UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
Instituto de Biofísica Carlos Chagas Filho
Concurso para Professor Adjunto
MS-020 – Biologia Computacional
Edital nº 953, de 20 de dezembro de 2019, publicado no DOU nº 248, de 24 de dezembro de 2019 - consolidado com as alterações dos Editais nº 9, de 9 de janeiro de 2020, nº 31, de 03 de fevereiro de 2020, nº 48, de 11 de fevereiro de 2020 e nº 116 de 25 de março de 2020

NOME: Maria Fernanda Ribeiro Dias

Tema 3 - Ferramentas estatísticas em Biologia Computacional

A Biologia Computacional se caracteriza pela elaboração de algoritmos matemáticos / computacionais para solucionar um problema biológico. A implementação desses algoritmos pode ser escrita em diferentes linguagens e para cada algoritmo é necessário métricas e programas estatísticos que visam avaliar o desempenho do mesmo.

Dentro da modelagem comparativa podemos citar o RMSD que avalia a diferença na posição dos átomos entre duas moléculas (modelo e molde), sendo útil para saber se o modelo criado segue parâmetros coerentes e pode ser utilizado em etapas seguintes na geração de uma proteína por exemplo.

Podemos ainda acrescentar à essa análise o gráfico de Ramachandran, que avalia os ângulos (PHI) Φ e (PSI) Ψ de um polipeptídeo, permitindo assim avaliar sua este

quiometria, por meio de todas as combinações possíveis desses ângulos

As duas métricas citadas são utilizadas em Biologia Estrutural e limitam-se a avaliação de estruturas 3D. Essas análises podem ser feitas usando a ferramenta PyMOL, por exemplo, um programa para visualização de moléculas 3D, ou ainda programas para análise de ancoramento molecular como o DoCking (AutoDock Vina).

Uma outra análise seria a análise de correlação entre duas variáveis, essa pode ser utilizada quando temos um conjunto de dados sendo representados por diferentes atributos, por exemplo, propriedades físico-químicas e se faz necessário avaliar se esses atributos estão diretamente ou inversamente correlacionados. Para essa análise podemos usar a correlação de Pearson, por exemplo, onde os atributos par a par são comparados, e gerados valores entre -1 e 1, valores próximos a -1 indica que os atributos se correlacionam de forma negativa; valores próximos a 1 indica que estes se correlacionam de forma positiva, logo um deles pode ser eliminado da representação do conjunto de dados, sem gerar danos para os mesmos; valores próximos a 0 indica que os atributos são independentes. Essas análises são utilizadas em problemas cujo

conjunto de dados apresenta alta dimensão, é necessário uma redução dimensional em sua representação. Problemas biológicos aplicados à análise por Inteligência Artificial necessitam passar por essa etapa de pré-processamento. Para fazer essas análises é possível utilizar a ferramenta R, um pacote estatístico com interface gráfica, que permite ao usuário utilizar bibliotecas prontas ou implementar algoritmos direcionados para o conjunto de dados em questão.

Ainda dentro da Inteligência Artificial, especificamente uma área de Aprendizagem de Máquina, diversas ferramentas estatísticas são usadas. Em Aprendizagem de máquina não Supervisionada, quando não se tem informações prévias sobre o conjunto de dados, pode-se usar métrica de distância, como análise de distância Euclidiana para separar os dados e extrair informações dos mesmo. Já a técnica de Aprendizagem supervisionada apresenta uma gama de ferramentas, implementadas em C, ou java ou PYTHON, para avaliar o desempenho de um classificador ou modelos de classificação que utilizaram diferentes parâmetros ou algoritmos.

Suponhamos que temos um algoritmo de predição cujo foco é classificar possíveis ligantes com potencial para inibição enzimática.

As análises estatísticas para avaliação de desempenho desse classificador iniciam com um processo de validação Cruzada, que utiliza o que chamamos de conjunto teste (ligantes e decoys, conhecidos e retirados de banco de dados). O procino de validação cruzada pode ocorrer de deferentes formas, uma delas é o uso K-fold-cross-validation onde utiliza-se os dados divididos em K partes, usando K-1 como conjunto de treinamento e 1 parte é separada como teste. Aplicando o classificador nesse conjunto de dados temos a formação do que chamamos de matriz de confusão, constando de: VP → Verdadeiros positivos; VN → verdadeiros negativos → e FP → falsos positivos; FN → falsos negativos; Esses constam de dados cujo classificador faz a predição de forma correta e dados que são preditos de forma errônea pelo classificador, respectivamente. Diante dessa classificação, podemos extrair métricas estatísticas como Recall, Especificidade e acurácia.

Recall → $\frac{VP}{VP+FN}$ → Avalia a quantidade de dados preditos como positivos diante de todos os positivos.

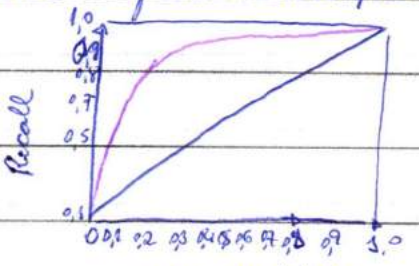
Especificidade → $\frac{VN}{VN+FP}$ → Avalia a quantidade de dados preditos como negativos diante todos os dados negativos.

Acurácia → $\frac{VP+VN}{VP+FP+VN+FN}$ Avalia a quantidade de dados preditos de forma correta diante o conjunto de todos os dados.

NOME: Nº Elias

No. fl. 5/19

As métricas de Recall e especificidade permitem a elaboração da análise de curva ROC (constitui análise global de desempenho do classificador). Essa análise traça um comparativo entre os classificações corretas de dados positivos e negativos, assim dizemos que a área sob a curva (AUC) demostra o desempenho de classificação em um determinado problema, Dado o exemplo do gráfico a seguir; podemos dizer que o classificador apresenta AUC : 0,90 na classificação de dados positivos.



Além da curva Roc é necessário ainda citar o fator de Enriquecimento que avalia a quantidade de dados classificados, com um determinado score a partir de um ponto de corte.

Embora essas métricas sejam muito utilizadas, é necessário que essas não sejam avaliadas sozinhas, mas em parceria com análise acurada do conjunto de dados. Em problemas biológicos é comum que tenhamos muitos dados relacionados a uma única classe, o que constitui o que chamamos de dados desbalanceados. Esse tipo de dado gera o que chamamos de overfitting (quando o preditor se adapta ao conjunto de dados específico, sendo muito eficiente no conjunto teste e apresentando valores de AUC próximos a 0,5 ao serem aplicados em conjuntos des

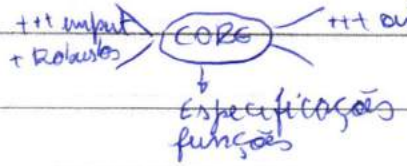
conhecidos. Esse problema pode ser tratado com a geração de dados hipotéticos no conjunto com menor quantidade de dados com a técnica SMOTE, por exemplo. Em minha tese de doutorado, essas métricas foram diversamente aplicadas e utilizados para o encontro de possíveis inibidores de proteases, utilizando como conjunto de dados de treinamento ligantes oriundos do banco de dados DUD e substratos naturais da protease do HIV, assim como peptídeos miméticos. O foco principal era focar na representação desses dados utilizando suas propriedades físico-químicas e analisar com diversas ferramentas estatísticas o quão eficaz eram essas representações e como elas podem ser refinadas, utilizando métodos matemáticos.

Tema 4 - Modelagem de Redes Metabólicas

A análise de Redes Metabólicas constitui uma subárea da Biologia de Sistemas e centra-se na análise de vias, formadas por reações químicas (de degradação ou síntese) que visam entender o funcionamento de um organismo ou complexo biológico. Esta área veio para suprir o seguinte questionamento "Por que precisamos olhar para apenas um componente (gene, proteína, DNA), se podemos olhar

para o Todo?" Essa pergunta permite que as pesquisas biológicas tomem novos rumos que se contrapõem à biologia reducionista e abrange análises um problema ou um organismo como um todo. Assim podemos citar duas metodologias; Top-down e bottom-up. Metodologias do tipo Top down geram um fluxo descendente de informação, onde dados de genômica ou transcriptômica são utilizados por exemplo para nutrir uma rede. já nas metodologias do tipo bottom-up, há um fluxo ascendente de informações, onde dados de metabólitos são utilizados para nutrir e montar a rede metabólica.

A montagem de redes metabólicas necessita de modelos matemáticos/computacionais que sejam robustos (avalia a estabilidade do método/funcionalidade, mesmo tendo condições de perturbação) e degenerativos (redundancia funcional). Uma arquitetura viável é do tipo bow-tie que se assemelha a a gravata borboleta, onde as entradas (várias) passam por um core especializado e menos degenerado



O protocolo para modelagem de redes metabólicas foi estabelecido em 2010 por Thiele e Palsson. Esse protocolo conta com 4 etapas e surgiu a partir de um genoma. São elas:

NOME: Wallas

No. fl. 8/14

1) Draft → Constitui a formação da rede utilizando dados disponíveis em bancos de dados Globais, como o KEGG. Essa etapa se constitui de um processo automático;

2) Curadoria manual → Consiste no uso de especialistas avaliando todas as reações presentes na rede, para isso é necessário conhecimento técnico e biológico sobre a via que se quer modelar, além da confirmação com dados descritos na literatura.

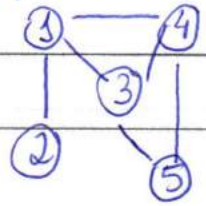
3) Aplicação de modelos matemáticos → Consiste na busca de ferramentas e métodos que melhor se aplicam na formação da rede. Análisa o escopo e abrangência da rede. Conversão computacional.

4) Avaliação da rede → Consiste na avaliação curada de gaps presentes na rede e comparação das rede que esta sendo formada com redes já formadas.

Todas as 4 etapas são interativas, podendo ser revisitadas e reavaliadas sempre que necesário, podendo fazer, por exemplo a inserção de informações em gap.

Uma forma de representar uma rede metabólica é por meio de grafos ou matrizes estequiométricas. Aplicando teoria dos grafos temos que $G = \{V, \epsilon\}$ onde V são Vértices que podem ser genes, proteínas etc e ϵ constitui arestas. Assim para um grafo G, temos: $V = \{1,2,3,4,5\}$

$\epsilon = \{(1,2), (1,3), (1,4), (4,5), (5,3), (3,4)\}$



Assim, cada vértice corresponde a um nó e cada linha corresponde a ~~conde~~ arestas ou interação entre um nó e outro. Supondo que cada nó seja uma proteína, ou uma reação química, podemos ter uma análise de fluxo direcional, avaliando os substratos gerados em determinadas reações.

A utilização de matrizes estequiométricas conta com análises de restrição onde cada reação forma uma coluna com índices que corresponde a valores positivos e negativos, sendo ~~produtos~~ os reagentes (-) e produtos (+), respectivamente.

Além dos gaps, um outro componente que se deve avaliar nas redes metabólicas são fatores externos, por exemplo fatores ambientais. Como o processo de produção é iterativo, a cada informação ou atualização da rede, um novo nó ou informação pode ser acrescentado e assim a rede metabólica é atualizada.

Uma das ferramentas muito utilizadas na implementação de redes é a COBRA (contrast-based Reconstructions Robusta Analysis), uma ferramenta que foi implementada em MATLAB, mas atualmente já apresenta implementação em Python. COBRA oferece ainda ferramentas de avaliação da rede, por exemplo ao modelar uma rede metabólica de procariotos, ~~erros~~

na avaliação o grau de i nó (K) indica quantas conexões a o mesmo possui com outros nós, sendo P(K) a destribuição probabilística desses nós em um grafo.

Por ser um protocolo isterativo, temos os modelos criados baseados em restrições, por meio de imposições de restrições físico-químicas e identificação de pressões seletivas, por exemplo fatores ambientais e o fluxo de reações pode ser representado por escalas estequiometricas, utilizando ainda otimização linear, baseada em estado estacionário, sem levar em conta parâmetros cinéticos.

Para análise da rede, podemos levar em conta ainda análise de modularidade, ontologia gênica e centralidade que constituem respectivamente em módulos (como subconjunto de redes) padronizados que juntamente com localização e funcionalidade celular constitui ontogia gênica, baseando ainda a análise de centralidade em métricas locais de grafos.

Tema 9. Fundamentos Teóricos de Algoritmos de Biologia Computacional e Bioinformática

Para embasar esse tema gostaria de citar Margarit Dayoff, que no final do

século XX, mais especificamente na decada de 70, iniciou a partir de análise comparativa de sequências de proteínas curadas manualmente, a produção de uma matriz de substituição para mutações, a matriz PAM. Margaret era matemática e utilizou para isso o conjunto de 1680 proteínas que compunham o 1º atlas proteicos, todas curadas manualmente pela própria Margaret. O que Margaret fez foi criar matrizes em que comparava duas sequências proteicas S_1 e S_2 e estabelecia pontuações para cada mutação encontrada, assim tinhamos os valores de 0 na diagonal das matrizes e de acordo com a taxa de mutação dos aminoácidos, esses iam ganhando valores mais positivos quando apresentavam altas taxas de mutação e negativos quando apresentavam valores baixo. Partindo das sequências S_1 e S_2, podemos dizer que ambas as sequências são 1 PAM divergentes se S_1 puder ser substituída por S_2, sem danos, ocorrendo uma mutação a cada 100 resíduos de aminoácidos. Ao fazer essas comparações, Margaret ainda gerou árvores de substituição em cada sequência e ia anotando as mutações ocorridos. Isso gerou duas métricas i) mutabilidade relativa; ii) frequência relativa; que corresponde ao número de vezes que um aminoácido foi substituído por outro e na posição do mesmo; e a

frequência de substituição, na qual foi gerado uma tabela com valores padrão para cada aminoácido. nessa tabela é possível observar que o triptofano sofre pouquíssima mutação, e essa informação é importante até hoje quando avaliamos sítios de interação entre proteínas de forma computacional pois a substituição deste aminoácido geralmente leva à inativação da proteína, por ser um aminoácido grande e interferir na estrutura proteica. A matriz PAM foi calculada em diferentes níveis tendo hoje PAM 30, PAM 70 e até 250 PAM.

20 anos após a formação dessas matrizes, surgia a matriz BLOSUM, gerada na década de 90 pelo casal Henikoff. As matrizes BLOSUM também eram calculadas a partir da comparação de sequências biológicas de proteínas utilizando pontuação, porém essa comparação era feita em blocos e utilizava sequências mais distantes e dissimilares. Tanto a matriz PAM, quanto a BLOSUM são utilizadas em algoritmos de alinhamento como o BLAST, Essas constituem métricas importantes para análise comparativa de genoma, por exemplo na comparação do genoma do SARS-COV-2 com outros coronavírus, inferindo informações acerca da origem

do vírus, assim como similaridades com outros organismos.

Outra base que fundamenta metodologias computacionais para aplicações em biologia é o uso de ferramentas que simulam questões biológicas a nível estrutural, como o uso de algoritmos genéticos, utilizados em algoritmos de docking molecular, esses são considerados algoritmos evolutivos por apresentarem abordagens darwinistas. Os algoritmos evolutivos surgiram por John Holland e são algoritmos estocásticos que no docking, permitiram o enriquecimento da técnica por permitir a inclusão de mais graus de liberdade relacionados à flexibilidade molecular. Esses algoritmos são otimizados por meio do que chamamos de mutações e recombinações e assim, inicialmente temos um cromossomo com possíveis soluções para o problema; esse vai sofrendo cross over e sendo apresentadas novas soluções. Essas soluções e a otimização do algoritmo podem ser avaliadas por meio dos valores de energia do docking ou da análise de curva ROC, já descrito no item 3.

Além dos ~~anteriores~~ algoritmos que deram início à análise de sequências lineares e análises estruturais, gostaria ainda de citar os algoritmos baseados em Inteligência Artificial (IA), área da computação, que desde o início do

século XXI com a intensificação de ~~dado~~ geração de dados vem sendo aplicada em problemas biológicos. Essa aplicação se intensifica conforme surgem mais processamentos e melhorias nas técnicas "ômicas". Um dos primeiros trabalhos a utilizar técnicas de IA, aprendizagem de maquina foi do Golub, em 1999 que utilizou análise de cluster como K-means e técnicas baseadas em distância como a euclidiana para separar dados de células de origem linfóide e mielóide. As análises feitas por Golub utilizaram uma plataforma chamada Genepattern, essa plataforma permite análise de genes utilizando técnicas de aprendizagem supervisionada e não supervisionada.

Na área de aprendizagem supervisionada, algoritmos como árvore de decisão foram aplicados para classificar sítios de ubiquitinação, utilizando informações físico-químicas de proteínas. Este projeto ocorreu na primeira década do século XXI e se constitui em um dos primeiros trabalhos desenvolvidos com aplicação de metodologia de inteligência Artificial em problemas biológicos. Hoje, temos uma gama de algoritmos como SVM, redes neurais e sua evolução para aplicação e resolução de problemas biológicos, sendo usados na avaliação de redes metabólicas e até mesmo na elucidação da estrutura terciária de uma proteína.