



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 no grande ~~maior~~ 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~~ 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 metabolito poderia ser considerada relevante para estudo em questão.

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

NOME: Lucinildo Silva Gurgueira

No. fl.

2

que pretende desenvolver sobre levantamentos previamente. Este levantamento teórico prévio mostra os 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 depende do tipo de experimento que este sendo realizado.

Para decidir qual seria a melhor abordagem em cada experimento geralmente conta-se com a ~~baseada~~ experiência de pesquisadores seniores ou é possível seguir diretrizes básicas para análise de seus dados. A primeira abordagem para se conhecer qual o tipo de dados estamos tratando é necessário fazer uma análise exploratória (~~análise~~) do conjunto de dados a ser tratado. 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 que os dados devem ser tomados para análise do conjunto.



NOME: Lucas Henrique Silveira Vergueiro

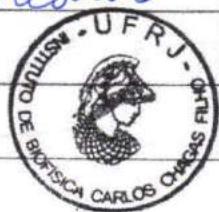
No. fl.

3

de dados. Após a anterior análise de dados é necessário a ~~o~~ realização de testes de одеренда à diferentes distribuições amostrais. As 'chutes' unisão sobre qual teste deve ser aplicado ~~o~~ 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 compor amostras ~~o~~ contra um valor de referência usando o teste T para amostras independentes. É empregado em condições onde o pesquisador quer compor dados de um novo experimento contra um valor de referência. Ainda podemos fazer comparações com amostras dependentes. Imagine que queremos estudar o crescimento de um cérebro 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 pareadas onde o interesse do teste é verificar a existência de possível diferença.



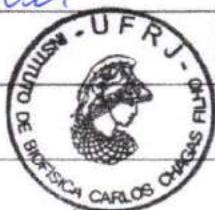
NOME: Fábio Mílton Selva Gergenra

No. fl.
4

entre amostras que não apresentam nenhuma característica de ~~de~~ dependência entre elas. Pode ser empregado na comparação de respostas de diferentes (~~tipos~~) drogas, por exemplo. Os testes anteriores são empregados em situações que envolvem duas amostras em comparação.

Pela razão onde existem duas ou mais amostras são empregados outros ferramentas estatísticas sendo a mais conhecida a análise de variâncias 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 múltipla. Esses ferramentas buscam relacionar um conjunto de variáveis explicativas com a variável que pode ser explicada pelos anteriores, ou seja, que depende das variáveis explicativas. Então, vamos ter a variável dependente como variável resposta e as demais como variáveis explicativas.



NOME: Henrique Silva Berguera

No. fl.

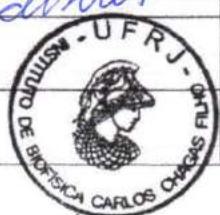
5

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 (~~possuir~~) apresentar algum impacto na resposta hemodinâmica como variáveis independentes. Usando a linguagem R de programação como referência podemos indicar o modelo geral a ser implementado:

(Yao)

$Y \leftarrow lm(\text{hypert} \sim \text{genes} + \text{genz}, \dots)$. Aqui apresentamos a variável 'hypert' como variável dependente e os genes como variáveis explicativas. Obviamente que estamos considerando o conjunto de dados como tendo uma distribuição normal.

Presto-me, fazemos sobre a distribuição do tipo de dados 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 evidentes de muito próximas. O formato da curva (σ) segue a silhueta de um sino.

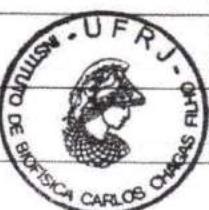


NOME: lucenilda silva eugêneira

No. fl.
6

Na, onde estes três parâmetros são todos iguais. Entretanto, os dados que não produzidos em experimentos de novos 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. Estes são os 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 pairedness 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 disto, temos ainda o teste de ordenância que levantam os pressupostos de normalidade de conjunto de dados.

Essas ferramentas são usadas para a extração de dados robustos

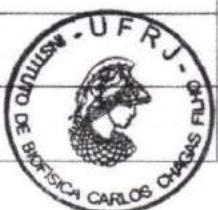


NOME: Bernardo Silveira Gurgelha

No. fl.
2

como genes, proteína, metabólitos e nossas drogas. São necessários para que se extraia informação relevante por conta do ambiente muito ruinoso. Esses ferramentas estabelecem critérios para mostras apresentar 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 múltipla a modelos logísticos. Dentro dessa abordagem podemos fazer uso de ferramentas como a regressão logística que pode classificar uma amostra de enzimática em dois estados: ativo e inativo ou um zero (0) e um (1) como variável dependente e os demais variáveis com explicativos. Com esta ferramenta é possível classificar um conjunto de dados em dois grupos distintos. Menez et al (2009) usando abordagem de regressão logística classificou um grupo de pacientes com Parkinson contra um grupo controle usando variáveis de ~~ambos~~ locomotiva. Previoamente os dados foram transformados para o



hiper exponencial usando a distância Euclídea. Neste caso, foi usado o grupo controle como referência (~~e os pacientes~~) para o artrose. Foi mostrado dois diferentes grupos organizados no diagrama de dispersão mostrando uma ferramenta computacional feita na suposição 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 (~~zeros~~) na distribuição do conjunto de dados.

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

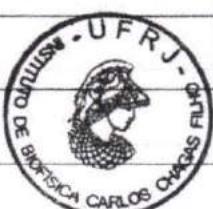


NOME: Henrique Sula Bergmeira

No. fl.

9

Os modelos lineares generalizados possuem muitas aplicações. Seu desenvolvimento ocorreu na década de 1970 para atender justamente à situações onde não era possível garantir uma distribuição normal de dados. Dessa forma, esses ferramentas podem ser empregados na descrição de muitos tipos de interesse como níveis drogas, diferentes sitios de ação farmacêutica entre outros na medida que é possível usar informações ordinárias, como os pesos de uma rede metabólica. Em situações onde existe atividode versus inatividade. De maneira geral ~~é~~ ^{São} ferramentas que permitem trabalhar com informações numéricas como carga genética quanto com a função que um gene pode exercer.



Quando se trata de dados ~~genéricos~~ omícos, por exemplo, estima-se que grande volume de dados com grande variação apresenta grande dificuldade de análise e, então, são empregados ferramentas como a análise de componentes principais juntamente com o objetivo de reduzir a dimensionalidade do conjunto de dados. A análise de componentes principais extrai informações relevantes considerando a variação dos dados e gerando um novo conjunto de coordenadas (~~base~~), os componentes principais que são uma combinação linear das demais variáveis. Esta ferramenta opera na direção da maior variação e é considerada como a primeira componente. A segunda componente é obtida após a remoção da variação da primeira componente, então, a variável com a segunda maior variação é considerada e assim até o limite de ~~satisfazer~~ interesse ou que o conjunto de dados permitem ou por alguma metodologia que determine a própria o número de componentes a serem empregados.

Podemos considerar ainda ferramentas de agrupamentos não supervisionados

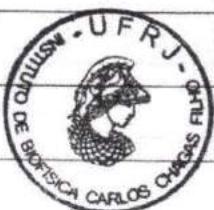


NOME: Lucenilda Silva Arguello

No. fl.

11

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



4 - Modelagem de Redes Biológicas

Muitas funções celulares são organizadas para uma rede global de funções que cruzam diversos canais. Estes estruturas podem ser modelados para entender o funcionamento de redes bioquímicas que conectam redes de proteína-proteínas, redes de genes entre outros.

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

$G = (V, E)$, onde V representa (~~os~~) os nós da rede e E representa a relações entre os nós. Cada rede possui topologia única a depender das relações entre o fenômeno de interesse. Geralmente, os nós podem ser representados de forma a mostrar alguma hierarquia entre elas. 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 tratando com dados de nós genes, por exemplo, seria o gene mais evidente através do nó que representa. Já a relação entre os nós pode ser avaliada através do cálculo de Pearson ou Spearman a força da

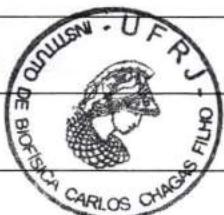


relações entre uma substância e outra podem ser estabelecidas. Em processos de descobertas de novos fármacos podem ser empregados em intervalos de correlação a partir dos quais podem considerar relevantes a partir de 95% dessas associações.

A modelagem de redes biológicas podem ser aplicadas à redes de proteínas por meio de regulação cíclica. Esta regula a classe de complexo de ~~(proteínas)~~ proteomas de enzima metabólicas que trabalham como ativador ou inhibidor dessas enzimas. Também podem conectar fosfoproteína ou fatores de transcrição proteómica para a transcrição de genes alvo.

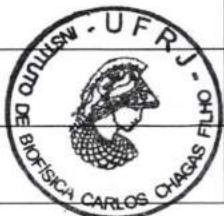
Também podemos ter redes gen-gene que vai expor a influência de cada um na função alvo. Esse tipo de rede pode ser empregado na descoberta de funções que não moduladoras pelo grupo de genes dentro de cada rede. Temos ainda rede composta por proteína e DNA, redes lipídicas.

Pela montagem das 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 vizinhos que relocação de um gene possa ter um nodo. Em processos de mineração de dados para a busca de fenótipos associados com genes é empregado a ~~ontologia~~ ontologia que segue a relocação de fenótipos já conhecidos para a descoberta de novos genes através da ~~(A)~~ rede biológica.

Atualmente, existem ferramentas computacionais que podem ser empregadas para a modelagem de redes, contudo, são extremamente caras e, ainda assim, são fornecidas poucos recursos 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 atrativas e de grande confiabilidade tornando-as adequada para o ambiente acadêmico. Portanto, somente o investimento no treinamento em programação, estatística e extensas bases em áreas como biologia molecular, física, química e matemática pode gerar um enorme salto (~~no processo de~~) na produção de conhecimento através do uso da faculdade que esses tecnólogos podem proporcionar.



NOME: Henrique Silva Lengueira

No. fl.

15

9 - Fundamentos teóricos de algoritmos de biologia computacional e Bioinformática

O desenvolvimento de algoritmos é um processo de desenhar, implementar, analisar e validar o mesmo. Em biologia computacional, desde os primórdios são empregados sequências para o armazenamento de informações genéticas. Cada sequência dessas (~~pequena~~) quase sempre contém informações sobre os (~~índices~~) 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 etapas que precisam ser computadas. Esses algoritmos devem realizar tarefas como o levantamento da estrutura de proteínas mostrando seus sites de interação com outras proteínas ou outras substâncias qualquer.

Para dados armazenados no "PUBMED" é possível usar algoritmos como FASTA que permitem uma rápida reprodução das informações. Contudo, diante do rápido desenvolvimento tecnológico atual onde (~~existem~~) metodologias de aprendizado profundo passaram a ser empregadas é necessário pensar sobre esse



NOME: Juvenildo Silva Bergueira

No. fl.

16

de armazenamento. Basicamente métodos de aprendizado profundo trabalham com dados dispersos em tabelas. Neste os linhas representam amostras e os 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 se 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.

