



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 surgem com a necessidade de entender as interações entre os diversos níveis de complexidade do organismo. As redes biológicas são sistemas complexos e possuem algumas propriedades: adaptações e auto organização. Portanto possuem características que permitem entender de uma forma mais clara todos os elementos que a envolvem. Por exemplo, existem diversos tipos de redes, as sinápticas, regulatórias e metabólicas, cada uma com sua característica intrínseca. Nas redes sinápticas é importante observar, entender quais os sinais (muitas vezes vindos do ambiente) são necessários para ativar um escuta de sinal 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 substitutos

e os reações químicas, ~~permitem observar a massa celular e a taxa de crescimento de célula.~~

Com o advento das plataformas de nanotecnologia e um grande acúmulo de dados e questões a serem entendidas surge a Biologia de Sistemas juntamente com essa área, surge também um avanço notável dos técnicos matemáticos e computacionais. Esses são essenciais para o entendimento das redes biológicas, pois eles 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. Para simular, prever o que acontece nas redes biológicas é necessário modelar ~~através~~ através dos métodos matemáticos e computacionais. Existem três tipos de modelos que vão permitir estipular hipóteses e observar o que acontece, nessas redes, auxiliando por exemplo a robustez da rede, ou seja, quais nós e arestas (genes/protn's - interações) são essenciais p/ o funcionamento da célula, um exemplo clássico de modelagem de redes é observar o comportamento dos mutações em redes



NOME: Graziela maria Dias

No. fl.

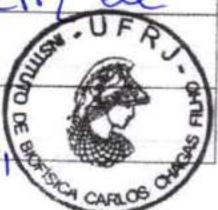
03

associados com câncer. Os modelos são definidos como lógico, contínuo e estocástico. 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 desenhadas a seguir.

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

Tres tipos de modelos são classificados como lógico: rede booleana, redes lógicas e petrinet.

As redes booleanas são consideradas os modelos mais simples, elas assumem somente 2 estados: on/off e off/on, ou seja, as interações irão ocorrer através de um funções booleana. Sua característica é que que todos os interações ocorrem de forma sincronizada, todos os eventos ocorrem de maneira simultânea para todos os variáveis. Embora, esse método tenha limitações,



NOME: Graciela Maria Dias

Nº fl.

04

diversos autores da comunidade científica têm mostrado processos em diversos modelos. Esse modelo é bem estudado em levaduras, apesar de momento na sequências de cíclios celulares. Vida das limitações é per isto evita computacionais e o número mínimos de conexões, consequentemente bastante distante da simplicidade, onde um organismo pode ter 20 conexões.

O segundo modelo é denominado como lógico, ele também é considerado primordial, como os genes bactériacos. Trata de eventos discretos, mas com a diferença que conseguem simular regras que 2 conexões, tornando o modelo mais sofisticado. Síntese diferencia é que os mudanças de estados podem ser feito de forma assimétrica.

O terceiro modelo é ~~uma rede~~ rede net, operando diferenças bastante relevantes quando comparadas as redes bactériacas e lógicas. Esse modelo pode ser usado para os genes metabólicos e regulatórios. Esse modelo é representado por um grafo bipartido, com componentes chamados fragmentos, transcrição e arcos. Geralmente, os fragmentos contêm os operadores e os fragmentos de receptor biológico. É também considerado que per um modelo não determinístico ex: regulas da vida de populações um B. probabilis



NOME: Graziela Maria Dros

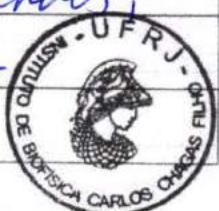
No. fl.

05

A simulação dinâmica é dada pelos trocadores e pelos atributos dos lugares, chamados de "fichas". Existem os dispositivos transversais onde são retiradas fichas de um lugar de extração e dispostas em outro lugar. Existe também variações extensões desse modelo que otimizam a simulação das redes. ex: redes pelas estruturas híbridas e estatística.

A segunda classificação dos modelos é denominada 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 não essenciais para o processo 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 aplicativos da EDO foi o estudo da replicação do operon lac. Outros exemplos também podem ser descritos: regulação das células da bactéria Escherichia coli crescentes, conhecida por apresentar 2 células morfológicamente diferentes e também é



bastante utilizado para estudo o ciclo celular em bactérias.

Outro método contínuo bastante utilizado é o μ FBA (análise de balanço de fluxos regulado). Esse método serve 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 suas taxas obtidas por programações lineares otimizando a função objetiva.

O FBA original é usado para as redes metabólicas, e baseado em uma matriz estequiométrica, e a partir desses pressupostos, estando estacionário, é possível observar o fluxo das reações. Esse método tem sido bastante utilizado para avaliar a taxa de crescimento, concentrações de substratos. Um diferença bastante interessante é que o μ FBA 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 seja devido os efeitos estocásticos recorrentes nas redes biológicas. Um dos métodos é o algo-



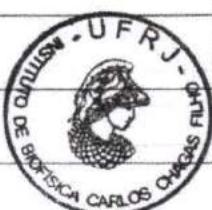
NOME: Graziela Maria Dias

No. fl.

07

ritmo que tem sido utilizados é Simulação, este é o modelo Gillespie (SSA). Esses modelos são de difícil aplicação pois são aplicados à cada reação individualmente. O SSA foi utilizado para estudar a regulação de fases considerando a fase líquida e fagocitica.

Além da descrição dos 3 métodos, outra abordagem que tem sido utilizada com sucesso são os técnicas de aprendizado de máquina. Um ex é identificação de redes regulatórias envolvidas na produção de antibióticos.



NOME: Grauegla Maria Diaz
Tópico III - Instrumentos estatísticos em biologia computacional

No. fl.

01

Com o advento das aplicações de nova geração, ~~com~~ a cada dia dados novos são gerados, criando uma grande desafio para biólogos estatísticos: armazenar, manter e verificar se os dados são propriamente estatísticos e conseguem explicar com confiabilidade os fenômenos biológicos.

Para fazer inferências de uma determinada população deve ser feita uma amostragem dessa população que por sua vez deve ser realizada estatisticamente. Existe 2 abordagens da estatística: a descritiva e inferencial.

A descritiva é o primeiro passo para os comitentes, é permitir a partitura desses dados em uma forma numérica dos dados. As principais medidas que representam uma amostra estatística são os modídos de resumo, variância e forma.

As medidas de resumo são a média, mediana e moda. A medida é privativa dos mediidores variáveis formadas e utilizadoras, assim ela é possivel obterem um resumo das ~~medidas~~ variáveis do domínio amostral, entendendo nôs resultados que essa elaboração é possivel



NOME: Grazielle 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 utilizada com frequência em dados categóricos e verifica o número de vezes que determinada categoria aparece. Outras medidas importantes são a variância que vai medir ~~desvio~~ quando 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 elle é adimensional. As medidas - forma também são importantes, para verificar o comportamento da curva em um gráfico, observando a simetria.

Os gráficos, também são uma importante ferramenta para 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 os técnicos são capazes de produzir, por exemplo microarranjos e RNAseq.

Antes de fazer a inferência estatística, ou seja, levar as hipóteses e aplicar testes, é fundamental analisar os dados para significados 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 microarray.

Após essa etapa, é possível passar a inferência estatística, com o levantamento de hipóteses e 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 α , o quanto de erro vai ser tolerado (nível mais comum 0,05), é feito o cálculo do p-valor, ou seja, qual é tão aquela distante o valor ~~esperado~~ da minha amostra está do valor esperado. Com esses parâmetros, é possível sózinho aceitar ou rejeitar a hipótese nula.

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



NOME: Graciela maria dias

No. fl.

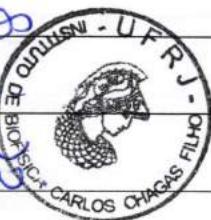
04

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

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

Outro exemplo de um trabalho que também está envolvida, é avançar através das Beta diversidade, quais as variáveis físicas químicas explicam determinado grupo taxonômico, obtido do pequenoamento de amplicon 16S.

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



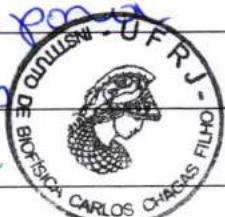
Se os dados seguem distribuição normal, os testes aplicados são os paramétricos. Um teste bastante utilizado para verificar a dist. normal é o Shapiro-Wilk. Outro pressuposto é obter a homocedasticidade, ou seja, se existe ou não variância entre as amostras, o teste aplicado é bivariado.

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

Os testes paramétricos compreendem no teste-t, teste-t pareado (p/ grupos dependentes), teste-t Welch (caso, não atinge 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 variação inter grupo e intra grupo.

Os testes não paramétricos ~~paramétricos~~ p/ 2 grupos independentes é o teste U Mann Whitney, para grupos pareados, teste Wilcoxon e para 3 grupos ou mais Kruskal-Wallis. Todas as TNP são baseadas no randomamento.



NOME:

No. fl.

06

e soma dos postos.

Outros testes que são essenciais na biologia, são os testes de comparação, devido a análise simultânea de diversos ~~variáveis~~ 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 puros, só 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 se um nº maior de amostras, pois ele reconhece os falsos desenhados no grupo de H_0 rejeitados, sendo mais eficaz no ajuste.



NOME: Grauela maria Dias

No. fl.

01

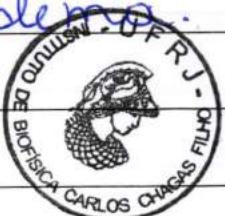
Tópico 09 - Fundamentos teóricos de algoritmos de biologia computacional e bioinformática

Os algoritmos são um conjunto de instruções 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 auxiliar a execução do alg, a solução do problema, é necessário auxiliar o tempo de execução.

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

Os principais algoritmos usados na biologia computacional são força bruta, programação dinâmica, algoritmos quentes e randomizados.

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



NOME: Gracilis Nuno Dias

produzido por *Bacillus phthirus*. Identificado no
strato da Epidermíta de mosquito, é bactéria
despendendo, portanto, efeitos de um periel mosquito
concreta, tanta é ação é expectativa que personagem
identificar o peptídeo na questão. Um oligonitino
não é peptídeo q tol ouviria é bactéria onde bactéria.

Outro oligonitino ~~que~~ bactéria na bactéria,
e a personagem direcionar ^(P) ele busca a melhor
solução dentro de problema isto é, ou seja,
ele busca a solução do problema, resolvendo por
meio de modo que não seja possível retomar as
problemas anterior, tornando a resolução do
problema educado. Os dois personagens que são
baseados no R.D & o ~~professor~~ Needlemann-Wisch
& Smith-Wolfgang, os personagens nas perspectivas
sofis olhamentos focal e focal respectivamente.

Pois a construção do olhamento é necessária
para uma construção de matrizes, onde virá para com-
partilhar ~~que~~ matrizes, matrizes e gaps. Na Bactéria
as matrizes de produção PAM e Glossum não
escrevem para o olhamento, os personagens das
personagens discussões apelavam para competidores.
Outro tipo de ~~que~~ olhamento pertence

é o olhamento multilevel, sendo privado
que o uso ~~de~~ de uma estrutura di-



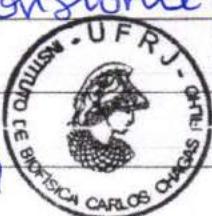
NOME: Graeida Maria Dias

03

mensional, gerando uma estratégia de alinhamento progressivo, onde é feito o alinhamento par-a-par e depois construído uma árvore guia, ~~onde~~ no qual vai ser computados os melhores alinhamentos. Uma das ferramentas mais utilizadas e citadas é o clustal.

Um exemplo bastante comum na minha área de atuação é fazer o alinhamento ~~de~~ de genes housekeeping a fim de identificar um perfil taxonômico de diversos espécies ~~bacterianas~~ de microorganismos, ou através da genómica comparativa, encontrar o seu genoma de um determinado organismo, linhagens diferentes da mesma espécie, os genes idênticos como esse-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 ótimos 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 limitação de manter a taxa molecular constante) e o Neighbor-Joining.

Para o alinhamento de sequências,



NOME: Graziela Maria Dias

No. fl.

04

o BLAST é uma das ferramentas mais utilizadas da biologia computacional, ele abriga diversos algoritmos, incluindo o Smith-Waterman fornecendo uma busca heurística das frequências.

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

O tipo guloso representa o grafo no seu modo mais simples, reúne 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 para montar dados oriundos do 454, utilizam o caminho hamiltoniano que torna o problema NP-difícil. Nesse método os nós são os reads e os arcos as sobreposições. Ex: Mira, Neubler

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



NOME: Graziela Maria Dues

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 percorrer todos, portanto a introdução dos métodos estocásticos, torna a solução mais viável, para encontrar esses padrões se usa com propositura a amostragem de Gibbs, os efeitos de Monkov onde os transições de um estado p/ outro são feitas de acordo com as distribuições condicionais completas

