IQB-723 – BIOINFORMÁTICA BÁSICA

Professor
Rafael Dias Mesquita (IQ-UFRJ) – rdmesquita@iq.ufrj.br

Carga horária: 30 horas
Disciplina teórico-prática

Créditos: 2

Vagas: 10

Objetivo

A disciplina tem por objetivo principal ensinar conceitos básicos de Bioinformática e treinar o aluno no uso de programas diversos disponíveis como serviços web sem necessidade de uso de Linux ou linha de comando.

Ementa

Os principais tópicos do curso são: (*i*) Introdução à Bioinformática; (*ii*) Ferramentas de software; (*iii*) Formatos de arquivos relacionados à Bioinformática; (*iv*) Bancos de dados na rede; (*v*) Buscas por similaridade; (*vi*) Domínios conservados; (*vii*) Alinhamento de sequências; (*viii*) Ortologia e derivados; (*ix*) Filogenia e árvores filogenéticas; (*x*) Modelagem estrutural.

Programa Analítico

1. Introdução à Bioinformática: conceitos básicos de um computador (*hardware*, sistema operacional e programas) e de Biologia Molecular (estrutura do DNA, tradução, *frames* de leitura, estrutura de proteínas e domínios conservados)
2. Ferramentas de software: Firefox – instalação e uso de complementos, editores de texto (Textpad) - conversão de codificação e de formato final de linha, quebra automática de linha, substituição de texto, seleção de colunas.
3. Formatos relacionados à Bioinformática (FASTA e Genbank): visualização e edição de sequências.
   PRÁTICA: Conversão de formato unix->win e win->unix de arquivos de texto.
   PRÁTICA: Busca e edição para remover o texto “rafael” e trocar o texto “mesquita” pela sequência original presente na proteína em formato fasta.
4. Bancos de dados no NCBI: PUBMED, livros, proteínas (nr, swiss-prot, refseq ...), domínios conservados (CDD), nucleotídeos (nt, mRNA, ORF, genes...), taxonômico etc. Acesso, busca, visualização e *download*.
5. Buscas por similaridade: matrizes de score à família Blast de programas, matrizes de score por posição à psi-blast.
   PRÁTICA: Buscar um gene ou mRNA, ORF ou uma proteína por uma palavra-chave. Baixar a sequência gênica, seu cDNA e a proteína codificada em formato fasta. Buscar 5

sequências similares de organismos próximos por blast, restringindo a busca por um taxid. Busca de ESTs que suportem o mRNA escolhido. *Download* das sequências similares identificadas em formato fasta (gene, ORF e proteína) e dos ESTs caso existam.

6. Domínios conservados, bancos de dados e arquitetura de domínios: CDD, PFAM, BLOCKS, SMART etc. Acesso, busca, visualização e *download*.
   PRÁTICA: Usar as sequências da aula anterior para buscar domínios conservados usando pelo menos dois bancos de dados e confrontar os resultados. Buscar outras sequências com a mesma arquitetura de domínios conservados. Verificar a anotação funcional e se os similares identificados estão dentre eles.
7. Alinhamento de sequências (clustalw, praline, muscle, t-coffee): Local x global, simples x múltiplo, parâmetros e matrizes de comparação. Avaliação de qualidade em alinhamentos.
   PRÁTICA: Usar as sequências das aulas anteriores para fazer um alinhamento simples do seu mRNA, ORF e do melhor similar. Alinhar todos os 6 mRNAs, ORFs (verificar qualidade) e identificar regiões que seriam boas para desenhar *primers* degenerados e espécie específico. Alinhar as 6 proteínas (verificar qualidade) e verificar se as regiões escolhidas para desenho de *primers* são conservadas a nível proteico e se as regiões de domínio conservado identificados anteriormente estão bem alinhadas.
8. Ortologia, paralogia, homologia e similaridade: bancos de dados de *clusters* de ortólogos – COG, KOG, KEGG, PRO, GO, Homologene etc. Identificação de ortólogos e de vias metabólicas no KEGG.
   PRÁTRICA: Verificar se sua proteína pertence a um *cluster* de ortólogos em algum banco de ortólogos e se as sequências similares que você selecionou pertencem ao mesmo *cluster*. Verificar se sua proteína pertence a alguma via metabólica, gerar um mapa colorido para a espécie de interesse.
9. Filogenia e dendogramas: bases da filogenia, métodos de matrizes de distância, máxima parcimônia e probabilidade, inferência bayesiana e seleção de modelo. Construção de árvores com e sem raiz.
   PRÁTICA: Inserir seu alinhamento múltiplo de proteínas e construir uma árvore filogenética usando um dos métodos explicados.
10. Modelagem por homologia: identificação e escolha de estrutura tridimensional de proteína, alinhamento e modelagem.
    PRÁTICA: Verificar se existe estrutura tridimensional para sua proteína e/ou algum similar que você já identificou. Caso contrário, encontre a sequência mais similar possível à sua com estrutura resolvida. Faça um alinhamento da sequência da estrutura com suas sequências e envie para modelagem pelo menos sequências homólogas de dois organismos diferentes.
11. Alinhamento tridimensional e avaliação da qualidade do modelo: RMSD, avaliação de voltas e de estrutura secundaria, estabilização dos modelos.
    PRÁTICA: Visualizar os modelos criados e alinhar a estrutura e os modelos gerados. Calcular o RMSD, avaliar a qualidade dos modelos gerados.

Literatura recomendada

- Verli, H. (2014). Bioinformática: da Biologia à Flexibilidade Moleculares. 1a edição, Porto Alegre, Brasil. Este livro está disponível para download gratuitamente: https://www.ufrgs.br/bioinfo/ebook/
- Lesk, A.M. (2012). Introduction to Genomics. Oxford University Press, 2nd edition, USA.
- Lesk, A.M. (2014). Introduction to Bioinformatics. Oxford University Press, 4th edition, USA.
- Bioinformatics – A practical guide to the analysis of genes and proteins. Baxevanis, AD and Ouellette, BFF. Wiley.3rd edition. 2005.
- Just the Facts: A Basic Introduction to the Science Underlying NCBI Resources. http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html
- The NCBI handbook: http://www.ncbi.nlm.nih.gov/books/NBK21101/