



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: Vitor Laima Callio

No. fl.
1

04- Modelagem de redes biológicas

O desenvolvimento, avanços e popularização da aplicação de tecnologias ômicas de alto rendimento aumentou consideravelmente a quantidade de dados e informações sobre fenômenos biológicos. A modelagem matemática destes fenômenos e suas interações são de grande interesse da ciência para compreender a estrutura geral dos sistemas in silico e utilizar seus resultados para auxiliar a pesquisa in vitro e in vivo. Modelos adequados para este grande quantidade de dados são necessários para simular e melhorar o conhecimento de complexos mecanismos moleculares.

Podemos destacar os modelos quantitativos e modelos lógicos para a modelagem de redes biológicas.

Os modelos quantitativos são modelos matemáticos onde o valor das variáveis são determinados através de análise numérica e parâmetros do sistema. A variável no modelo quantitativo representa uma determinada espécie molecular, como a concentração de metabólitos,

NOME: Vitor Lima Coelho

No. fl.

2

quantidade de moléculas de mRNA ou atividade de um gene, por exemplo. Este tipo de modelo matemático controla de forma dinâmica a diminuição (por exemplo, degradação, exportação, inibição) e o aumento (por exemplo, síntese, importação, ativação) das espécies moleculares modeladas. Este processo é acompanhado através da alteração das taxas medidas em determinados intervalos de tempo.

Os modelos lógicos são modelos matemáticos onde as variáveis são valores discretos determinados pelo combinação lógica dos valores de outras variáveis. As variáveis podem ser diferentes elementos como a atividade de um gene, presença de uma proteína, o estado de uma célula, entre outros. Os modelos lógicos são usados ou aplicados a conjuntos de variáveis qualitativas representadas por números inteiros discretos.

Por trás dos modelos matemáticos utilizados para modelagem de redes biológicas, existe uma representação de rede que é adequada para o domínio do problema e dados relacionados. O modelo adequado para determinado rede biológica e sua representações são fatores críticos para o sucesso da modelagem do sistema biológico. Podemos citar as seguintes representações de redes biológicas:



NOME: Vitor Laima Coelho

No. fl.

3

redes de interação, fluxo de atividades, mapas de duração de processos e mapas de atidode-velocimento.

As redes de interação são redes construídas a partir de listas de interação física ou funcional, por exemplo, as redes de interações proteína-proteína. As redes de interação fornecem uma visão abrangente de reguladores de processos específicos. Além disso, este tipo de rede é útil para analisar a estrutura de sistemas como um todo ou o resultado de uma perturbação. Por outro lado, devido a falta da percepção mecanicista e sua natureza estática tornam a rede de interação inadequada para representar modelos dinâmicos.

A representação de redes por fluxos de atividades é usada quando os detalhes de uma reação química não são conhecidos ou não são essenciais para a compreensão da biologia do problema. Este tipo de rede é muito utilizado para modelar redes de sinalização celular ou redes regulatórias. Fluxos de atividades são representações naturais para modelos qualitativos e, em particular, modelos lógicos.



Os mapas de descrição de processos são representados por grafos bipartidos, compostos por dois tipos de nós: as variáveis cuja a evolução se deseja acompanhar; e os processos que diminuem ou aumentam os valores dessas variáveis. Este tipo de apresentação é aplicado para modelagem de transferência de massa, com ações direcionadas e é composto por redes sequenciais. Os mapas de descrição de processos são aplicados para descobrir o metabolismo central ou reações metabólicas associadas à mineralização ou regulação química, como por exemplo, os mapas de rias do KEGG e Reactome.

Os mapas de entidade-relacionamento modelam redes biológicas baseados em três circuitos: a entidade (por exemplo, um gene), declarações sobre essas entidades (por exemplo, interações ou estado de metiloces) e a influência das entidades em declaração (estímulo). Estas influências possuem direcionalidade, isto é, " X influencia Y " é diferente de " Y influencia X ". Os mapas de entidade-relacionamento oferecem granularidade de representação não adequada para mecanismos e eventos moleculares (por exemplo, ciclo celular e apoptose) e são construídos através do acúmulo de reloges independentes.

Como dito anteriormente, a escolha de modelos e representações de rede para modelagem



NOME: Vitor Laima Calho

No. fl.

5

de redes biológicas é de suma importância para o sucesso do estudo. Alguns fatores devem ser considerados nessa tomada de decisão, como por exemplo; qual o conhecimento disponível sobre a direcionalidade das regulagens e os mecanismos envolvidos; qual a natureza dos dados disponíveis: experimentais quantitativas temporais sobre concentração ou nível de expressão genética. Dependendo das respostas para essas perguntas, deve-se adotada uma abordagem bottom-up ou baseada em conhecimento, utilizando informações obtidas da literatura e bancos de dados biológicos com análises previamente executadas ou inferir o ponto de partida diretamente a partir de conjunto de dados experimentais.

Independentemente da abordagem, modelo ou representações de rede ~~estimadas~~ usadas para a modelagem de redes biológicas, é importante considerar também a qualidade das fontes de informação e o nível de ~~correlação~~ confiabilidade dos dados, com subsequente validação experimental.



NOME: Vitor Henrique Lachus

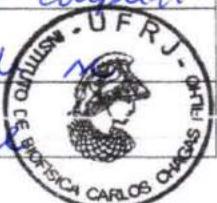
No. fl.

6

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

Com o início da "Era Genômica" no início dos anos 2000 e o advento das tecnologias de sequenciamento de nova geração nos anos subsequentes, a quantidade de sequências biológicas disponíveis publicamente aumentou exponencialmente. Consequentemente, aumentou-se também a quantidade de dados e meta-dados diretos ou indiretamente relacionados. Analisar e ~~reduzir~~maneira eficiente esta grande quantidade de informação é um dos grandes desafios das ciências da vida e requer cada vez mais métodos computacionais mais eficientes e específicos. Projetar algoritmos para estes problemas que sejam eficientes, corretos e efetivos é uma área em constante desenvolvimento na Biologia Computacional e Bioinformática. A diferentes aspectos teóricos não são todos levados em consideração durante sua elaboração.

Um fundamento teórico importante a ser considerado durante o projeto de ~~até~~ algoritmos é a sua complexidade de tempo. A ordem de complexidade de um algoritmo determinará seu desempenho em função da quantidade de dados de entrada. A ordem de complexidade é representada por $O(\underline{\underline{f(x)}})$ (big-O) e pode ter de diferentes ordens de grandeza. O caso ideal



na é igual na maior parte das aplicações i que o tempo de execução seja constante independentemente do valor de n (dados de estuda). Em casos mais realistas, a complexidade pode ser, por exemplo, linear, quadrática, logarítmica e exponencial. A implementação de algoritmo deve considerar tanto os casos médios, pior caso e melhor caso de consumo de tempo, buscando a solução mais apropriada e menos custosa de pior caso. P complexidade de memória, ou seja, o consumo de memória do algoritmo em função de n também é outro aspecto a ser considerado e que muitas vezes é desconsiderado diante a aumento das capacidades de memória primária e ~~e~~ secundária de modernos.

Diferentes abordagens de projetos de ~~est~~ algoritmos são empregadas em ^{diversas} diferentes aplicações em biociências como: algoritmos de busca exaustiva, algoritmos baseados em dividir e conquistar, algoritmos de programação dinâmica, greedy e branch-and-bound.

Os algoritmos de busca exaustiva se baseiam na redução de um problema considerando todo espaço de solução. Algoritmos de busca exaustiva foram aplicados na busca por motivos de DNA, isto é, sequências de DNA que se repetem por todo o genoma e podem ter proximamente alguma



NOME: Vitor Leiria Coelho

No. fl.

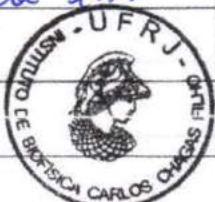
8

função regulatória. Neste caso, o espaço de busca é o genoma intiro, evidenciando que isto abordagem é impraticável para grande parte dos problemas reais na biologia computacional.

A abordagem de dividir-conquistar consiste num método para otimizações que subdivide o problema em sub-problemas resolvendo-os individualmente e agrupando-os ao final do execução para construir a solução ótima geral. Pode ser aplicado para problemas de identificação de reloções de ~~co-~~expressão de genes em grafos.

Os algoritmos de programação dinâmica se basiam na enumeração de todas soluções do espaço de busca para obter a solução que maximiza (ou minimiza) a função objetivo. Esta abordagem tem a aplicação clássica no alinhamento de sequências biológicas, onde as possíveis soluções são avaliadas a partir de pontuação para matches (correspondência), delgés e inserções para maximizar os bits de alinhamento.

Os algoritmos de greedy se basiam na busca de uma ou mais soluções ótimas locais. Esta abordagem ^{é aplicada} onde a simples detecção de um sinal biológico é suficiente para responder a pergunta em vez da enumeração da melhor solução que gerou o sinal sinal, por exemplo.

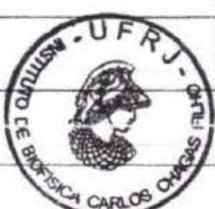


NOME: Vitor Leima Corlho

No. fl.
9.

A abordagem de branch-and-bound se baseia na busca de uma solução ótima para um função objetivo que utiliza intervalos mínimos e máximos de aceitação para percorrer determinado caminho para um solução. Caso seja detectado que não há perspectiva de atingir os limites, a solução é descartada. A pesquisa AstoFunk (Corlho e Sonneth, 2018) utiliza esta técnica em conjunto com a programação dinâmica para entre cálculos em células da matriz de alinhamento que não levarão a uma solução dentro dos parâmetros mínimos de aceitação de alinhamento de um domínio proteico, representado por um modelo curto de Markov.

A escolha de uma abordagem adequada de projeto de algoritmos de Biologia Computacional e Bioinformática é crucial para o sucesso do método e popularização da técnica para utilização na pesquisa em biociências. No entanto, é importante também que o pesquisador, desenvolvedor ou projetista do método esteja alinhado com os aspectos biológicos do problema, produzindo uma solução que faça sentido biologicamente.



D3 - Ferramentas estatísticas em biologia computacional

A complexidade do estudo dos fenômenos biológicos e seus mecanismos requer o trabalho conjunto da Biologia, computação, química, matemática, estatística, entre muitas outras. As ferramentas estatísticas de Biologia Computacional possuem grande relevância nesse estudo tanto para descobrir novos insights biológicos quanto para garantir a confiabilidade dos descobertas.

Métodos ~~de~~ de aprendizado de máquina são portamente baseados em estatística e computação, com diversas aplicações em Biociências. Métodos de aprendizado de máquina supervisionado são aplicados por exemplo na identificação de padrões e classificação de amostras. Já métodos de aprendizado de máquina não-supervisionados podem ser aplicados para inferir o sucesso de um experimento com base no agrupamento de amostras de mesma condição.

Outras ferramentas, como Hmmer e Blastp, se baseiam em cálculos estatísticos e probabilísticos probabilísticos para identificar domínios protéticos ou famílias de proteínas.

Ferramentas estatísticas além de serem utilizadas para descobertas de eventos biológicos, também auxiliam outras ferramentas e áreas para garantir a confiabilidade de seus resultados.

As ferramentas Combatch e Noisq implementam



NOME: Vitor Lemos Coelho

No. fl.

11

tistas estatísticos para inferir o efeito de fatores externos em experimentos de análise de expressão química, considerando posteriormente efeitos de batch com fatores conhecidos ou desconhecidos, respectivamente.

Os estudos quantitativos de dados de RNA-Seq (sequenciamento de RNA) impregnam diversas ferramentas que implementam internamente métodos estatísticos e testes probabilísticos desde a análise de expressão diferencial até a análise funcional. Os métodos de análise de expressão diferencial avaliam estatisticamente quando a variação de expressão genética entre duas condições é confiável. A quantidade de amostras é um fator fundamental para a significância estatística das análises.

Os métodos de análise de expressão diferencial podem ser paramétricos (EdgeR, DESeq), ou seja, assumem uma distribuição dos níveis de expressão; ou não paramétricos (Noise, GFO). Estes últimos são geralmente aplicados quando a quantidade de réplicas para cada condição é reduzida ou inexistente.

A partir dos resultados da análise de expressão diferencial, obtemos uma lista de genes que apresentaram diminuição ou aumento estatisticamente significativo. O passo seguinte é inferir qual o papel biológico desses genes nas condições analisadas. O método GSEA é utilizado nesse passo.



NOME: Mílton Límo Coelho

No. fl.

12

identificar quais termos de ontologias (processos biológicos, função molecular, componente celular) está isto é, totalmente representado dentro todos os termos detectados para cada condição. Esta análise é denominada de análise de enriquecimento de termos GO, e é uma etapa importante da análise funcional a partir dos dados de RNA-Seq.

A quantidade de de ferramentas estatísticas aplicadas à bioinformática é bem ampla, sendo aplicadas nos mais diversos ramos das ciências biomédicas assim como na sua integração delas. Estas ferramentas são aplicadas tanto na descoberta de conhecimento quanto na avaliação da confiabilidade dos resultados. Embora as ferramentas estatísticas estejam em constante evolução, a avaliação crítica do pesquisador é fundamental antes de qualquer resultado gerado por uma ferramenta.

