Dados Biológicos: Expressão Gênica

Estagiário PAE: Pablo Andretta Jaskowiak
Professor: Ricardo J. G. B. Campello

SCC0173 | Mineração de Dados Biológicos

Créditos

- Partes destes slides são baseadas em materiais de
  - Ivan Gesteira Costa Filho
    - http://www.cin.ufpe.br/~igcf/
  - Marcílio Carlos Pereira de Souto
    - http://www.dimap.ufrn.br/~marcilio/
- Agradecimento
  - Bruno Feres pelos comentários e sugestões

Sumário

- Introdução
- Conceitos Biológicos
- Análise de Expressão Gênica
- Tecnologia de Microarray

Introdução

- Pré-Processamento
- Seleção de atributos
- Classificação
- Agrupamento

Conceitos Biológicos

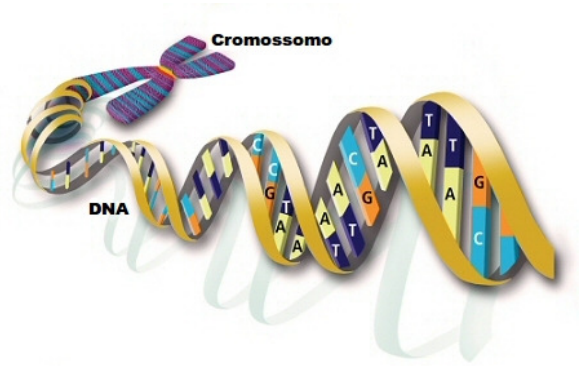
- Compreendendo a vida a nível celular
  - Como a informação genética é herdada?
  - Como a informação genética Influencia processos celulares?
  - Como genes interagem para realização de processos celulares?

Conceitos Biológicos

- Duas moléculas importantes
  - Ácido Desoxirribonucleico (DNA)
  - Proteínas

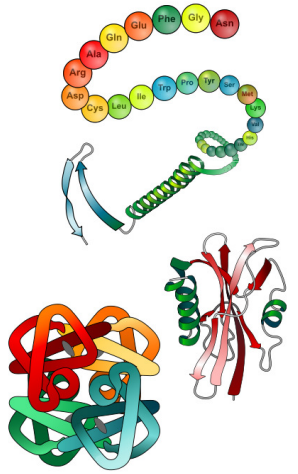
Conceitos Biológicos

- Informação genética
  - Ácido Desoxirribonucleico (DNA)
  - Cadeia de bases
    - Adenina
    - Timina
    - Citosina
    - Guanina



Conceitos Biológicos

- Proteínas
  - Cadeia de aminoácidos
    - 20 diferentes aminoácidos
  - Entidades funcionais das células
  - Responsáveis
    - Catalisar reações químicas
    - Formação de estruturas

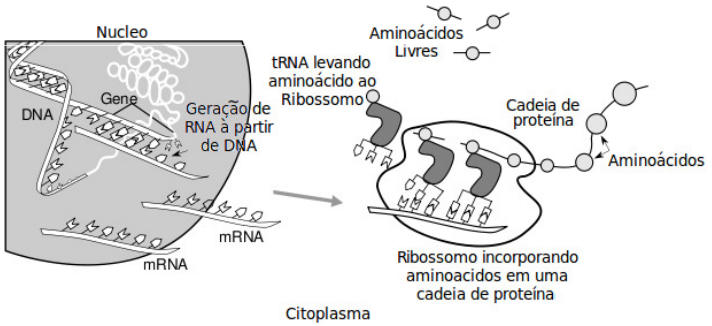


Conceitos Biológicos

Como informação contida no DNA se transforma em proteínas?

Conceitos Biológicos

- Processo de expressão gênica
  - Dogma Central da Biologia Molecular



Conceitos Biológicos

- Células de um mesmo organismo
  - Mesmo DNA genômico
- Identidades celulares distintas
  - Células especializadas
  - Diferenças na expressão gênica
    - Transcrição e Tradução
- Transcrição ou não de um gene
  - Determinada pela presença/ausência de outros produtos dos genes (especialmente proteínas)

Conceitos Biológicos

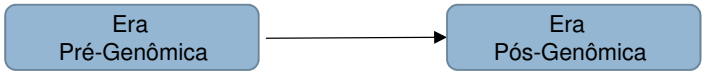
- Genes interagem em redes complexas
  - Gene A ativa gene B
  - Gene B desliga gene C
  - Gene C aumenta expressão do gene A
- Perturbações em um único gene
  - Mudança no nível de expressão de diversos genes

Conceitos Biológicos

- Células e tecidos com funções normais
  - Genes expressos de forma regulada
- Expressão alterada de um ou mais genes
  - Podem alterar o homeostase do organismo
  - Possível surgimento de doenças

Análise de Expressão Gênica

Análise de Expressão Gênica



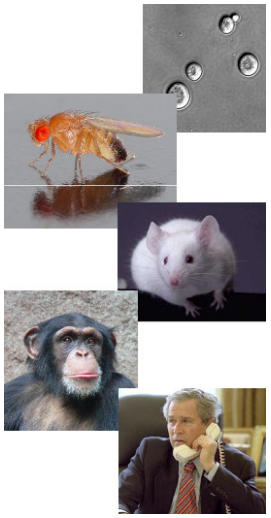
- Próximo passo após o sequenciamento
  - Compreensão das conexões entre
    - Seqüências de DNA e características fenotípicas dos organismo
  - Passo complexo!
  - Proteínas e genes interagem em redes altamente conectadas
- Tradicionalmente
  - Biologia Molecular trabalha com o paradigma
    - Um gene - uma função

Análise de Expressão Gênica

- Técnicas de Low Throughput
  - Quatro Principais
    - Eastern Blot (DNA)
    - Northern Blot (RNA)
    - Southern Blot (Proteína)
    - PCR
  - Poucos genes por experimento
  - Alta precisão
  - Baixa quantidade de ruído
  - Não permite uma visão geral de um processo

Análise de Expressão Gênica

- Organismos Complexos
  - Levedura 6,000 genes
  - Drosophila 13,500 genes
  - Camundongo 22,000 genes
  - Chimpanzé 20,000 genes
  - Homo Sapiens 23,000 genes



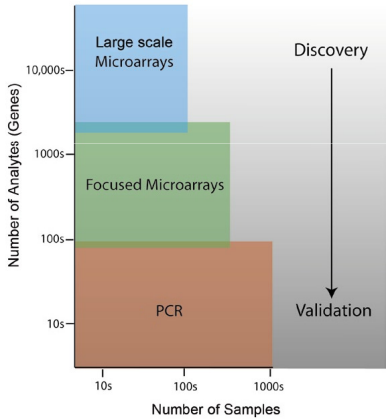
Análise de Expressão Gênica

- Como estudar tais organismos?
- Como atribuir funções a genes?
- Como compreender interações entre genes?
- Como compreender seu funcionamento como um todo?

Análise de Expressão Gênica

- Tecnologias de High Troughput
  - SAGE (Serial Analysis of Gene Expression)
  - MPSS (Massively Parallel Signature Sequencing)
  - **Microarrays**

- Métodos recentes
  - Abordagem diferenciada

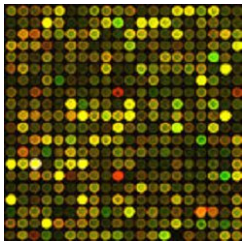


Tecnologia de Microarray

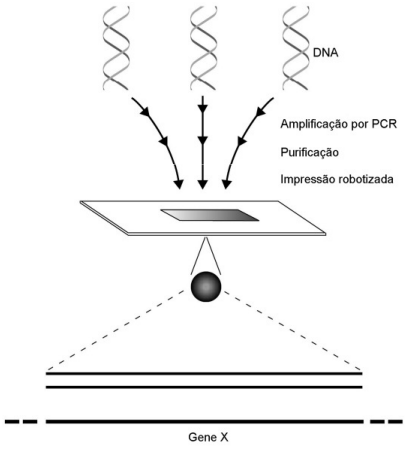
- Análise em alta escala
  - Nível de genoma
  - Visão global de processos ou amostras de tecidos
  - Medição do nível de expressão de vários genes
- Baixo custo
  - Financeiro
  - Tempo

Tecnologia de Microarray

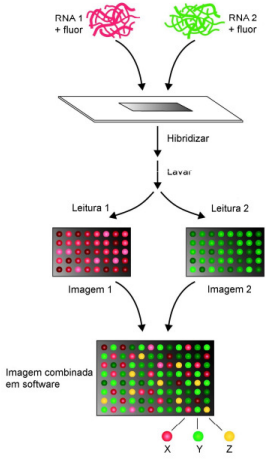
- Milhares de *spots* representando genes
- Intensidade de cada *spot*
  - Quantidade de cDNA hibridizado
  - Quantidade de mRNA transcrito na amostra



Tecnologia de Microarray



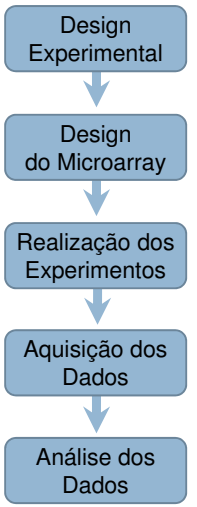
Fabricação



Experimento

Tecnologia de Microarray

- Cinco Principais Etapas



Tecnologia de Microarray

- Cinco Principais Etapas

Design Experimental

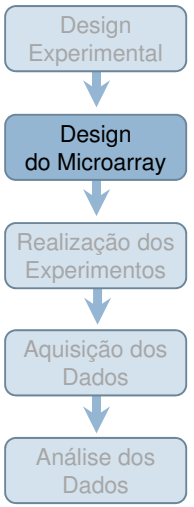
- Qual o objetivo do experimento biológico?
  - Compreensão de processos celulares
    - Expressão em relação ao tempo
  - Diagnóstico de câncer
    - Expressão de diversos pacientes e tipos de câncer
  - Desenvolvimento de drogas
    - Expressão de uma células tratadas com diferentes medicamentos
    - Expressão de uma célula com várias dosagens de medicamentos

Design Experimental

- Experimento será replicado?
  - Réplica técnica
    - Mesmo material biológico é utilizado
    - Redução de *outliers* e ruído
  - Réplica biológica
    - Diferente materiais biológicos sob as mesmas condições
    - Capturar variabilidade biológica

Tecnologia de Microarray

- Cinco Principais Etapas



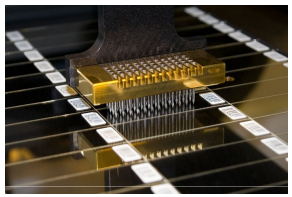
Design do Microarray

- Quais genes serão incluídos no estudo?
- Uso de múltiplas sondas por gene
  - Minimizar ruído
- Qual plataforma será utilizada?

Design do Microarray

- Plataformas comerciais
  - Diferentes características quanto à
    - Customização
    - Preço
    - Padronização
- Duas principais
  - cDNA
  - Affymetrix

Design do Microarray

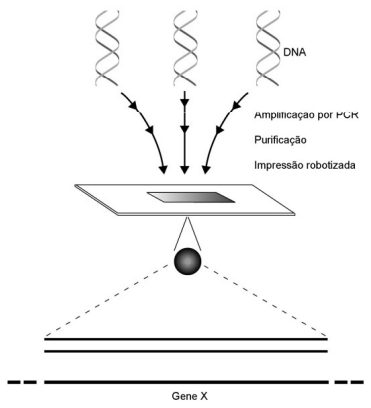


- cDNA
  - Robô deposita sondas em vidro
  - Laboratório deve possuir equipamento
  - Boa customização

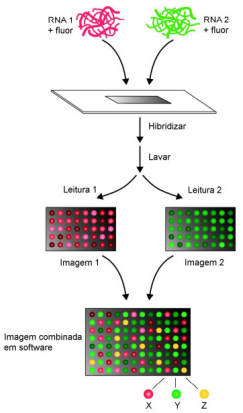
- Affymetrix
  - Plataformas já definidas
    - Padronizado
  - Arrays são comprados prontos

Tecnologia de Microarray

- cDNA



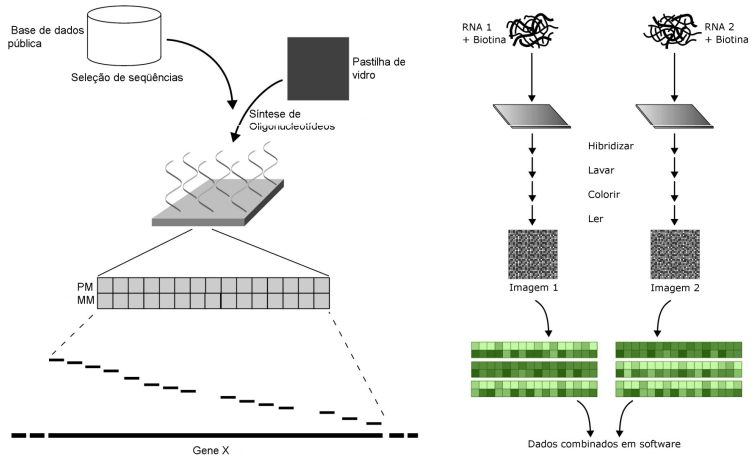
Fabricação



Experimento

Tecnologia de Microarray

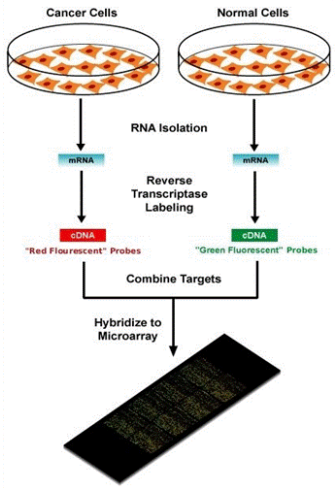
- Affymetrix



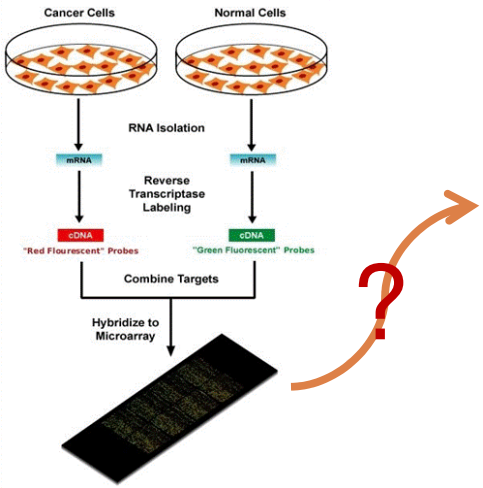
Tecnologia de Microarray

- Cinco Principais Etapas

Realização do Experimento



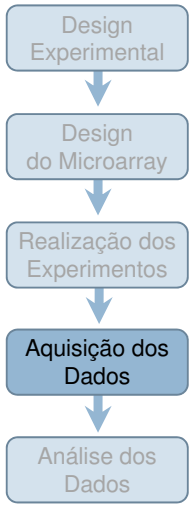
Realização do Experimento



	Amostra 1	Amostra ...	Amostra n
Gene 1	0.92	-0.32	0.23
Gene 2	1.24	8.29	2.34
Gene 3	0.99	6.32	8.09
Gene 4	1.11	4.32	5.64
Gene 5	0.09	0.00	0.01
Gene 6	0.98	0.12	19.0
Gene ...	-0.98	-0.01	1.02
Gene m-1	7.06	5.04	2.06
Gene m	-0.09	0.12	4.56

Tecnologia de Microarray

- Cinco Principais Etapas

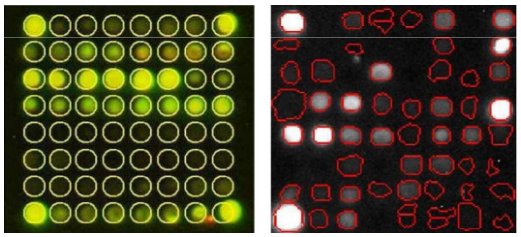


Aquisição de Dados

- Extração dos valores de expressão
  - Identificação dos spots

Aquisição dos Dados

- Imagem passa por processo de segmentação

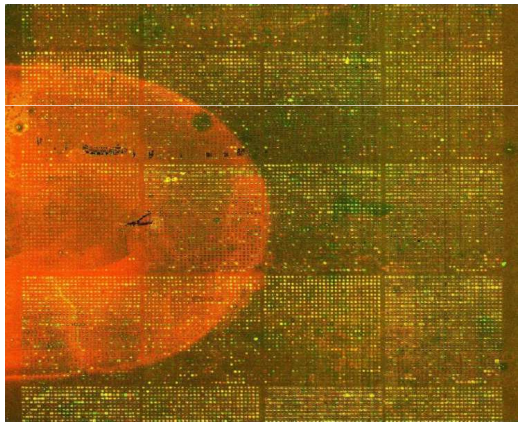


Aquisição de Dados

- Extração dos valores de expressão
  - Cálculo da intensidade dos sinais
  - Normalização de valores entre arrays
  - Detecção de ruído

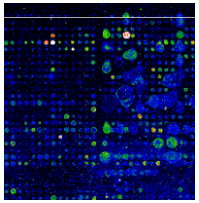
Aquisição dos Dados

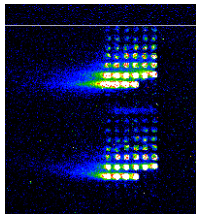
- Nem tudo é tão bonito quanto parece...

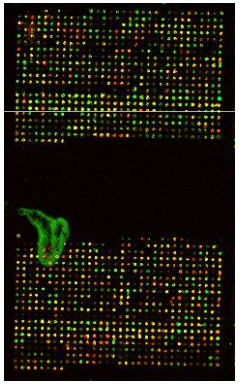


Aquisição dos Dados

- Nem tudo é tão bonito quanto parece...







Matriz de Dados

- Finalmente! A tão desejada matriz de dados!

Nossos Problemas Acabaram!!!

	Amostra 1	Amostra ...	Amostra n
Gene 1	0.92	-0.32	0.23
Gene 2	1.24	8.29	2.34
Gene 3	0.99	6.32	8.09
Gene 4	1.11	4.32	5.64
Gene 5	0.09	0.00	0.01
Gene 6	0.98	0.12	19.0
Gene ...	-0.98	-0.01	1.02
Gene m-1	7.06	5.04	2.06
Gene m	-0.09	0.12	4.56

Matriz de Dados

- Finalmente! A tão desejada matriz de dados!

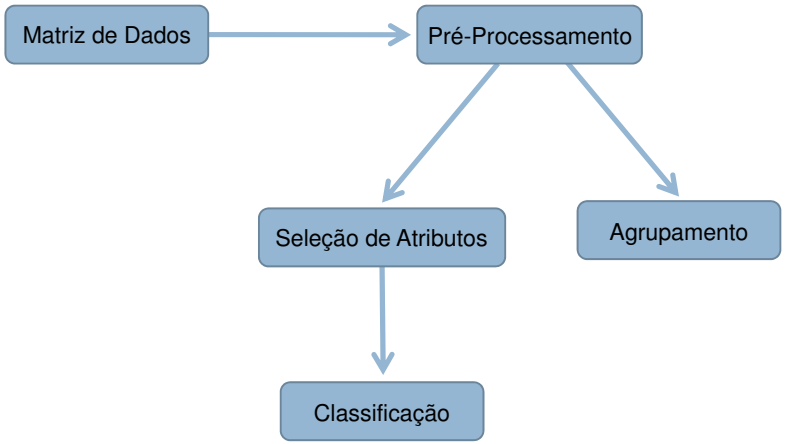
Nossos Problemas Acabaram???

	Amostra 1	Amostra ...	Amostra n
Gene 1	0.92	-0.32	0.23
Gene 2	1.24	8.29	2.34
Gene 3	0.99	6.32	8.09
Gene 4	1.11	4.32	5.64
Gene 5	0.09	0.00	0.01
Gene 6	0.98	0.12	19.0
Gene ...	-0.98	-0.01	1.02
Gene m-1	7.06	5.04	2.06
Gene m	-0.09	0.12	4.56

Tecnologia de Microarray

- Cinco Principais Etapas

Análise dos Dados



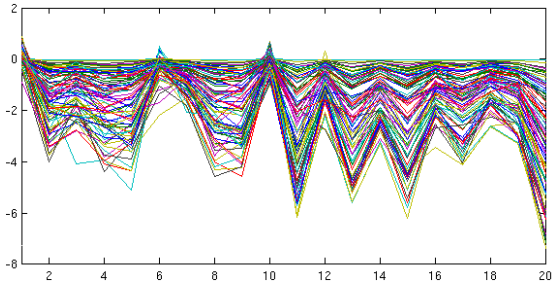
Análise de Dados

- Dados com características especiais
  - Grande quantidade de genes
  - Quantidade relativamente pequena de amostras
  - Ruídos
  - *Outliers*
  - Valores ausentes
- Pré-Processamento é essencial!

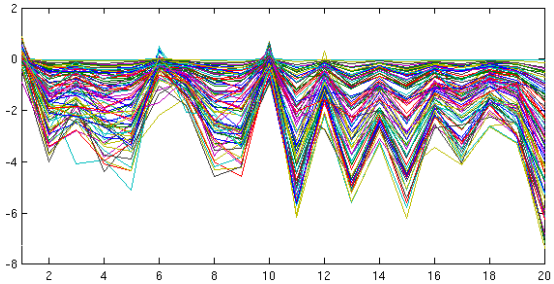
Análise de Dados

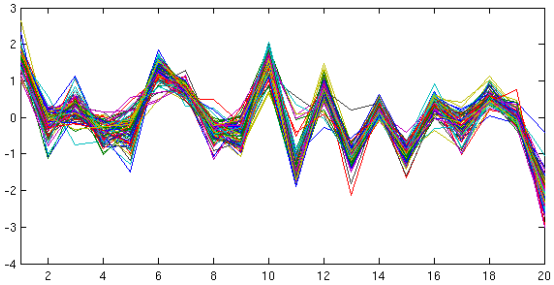
- Pré-Processamento
  - Tratar valores ausentes
  - Outliers
  - Genes sem alteração significativa de nível de expressão
  - Normalização / Padronização

Análise de Dados



Análise de Dados





Análise de Dados

- O que fazer com os dados?
  - Dependente do objetivo do experimento biológico
    - Compreensão de processos celulares
      - Expressão em relação ao tempo
    - Diagnóstico de câncer
      - Expressão de diversos pacientes e tipos de câncer
    - Desenvolvimento de drogas
      - Expressão de uma células tratadas com diferentes medicamentos
      - Expressão de uma célula com várias dosagens de medicamentos

Análise de Dados

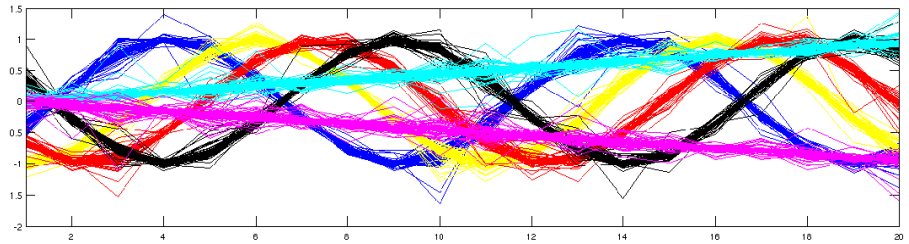
- Agrupamento
  - Genes
  - Amostras
- Seleção de atributos e Classificação
  - Amostras

Análise de Dados

- Agrupamento de dados
  - Genes
    - Séries temporais
    - Identificação de genes co-expressos
    - Mesmo nível de expressão ao longo do tempo
    - Regulados pelos mesmos genes
    - Identificação de funcionalidades

Análise de Dados

- Agrupamento de dados
  - Genes



Análise de Dados

- Agrupamento de dados
  - Genes
- Trabalho de Eisen *et al.*
  - Trabalho pioneiro
  - Agrupamento de genes de levedura
  - 6621 genes originais – 2467 genes restantes com variabilidade
  - Identificação de genes com funções similares
  - Sugere funções para genes com funções ainda desconhecidas

Eisen, M. B.; Spellman, P. T.; Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns
Proceedings of the National Academy of Sciences of the United States of America, 1998, 95, 14863-14868

Análise de Dados

- Agrupamento de dados
  - Amostras
    - Identificação de subtipos de doenças
  - Trabalho de Golub *et al.*
    - Distinção entre grupos de amostras de leucemia
      - AML (11 amostras)
      - ALL (27 amostras)
      - 6817 genes humanos

Golub, T. R. et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring
Science, 1999, 286, 531-537

Análise de Dados

- Classificação
  - Poucos objetos
  - Alta dimensionalidade
- Seleção de atributos!

	Amostra 1	Amostra ...	Amostra n
Gene 1	0.92	-0.32	0.23
Gene 2	1.24	8.29	2.34
Gene 3	0.99	6.32	8.09
Gene 4	1.11	4.32	5.64
Gene 5	0.09	0.00	0.01
Gene 6	0.98	0.12	19.0
Gene ...	-0.98	-0.01	1.02
Gene m-1	7.06	5.04	2.06
Gene m	-0.09	0.12	4.56

Análise de Dados

- Seleção de atributos
  - Grande parte das aplicações envolvem filtros
    - Métodos de rank – rapidez
- Classificação
  - Grande diversidade de métodos
  - kNN, SVMs, Naïve Bayes, Árvores de Decisão, ...
- Comparação de métodos
  - Dudoit *et al.*
  - Li *et al.*

*Dudoit *et al.* Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data Journal of the American Statistical Association, 2002, 97, 77-87

*Li, T. *et al.* A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression Bioinformatics, 2004, 20, 2429-2437

Dúvidas?