Prática 9 - Análise de Dados de *Microarray*

Ricardo de Marco Alessandro S. Nascimento

16 de Novembro de 2016

1 Introdução

Esta atividade tem como objetivo familiarizar o aluno com técnicas de analise de dados de *microarray*. Para isso, iremos utilizar o programa **GenePilot** que possui vários recursos. Note que diferentes estratégias podem ser utilizadas para análises deste tipo de dados e que cada pacote de programas podem ter diferentes recursos. Neste caso, iremos utilizar dados provenientes de experimentos de *microarray* utilizando diferentes tecidos tumorais.

2 Etapa 1 - Carregamento dos Dados

Nesta etapa, iremos realizar o carregamento dos dados de *microarray* no programa **GenePilot**. O arquivo de entrada tem formato ASCII (txt) e contêm uma tabela com os valores de leitura de intensidade resultante dos diferentes experimentos.

1. Faça o download do arquivo NCI_T_matrix_1375.txt.
2. Abra este arquivo em um editor de planilhas e examine seu conteúdo. O que significam as diferentes colunas e linhas deste arquivo? Obs: Antes de abrir o programa GenePilot, é preciso alterar a data do seu computador para 01/03/2008. Após abrir o programa, retorne a data para a atual, para evitar problemas com a navegação na Web.
3. Abra o programa **gene pilot**.
4. Vá em "file $->$ new datase $->$ from template";
5. Selecione o arquivo "*NCI_T_matrix_1375.txt*". Isso permite com que os dados existentes neste arquivo sejam carregados no programa.

3 Pré-processamento dos Dados

Esta etapa tem como objetivo realizar uma filtragem de conjuntos de dados somente com sinais de baixa intensidade e, portanto, pouca significância. Também será realizada uma seleção de dados com alta variação, sendo descartados dados monótonos e pouco informativos. Além disso, é realizada uma normalização dos dados provenientes das diversas amostras bem como o preenchimento de valores ausentes.

1. Vá ao menu "*action* $->$ *pré-processing*"

2. Escolha na opção "*percentage present*" coloque o valor 95. Isto significa que somente genes que contenham 95% dos dados existentes serão computados (dados podem estar faltando devido a problemas experimentais que impeçam a leitura adequada de intensidade);

3. Escolha a opção "*standard deviation*" e selecione o valor de 1. Isto fará que somente as linhas (genes) que possam grande variação nos seus níveis de expressão entre amostras sejam selecionados.

4. Escolha a opção " *at least* __ *Observations...*" colocando um *"tick"* ao lado desta opção. Isso fará que somente genes que contenham pelo menos um caso de valor de intensidade igual ou maior que 2 serão computados (filtragem de dados de baixa intensidade)

5. Aperte o botão "*calculate remaining*" para ver quantos genes foram filtrados por este procedimento. **Que porcentagem dos genes foram filtrados? Que características você espera para os genes remanescentes? (Q1)**

6. Aperte o botão "do pre-processing";

4 Agrupamento Hierárquico

O agrupamento hierárquico realiza agrupamentos sucessivos de amostras e genes de acordo com a distância medidas. Isto irá gerar um dendrograma unindo os genes e as amostras de modo que é possível deduzir diversos grupos com perfil semelhante de variação.

1. Vá ao menu "action $->$ hierarchical";

2. Aperte o botão "run";

3. Abrirá uma janela contendo dois espaços. O mais à esquerda mostra um perfil geral dos dendrogramas gerados e do "heatmap" representando os valores de expressão gênica. A área à direita mostra detalhes do heatmap, que variam de acordo com a sua seleção. Tente clicar em porções do dendrograma e veja que esta janela mostra um zoom da área selecionada com os nomes dos gene. Clique duas vezes em um dos nomes dos genes e você verá detalhes como a classificação de GO atribuída a este gene.

4. Avalie o dendrograma gerado para o agrupamento das diferentes amostras. **O perfil de expressão gênica é capaz de separar os diferentes tipos de câncer? Existem conjuntos de genes que possuem padrão de expressão especifica (aumentado ou diminuído) dependente do tipo de tumor? (Q2)**

5. Feche a janela do *hierarchical clustering* e volte para a janela original.

5 Agrupamento por k-médias

O agrupamento por k-médias realiza o agrupamento de dados baseado em um numero arbitrário de grupos definido pelo usuário.

1. Vá no menu "$action -> K\text{-}means$";
2. Na nova janela aberta clique em *run*.
3. Uma nova janela se abrirá contendo botões no canto direito superior que indicam todos os agrupamentos (*"clusters"*) formados. Ao clicar em cada um destes botões o resto da tela exibirá informações sobre este.
4. Examine os grupos formados. **Eles refletem alguma divisão entre genes predominantemente expressos em alguns tipos de tumores? (Q3)**
5. Analisando os seus dados, você acha que deveria haver mais ou menos grupos para uma boa divisão? Feche a janela da sua analise de k-mean atual e refaça a analise de k-mean escolhendo um número diferente de grupos. A**nalise o resultado e veja se o agrupamento de fato se tornou melhor. (Q4)**