



CODIGO DA PROVA: RP 001 / 0010



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

tópico 7

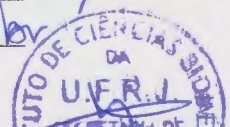
FOLHA DE RESPOSTA

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.

A comunidade científica vem ao longo dos anos ampliando o conhecimento a cerca da estrutura dos genomas, desde os anos 50 do séc 20, mas principalmente a partir dos anos 70 do mesmo século, com o surgimento das tecnologias de sequenciamento de biomoléculas. Entender a estrutura dos genomas, mostra-se fundamental para compreender a complexidade biológica das células em termos funcionais. Já se sabe que o conhecimento que a estrutura e organização dos genomas se deve a interação entre proteínas e DNA. Algumas dessas proteínas estão diretamente associadas a regular o DNA, ou mesmo o próprio ambiente pode influenciar essa regulação. Essa interação entre fatores ambientais e proteínas capazes de promover alterações conformacionais sem comprometer a integridade do DNA é chamada de epigenética. Essa área das ciências biológicas destaca que alguns SINAIS podem ser importantes para analisar como o DNA pode ser regulado. três delas são amplamente discutidas na literatura como: a metilação do DNA, modificação nas proteínas histonas e remodelação da cromatina. A metilação da molécula de DNA pode comprometer diretamente a capacidade das células em expressar

(1)

Determinados genes. Isso se deve ao fato de que a incorporação de grupos metil (CH_3) em geral nas citosinas do DNA é capaz de silenciar genes que podem ter papéis fundamentais na integridade das células. Esses tipos de alterações na estrutura química do DNA podem levar ao surgimento de várias patologias ou distúrbios crônicos. Algumas estratégias no intuito de analisar e investigar essas questões vem sendo orientadas por metodologias interdisciplinares como a bioinformática. Um exemplo de aplicação metodológica é a técnica de ~~bisulfito-seq~~ bisulfito-seq, capaz de observar o grau de metilação a partir do tratamento com bisulfito de sódio no DNA. Inicialmente uma etapa experimental é realizada para tratar o DNA metilado, fazendo o bisulfito alterar as citosinas não metiladas em timinas, mantendo as metiladas como citosinas. Após essa etapa experimental o DNA pode ser sequenciado por métodos como o sequenciamento de Sanger. Etapas computacionais de avaliação de qualidade são feitas e de remoção de adaptadores, como pela ferramenta FASTQC. Essa ferramenta apresenta relações de qualidade dos dados brutos gerados pelo sequenciamento. Após o sequenciamento e o controle de qualidade, as sequências podem ser alinhadas contra genomas de referência a partir de softwares específicos, como o Bismark, no intuito de promover a identificação dessas regiões. Após todas essas passadas é possível entender ~~uma~~ uma infinidade de problemas celulares dando a identificação de um determinado padrão de metilação no genoma. A integração de epigenômica, genômica e transcriptômica pode ainda revelar mais profundamente como a complexidade das células explica o surgimento e promoção de doenças. Ferramentas como o Galaxy e o pacote de R, bioconductor, podem auxiliar esse entendimento.



Dada a metilção do DNA, alterações nas proteínas histonas também configuram o espectro de análises epigenômicas. Essas proteínas podem também sofrer alterações químicas como metilção e acetilção e modificar as interações com a molécula de DNA, isso leva uma possível alteração no controle de determinadas regiões do DNA, aumentando ou diminuindo sua expressão. Algumas técnicas em bioinformática são capazes de analisar essas proteínas histonas e os locais de interação. Os passos iniciais giram em torno da precipitação dessas proteínas por métodos experimentais. O DNA associado é separado e posteriormente sequenciado em geral pelo método de Sanger. Após as etapas que envolvam a qualidade dos dados pelo FastQC e trimagem pelo trimomatics, as sequências podem ser alinhadas contra genomas de referência usando o HISAT2 ou bowtie. Em seguida outros métodos computacionais auxiliam a identificação das regiões associadas às histonas, por ferramentas MACS e annotação de domínios funcionais pela ferramenta Funmer. Essa metodologia é pipeline de análises, que soma o método de ChIP-seq (Chromatin Immunoprecipitation). Esses dados integrados a dados de RNA-seq, podem ainda potencializar como esses regiões, exploram a dinâmica de expressão, dada a observação de alterações nas proteínas histonas.

Alterações na conformação do genoma por proteínas histonas ou em modificações químicas no genoma, também estão associadas a fatores ambientais. Portanto, entender como o ambiente afeta o dinamismo da expressão dos genes vem sendo um desafio para a comunidade científica, mas que vem ganhando cada vez mais relevância, dada a importância da bioinformática

Tópico 2:

~~Controle da expressão gênica~~
~~Controle da expressão gênica~~

Praticamente todas as funções celulares ocorrem dada a expressão de genes funcionais importantes, que podem estar sujeitos a eventos metabólicos, de comunicação, diferenciação etc. Caso a expressão desses genes possa estar comprometida, a integridade celular também estará. Portanto, é crucial para a célula a capacidade de manter um controle rígido sobre como, onde e quando os genes podem ser transcritos. A regulação gênica passa por várias etapas, que vão desde a transcrição, impedindo ou facilitando na partir de regiões promotoras, a transcrição, após a transcrição se os RNAs transcritos serão degradados ou ~~degradados~~ processados (splicing). Após o processamento se o RNA sofre por acetilação ou fusão do cap 5', por fim, se ~~degrada~~ a tradução e eventos pós-translationais ocorrerem. Avaliar e estabilizar o controle da expressão gênica tem se tornando mais amplo com a inserção de métodos computacionais. Dado o sequenciamento das sequências algumas ferramentas são capazes de apontar regiões promotoras e enhancers em genes e assim ampliar o conhecimento dessas regiões de controle. Algumas exemplos de ferramentas capazes de realizar essa identificação são o MEME e JASPAR. Essas técnicas fornecem a capacidade de avaliar a capacidade de controle de expressão observando se a organização molecular de promotores e enhancers podem estar complementadas, a partir de alinhamentos múltiplos contra bases de dados específicos, como o NCBI.

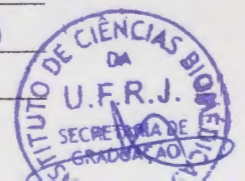
Outras formas de analisar o controle de expressão gênica tornam-se possíveis por ~~estes~~ metodologias de inferência bayesiana e por redes lógicas co-expressão gênica.

Um exemplo de ferramenta computacional capaz de re-criar redes de regulação gênica é o COPASI, que utiliza o algoritmo de Gillespie, baseado no método de Monte-Carlo. Essa ferramenta é capaz de determinar um conjunto de possíveis interações ~~entre~~ entre DNA e proteínas a partir de cálculos matemáticos (diferenças) orientados por reações químicas previamente descritas na literatura. A mesma ferramenta disponibiliza a capacidade de determinar a configuração de vias bioquímicas, que podem ser cruzadas às informações contidas bases de dados como o KEGG.

Tópico 9.

~~Resumo de Tópico 9~~

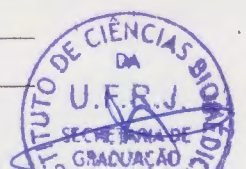
Desde o surgimento das tecnologias de sequenciamento de próxima geração (NGS), uma explosão de dados computacionais ofereceu a comunidade científica a possibilidade de uma análise exploratória mais abrangente. Análises genômicas permitem comparar regiões importantes dos cérebros e assim identificar regiões raras. Essas análises comparativas, também descritas por genômica comparativa, muitas vezes podem destacar genes mutantes associados a ocorrência de diversas patologias. Na bioinformática esse tipo de estratégia pode ser realizada por softwares como o MASSIVE progressive mapper, um alinhador de genomas capaz de destacar rearranjos moleculares e fragmentação dos genes no genoma. Alguns estudos tem direcionado essa abordagem apenas às regiões codificantes, descritas como exoma. Uma vez tendo apenas as regiões codificantes sequenciadas por métodos com Illumina, Ion torrent, Pacbio ou MinION, estes exomas podem ser comparados entre genomas controle, e determinar regiões que representam mutações. Análises de sequenciamento de gene único também podem ser analisadas por ~~diversas~~ técnicas, de alinhamento múltiplo e determinar mutações específicas. Alguns exemplos de alinhadores são: Clostral, MUSCLE, Jaffe e MAFT. Ainda no campo genômico algumas análises estatísticas podem ser aplicadas para avaliar se certas mutacionais podem estar se expandindo ou não em populações. Para tal abordagem é necessário o teste de neutralidade feito em um conjunto de sequências. Um exemplo de teste de neutralidade pode ser descrito pelo método de Tajima-D que permite avaliar pelo número total



de nucleotídeos nos genes, a relação de
sítios de segregação e assim, estimar se
as mutações são, neutras ou não.

Outras análises podem ser aplicadas
para compreender o surgimento de patologias
ou se estudar o aumento da transcrição de
genes. A técnica de RNA-seq permite
avaliar se possíveis genes podem estar,
por exemplo, implicados no surgimento de
câncer. Inicialmente algumas etapas experimentais
são realizadas como extração do RNA por kits,
em seguida, a criação de bibliotecas de ~~RNA~~
cDNA pela atividade da transcriptase reversa.
Após a ligação aos adaptadores e sequenciamento,
as leituras podem ser alinhadas contra o genoma
de referência usando HTSeq, um alinhador e
em seguida, passo por etapas de quantificação
das transcritos, feita pela ferramenta RSEM.
Esta ferramenta usa o método de Expression-
Maximization para a quantificação. As etapas
de expressão diferencial podem ser conduzidas
pelo DESeq2 ou EdgeR. O DESeq2 é um
pacote do R que utiliza modelos lineares
generalizados. Esta ferramenta normaliza os
dados de sequenciamento para possíveis problemas
nas etapas experimentais. Por fim ele vê
a relação a partir dos modelos lineares generalizados
dos dados normalizados com os transcritos.
A variância é feita por estatística
bayesiana e a diferenciação de transcritos
través ~~de~~ amostras e feita por teste
de hipótese (χ^2). Assim é possível avaliar
se a super expressão de genes podem ~~ter~~
impacto na promoção de doenças como o
câncer. Entende-se como a expressão diferencial
implica na promoção de doenças, aumentando
o conhecimento o impacto da expressão genica
em patologias humanas.

7



Além das técnicas em genômica e transcriptômica a bioinformática também contribui para ampliar o conhecimento sobre a estrutura e dinâmica das proteínas. Grande parte das patologias em humanos se deve ao fato de surgimento de proteínas mal formadas seja por mutações ou problemas na regulação gênica. Um campo da biologia computacional chamada de biologia ~~estrutural~~ ^{estrutural} é capaz de simular a geometria espacial de proteínas, sua especificidade de interação molecular (Docking) e a dinâmica molecular desses biomoléculas.

Obter modelos 3D de proteínas pode ser possível de várias formas. Por métodos experimentais como a cristalografia de raios X ou ressonância magnética. Por métodos computacionais, é possível por modelagem comparativa, threading e de novo. Por modelagem a sequência a ser modelada precisa ter ao menos 30% de identidade/similaridade contra proteínas de ~~referência~~ referência, no caso o template. Ferramentas como o Modeller ou SWISS Model são capazes de gerar estruturas 3D utilizando bancos de dados, que são cálculos matemáticos capazes de avaliar números de interações intermoleculares, distâncias em Angstroms e forças. O Modeller usa o campo de força CHARMM e o SWISS Model o ~~CHARMM~~ PROTONS. Uma vez os modelos feitos eles podem ser validados por gráficos de Ramachandran, que avaliam a qualidade estereométrica dos ângulos e verifica se há choque eletroquímico pelas nuvens eletrônicas. Outra forma de validação pode ser pelos DOPE score que apresenta uma lista de valores de energia para cada modelo gerado. O método de threading usa as estruturas secundárias para a reconstrução da estrutura da proteína alvo. Algumas bases de dados como o I-TASSER são capazes de determinar essas estruturas

Uma vez que não exista template e nem estruturas secundárias para serem usadas de molde, a técnica de novo pode ser aplicada para determinar os modelos 3D. Esses algoritmos se baseiam na configuração das aminoácidos e consideram os tipos de interações químicas capazes para promover o enovelamento. Bases de dados como o Robetta podem fornecer estruturas 3D a partir desse método. As validações das proteínas podem seguir o mesmo padrão dos gráficos de Ramachandran e pontuação Dipe, mas também são gráficos de RMSD que calculam a raiz quadrada média dos carbonos α na estrutura.

Uma vez que as estruturas tenham sido determinadas, elas podem sofrer várias análises posteriores. Se for uma proteína mutante analisar em softwares como PyMOL ou Chimera pode revelar a causa da mal funcionamento da proteína. Se forem proteína "Sarcavets" mas associadas a algum tipo de patologia, técnicas como o Docking molecular podem avaliar a interrupção da função desta proteína (identificando ligantes que interage multivalentemente com a proteína e inibe sua função). Softwares como o Dock6 ou MOE são capazes de avaliar o número de interações intermoleculares exercidas entre o complexo proteína-ligante e pela variação de energia livre apontar os ligantes mais favoráveis. Alguns softwares utilizam os campos de força para determinar essas interações, mas algumas metodologias mais recentes utilizam IA a ~~propósito~~ ^{propósito} de redes neurais para calcular essas interações e forças. Alguns procedimentos devem ser realizados na estrutura antes do Docking, como avaliar as cargas das aminoácidos, principalmente da região de interesse da proteína determinada por grids. Algumas ferramentas como o ProPKA podem fazer essa validação ou pelo Chimera.

Etapas de dinâmica molecular também auxiliam o entendimento sobre interações e sítios de ligação celular associados a casos de patologias em humanos. Ferramentas como o GROMACS e Amber são capazes de calcular a posição inicial dos átomos e as forças covalentes e eletrostáticas que influenciam no comportamento dos átomos. Os cálculos avaliam a diferença inicial e final para calcular a trajetória, e assim simular a dinâmica destas biomoléculas.

Outras abordagens, computacionais para entender como interações entre proteínas podem estar associadas a promoção de doenças pode ser realizada por técnicas de PPI - protein-protein interaction. Algumas plataformas computacionais como o STRING fornecem gráficos de redes de interação a partir da seleção de um alvo específico. O STRING se baseia em bases de dados com sequências anotadas por dados experimentais, como UniProt. Porém, técnicas com data-mining também são empregadas para consultar diretamente na literatura sobre experimentos que descrevem alguma rede de interação.

Em conjunto, esses métodos vem contribuindo para entender mecanismos de doenças hereditárias, controle de patógenos e descoberta de possíveis tratamentos contra doenças.