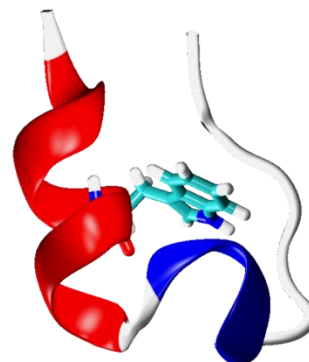
Atenção: Para evitar que as turmas seguintes sobrescrevam seus arquivos de saída crie uma pasta de trabalho, dentro da pasta rosetta3, com um nome personalizado. Para todos os efeitos usaremos nomes genéricos “exemp1” e “exemp2”.

Exemplo 1: Trp-Cage

Passo 1: Geração dos fragmentos

OBJETIVO: Criação dos arquivos de fragmentos para a sequência alvo.

O primeiro passo para aplicação do Rosetta é a geração dos arquivos contendo os fragmentos de proteínas conhecidas. Para isso precisamos de no mínimo dois arquivos: o arquivo **fasta** contendo a sequência e o arquivo **psipred** contendo a predição de estrutura secundária para a sequência alvo.



Arquivo fasta

O arquivo fasta contém a sequência alvo e pode ser obtido de diversos bancos de sequência, ou mesmo editado pelo usuário.

Arquivo psipred

A melhor maneira de se obter esses arquivos é sempre utilizar o servidor do PSIPRED (<http://bioinf.cs.ucl.ac.uk/psipred/>). Observe a predição da estrutura secundária:

```
# PSIPRED HFORMAT (PSIPRED V2.6 by David Jones)
Conf: 92132111388667888968
Pred: CCEEEEECCCCCCCCCCCCC
AA:  NLYIQWLKDGGPSSGRPPPS
      10      20
```

A linha “Conf:” indica a confiança na predição para aquela posição na sequência. A linha “Pred:” é um código para o tipo de estrutura secundária predita: C = coil, E = extended (fita) e H = helix (hélice).

– Como você acha que a predição da estrutura secundária irá afetar os modelos finais?

O Rosetta oferece a possibilidade de utilizar outros métodos de predição de estrutura secundária adicionalmente ao psipred.

Geração dos fragmentos

Teoricamente é possível instalar localmente o pacote para geração de fragmentos, porém, isso não é a prática mais recomendada pelos seguintes motivos.

A geração de fragmentos depende de um banco de estruturas: tal banco ocupa bastante espaço em disco ao ser baixado pelo usuário.

A partir do momento que um usuário baixa em seu computador uma cópia do banco ele está desatualizado.

A geração de fragmentos é demorada, levando muitas vezes mais tempo do que a própria predição da estrutura e por isso, necessita de um computador poderoso.

O programa de geração de fragmentos necessita que os programas de predição de estrutura secundária: psipred, jufo e prof estejam instalados. A instalação desses programas também tem suas complicações.

Por essas e outras razões atualmente a maneira mais prática de gerar os fragmentos é utilizar o portal Robetta mantido pelo próprio grupo do Rosetta. Isso garante sempre o uso de bancos atualizados, programas atualizados e a execução automática dos programas de predição de estrutura de proteínas.

<http://robetta.bakerlab.org/fragmentsubmit.jsp>

www.bakerlab.org

ROBETTA
Full-chain Protein Structure Prediction Server

Structure Prediction Fragment Libraries Alanine Scanning
[Queue] [Submit] [Queue] [Submit] [Queue] [Submit]
[Register / Update] [D 5 Qs] [News] [Software] [Login]

Submit a job to the Fragment Server
*Please submit one job at a time

Required
1 Registered Username: or Registered Email Address:
2 Target Name:
3 Paste Fasta
or Upload Fasta: Choose File No file chosen
4 Submit

Optional
Identifier:
Exclude Homologues: ☒
Rosetta NMR (click links below for input format)
Chemical Shifts: Choose File No file chosen
NOE Constraints: Choose File No file chosen
Dipolar Constraints: Choose File No file chosen
Submit

Robetta is available for NON-COMMERCIAL USE ONLY at this time
[Terms of Service]
Copyright © 2004-2007 University of Washington

1. Digite um nome de usuário registrado ou email em um dos campos “Registered...”. Podem usar o usuário “**vemmsb**” criado para a escola.
2. Coloque no campo “Target Name” um código de quatro letras.
3. Cole o conteúdo do arquivo fasta no campo “Paste Fasta”. Inclua a linha de cabeçalho (iniciada por >).
4. Aperte o botão “Submit”.
5. Agora basta acompanhar na fila (queue) o andamento dos seus fragmentos.

A opção “Exclui Homologues” exclui da geração de fragmentos estruturas com sequência homóloga à sequência alvo.

Para esse tutorial já geramos os fragmentos e os arquivos se encontram no computador.

Atenção: Nesse momento é interessante você submeter alguma sequência de interesse para geração de fragmentos que poderão ser utilizados durante o último dia de prática.

Atenção²: Estão sendo fornecidos fragmentos e resultados excluindo-se ou não homólogos. O tutorial segue utilizando os resultados sem homólogos. Caso você deseje utilizar homólogos basta alterar o sufixo “_nh” para “_h” do nome das pastas, por exemplo: de “trpc_frgs_nh” para “trpc_frgs_h”.

O Robetta fornece os seguintes arquivos:

```
aat000_03_05.200_v1_3
aat000_09_05.200_v1_3
t000_.check
t000_.checkpoint
t000_.dat
t000_.fasta
t000_.homolog_nr
t000_.homolog_val1
t000_.jufo_ss
t000_.psipred
t000_.psipred_ss2
t000_.rdb
```

Esses arquivos foram baixados diretamente do servidor:

<http://robetta.bakerlab.org/downloads/fragments/16561/>

Agora crie e entre no diretório de trabalho:

```
$ mkdir ~/exemp1
$ cd ~/exemp1
```

Copie os arquivos da geração de fragmentos para o diretório de trabalho (ou baixe do servidor):

```
$ cp ~/abinitio/trpc_frgs_nh/* ~/exemp1
```

Existem diversos arquivos intermediários do processo de criação além dos **dois arquivos contendo os fragmentos de três aminoácidos**:

```
$ less aat000_03_05.200_v1_3
```

e nove aminoácidos:

```
$ less aat000_09_05.200_v1_3
```

Cada arquivo contém 200 fragmentos diferentes para cada uma das posições possíveis da sequência alvo. Observe as informações contidas nos fragmentos, como, por exemplo, código

pdb da estrutura de origem, tipo de aminoácido, tipo de estrutura secundária e ângulos de torção do esqueleto peptídico.

Existe um arquivo muito interessante:

t000_.homolog_val1

1. Liste o conteúdo desse arquivo.
2. Abra a estrutura listada nesse arquivo no PDB.

Passo 2: Construção dos modelos

OBJETIVO: Geração de 20 modelos para trp-cage utilizando o protocolo "abfastrelax".

Os arquivos de entrada para geração dos modelos são semelhantes aos arquivos de entrada para geração dos fragmentos. A diferença é que para a construção dos modelos os arquivos contendo os fragmentos são necessários.

Execute o Rosetta para gerar os modelos para essa sequência.

O nome (com o caminho) do executável que gera modelos ab initio com refinamento é:

\$ ~/rosetta3/rosetta_source/bin/AbinitioRelax.linuxgccrelease

Monte então uma linha de comando com os seguintes parâmetros:

```
-in::file::fasta t000_.fasta
-database ~/rosetta3/rosetta_database
-in:file:frag9 aat000_09_05.200_v1_3
-in:file:frag3 aat000_03_05.200_v1_3
-out:pdb
-out:nstruct 20
-abinitio:fastrelax
-run:use_time_as_seed
```

Agora é só aguardar de sete a oito minutos... Pronto, você acabou de criar vinte modelos para a sequência da proteína trp-cage utilizando o Rosetta.

Entendendo a linha de comando:

-in::file::fasta t000_.fasta	-> passando o arquivo .fasta
-database ~/rosetta3/rosetta_database	-> localização dos arquivos de parâmetros do Rosetta
-in:file:frag9 aat000_09_05.200_v1_3	-> passando arquivo de fragmentos com 9 aas
-in:file:frag3 aat000_03_05.200_v1_3	-> passando arquivo de fragmentos com 3 aas
-out:pdb	-> pdb como tipo dos arquivos de saída
-out:nstruct 20	-> gerar 20 trajetórias diferentes, logo, 20 estruturas
-abinitio:fastrelax	-> fazer o relaxamento rápido (em alta resolução)

Passo 3: Análise dos resultados

OBJETIVOS: Observar os modelos gerados. Calcular o RMSD em relação à estrutura experimental.

Nessa etapa o ideal é que você use o programa de visualização de sua preferência. O objetivo é comprar a estrutura dos modelos gerados com a estrutura determinada experimentalmente depositada no PDB. Neste tutorial usaremos o programa VMD para tal tarefa, além de calcular o RMSD. Note que, em uma aplicação “real” o usuário não tem certeza da estrutura nativa daquela sequência, dessa forma essa etapa de calcular o RMSD é puramente ilustrativa.

Baixe a estrutura nativa 1l2y do pdb usando um navegador ou digitando:

```
$ wget -q -O 1l2y.pdb  
"http://www.pdb.org/pdb/download/downloadFile.do?fileFormat=pdb&compression=NO&structureId="1l2y"
```

Foram criados vários arquivos do tipo pdb (“S_00*.pdb”) numerados de um a 20.

Abra todos os modelos gerados e o arquivo com a estrutura experimental no vmd.

```
$ vmd 1l2y.pdb -f S_00*.pdb
```

A opção “-f” faz com que todos os arquivos especificados na linha de comando depois dela sejam carregados em uma única “molécula”. Isso facilita a comparação entre as estruturas geradas utilizando-se o cálculo do RMSD de uma trajetória. Calcule o RMSD em relação à estrutura experimental:

1. Abra o item: “Extensions>Analysis>RMSD Trajectory Tool”
2. Na janela que apareceu escolha a opção “backbone”.
3. No campo “Ref:” escolha a opção “Selected” e clique na linha do “1L2Y.pdb”.
4. Clique em “Align” para alinhar as estruturas.
5. Clique em RMSD para calcular.
6. Ainda nessa janela, abra a opção “File>Plot Data”.

Analise os resultados. Qual dos modelos gerados foi o que melhor se aproximou da estrutura experimental? Se você possuir experiência no uso do VMD, tente observar o posicionamento do resíduo de TRP desse modelo.

Exemplo 2: Ubiquitina – 1ubq

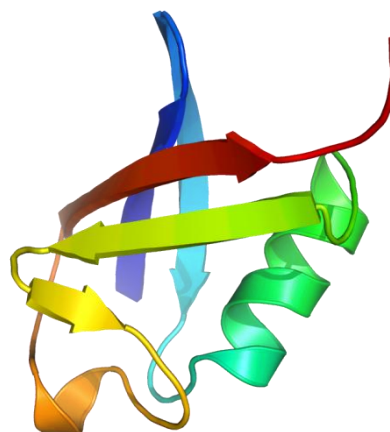
Neste exemplo tentaremos executar uma predição de uma estrutura mais complexa para comparar com o resultado do modelo gerado usando **modelagem comparativa**.

Passo 1: Geração dos fragmentos

OBJETIVO: Geração dos arquivos de fragmentos.

Os fragmentos já foram fornecidos. Crie uma cópia de trabalho do diretório contendo os arquivos de entrada (ou baixe de <http://rosetta.bakerlab.org/downloads/fragments/16562/>);

```
$ mkdir ~/exemp2  
$ cd ~/exemp2  
$ cp ~/abinitio/ubiq_frgs_nh/* ~/exemp2/
```



Note que além do arquivo fasta e do arquivo com a predição de estrutura secundária gerado pelo psipred, esse diretório contém o resultado da predição de estrutura secundária gerado pelos programas SAM-T99 (.rdb) e JUFO (.jufo). O programa de geração de fragmentos é capaz de usar mais de uma predição para melhorar a escolha dos fragmentos.

Observe que os arquivos eles são muito maiores do que os gerados no exemplo anterior.

```
$ less aat000_03_05.200_v1_3
```

Abra o arquivo de homólogos excluídos **t000_.homolog_vall**.

Passo 2: Construção dos modelos

OBJETIVO: Geração de modelos com o protocolo “abfastrelax”.

Leia na seção “Passo 2: Construção dos modelos”, página 6, como montar a linha de comando. **Estamos tratando de um alvo maior então gere apenas 2 modelos.**

A opção “**-abinitio:fastrelax**” ativa a simulação com todos os átomos explícitos (full-atom) e utilizam um campo de forças com mais detalhes. Além disso, o posicionamento das cadeias laterais é otimizado com base em uma biblioteca de rotâmeros.

Abra uma nova janela de terminal e dentro do diretório de trabalho do segundo exemplo execute:

```
$ ls S_00*.pdb
```

Para acompanhar a geração dos modelos.

Passo 3: Análise dos resultados

Abra um navegador de sua escolha e entre no PDB (www.pdb.org) para baixar o arquivo com a estrutura determinada experimentalmente, por cristalografia de raios-X, da ubiquitina - 1ubq (<http://www.pdb.org/pdb/explore/explore.do?structureId=1UBQ>).

Como no passo 3 do exemplo 1 (página 7) abra os arquivos com os modelos de alta resolução e o arquivo baixado do pdb.

```
$ vmd 1ubq.pdb -f S_00*.pdb
```

Calcule o RMSD entre as estruturas.

Em aplicações práticas, onde não se sabe a estrutura nativa de antemão, a análise dos resultados vai muito além do cálculo do RMSD. Em tais situações, o pesquisador deve recorrer ao máximo de informações disponíveis sobre a função, a família, estruturas homólogas, etc. Para assim avaliar a qualidade dos modelos.

Passo 4: Selecionando modelos

OBJETIVOS:

(1) Agrupamentos dos 100 modelos fornecidos utilizando o “maxcluster”.

(2) Calcular o RMSD entre os 100 modelos e a estrutura experimental.

(3) Comparar o menor RMSD e o RMSD do “melhor modelo” segundo o “maxcluster” com a estrutura experimental.

Atenção: Os protocolos de predição ab initio de estruturas de proteínas costumam gerar centenas e até milhares de estruturas diferentes para cada sequência. Por exemplo, o protocolo do Rosetta aplicado no CASP, além de usar trajetórias de otimização mais longas (com 120x mais passos) geram mais de 1500 estruturas por sequência alvo.

Sem conhecimento sobre a estrutura nativa como escolher um modelo dentre tantos?

- Agrupamento de estruturas. A idéia é que as estruturas mais comumente encontradas ao final das trajetórias devem ser as mais acessíveis, assim, mais próximas da nativa.

Para realizar o agrupamento dos modelos utilizamos o programa **maxcluster**. Atualmente existem diversas ferramentas para isso. Se você estiver familiarizado com alguma a utilize então.

```
$ ~/rosetta3/rosetta_source/maxcluster -h
```

Os modelos gerados aqui **não** são suficientes para se realizar um agrupamento satisfatório.

Foram fornecidos 100 modelos para a realização desse passo. Copie esses modelos para sua pasta de trabalho:

```
$ cp ~/abinitio/runubq_nh/*.pdb ~/exemp2/
```

Crie um arquivo texto contendo uma lista com o nome dos arquivos .pdb gerados. Depois passe esse arquivo para o maxcluster usando a opção “-l”.

```
$ cd ~/exemp2
```

```
$ ls -1 S_00*.pdb > lista.txt
```

```
$ ~/rosetta3/rosetta_source/maxcluster -l lista.txt
```

1. Observe os resultados.
2. Anote o número da estrutura centróide do grupo com o maior número de pares na saída do maxcluster, seção:

```
INFO : =====
INFO : Nearest Neighbour clustering
INFO : =====
INFO : Centroids
INFO : =====
INFO : Cluster Centroid Size Spread
```

3. Abra **todos** os modelos no vmd e calcule o RMSD em relação à experimental.

```
$ vmd 1ubq.pdb -f S_00*.pdb
```

4. Compare os resultados.
5. Compare o centróide do maior cluster com a estrutura modelada usando o Modeller.

Repita essa análise para o Trp-cage:

```
$ cd ~/exemp1
```

```
$ cp ~/abinitio/runntrpq_nh/*.pdb ~/exemp1/
```

```
$ ls -1 S_00*.pdb > lista.txt
```

```
$ ~/rosetta3/rosetta_source/maxcluster -l lista.txt
```

Exemplo 3: Outra sequência

Neste exemplo não forneceremos nenhum arquivo de entrada. O objetivo é a construção de um modelo de alta resolução para uma sequência de interesse do usuário.

Passo 1: Geração dos fragmentos

Passo 2: Construção dos modelos

Passo 3: Análise dos resultados

Referências

- [1] Simons KT, Kooperberg C, Huang E, Baker D (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* 268, 209-25.
- [2] Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* 34, 82-95.
- [3] Bradley P, Misura KM, Baker D (2005). Toward high-resolution de novo structure prediction for small proteins. *Science* 309, 1868-71.
- [4] Humphrey, W.; Dalke, A. & Schulten, K. (1996). VMD -- Visual Molecular Dynamics *Journal of Molecular Graphics*, 14, 33-38.