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 Coelho | 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 bióligos. 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 ~~auxi~~ auxiliar a pesquisa in vitro e in vivo. Modelos apropriados para esta 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~~ é determinados através de análise numérica e parâmetros do sistema. A variável no modelo quantitativo representa uma determinada espécie molecular, como a concentração de metabólitos,

quantidade de moléculas de mRNA ou atividade de um gene, por exemplo. Este tipo de modelo matemático controla de forma dinômica 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 dos taxas medidas em determinados intervalos de tempo.

Os modelos lógicos são modelos matemáticos onde as variáveis são valores discretos determinados pela 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 representados por números inteiros discretos.

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

redes de interação, fluxo de atividades, mapas de descrição de processos e mapas de entidade-relacionamento.

As redes de interação são redes construídas a partir de listas de interação física ou funcional, por exemplo, as redes de interação proteína-proteína. As redes de interação fornecem uma visão abrangente de regulação 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 vias de sinalização celular ou redes regulatórias. Fluxos de atividades são representações naturais para modelos protocolos e, em particular, modelos lógicos.

NOME: Vitor Lema Coelho

No. fl. 4

Os mapas de descrição de processos são representados por grafos bipartidos, compostos por dois ~~nós~~ 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 representação é aplicado para modelagem de transferência de massa, com arestas direcionadas e é composto por redes ~~separáveis~~ sequenciais. Os mapas de descrição de processos são aplicados para descrever o metabolismo central ou reações metabólicas associadas à sinalização ou regulação gênica, como por exemplo, os mapas de vias do KEGG e Reactome.

Os mapas de entidade-relacionamento modelam redes biológicas baseados em três conceitos: a entidade (por exemplo, um gene), declarações sobre essas entidades (por exemplo, interação ou estado de mediação) 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~~ ~~granulado~~ granularidade de representação adequada para mecanismos e eventos moleculares (por exemplo, ciclo celular e apoptose) e são construídos através do acúmulo de relações independentes.

Como dito anteriormente, a escolha do modelo e representação de rede para ~~modelagem~~ modelagem

de redes biológicas é de suma importância para o sucesso do estudo. Alguns fatores devem ser considerados nesta tomada decisão, como por exemplo; qual o conhecimento disponível sobre a direcionalidade das regulações e os mecanismos envolvidos; qual a natureza dos dados disponíveis: experimentos quantitativos ~~temporais~~ temporais sobre concentrações ou níveis de expressão gênica. Dependendo das respostas para essas perguntas, deverá ser adotada uma abordagem botton-up ou baseada em conhecimento, utilizando informações obtidos 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 representação de rede ~~definidos~~ escolhido para a modelagem de redes biológicas, é importante considerar também a qualidade das fontes de informação e o nível de ~~confiabilidade~~ confiança dos dados, com subsequente validação experimental.

NOME: Vitor Haimo Coelho

No. fl. 6

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

Com o início do "Era Genômica" no início dos anos 2000 e o advento das tecnologias de sequenciamento de nova geração nos anos subsequentes, a quantidade de sequências biológicos disponíveis publicamente aumentou exponencialmente. Consequentemente, aumentou-se também a quantidade de dados e meta-dados direto ou indiretamente relacionados. Analisar e ~~relacionar~~ relacionar 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 - e diferentes aspectos teóricos ~~não~~ são ~~levados~~ levados em consideração durante sua elaboração.

Um fundamento teórico importante a ser considerado durante o projeto de ~~algo~~ 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 n. A ordem de complexidade é representada por O (~~big-O~~) (big-O) e pode ser de diferentes ordens de grandeza. O caso ideal

~~nao~~ é irreal na maior parte das aplicações é que o tempo de execução fosse constante independentemente do valor de n (dados de entrada). Em casos mais realísticos, a complexidade pode ser, por exemplo, linear, quadrática, logarítima e exponencial. A implementação do algoritmo deve considerar então os casos médios, pior caso e melhor caso de consumo de tempo, buscando a solução mais apropriado e menos ~~custosa~~ custosa de pior caso. A complexidade de memória, ou seja, o consumo de memória do algoritmo em função de n tambem é outro aspecto a ser considerado e que muitas vezes é desconsiderado devido a aumento das capacidades de memória primária ~~e~~ ~~sega~~ secundária de servidores.

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

Os algoritmos de busca exaustiva se baseiam ~~em~~ na resolução 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 presumivelmente alguma

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

A abordagem de dividir-conquistar consiste num método para otimização que subdivide o problema em sub-problemas resolvendo-os individualmente e agregando-os ao final da execução para construir a solução ótima geral. Pode ser aplicado para problemas de identificação de relação de ~~co~~ coexpressão de genes em grafos.

Os algoritmos de programação dinâmica se baseiam 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 calculadas a partir de pontuação para matches (correspondência), deleções e inserções para maximizar o bitscore do alinhamento.

Os algoritmos de greedy se baseiam na ~~busca~~ busca de uma ou mais soluções ótimas locais. Esta abordagem (1ª aplicada) onde a simples detecção de um sinal biológico é suficiente para responder a pergunta em vez da enumeração do melhor solução que se- ~~sinal~~ sinal, por exemplo.

NOME: Vitor Leima Coelho

No. fl. 9.

A abordagem de brand-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. O fundamento AstroFunk (Coelho e Sommeth, 2018) utiliza esta técnica em conjunto com a programação dinâmica para evitar cálculos em células da matriz de alinhamento que não levarão a uma solução dentro dos parametros mínimos de aceitação de alinhamento. de um domínio proteico, representado por um modelo oculto de Markov.

A escolha de uma abordagem apropriada 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.

NOME: Vitor Leima Carlos | No. fl. 10

03 - Ferramentas estatísticas em biologia computacional

A complexidade do estudo dos fenômenos biológicos e seus mecanismos requer o trabalho conjunto das 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 das descobertas.

Métodos ~~como~~ de aprendizado de máquina são fortemente baseados em estatística e computação, com diversas aplicações em Biociências. Metodos 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 Limma e Astafunk, se baseiam em cálculos estatísticos e ~~probabilísticos~~ probabilísticos para identificar domínios proteicos ou famílias de proteínas.

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

As ferramentas Combatch e Noiseq implementam

testes estatísticos para inferir o efeito de fatores externos em experimentos de análise de expressão gênica, corrigindo posteriormente efeitos de batch em fontes conhecidos ou desconhecidos, respectivamente.

Os estudos quantitativos de dados de RNA-Seq (Sequenciamento de RNA) empregam diversas ferramentas que implementam internamente métodos estatísticos e ~~poeba~~ 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 gênica entre duas condições é confiável. A quantidade de amostras é um fator fundamental para a significância estatística das análises. Os metodos de análise de expressão diferencial podem ser paramétricos (EdgeR, DESeq), ou seja, assumem uma distribuição dos níveis de expressão; ou não paramétricos (NOISeq, GFOLD). Estes últimos são geralmente aplicados quando a quantidade de réplicas para cada condição é reduzida ou inexistente.

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

identificar quais termos de ontologias (processos biológicos, função molecular, componente celular) está estatisticamente super representado dentre 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 de dados de RNA-Seq.

A quantidade de ferramentas estatísticas aplicadas à biociencias é bem ~~ca~~ ampla, sendo aplicada nos mais diversos ramos das ciências ômicas 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.