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 a estrutura e organização dos genomas se deve a inter-conexão entre proteínas e DNA. Algumas dessas proteínas estão diretamente associadas a sequestrar 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 epigenômica. Essa área das ciências biológicas destaca que alguns sinais podem ser apontados para analisar como o DNA pode ser regulado. Três deles são amplamente discutidos 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

Comissão Organizadora do Concurso
Gabinete da Direção

determinados genes. Isso se deve ao fato de que a incorporação de grupos metil (CH_3) em geral nos citosinas do DNA é capaz de silenciar genes que podem ter papeis 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 [illegible] bissulfito-Seq, capaz de observar o grau de metilação a partir do tratamento com bissulfito de sódio no DNA. Inicialmente uma etapa experimental é realizada para tratar o DNA metilado, fazendo o bissulfito 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. Esta ferramenta apresenta relatórios 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 todos esses passos é possível entender uma infinidade de problemas celulares dado 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 do R, bsseq conductor, podem auxiliar esse aprendizado.

Além da metilação do DNA, alterações nas proteínas Histonas também configuram o espectro de análises epigenômicas. Essas proteínas podem também sofrer alterações químicas como metilação e acetilação 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 de avaliação da qualidade dos Dados pelo FastQC e trimagem pelo trimmomatics, 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 as Histonas, pela ferramenta MACS e anotação de domínios funcionais pela ferramenta HMMER. Essa metodologia e pipeline de análises, descrevem o método de CHIP-Seq (CHromatin ImmunoPrecipitation). Esses dados integrados a dados de RNA-Seq, podem ainda potencializar como essas regiões explicam a dinâmica de expressão, dado a observação de alterações nas proteínas Histonas.

Alterações na conformação do genoma por proteínas, Histonas ou em modificações químicas no genoma, também estão atreladas a fatores ambientais. Portanto, entender como o ambiente altera o dinamismo da expressão dos genes vem sendo um desafio para a comunidade científica, mas que vem avançando cada vez mais claramente, dada a presença da bioinformática

tópico 2.

[redacted]

Praticamente todas as funções celulares ocorrem dada a expressão de genes funcionais importantes, que podem estar associados 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 a partir de regiões promotoras, a transcrição, após a transcrição se os RNAs transcritos serão degradados ou [redacted] processados (splicing). Após o processamento se o RNA sofre poliadenilação ou inserção do cap 5'; por fim, se [redacted] a tradução e eventos pós traducionais ocorrerão. Avaliar e analisar o controle da expressão gênica vem se tornando mais amplo com a inserção de métodos computacionais. Dado o sequenciamento dos genomas algumas ferramentas são capazes de apontar regiões promotoras e enhancers em genes e assim ampliar o conhecimento dessas regiões de controle. Alguns 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 comprometidas, a partir de alinhamentos múltiplos contra bases de dados específicos, como o NCBI.

Outras formas de analisar o controle de expressão gênica tornam-se possível por método bayesiana e por redes de co-expressão gênica.

Um exemplo de ferramenta computacional capaz de re-criar redes de regulação gênica é o COPASI, que utiliza o algoritmo de Gibbs, baseado no método de Monte-Carlo. Essa ferramenta é capaz de determinar um conjunto de possíveis interações entre DNA e proteína a partir de cálculos matemáticos diferenciais orientados por reações químicas previamente descritas na literatura. A mesma ferramenta disponibiliza a especificidade de determinar a configuração de vias bioquímicas, que podem ser buscadas as informações contida bases de dados como o KEGG.

Tópico 9.

[redacted]

Desde o surgimento das tecnologias de sequenciamento de próxima geração (NGS), uma explosão de dados computacionais ofereceu à comunidade científica a possibilidade de uma análise exploratória mais abrangente. Análises genômicas permitem comparar regiões importantes dos cêlulas e assim identificar regiões gênicas. Essas análises comparativas, também descritas por genômica comparativa, muitas vezes podem destacar genes mutantes associados à ocorrência de diversas patologias. Na bioinformática esse tipo de estratégia pode ser realizada por softwares como o Massive progressive Mauve, um alinhador de genomas capaz de destacar rearranjos moleculares e configuração dos genes no genoma. Alguns estudos têm 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, esses exomas podem ser comparados entre genomas controle, e determinar regiões que apresentam mutações. Análises de sequenciamento de gene único também podem ser analisados por [redacted] técnicas, de alinhamento multiplo e determinar mutações específicas. Alguns exemplos de alinhadores são: Clustal, Muscle, Seaffe e MAFT. Ainda no campo genômico alguma analise estatísticas podem ser aplicadas para avaliar se perfis mutacionais podem estar se expandindo ou não em populações. Para tal abordagem, é necessário o teste de neutralidade feito em um conjunto de sequencias. 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 ao 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 cDNA pela atividade da transcriptase reversa. Após a ligação dos adaptadores e sequenciamento, as leituras podem ser alinhadas contra genomas de referência usando HTSeq, um alinhador e em seguida passam por etapas de quantificação dos transcritos, feita pela ferramenta RSEM. Essa 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. Essa ferramenta normaliza os dados de sequenciamento para possíveis problemas nas etapas experimentais. Por fim ele vê a variância a partir dos modelos lineares generalizados dos dados normalizados com os transcritos. A variância é feita por estatística bayesiana e a diferenciação de transcritos entre anotados 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. Entender como a expressão diferencial implica na promoção de doenças, aumentando o conhecimento o impacto da expressão gênica em patologias humanas.

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 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 ou ressonância magnética. Por métodos computacionais, é possível por modelagem comparativa, threading e de novo. Para modelagem a sequência a ser modelada precisa ter ao menos 30% de identidade/similaridade contra proteínas de referência, no caso o template. Ferramentas como o Modeller ou Swiss Model são capazes de gerar estruturas 3D utilizando campos de força, que são cálculos matemáticos capazes de avaliar números de interações inter moleculares, distâncias em Angstroms e forças. O Modeller usa o campo de força CHARMM e o Swiss model o GROMOS. Uma vez os modelos feitos eles podem ser validados por gráficos de Ramachandran, que avaliam a qualidade estereoquímica dos átomos e verifica se há choque eletrostáticos 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 das 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 existe 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 dos 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 DOPE, mas também pelos gráficos de RMSD que calculam a Raiz quadrada média dos desvios 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 do mal funcionamento da proteína. Se forem proteína "saudáveis" mas associadas a algum tipo de patologia, técnicas como o Docking molecular podem avaliar a intervenção de fármacos dessa proteína identificando ligantes que interage molecularmente com a proteína e inibe sua função. Softwares como o Dock6 ou MOE são capazes de avaliar o número de interações inter moleculares exercidas entre o complexo proteína-ligante e pela variação de energia livre apontar os ligantes mais favoráveis. Alguns softwares utilizam de campos de força para determinar essas interações, mas algumas metodologias mais recentes utilizam IA a partir de redes neurais para calcular essas interações e forças. Alguns procedimentos devem ser realizados na estrutura antes do Docking, como avaliar as cargas dos Amino ácidos, principalmente da região que interesse da proteína determinada por grids. Algumas ferramentas como o pdb4amber podem fazer essa validação ou pelo Chimera.

Etapas de dinâmica molecular também auxiliam o entendimento sobre interação e sinalização celular associados a casos de patologias em humanos. ferramentas como o GROMACS e Amber são capazes de calcular a posição inicial dos átomos e a 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 dessas 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 grandes bancos de redes de interação da célula de um alvo específico. O String se baseia em bases de dados com sequências anotadas por dados experimentais, como o UniProt. Porem, técnicas com data-mining também são implementadas para consultar diretamente na literatura sobre experimentos que descrevam aquela rede de interação.

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