**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: LUCENILDO SILVA CERQUEIRA

No. fl. 1

3- Ferramentas estatísticas em biologia computacional

As ferramentas estatísticas estão presentes na grande maioria dos mais diversos campos científicos. Inicialmente, os problemas abordados por tais ferramentas possuíam como características a possibilidade de controlar os experimentos de forma a isolar o máximo possível a variável de interesse. Contudo, a medida que o desenvolvimento tecnológico foi avançando foi possível desenvolver desenhos experimentais com grandes volumes de dados. No campo da biologia computacional, as ferramentas estatísticas, que eram empregadas no que podemos classificar de ciência tradicional, acompanharam esse desenvolvimento para facilitar a tomada de decisão sobre qual gene, proteína ou metabólito poderia ser considerada relevante para estudo em questão.

Geralmente, no emprego de qualquer ferramenta é importante que os pressupostos teóricos do projeto

que pretende desenvolver sejam levantados previamente. Este levantamento teórico prévio mostra as abordagens que são aplicadas pelos diferentes grupos de estudos pelo mundo. Os dados que são produzidos nos experimentos envolvendo material biológico costumam apresentar estruturas que não apresentam uma distribuição, uma dispersão em relação a um valor de referência. São dados que apresentam diferentes distribuições e dependi do tipo de experimento que esta sendo realizado.

Para decidir qual seria a melhor abordagem em cada experimento geralmente conta-se com a ~~experiência~~ experiência de pesquisadores seniors ou é possível seguir diretrizes básicos para analise de seus dados. A primeira abordagem para se conhecer qual o tipo de dados estamos tratando é necessário fazer uma análise exploratória (~~de~~) do conjunto de dados a ser tratatado. Nessa etapa são elaborados figuras que podem mostrar o comportamento dos dados, como estão distribuídos num plano cartesiano. Também é importante que sejam elaborados tabelas e histogramas os quais mostram quais decisões devem ser tomadas para análise do conjunto

de dados. Após criteriosa análise de dados é necessário a (~~dado~~) realização de testes de aderência à diferentes distribuições amostrais. Os 'chutes' iniciais sobre qual teste deve ser aplicado (~~ser~~) ocorrem após a etapa anterior.

Em estatística podemos considerar que existem dados que seguem uma distribuição normal, por exemplo, a concentração de hemácias. Enquanto, outros seguem uma distribuição não normal. Geralmente, dados biológicos seguem uma distribuição não normal.

Para dados que seguem uma distribuição normal é possível comparar amostras (~~ex~~) contra um valor de referência usando o teste t para amostras independentes. É empregado em condições onde o pesquisador quer comparar dados de um novo experimento contra um valor de referência. Ainda podemos fazer comparações com amostras dependentes. Imagine que queira avaliar o crescimento de uma célula em dois momentos distintos, pré e pós intervenção. Esse é o teste t para amostras pareadas. Ainda dentro dessa abordagem temos o teste t para amostras não pareados onde o interesse do teste é verificar a existência de possível diferença

entre amostras que não apresentam qualquer característica de dependência entre elas. Pode ser empregado na comparação de resposta de diferentes (~~doses~~) drogas, por exemplo. Os testes anteriores são empregados em situações que tenham duas amostras em comparação.

Para casos onde existem duas ou mais amostras são empregadas outras ferramentas estatísticas sendo a mais conhecida a análise de variância ou ANOVA. Esta ferramenta usa a razão F entre a variância explicada e a variância residual. A tomada de decisão ocorre quando a razão F é superior a um valor estabelecido de acordo com os graus de liberdade do experimento.

Dentro dessa classe onde se considera o conjunto de dados como normal ainda temos a análise de regressão linear simples e a multipla. Essas ferramentas buscam relacionar um conjunto de variáveis explicativas com a variável que pode ser explicada pelas anteriores, ou seja, que depende das variáveis explicativas. Então, vamos ter a variável dependente como variável resposta e as demais como variáveis explicativas

ou variáveis independentes. Em experimentos onde seja conhecido um fenótipo, por exemplo, hipertensão arterial podemos ter este como a variável dependente e o conjunto de genes que podem (~~ter~~) apresentar algum impacto na resposta hemodinâmica como variáveis independentes. Usando a linguagem R de programação como referência podemos induzir o modelo geral a ser implementado: (~~Y ~ a~~)

Y ← lm (hipert ~ gene$_1$ + gene$_2$ ...). Aqui apresentamos a variável 'hipert' como variável dependente e os genes como variáveis explicativas. Obviamente que estamos considerando o conjunto de dado como tendo uma distribuição normal.

Previamente, falamos sobre a distribuição do tipo de dado que estamos trabalhando como possuindo uma distribuição considerada normal ou não normal. Esta decisão sobre qual (~~distribuição~~) distribuição estamos trabalhando é tomada baseado no formato da curva de distribuição e também através de métricas como média, mediana e moda coincidentes ou muito próximas. O formato da curva (~~é~~) segue a silhueta de um sino.

 , onde estes três parâmetros são localizados.

Entretanto, os dados que são produzidos em experimentos de novas drogas, por exemplo, em espectrometro de massa, não possuem essa característica de distribuição normal. Neste caso, existe ~~outro~~ outra abordagem que auxilia na tomada de decisão. Estas são as estatísticas não paramétricas como dito anteriormente. Podemos empregar o teste de Wilcoxon quando o experimento possui características de dependência entre amostras. Neste último caso, é empregado em situações de operação onde seriam aplicado um teste t para amostras pareadas. Temos o teste de Kruskall-Wallis que é empregado em experimentos com duas ou mais amostras independentes entre elas. Existe ainda o teste de Mann-Whitney que pode ser empregado em situações onde existe alguma dependência e pode ser aplicado em duas ou mais amostras. Dentro deste caso temos ainda os testes de aderência que levantam os pressupostos de normalidade do conjunto de dados.

Essas ferramentas são usadas para a extração de dados relevantes

como genes, proteína, metabólitos e novas drogas. São necessários para que se extraia informação relevante por conta do ambiente muito ruidoso. Essas ferramentas estabelecem critérios para mostrar a presença de alguma proteína de interesse.

Uma outra abordagem que pode ser empregada envolve o emprego de modelos lineares generalizados (GLM) que auxilia na previsão, classificação de novos dados. Este conjunto de técnicas empregam desde da regressão linear simples e multipla a modos logísticos. Dentro dessa abordagem podemos fazer uso de ferramentas como a regressão logística que pode classificar uma atividade enzimática em dois estados: ativo e inativo ou em zero(∅) e um (1) como variável dependente e os demais variáveis com explicativas. Com esta ferramenta é possível classificar um conjunto de dados em dois grupos distintos. Muniz et al (2009) usando abordagem de regressão logística classificou um grupo de pacientes com Parkinson contra um grupo controle usando variáveis de ~~(locomoção)~~ locomoção. Previamente os dados foram transformados para o

hiper espaço usando a distância Euclidiana. Neste caso, foi usado o grupo controle como referência (~~e os parentes~~) para o centróide. Foi mostrado dois diferentes grupos organizados no diagrama de dispersão mostrando uma ferramenta computacional útil na separação de diferentes conjuntos de dados.

Ainda é possível aplicar a regressão de POISSON para dados de contagem quando se considera um período de tempo. Os dados dessa distribuição apresentam um formato hiperbólico da contagem de eventos no período considerado. Temos ainda a binomial negativa que também é empregada em dados de contagem durante um período, contudo, agora considera-se a presença de um elevado número de zeros (~~nos dados~~) na distribuição do conjunto de dados.

Podemos fazer uso ainda de modelos de regressão de modelos mistos onde é considerado a influência de variáveis que (~~não~~) estão influenciando o experimento de forma indireta, por exemplo, resposta temporal quando são avaliados diferentes (~~concentrações~~) concentrações de uma droga na corrente sanguínea ao longo do experimento. Estas respostas precisam serem incluídas na modelagem.

NOME: Jucimildo Silva Cerqueira



Os modelos lineares generalizados possuem muitas aplicações. Seu desenvolvimento ocorreu na década de 1980 para atender justamente à situações onde não era possível garantir uma distribuição normal de dados. Dessa forma, essas ferramentas podem ser empregados na descoberta de novos alvos de interesse como novas drogas, diferentes sítios de ação enzimática entre outros na medida que é possível usar informações ordinais, como as passos de uma rede metabólica. Em situações onde existe atividade versus inatividade. De maneira geral ~~é como~~ são ferramentas que permite trabalhar com informações numéricos como carga genética quanto com a função que um gene pode exercer.

NOME: Fernando Vieira Carqueira

No. fl. 10

Quando se trata de dados ~~genômicos~~ ômicos, por exemplo, estamos falando de um grande volume de dados. Esse grande volume de dados apresenta grande dificuldade de análise e, então, são empregadas ferramentas como a análise de componentes principais justamente com o objetivo de reduzir a dimensionalidade do conjunto de dados. A análise de componentes principais extrai informações relevantes considerando a variância dos dados e gerando um novo conjunto de coordenadas (~~pontos~~), os componentes principais que são uma combinação linear das demais variáveis. Esta ferramenta opera na direção de maior variância e esta é considera como a primeira componente. A segunda componente é obtida após a remoção da variância da primeira componente e, então, a variável com a segunda maior variação é considerada e assim até o limite de ~~interesse~~ interesse ou que o conjunto de dados permitem ou por alguma metodologia que determine a priori o número de componentes a ser empregado.

Podemos considerar ainda ferramentas de agrupamentos não supervisionados

NOME: Lucinildo Silva Cerqueira

No. fl. 11

como o K-means. Esta ferramenta pode ser empregada em contextos onde não se sabe exatamente o comportamento dos variáveis. Portanto, pode agrupar dados com características semelhantes baseado na média dos vizinhos. Inicialmente, os valores mais próximos são ~~agrupados~~ agrupados de forma não supervisionada e pode ser empregado para a classificação de comportamentos ou respostas semelhantes de conjunto de genes por exemplo.

NOME:



# 4 - Modelagem de Redes Biológicas

Muitas funções celulares são orquestradas para uma rede global de funções que cruzam diversos comados. Estas estruturas podem ser modelados para entender o funcionamentos de redes bioquímicas que conectam redes de proteína-proteínas, redes de genes entre outros.

A modelagem de redes biológicos (~~regras~~) pode ser representada como:

$Y = G(V, E)$, onde $V$ representa (~~os~~) os nodes da rede e $E$ representa a relação entre os nodos. Cada rede possui topologia única a depender dos relações entre o fenômeno de interesse. Geralmente, os nodos podem ser representados de forma a mostrar alguma hierarquia entre eles. Para isso, podem ser usadas informações de contagem, por exemplo, o número mais elevado é representado pelo conjunto de dados que apresentou maior frequência no conjunto de dados. Se estiver trabalhando com dosdartas de raios genes, por exemplo, seria o gene mais evidente através do nodo que o representa. Já a relação entre as entidades pode ser avaliada através da correlação de Pearson ou Spearman a força da

relação entre uma substância e outra podem ser estabelecida. Em processos de descobertas de novos fármacos podem ser empregados em intervalos de correlação a partir dos quais podem considerar relevante a partir de 95% dessas associações.

A modelagem de redes biológicos podem ser aplicados à redes de proteínas por meio de regulação alostérica. Esta regula a classe de conexão de (~~proteínas~~) proteomas de enzima metabólicas que trabalham como ativador ou inibidor dessas enzimas. Também podem conectar fosfoproteína ou fatores de transcrição proteômica para a transcrição de genes alvo.

Também podemos ter redes gen-gene que vai expor a influência de cada um na função alvo. Esse tipo de rede pode ser empregado na descoberta de funções que são modulados pelo grupo de genes dentro de cada rede. Temos ainda rede composta por proteína e DNA, redes lipídicas.

Para a montagem dos redes é observada a quantidade de vezes que uma substância apresenta dentro da massa de dados. Se o interesse é fazer o levantamento de quais genes estão presentes conta-se a quantidade de vezes

que esta molécula apresenta e em seguida como cada gene se relaciona com os demais através da contagem do número de vezes que relação de um gene passa por um nodo. Em processos de mineração de textos para a busca de fenótipos associados com genes é empregado a ~~(estologia)~~ ontologia que segue a relação de fenótipos já conhecidos para a descoberta de novos genes através da rede biológica.

Atualmente, existem ferramentas computacionais que podem ser empregados para a modelagem de redes, contudo, são extremamente caras e, ainda assim, são fornecidos poucos acesso aos algoritmos que são utilizados. Dessa maneira, o emprego de ferramentas de código aberto, como a linguagem Python e R se apresentam como alternativas extremamente atraentes e de grande confiabilidade tornando-as adequada para o ambiente acadêmico. Portanto, somente o investimento no treinamento em programação, estatística e extensa base em áreas como biologia molecular, física, química e matemática pode gerar um enorme salto (na ~~próxima área~~) na produção de conhecimento através do acesso facilitado que essas tecnologias podem proporcionar.

NOME: Lucindo Silva Junqueira

No. fl. 15

9 - Fundamentos teóricos de algoritmos de biologia computacional e Bioinformática

O desenvolvimento de algoritmo é um processo de desenhar, implementar, analisar e validar o mesmo. Em biologia computacional, desde os primórdios são empregados sequência para o armazenamento de informações genéticas. Cada sequência dessas guardam notações sobre os nucleotídeos que formam uma proteína, por exemplo. Este é um formato onde os algoritmos usados para a decodificação de substâncias fazem buscas recorrentes como Needleman-Wunch que diminui o número de células que precisam ser computadas. Esses algoritmos devem realizar tarefas como o levantamento da estrutura de proteínas mostrando seus sítios de interação com outras proteínas ou outra substância qualquer.

Para dados armazenados no "PUBMED" é possível usar algoritmos como FASTA que permitem uma rápida recuperação da informações. Contudo, devido ao rápido desenvolvimento tecnológico atual onde os métodos metodologios de aprendizado profundo passaram a ser empregados é necessário pensar sobre esse

de armazenamento. Basicamente métodos de aprendizado profundo trabalham com dados dispostos em tabelas. Onde as linhas representam amostras e as colunas representam características. A partir dessas estruturas que é possível fazer uso dessas informações. Então, os algoritmos devem ser pensados de forma que torne a extração das informações de forma mais rápida conforme o tipo de abordagem que vai ser aplicada. Considerando que já existe uma vasta quantidade de informações no formato sequencial é interessante que os algoritmos possam entregar os dados já disponíveis em formato de data-frame para a etapa de modelagem e extração de informações.