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: Graciela Maria Dias

No. fl. 01

Topico 4 - Modelagem de redes biológicas

As redes biológicas surgiram com a necessidade de entender as interações entre os diversos níveis de complexidade do organismo. As redes biológicas são sistemas complexos e possui algumas propriedades: adaptação e auto organização. Portanto possuem características que permitem entender de uma forma mais clara todos os comados que a envolvem. Por exemplo, existem diversos tipos de redes, as sinalizadoras, regulatórias e metabólicas, cada uma com sua característica intrínseca. Nas redes sinalizadoras é importante observar, entender quais os sinais (muitos vezes vindo do ambiente) são necessários para ativar um cascata de sinais no interior da célula. Nas redes regulatórias é importante entender quais fatores de transcrição são responsáveis pelo início da transcrição e por fim e não menos importante as redes metabólicas/bioquímicas, onde é possível observar os substratos

e as reações químicas [redacted] pondo possível observar a massa celular e a taxa de crescimento da célula.

Com o advento das plataformas de nova geração e um grande acúmulo de dados e questões a serem entendidas surge a Biologia de Sistemas, junto com essa área, surge também um aprimoramento das técnicas matemáticas e computacionais. Essas serão essenciais para o entendimento das redes biológicas, pois elas fornecem um grau de abstração necessário. Por exemplo, a teoria de grafos tem sido amplamente usada para especificar uma rede, os nós geralmente representam os genes/proteínas e as arestas as reações químicas ou interações. [redacted] Para simular, prever o que acontece nas redes biológicas é necessário modelar [redacted] através dos métodos matemáticos e computacionais. Existem três tipos de modelos que irão permitir estipular hipóteses e observar o que acontece nessas redes, avaliando por exemplo a robustez da rede, ou seja, quais nós e arestas (genes/ptn's – interações) são essenciais p/ o funcionamento da célula, um exemplo clássico de modelagem de redes é observar o comportamento das mutações em redes

associados com câncer. Os modelos são definidos como lógico, contínuo e estocásticos. Esses modelos geralmente estão associados às redes regulatórias, enquanto que para as redes metabólicas, existem algumas diferenças, por exemplo uma técnica bastante utilizada para as redes metabólicas é o FBA, análise de balanço de fluxo que também serão descritos a seguir.

Os modelos lógicos consistem em uma análise qualitativa das redes. Observa-se os eventos disparados que ocorrem na célula em um instante t. Esses modelos podem ser usado para acompanhar o desenvolvimento embrionário de algum organismo, foi usado com sucesso nos ouriços-do-mar.

Três tipos de modelo são classificados como lógico: rede booleana, redes lógicas e petrinet.

As redes booleanas são considerados os modelos mais simples, elas assumem somente 2 estados: inativo e ativo, ou seja, as interações irão ocorrer através de uma função booleana. Outra característica é que que todas as interações ocorrem de forma síncrona, todos os eventos ocorrem de maneira simultânea para todas as variáveis. Embora, esse método tenha limitações,

NOME: Graciela Maria Dias

No. fl. 04

diversos autores da comunidade científica têm mostrado sucesso em diversos modelos. Esse modelo é bem estudado em leveduras, principalmente na regulação do ciclo celular. Uma das limitações é seu alto custo computacional e o número mínimo de conexões, característica bastante distante da realidade, onde um gene/proteína pode ter 20 conexões.

O segundo modelo é denominado como lógico, ele também é considerado simples, como as redes booleanas trata de eventos discretos, mas com a diferença que consegue simular mais que 2 conexões, tornando o modelo mais sofisticado. Outra diferença é que as mudanças de estados pode ser feito de forma assincrona.

O terceiro modelo, [redacted] petri net, apresenta diferenças bastante relevantes quando comparados as redes booleanas e lógicas. Esse modelo pode ser usado para as redes metabólicas e regulatórias. Esse modelo é representado por um grafo bipartido, com componentes chamados lugares, transições e arcos. Geralmente, os lugares constituem os genes/ptn's e as transições as reações bioquímicas. É também caracterizado por ser um modelo não determinístico

Ex: regulação da via de esporulação em B. subtilis

NOME: Graciela Maria Dias

No. fl. 05

A simulação dinâmica é dirigida pelas transições e pelos atributos dos lugares, chamado de "fichas". Existem os disparos das transições onde serão retiradas fichas de um lugar de entrada e dispostas em outro lugar. Existe também várias extensões desse modelo que diminuem a simulação das redes, ex: redes petri estendidas híbridas e estocástica.

A segunda classificação dos modelos é denominado contínuos. Os modelos contínuos diferem dos lógicos por tratarem de variáveis contínuas e não discretas, havendo algumas limitações para o uso do modelo, como por exemplo a introdução de parâmetros cinéticos são essenciais para o sucesso do modelo. Os modelos mais clássicos que compõe essa classificação são as equações diferenciais ordinárias (EDO) Matematicamente ela trata de funções de uma única variável independente e suas derivadas.

Um dos primeiros aplicações da EDO foi o estudo da regulação do operon lac. Outros exemplos também podem ser citados: regulação do ciclo celular da bactéria Caulobacter crescentus, conhecida por apresentar 2 células morfologicamente diferentes e também é

bastante utilizado para estudo o ciclo celular em leveduras.

Outro método contínuo bastante utilizado é o rFBA (analise de balanço de fluxos regulado. Esse método veio para compensar a falta de dados experimentais para a aplicação da EDO. Assumindo que a rede se encontra no estado estacionário, é possível transformar as EDO em equações lineares e seus todos obtidos por programação linear otimizando a função objetiva.

O FBA original é usada para as redes metabólicas, e baseado em uma matriz estequiometrica, e a partir desses pressupostos, estado estacionario, e possível observar o fluxo das reações. Esse método tem sido bastante utilizado para avaliar a taxa de crescimento, concentrações de substratos. Uma diferença bastante interessante é que o rFBA adiciona entidades de regulação booleana, tornando um modelo lógico e contínuo. Esses métodos são bastante utilizados em E. coli.

Por fim, existe os modelos estocásticos, são modelos ainda pouco usados, embora pela clara os efeitos estocásticos recorrentes nos redes biologicos. Um dos métodos/ algo-

ritmo que tem sido utilizados é Simulação, estocástico Gillespie (SSA). Esses modelos são de difícil aplicação pois são aplicados à cada reação individualmente. O SSA foi utilizado p/ estudar a regulação do fago considerando a fase lítica e lisogênica.

Além da descrição dos 3 métodos, outra abordagem que tem sido utilizada com sucesso são as técnicas de aprendizado de máquina. Um ex é identificação de redes reguladoras envolvido c/ a produção de antibióticos.

NOME: Graciela Maria Dias

No. fl. 01

Tópico III - Ferramentas estatísticas em biologia computacional.

Com o advento das plataformas de nova geração, a cada dia dados novos são gerados, criando um grande desafio para biologia computacional: armazenar, minerar e verificar se os dados são significativos e conseguem explicar com confiabilidade as perguntas biológicas.

Para fazer inferências de uma determinada população, é preciso ter uma amostragem dessa população que por sua vez serão analisados pela estatística. Existe 2 abordagens da estatística: a descritiva e inferencial.

A descritiva é o primeiro passo para as análises, é possível a partir delas, fazer uma análise exploratória e uma sumarização dos dados. As principais medidas que permitem essa análise exploratória são as medidas de resumo, variância e forma.

As medidas de resumo são a média, mediana e moda. A média é uma das medidas mais famosas e utilizadas, com ela é possível observar um resumo dos valores do tamanho amostral, entretanto não reconhe com clareza, a presença dos outliers

NOME: Graciela Maria Dias

No. fl. 02

A mediana é a posição central de uma amostragem, para se obter esse valor é preciso ordenar e dividir a amostra, é indicado para dados com muito outliers. A moda é utilizado com frequência em dados categóricos e verifica o número de vezes que determinada categoria aparece. Outras medidas importantes são a variância que vai medir o quanto os dados estão dispersos com relação a média e o desvio padrão que consiste na raiz quadrada da variância, tornando a variância interpretável, já que ele é adimensional. As medidas-forma também são importantes, para verificar o comportamento da curva em um gráfico, observando a simetria.

Os gráficos, tambem são uma importante ferramenta para a análise exploratória, um dos mais completos é o boxplot, onde é possível verificar pelo menos 5 medidas citadas acima.

Na biologia computacional é de extrema importância essa análise exploratória, diante da infinidade de dados que as técnicas são capazes de produzir, por exemplo microarranjos e RNAseq.

Antes de fazer a inferência estatística, ou seja, levantar hipóteses e aplicar testes p/ avaliar se os dados são significativos ou

ou não, é preciso muitas vezes reduzir a dimensionalidade. Essa redução pode ser feita pela análise principal de componente (PCA) e PCOA (análise dos coordenados principais). Esses métodos são frequentemente usados como etapas de pré-processamento e controle de qualidade dos dados oriundos de RNA seq e microarranjo.

Após essa etapa, é possível fazer a inferência estatística, com o levantamento de hipóteses é inerente que erros ocorrem. Existem 2 tipos de erros, erro tipo I, que ocorre quando rejeita a H_0, sendo verdadeira, esse erro é denominado α, nível de significância, e o erro tipo II que ocorre quando se aceita a H_0, sendo ela falsa (β). Após uma escolha do nível de significância, ou seja, o quanto de erro vai ser tolerado (nível mais comum 0,05), é feito o cálculo do p-valor, ou seja, avalia o tão quão distante o valor [redacted] da minha amostra esta do valor esperado. Com esses parâmetros, é possível assim aceitar ou rejeitar a hipótese nula.

Geralmente, na biologia os testes são aplicados para observar por exemplo, se existe diferenças significativas em padrões da expressão gênica, diferença na comuni-

dade microbiana, dados que podem ser obtidos da transcriptoma ou metagenômica.

Um exemplo onde é possível aplicar os testes estatísticos, é por exemplo observar as diferenças na comunidade microbiana ou perfil funcional. de metagenomas isolados de ambientes contaminados com petróleo e áreas que não sofreram o impacto, atualmente estou fazendo essas análises. Como na transcriptoma, as análises de redução de dimensionalidade são essenciais, além do PCA e PCoA (este, diferente do PCA, trata de uma distância de dissimilaridade(s)), tem também o NMDS (escalonamento métrico não dimensional), este não assume uma linearidade dos dados, e usa uma distância de Bray Curtis, e tenta a medida que adiciona novas amostras, reduzir o espaço entre as variáveis.

Outro exemplo de um trabalho que também estou envolvida, é avalio através da Beta diversidade, quais as variáveis física química explicam determinado grupo taxonômico, obtido do sequenciamento do amplicon 16S.

Em geral, para a aplicação dos testes, é necessário observar a distribuição do dados, ou seja, se é normal ou não

Se os dados seguem distribuição normal, os testes aplicados são os paramétricos. Um teste bastante utilizado p/ verificar a dist. normal é o Shapiro-Wilker. Outro pressuposto e avaliar a homocedasticidade, ou seja, se existe ou não variância entre as amostras, o teste aplicado é o Levene.

Na maioria das vezes, os dados de expressão gênica não seguem distribuição normal, necessitando de testes não paramétricos, ou manipulações no pré-processamento, como a normalização, ou transformação (ex.: transformar os dados em uma distribuição logaritmica normal).

Os testes paramétricos compreendem no teste-t, teste-t pareado (p/ grupos dependentes), teste-t Welch (caso não atinja o pressuposto da homocedasticidade). Esses testes são utilizados para comparar 2 grupos.

Para três grupos ou mais, utiliza-se o teste ANOVA, através da razão F que avalia a variância inter grupo e intra grupos.

Os testes não paramétricos (TNP) p/ 2 grupos independentes é o teste U Mann Whitney, para grupos pareados, teste Wilcoxon e para três grupos ou mais Kruskal-Wallis. Todos os TNP são baseados no ranqueamento

e soma dos postos.

Outros testes que são essenciais na bioloqua, são os testes de correção, devido a análise simultânea de diversos variáveis e um número limitado de amostras, quando um número alto de falsos positivos. Os testes mais comuns são o teste de Bonferroni e FDR. O teste de Bonferroni, é um dos testes mais simples, são indicados para um pequeno número de amostras, já que é bastante conservador, ou seja, reconhece os falsos positivos, mas também descarta os verdadeiros positivos.

O FDR é mais indicado p/ um nº maior de amostras, pois ele reconhece a proporção de falsos descobertos no grupo de Ho rejeitados, sendo mais eficaz no ajuste.

NOME: Graciela Maria Dias

No. fl. 01

Tópico 09 - Fundamentos teóricos de algoritmos de biologia computacional e bioinformática

Os algoritmos são um conjunto de tarefas dadas ao computador, de forma que elas possam ser interpretadas e computadas em um determinado tempo. Uma propriedade essencial que deve ser avaliada é a complexidade e eficiência. Por exemplo, algoritmos $n!$ e 2^n são proibitivos quando se tem uma grande quantidade de dados. Portanto, além de avaliar a execução do alg, a solução do problema, é necessário avaliar o tempo de execução.

Os algoritmos computacionais não foram inventados p/ solucionar os problemas biológicos, pelo contrário, os algoritmos na biologia, já foram aplicados em diversos tipos de problemas computacionais e em outras áreas, como a física e astronomia.

Os principais algoritmos usado na biologia computacional são força bruta, programação dinâmica, algoritmos gulos e randomizados.

O algoritmo de força bruta tenta buscar todas as soluções possíveis para um problema. Uma aplicação desse algoritmo é a busca dos NRPS, ex. a tirocidina

produzido por Bacillus stratis. Identificá-lo através da Espectometria de massa, é bastante dispendioso, portanto através de um perfil massa/carga, tenta se achar os espectros que possam identificar o peptídeo em questão. Um algoritmo mais sofisticado p/ tal análise é branch and bound.

Outro algoritmo essencial na biologia, é a programação dinâmica (P.D), ele busca a melhor solução dentro de sub-problemas ótimos, ou seja, ele busca a divisão do problema, resolvendo por partes de modo que não seja possível retornar ao problema anterior, tornando a resolução do problema eficaz. Os dois algoritmos que são baseados na P.D é o Needleman-Wusch e Smith-Waterman, os mesmos são responsáveis pelo alinhamento global e local, respectivamente.

Para a construção do alinhamento é necessário uma construção de matriz, onde vai ser computados matches, mismatch e gaps. Na Biologia as matrizes de pontuação PAM e Blossum são essenciais para o alinhamento, as informações das propriedades físicas-químicas são computadas.

Outro tipo de alinhamento possível é o alinhamento multiplo, sendo inviável o uso de um matriz di-

mensional, gerando uma estratégia de alinhamento progressivo, onde é feito o alinhamento par-a-par e depois construído uma arvore guia, no qual vai ser computado os melhores alinhamentos. Uma dos programas mais utilizados e estados é o clustal.

Um exemplo bastante comum na minha área de atuação é fazer o alinhamento de genes housekeeping a fim de identificar um perfil taxonômico de diversos espécies de microorganismos, ou através da genômica comparativa, encontrar o core genoma de um determinado gênero ou linhagens diferentes da mesma espécie, os genes identificados como core-genoma são alinhados para uma posterior inferência filogenética. Na filogenia, diversos algoritmos são utilizados, de forma que eles podem ser exatos, ou seja, testam todos os árvores possíveis, ou são heurísticos, no qual se tenta encontrar uma melhor solução possível sem avaliar todas as alternativas. Os algoritmos mais populares são o UGMA (embora tenha a limitação de manter a relogio molecular constante) e o Neighbor-Joining.

Para o alinhamento de sequencias,

NOME: Graciela Maria Dias

No. fl. 04

o BLAST é uma das ferramentas mais utilizadas da biologia computacional, ele abriga duas os algoritmos, incluindo o Smith-Waterman tornando uma busca heuristica das sequencias.

Os algoritmos que usam a teoria dos grafos são utilizados na montagem dos genomas, eles podem ser de 3 tipos gulosos, OLC (overlap-consensus e layout) e de Bruijn.

O tipo guloso representa o grafo no seu modo mais simples, reconhe os overlaps na extremidades e ocorre a sobreposição, um exemplo de ferramenta que utiliza essa abordagem é o CAP3 e SSAKE.

O OLC foi proposto p/ montar dados oriundos do 454, utilizam o caminho hamiltoniano o que torna o problema NP-difícil. Nesse metodo os nós são os reads e os arestas os sobreposições. Ex: Mira, Newbler

O grafo de Bruijn usa a estrategia de K-mers, onde os K-mers são os nós e os arestas os K-mers-1. O caminho feito é o euleriano, ou seja, é preciso passar pelos arestas uma unica vez, enquanto o hamiltoniano, o caminho é feito pelos nós. O caminho euleriano torna o problema muito mais fácil

NOME: Graciela Maria Dias

No. fl. 05

Por fim, existe os algoritmos randomizados que são aplicados para um problema bastante difícil na Biologia, encontrar os motivos.

O número de variáveis são tão grandes que se torna inviável prevê-los, portanto a introdução dos métodos estocásticos, torna a solução mais viável, para encontrar esses padrões se usa com frequência a amostragem de Gibbs, são cadeias de Markov onde as transições de um estado p/ outro são feitas de acordo com as distribuições condicionais completas