Introdução à Bioinformática

Trabalho Prático 02
Cálculo de Contatos entre Proteínas e Ligantes

Aluno: Gustavo Campos Ferreira Guimarães – 2009052247

Introdução

O presente trabalho foi desenvolvido durante a disciplina de Introdução à Bioinformática, ministrada pela professora Sabrina Silveira. O objetivo principal, como interpretado pelo aluno, é extrair informações de arquivos PDB, representando proteínas e respectivos ligantes, seguido do cálculo de prováveis contatos ocorridos entre essas moléculas e posterior análise dos resultados obtidos.

A fim de atingir o objetivo proposto, escolheu-se a linguagem Python em sua versão 2.7.5, com breves utilizações dos pacotes externos *numpy* e *gzip.*

Decisões de Implementação

1. Algoritmo Utilizado

Duas estratégias são largamente citadas na literatura quando o requisito é o cálculo de contatos. A mais complexa, e não coincidentemente mais exata, é a estratégia geométrica. Tal técnica implementa diagramas de Voronoi e a triangulação de Delaunay a fim de calcular, além das distâncias entre átomos, a influência de todos os átomos vizinhos numa possível ligação. Isso garante resultados melhores à medida que considera toda a vizinhança de um determinado átomo ao buscar contatos.

Já a técnica adotada neste trabalho, bem mais simples, é denominada *cutoff*. Trata-se de uma estratégia baseada no cálculo da distância euclidiana simples num espaço euclidiano de três dimensões. Define-se um valor de distância arbitrário, chamado de *cutoff*, e considera-se prováveis contatos todos os pares de átomos cuja distância for menor do que este valor. Essa técnica produz resultados muito rapidamente e é de implementação bastante simples, mas por desconsiderar as interações entre múltiplos átomos, emite resultados menos confiáveis que a estratégia geométrica.

2. Decodificação do Arquivo PDB

A fim de extrair as informações contidas nos arquivos PDB, foi consultada a especificação do formato no website original. A versão corrente do padrão PDB pode ser consultada na seguinte URL:

http://www.wwpdb.org/documentation/format33/v3.3.html

Essa versão data de Julho de 2011, com as últimas correções adicionadas em Novembro de 2012. Estamos interessados especificamente nas coordenadas geométricas dos átomos e heteroátomos, cuja especificação encontra-se na subseção “Coordinate Section”, itens “ATOM” e “HETATM”.

A fim de agilizar a produção de resultados, optou-se por ler os arquivos PDB compactados, na forma como foram fornecidos. Essa escolha baseia-se em evidências, ainda que anedóticas, de que o overhead de processamento obtido pela descompactação em tempo real dos arquivos é menor do que o overhead de entrada/saída na leitura dos arquivos descompactados. A melhoria de performance, no entanto, é marginal e não deve ser considerada relevante.

3. Filtragem do Arquivo PDB

Adotando instruções de sala de aula, nem todos os átomos das proteínas fornecidas foram considerados para o cálculo de contatos. Ao invés disso, optou-se por utilizar apenas os carbonos alfa de cada aminoácido para este cálculo. Essa abstração reduz a quantidade de dados a ser analisados, mas impede a análise mais profunda dos tipos de ligação executados, uma vez que os carbonos alfa situam-se, em geral, muito distantes das pontas dos aminoácidos e portanto "escondidos" na cadeia das proteínas.

Assumiu-se, portanto, que o cálculo dos contatos com os carbonos alfa seria uma aproximação rudimentar do cálculo dos contatos entre resíduos e átomos do ligante, ao invés de contatos entre átomos propriamente ditos. Baseando-se nessa suposição, os índices mencionados na análise dos dados referem-se aos resíduos, e não aos átomos propriamente ditos.

Além da filtragem dos carbonos alfa, notou-se uma grande quantidade de átomos de água nos arquivos PDB, provavelmente provenientes de um procedimento de solvatação. Todas as ocorrências de água também foram filtradas, restando apenas os heteroátomos pertencentes ao ligante principal.

Detalhes de Implementação

O software consiste de apenas 2 módulos Python, sua implementação sendo simples ao ponto da banalidade. Boa parte dessa simplicidade pode ser atribuída às inúmeras facilidades proporcionadas pela linguagem, em especial:

- Substring através de "slicing"
- Manipulação compreensão de listas
- Excelente suporte a expressões regulares

A seguir encontram-se as descrições das funções de cada módulo.

gontacter.py

O módulo principal é, essencialmente, o que se espera de um módulo "main" em um software qualquer. Este módulo apenas faz a interpretação dos argumentos de chamada e executa os passos da estratégia definida sequencialmente. Em especial:

- Obtém a lista de todos os arquivos pdb.gz num diretório
- Lê estes arquivos, inicializando as estruturas de dados implementadas
- Calcula a matriz de distâncias
- Imprime na tela os resultados obtidos
- Acumula os dados obtidos em matrizes de frequência individuais
- Imprime na tela os resultados acumulados

pdb.py

Este módulo contém quatro classes, sendo três delas apenas invólucros para os dados lidos dos arquivos PDB. Essas classes: Atom, HAtom e Contact, representam respectivamente um átomo da proteína, um heteroátomos do ligante e um contato estabelecido, e não possuem comportamentos.

A quarta classe, por outro lado, chamada de PDB, é responsável por toda a funcionalidade propriamente dita do software. Os métodos internos *_parse_atom e _parse_hatom* criam listas internas de Atom's e HAtom's que serão depois usadas para os cálculos, enquanto o método calc_distance_matrix é o responsável pela lógica de cálculo da matriz de distâncias. Algumas pequenas otimizações foram feitas para se tornar o software mais rápido, em especial as distâncias calculadas que mostram-se menores que o *cutoff* padrão são inseridas em tempo real numa lista de Contact's, evitando que se precise varrer a matriz novamente depois. Ao fim da execução, além da matriz com todas as distâncias entre carbonos alfa e átomos do ligante, obtém-se imediatamente a lista com todos os prováveis contatos.

Resultados Obtidos

A saída do software, em seu estado apresentado em sala de aula, é a seguinte:

```
Lendo arquivo 1PXI.pdb.gz
Lidos 294 carbonos alpha and 14 heteroatomos
Computando matriz de distancias...
Encontrados 32 contatos para valor de cutoff 6.0

Contatos encontrados:
################################################################################
CA  102 RES GLY:  13 =   3 CTS, 5.0747 AVG_D
CA  239 RES ALA:  31 =   5 CTS, 5.1923 AVG_D
CA  619 RES PHE:  82 =   4 CTS, 4.5251 AVG_D
CA  630 RES LEU:  83 =   5 CTS, 5.3773 AVG_D
CA 1027 RES ASN: 132 =   3 CTS, 4.6425 AVG_D
################################################################################

Lendo arquivo 1PXJ.pdb.gz
Lidos 294 carbonos alpha and 14 heteroatomos
Computando matriz de distancias...
Encontrados 33 contatos para valor de cutoff 6.0

Contatos encontrados:
################################################################################
CA  239 RES ALA:  31 =   7 CTS, 5.2494 AVG_D
```

```
CA  619 RES PHE:  82 =   6 CTS, 4.8314 AVG_D
CA  630 RES LEU:  83 =   5 CTS, 5.0886 AVG_D
CA 1027 RES ASN: 132 =   3 CTS, 4.8603 AVG_D
################################################################################

Lendo arquivo 1PXK.pdb.gz
Lidos 290 carbonos alpha and 17 heteroatomos
Computando matriz de distancias...
Encontrados 49 contatos para valor de cutoff 6.0

Contatos encontrados:
################################################################################
CA   81 RES ILE:  10 =   6 CTS, 5.4757 AVG_D
CA  239 RES ALA:  31 =   6 CTS, 5.5174 AVG_D
CA  592 RES PHE:  82 =   7 CTS, 4.8965 AVG_D
CA  603 RES LEU:  83 =   9 CTS, 5.0335 AVG_D
CA  611 RES HIS:  84 =   6 CTS, 4.5818 AVG_D
CA  621 RES GLN:  85 =   4 CTS, 4.6762 AVG_D
CA 1000 RES ASN: 132 =   3 CTS, 5.5724 AVG_D
CA 1090 RES ALA: 144 =   3 CTS, 5.2548 AVG_D
################################################################################

Lendo arquivo 1PXL.pdb.gz
Lidos 290 carbonos alpha and 24 heteroatomos
Computando matriz de distancias...
Encontrados 67 contatos para valor de cutoff 6.0

Contatos encontrados:
################################################################################
CA   81 RES ILE:  10 =   4 CTS, 5.5563 AVG_D
CA   89 RES GLY:  11 =   4 CTS, 5.8355 AVG_D
CA  239 RES ALA:  31 =   4 CTS, 5.2140 AVG_D
CA  592 RES PHE:  82 =   5 CTS, 4.7030 AVG_D
CA  603 RES LEU:  83 =   7 CTS, 4.8736 AVG_D
CA  611 RES HIS:  84 =   6 CTS, 5.0383 AVG_D
CA  621 RES GLN:  85 =  11 CTS, 4.9615 AVG_D
CA  630 RES ASP:  86 =   9 CTS, 5.4187 AVG_D
CA 1000 RES ASN: 132 =   3 CTS, 5.7284 AVG_D
CA 1090 RES ALA: 144 =   5 CTS, 5.3701 AVG_D
CA 1095 RES ASP: 145 =   3 CTS, 5.2534 AVG_D
################################################################################

Lendo arquivo 1PXM.pdb.gz
Lidos 291 carbonos alpha and 21 heteroatomos
Computando matriz de distancias...
Encontrados 59 contatos para valor de cutoff 6.0

Contatos encontrados:
################################################################################
CA   81 RES ILE:  10 =   4 CTS, 5.4578 AVG_D
CA  239 RES ALA:  31 =   6 CTS, 5.2955 AVG_D
CA  597 RES PHE:  82 =   5 CTS, 4.6805 AVG_D
CA  608 RES LEU:  83 =   7 CTS, 4.9298 AVG_D
CA  616 RES HIS:  84 =   4 CTS, 4.8285 AVG_D
CA  626 RES GLN:  85 =   7 CTS, 4.5393 AVG_D
CA  635 RES ASP:  86 =   6 CTS, 5.0249 AVG_D
CA 1005 RES ASN: 132 =   6 CTS, 5.2328 AVG_D
CA 1095 RES ALA: 144 =   5 CTS, 5.4055 AVG_D
CA 1100 RES ASP: 145 =   3 CTS, 5.5353 AVG_D
################################################################################

Lendo arquivo 1PXN.pdb.gz
Lidos 295 carbonos alpha and 22 heteroatomos
Computando matriz de distancias...
Encontrados 72 contatos para valor de cutoff 6.0

Contatos encontrados:
################################################################################
CA   81 RES ILE:  10 =   7 CTS, 5.4894 AVG_D
```

```
CA   89 RES GLY:  11 =   5 CTS, 4.8422 AVG_D
CA  136 RES VAL:  18 =   3 CTS, 5.7245 AVG_D
CA  239 RES ALA:  31 =   6 CTS, 5.4972 AVG_D
CA  628 RES PHE:  82 =   5 CTS, 4.6409 AVG_D
CA  639 RES LEU:  83 =   8 CTS, 4.9623 AVG_D
CA  647 RES HIS:  84 =   6 CTS, 5.1202 AVG_D
CA  657 RES GLN:  85 =   8 CTS, 4.4789 AVG_D
CA  666 RES ASP:  86 =   8 CTS, 5.0412 AVG_D
CA 1036 RES ASN: 132 =   7 CTS, 5.4355 AVG_D
################################################################################

Lendo arquivo 1PXO.pdb.gz
Lidos 295 carbonos alpha and 32 heteroatomos
Computando matriz de distancias...
Encontrados 94 contatos para valor de cutoff 6.0

Contatos encontrados:
################################################################################
CA   81 RES ILE:  10 =  14 CTS, 5.3160 AVG_D
CA   89 RES GLY:  11 =   3 CTS, 5.1626 AVG_D
CA  239 RES ALA:  31 =   6 CTS, 5.2611 AVG_D
CA  608 RES PHE:  80 =   3 CTS, 5.6479 AVG_D
CA  628 RES PHE:  82 =   7 CTS, 4.9957 AVG_D
CA  639 RES LEU:  83 =   9 CTS, 5.2498 AVG_D
CA  647 RES HIS:  84 =   9 CTS, 5.1725 AVG_D
CA  657 RES GLN:  85 =  15 CTS, 5.3355 AVG_D
CA  666 RES ASP:  86 =  10 CTS, 5.2118 AVG_D
CA 1027 RES GLN: 131 =   3 CTS, 4.8580 AVG_D
CA 1036 RES ASN: 132 =   5 CTS, 5.3518 AVG_D
CA 1126 RES ALA: 144 =   5 CTS, 4.8849 AVG_D
CA 1131 RES ASP: 145 =   3 CTS, 5.0794 AVG_D
################################################################################

Lendo arquivo 1PXP.pdb.gz
Lidos 295 carbonos alpha and 23 heteroatomos
Computando matriz de distancias...
Encontrados 65 contatos para valor de cutoff 6.0

Contatos encontrados:
################################################################################
CA   81 RES ILE:  10 =   8 CTS, 5.3629 AVG_D
CA  239 RES ALA:  31 =   5 CTS, 5.4886 AVG_D
CA  628 RES PHE:  82 =   5 CTS, 5.0251 AVG_D
CA  639 RES LEU:  83 =   7 CTS, 5.0324 AVG_D
CA  647 RES HIS:  84 =   4 CTS, 4.8269 AVG_D
CA  657 RES GLN:  85 =   9 CTS, 4.9235 AVG_D
CA  666 RES ASP:  86 =   7 CTS, 5.4430 AVG_D
CA 1036 RES ASN: 132 =   6 CTS, 5.3695 AVG_D
CA 1126 RES ALA: 144 =   5 CTS, 5.5452 AVG_D
CA 1131 RES ASP: 145 =   3 CTS, 5.5491 AVG_D
################################################################################
```

Como pode-se notar, o programa analisa cada arquivo PDB separadamente, imprimindo a quantidade de carbonos alfa encontrados, bem como de heteroátomos. A partir dessa saída, podemos inferir que há diferenças entre as proteínas fornecidas, ainda que sutis, e os ligantes são bastante diferentes.

Um comentário em tempo: poder-se-ia, possivelmente, obter melhores resultados se as proteínas houvessem sido alinhadas entre si antes de feita a análise. Isso garantiria que a coincidência de resíduos em determinadas posições referir-se-ia aos mesmos resíduos, e não a coincidências de posição.

As listas de contatos encontrados mostradas estão acumuladas. A seguinte linha, exemplificando:

```
CA  657 RES GLN:  85 =  15 CTS, 5.3355 AVG_D
```

Significa que o carbono alfa 666, pertencente ao resíduo Glicina na posição 85, executou 15 contatos com o ligante fornecido, com uma média de distância de 5.3355 Angstrons.

A fim de obter os resultados mostrados acima, algumas restrições foram aplicadas. Em especial, ao imprimir os contatos acumulados, descartou-se os resíduos que faziam apenas um ou dois contatos, a fim de facilitar a interpretação.

A análise obtida foi a seguinte:

```
 10|ILE fez  46 contatos em  8 arquivos
 11|GLY fez  20 contatos em  7 arquivos
 12|GLU fez   4 contatos em  2 arquivos
 13|GLY fez  10 contatos em  5 arquivos
 18|VAL fez   9 contatos em  4 arquivos
 31|ALA fez  46 contatos em  8 arquivos
 80|PHE fez  13 contatos em  7 arquivos
 81|GLU fez   9 contatos em  7 arquivos
 82|PHE fez  45 contatos em  8 arquivos
 83|LEU fez  58 contatos em  8 arquivos
 84|HIS fez  37 contatos em  7 arquivos
 85|GLN fez  55 contatos em  6 arquivos
 86|ASP fez  42 contatos em  6 arquivos
 89|LYS fez   3 contatos em  1 arquivos
131|GLN fez  11 contatos em  6 arquivos
132|ASN fez  37 contatos em  8 arquivos
144|ALA fez  29 contatos em  8 arquivos
145|ASP fez  15 contatos em  5 arquivos
```

Dessa análise, podemos inferir que:

- ILE 10
- ALA 31
- PHE 82
- LEU 83
- ASN 132
- ALA 144

São provavelmente os contatos mais importantes, apresentando tendência a fazer parte da estrutura ou função da proteína, já que aparecem em todos os arquivos fazendo contatos nas mesmas posições.

Além desses, os seguintes:

- GLY 11
- PHE 80

- GLU 81
- HIS 84

Também são bons candidatos a fazer parte de estruturas chave da proteína.

Em especial, a aparição dos resíduos contíguos de 81 a 86 indica uma forte tendência de que essa porção da proteína esteja intimamente ligada à manutenção de sua estrutura ou função.

Como exceções:

- GLU 12
- LYS 89

Fizeram um número muito pequeno de contatos e só apareceram em um arquivo cada, indicando que são, possivelmente, mutações não conservativas na estrutura das proteínas, que podem ser responsáveis por modificar sua função tanto para melhor: possibilitando o melhor encaixe de um ligante específico, quanto para pior, dificultando a ligação de um ligante útil.