SCC5836 - Visualização Computacional
Profa. Dra. Rosane Minghim

Visualization-based cancer microarray data classification analysis

Minca Mramor , Gregor Leban , Janez Demšar e Blaž Zupan

Tamires Brito da Silva
8124382

Motivação

Motivação

- Métodos para analisar dados de *microarray* de câncer frequentemente enfrentam dois desafios distintos
  - Classificar novas amostras de tecido
  - Fornecer uma visão dos padrões e interações de genes escondidos nos dados

Motivação

- A visualização de dados pode fornecer uma excelente abordagem para a descoberta de conhecimento e análise de dados com rótulo de classe
- **VizRank**
  - Pontua e classifica visualizações baseadas em pontos de acordo com o grau de separação de instâncias de dados de diferentes classes

Proposta do Artigo

- Extensão do VizRank
  - Técnicas para descobrir *outliers*
  - Ranquear *features* (características) (genes)
  - Executar classificação
- Demonstrar que a abordagem proposta é bem adequada para a análise de *microarrays* de câncer

Contextualização

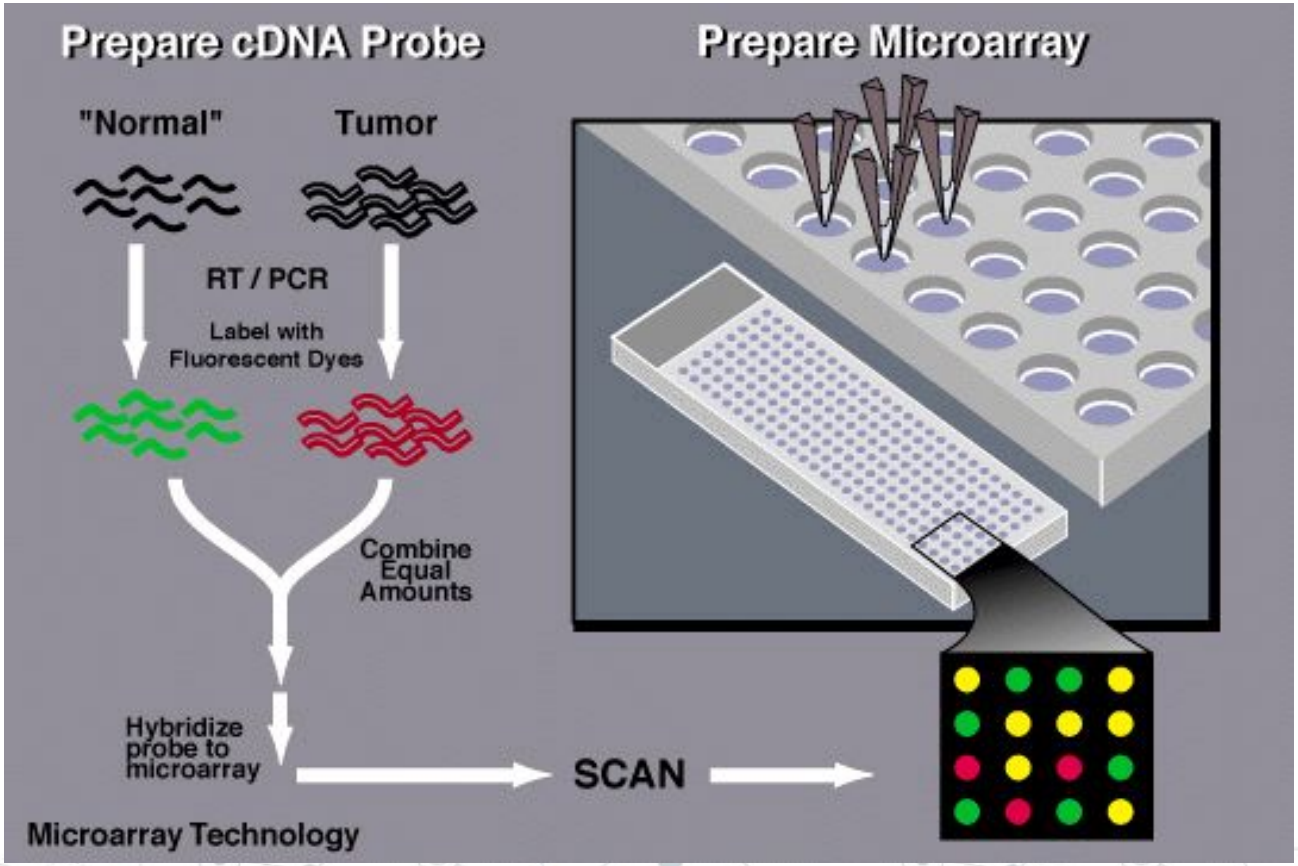
Contextualização

- Distinguir e classificar malignidades humanas dependem de uma variedade de aspectos clínicos, moleculares e parâmetros morfológicos
- O diagnóstico preciso do câncer continua a ser uma tarefa desafiadora
- *Microarrays* de DNA

Contextualização

Microarray de DNA

- Determinar se o DNA contém mutação de genes



FONTE: https://www.genome.gov/10000533/dna-microarray-technology/

Contextualização

- Muitas pesquisas foram feitas para diagnóstico de câncer (Golub et al., 1999; Bhattacharjee et al., 2001; Khan et al., 2001; Shipp et al., 2002)
- Vários métodos estatísticos e de mineração de dados para *microarray* análise de dados evoluíram (Allison et al., 2006; Asyali et al., 2006; Pham et al., 2006)

Contextualização

- Inicialmente eram usado métodos sem supervisão
  - Cluster e análise de componentes principais
- Modelos de dados de câncer são problemas de mineração de dados supervisionado
  - Máquinas de vetores de suporte *(support vector machines SVMs)* (Statnikov et al., 2005)
  - Redes neurais artificiais *(artificial neural networks ANN)* (Khan et al., 2001)
  - *k-nearest neighbors* (k-NN)(Golub et al., 1999)

Contextualização

- Os dois aspectos mais importantes de mineração de dados preditivos
  - Precisão nas previsões
  - Ganho de *insight* (visão)
- Nem todos os métodos cobrem esses dois aspectos igualmente bem
- Visualização de dados oferece meios para expor graficamente padrões interessantes

Visualização

Visualização

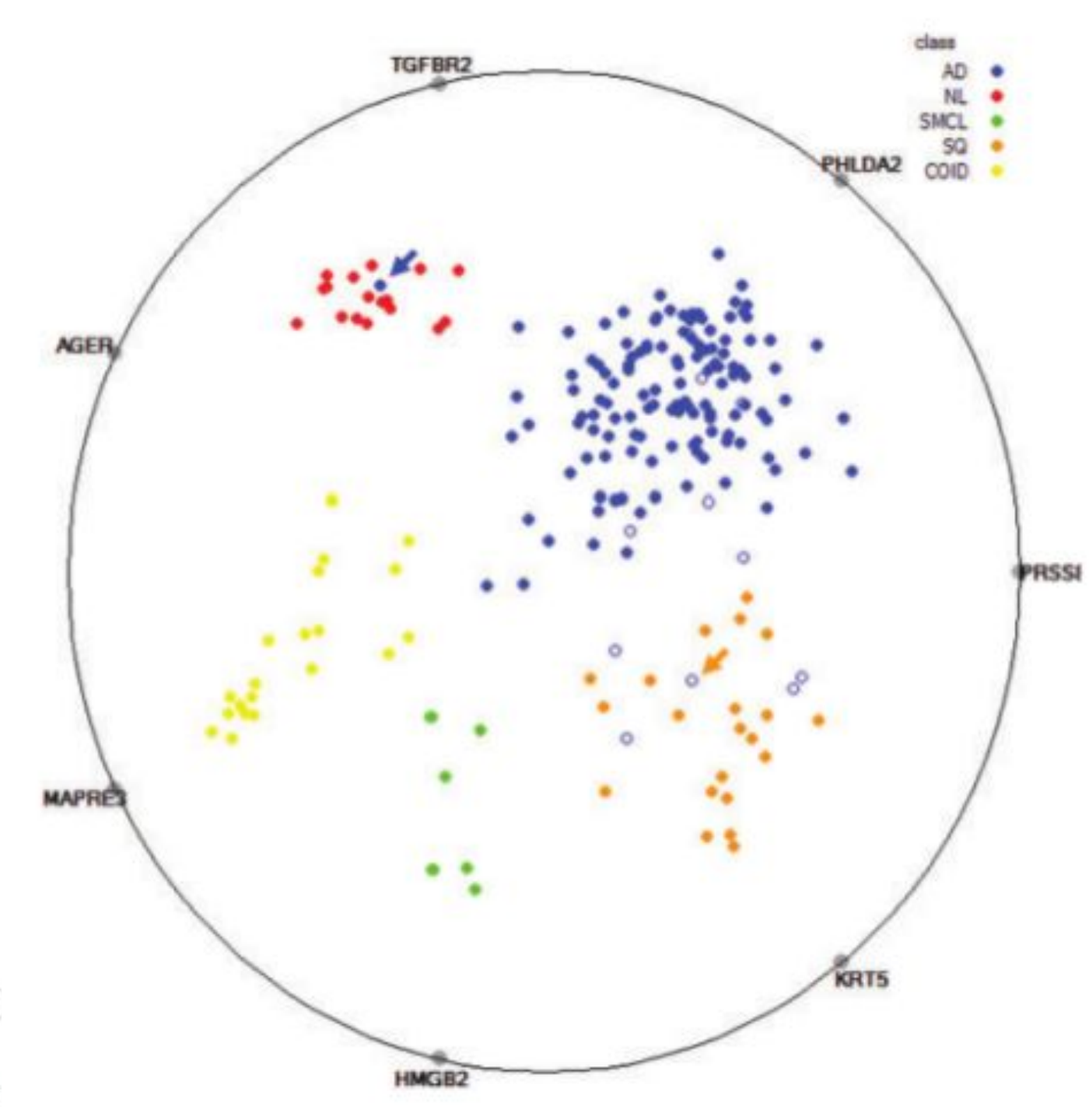
- Radviz (Hoffman et al., 1997)
  - Método de visualização não linear
  - *Features* são visualizadas como ponto-âncora igualmente espaçados ao redor do perímetro de um círculo unitário
  - Os valores das *features* são normalizados entre 0 e 1
  - As instâncias de dados são mostradas como pontos dentro do círculo

Visualização

- Radviz (Hoffman et al., 1997)
  - Abstração física
    - Cada ponto é mantido no lugar com molas presas na outra extremidade dos ponto-âncoras da *feature*
    - A rigidez de cada mola é proporcional ao valor da *feature* correspondente
    - Ponto termina na posição em que as forças da mola estão em equilíbrio
  - As instâncias de dados que estão mais próximas de um conjunto de *features* tem maior valor para essas *features* do que as outras

Visualização - Exemplo 1

- Radviz (Hoffman et al., 1997)
- *Dataset* de câncer de pulmão (Bhattacharjee et al., 2001)
  - 5 classes
  - 203 amostras de tecido (pontos)
- Posição de cada ponto depende do valor da sua *feature* (expressões genéticas)
- 7 genes é o suficiente para separar esse *dataset*

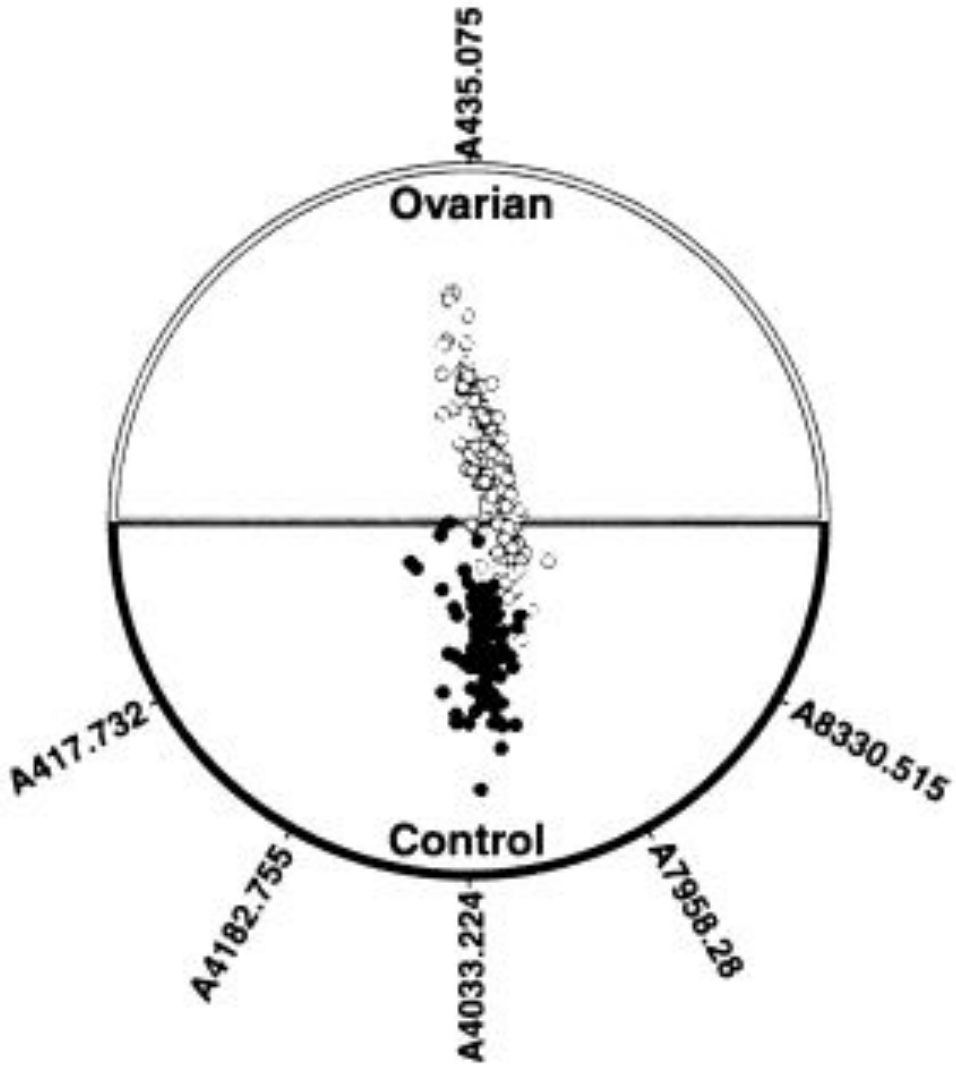


- adenocarcinoma (AD)
- pulmão normal (NL)
- câncer de pulmão de pequenas células (SMCL)
- carcinoma de células escamosas (SQ)
- carcinóide pulmonar (COID)

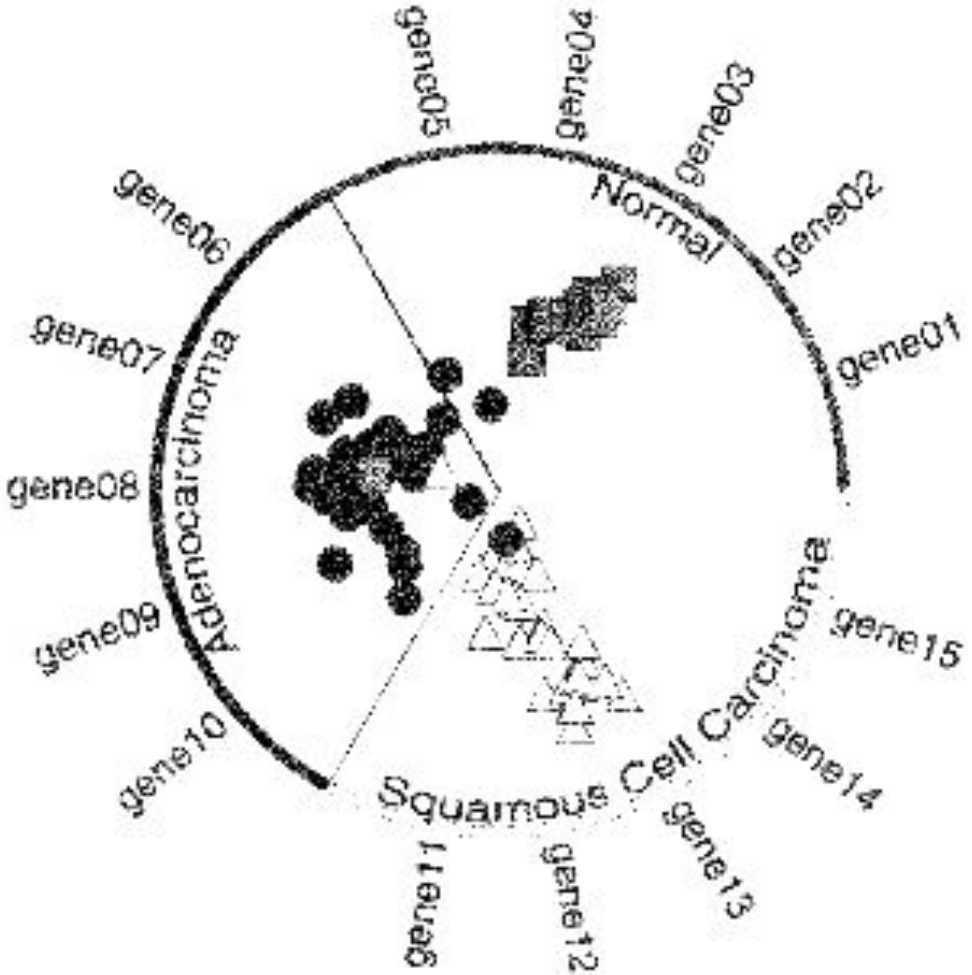
Visualização - Exemplo 2

- (McCarthy et al., 2004)
  - Seleção de características através da aprendizagem de redes neurais para reduzir o número de genes na visualização
  - *Cluster* de *features* (*cluster* de âncoras de *features* correlacionados)

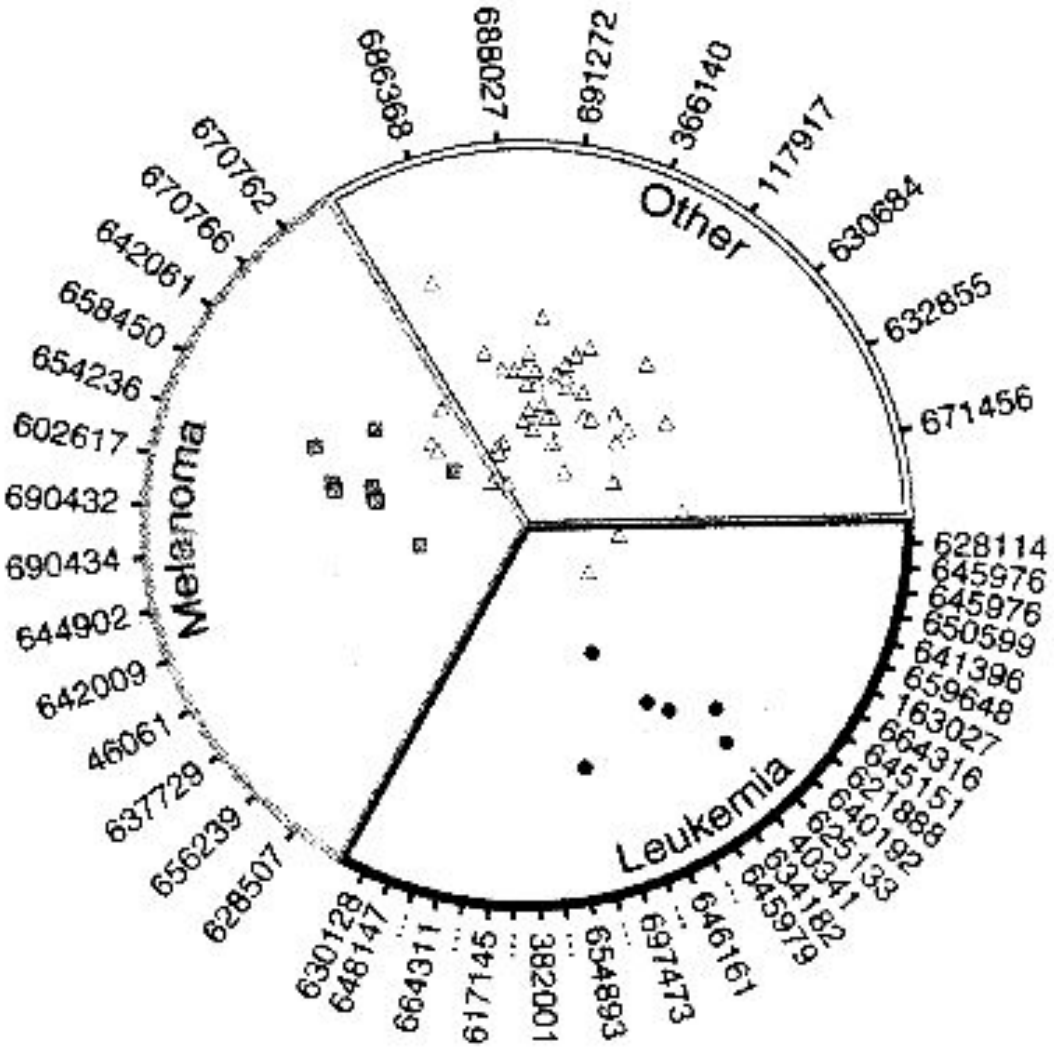
Visualização - Exemplo 2



Visualização - Exemplo 2

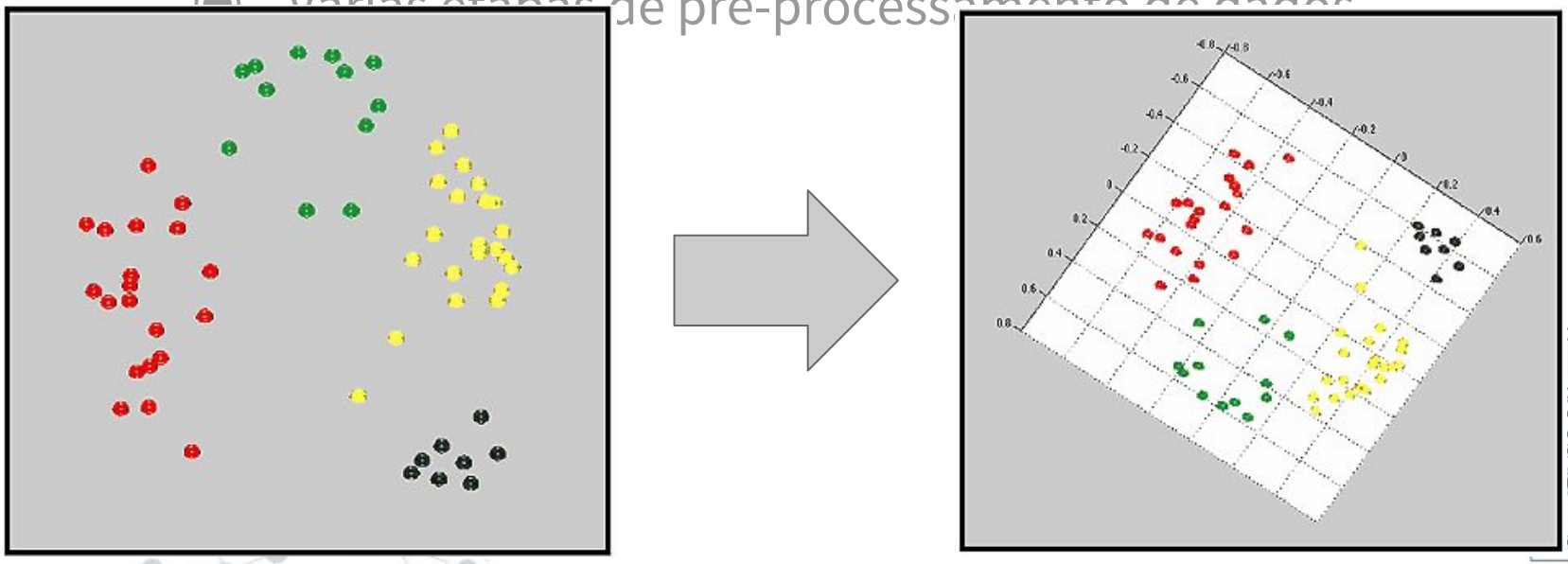


Visualização - Exemplo 2



Visualização - Exemplo 3

- (Khan, et al., 2001) resumiu os resultados da análise em uma visualização planar que mostra uma clara separação dos casos diagnósticos
  - Dados não podem ser rastreados até os genes originais
    - Plotagem obtida por escalonamento multidimensional
    - Várias etapas de pré-processamento de dados

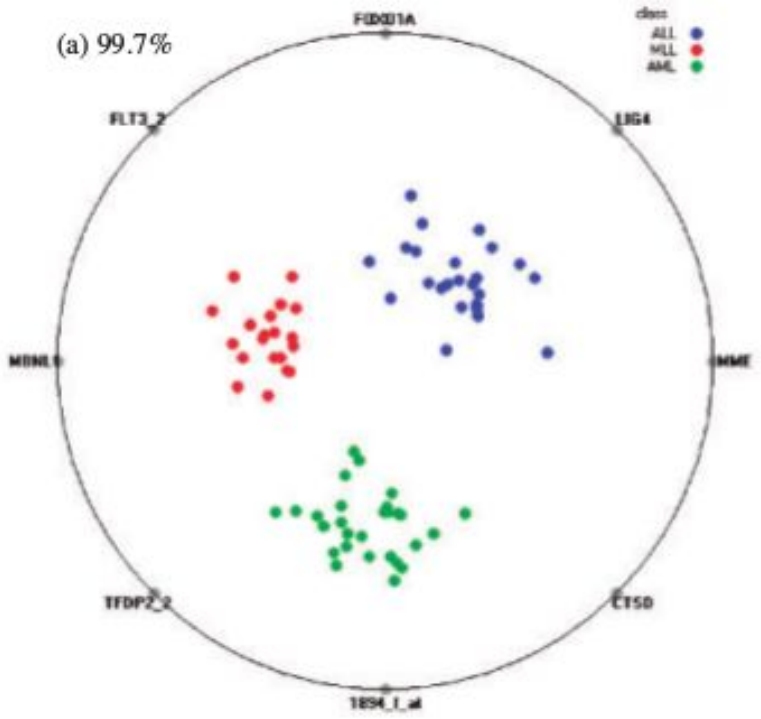


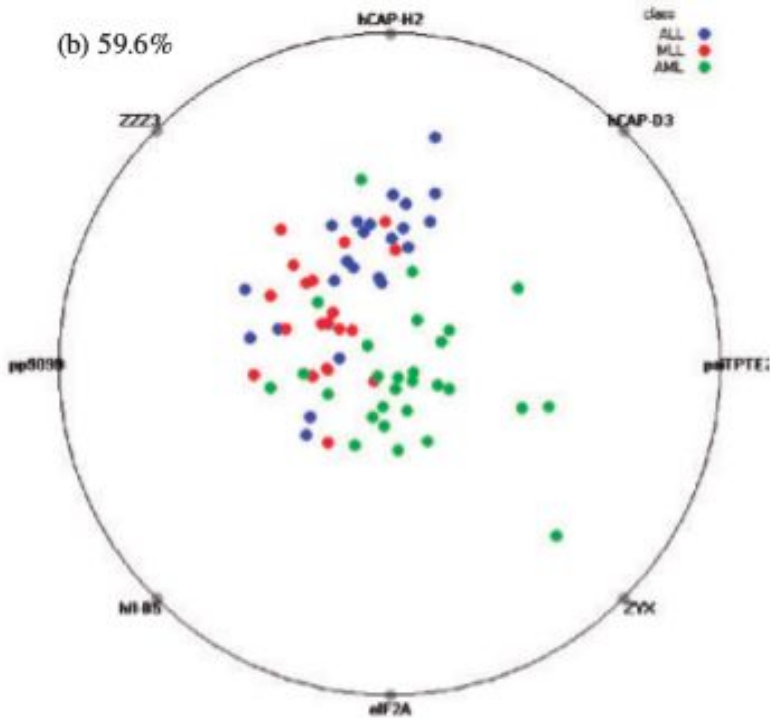
Visualização

- Não é trivial achar uma projeção limpa e que separe as classes
  - Existem milhões de projeções possíveis no Radviz
- VizRank (Leban et al., 2006)
  - Método para ordenar projeções visuais de dados com classificação de classe por seu potencial de interesse
  - Se concentra em um pequeno subconjunto de visualizações que são mais prováveis fornecer a melhor visão sobre os dados
    - Análise de dados não procura aleatoriamente entre milhões de possíveis projeções

Visualização - Exemplo 4

- VizRank (Leban et al., 2006)
  - Define o interesse da projeção estimando o quanto as instâncias de dados da mesma classe são agrupadas e separadas das instâncias de outras



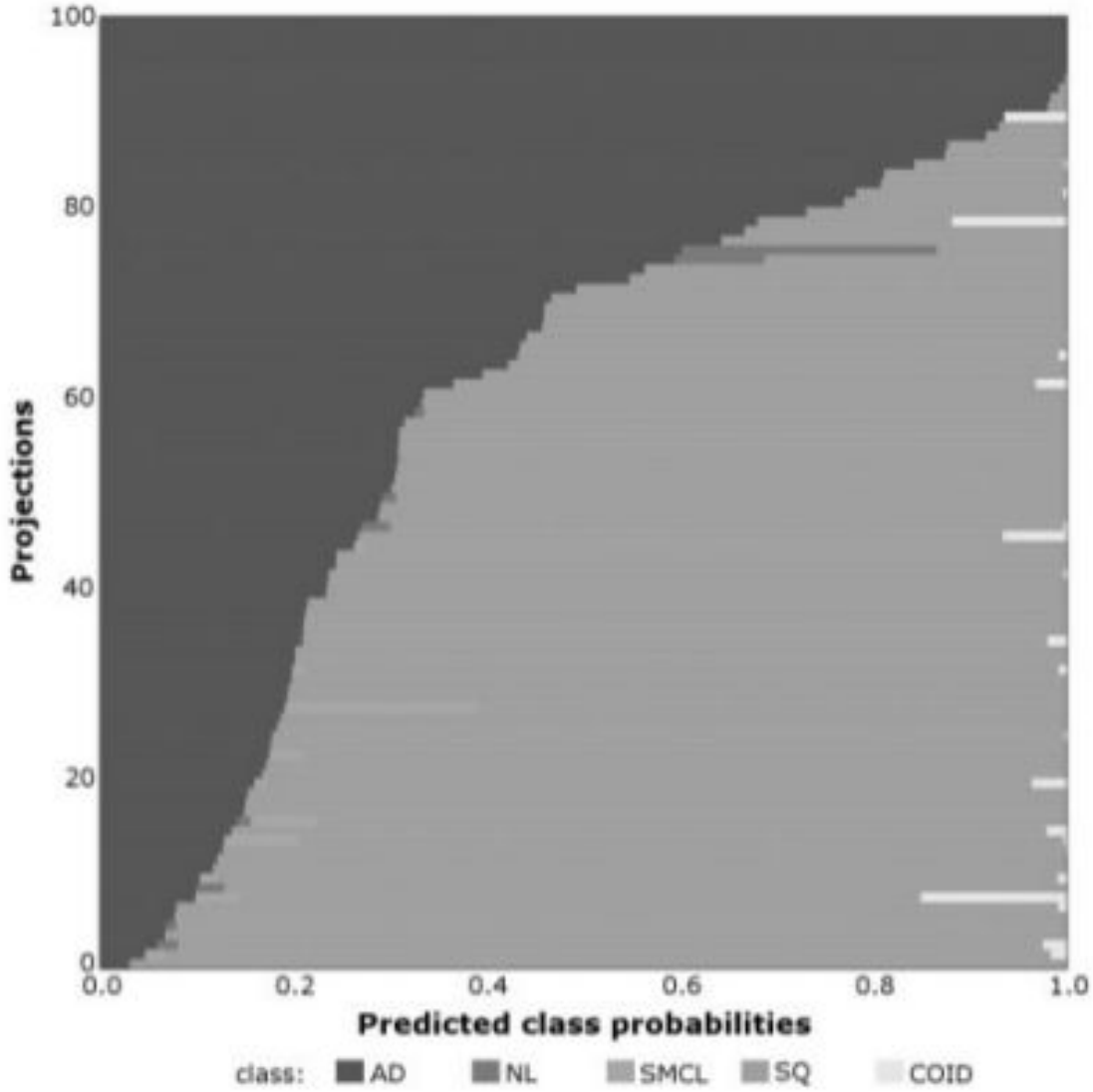


Visualização - Extensão VizRank

- Classificação
  - As projeções achadas pelo VizRank podem ser usadas para classificar novas amostras
  - A posição da amostra na projeção é determinada por sua expressão dos genes utilizadas na projeção
  - A amostra é então classificada para a classe predominante de amostras k-mais próximas da visualização original
  - O algoritmo de classificação é, portanto, o mesmo que o utilizado no ranking das projeções

Visualização - Extensão VizRank

- Classificação

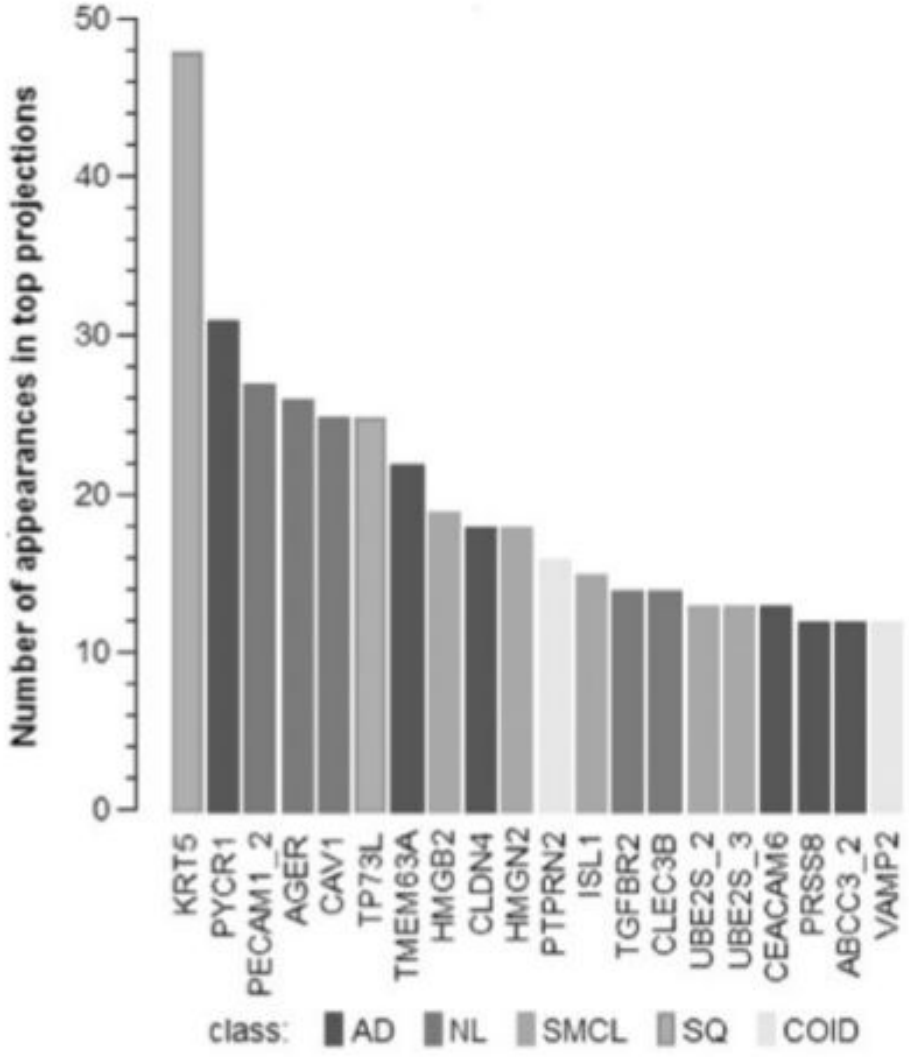


Visualização

- Ranqueamento de *features*
  - Espera-se que os genes que aparecem nas projeções mais bem classificadas sejam aqueles que detêm mais informações para discriminação de classe
  - Pontuação utilidade genética
    - Número de aparições do gene em P melhores projeções ranqueadas

Visualização

- Ranqueamento de *features*



Visualização - Extensão VizRank

- Detecção de *outliers*
  - A identificação e análise dos outliers nas melhores projeções podem revelam características interessantes dos dados
    - Casos especiais de uma doença específica
    - Amostras diagnosticadas erroneamente
  - Técnica automática que suporta análise exploratória de dados e examina um único caso selecionado
    - Relata suas probabilidades de classe previstas usando um conjunto de visualização mais bem classificadas

Conclusão

Conclusão

- Método para analisar dados de expressão gênica
  - Fornece um modelo de classificação confiável
  - Fornece uma visão valiosa dos dados na forma de visualizações informativas
- O método proposto de ranqueamento e classificação de projeções
  - Pode encontrar visualizações simples de conjuntos de dados de expressão gênica de câncer que usam um subconjunto muito pequeno

Conclusão

- Devido ao potencial na análise exploratória de dados, tempos de execução curtos e interface interativa
  - Visualização de dados suportada com técnicas eficientes de pesquisa de projeção deve complementar outras técnicas estabelecidas na análise de *microarrays* de câncer e se tornar parte das ferramentas de análise padrão

Conclusão

- VizRank e Radviz são implementados como parte do Orange data mining suite. Disponível em http://www.ailab.si/orange
- Os dados utilizados estão disponíveis em http://www.ailab.si/supp/bi-cancer
- Citado por 74 artigos no Google Scholar e 37 no Web of Science
  - A maioria em conferências/jornais de Bioinformática/ Biomedicina
  - Mas também foi citado em conferências de visualização

Perguntas

Perguntas

1. O que é um microarray de DNA?
2. Como a visualização pode ajudar no diagnóstico de câncer?
3. Como é a procura de projeções do VizRank?

Obrigada!

Perguntas?

tamiresbs@usp.br