Tutorial – Introdução a anotação e comparação de genomas

Tiago Mendes – Doutorando em Bionformática

Hoje iremos trabalhar com dois programas free desenvolvidos pelo Sanger institute: Artemis e ACT. Artemis foi desenvolvido com a finalidade de permitir a visualização gráfica de um genoma, bem como uma ferramenta de anotação que permite visualização, edição e análise de sequências gênicas em suas seis janelas de leitura. ACT é um programa que permite comparação de genomas.

Ao longo do tutorial existem 12 exercícios (digitadas em vermelho) que deverão ser enviados individualmente para o email: mendesfarmacia@gmail.com até domingo (14/11/2015) as 23:59 .

Como os dois programas são escritos em Java, para instalação do Artemis e do ACT é necessário a instalação previa do Java (já realizada nas máquinas do laboratório) com instruções que podem ser obtidas em: http://java.com/en/download/manual.jsp

Exercício 1: Quais as vantagens da utilização dos métodos de sequenciamento de nova geração sobre o método de sequenciamento de sanger

Exercício 2: Explique as diferenças entre a montagem de um genoma pelo método de novo e pelo método de referência.

Obtenção dos arquivos:

1- Abra o terminal (Aplicativos → Acessórios → Terminal)
2- Entre no diretório Documentos (cd Documentos)
3- Crie a pasta artemis (mkdir artemis)
4- Entre na página: http://pinguim.fmrp.usp.br/anotacao/
5- Faça Download dos arquivos e salve na pasta artemis criada:
    a. organismo1.fasta

b. organismo2.embl
c. organismo3.embl
d. org1.vs.org2.formatado
e. org1.vs.org3.formatado

6- Entre na pasta artemis (cd artemis)
7- Digite 'ls –l' e confira se os cinco arquivos do item 5 estão presente no diretório artemis.

Instalação Artemis e ACT em Linux:

1- Entre no site do Artemis (http://www.sanger.ac.uk/resources/software/artemis/) e clique na aba Download
2- Em FTP Download clique Artemis for Unix
3- Salve o arquivo artemis.tar.gz no diretório Documentos
4- Abra o terminal e entre no diretório Documentos (cd Documentos)
5- Extrai o arquivo (tar -vzxf artemis_linux.tar.gz)

Instalação do Artemis e ACT em Windows:

1- Entre no site do Artemis (http://www.sanger.ac.uk/resources/software/artemis/) e clique na aba Download
2- Em FTP Download clique Artemis for Windows
3- Salve o arquivo artemis.jar no Desktop
4- Clique duas vezes no arquivo artemis.jar
5- Clique em Browser e selecione o diretório Documentos

Anotação manual de genes codificadores de proteínas

1. Para executar o programa Artemis digite: ./art e aperte enter
2. Clique em Browse e selecione o diretório artemis
3. Clique em 'Options' e selecionne '1 – Standard'

4. Clique em File → Open... → Abra o arquivo 'organismo1.fasta'

5. Uma vez aberto o arquivo, procure uma ORF (Open Read Frame) e com o mouse selecione um trecho dessa ORF. Obs: Não selecione nenhum tracinho referente a um stop códon, selecione de preferência o meio da ORF;

Exercício 3: Quais as características de uma CDS (sequência codificante para uma proteína completa)?

Exercício 4:

6. Clique em Create → Feature From Base Range. Na janela que apareceu clique em 'Apply' e OK;

7. Clique no retângulo azul criado e clique em 'Edit' → Extend Selected Features → To Previous Stop Códon. A sequência deve crescer para a esquerda (até um stop códon);

8. Clique novamente no mesmo retângulo azul e vá em 'Edit' → 'Extend selected feature' → 'To next stop códon and fix'/.Pronto, agora seu gene termina em um stop códon, mas será que o começo está correto?

9. Clique novamente no mesmo retângulo azul e clique em Edit → Trim Selected Features → To Next Met). A sequência vai ser corrigida para começar em uma metionina. Agora sim! Seu gene tem metionina inicial e stop códon, 'pode' codificar algo...

10. Feito isso, procure outras 3 ORFs disponível e repita do passo 4 ao 9;

11. Uma vez criado genes, clique em 'Select' → 'All CDS features'

12. Clique em 'Edit' → 'Automatically create genes name'

    a. 'Enter the start character...': coloque o nome Gene_

b. ‘start count at’, coloque: 1

c. ’increment number by’, coloque: 1

d. ‘enter a qualifier name to use’, coloque: locus_tag

e. ‘number of digits...’, coloque:2

f. ‘append “c”...., clique em No. Pronto, os genes estão nomeados!!!

Exercício 5: Para identificarmos qual possível proteína cada CDS codifica, será utilizado BLAST (alinhamento local contra um banco de dados). Qual programa seria mais indicado para esta identificação: Blastn (utilizando a sequência gênica) ou Blastp (utilizando a sequência da proteína predita)? Justifique.

13. Clique no primeiro gene, vá em ‘View’ → ‘Aminoacid of selection as fasta’ → copie a sequência que aparecer em formato .fasta e faça uma busca por similaridade no BLASTp no site do NCBI;

14. Anote o resultado, para isso, selecione o gene de onde veio a sequência e aperte a tecla E (Edit/Selected Features in Editor). Vai aparecer uma janela, nessa janela que serão anotadas todas as informações sobre o gene;

15. Na janela aberta, clique na setinha preta do campo ‘Add Qualifier’ e adicione os campos product; curation e similarity. Em cada escolha, clique em Add Qualifier e confira se o campo foi criado dentro da janela;

16. Com o resultado do BLASTp feito , preencha os campos acima adicionando;

    a. Curation: o nome do anotador

    b. Product: o nome do produto codificado pelo gene, de acordo com o BLASTp;

c. Similarity: campo mais importante. Preencher da seguinte maneira – Similar to (organismo que deu maior similaridade); nome do produto da maior similaridade; tamanho dessa proteína em aminoácidos; e-value: valor de e encontrada nessa maior similaridade; % de similaridade entre as duas proteínas (% id) in (número de aminoácidos encontrados na similaridade) (exemplo no slide!!!);

d. Clique em 'Apply' e 'Ok'

17. Feito isso, vá nos outros genes e repita do passo 13 ao 19;

18. Salve o arquivo com o nome de organismo1.embl no diretório artemis clicando em File → Save An Entry As → EMBL Format → organismo1.fasta

19. Pronto, curadoria realizada com sucesso!!! Agora vocês já podem anotar um genoma de verdade, com todos seus elementos! ☺

20. Feche o programa, abra o arquivo .embl gerado (more organismo1.embl) e analise sua estrutura.

Exercício 6: Qual os produtos gênicos prováveis para os seguintes genes:

a) Gene_01:
b) Gene_02:
c) Gene_03:
d) Gene_04:

Observação do cromossomo 1 de Trypanosoma brucei anotado

1- Clique no link: http://www.ncbi.nlm.nih.gov/mapview/
2- Clique em Prozoan → Trypanosoma brucei (Build 1.1) → clique no cromossomo 1
3- Clique em Download/View Sequence/Evidence
4- Clique em Save to Disk

5- Clique em Send → File → Mude o formato para GenBank (full) → create file → download
6- Mova o aquivo para o diretorio artemis (mv ../../Download/sequence.gb .) – o ponto final é muito importante!!
7- Abra o artemis (./art)
8- Clique em 'Options' e selecionne '4 – Mold, Protozoan,...'
9- Clique em File → Open... → Abra o arquivo 'sequence.gb'
10- Se aparecer a mensagem – there are warnings while reading – view now? à Clicar em 'No'

Exercício 7: Copie e descreve a característica de três estruturas anotadas neste cromossomo.

Comparação de três genomas com ACT

Para rodar o ACT precisa-se de pelo menos 3 arquivos. Dois arquivos com as seqüências a serem comparadas que podem ser anotadas (formato anotado do Genbank, por exemplo) ou não (fasta) e um arquivo de comparação entre as duas seqüências. Nós já temos os arquivos das seqüências (organismo1.embl, organismo2.embl e organismo3.embl) e para gerar os arquivos comparativo foi utilizado o programa tblastx (org1.vs.org2.formatado e org1.vs.org3.formatado) com o seguinte formato:

Exercício 8: Como tblastx faz alinhamento entre duas sequências?

O arquivo comparativo deve possuir os seguintes campos:

Col.12: score
Col.3: % identidade
Col.7: Query start
Col.8: Query end
Col.1: Query sequence name
Col.9: Subject start

Col.10: Subject end

Col.2: Subject sequence name

1. Abrir o programa ACT digitando ./act e apertando enter
2. Clicar em file → Open e complentar os campos com as seguintes informações:

    Sequence file 1: organismo2.embl

    Comparison file 1: org1.vs.org2.formatado

    Sequence file 2: organismo1.embl

3. Clicar em "More files..."

    Comparison file 2: org1.vs.org3.formatado

    Sequence file 3: organismo3.embl

4. Clicar em 'Apply'
5. Comparar as CDS e regiões não codificadoras nos três genomas (linhas vermelhas match sem inversão, linhas azuis match com inversão).

Exercício 9: Conceitue sintenia gênica e como é esperado a sintenia entre dois organismos filogeneticamente próximos e mais afastados?

Exercício 10: Comparando o genoma do Organismo2 e o Organismo3, existe diferença entre o quantidade de genes? Quais genes estão ausentes ou duplicados entre estes organismos? Há alguma inversão ou alteração na sintenia gênica?

Exercício 11: Entrar no link: http://biodados.icb.ufmg.br/pinguim/anotacao/ e identificar seu número (em frente ao seu nome) no arquivo Lista_de_alunos.pdf e baixar o arquivo sequencia_X.txt no link: http://biodados.icb.ufmg.br/pinguim/anotacao/Sequencias_exercicios/ , onde X é o seu número na lista de alunos. Anotar o gene presente nesta sequência utilizando o programa Artemis.

Exercício 12: Escreva um parágrafo sobre a função da sua proteína identificada no exercício 11 citando duas referências bibliográficas.