



CODIGO DA PROVA:

RP001 / 0012



UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS
CONCURSO:

FOLHA DE RESPOSTA

Importante: O código da prova só será colocado na entrega da prova ao fiscal. As provas serão escaneadas e enviadas aos membros da banca avaliadora sem o nome do candidato.

9 - Aplicações de bioinformática no estudo de patologias humanas

No estudo de patologias humanas, a bioinformática se destaca em diferentes frentes. A frente de armazenamento e disponibilização de dados multiômicos em bancos de dados biológicos públicos constitui um importante ponto de partida para estudos das áreas biológicas, médicas e farmacêuticas. Diferentes tipos de dados podem ser disponibilizados em seus formatos padronizados, facilitando seu uso e integrado em estudos. Dentre os bancos de dados disponíveis relacionados a patologias humanas, podemos citar:

- (i) dbSNP, que contém informações sobre diferentes tipos de mutação de único nucleotídeo que foram detectadas em estudos de genome-wide association (GWAS). Esse banco de dados possui associações funcionais e informações relacionadas ao perfil patogênico ou benéfico de tais mutações, fornecendo insights envolvendo sua associação a doenças genéticas, por exemplo;
- (ii) ClinVar, inclui também informações relacionadas ao perfil dessas mutações;
- (iii) OMIM: banco de dados completo de mutações com herança mendeliana.

Além de inúmeros outros bancos de dados com informações de associação de mutações a patologias humanas, como o SwissVar⁵ e HumanVar.

Outra importante aplicação da bioinformática constitui ferramentas que permitem analisar grandes conjuntos de dados genômicos, como aqueles provenientes de NGS (sequenciamento de nova geração) em estudos de GWAS. Esses algoritmos permitem alinhar inúmeros genomas sequenciados, permitindo uma análise comparativa aprofundada que viabiliza a identificação de variantes de nucleotídeo únicos e também de estudos de associação e modelagem de risco. Ferramentas em R e Python, como o biopython e o Bioconductor fornecem bibliotecas/software otimizados para esse tipo de tarefa. Com relação aos estudos de associação, eles permitem relacionar estatisticamente um genótipo sequenciado com um fenótipo ~~o~~ específico, como, por exemplo, a ocorrência de uma patologia. Para tanto, são utilizados algoritmos que aplicam inferências estatísticas, como o teste de χ^2 ou permitem corrigir o viés de múltiplas comparações, como a ~~teste~~ correção de Bonferroni.

A bioinformática também disponibiliza ferramentas para modelar o risco associado a determinado perfil (mutações genéticas). Nesse caso, podem ser utilizados modelos de regressão linear múltiplas, quando se está lidando com variáveis numéricas ou modelos de regressão logística, no caso de ~~uma~~ variáveis categóricas. Uma vez construído, esse modelo pode ser ~~o~~ avaliado para compreender o quanto da variância, nos dados influencia, no risco final de determinada doença, utilizando, para tanto, a métrica R^2 .

Além da modelagem de risco, baseada em um perfil genético, é possível ~~o~~ construir modelos para prever ~~o~~ determinada variante

genética, como as mutações missense, poderiam levar ao desenvolvimento de patologias humanas. Esse é o caso dos algoritmos de predição funcional, que permitem prever o ~~perfil~~ impacto funcional de determinada mutação, isto é, se é neutra ou deletéria para o funcionamento adequado de um gene ou proteína codificada. Dentre essas ferramentas, modelos como o SIFT e o PolyPhen são bem-estabelecidos para este fim. Outros algoritmos, como o MutPred, por exemplo, permitem ainda obter insights sobre a possível causa do impacto funcional predito. Dentre os principais parâmetros considerados por esses algoritmos em suas predições estão a conservação evolutiva, na qual aminoácidos mais conservados tendem a ter maior importância funcional devido à possível maior pressão seletiva associada. Outros parâmetros ~~que~~ amplamente considerados por algoritmos de predição funcional envolve a alteração de energia livre (termodinâmica), impacto do perfil físico-químico do aminoácido substituído (e.g. tamanho, carga, hidrofobicidade), sítio de ocorrência da mutação (e.g. sítio de ligação a metais e cofatores enzimáticos).

Recentemente, com o advento da inteligência artificial, esses algoritmos têm conseguido resultados de performance preditiva superiores, graças à capacidade aprimorada de modelos como os Large Language Models de aprender/modelar dados ~~textuais~~ como as sequências nucleotídicas e de aminoácido, em padrões não lineares complexos.

Além disso, outra área ~~bastante~~ de atuação bem presente da bioinformática no estudo de patologias humanas inclui a biologia de sistemas, que permite construir redes, como as redes regulatórias gênicas, redes de interação proteína-proteína, redes metabólicas e

redes de sinalização / transdução de sinal. Com o advento de tais ferramentas é possível integrar dados multiômicos e compreender relações entre diferentes membros dessa rede (mó). A partir dessa compreensão, é possível identificar mecanismos centrais para o funcionamento dessa rede, e assim, identificar potenciais alvos terapêuticos a serem modulados. Isso pode ser aplicado no desenvolvimento de terapias antimicrobianas, por exemplo, permitindo a identificação de um gene/proteína central na produção de um fator de virulência, por exemplo.

Ainda considerando aplicações da bioinformática para o estudo de patologias humanas, podemos citar ~~as comparações~~ as análises comparativas de expressão gênica. Ferramentas como o DeSeq2 permitem comparar dados de expressão gênica entre dois cenários distintos, como, por exemplo, a ~~identificação~~ um estado fisiológico com um estado patológico, ou mesmo um perfil de célula saudável com o perfil de expressão gênica de uma célula cancerígena. Com isso é possível identificar genes sub-expressos e super-expressos que poderiam ocasionar ou contribuir para dada patologia. A partir da análise diferencial destes perfis de expressão, é possível identificar biomarcadores para auxiliar no estabelecimento, prognóstico, diagnóstico precoce, e até direcionamento para terapias mais efetivas. Isso é particularmente relevante em cenários de câncer.

Os estudos de expressão genética também se beneficiam de ferramentas de bioinformática voltadas à visualização e análise de dados, permitindo a obtenção de insights relevantes ~~sobre~~ sobre dados de microarray e RNA-seq. Com essas ferramentas é possível agrupar genes com base em sua expressão ou até mes-

mo identificar genes correlacionados, auxiliando na compreensão de mecanismos associados a doença e como esses genes se relacionam, por exemplo, em uma rede regulatória gênica. Essa abordagem ~~está~~ também permite identificar padrões característicos de desregulação da expressão gênica em células de câncer, direcionando a terapias mais efetivas.

Por fim, com o advento de tecnologias modernas de aprendizagem profunda, como as redes neurais convolutionais, tem sido cada vez mais frequente o desenvolvimento de modelos para detecção / auxiliar na identificação de padrões característicos de certos tipos de câncer a partir de exames de imagens, como RMN e tomografias computadorizadas. Essas aplicações auxiliam médicos durante o diagnóstico e ~~o tratamento~~ obtiveram sucesso em tarefas de detecção de melanomas e cânceres de mama. Em alguns casos esses modelos superaram inclusive a performance individual de único médico, demonstrando o potencial desses instrumentos de bioinformática no contexto da saúde. Essa elevada precisão decorre da capacidade de modelos de redes neurais convolutionais em extrair padrões complexos de imagem, como as de exame de imagens, gerando representações simplificadas favoráveis ao aprendizado.

A bioinformática permite ainda realizar análises diferenciais de expressão gênica no sentido de mapear a superexpressão de genes associados ao perfil de resistência de células tumorais a medicamentos quimioterápicos.

7 - Bioinformática aplicada a análise epigenômica.

As ferramentas de bioinformática permitem, inicialmente, a interpretação de dados epigenômicos, como aqueles provenientes de CHIP-seq e bisulfito-sequeenciamento. A epigenética consiste em mecanismos que ocorrem sobre o DNA sem que sua sequência de nucleotídeos propriamente dita, seja alterada. Esse tipo de modificação afeta o perfil de expressão gênica, podendo silenciá-los ou ~~ativá-los~~ ativá-los. Dentre os principais mecanismos epigenéticos estão as modificações covalentes em proteínas histonas, metilação de ~~citosinas~~ citosinas (5mC) e ~~interferências por RNAs não codificantes (ncRNAs)~~ ncRNAs.

~~O DNA se compacta, dentre:~~

A interação do DNA com proteínas histonas é essencial para promover sua compactação, em estruturas consideravelmente menores. O nível de organização (compactação) da cromatina, que depende da interação com proteínas histonas, pode resultar em regiões de heterocromatina e eucromatina, sendo a primeira forma densamente compactada e ligada a histonas, enquanto a última, está pouco ligada a histonas e, conseqüentemente, pouco compactada (mais disponível). Regiões mais disponíveis podem ~~interagir~~ interagir mais livremente com o complexo de transcrição e permitir a expressão gênica. Assim, modificações em proteínas histonas ~~alteram~~ alteram a organização do DNA, reduzindo a afinidade pela dupla hélice negativa, tornando-o menos compacto.

Já modificações diretas em bases nucleotídicas, como a metilação, faz com que o DNA não consiga ser decodificado, silenciando-o ~~por meio de RNAs não codificantes~~.

Considerando esses mecanismos citados, dois métodos experimentais são amplamente utilizados para monitorar/perfilar esses tipos de alterações: (1) ChIP-seq, que envolve a imunoprecipitação de cromatina seguida pelo sequenciamento, e (2) bisulfito sequenciamento, que envolve um ensaio com bisulfito, o qual interage apenas com citosinas não-metiladas, convertendo-as em última instância para timinas.

Nesse contexto, a bioinformática possui um papel central na análise e interpretação desses resultados. Com relação ao ChIP-seq ~~por~~ e o bisulfito-sequenciamento, ambas as abordagens envolvem sequenciamento de nova geração (NGS), dependendo, portanto, de algoritmos de bioinformática para montar os pequenos e longos fragmentos sequenciados. Para tanto, algoritmos como o de overlap sequence, permitem ~~de~~ montar contigs maiores a partir da sobreposição de fragmentos.

Os algoritmos de montagem de genoma também permitem alinhar as regiões sequenciadas com um genoma de referência, viabilizando a construção de um mapa/perfil epigenético, ou epigenômico. Assim, é possível identificar regiões densamente condensadas (ou silenciadas) ou até mesmo comparar dois perfis epigenéticos para compreender quais mudanças ocorreram após determinada exposição ambiental/temporal. Com isso, é possível estabelecer correlações e causalidade ~~de~~ entre ~~o~~ o perfil (mudanças no perfil) e determinado evento.

Com relação ~~ao~~ a montagem das sequências de bisulfito-sequenciamento é importante destacar que alguns métodos/algoritmos precisam fazer considerações adicionais para aumentar a precisão e eficiência do alinhamento. Isso se

faz necessário, pois a técnica do bisulfito sequenciamento em si converte bases utxmg (U) em Timina (T), em última instância. Com isso, o perfil de utxina é ~~afetado~~ subrepresentado, dificultando sua montagem. Apesar de ferramentas tradicionais de montagem conseguirem realizar este tipo de tarefa, outros algoritmos mais focados nesta técnica foram desenvolvidos. Um exemplo disso são os alinhadores de ~~alinhadores~~ ou até mesmo alinhadores que consideram apenas três bases, como o BioMark.

É importante mencionar que ^{também} existem bibliotecas especializadas na ~~análise~~ análise de dados de CHIP-seq, como a MACS.

Outra questão importante com relação à análise de dados de CHIP-seq, é que a interpretação das leituras de bases nitrogenadas pode ser dificultada pela técnica propriamente dita. Uma questão inerente ao CHIP-seq é o fato da sonicação, por exemplo, favorecer múltiplas fragmentações do DNA, dificultando a interpretação direta de seus resultados.

Além disso, é necessário a aplicação de algoritmos robustos para interpretar os sinais de CHIP-seq, permitindo distinguir sinais de ruído. Existem, nesse sentido, algoritmos que viabilizam a distinção do sinal em relação a distribuição de ruídos (sinal de background) conhecidas.

Ferramentas de bioinformática possuem aplicativos relevantes na análise e caracterização do perfil epigenético de células cancerígenas. Uma vez mapeado esse perfil, é possível direcionar terapias para ~~modificar o perfil~~ de expressão gênica mediado pela epigenética.

melhor e adequarem ao perfil epigenômico ma-
peado.

Por fim, destes bancos de dados de informações epi-
genômicas, como o método.

2- Controle da expressão gênica com ênfase em
métodos computacionais.

~~Uma~~ ^{presença de} Uma importante aplicação de bioinformá-
tica ~~sobre~~ ~~consiste~~ consiste na identificação
de regiões/motivos reguladores no DNA.
~~Além de fornecer bases de dados~~. Estes
algoritmos permitem identificar regiões
^{regiões} conhecidas no DNA, como o FIMO, ou predi-
zer regiões regulatórias, como enhancers,
a partir de perfis já conhecidos, considerando
a conservação evolutiva de tais regiões.

Além disso, a bioinformática, através da
biologia de sistemas, permite construir
e conectar diferentes genes, em uma rede
regulatória gênica. Com isso, é possível analisar
como o perfil de expressão gênica é co-dependente
e, assim, identificar alvos chave para modular
em intervenções terapêuticas.

Uma série de ferramentas computacionais tam-
bém podem ser aplicadas ao estudo de RNAs
não codificantes, os quais constituem importantes
reguladores da expressão gênica. Ferramentas
como o ~~MIRANDA~~ ^{MIRANDA}, permitem identificar
regiões em mRNA que são complemen-
tares ao miRNA de interesse. Isso se faz
particularmente importante para identificar
potenciais genes silenciados, dado que
a interação miRNA-mRNA, por complemen-
tariedade, ~~de~~ ^{termina} o silenciamento do
genótipo codificado por aquele RNA. ~~Além~~

Além disso, softwares como o deepRNA utilizam
redes neurais artificiais para identificar sequên-
cias de nucleotídeos com potencial para se
enovelar em ~~em~~ formas de alça ou loop

características dos RNAs não-codificantes, permitindo a predição de novas regiões de ~~qualificação~~ RNA não-codificante.

Também é possível a utilização de ferramentas computacionais para ~~identificação~~ predição de modificações pós-transcricionais específicas envolvidas na inativação ou liberação de proteínas. Como exemplo, ~~temos~~ podemos citar o deepPhos, que realiza predição de ~~sítios~~ sítios de fosforilação.

Além dos mecanismos de regulação da expressão gênica baseados em epigenética, mencionados e discutidos com detalhes no tópico 7, ferramentas preditivas permitem prever sítios de modificação em histonas, ~~temos~~ e predição de sítios metilados no DNA ~~(propensos)~~ (propensos). Isso fornece insights valiosos sobre os mecanismos de regulação da expressão gênica, o que pode potencialmente ser explorado no desenvolvimento de terapias guiadas.

Estudos de metagenômica fornecem ainda dados de entrada para ferramentas computacionais voltadas à identificação de ~~regiões~~ unidades genéticas coordenadas, como os operons. Isso pode ser utilizado a favor de modular essas unidades a fim de inibir o desenvolvimento do patógeno.

Além disso, ferramentas computacionais podem ainda ser usadas no estudo da expressão gênica bacteriana/levedura, a fim de otimizar a produção de medicamentos, bioprodutos ou biocombustíveis, com ampla aplicação no campo da biotecnologia.

Por fim, o controle da expressão gênica pode ser explorado através de estudos ~~de~~ comparativos de expressão gênica a partir de dados de microarray ou RNAseq. A ~~análise~~ análise de tais resultados é possível mapear genes de interesse, como, por exemplo, oncogenes e genes envolvidos na sobrevivência celular.

Como a PI3K ou mTOR, e com isso desen-
volvendo terapias guiadas para o tratamento de
câncer.

Por fim, as ferramentas computacionais ofere-
cem um amplo arsenal de ferramentas para a
construção, análise e simulação de perturba-
ções em redes regulatórias gênicas. Isso
pode ser modelado utilizando conceitos
mais simples, como redes booleanas (ligado ou
desligado), Bayesianas, envolvendo probabi-
lidade de ocorrência, até modelos baseados
em equações diferenciais. É importante mencionar
que os conexões dessas redes e a contribuição
de cada nó depende da integração de dados
experimentais e ajustes computacionais a partir
do feedback experimental. Uma vez validando
essa rede e verificando se ela reflete de fato
a relação entre os genes relacionados, é possí-
vel utilizá-las para ~~testar~~ simulações
de perturbações no seu equilíbrio a fim de
testar quais nós são mais relevantes e influen-
ciam os demais nós (genes). Com isso, é possível
identificar genes-chave de determinada via.