ESTUDO DE COEFICIENTES DE CORRELAÇÃO PARA MEDIDAS DE PROXIMIDADE EM DADOS DE EXPRESSÃO GÊNICA

Pablo Andretta Jaskowiak
Orientador: Prof. Dr. Ricardo J. G. B. Campello

Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
São Carlos, 2 de março de 2011

Sumário

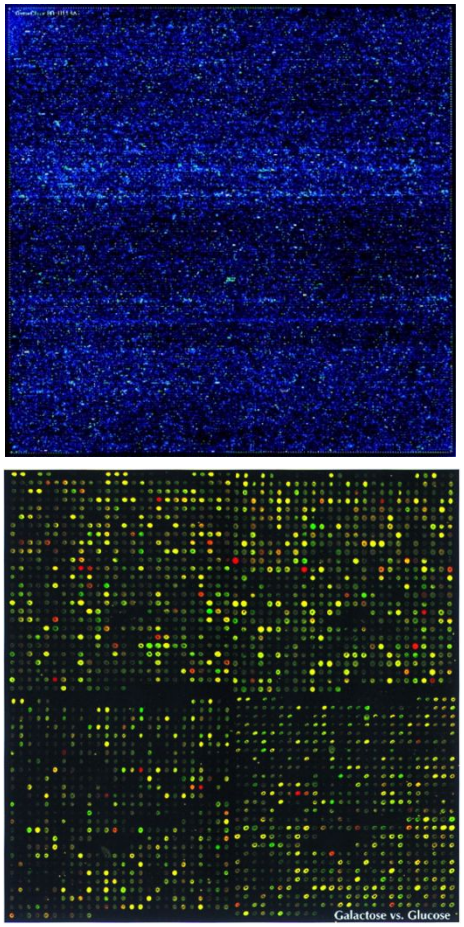
- Introdução
- Análise de Dados de Expressão Gênica
- Medidas de Proximidade em Dados de Expressão Gênica
- Resultados
  - Agrupamento
  - Seleção de Atributos e Classificação
- Conclusões e Contribuições

Introdução

- Próximo passo após o sequenciamento
  - Compreensão das conexões entre
    - Seqüências de DNA e características fenotípicas dos organismos
  - Proteínas e genes interagem em redes altamente conectadas
- Tradicionalmente
  - Biologia Molecular trabalha com o paradigma
    - Um gene - uma função
- Novas tecnologias
  - Medição dos níveis de expressão de genes a nível genômico

Introdução

- Tecnologia de *microarray*
  - Análise em alta escala
  - Custo relativamente baixo
- Duas principais tecnologias
  - Oligonucleotideos (Affymetrix)
  - cDNA



Figuras:

Wilson, K. E. et al. (2004). Functional genomics and proteomics: application in neurosciences. J. of Neurology, Neurosurgery & Psychiatry.
Lashkari, D. A. et al. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. PNAS of the United States of America.

Introdução

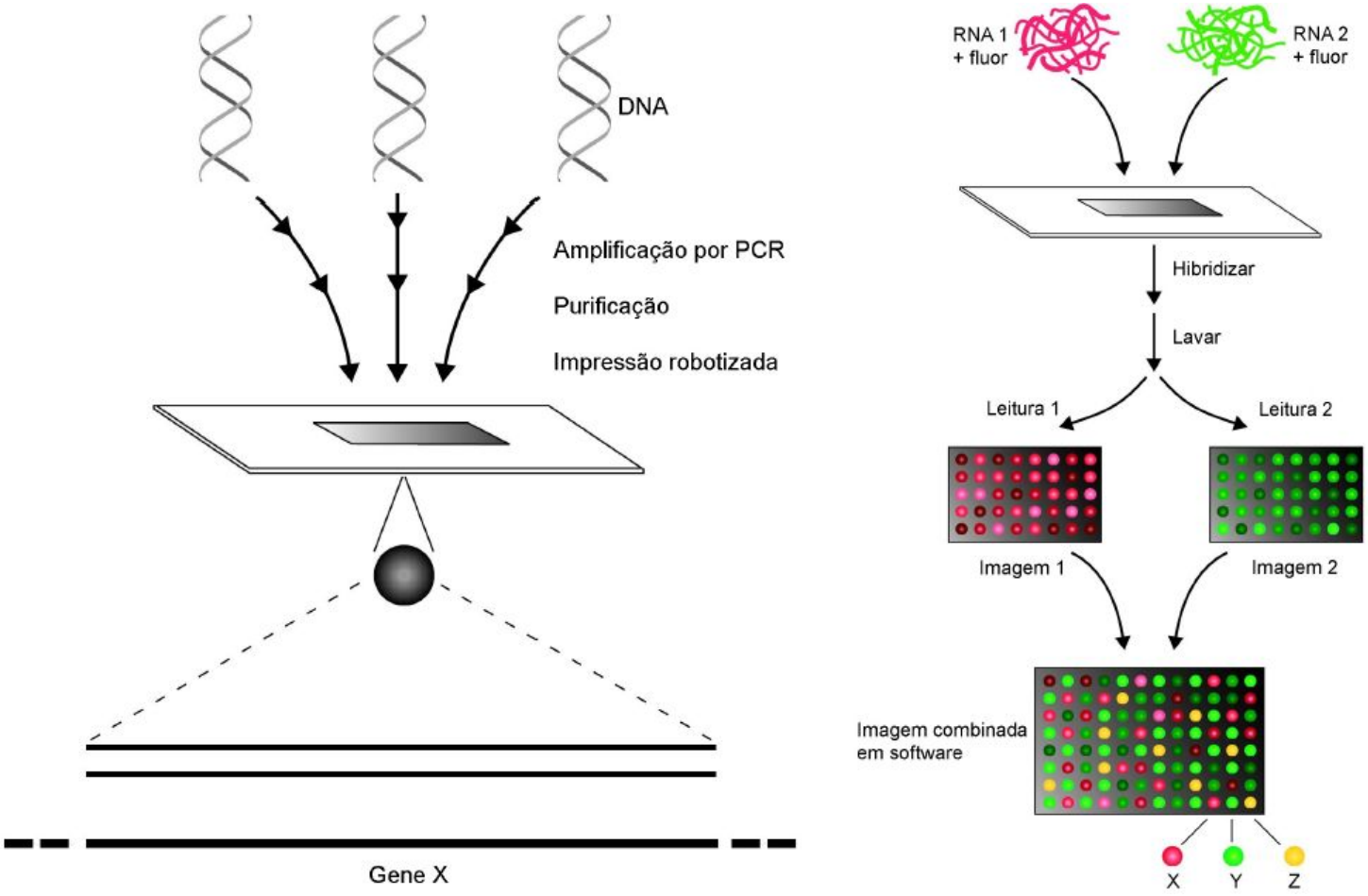
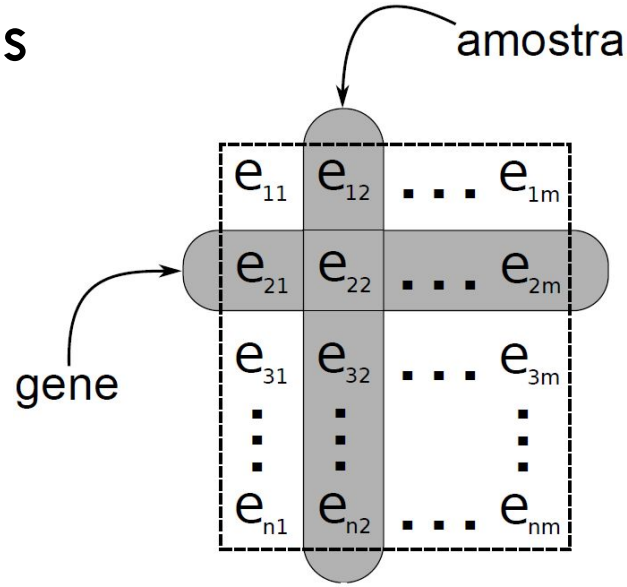


Figura adaptada de:
Harrington, C. A., Rosenow, C., e Retief, J. (2000). Monitoring gene expression using DNA microarrays. Current Opinion in Microbiology.

Introdução

- Dados com características especiais
  - Grande quantidade de genes
  - Pequena quantidade de amostras
  - Ruído
  - *Outliers*
  - Valores ausentes



Medição dos Níveis de Expressão

↓

Análise dos Dados Obtidos

Figura adaptada de:
Zhang, A. (2006). Advanced analysis of gene expression microarray data.

Análise de Dados de Expressão Gênica

- Três principais tarefas envolvidas na sua análise
  - Agrupamento de dados
    - Agrupamento de amostras
    - Agrupamento de genes
  - Seleção de atributos
    - Seleção de genes
  - Classificação
    - Classificação de amostras
- Métodos de diferentes áreas têm sido aplicados
  - Aprendizado de Máquina, Estatística, Mineração de Dados
  - Diversos métodos baseados em proximidade

Análise de Dados de Expressão Gênica

- Dados de Expressão Gênica
  - Considerando dois
    - Genes
    - Amostras
  - Duas sequências numéricas

$$a = (a_1, a_2, \ldots a_p)$$

$$b = (b_1, b_2, \ldots b_p)$$

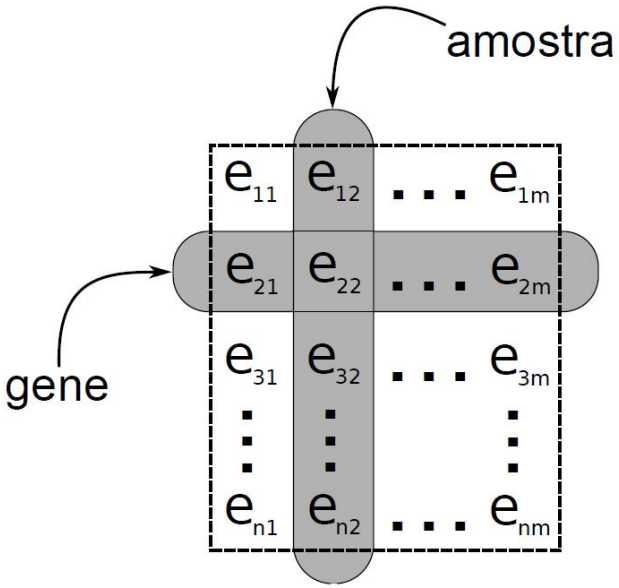


Figura adaptada de:
Zhang, A. (2006). Advanced analysis of gene expression microarray data.

Motivação

- Similaridade em tendência ou forma
  - Coeficientes de correlação
- Correlação de Pearson medida predominante
  - Em menor proporção correlação de Spearman
- Diferentes medidas existentes na literatura
  - Medidas específicas têm sido propostas
- Poucos trabalhos têm se preocupado com sua avaliação
  - Diferentes cenários possíveis

Objetivos

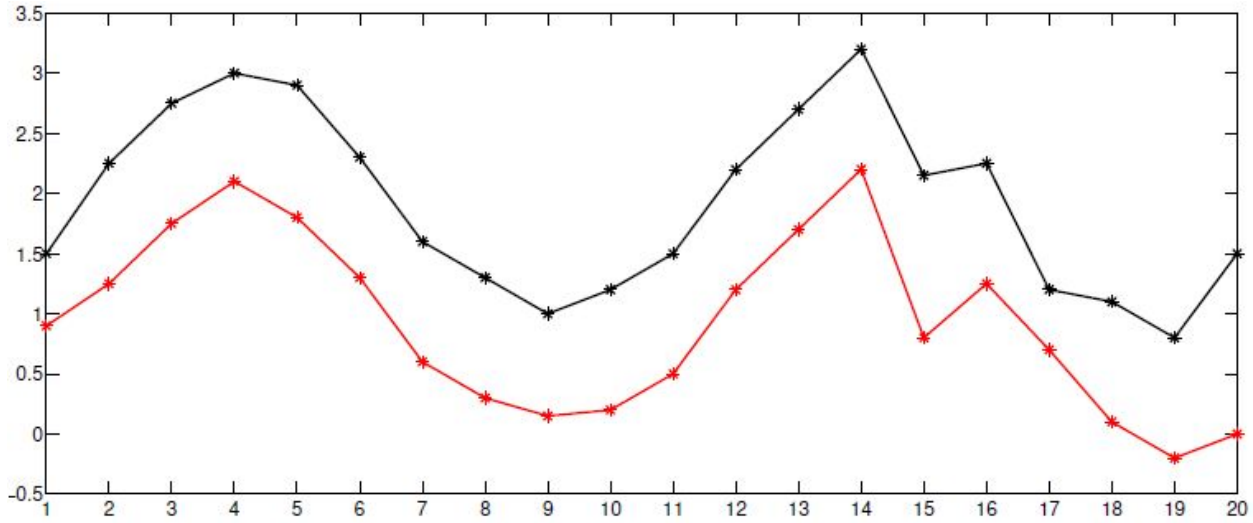
- Avaliar de maneira experimental
  - Diferentes coeficientes de correlação
  - Principais tarefas de análise de dados de *microarray*
  - Diferentes tipos de bases: cDNA e Affymetrix

- Tarefas consideradas
  - Agrupamento
    - Amostras
    - Genes
  - Seleção de genes para classificação de amostras
  - Classificação de amostras (sem seleção de genes)

Medidas de Proximidade

Medidas de Proximidade

- Dados de Expressão Gênica
  - Similaridade em forma ou tendência



  - Coeficientes de Correlação
    - Medidas comumente utilizadas

Medidas de Proximidade

- Considerando duas sequências ***a*** e ***b***
  - Genes
  - Amostras

Correlação	Sensibilidade	Complexidade
Pearson	Magnitudes	$O(n)$
Spearman	*Ranks*	$O(n \log n)$
Kendall	*Ranks*	$O(n \log n)$
Goodman-Kruskal	*Ranks*	$O(n \log n)$
Goodman-Kruskal Ponderado	Magnitudes e *ranks* de ambas as sequências	$O(n^2)$
Rank-Magnitude*	*Ranks* de ***a*** e magnitudes de ***b***	$O(n \log n)$
Jackknife	Magnitudes	$O(n^2)$

Tabela adaptada de:
Campello, R. J. G. B. e Hruschka, E. R. (2009). On comparing two sequences of numbers and its applications to clustering analysis. Information Sciences.

Medidas de Proximidade

- Medidas baseadas em Coeficientes de Correlação
  - Propostas para agrupamento de genes
  - Genes vistos como séries temporais

- Consideradas para comparação
  - YR1[1]
  - YS1[1]
  - Dissimilaridade *Short Time-Series*[2]

[1]Son, Y. S. e Baek, J. (2008). A modied correlation coefficient based similarity measure for clustering time-course gene expression data. Pat. Recognition Letters.
[2]Möller-Levet *et al.* (2005). Clustering of unevenly sampled gene expression time-series data. Fuzzy Sets and Systems.

Resultados

Agrupamento

Metodologia de Avaliação

- Medidas avaliadas em quatro algoritmos
  - *Single-Linkage*
  - *Average-Linkage*
  - *Complete-Linkage*
  - *k-medoids*

- Metodologia de avaliação diferenciada
  - Agrupamento de amostras
  - Agrupamento de genes

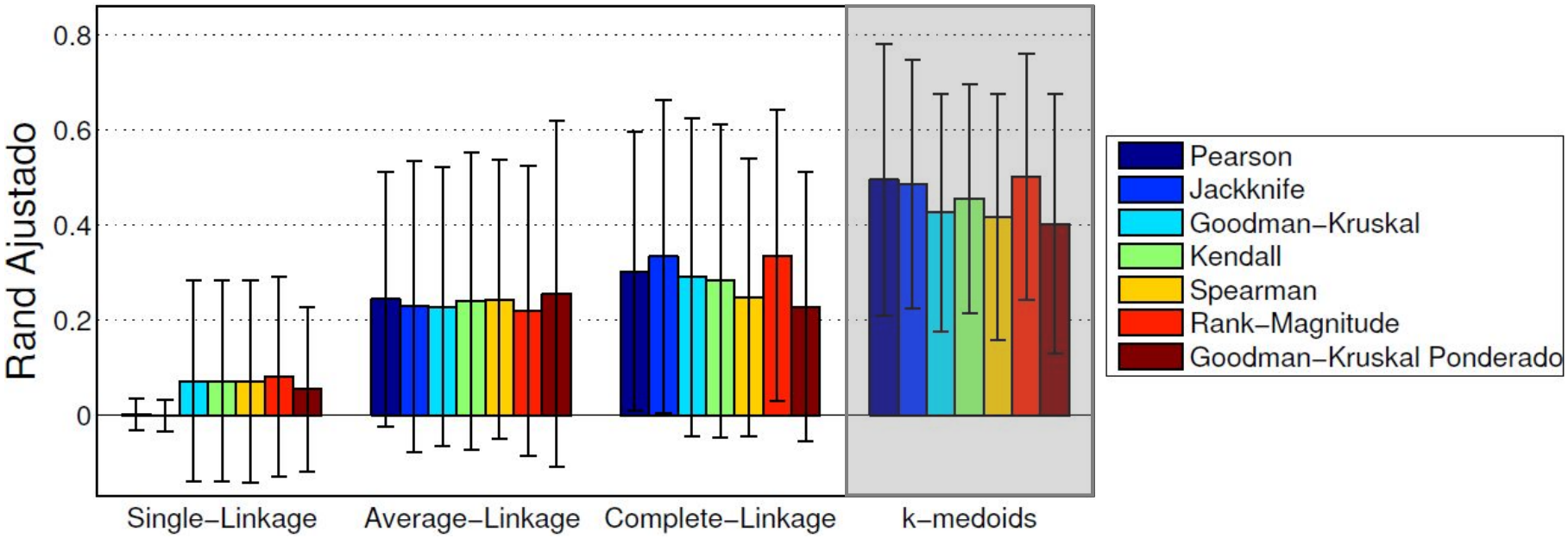
Metodologia de Avaliação - Amostras

- Bases de dados
  - 35 bases de *benchmark*[1] - agrupamento de câncer
- Três cenários de avaliação
  - Número de grupos fixo
    - Comparação entre partições com melhor Rand Ajustado
  - Número de grupos variável
    - Comparação entre partições com melhor Rand Ajustado
  - Número de grupos estimado
    - Melhor partição eleita pelo critério da Silhueta
    - Comparação entre os valores de Rand Ajustado das partições

[1] de Souto, M., Costa, I. G., de Araujo, D., Ludermir, T., e Schliep, A. (2008). Clustering cancer gene expression data: a comparative study. BMC Bioinformatics.

Agrupamento de Amostras - Resultados

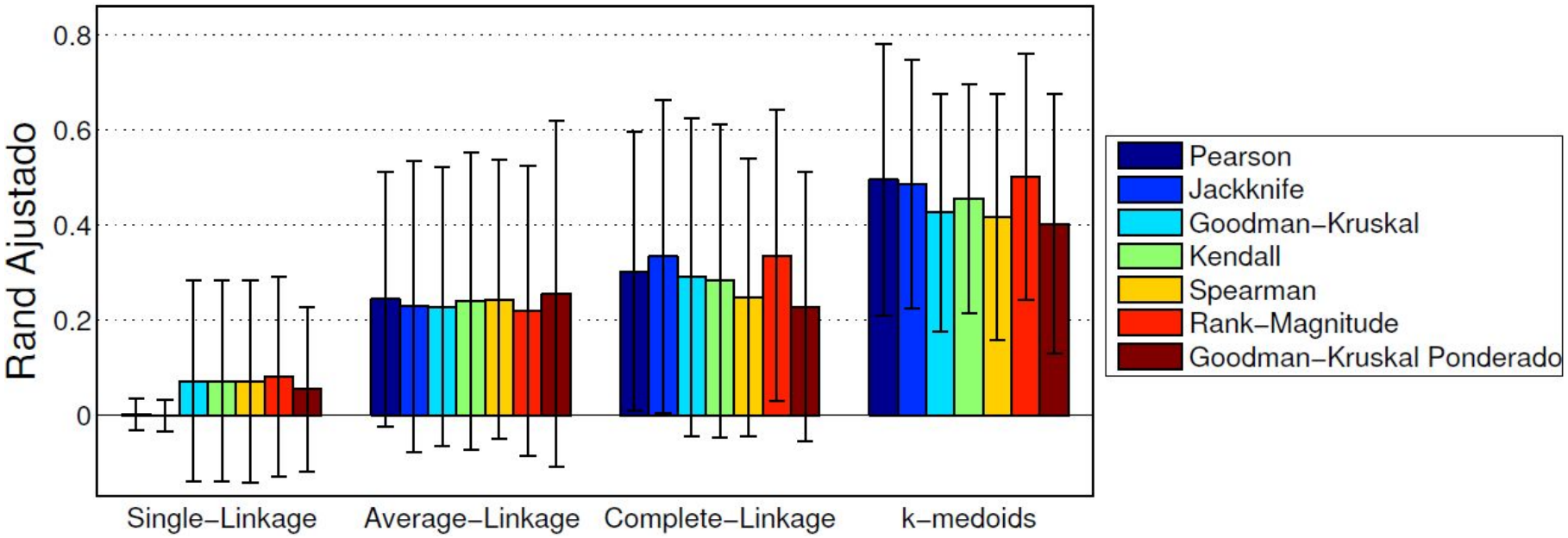
Número de grupos fixo



- Melhores resultados
  - Algoritmo *k-medoids*

Agrupamento de Amostras - Resultados

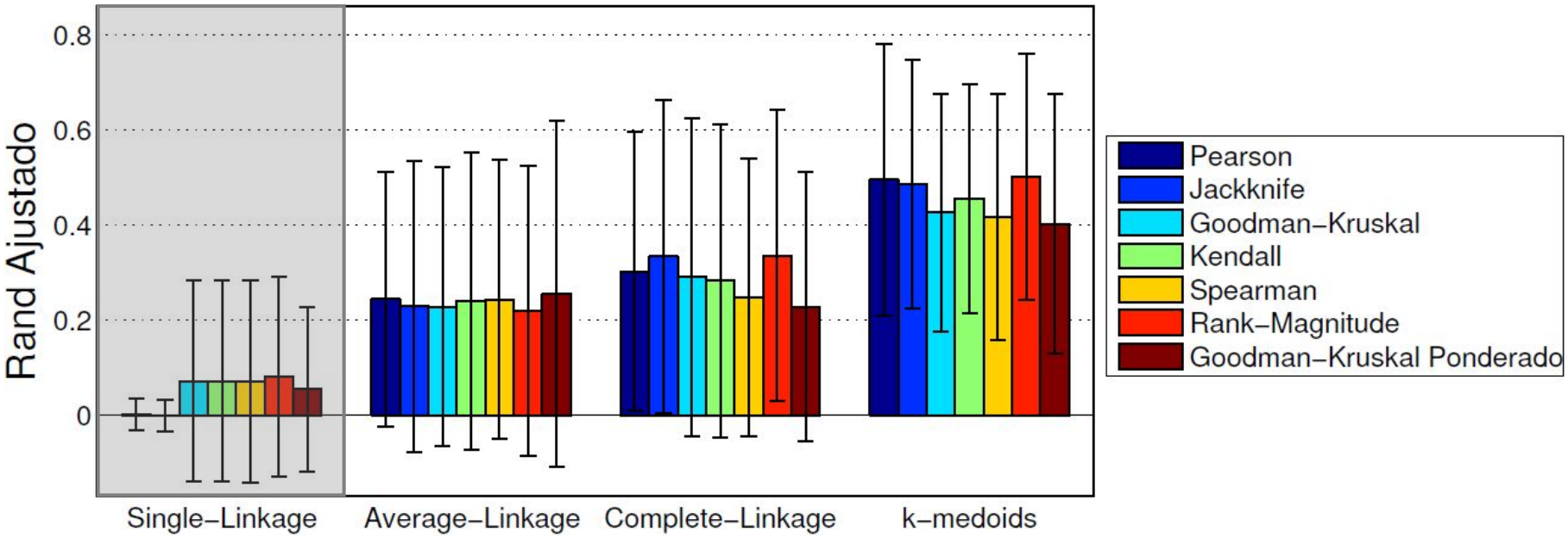
Número de grupos fixo



- Resultados superiores ou competitivos
  - Pearson, Jackknife e Rank-Magnitude

Agrupamento de Amostras - Resultados

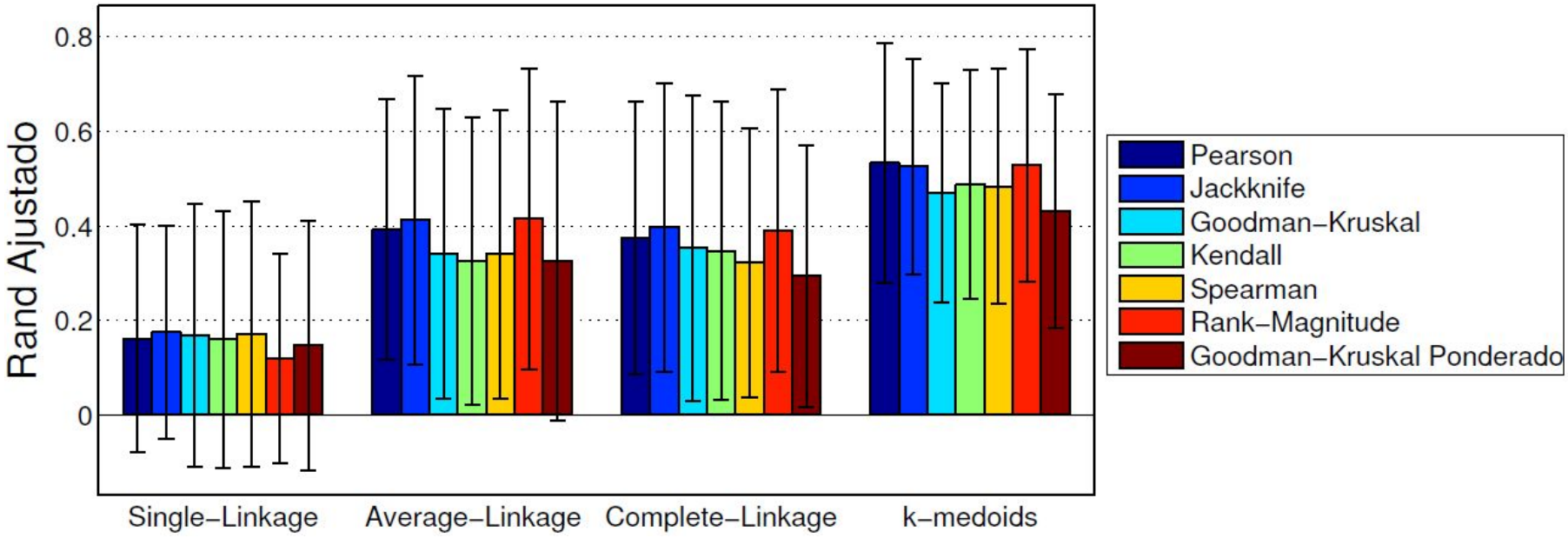
Número de grupos fixo



- Dentre todos os algoritmos
  - Piores resultados obtidos com o algoritmos *Single-Linkage*

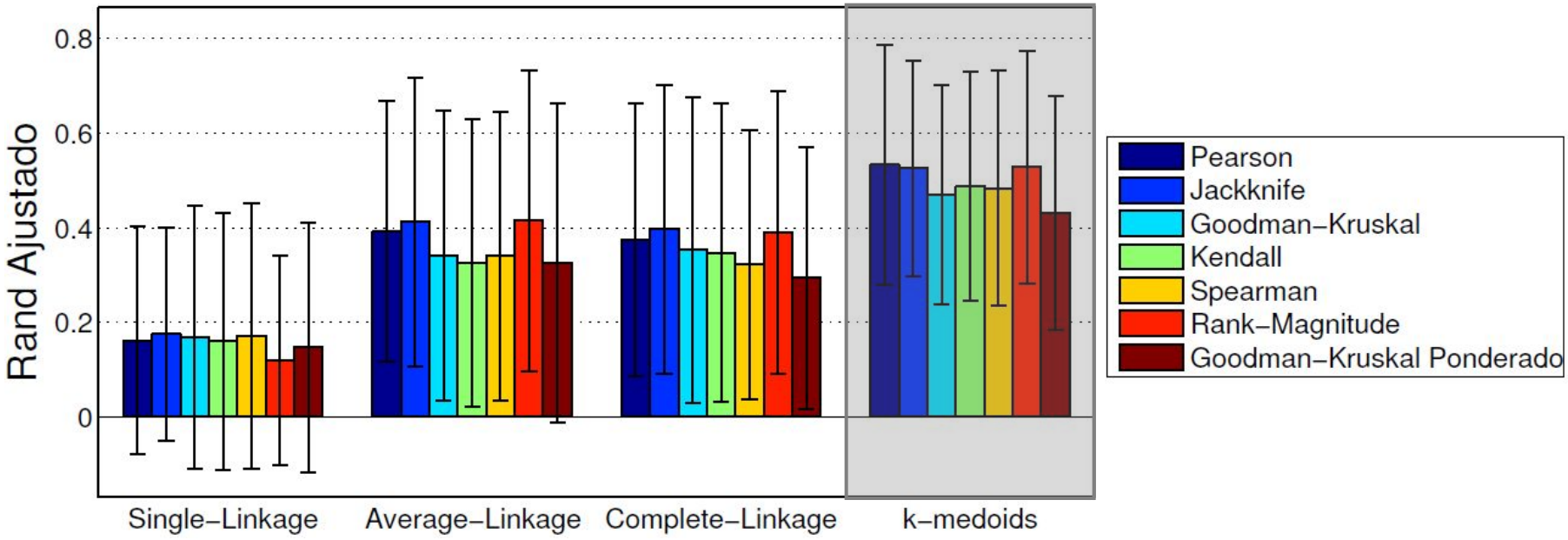
Agrupamento de Amostras - Resultados

Número de grupos variável



Agrupamento de Amostras - Resultados

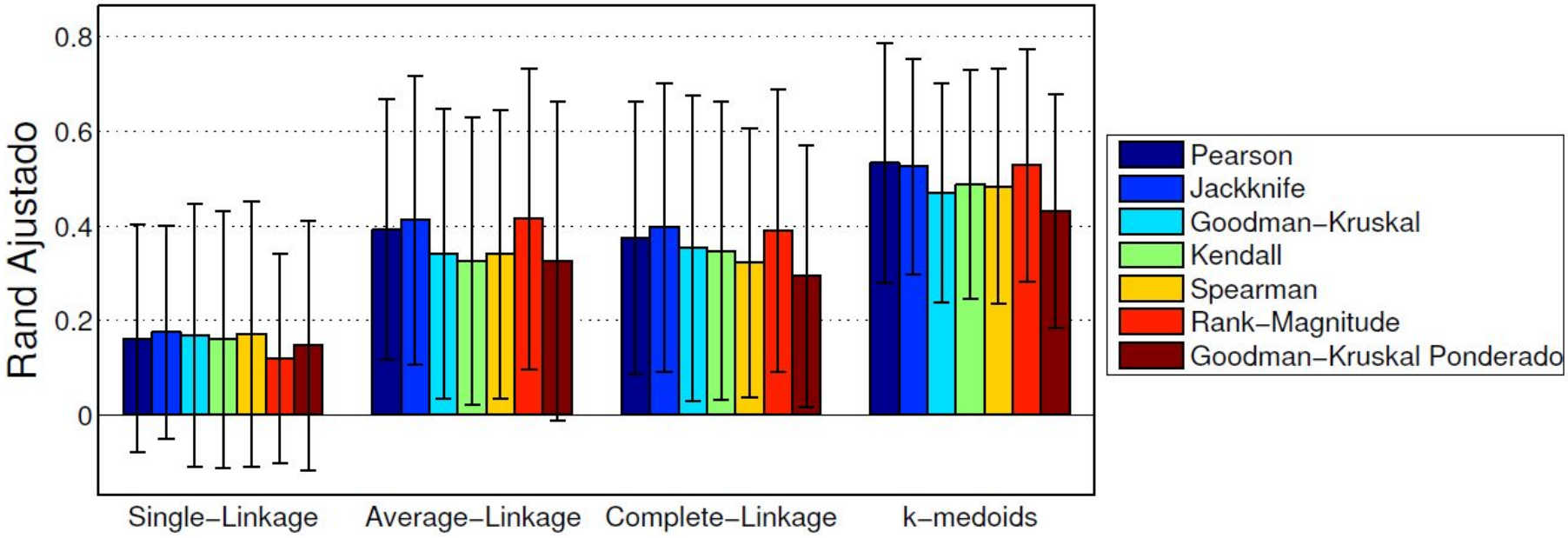
Número de grupos variável



- Melhores Resultados
  - Algoritmo *k-medoids*

Agrupamento de Amostras - Resultados

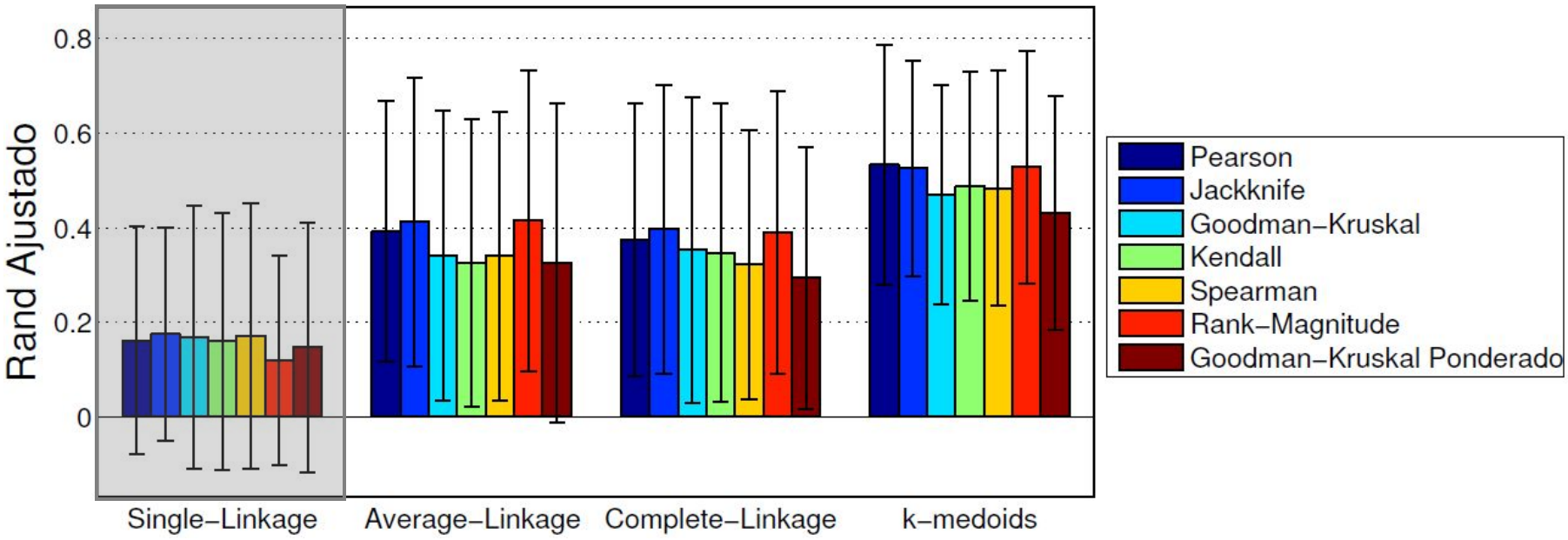
Número de grupos variável



- Resultados superiores ou competitivos
  - Pearson, Jackknife e Rank-Magnitude
- Medidas baseadas em *ranks*
  - Resultados similares

Agrupamento de Amostras - Resultados

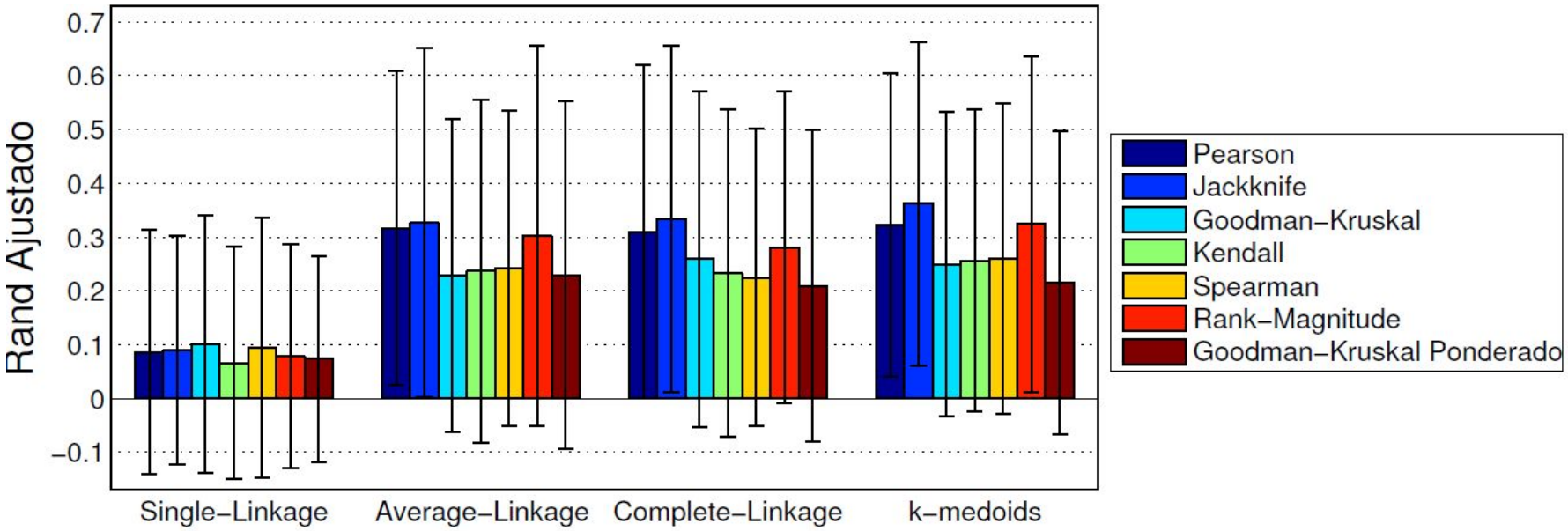
Número de grupos variável



- Valores de Rand Ajustado
  - Superiores aos obtidos com número de grupos fixo
- *Single-Linkage* apresentou os piores resultados

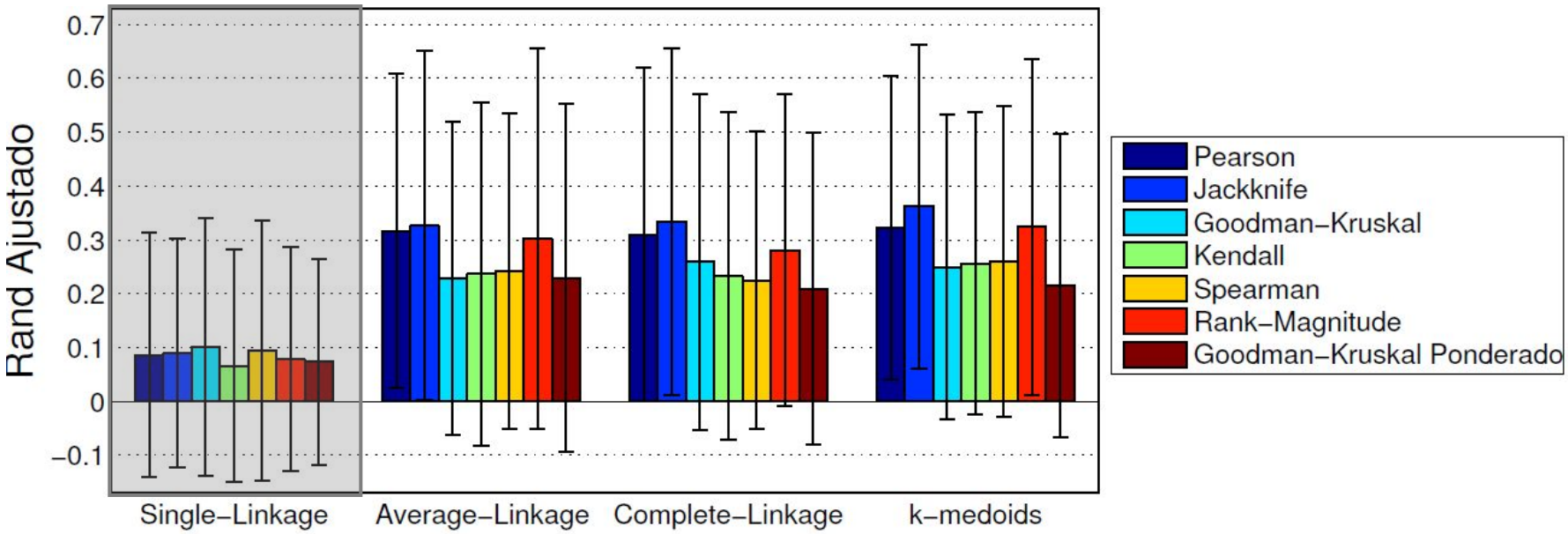
Agrupamento de Amostras - Resultados

Número de grupos estimado - Silhueta



Agrupamento de Amostras - Resultados

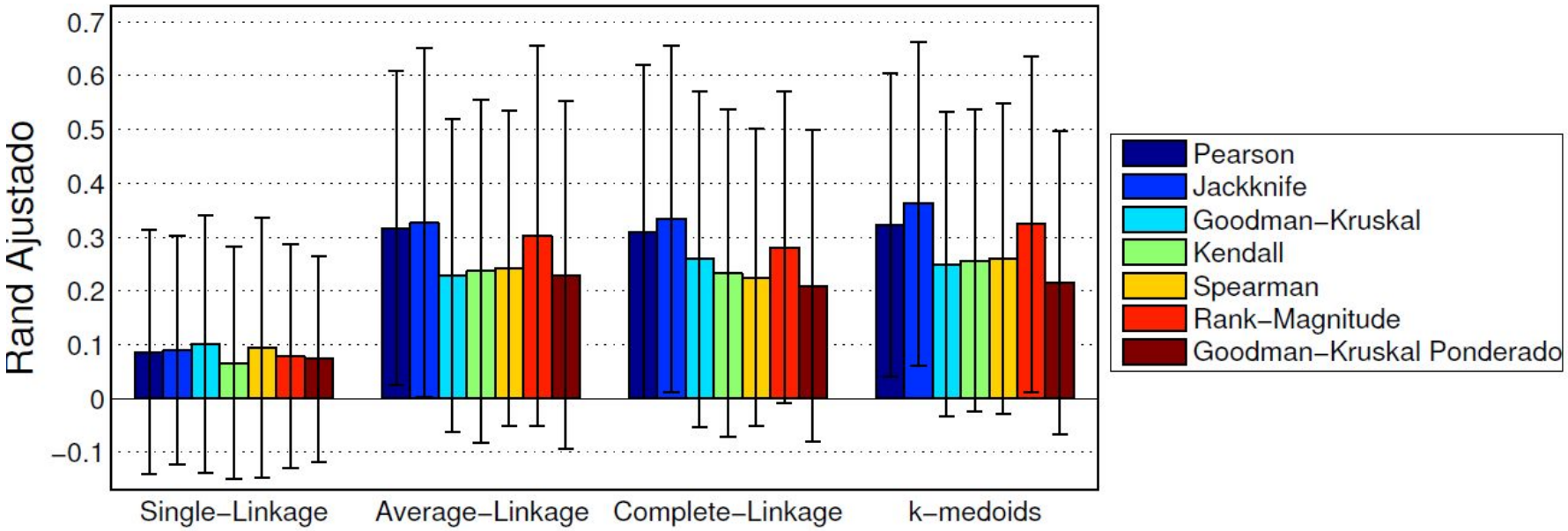
Número de grupos estimado - Silhueta



- Menores diferenças entre algoritmos de agrupamento
  - Piores resultados obtidos novamente com o *Single-Linkage*

Agrupamento de Amostras - Resultados

Número de grupos estimado - Silhueta



- Resultados superiores ou competitivos
  - Pearson, Jackknife e Rank-Magnitude

Agrupamento de Amostras - Resumo

- Jackknife, Pearson e Rank-Magintude
  - Melhores resultados em média
- *k-medoids*
  - Resultados superiores ou competitivos
- *Single-Linkage*
  - Piores resultados independentemente da medida utilizada
- Melhores medidas em média não são *sempre* a melhor opção
- Medidas baseadas em *ranks*
  - Goodman-Kruskal e Kendall são boas alternativas à Spearman

Resultados

Agrupamento de Genes

Metodologia de Avaliação

- Comparação realizada em 17 bases de dados
  - Bases de dados de séries temporais
  - Ausência de rótulos externos
- Para cada par algoritmo – correlação
  - Melhor partição eleita pelo critério da Silhueta
- Avaliação dos resultados
  - Heurística baseada na *Gene Ontology* (GO)
  - Análise de enriquecimento de grupos de genes
  - Comparação dos níveis de enriquecimento obtidos

Agrupamento de Genes - Resultados

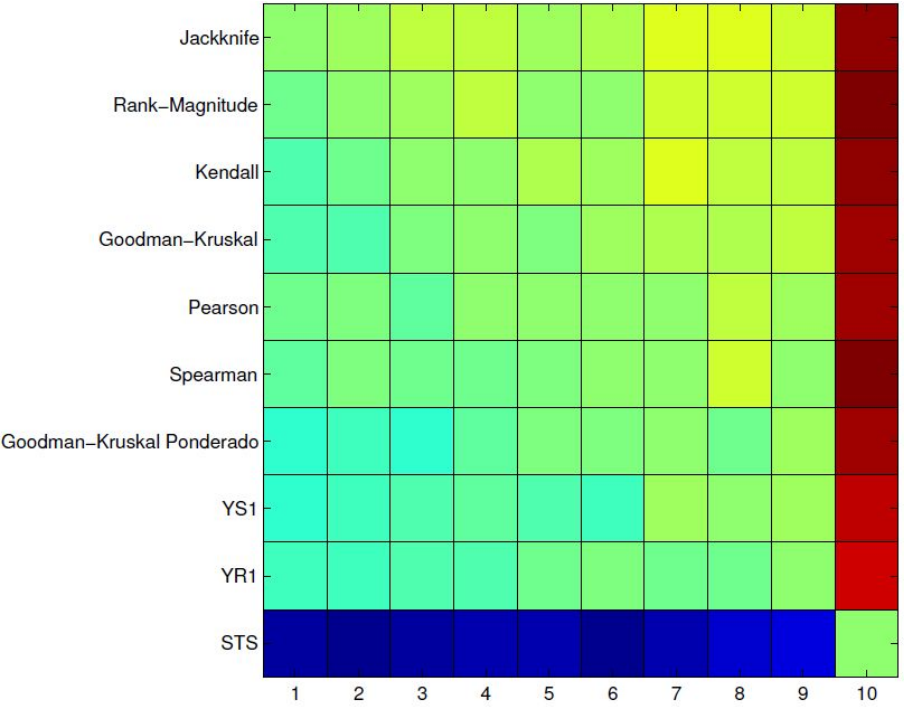
- Dentre os algoritmos utilizados
  - Pior desempenho observado com o *Single-Linkage*

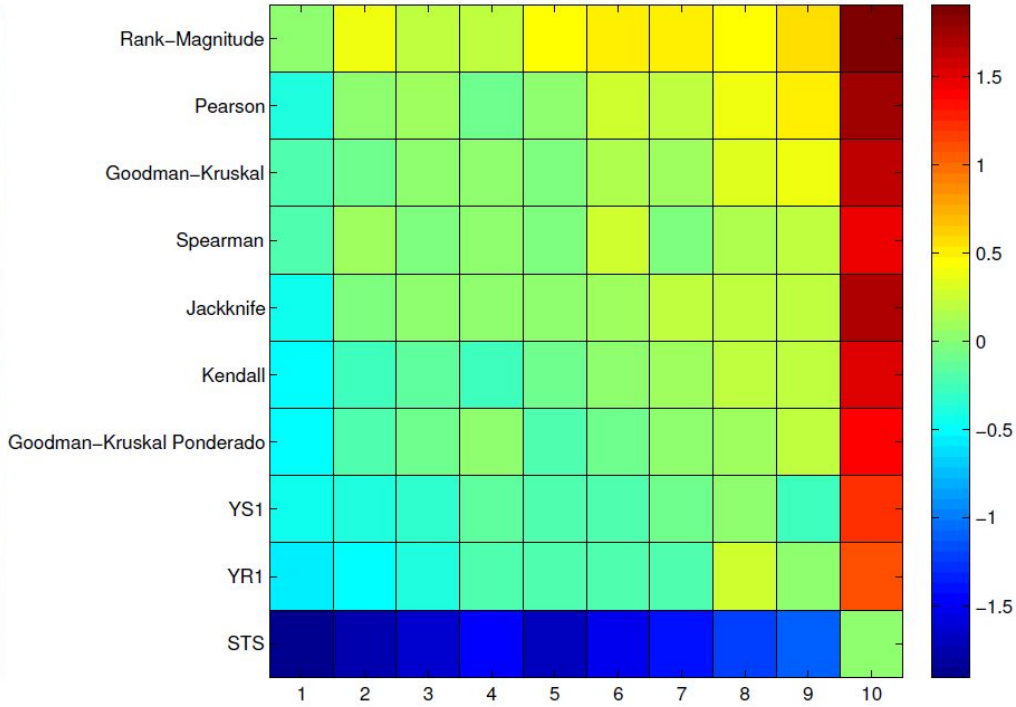
	Single	*Average*	*Complete*	*k-medoids*
Single	-	142/21/1537	135/8/1557	112/1/1587
Average	1537/21/142	-	955/16/729	831/12/857
Complete	1557/8/135	729/16/995	-	669/11/1020
k-medoids	1587/1/112	857/12/831	1020/11/669	-

Vitórias / Empates / Derrotas

Agrupamento de Genes - Resultados

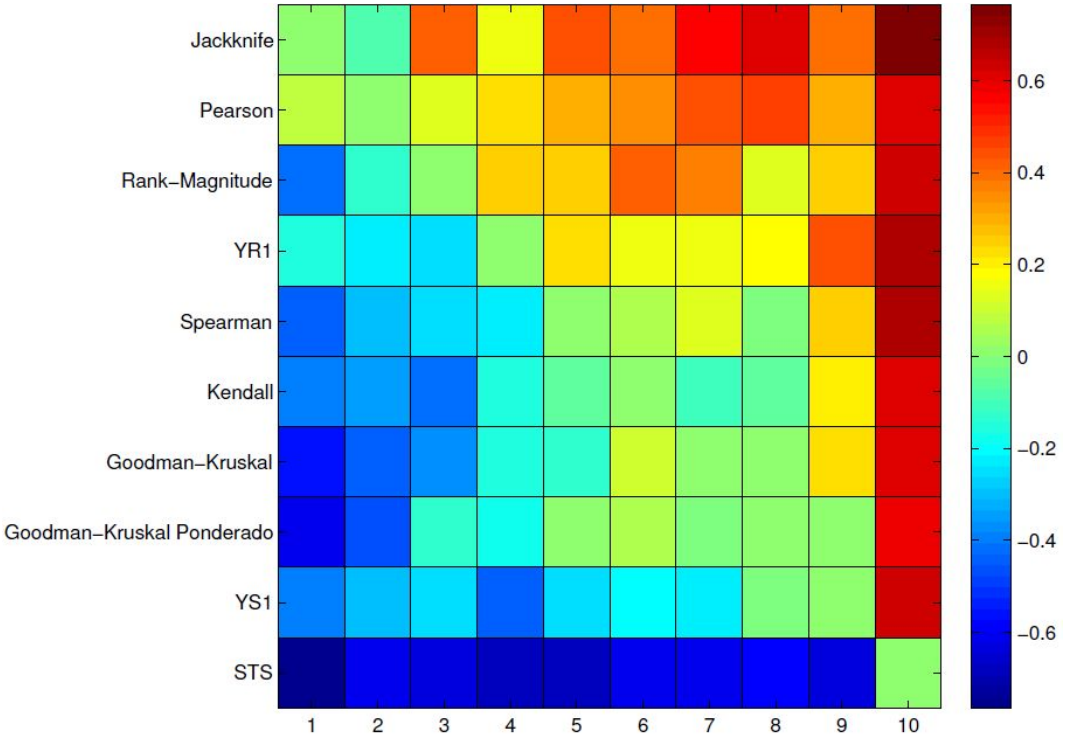
- Menores diferenças entre as medidas avaliadas
  - *Average-Linkage*
  - *Complete-Linkage*





Agrupamento de Genes - Resultados

Comparação entre os níveis de enriquecimento obtidos



- *k-medoids*
  - Melhores resultados: Jackknife, Pearson e Rank-Magnitude

Agrupamento de Genes - Resumo

- Independentemente do algoritmo utilizado
  - Melhores resultados - resultados competitivos
    - Jackknife
    - Pearson
    - Rank-Magnitude
  - Piores resultados
    - Medidas baseadas em correlação
      - YR1 e YS1
    - Dissimilaridade STS

Resultados

Seleção de Atributos e Classificação

Metodologia de Avaliação

- Seleção de atributos
  - *Simplified Silhouette Filter* (SSF)[1]
    - 1NN
    - Naïve Bayes
- Classificação
  - *k-Nearest Neighbors* (*k*NN)
    - 1NN

[1]Covões, T. F., Hruschka, E. R., Castro, L. N., e Santos, A. M. (2009). A cluster-based feature selection approach. In HAIS '09.

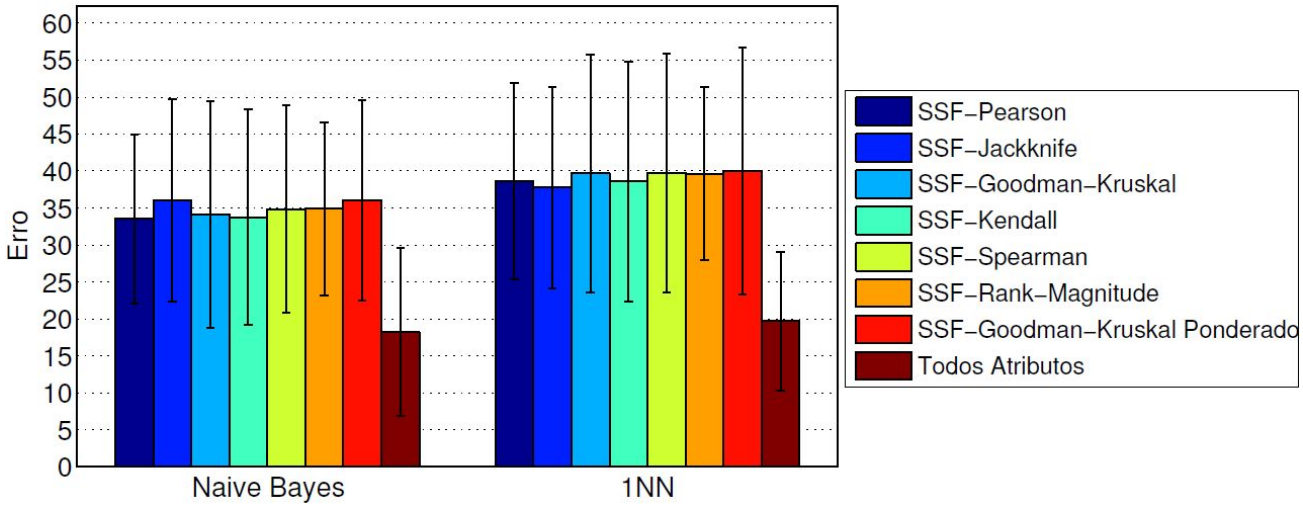
Metodologia de Avaliação

- Correlações comparadas quanto às acurácias obtidas
- Estimação de erro
  - Validação cruzada de 10 pastas estratificada
  - Seleção de atributos somente na pasta de treinamento
- 35 bases de *benchmark*[1] - câncer

[1] de Souto, M., Costa, I. G., de Araujo, D., Ludermir, T., e Schliep, A. (2008). Clustering cancer gene expression data: a comparative study. BMC Bioinformatics.

Seleção de Atributos - Resultados

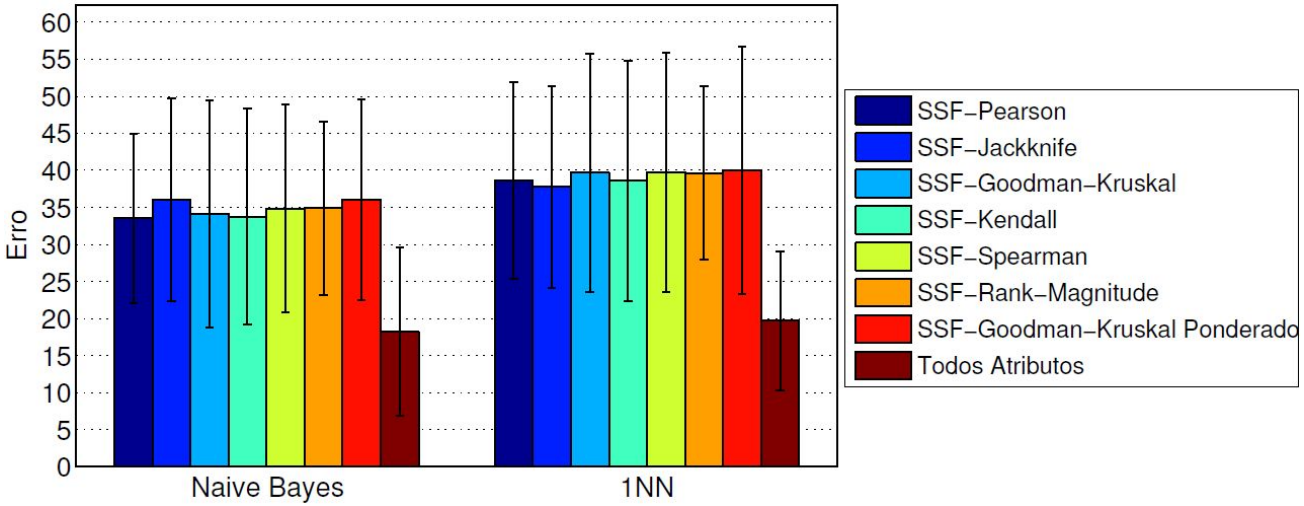
- Variantes SSF x Todos atributos
  - A utilização de todos atributos levou a menores erros
    - Para ambos: *k*NN e Naïve Bayes



  - Variantes utilizadas selecionaram poucos atributos
    - No geral por volta de 1% dos atributos ou menos

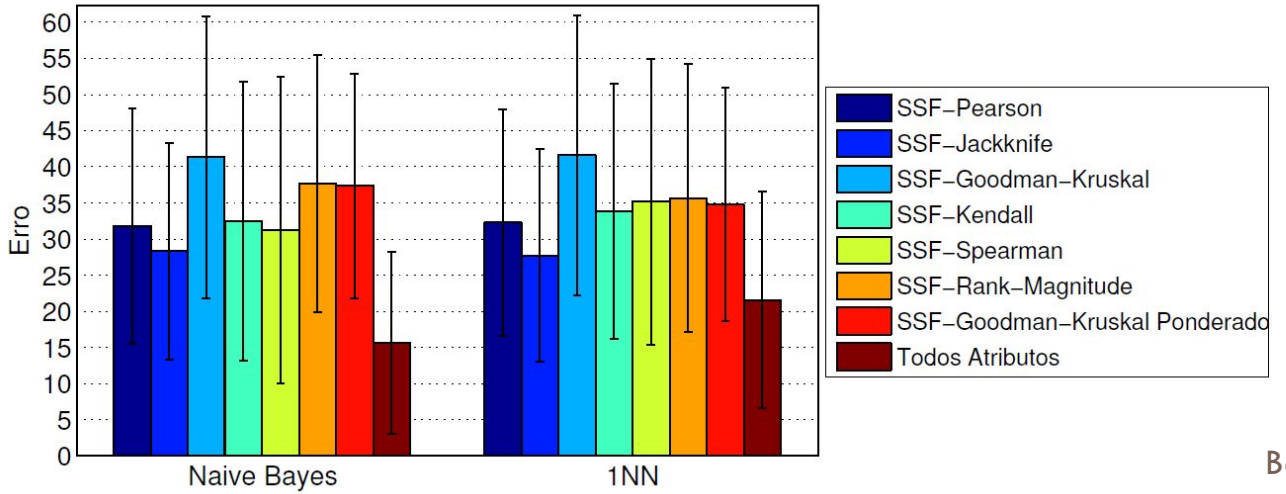
Seleção de Atributos - Resultados

- Comparação somente entre variantes
  - Bases cDNA
    - Apenas pequenas diferenças entre as medidas comparadas



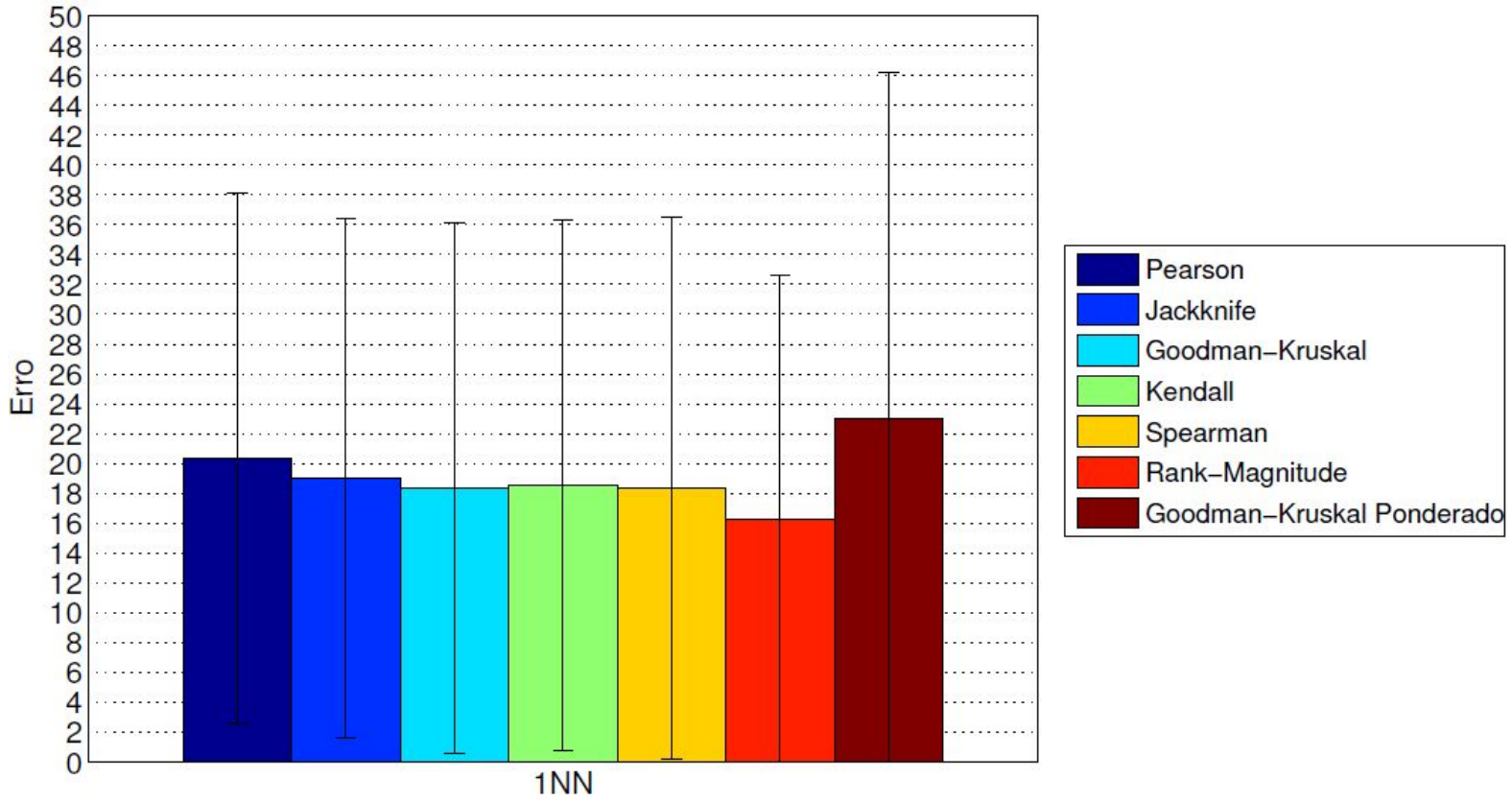
Seleção de Atributos - Resultados

- Bases Affymetrix
  - Menores erros produzidos com a correlação Jackknife
  - Jackknife levou a menores erros que Pearson
  - Entre as correlações baseadas em *ranks*, destacaram-se
    - Kendall
    - Spearman



Classificação - Resultados

- Pequenas diferenças entre as medidas comparadas



Bases Affymetrix

Classificação - Resumo

- Resultados em média
  - Pequenas diferenças entre correlações comparadas
- Resultados individuais
  - Diferenças de até 20% nos valores de erros observados
- Em bases Affymetrix
  - Testes estatísticos indicam diferença favorável para a correlação Rank-Magnitude frente à Pearson e Goodman-Kruskal Ponderado

Conclusões e Contribuições

Conclusões

- Agrupamento
  - Pearson, Jackknife e Rank-Magnitude foram, no geral, superiores às demais medidas comparadas
  - Rank-Magnitude torna-se uma alternativa interessante
  - Correlações baseadas em *ranks*
    - Goodman-Kruskal e Kendall – resultados competitivos aos apresentados pela correlação de Spearman
  - Agrupamento de genes
    - Medidas específicas produziram piores resultados

Conclusões

- Seleção de atributos
  - Bases Affymetrix
    - Melhores resultados obtidos com a correlação Jackknife
- Classificação
  - Pequenas diferenças entre as medidas, na média

Conclusões

- Não houve medida superior em todos os cenários
- Avaliação prévia de um conjunto de correlações mostra-se uma estratégia mais apropriada
- Quando uma avaliação preliminar mostra-se inviável
  - Pearson, Jackknife e Rank-Magnitude
- Medidas baseadas em *ranks* pouco utilizadas mostraram resultados competitivos aos da correlação de Spearman

Contribuições

- Estudo de diferentes coeficientes de correlação
- Comparação das medidas em diferentes cenários
- Revisão bibliográfica
  - Fundamentação biológica
  - Tecnologia de *microarray*
  - Análise de dados de expressão gênica

Publicações

- Até o presente momento
  - Jaskowiak, P. A., Campello, R. J. G. B., Covões, T. F., Hruschka, E. R. (2010). A Comparative Study on the Use of Correlation Coefficients for Redundant Feature Elimination. Em 11th Brazilian Symposium on Neural Networks – **SBRN 2010**, páginas 13-18.

- Em elaboração
  - Comparing Correlation Coefficients as Proximity Measures for Cancer Classification in Gene Expression Data.
  - Proximity Measures for Clustering Gene Expression Time-Series: A Comparative Study.
  - Comparing Correlation Coefficients as Proximity Measures for Clustering Gene Expression Profiles of Cancer.

Agradecimentos

CNPq – Apoio Financeiro

ICMC – Recursos Físicos

Instituto de Ciências Matemáticas e de Computação - USP
Avenida Trabalhador são-carlense, 400 - Centro
CEP: 13566-590 - São Carlos - SP

Medidas de Proximidade

- Pearson

$$\rho(\mathbf{a},\mathbf{b}) = \frac{\sum_{i=1}^{p}(a_i-\bar{a})(b_i-\bar{b})}{\sqrt{\sum_{i=1}^{p}(a_i-\bar{a})^2}\sqrt{\sum_{i=1}^{p}(b_i-\bar{b})^2}}$$

- Kendall

$$\tau(\mathbf{a},\mathbf{b}) = \frac{S_+ - S_-}{p(p-1)/2}$$

- Goodman-Kruskal

$$\gamma(\mathbf{a},\mathbf{b}) = \frac{S_+ - S_-}{S_+ + S_-}$$

Medidas de Proximidade

- Goodman-Kruskal Ponderado

$$\hat{\gamma}(\mathbf{a},\mathbf{b}) = \frac{\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\hat{w}_{ij}}{\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}|w_{ij}|}$$

$$\hat{w}_{ij} = \begin{cases} min\{\hat{w}_{ij}^{\mathbf{a}}/\hat{w}_{ij}^{\mathbf{b}}, \hat{w}_{ij}^{\mathbf{b}}/\hat{w}_{ij}^{\mathbf{a}}\} & \hat{w}_{ij}^{\mathbf{a}}\ \hat{w}_{ij}^{\mathbf{b}} > 0 \\ max\{\hat{w}_{ij}^{\mathbf{a}}/\hat{w}_{ij}^{\mathbf{b}}, \hat{w}_{ij}^{\mathbf{b}}/\hat{w}_{ij}^{\mathbf{a}}\} & \hat{w}_{ij}^{\mathbf{a}}\ \hat{w}_{ij}^{\mathbf{b}} < 0 \\ 1 & \hat{w}_{ij}^{\mathbf{a}} = \hat{w}_{ij}^{\mathbf{b}} = 0 \\ 0 & \text{demais casos} \end{cases}$$

$$w_{ij} = \begin{cases} w_{ij}^{\mathbf{a}}/w_{ij}^{\mathbf{b}} & \text{se } w_{ij}^{\mathbf{b}} \neq 0 \\ 1 & \text{se } w_{ij}^{\mathbf{a}} = 0 \text{ e } w_{ij}^{\mathbf{b}} = 0 \\ 0 & \text{demais casos} \end{cases}$$

$$\hat{w}_{ij}^{\mathbf{a}} = \begin{cases} \frac{a_i - a_j}{a_{max} - a_{min}} & \text{se } a_{max} \neq a_{min} \\ 0 & \text{outro caso} \end{cases}$$

$$w_{ij}^{\mathbf{a}} = sign(a_i - a_j)$$

$$\hat{w}_{ij}^{\mathbf{b}} = \begin{cases} \frac{b_i - b_j}{b_{max} - b_{min}} & \text{se } b_{max} \neq b_{min} \\ 0 & \text{outro caso} \end{cases}$$

$$w_{ij}^{\mathbf{b}} = sign(b_i - b_j)$$

Medidas de Proximidade

- Rank-Magnitude

$$\hat{r}(\mathbf{a},\mathbf{b}) = \frac{2\sum_{i=1}^{p} R(a_i)b_i - r_m^{max} - r_m^{min}}{r_m^{max} - r_m^{min}}$$

$$r_m^{min}(\mathbf{a},\mathbf{b}) = \sum_{i=1}^{p}(n+1-i)\bar{b}_i$$

$$r_m^{max} = \sum_{i=1}^{p} i\bar{b}_i$$

$$r(\mathbf{a},\mathbf{b}) = \frac{\hat{r}(\mathbf{a},\mathbf{b}) + \hat{r}(\mathbf{b},\mathbf{a})}{2}$$

Medidas de Proximidade

- Jackknife

$$\varrho(\mathbf{a},\mathbf{b}) = \min\{\rho^1(\mathbf{a},\mathbf{b}), \rho^2(\mathbf{a},\mathbf{b}), \rho^3(\mathbf{a},\mathbf{b}), \ldots, \rho^p(\mathbf{a},\mathbf{b}), \rho(\mathbf{a},\mathbf{b})\}$$

Medidas de Proximidade

- Son e Baek

$$YR1(\mathbf{a},\mathbf{b}) = \omega_1 R(\mathbf{a},\mathbf{b}) + \omega_2 A(\mathbf{a},\mathbf{b}) + \omega_3 M(\mathbf{a},\mathbf{b})$$

$$YS1(\mathbf{a},\mathbf{b}) = \omega_1 S(\mathbf{a},\mathbf{b}) + \omega_2 A(\mathbf{a},\mathbf{b}) + \omega_3 M(\mathbf{a},\mathbf{b})$$

$$inclinação(\mathbf{x},i) = \frac{x_{i+1} - x_i}{t_{i+1} - t_i}$$

$$L(\mathbf{x},i) = \begin{cases} 1, & \text{se } inclinação(\mathbf{x},i) > 0 \\ -1, & \text{se } inclinação(\mathbf{x},i) < 0 \\ 0, & \text{se } inclinação(\mathbf{x},i) = 0 \end{cases}$$

$$A(\mathbf{a},\mathbf{b}) = \sum_{i=1}^{n-1} \frac{I(L(\mathbf{a},i) = L(\mathbf{b},i))}{n-1}$$

$$M(\mathbf{a},\mathbf{b}) = \begin{cases} 1 & \text{se } t_{\mathbf{a}}^{min} = t_{\mathbf{b}}^{min} \text{ e } t_{\mathbf{a}}^{max} = t_{\mathbf{b}}^{max} \\ 0.5 & \text{se } t_{\mathbf{a}}^{min} = t_{\mathbf{b}}^{min} \text{ ou } t_{\mathbf{a}}^{max} = t_{\mathbf{b}}^{max} \\ 0 & \text{se } t_{\mathbf{a}}^{min} \neq t_{\mathbf{b}}^{min} \text{ e } t_{\mathbf{a}}^{max} \neq t_{\mathbf{b}}^{max} \end{cases}$$

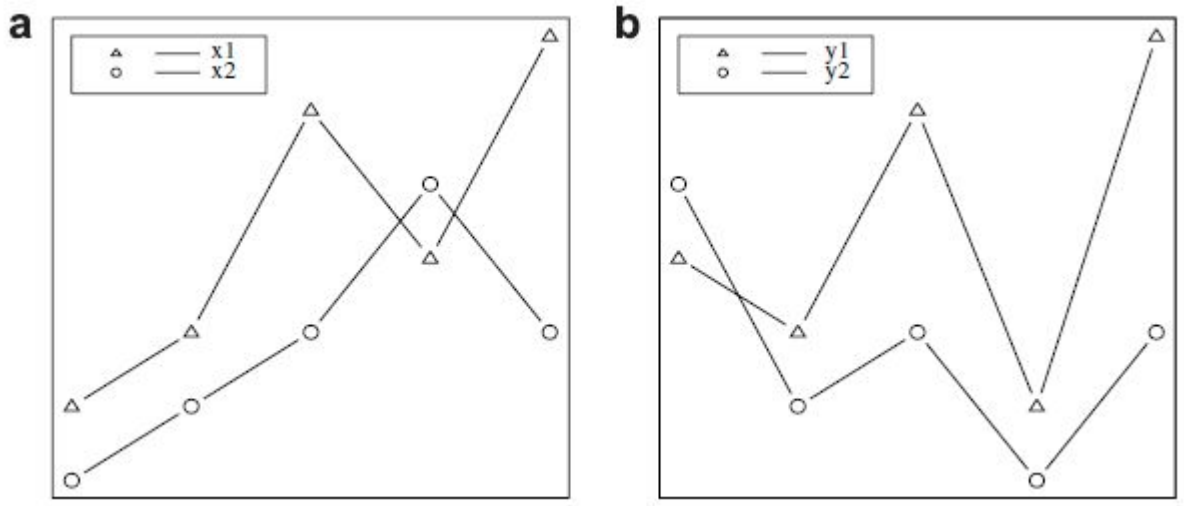
Medidas de Proximidade

- Dissimilaridade *Short Time-Series*

$$STS(\mathbf{a},\mathbf{b}) = \sqrt{\sum_{i=1}^{p-1}\left(\frac{b_{i+1}-b_i}{t_{i+1}-t_i} - \frac{a_{i+1}-a_i}{t_{i+1}-t_i}\right)^2}$$

Medidas de Proximidade

- Son e Baek - Motivação



Permutação entre posições 1 e 4