Inferência para redes de proteínas através de dados genômicos múltiplos: uma abordagem supervisionada

Carla Monteiro

Based on paper
Protein network inference from multiple genomic data: a supervised approach
by Yamanishi et al., 2004

- Motivação
- Problema
- Redes de Proteínas
- Tipos de dados genômicos
- Representação dos dados via Kernel
- Abordagens não supervisionadas
- Abordagem supervisionada
- ROC curves
- Exemplo: S.cerevisiae

❖ Motivação

Hipótese:

A maioria das funções biológicas envolve as interações entre várias proteínas, o que aumenta a complexidade de sistemas vivos.

❖ Problema

Inferir redes completas de proteínas de um organismo.

❖ Redes de proteínas

É um grafo tal que:

Vértices → proteínas

Arestas (*edge*)→ relações binárias entre as proteínas (interações)

▼Há três tipos de relações binárias:

(i) as proteínas interagem fisicamente;

(ii)as proteínas são enzimas que catalizam duas reações químicas sucessivas em uma sequência (*pathway*) e

(iii) uma das proteínas regula a expressão da outra.

(Isto deve ser levado em consideração no estudo do comportamento do sistema biológico)

❖Tipos de dados genômicos

▼ *Protein network data(gold standard)*
(Kanehisa et al., 2004)

Uma parte da rede de proteínas (como definida anteriormente).

Ex: S. cerevisiae.

Conjunto de dados: KEGG/PATHWAY.

Vértices: 769
Arestas: 3702

(Considerada uma parte confiável da rede global de proteínas a ser inferida)

▼ Expression data

Os dados de expressão aqui correspondem a 157 experimentos:

77experimentos (Spellman et al.,1998) e
80 experimentos(Eisen et al.,(1998).

Cada proteína é associada a um vetor de dimensão 157.

▼ Protein interaction data
(Ito et al.,2001; Uetz et al., 2000)

5470 pares de interações de proteínas obtidas de vários experimentos yeast two-hybrid (Y2H)

(Introduz muitos falso-positivos)

▼ Localization data
(Huh et al.,2003)

Este conjunto de dados descreve as localizações das proteínas em 23 posições intracelular (mitochondrion, Golgi e nucleus).

Para cada proteína é designada um string de 23 bits codificado como segue:

1 se a proteína aparece em uma dada posição intracelular e

0 se a proteína não aparece em uma dada posição intracelular

▼ Phylogenetic profile

Foram construídos através de *ortholog clusters in KEEG dataset*, os quais descrevem os conjuntos orthologous de proteínas em 145 organismos.

(11 eukaryotes, 16 archaea e 118 bacterias)

Para cada phylogenetic profile é designado um string de 145 bits codificado como segue:

1 se a orthologous proteína esta presente

0 se a orthologous proteína não esta presente

(Orthologous proteinas apresentam a mesma função em organismos diferentes)

❖ Representação dos dados via Kernel

Transformação dos diferentes conjuntos dados em uma função positiva definida, chamada *Kernel*.

A função, $K(x,y)$, de Kernel é definida t.q.

$K(x_i, x_j) = K(x_j, x_i)$ para qualquer duas proteínas e

$\sum_{i=1}^{n} a_i a_j K(x_i, x_j) \geq 0$ para n inteiro ;

$(x_1, \ldots, x_n)$ um conjunto de proteínas e
$(a_1, \ldots, a_n)$ um conjunto de números reais

Kernel → "medida de similaridade"

Exemplos de datasets :

1. vetores para cada proteína

(I) expression data;
(II) localization data e
(III) phylogenetic profiles

a)Gaussian RBF kernel:

$$K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma^2)$$

b)Linear kernel:

$$K(x_i, x_j) = x_i . x_j$$

K matriz NXN com elementos $K(x_i, x_j)$
N - número de proteínas

2. Grafo de proteínas

(I) gold standard protein network
(II) noisy protein interations data

a)diffusion kernel

$$K = \exp(\beta H)$$, onde

$\beta > 0$ é o parâmetro e

$$H_{i,j} = \begin{cases} 1 & para \quad i \sim j, \\ -d_i & para \quad i = j, \\ 0 & caso\, contrario \end{cases}$$

são os elementos da matriz H Laplaciana (oposta) do grafo.

$i \sim j$ significa que i-ésima e j-ésima proteínas são ligadas por uma aresta

d_i é o número de proteínas ligadas a proteína i

▼ Motivação para o uso de Kernel

(I) Todos os tipos de conjuntos de dados (vetor, string grafo) são codificados de mesma forma (Matriz).

(II) Permite utilizar métodos de Kernel.

▼ Conjuntos de dados integrados

Os conjuntos de dados apresentados para predizer a rede global de proteínas foram representados pelos seus Kernels $\boldsymbol{K}$.

Considere P conjuntos de dados, então $\boldsymbol{K}_1$,..., $\boldsymbol{K}_P$ kernels representam as similaridades das proteínas com respeito ao p-ésimo conjunto.

Estes conjuntos de dados podem ser combinados formando um novo kernel através da soma de todos os kernels (*data integration*).

$$K = \sum_{p=1}^{P} K_p$$

(Kernel permite combinar dados heterogêneos)

❖Abordagens não supervisionadas para predição de rede de proteínas

Esta abordagem não agrega informações a priori da rede global de proteínas a ser inferida.

▼ Abordagem direta

Considerando o problema de predizer a rede global de proteínas da S.cerevisiae dos conjuntos de dados apresentados.

Um método direto de inferência, sob a suposição que proteínas conectadas são mais similares nestes conjunto é apresentado é apresentado a seguir.

As arestas do grafo é inferida quando o valor de $K(x_i, x_j)$ é grande.

Dependendo da escolha deste valor(K), irá cobrir as situações onde as proteínas selecionadas
terão:
- expressões correlacionadas;
- perfios similares;
- localizações similares ou
- todas simultaneamente

Para uma escolha K fixa, a rede global de proteína pode ser construída progressivamente, começando de vértices isolados adicionando arestas entre pares de proteínas com valores de kernel decrescente.

▼Abordagem spectral clustering

Considere ainda o problema de predizer a rede global de proteínas da S.cerevisiae dos conjuntos de dados apresentados.

Quando o interesse no conjunto de proteínas é agrupar em clusters, a idéia do spectral clustering(SC) é mapear este conjunto dentro de um *feature space* onde os clusters são mais fáceis de serem detectados antes de aplicar a teoria clássica de análise de clusters.

O *feature space* é definido como o espaço gerado dos primeiros autovetores da matriz de similariedades entre as proteínas(K).

Uma alternativa para analisar o spectral clustering é através de *Kernel principal component analysis*(KPCA), pois este está relacionado ao spectral clustering como segue:

Em KPCA o espaço gerado pelas primeiras componentes principais - principal components(PCs) - é um espaço fácil de detectar clusters como o feature space em SC.

A KPCA pode ser resumida como segue.

Seja N proteínas $\chi = \{x_1, \ldots, x_N\}$,

e uma função de kernel $K: \chi^2 \to \Re$

Então, temos o conjunto de funções

$$H = \{f(x) = \sum_{i=1}^{N} \alpha_i K(x_i, x), (\alpha_1, \ldots, \alpha_N) \in \Re^N\}$$

provado que a norma é

$$\|f\|_H = \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j)$$

A projeção sobre a primeira direção principal é definida, a menos de um escalar, como a função

$f^1 \in H$ que minimiza $\|f^1\|_H$ sob a restrição

$$\sum_{i=1}^{N} f^1(x_i)^2 = 1$$

As projeções sobre as seguintes direções principais são definidas recursivamente do mesmo modo, mas com a adicional restrição de ortogonalidade:

$$\sum_{i=1}^{N} f^{(l)}(x_i) f^{(m)}(x_i) = 0 \quad \text{se } l < m$$

Quando $K(x_i, x_j) = x_i, .x_j$ *para vetores* $\rightarrow$ *PCA*

Para utilizar a abordagem de Spectral Clustering represente a proteína x_i *pelo vetor*

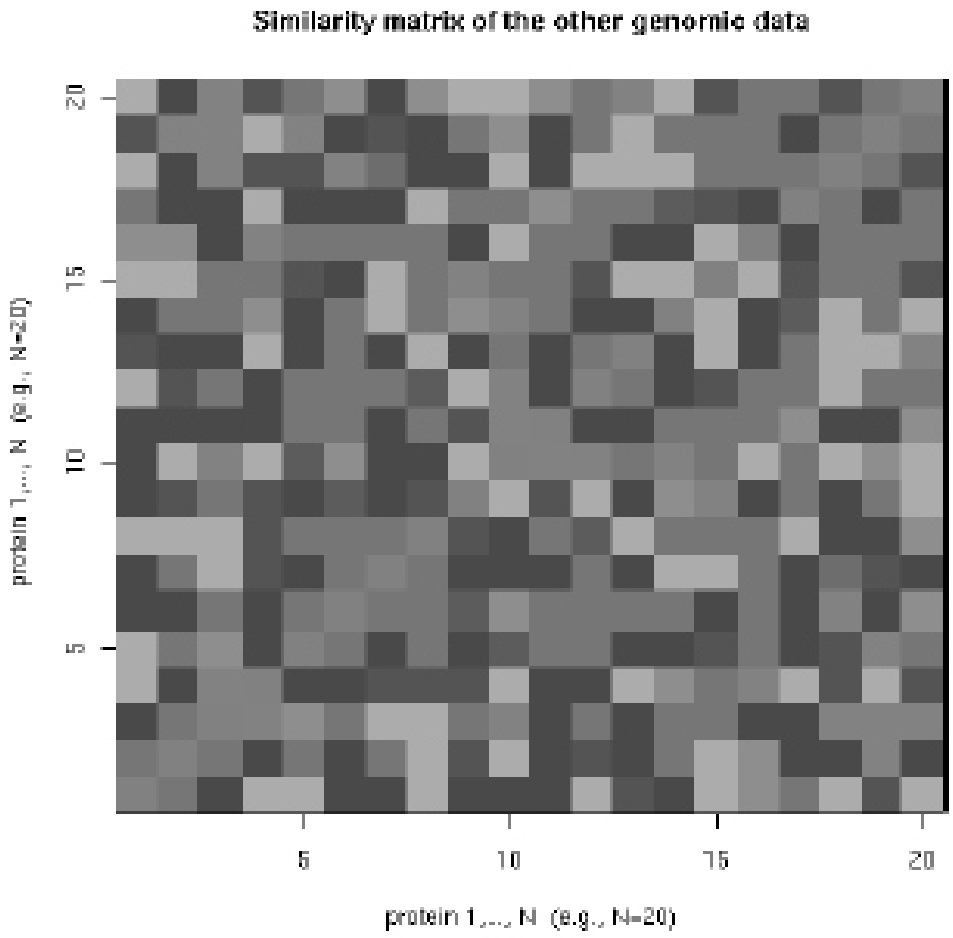
$$[f^{(1)}(x_i),\ldots,f^{(L)}(x_i)]^T \quad L < N$$

L - referente aos L primeiros CPs
N - Número de proteínas
e analise os dados pelo método clássico de Clustering baseado nessas representações.

Resumo:
Projete todas as proteínas no subespaço(feature space) definido pelos primeiros PC's obtidos por KPCA e então selecione pares de similar proteínas neste subespaço, como feito na abordagem direta.

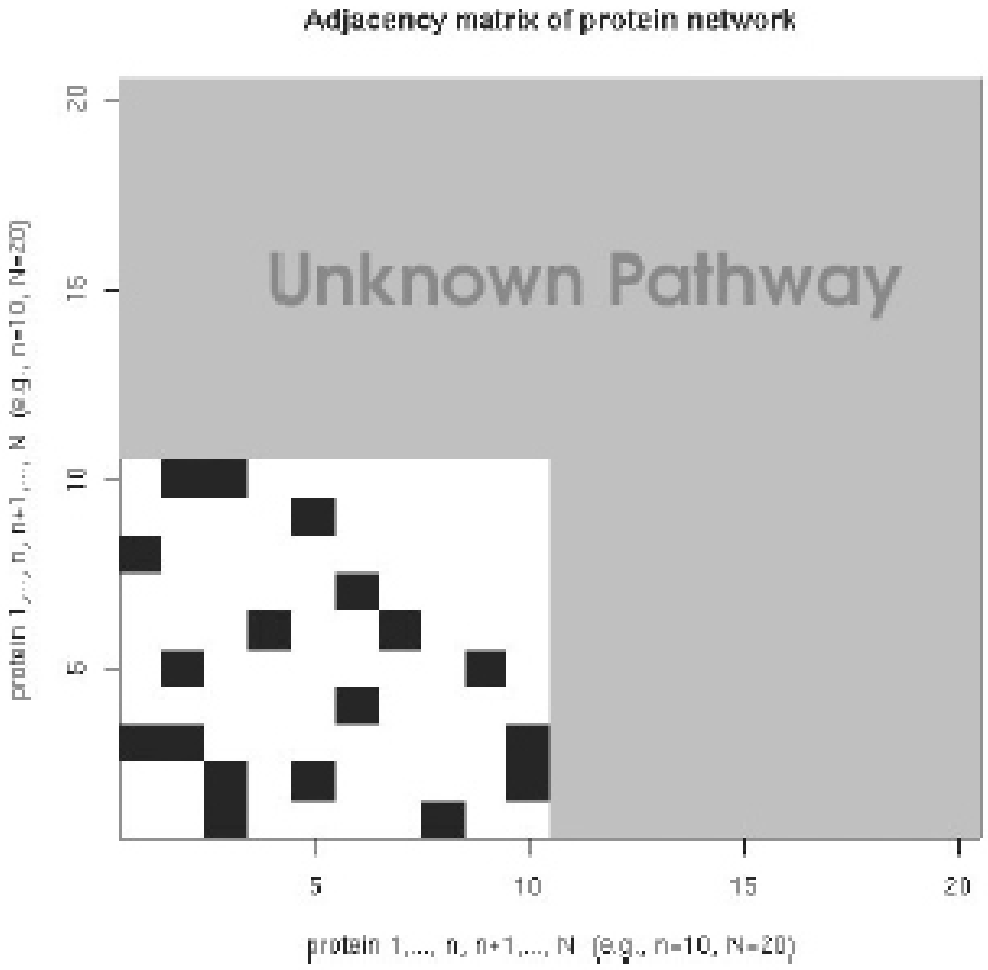
❖Abordagem supervisionada para predição de rede de proteínas

Figura 1



A figura 1 representa um conjunto de dados de N proteínas (very noisy) que pretendemos inferir para a rede global de proteína.

Figura 2



A figura 2 representa um conjunto de dados de n proteínas que com certa confiança faz parte da rede de proteínas a ser inferida. n<N (Conhecimento a priori)

As abordagens diretas e spectral não são supervisionadas, no sentido que estas não usam informação a priori como no figura 2, mas inferem a rede de proteínas a partir dos dados na figura 1.

A abordagem supervisionada irá inferir a rede de proteínas baseada nos dados das duas figuras.

O método supervisionado a ser apresentado considera uma pequena modificação do método spectral clustering (não supervisionado)

Na abordagem spectral, cada x é primeiro representado por um vetor

$$f(x) = [f^{(1)}(x), \ldots, f^{(L)}(x)]^T \quad L < N$$

onde $f^{(l)}(x)$ é a projeção de x sobre o l-ésimo componente principal.

O objetivo desta projeção é definir um *feature space* onde pares de proteínas interagindo tem projeção similares. O que torna possível inferir as interações de similariedade neste espaço.

Se x_i interage com x_j então gostaríamos que $f(x_i)$ fosse similar a $f(x_j)$ e assim $f^{(l)}(x_i)$ fosse similar a $f^{(l)}(x_j)$ $l = 1,\ldots,L$.

As projeções no *feature space ideal* não podem ser computados uma vez que não conhecemos a rede completa das proteínas.

Para melhorar a representação na abordagem spectral, foi proposto restringir o feature space ideal a um *feature space ideal ajustado*(pelo menos na parte conhecida a priori)

Sejam

$\{x_1,\ldots,x_n\}$ as n proteínas em “gold standard”
e
$\{x_{n+1},\ldots,x_N\}$ as proteínas a serem inferidas na rede

como mostrado na figura 2.

Seja K_1 o kernel relativo as n proteínas (as conhecidas a priori em “gold standard”) em um particular conjunto de dados.

Seja K_2 o diffusion kernel relativo a rede de proteína conhecida a priori (n proteínas).

K_1 e K_2 são matrizes nxn.

Para qualquer função f definida sobre $\{x_1,\ldots,x_n\}$ defina $\|f_1\|$ e $\|f_2\|$ como normas de K_1 e K_2.

É de interesse encontrar uma função f (feature) tal que

$\|f_1\|$ seja pequeno como na abord. spectral

e

$\|f_2\|$ seja pequeno simultaneamente como numa representação ideal.

Para garantir o que foi definido, foi proposto seguinte:

ache duas funções f_1 e f_2 tais que

$$\sum_{i=1}^{N} f_k(x_i)^2 = 1 \quad \text{para k = 1 , 2}$$

e que maximiza a função (1)

$$corr(f_1, f_2) \times \frac{1}{\sqrt{1+\lambda_1 \|f_1\|^2}} \times \frac{1}{\sqrt{1+\lambda_2 \|f_2\|^2}}$$

onde λ_1 , $\lambda_2 > 0$ são parâmetros e
$corr(f_1 , f_2)$ *é a correlação entre* f_1 e f_2

Subseqüêntes features podem ser definidos recursivamente por minimizar a função (1) com restrições adicionais de ortogonalidade.

A razão principal de usar a função (1) é que pode ser mostrado que resolver o problema utilizando-a, é equivalente a resolver o problema de autovalor generalizado como segue (2):

$$\begin{pmatrix} 0 & K_1K_2 \\ K_1K_2 & 0 \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

$$= \rho\begin{pmatrix} (K_1+\lambda_1 I)^2 & 0 \\ 0 & (K_2+\lambda_2 I)^2 \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

Sucessivas soluções da equação (1) pode ser escrita como

$$f_1 = K_1\alpha_1 \text{ e } f_2 = K_2\alpha_2$$

onde α_1 e α_2 são autovetores da equação (2) com autovalores ρ decrescente.

Este problema é chamado de
Kernel canonical correlation analysis(KCCA)

▼*Kernel canonical correlation analysis (KCCA)*

Sejam $(\alpha_1^{(1)},\ldots,\alpha_1^{(L)})$ as primeiras L soluções da equação (2)- obtidas através de valores decrescentes de ρ.

Então os L features de interesse do K_1 obtidos são

$f^{(l)} = K_1\alpha_1^{(l)}$ para $l = 1,\ldots, L$

(*é esperado que eles se ajustem aos feature ideais no gold standard dataset*)

Estes features podem agora ser generalizados para qualquer proteína x como segue:

$$f^{(l)}(x) = \sum_{k=1}^{n} \alpha_1^{(l)}(x_k) K(x_k, x) \quad (3)$$

O conjunto destes features serão então uilizados para inferir sobre as interações de proteínas.

Tanto no método spectral quanto no supervisionado KCCA, cada proteína x é mapeada para um feature space (vetor de dimensão *L*):

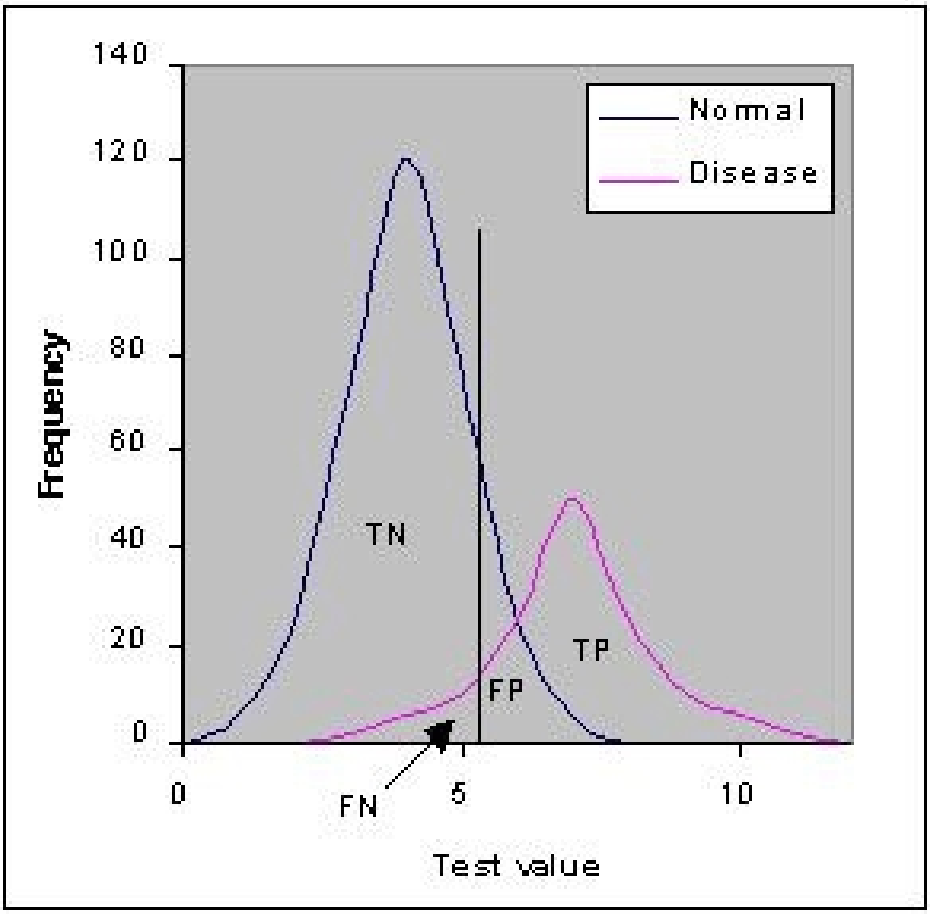
$$u = (u_1, \ldots, u_L)^T = [f^1(x), \ldots, f^L(x)]^T$$

Para avaliar as similariedades das proteínas x e y neste feature space, quantifica-se as similariedades entre os pontos u e v através da correlação :

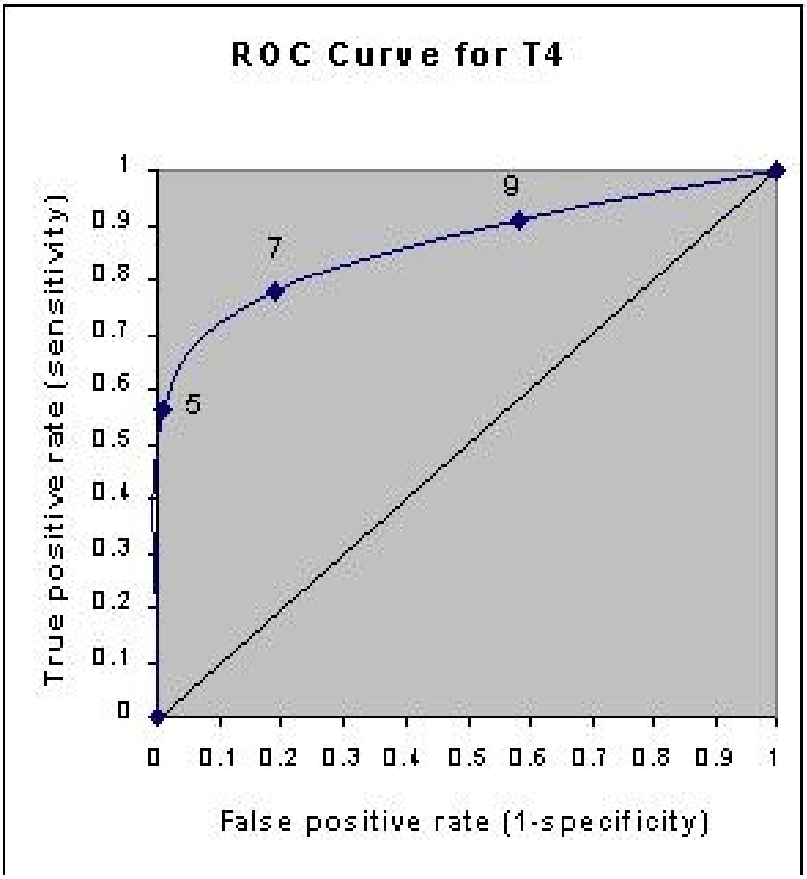
$$corr(u,v) = \frac{\hat{\text{cov}}(u,v)}{\sqrt{\hat{\text{var}}(u)\,\hat{\text{var}}(v)}}$$

❖ Receiver Operating Characteristic curve - ROC curve

Exemplo: Teste de Diagnóstico



A ROC curve é um gráfico da proporção de verdadeiro-positivo contra falso-positivo para possíveis diferentes cutpoints de um teste(diagnóstico).



Este gráfico mostra o seguinte:
(I) Quanto mais a curva segue a borda esquerda e está perto do topo, mais preciso o teste.
(II) Quanto mais perto a curva está da diagonal(45 graus), menos preciso é o teste.
(III) A área abaixo da curva é a medida de precisão do teste.

❖ Exemplo: S.cerevisiae

Todos os conjuntos de dados são transformados em Kernels como segue:

1) . the gold standard protein network- K_{gold}

. the noisy protein interaction- K_{ppi}

(*difussion kernel*, β =*1*)

2) . gene expression data- K_{exp}

(*Gaussian RBF kernel*, $\sigma^2 = 5$)

3) . location data- K_{loc}

. phylogenetic profile- K_{phy}

(*simple linear kernel*)

(Todos os kernels são normalizados , isto é, 1 na diagonal e centrado no feature space.)

☹ Os métodos direto e spectral foram testado para cada conjunto de dados, representado por kernel, e também para o conjunto integrado, representado pela soma dos kernels.

Para o método spectral, fixou-se *L= 50* para definir o feature space.

A precisão de ambos métodos foi verificada através do "*gold standard dataset* ".

Cada método construiu-se uma rede de proteínas, começando de isolados vétices(proteínas) e então adicionando arestas entre pares de proteínas ordenadas pelas similariedades.

Em cada adição, foi codificado o número de :

verdadeiro-positivos - arestas preditas corretamente, pois estavam presentes no gold standard dataset e

falso-negativos - arestas preditas erroneamente, pois não estavam presentes no gold standard dataset.

Roc Curvas foram construídas para os dois métodos.

Figura 3

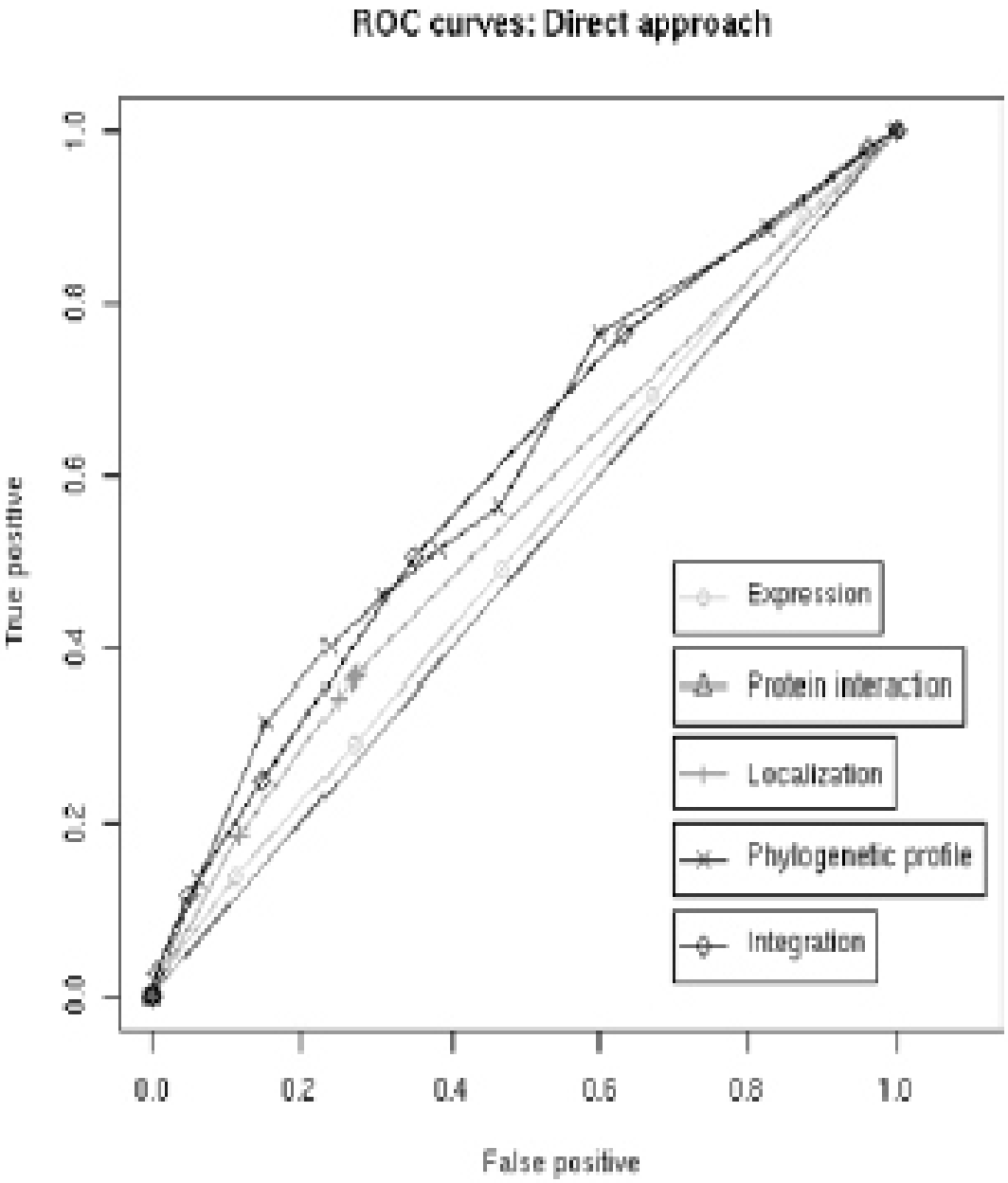
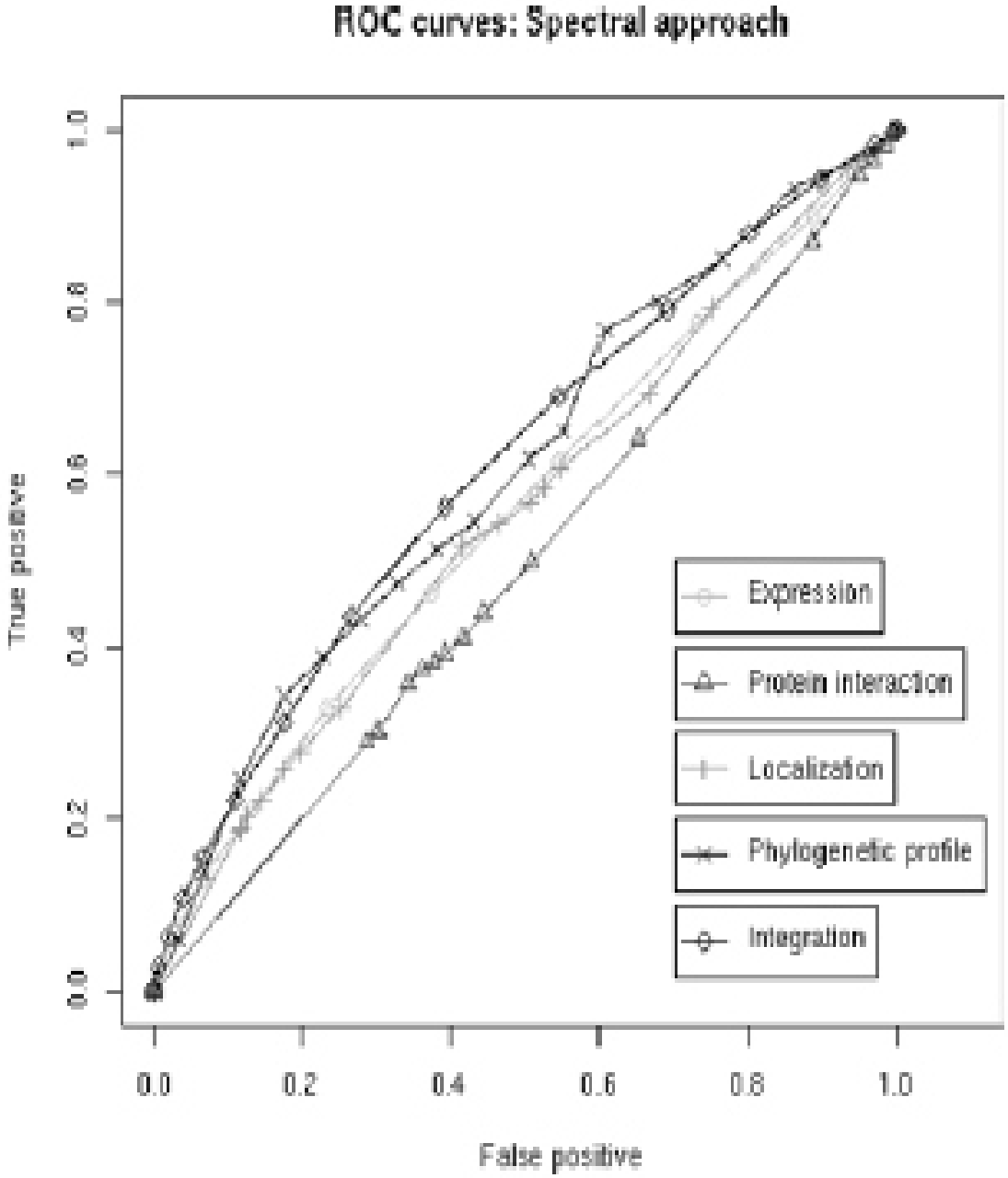

ROC curves: Direct approach
True positive
False positive
0.0
0.2
0.4
0.6
0.8
1.0
Expression
Protein interaction
Localization
Phylogenetic profile
Integration

Figura 4


ROC curves: Spectral approach
True positive
False positive
0.0
0.2
0.4
0.6
0.8
1.0
Expression
Protein interaction
Localization
Phylogenetic profile
Integration

A precisão geral dos dois métodos é muito limitada. Pouca informação é capitada pelo método direto e spectral(menos ruim).

Pouco melhor quando utilizando os dados integrados, mas ainda sem expresividade.

☺ O método supervisionado foi testado como segue:

- pararâmetros: $\lambda_1 = \lambda_2$ =0.1 e

- número de features: $L = 50$

Para cada conjunto de dados (kernels) várias combinações foram feitas para serem ajustadas ao gold standard dataset.

Para avaliar a precisão do método um *10-fold experiment* foi realizado.

Em cada uma das 10 iterações o conjunto de 769 proteínas em gold standard dataset foi dividido em:

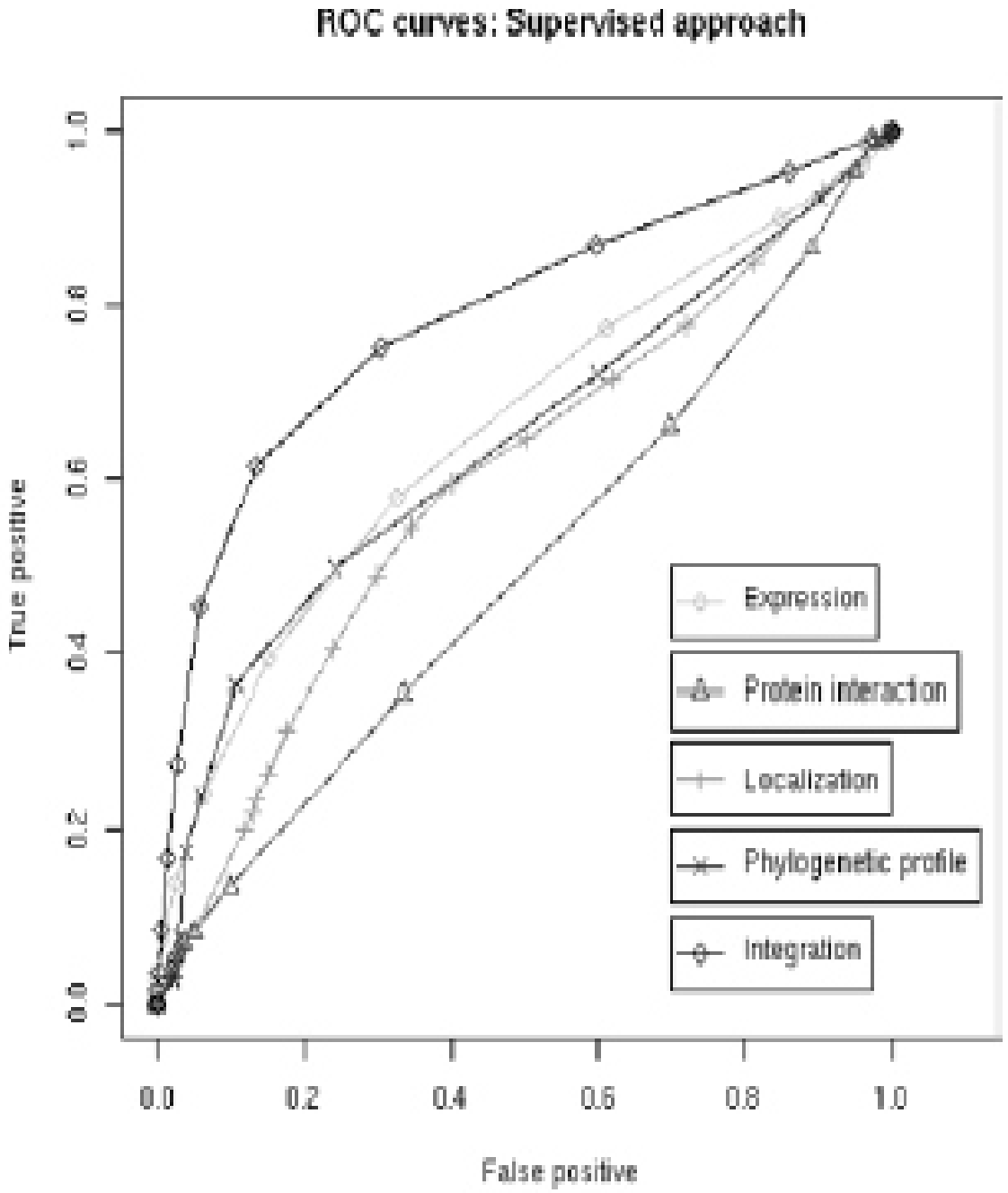
training set : 769 x 0,9 proteínas
test set : 769 x 0,1 proteínas

O feature space foi ajustado *no training set*
e
a inferência da interação foi desempenhada nas possíveis interações involvendo as proteínas no *test set*.

Novamente um gráfico foi construído e progressivamente e foi gravado o número de interações verdadeira-positivas como uma função de falso- positivas

As curvas ROC sobre a média geral das 10 iterações é apresentada na figura 5.

Figura 5

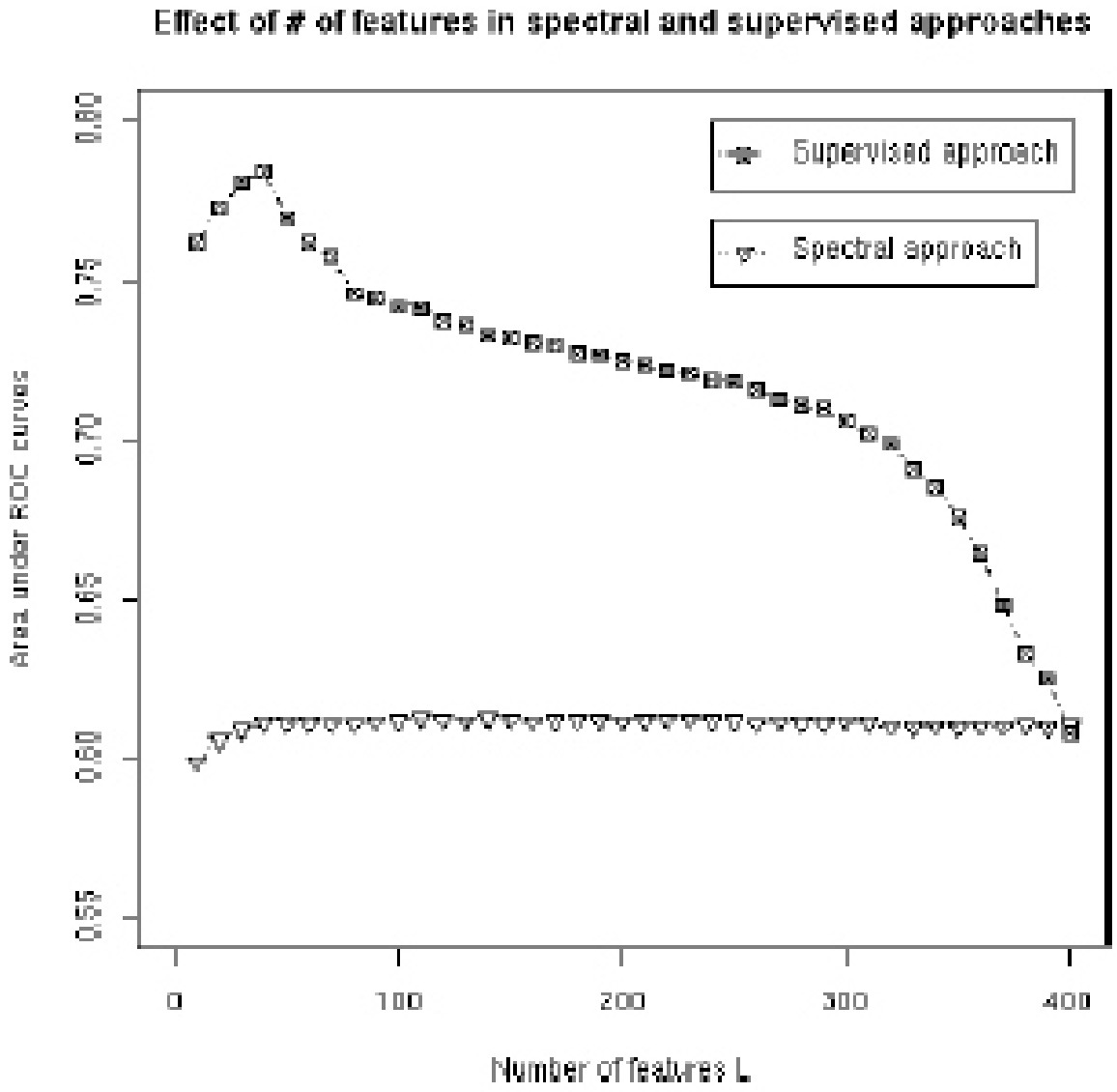

ROC curves: Supervised approach
True positive
False positive
0.0
0.2
0.4
0.6
0.8
1.0
Expression
Protein interaction
Localization
Phylogenetic profile
Integration

Em comparação com a abordagem direta e spectral, a supervisionada apresentou uma grande melhora (melhores predições)

Finalmente foi investigado o efeito que o número de features L tinha sobre a abordagem spectral e supervisionada.

A figura 6 mostra a área abaixo da curva ROC nas duas abordagens.

Figura 6



(*Máximo L=40 supervisioned approach.*)

Bibliografia

Yamanishi, Y., Vert, J.-P. And Kanehisa, M.(2003) Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics (in ISMB 2004)*, **20**, 363l-370l.

Johnson, R. A., and Wichern, D. W.(2002) Applied multivariate statistical analysis, 5th ed. Prentice Hall, Upper Saddle River, NJ

website

http:/gim.unmc.edu/dxtests/ROC1.htm (ROC curves)