Processamento das Sequências Obtidas com Equipamentos de Sequenciamento de Nova Geração (NGS)

Diego Frias e Mauricio Souza Menezes

[1]Departamento de Ciências Exatas e da Terra, Campus I
Universidade do Estado da Bahia (UNEB)
Salvador, Bahia, Brasil.

mauriciosm95@gmail.com

Resumo. *Este documento tem como objetivo descrever o contexto da pesquisa, com vistas a definir melhor o escopo do projeto. Possíveis questões de pesquisa são elaboradas ao longo do texto, para uma posterior seleção e generalização.*

Abstract. *This document aims to describe the context of the research, in order to better define the scope of the project. Possible research questions are elaborated throughout the text, for later selection and generalization.*

1. Questões de Pesquisa

Começamos listando as questões de pesquisa divididas em duas partes:

- Host Independent Case
    1. Quão diferentes são os cladogramas construídos com distâncias obtidas a partir da diferença do uso de códons (somente ID's), das árvores filogenéticas construídas a partir das mutações de nucleotídeos, levando em conta apenas a topologia, ou seja, os agrupamentos (clusters) obtidos?
    2. Sendo diferentes, como interpretar as diferenças?
    3. Sendo similares, o método Codon Based Unsupervised Classification (CBUC) poderia ser uma alternativa para genotipagem rápida de vírus? Quão rápida?
- Host Dependent Case
    1. Quão diferentes são os cladogramas construídos com distâncias dependentes e independentes do hospedeiro?
    2. Como evoluiu o fitness viral de tradução ao longo da pandemia de COVID-19? As cepas emergentes tinham mais ou menos fitness?

2. Hipóteses/Considerações de Partida

1. Em sequências codificantes, todo nucleotídeo pertence a um códon (podendo estar na primeira, segunda ou terceira posição dele).
2. Toda mutação de nucleotídeo numa sequência codificante muda o códon ao qual o nucleotídeo pertence.
3. Algumas mutações de códons mudam o aminoácido (as não sinônimas), já as outras (as sinônimas) não.

SOBRE MUTAÇÕES DE AMINOÁCIDOS (ENFOQUE ATUAL)

4. Algumas mudanças de aminoácidos (mutações não sinônimas) produzem uma mudança significativa da estrutura 3D e/ou interatividade da proteína codificada.
5. Algumas mudanças da estrutura 3d e/ou da interatividade da proteína codificada podem ter influência favorável/desfavorável para a espécie. No caso de vírus, por exemplo, pode favorecer o escape à resposta imune do hospedeiro.

   SOBRE O IMPACTO DO USO DE CÓDONS (BASE DO NOVO ENFOQUE)

6. Toda mudança de códon (sinônima ou não sinônima) altera (em maior ou menor grau) a logística do processo de tradução de proteína codificada. Mesmo que com tolerância à variabilidade natural de tecido a tecido, na célula existe um equilíbrio entre a abundância de tRNAs específicos para cada códon no pool de tRNA e a frequência com que o códon aparece nos mRNAs sendo traduzidos. Como regra, quando o códon é muito frequente, há mais abundância dos seus tRNA cognatos e vice-versa, quando o códon não é muito frequente, seus tRNA cognatos são menos abundantes (economia de recursos). **É necessário citar trabalhos sobre isso (ver o artigo sobre HIV)**
7. A taxa de produção de proteínas é influenciada positiva ou negativamente pela alteração do balanço entre a frequência dos códons no poll de mRNA e a abundância de tRNAs. Os tRNAs não tem a mesma proporção, tem espécies de tRNA mais frequentes que outras. Por isso a velocidade de elongação da cadeia polipeptídica que compõe uma proteína não é uniforme: os códons com mais tRNA são traduzidos mais rápidos que os códons com menos tRNA. Em outras palavras, quando um mRNA possui um códon lento ocorre uma pausa toda vez que esse códon chega ao sítio de tradução (por ficar esperando a chegada do tRNA escasso), aumentando o tempo médio de tradução da proteína codificada por ele.
8. Se códons rápidos são substituídos por códons lentos, a demora na síntese dessa proteína pode ter um impacto negativo no funcionamento celular, podendo até causar a morte celular. No caso contrário, a substituição de códons lentos por rápidos aumenta a velocidade média de síntese de proteínas, o que pode ser benéfico para um vírus que precisa se replicar dentro de uma célula hospedeira.
9. O balanço perfeito, no qual há uma correspondência total entre a frequência de codons no pool de mRNA e a abundância relativa das distintas espécies de tRNA no pool de tRNAs, pode ser considerado como o ótimo para maximizar a taxa de produção (síntese) de proteínas na célula.
10. Mudanças no uso de códons que se aproximem do balanço perfeito podem ser consideradas favoráveis enquanto as que se afastam do balanço perfeito como desfavoráveis.
11. A teoria prevalente é que a evolução do uso de códons e da abundância de tRNAs (medida em função do número de cópias dos genes que codificam as distintas espécies de tRNA no genoma) tem conduzido a otimizar o balanço (co-evolução otimizante). **É necessário citar trabalhos sobre isso (ver o artigo sobre HIV)**
12. Quando um vírus infecta uma célula, suas proteínas vão ser produzidas num ambiente que foi otimizado (balanceado) para a célula hospedeira. Neste contexto, na medida que o uso de códons das proteínas do vírus for mais semelhante ao da célula hospedeira, espera-se que suas proteínas sejam sinterizadas com maior ve-

locidade. Da mesma forma, quanto mais diferente seja o uso de códons do vírus, suas proteínas serão sintetizadas com menor velocidade.

13. Quanto maior a taxa de síntese das proteínas virais, maior sera: **Buscar referências- Discutir com Vagner**
    (a) A taxa de produção de virions -> carga viral -> virulência.
    (b) O dano à célula hospedeira aumentando a taxa de apoptosis -> morbidade/mortalidade do vírus.
14. É possível medir a capacidade relativa de replicação viral numa determinada célula hospedeira de forma indireta, comparando o uso de códons dos genes virais com o do hospedeiro e de forma direta comparando o uso de códon viral com a abundância de tRNA no hospedeiro.

LINK COM CBUC

15. O CBUC gera uma lista de códons mutantes ordenada de acordo com a variabilidade do códon, observada num conjunto de treinamento. A cada sequência no dataset de entrada atribui-se uma lista com os códons específicos que é chamada de id.
16. Como é possível calcular a distância (similaridade) entre dois IDs, é possível construir uma matriz de distâncias e usá-la para construir um cladograma (dendrograma e árvore filogenética). Contudo, nesta abordagem o agrupamento é independente do hospedeiro.
17. Contudo, no caso de vírus é possível derivar distâncias que levem em conta as características do hospedeiro. Pode se usar a via direta estabelecendo a correlação com a frequência de genes de tRNA, ou de forma indireta estabelecendo a correlação com o uso de códons do hospedeiro. Em princípio, cada códon tem uma frequência conhecida no hospedeiro e uma proporção de genes de tRNA cognato. Se no vírus ocorre a troca de um códon A por um códon B, qué efeito isso causaria no fitness de tradução? Isso depende de que? Nós podemos calcular o grau de fitness padrão do hospedeiro pela correlação entre uso de códons e abundância de tRNA. Também podemos calcular a variação do fitness para cada mudança de pár de códons (metade de uma matriz de 61X61). Esta matriz seria simétrica?
18. **Se for simétrica: Vou responder matematicamente antes.**
19. Como fazer isto? A distância entre 2 sequências A e B é definida como um vetor de tamanho IDsize contendo a mudança de fitness associada à mudança de cada códon na seq. A para o códon da seq, B. Alguma norma (L1 ou L2) desse vetor pode servir como distância. Há então que estudar o efeito da escolha da norma.
20. É interessante então, avaliar o efeito do hospedeiro na classificação de sequências virais. Comparar cladogramas com distâncias entre IDs, e com distâncias derivadas do fitness de tradução. Explicar/discutir as diferenças. Com qual das 2 distâncias obtém-se cladogramas mais "diferentes" das árvores filogenéticas.
21. No caso de distâncias baseadas no fitness de tradução, os agrupamentos nos cladogramas ocorrem de acordo com a taxa potencial **(vide ponto 13)** de replicação viral. Quão diferentes são essa taxas?. As variantes emergentes tem maior ou menor taxa potencial de replicação que as antecessoras?

NEW CBUC ALGORITHM

1. Com dataset codificante alinhado (colunas=nucleotideos, linhas=strain), traduzir todas as sequencias a códons (colunas = códons). Tendo N nucleotídeos, teremos C=N/3 códons.
2. Varrer as colunas contando o número de codons diferentes: V(c), c=1,2,..,,C nas sequências.
3. Organizar as colunas de maior a menor V, registrando a posição original "c" de cada uma delas.
4. Eliminar as colunas com V==1
5. Contar as colunas restantes = tamanho do ID (IDsize).
6. Obter Vmax (maior variabilidade observada no dataset) e número máximo de classes primárias possíveis (nPclass = produto de todas as variabilidades).
7. Se Vmax>2 então é possível reduzir o tamanho do ID, introduzindo um limiar 1<Vlim<Vmax. Dado um Vlim, eliminar todas as colunas com V<Vlim e recalcular o tamanho do ID (IDsize).
8. Trasformar os vetores linha em strings gerando IDs.
9. Descoberta de classes primárias: Gerar lista de IDs distintos (unique/set). Imprimir número de classes primárias: P
10. Anotação primária: Atribuir uma classe primária p=1,2,...,P a cada sequência do dataset, a partir do seu ID.
11. Fazer histograma das classes.

3. Plano de Atividades

Aqui VOCÊ precisa elencar TODAS as "atividades", passo a passo em cada questão de pesquisa, na ordem, para poder fazer o planejamento do projeto.
Construa uma árvore para melhor entendimento dos ramos da pesquisa
No final escolheremos um 'path' executável no seu TCC
Note que:
Cada pergunta gera ≥ 1 atividades.
Também onde disse 'precisa(mos)' há atividades implícitas a serem realizadas.