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: GIORDANO BRUNO SOARES SOUZA

No. fl. 1

Questão 1) → Ferramentas estatísticas em biologia computacional

A estatística é um campo do conhecimento orientado a descrever populações (ou amostras) e tentar inferir a probabilidade de de eventos. Não é surpresa que esteja intimamente atrelada à história evolutiva dos organismos e à descrição dessas populações. Grandes avanços no campo da estatística foram realizados por geneticistas como Galton e Fisher que ~~fazem~~ criaram novos modelos e ferramentas para descrever as populações e tentar prever alterações nestas.

De maneira geral, a estatística pode ser dividida em uma parte descritiva que busca caracterizar as amostras. As características de cada elemento dessa população podem ser denominadas variáveis. Variáveis assumem valores qualitativos como as categóricas - associadas à rótulos; e as ordinais - que representam a ordenação da amostra. As variáveis podem assumir ainda valores discretos ou contínuos.

Em populações biológicas, exemplos de variáveis contínuas são massa e altura dos indivíduos, ~~variáveis discretas são a contagem de alelos em um locus~~ podem ser assim como a ~~o~~ contagem de alelos de um locus pode assumir caráter discreto em um ~~indivíduos~~ indivíduo. Alelos são abstrações para representar a herança genética de indivíduos e, dessa forma, nos aju

dam a interpretar e inferir as características e a história evolutiva das populações. O ato de tentar prever essas características e as mudanças temporais constituem a segunda parte da estatística, denominada indutiva ou preditiva.

Como no caso de Galton que tentava inferir o impacto da genética na altura dos indivíduos, outras abordagens são possíveis. Pode-se querer inferir quais fatores genéticos e socioeconômicos estão associados à uma doença como o câncer gástrico. Neste caso, propomos um teste de hipótese onde a hipótese nula representa que não há associação entre um genótipo ou a renda com a doença e a alternativa que indica a associação. Para testar essas hipóteses recorremos aos testes estatísticos, estes podem ser paramétricos quando a distribuição da variável é conhecida e normal, ou não-paramétrico para os demais casos. Entre os testes estatísticos mais utilizados, podemos usar o teste de Kolmogorov-Smirnov para identificar se o status sócio-econômico está associado à doença ou a regressão para inferir a relação entre a variável dependente e a resposta. Esses testes citados dependem de que as variáveis testadas sejam independentes, o que nem sempre é o caso. Em um teste de associação onde as variáveis são tomadas ao longo do tempo, essa assunção é violada. Em um estudo para associar a cor da pele (constitutiva e a responsiva ao UV) a dados genéticos e socioeconômicos usando uma coorte familiar, a assunção de independência é violada uma vez que os indivíduos são aparentados e compartilham grande parte do genoma. Nestes casos onde a colinearidade está presente, é conveniente usar outras ferramentas como os modelos lineares mistos, e estes são capazes

de criar diferentes interceptos para amostragens temporais - um intercepto para cada medição no tempo ou familiares - um intercepto por família.

Fica claro neste ~~um~~ exemplo que é necessário conhecer o grau de compartilhamento genético para utilizar os métodos estatísticos apropriados. ~~Um~~ Uma forma conveniente de se avaliar a diversidade genética das populações e identificar estruturas crípticas é realizar o cálculo das frequências alélicas das populações e aplicar o modelo do Equilíbrio de Hardy-Weinberg. Neste modelo, a ~~pe~~ probabilidade conjunta das frequências alélicas de duas populações é estimada e comparada a observada, diferenças significativas indicam a estrutura populacional. Esse mesmo princípio pode ser utilizado para inferir a divergência genética de grupos populacionais a partir da análise de variância. Esse cálculo denominado Índice de fixação (F_{ST}) na genética de populações nos permite identificar o grau de divergência entre os loci das populações com grande serventia nas áreas evolutiva e biomédica. Em um estudo realizado em populações nativas do Peru, conseguimos inferir polimorfismo possivelmente sob seleção natural devido ao grau de divergência ~~dos~~ das frequências ~~em loci específicos~~ e identificamos um ~~o~~ conjunto de marcadores com estruturação e associados à doenças em outros estudos na literatura científica. Isso pode indicar que alguns desses achados são falso-positivos e ~~está~~ a associação causal não ocorre devido ~~à~~ ao ~~fat~~ genótipo mas a algum fator associado. ~~À exemplo~~ Por exemplo, no estudo de associação ~~com~~ para o câncer gástrico, havia associação entre a ancestralidade ameríndia e fatores sócioeconômicos ~~à presença~~ ao desfecho clínico, mas a nenhum loci foi

indevidamente associado devido à incorporação da ancestralidade no modelo. ~~Para~~ ~~Em~~ Em alguns casos, indivíduos muito diferentes ~~em~~ geneticamente nos grupos caso ou controle podem ser excluídos das análises. Para exclusão ou para incorporação ao modelo, as análises de redução de dimensionalidade, tais como, análise de componentes principais são usualmente empregados.

A identificação das bases moleculares das doenças pode levar à melhoras no prognóstico, diagnóstico e tratamento, mas devido ao caráter complexo da etiologia dos desfechos é conveniente incorporar a incerteza nos modelos de predição e decisão. Em geral, isso pode ser alcançado através da estatística Bayesiana. Esta, se assenta sobre o teorema de Bayes e estima a probabilidade de um evento a partir da incorporação do conhecimento prévio ou da incerteza ao modelo. Nesta forma, temos uma distribuição de probabilidades a priori - o que sabemos, uma distribuição de probabilidades a posteriori - o que desejamos conhecer e uma verossimilhança associada que nos indica quais valores são mais prováveis. Análises Bayesianas são utilizadas em modelos de decisão, como ajustar a dose de um medicamento de acordo com o perfil genético do indivíduo ou até mesmo serem incorporadas em modelos mais complexos como o ABC (Computação Bayesiana Aproximada) que facilita e permite inferências evolutivas. ~~Em~~ Nos estudos evolutivos, essa ferramenta é utilíssima pois desconhecemos vários dos parâmetros do modelo e, por isso, é conveniente representar essa incerteza.

NOME: GIORDANO BRUNO SOARES SOUZA

No. fl. 5

O uso das ferramentas estatísticas pela biologia computacional é essencial para descrever as populações, inferir sua história evolutiva e as implicações desta na expressão de fenótipos. Durante minha ~~tra~~ trajetória acadêmica, este tem sido o meu principal objetivo. Através da descrição da diversidade de populações e loci, da tentativa de identificar as bases moleculares de fenótipos como o câncer gástrico e a cor da pele, assim como a associação destes à fatores sócioeconômicos e no desenvolvimento de ferramentas para sumarizar dados biológicos tenho continuamente aplicado vários dos conceitos e ferramentas necessários para aumentar o ~~nosso~~ nosso conhecimento sobre a evolução e o impacto da diversidade.

NOME: GIORDANO BRUNO SOARES SOUZA

No. fl. 6

Q2) modelagem de redes biológicas

Redes são abstrações que nos permitem identificar os componentes, entender o comportamento dinâmico e a interação entre os ~~seus~~ elementos. Constituem um método ~~apropriado~~ conveniente para modelar a relação entre entidades por não requerer regularidade e por serem objetos combinatórios flexíveis. ~~São constituídos~~ São representadas como grafos ($G=\{V,E\}$) constituídos por vértices que representam os elementos e arestas que representam os relacionamentos. Estas podem ser direcionadas (sentido) ou não-direcionadas, e são ainda ~~podem além de~~ capazes de expressar confiança ou força através dos pesos, ou ativação e repressão através de sinais.

Duas aplicações de redes envolvem a visualização da relação entre variantes causais ~~de~~ diferentes fenótipos e que nos permitem identificar elementos comuns às doenças e também os específicos. Outra aplicação refere-se à ~~fo~~ dissolução de redes familiares para cálculo de ancestralidade. Neste caso, uma rede é modelada tendo as arestas representando o grau de relacionamento entre os indivíduos e, recursivamente, grafos são formados ~~tem~~ com o intuito de agrupar a maior quantidade de indivíduos não aparentados.

Outro tipo de rede bastante utilizado são as redes bayesianas onde as arestas representam probabilidades e os grafos são direcionados. Equações diferenciais podem ser usadas para modelar sistemas dinâmicos no espaço

indivíduos e para estimar o grau de dissimilaridade genômica entre indivíduos de espécies próximas.

A teoria de sistemas está intimamente ~~co~~ conectada à biologia de sistemas por buscar modelar e prever os resultados de perturbações. Um sistema é descrito como um grupo coeso, interrelacionado, interdependente, situado no espaço e no tempo que relaciona-se com o ambiente e possui ~~uma~~ estrutura e ~~um~~ propósito. A modelagem dos sistemas nos permitem compreender o seu funcionamento e identificar sinergia e propriedades emergentes. Uma das aplicações desta

e no tempo. A representação das interações e topologias podem auxiliar a entender a dinâmica do sistema modelado. Algumas medidas importantes para caracterizar redes consistem em: Conectividade - ~~um~~ número de arestas necessárias para remover e separar um ~~gra~~ sub-grafo; ~~central~~ centralidade que atesta a importância do nó. homogeneidade que representa a quantidade média de ligações; distância que é calculada pela soma das arestas e intermediação que representa o nó pelo qual passam a maior parte dos caminhos mais curtos.

NOME: GIORDANO BRUNO SOARES SOUZA

No. fl. 8

Q3) Fundamentos teóricos de algoritmos de biologia computacional e bioinformática.

A biologia computacional é um ~~assenta-se~~ campo interdisciplinar que envolve as áreas da ciência da computação, ciências biológicas e estatística. Os fundamentos teóricos ~~que~~ sobre os quais se assenta a biologia ~~de~~ computacional e a bioinformática envolvem: 1) teoria da informação; 2) teoria dos sistemas; 3) teoria do controle; 4) matemática discreta; 5) estatística; 6) genética de populações e evolução; 7) filogenia e sistemática. ~~Através~~ É através da interação entre essas ~~as~~ fontes que podemos modelar problemas e aplicar ~~técnicas~~ ~~de~~ algoritmos para solucionar as questões postas pelos pesquisadores.

A teoria da informação estabelece o arcabouço teórico para se entender como as mensagens são emitidas e decodificadas, além de estabelecer a entropia como uma medida de incerteza associada a um experimento probabilístico. A entropia pode ser medida entre duas variáveis ~~indicando~~ indicando independência ou assimilaridade. Essa também relaciona-se à complexidade de uma sequência, por exemplo, e pode indicar o grau de compressão possível. O conceito de entropia tem sido aplicado em ~~a~~ diferentes ~~problemas~~ algoritmos para estabelecer o grau de similaridade entre sequências, como a identificação de assinaturas genômicas ou comparação de sequências sem alinhamento. Em trabalho recente, ~~identifiquei~~ utilizei ~~este~~ ferramentas associadas à este conceito em ~~para~~ rotinas de controle de qualidade para confirmar que dados de sequenciamentos realizados em tempo e espaço distintos eram provenientes do mesmo

indivíduo e para estimar o grau de dissimilaridade genômica entre indivíduos ~~dos~~ de espécies próximas.

A teoria de sistemas ~~te~~ busca modelar e prever o impacto das ~~ob~~ perturbações em um sistema. Este pode ser descrito como um grupo de elementos coeso, interrelacionado, interpendente, situado no tempo e no espaço, que relaciona-se com o ambiente e possui estrutura e propósito. Essas características que descrevem o sistema são amplamente encontradas na biologia e, por isso, a teoria de sistemas fornece subsídios para a modelagem de sistemas biológicos afim de prever suas funcionalidades. Um exemplo de aplicação ocorre na modelagem da seleção natural poligênica onde ~~uma~~ a presença ~~rede~~ de vários genes com indícios de seleção em subgrafos de uma via metabólica podem indicar a ocorrência de um processo de adaptação. Em trabalho com ~~A~~ nativos americanos, conseguimos identificar agrupamentos de genes relacionados à metabolização de fármacos. ~~como~~

~~Ba~~

Quando se ~~vo~~ foca nos próprios algoritmos, há diferentes ~~abordagens~~ fundamentos e abordagens que podem ser aplicadas para a resolução de uma tarefa. Em geral, os algoritmos podem ser divididos em: I) programação dinâmica; II) programação linear; III) algoritmos gulosos; IV) dividir e conquistar; V grafos. A essência de cada categoria e exemplos de utilização na minha carreira acadêmica são descritos ~~nos~~ nos parágrafos a seguir.

Os algoritmos baseados em Hidden Markov ~~são~~ constituem modelos probabilísticos generativos onde se representa ~~os~~ uma cadeia de estados com os seguintes parâmetros: ~~I~~ inicial, transição e emissão. O símbolo e o estado ~~não~~ só dependem do próprio estado. ~~um~~ Uma distribuição de probabilidades é associada a cada caminho e o algoritmo de Viterbi escolhe o melhor caminho. São aplicados para gerar parâmetros iniciais em programas de inferência populacional.

Programação dinâmica ~~consi~~ consiste em dividir os problemas em soluções menores. São utilizados em alinhamentos de ~~dois~~ sequências.

Algoritmos gulosos resolvem problemas peça-a-peça e só se importam com a melhor solução local e imediata.

Algoritmos de dividir e conquistar quebram os problemas em instâncias menores e resolvem recursivamente. Bastante utilizados em ordenação e com aplicações na seleção de primers para PCR multiplex.

Programação ~~linear~~ linear é um processo de otimização onde as restrições e otimizações são lineares.

Algoritmos de grafos podem auxiliar na identificação de subconjuntos através de decomposição modular e podem ser aplicados em árvores de decisão: