**Seleção de características e
predição intrinsecamente multivariada
em identificação de
redes de regulação gênica**

David Correa Martins Junior

TESE APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
DOUTOR EM CIÊNCIAS

Programa: Ciência da Computação

Orientador: Prof. Dr. Roberto Marcondes Cesar Junior

Co-Orientador: Prof. Dr. Junior Barrera

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da
FAPESP (proc. 04/03967-0) e da CAPES (proc. BEX-3696-06-5)

São Paulo, junho de 2009

*Seleção de características e
predição intrinsecamente multivariada
em identificação de
redes de regulação gênica*

Este exemplar corresponde à redação final
da tese devidamente corrigida
e apresentada por David Correa Martins Junior
e aprovada pela Comissão Julgadora.

São Paulo, 29 de junho de 2009

Banca Examinadora:

- Prof. Dr. Roberto Marcondes Cesar Junior (orientador) - IME-USP
- Prof. Dr. Hugo Aguirre Armelin - IQ-USP
- Prof. Dr. Sandro José de Souza - ILPSC
- Prof. Dr. Ricardo Zorzetto Nicolliello Vêncio - FMRP-USP
- Prof. Dr. Junior Barrera - FFCLRP-USP

Agradecimentos

Primeiramente, ofereço meus mais sinceros agradecimentos ao grande amigo, o professor Roberto M. Cesar Jr, por ter orientado toda a minha pós-graduação (mestrado e doutorado) e confiado no meu trabalho. Sua dedicação à pesquisa, bem como a energia e a motivação que ele transmite, foram catalisadores indispensáveis para o pleno alcance de todos os objetivos relacionados com esta tese.

A outro grande amigo, o professor Junior Barrera, por ter co-orientado este trabalho. Sua sabedoria, sua experiência e sua presença em momentos importantes foi essencial para o enriquecimento de toda pesquisa relacionada a esta tese. Agradeço também ao seu esforço em estabelecer os meus primeiros contatos com o professor Edward R. Dougherty para a realização de um estágio de doutorado sanduíche no exterior.

I'am very grateful to professor Edward R. Dougherty (a.k.a. Ed) by advising me at the Genomics Signal Processing Laboratory (GSP Lab) in Texas A&M University in a one-year PhD research stage. This work was plenty benefited by all his thoughts, wisdom, motivation and energy. Also, it is worth to mention his great effort to make this stage possible and to provide all necessary conditions to the harmonic development of the research. I also would like to thank all his students in GSP Lab for the friendship and the help with issues discussed in lab.

Também não poderia deixar de agradecer ao professor Ulisses Braga Neto por ter me beneficiado sobremaneira durante o estágio de doutorado no Texas A&M University, não apenas com a pesquisa em si, mas também com inúmeras outras questões, especialmente por ter me ajudado com a questão da língua inglesa em diversos momentos. Desnecessário dizer que sua colaboração foi fundamental para o sucesso desse estágio.

Aos demais colaboradores dessa pesquisa. Ao professor Ronaldo Hashimoto e ao Evaldo Araújo pela colaboração em questões envolvendo predição intrinsecamente multivariada. Ao professor Hernando Portillo pelas análises biológicas dos resultados das PGNs em dados de malária. Ao Ricardo Vêncio e ao Carlos Bragança Pereira pela análise estatística dos dados da malária. To professor Michael Bittner for the biological analysis of intrinsically multivariate prediction results in melanoma data. Ao Fabrício Lopes pela colaboração no desenvolvimento do software de seleção de características e pela cuidadosa revisão de algumas partes da tese, bem como pelas suas valiosas sugestões a respeito.

Ao Marcelo Ris pela colaboração na criação de um novo algoritmo *branch-and-bound* de seleção de características. Ao Luís Gustavo Nonato e ao Mario Liziér pela colaboração envolvendo o trabalho de geração de malhas sobre texturas com o auxílio de operadores janela (W-operadores).

A todos os colegas do laboratório BIOINFO da Rede Vision do IME-USP pela amizade e pela ajuda em questões técnicas do laboratório. Agradecimento especial à Ana Carolina Simões pela leitura e correção de partes do texto.

Aos ImescosTM pela sólida amizade construída ao longo de 10 anos.

À FAPESP pelo apoio financeiro concedido ao doutorado no Brasil e à CAPES pelo suporte ao estágio de doutorado com duração de 1 ano no Texas A&M University.

Finalmente, sou eternamente grato aos meus pais Beth e David e à minha irmã Vanessa que, apesar das dificuldades, sempre me incentivaram e me apoiaram com relação às minhas escolhas, as quais nem sempre envolveram o caminho mais simples e direto. Esta tese é integralmente dedicada a eles.

Resumo

Seleção de características é um tópico muito importante em aplicações de reconhecimento de padrões, especialmente em bioinformática, cujos problemas são geralmente tratados sobre um conjunto de dados envolvendo muitas variáveis e poucas observações. Este trabalho analisa aspectos de seleção de características no problema de identificação de redes de regulação gênica a partir de sinais de expressão gênica. Particularmente, propusemos um modelo de redes gênicas probabilísticas (PGN) que devolve uma rede construída a partir da aplicação recorrente de algoritmos de seleção de características orientados por uma função critério baseada em entropia condicional. Tal critério embute a estimação do erro por penalização de amostras raramente observadas. Resultados desse modelo aplicado a dados sintéticos e a conjuntos de dados de *microarray* de *Plasmodium falciparum*, um agente causador da malária, demonstram a validade dessa técnica, tendo sido capaz não apenas de reproduzir conhecimentos já produzidos anteriormente, como também de produzir novos resultados.

Outro aspecto investigado nesta tese é o fenômeno da predição intrinsecamente multivariada (IMP), ou seja, o fato de um conjunto de características ser um ótimo caracterizador dos objetos em questão, mas qualquer de seus subconjuntos propriamente contidos não conseguirem representá-los de forma satisfatória. Neste trabalho, as condições para o surgimento desse fenômeno foram obtidas de forma analítica para conjuntos de 2 e 3 características em relação a uma variável alvo.

No contexto de redes de regulação gênica, foram obtidas evidências de que genes alvo de conjuntos IMP possuem um enorme potencial para exercerem funções vitais em sistemas biológicos. O fenômeno conhecido como canalização é particularmente importante nesse contexto. Em dados de *microarray* de melanoma, constatamos que o gene DUSP1, conhecido por exercer função canalizadora, foi aquele que obteve o maior número de conjuntos de genes IMP, sendo que todos eles possuem lógicas de predição canalizadoras. Além disso, simulações computacionais para construção de redes com 3 ou mais genes mostram que o tamanho do território de um gene alvo pode ter um impacto positivo em seu teor de IMP com relação a seus preditores. Esta pode ser uma evidência que confirma a hipótese de que genes alvo de conjuntos IMP possuem a tendência de controlar diversas vias metabólicas cruciais para a manutenção das funções vitais de um organismo.

Abstract

Feature selection is a crucial topic in pattern recognition applications, especially in bioinformatics, where problems usually involve data with a large number of variables and small number of observations. The present work addresses feature selection aspects in the problem of gene regulatory network identification from expression profiles. Particularly, we proposed a probabilistic genetic network model (PGN) that recovers a network constructed from the recurrent application of feature selection algorithms guided by a conditional entropy based criterion function. Such criterion embeds error estimation by penalization of rarely observed patterns. Results from this model applied to synthetic and real data sets obtained from *Plasmodium falciparum* microarrays, a malaria agent, demonstrate the validity of this technique. This method was able to not only reproduce previously produced knowledge, but also to produce other potentially relevant results.

The intrinsically multivariate prediction (IMP) phenomenon has been also investigated. This phenomenon is related to the fact of a feature set being a nice predictor of the objects in study, but all of its properly contained subsets cannot predict such objects satisfactorily. In this work, the conditions for the rising of this phenomenon were analytically obtained for sets of 2 and 3 features regarding a target variable.

In the gene regulatory networks context, evidences have been achieved in which target genes of IMP sets possess a great potential to execute vital functions in biological systems. The phenomenon known as canalization is particularly important in this context. In melanoma microarray data, we verified that DUSP1 gene, known by having canalization function, was the one which composed the largest number of IMP gene sets. It was also verified that all these sets have canalizing predictive logics. Moreover, computational simulations for generation of networks with 3 or more genes show that the territory size of a target gene can contribute positively to its IMP score with regard to its predictors. This could be an evidence that confirms the hypothesis stating that target genes of IMP sets are inclined to control several metabolic pathways essential to the maintenance of the vital functions of an organism.

Sumário

Lista de símbolos	v
Lista de abreviaturas e termos	ix
1 Introdução	1
1.1 Comentários iniciais	1
1.2 Objetivos	3
1.3 Contribuições	4
1.4 Organização do texto	7
2 Revisão e conceitos básicos	11
2.1 Sinais de expressão gênica	11
2.2 Reconhecimento de padrões em análise de sinais de expressão	13
2.2.1 Classificador Bayesiano	15
2.2.2 O problema da dimensionalidade	17
2.3 Seleção de características	19
2.3.1 Métodos <i>wrapper</i>	20
2.3.2 Funções critério	24
2.4 Redes de regulação gênica	31
2.4.1 Lista de partes	32

2.4.2	Modelos topológicos	32
2.4.3	Modelos de lógicas de controle	34
2.4.4	Modelos dinâmicos	37
2.4.5	Identificação de redes gênicas	41
3	Identificação de redes gênicas probabilísticas	45
3.1	Introdução	45
3.2	Redes gênicas probabilísticas	46
3.3	Projeto de redes gênicas probabilísticas por entropia condicional	47
3.4	Validação por meio de simulações	50
3.5	Resultados experimentais em dados de <i>Plasmodium falciparum</i>	53
3.5.1	Normalização e quantização	54
3.5.2	Dados de HB3	56
3.5.3	Dados de 3D7, Dd2 e HB3	58
3.6	Conclusão	71
4	Predição intrinsecamente multivariada: derivação analítica	77
4.1	Introdução	77
4.2	Motivação	79
4.3	Definição de predição intrinsecamente multivariada	81
4.4	Lógicas de predição	82
4.5	Poder preditivo	86
4.5.1	Preditores com distribuição uniforme	88
4.5.2	Preditores com distribuição não-uniforme	90
4.6	Resultados em expressões gênicas de melanoma	105
4.7	Discussão	110

5	Predição intrinsecamente multivariada: redes simuladas	113
5.1	Considerações preliminares	113
5.2	Modelo de redes	115
5.2.1	Hipóteses e restrições assumidas	115
5.2.2	Modelo proposto de redes	116
5.3	Descrição dos experimentos	117
5.3.1	Número de lógicas XOR/NXOR	118
5.3.2	Tamanho de território	119
5.3.3	Discussão	126
5.4	Conclusão	129
6	Conclusão	131
6.1	Considerações finais	131
6.2	Perspectivas futuras	133

Lista de símbolos

t	Instante de tempo
$\mathbf{X}$	Vetor de características
$\mathbf{x}$	Padrão ou instância observada de $\mathbf{X}$
X	Característica; variável aleatória
x	Valor da característica ou variável aleatória X
Y	Variável aleatória correspondente aos rótulos das classes
y	Rótulo da classe; valor da variável aleatória Y
n	Dimensionalidade (tamanho) do espaço de características
c	Número de classes
i, j	Índices
P	Probabilidade
$\psi(\cdot)$	Classificador Bayesiano
y^*	Rótulo atribuído pelo classificador Bayesiano
$\varepsilon(\cdot)$	Erro de classificação ou de predição
$\hat{\varepsilon}(\cdot)$	Erro resultante de um classificador projetado sobre um número finito de amostras
$\mathbf{Z}$	Subconjunto de características de $\mathbf{X}$
$\mathbf{z}$	Instância de $\mathbf{Z}$
d	Dimensão de um subconjunto de características
R	domínio de valores de características
$ \cdot $	cardinalidade
$H(\cdot)$	Entropia
$IM(\cdot)$	Informação mútua
$CoD(\cdot)$	Coeficiente de Determinação

M	Número de instâncias possíveis (possibilidades de representação de um padrão) de um vetor de características
N	Número de instâncias não observadas ou número de instâncias raramente observadas de um vetor de características (Seção 2.3.2)
s	Número de amostras de um conjunto de treinamento
α	Parâmetro para o critério de penalização de instâncias não observadas
β	Parâmetro para o critério de penalização de instâncias raramente observadas
$f(\cdot)$	Frequência absoluta (número de observações)
$\Delta(\cdot)$	Distribuição de probabilidades utilizada na penalização das instâncias raramente observadas aplicada à entropia condicional média (Equação 2.15)
$\rho(\cdot)$	conjunto dos preditores diretos de um nó em uma rede Bayesiana
$\mathcal{F}$	Função de controle lógico
$\mathbf{Y}$	Conjunto de genes alvo da predição em uma PGN
$\Pi_{\mathbf{Y} \mathbf{X}}$	Matriz de probabilidades entre transição de estados de uma cadeia de Markov
$\mathcal{G}$	Rede gênica probabilística (PGN)
Φ	Vetor de funções (função de transição de estado)
ϕ	Uma das componentes da função de transição de estados Φ
Q	Critério de qualidade para genes como preditores em uma PGN obtida por simulação (Seção 3.4)
$E[\cdot]$	Esperança
$\sigma[\cdot]$	Desvio padrão
$\mathbf{G}$	Matriz de expressões gênicas
g	Gene (perfil de expressão de g)
$\eta \cdot$	Transformação normal
A	Vetor de postos de um sinal (índice de ordenação de cada valor no sinal)
$\mathbf{a}$	Vetor de coeficientes de combinação linear
ω	Vetor de função estocástica
λ, δ	Restrições para que um conjunto de características seja de predição intrinsecamente multivariada de acordo com a Equação 4.1
$I(\cdot)$	Pontuação de predição intrinsecamente multivariada (IMP score)

α_{ij}	Distribuição conjunta de probabilidades $P(X_1 = i, X_2 = j)$
π_{ij}	Probabilidade condicional de uma variável alvo assumir valor verdadeiro ($Y = 1$) dado que $X_1 = i$ e $X_2 = j$ ($P(Y = 1 X_1 = i, X_2 = j)$)
$F(\cdot)$	Função que devolve o mínimo entre dois valores pertencentes ao intervalo $[0, 1]$ tal que a soma deles seja igual a 1
p	Poder preditivo
γ	Covariância
T	Território de um nó alvo X em um grafo
T	Um dos nós que faz parte do território do nó alvo X em um grafo
μ	Média
$C(\cdot)$	Coefficiente de incerteza

Lista de abreviaturas e termos

CoD	Coefficiente de Determinação (“Coefficient of Determination”)
DFT	Transformada Discreta de Fourier (“Discrete Fourier Transform”)
ECM	Entropia Condicional Média
IMP	Predição Intrinsecamente Multivariada (“Intrinsically Multivariate Prediction”)
<i>IMP score</i>	Medida do teor de predição intrinsecamente multivariada de um conjunto de variáveis
MCI	Melhores Características Individuais
PBN	Rede Booleana Probabilística (“Probabilistic Boolean Network”)
PGN	Rede Gênica Probabilística (“Probabilistic Genetic Network”)
PTA	Adiciona l - Remove r (“Plus l - Take Away r ”)
SAGE	Análise Serial de Expressão Gênica (“Serial Analysis of Gene Expression”)
SFS	Busca Seqüencial para Frente (“Sequential Forward Search”)
SBS	Busca Seqüencial para Trás (“Sequential Backward Search”)
SFFS	Busca Seqüencial Flutuante para Frente (“Sequential Floating Forward Search”)
SFBS	Busca Seqüencial Flutuante para Trás (“Sequential Floating Backward Search”)

Lista de Figuras

1.1	Esquema sobre a relação entre os tópicos principais desta tese.	9
2.1	Esquema simplificado da dinâmica celular.	12
2.2	Exemplo de imagem de <i>microarray</i>	13
2.3	Gráfico das taxas de erro em função da dimensionalidade com número fixo de amostras ilustrando o problema da dimensionalidade. A curva do erro Bayesiano (erro do classificador ótimo) é dada por $\varepsilon_Y(\mathbf{X})$, enquanto a curva do erro esperado ao aplicar um classificador projetado a partir de um número finito de amostras é dada por $\hat{\varepsilon}_Y(\mathbf{X})$	18
2.4	Fluxograma de um algoritmo <i>wrapper</i> genérico. Os retângulos representam procedimentos que realizam uma certa atividade, os triângulos representam fontes de informação e as setas indicam fluxo de informação.	21
2.5	O histograma da esquerda configura uma situação em que Y é bem predito por $\mathbf{x}$ porque a massa de probabilidades condicionais está bem concentrada em $Y = 1$ (entropia condicional baixa). Já para o histograma da direita, a massa de probabilidades está melhor distribuída ao longo das classes, o que faz com que o padrão $\mathbf{x}$ não seja um bom preditor de Y (entropia condicional alta).	26
2.6	Fluxograma adaptado de [Jong, 2002] que ilustra o processo de análise de redes de regulação gênica. Os retângulos representam procedimentos que realizam uma certa atividade, os triângulos representam fontes de informação e as setas indicam fluxo de informação.	31

- 2.7 Grafo orientado representando uma rede de regulação gênica. Os arcos indicados por $\rightarrow$ e $\neg$ representam interações de ativação e inibição, respectivamente. 33
- 2.8 Circuito lógico Booleano entre três genes de entrada produzindo como resultado o valor do quarto gene (Adaptado de [Schlitt and Brazma, 2007]). O quarto gene é ativado apenas quando os genes 1 e 2 estão ativos e o gene 3 está inativo, ou seja, $X_4 = [(X_1 AND X_2) AND \bar{X}_3]$ 35
- 2.9 Rede Bayesiana com todas as probabilidades condicionais associadas e a correspondente expressão da probabilidade conjunta de $\mathbf{X} = [X_1, X_2, X_3, X_4, X_5]$ 36
- 2.10 Exemplo de uma rede Booleana consistindo de 3 genes X_1, X_2, X_3 . (A) regras lógicas de transição de estado; (B) grafo de relações, dependências ou predições; (C) tabela de todos os possíveis estados (X_1, X_2, X_3) e as respectivas transições de estado para o instante de tempo posterior; (D) grafo representando as transições de estado. (Adaptado de [Schlitt and Brazma, 2007]). 38
- 3.1 Procedimento de obtenção das frequências absolutas do gene alvo Y dada cada uma das instâncias dos candidatos a preditores (X_1, X_2) a partir de uma matriz de expressões temporais ternárias composta por 48 instantes de tempo em que os genes $X_1, X_2, \dots, X_n$ são dipostos em linhas e os tempos $(t = 1, \dots, 48)$ em colunas. A figura da esquerda mostra o início do procedimento (coleta da primeira amostra $(x_1[1], x_2[1], y[2])$) e a da direita mostra o final do procedimento (coleta da última amostra $x_1[47], x_2[47], y[48]$). As cores verde, preto e vermelho representam, respectivamente, os valores -1 (subexpressão), 0 (normal) e 1 (superexpressão). Uma sequência de slides que ilustra esse processo encontra-se em <http://www.vision.ime.usp.br/~davidjr/doutorado/tese/PGNdemo.pps>. . . 49
- 3.2 Exemplo ilustrativo de arquitetura de PGN obtida utilizando-se alvos (vermelhos), cada um com o seu melhor par de preditores. Note que um alvo pode fazer parte do melhor subconjunto de preditores de um outro alvo. . . 50
- 3.3 Rede artificial de 12 genes e seus respectivos preditores. 51

3.4	Esquema do procedimento dos cálculos das informações mútuas, qualidades e suas respectivas ordenações.	52
3.5	Rede dos 50 melhores preditores individuais para cada um dos quatro alvos da glicólise (amarelo) e quatro alvos de apicoplasto (verde). No total, foram analisados 6532 genes.	59
3.6	Rede dos 20 melhores pares de preditores para cada um dos quatro alvos da glicólise (amarelo) e quatro alvos de apicoplasto (verde). No total, foram analisados 21.330.246 pares de genes (combinação de 6532, 2 a 2).	60
3.7	Histogramas de correlação de Pearson (pares de sinais), correlação de Spearman (pares de sinais) e concordância de Kendall (três sinais), contendo o número de genes para cada intervalo de correlação (ou concordância). Todos eles apresentam pelo menos 70% da massa concentrada no intervalo de 0.7 a 1, o que sugere que os três conjuntos podem ser unidos.	62
3.8	PGN dos 18 melhores preditores individuais dos alvos da glicólise formando uma única componente conexa (os nós amarelos representam os genes da glicólise). As setas vermelhas indicam dois outros genes da glicólise obtidos como preditores, fato que não ocorre quando se considera cada PGN de 3D7, Dd2 e HB3 separadamente.	64
3.9	PGN dos 10 melhores pares de preditores dos alvos da glicólise formando uma única componente conexa (os nós amarelos representam os genes da glicólise). As setas vermelhas indicam dois outros genes da glicólise obtidos como preditores, fato que não ocorre quando se considera cada PGN de 3D7, Dd2 e HB3 separadamente.	65
3.10	PGN dos 4 melhores preditores individuais dos alvos do apicoplasto formando uma única componente conexa (os nós verdes representam os genes do apicoplasto).	66
3.11	PGN dos 2 melhores pares de preditores dos alvos do apicoplasto formando uma única componente conexa (os nós verdes representam os genes do apicoplasto).	67

- 3.12 44 melhores preditores individuais dos 10 alvos da glicólise. 44 é o número máximo de preditores individuais em que os componentes formados pelos alvos da glicólise e do apicoplasto permanecem desconectados. 68
- 3.13 44 melhores preditores individuais dos 24 alvos do apicoplasto. 44 é o número máximo de preditores individuais em que os componentes formados pelos alvos da glicólise e do apicoplasto permanecem desconectados. 69
- 3.14 Esquema que mostra o gene OPFJ12793 (40S ribosomal protein, putative) fazendo a ponte de ligação entre os componentes formados por 86 melhores preditores individuais dos 10 genes alvo da glicólise e dos 24 genes alvo do apicoplasto. 86 é o número máximo de preditores individuais que conecta os dois componentes por no máximo 1 gene. 70
- 3.15 (A) Esquema ilustrando o procedimento de geração de PGNs a partir de 24 apicoproteínas selecionadas aleatoriamente dentre as 550 no total (15 melhores preditores individuais para cada alvo escolhido, obtendo 280 genes em média). Este procedimento foi executado 5000 vezes resultando numa média de 10.32% dos 280 genes como sendo de apicoproteínas. (B) Esquema que ilustra os experimentos de PGNs e DFT utilizando os 24 genes de apicoplasto como alvos. Dos 289 genes em fase com os alvos do apicoplasto (DFT), 13.14% foram de apicoproteínas. Já com relação aos genes resultantes da PGN com 60 melhores preditores individuais (280 genes), 14.28% deles foram de apicoproteínas. Finalmente, para PGN com 12 melhores pares (273 genes), as apicoproteínas compuseram 12.82% do total de genes resultantes. 72

- 3.16 Histograma de porcentagens da intersecção com o conjunto das apico-
proteínas de 5000 execuções do método de PGN com 15 melhores preditores
usando 24 apicoproteínas sorteadas como sementes. As setas indicam as
barras às quais os experimentos que utilizam genes de apicoplasto pertencem
se elas fossem uma das 5000 execuções. Note, portanto, que os
genes de apicoplasto como alvos para os três métodos comparados levam
a porcentagens significativamente acima da média. Tal média foi obtida
justamente através das 5000 execuções usando 24 apicoproteínas como se-
mentes, sendo que os genes de apicoplasto não fazem parte do conjunto das
apicoproteínas. 73
- 3.17 PGN construída a partir de 10 genes alvo da glicólise (nós amarelos) e 20
melhores preditores individuais de cada alvo. Quadrados representam genes
com sinais de expressão senoidais enquanto círculos representam genes com
sinais não senoidais. As setas indicam dois genes alvo da glicólise que não
possuem sinal senoidal e que foram incluídas na PGN. 75
- 3.18 PGN construída a partir de 27 genes alvo do apicoplasto (nós verdes) e 7
melhores preditores individuais de cada alvo. Quadrados representam genes
com sinais de expressão senoidais enquanto círculos representam genes com
sinais não senoidais. 76
- 4.1 Ilustração do conceito de predição intrinsecamente multivariada para o caso
de 3 preditores e 1 alvo: (a) Preditores tomados isoladamente não predi-
zem o comportamento do alvo; (b) Preditores tomados aos pares também
não realizam uma boa predição; (c) A tripla (X_1, X_2, X_3) é um excelente
preditor de Y 78
- 4.2 Tabelas de probabilidades conjuntas para as lógicas AND e XOR, mos-
trando o porquê dessas lógicas serem IMP. CoD_1 , CoD_2 e CoD_{12} referem-se
respectivamente a $\text{CoD}_Y(X_1)$, $\text{CoD}_Y(X_2)$ e $\text{CoD}_Y(X_1, X_2)$ 85
- 4.3 Gráfico de não-linearidade $F'(x)$ 87
- 4.4 CoD_{12} para as lógicas AND e XOR como função do poder preditivo p , com
preditores uniformes. 91

- 4.5 Valor crítico de p para IMP como uma função de δ para as lógicas AND e XOR com preditores uniformes, mostrando que a condição IMP é sempre mais restritiva para a lógica AND do que para a lógica XOR. 91
- 4.6 Gráficos representando $\max\{\text{CoD}_1, \text{CoD}_2\}$ (primeira linha), CoD_{12} (segunda linha) e a junção de $1 - \max\{\text{CoD}_1, \text{CoD}_2\}$ com o CoD_{12} para a lógica AND e três valores distintos de p (0.90, 0.95 and 0.99) assumindo que X_1 e X_2 são independentes. As probabilidades P_1 e P_2 são dadas pelos eixos X e Y respectivamente. Nos gráficos das duas primeiras linhas, cores escuras correspondem a valores menores. Na última linha de gráficos, a cor vermelha indica que $\max\{\text{CoD}_1, \text{CoD}_2\}$ e CoD_{12} possuem valores baixos, a cor verde indica que esses dois valores são altos e, finalmente, quanto mais próximo de amarelo, menor será o valor do primeiro e maior será o valor do segundo. Assim, as regiões de interesse para geração de conjuntos IMP são compostas pela intersecção das regiões escuras nos gráficos da primeira linha com as claras nos da segunda linha. Essas regiões de intersecção são indicadas pela cor amarela nos gráficos da terceira linha. 94
- 4.7 Gráficos representando $\max\{\text{CoD}_1, \text{CoD}_2\}$ (primeira linha), CoD_{12} (segunda linha) e a junção de $1 - \max\{\text{CoD}_1, \text{CoD}_2\}$ com o CoD_{12} para a lógica XOR e três valores distintos de p (0.90, 0.95 and 0.99) assumindo que X_1 e X_2 são independentes. As probabilidades P_1 e P_2 são dadas pelos eixos X e Y respectivamente. Nos gráficos das duas primeiras linhas, cores escuras correspondem a valores menores. Na última linha de gráficos, a cor vermelha indica que $\max\{\text{CoD}_1, \text{CoD}_2\}$ e CoD_{12} possuem valores baixos, a cor verde indica que esses dois valores são altos e, finalmente, quanto mais próximo de amarelo, maior será o valor do primeiro e menor será o valor do segundo. Assim, as regiões de interesse para geração de conjuntos IMP são compostas pela intersecção das regiões escuras nos gráficos da primeira linha com as claras nos da segunda linha. Essas regiões de intersecção são indicadas pela cor amarela nos gráficos da terceira linha. 95

- 4.8 Lógicas de 1,3-mintermo e seus respectivos gráficos simétricos da junção dos valores de CoD_{12} e $1 - \max\{\text{CoD}_1, \text{CoD}_2\}$ em canais verde e vermelho. O poder preditivo é $p = 0.99$ e as lógicas correspondentes a cada gráfico estão referenciadas pela notação de palavra binária. 96
- 4.9 “*Scatter plots*” do *IMP score* (I) contra entropia (H) com 970.299 pontos do espaço de parâmetros (P_1, P_2, P_3) para $p = 0.90$, $p = 0.95$ e $p = 0.99$ e lógica 00000001, (lógica IMP). As correlações respectivas estão embaixo de cada um dos gráficos. 98
- 4.10 “*Scatter plots*” do *IMP score* (I) contra entropia (H) com 970.299 pontos do espaço de parâmetros (P_1, P_2, P_3) para $p = 0.90$, $p = 0.95$ e $p = 0.99$ e lógica 00011001 (lógica semi-IMP). As correlações respectivas estão embaixo de cada um dos gráficos. 100
- 4.11 Gráficos dos CoDs e *IMP score* como funções de p para $P_1 = P_2 = 0.5$ e valores não-negativos de covariância ($\gamma \geq 0$). 103
- 4.12 Gráficos dos CoDs e *IMP score* como funções de p para $P_1 = P_2 = 0.5$ e valores não-positivos de covariância ($\gamma \leq 0$). 104
- 4.13 Gráficos representando a junção dos valores de CoD_{12} e $1 - \max(\text{CoD}_1, \text{CoD}_2)$ nos canais verde e vermelho para as lógicas AND e XOR e para três valores distintos de p e cinco valores distintos de γ . As probabilidades P_1 e P_2 são dadas pelos eixos X e Y respectivamente. Regiões amarelas correspondem a regiões de interesse para surgimento de conjuntos IMP. 106
- 4.14 Gráficos com canais verde e vermelho representando os valores de CoD_{12} e $1 - \max(\text{CoD}_1, \text{CoD}_2)$ respectivamente para os conjuntos IMP de DUSP1 dados na Tabela 4.9. Cada gráfico foi gerado tomando como entrada a lógica, o poder preditivo p e a covariância γ obtidas. Os pontos azuis indicam as localizações (P_1, P_2) obtidas. 108
- 5.1 Fluxograma ilustrando o mecanismo de construção de redes. 118
- 5.2 Exemplo de rede construída pelo modelo descrito com $n = 8$ 119

- 5.3 Diagramas apresentando os valores das medianas dos *IMP scores* máximos para conjuntos representados pelo número de lógicas XOR/NXOR para $p = 0.95$, $p = 0.9$, $p = 0.8$ e $p = 0.6$. As barras verticais indicam a região com 50% de concentração de pontos para cada conjunto, enquanto as extremidades das barras indicam onde o quartil inferior termina e onde o quartil superior começa. As correlações entre o número de lógicas XOR/NXOR e o *IMP score* máximo são: (a) 0.4269, (b) 0.4806, (c) 0.5871 e (d) 0.5856. 120
- 5.4 Diagramas apresentando os valores das medianas dos *IMP scores* máximos para conjuntos representados pelo número de lógicas de 1,3-mintermo para $p = 0.95$, $p = 0.9$, $p = 0.8$ e $p = 0.6$. As barras verticais indicam a região com 50% de concentração de pontos para cada conjunto, enquanto as extremidades das barras indicam onde o quartil inferior termina e onde o quartil superior começa. As correlações entre o número de lógicas de 1,3-mintermo e *IMP score* máximo são: (a) -0.0086, (b) -0.0414, (c) -0.1295 e (d) -0.1647. 121
- 5.5 Diagramas apresentando os valores das medianas dos *IMP scores* máximos para conjuntos representados pelo número de lógicas de grau 3 para $p = 0.95$, $p = 0.9$, $p = 0.8$ e $p = 0.6$. As barras verticais indicam a região com 50% de concentração de pontos para cada conjunto, enquanto as extremidades das barras indicam onde o quartil inferior termina e onde o quartil superior começa. As correlações entre o número de lógicas de grau 3 e *IMP score* máximo são: (a) -0.1421, (b) -0.1563, (c) -0.1115 e (d) -0.0318. 122
- 5.6 Tamanho de território de cada nó na rede exemplo ilustrada na Figura 5.2. 123
- 5.7 As melhores redes classificadas dentre as 10.000 redes simuladas para $p = 0.95$ e $p = 0.90$ no experimento apresentado na Seção 5.3.1. Os alvos com os mais altos *IMP scores* são indicados pelas setas vermelhas. Os tamanhos de território para esses dois conjuntos IMP são: (a) 13; (b) 11 124
- 5.8 As melhores redes classificadas dentre as 10.000 redes simuladas para $p = 0.80$ e $p = 0.60$ no experimento apresentado na Seção 5.3.1. Os alvos com os mais altos *IMP scores* são indicados pelas setas vermelhas. Os tamanhos de território para esses dois conjuntos IMP são: (a) 10; (b) 9. 125

-
- 5.9 Histogramas de amostras com um dado tamanho de território e *IMP score* dentre os 5% maiores para lógicas XOR/NXOR de predição do nó 1 e $p = \{0.95, 0.9, 0.8, 0.6\}$ 127
- 5.10 Histogramas de amostras com um dado tamanho de território e *IMP score* dentre os 5% maiores para lógicas 1,3-mintermo de predição do nó 1 e $p = \{0.95, 0.9, 0.8, 0.6\}$ 128
- 5.11 Evolução das médias das covariâncias (em módulo) entre os preditores do gene 1 ao longo dos tamanhos de território para $p = \{0.95, 0.90, 0.80, 0.60\}$. As respectivas correlações das covariâncias e dos tamanhos de território são -0.5105, -0.5919, -0.6510 e -0.5960. 129
- 5.12 Relação do *IMP score* com as demais condições analisadas (lógicas, poder preditivo, tamanho de território e covariância entre preditores). 130

Lista de Tabelas

- 2.1 Exemplo de tabelas de distribuição de probabilidades conjuntas e condicionais para $\mathbf{X} = \{X_1, X_2\} = \{0, 1\}^2$ e $Y = \{0, 1\}$. Note que, pela tabela de probabilidades conjuntas sabe-se que $P(X_1 = 0, X_2 = 0) = 0.4$, $P(X_1 = 0, X_2 = 1) = 0.25$, $P(X_1 = 1, X_2 = 0) = 0.1$ e $P(X_1 = 1, X_2 = 1) = 0.25$. Essas informações não estão presentes na tabela de probabilidades condicionais. 17
- 2.2 Tabela de frequências absolutas de (X_1, X_2, X_3, Y) e suas respectivas entropias condicionais e erros de predição, com e sem penalização de instâncias raramente observadas para $\alpha = 1$. As frequências absolutas para $Y = 0$ e $Y = 1$ são indicadas respectivamente por $f(Y = 0)$ e $f(Y = 1)$. $f(\mathbf{x})$ e $f'(\mathbf{x})$ são as frequências absolutas da instância $\mathbf{X} = \mathbf{x}$ sem e com penalização, respectivamente. $H(Y|\mathbf{x})$ e $H'(Y|\mathbf{x})$ são as entropias condicionais sem e com penalização, respectivamente. $\varepsilon_Y(\mathbf{x})$ e $\varepsilon'_Y(\mathbf{x})$ são os erros *a posteriori* sem e com penalização, respectivamente. A última linha da tabela, denotada por “Total”, indica as frequências absolutas totais e as médias ponderadas das entropias condicionais e dos erros *a posteriori*. Para este exemplo, $H(Y) = 0.9932$, $\varepsilon_Y = 0.4516$ e os CoDs sem e com penalização valem 0.3577 e 0.2892, respectivamente. 29

- 2.3 Tabela de freqüências absolutas de (X_1, X_2, X_3, Y) e suas respectivas entropias condicionais e erros de predição, com e sem penalização de instâncias observadas uma única vez para $\beta = 0.8$. As freqüências absolutas para $Y = 0$ e $Y = 1$ e $\mathbf{X} = \mathbf{x}$ são indicadas respectivamente por $f(Y = 0)$, $f(Y = 1)$ e $f(\mathbf{x})$. $H(Y|\mathbf{x})$ e $H'(Y|\mathbf{x})$ são as entropias condicionais sem e com penalização, respectivamente. $\varepsilon_Y(\mathbf{x})$ e $\varepsilon'_Y(\mathbf{x})$ são os erros *a posteriori* sem e com penalização, respectivamente. A última linha da tabela, denotada por “Total”, indica as freqüências absolutas totais e as médias ponderadas das entropias condicionais e dos erros *a posteriori*. Para este exemplo, $\varepsilon_Y = 0.4516$ e os CoDs sem e com penalização valem 0.3517 e 0.3428, respectivamente. 30
- 3.1 Cinco melhores genes individuais obtidos de acordo com a informação mútua (IM , na tabela) para cada gene da rede disposto como alvo a partir de um conjunto de 48 amostras simuladas através da rede de predição descrita na Figura 3.3 com $P(Y|\mathbf{X} = \mathbf{x}) = 0.8$ para todo $\mathbf{x} \in \mathbf{X}$ 53
- 3.2 Cinco melhores genes em termos de qualidade (Q , na tabela) para os cinco melhores pares obtidos de acordo com a informação mútua para cada gene alvo a partir de um conjunto de 48 amostras simuladas através da rede de predição descrita na Figura 3.3 com $P(Y|\mathbf{X} = \mathbf{x}) = 0.8$ para todo $\mathbf{x} \in \mathbf{X}$. . 54
- 3.3 Cinco melhores genes em termos de qualidade (Q , na tabela) para as cinco melhores triplas obtidas de acordo com a informação mútua para cada gene alvo a partir de um conjunto de 48 amostras simuladas através da rede de predição descrita na Figura 3.3 com $P(Y|\mathbf{X} = \mathbf{x}) = 0.8$ para todo $\mathbf{x} \in \mathbf{X}$. . 55
- 4.1 Lógica ou-exclusivo (XOR) como exemplo de predição intrinsecamente multivariada. 78
- 4.2 Lógica AND como exemplo de função canalizadora. 80
- 4.3 Tabela de distribuição conjunta de probabilidades para a lógica AND assumindo independência entre X_1 e X_2 92
- 4.4 Tabela de distribuição conjunta de probabilidades para a lógica XOR assumindo independência entre X_1 e X_2 92

- 4.5 Número de conjuntos IMP e entropia média das distribuições dos preditores para cada subespaço do espaço de parâmetros (P_1, P_2, P_3) definido por λ e δ para $p = 0.90$, $p = 0.95$ e $p = 0.99$ e lógica 00000001. Os eixos X e Y correspondem aos valores de δ e λ respectivamente. As legendas estão dispostas na última coluna (o número de conjuntos IMPs é dado por potências de 2 e as entropias estão no intervalo de 0 a 1). 98
- 4.6 Número de conjuntos IMP e entropia média das distribuições dos preditores para cada subespaço do espaço de parâmetros (P_1, P_2, P_3) definido por λ e δ para $p = 0.90$, $p = 0.95$ e $p = 0.99$ e lógica 00011001 (semi-IMP). Os eixos X e Y correspondem aos valores de δ e λ respectivamente. As legendas estão dispostas na última coluna (o número de conjuntos IMPs é dado por potências de 2 e as entropias estão no intervalo de 0 a 1). 99
- 4.7 Tabela de distribuição de probabilidades conjuntas para a lógica AND adaptada com a covariância γ 100
- 4.8 Tabela de distribuição de probabilidades conjuntas para a lógica XOR adaptada com a covariância γ 101
- 4.9 Lista de conjuntos IMP que têm DUSP1 como alvo. Os poderes preditivos e as covariâncias estão indicados nas colunas p e γ respectivamente. Os respectivos gráficos da junção dos valores de CoD_{12} e $1 - \max(\text{CoD}_1, \text{CoD}_2)$ nos canais verde e vermelho estão ilustrados na Figura 4.14. Esses gráficos foram gerados com base na lógica de predição, γ , p e o par (P_1, P_2) (o último indicado por um ponto azul nos gráficos). 109
- 4.10 Genes alvo com o maior número de conjuntos IMP para os pares $(\lambda, \delta) = \{(0.2, 0.7), (0.2, 0.8), (0.3, 0.7), (0.3, 0.8)\}$ 110
- 4.11 Distribuição dos genes IMP classificados como lógicas de 1,3-mintermo e 2-mintermo (XOR e NXOR) para cada valor de poder preditivo apresentado nos resultados com $\lambda = 0.2$ e $\delta = 0.8$ ($p = 0.903, 0.936, 0.968, 1$). 110
- 5.1 Tabelas de distribuição de probabilidades condicionais para lógicas de predição de grau de entrada 1 (ativação e inibição) com poder preditivo p 115

5.2	Médias (μ) e desvios padrões (σ) dos <i>IMP scores</i> para cada tamanho de território (t) e para três quantidades de amostras (30.000, 60.000, 120.000) considerando lógicas XOR/NXOR como preditoras do nó 1 e poder preditivo $p = 0.95$	126
-----	--	-----

Capítulo 1

Introdução

1.1 Comentários iniciais

O reconhecimento de padrões é um problema com aplicações nos mais diversos campos do conhecimento, sendo que em bioinformática, por si só, muitas questões podem ser respondidas com seu auxílio. Dentre elas pode-se destacar: análise de seqüências de DNA, classificação estrutural de proteínas, diagnóstico de tumores de tecido, identificação de redes de regulação gênica, análise filogenética, entre outros. Uma particularidade existente na maioria das aplicações que exigem reconhecimento de padrões em bioinformática é o fato dos conjuntos de dados apresentarem inúmeras medidas para poucas observações (tipicamente dezenas de amostras com milhares de medidas). Porém, a situação estatisticamente desejável para um projeto de classificador de padrões por aprendizado computacional a partir de conjuntos de treinamento é justamente a oposta (milhares de amostras para dezenas de medidas). Portanto, selecionar um subespaço do espaço total de características que represente de forma satisfatória os objetos observados (redução de dimensionalidade) é uma tarefa bastante importante e recorrente em reconhecimento de padrões aplicado a problemas de bioinformática, embora não trivial [Duda et al., 2000, Bittner et al., 2000, Kim et al., 2000a, Kim et al., 2000b, Barrera et al., 2006a, Barrera et al., 2007].

A seleção de características é uma das principais abordagens de redução de dimensionalidade que seleciona um subconjunto das características originais para representação dos padrões observados. Em análise de sinais de expressão gênica (concentrações de mRNA), existem dois objetivos principais em realizar seleção de características

[Dougherty et al., 2005]. Um deles é eliminar genes irrelevantes de um modelo de classificação (ou predição) de forma a melhorar o seu desempenho. Outro objetivo é descobrir a estrutura das redes gênicas ou dos mecanismos responsáveis por um determinado fenômeno biológico (progresso ou repressão de uma doença, por exemplo).

Em reconhecimento de padrões, a informação é usualmente dada na forma de vetores aleatórios com um grande número de atributos. Um algoritmo de reconhecimento de padrões opera diretamente sobre as variáveis considerando-as como características. Em análise de dados de expressão gênica, seleção de características significa essencialmente seleção de genes.

A importância da seleção de características é bem documentada na literatura de aprendizado computacional, já que o desempenho de algoritmos de classificação ou predição degrada rapidamente não só com a inserção de características irrelevantes, mas também com a inclusão de características relevantes, caso elas sejam correlacionadas. Essa degradação é preocupante em casos realistas nos quais não se conhece os parâmetros das distribuições de probabilidades. Nesses casos, o algoritmo de predição deve estimar os parâmetros com base nos dados disponíveis. Caso o número de características representando os padrões seja alto, a variância do grande número de parâmetros correspondentes também é alta. Assim, o conjunto de características ótimo estimado depende do tamanho do conjunto de treinamento, sendo maior para conjuntos de treinamento com maior número de amostras.

O controle celular é resultado de atividade multivariada entre genes. Assim, para propósitos terapêuticos e criação de novas drogas, há uma forte motivação em modelar essa interação multivariada. Genes e proteínas agem em conjunto, formando intrincadas redes regulatórias. Redes de regulação gênica descrevem esses processos regulatórios e a reação molecular de uma célula a vários estímulos. Técnicas que medem concentrações de RNA e proteínas permitem novas abordagens de análise de tais redes. A análise desses dados requer técnicas sofisticadas, sendo que novos métodos vem sendo desenvolvidos com esse propósito [Kelemen et al., 2008].

Modelagens com diversos níveis de detalhe são empregadas para descrição de redes regulatórias. Nesse sentido, a inferência de parâmetros de uma rede regulatória a partir de dados experimentais é um dos grandes desafios existentes em bioinformática. O fato do número de parâmetros ser muito maior que o número de amostras disponíveis para esse propósito deu início ao uso de fontes de dados heterogêneas e inclusão de conhecimento

biológico na estimação de parâmetros para inferência de redes.

Na literatura, existem duas abordagens principais de modelagem de redes regulatórias [Dougherty et al., 2005]. A primeira baseia-se na construção de modelos quantitativos detalhados de redes bioquímicas para reações celulares de interesse, empregando equações diferenciais e suas variações [Jong, 2002]. Essa abordagem costuma exigir um conjunto considerável de amostras, além de determinados parâmetros característicos das reações. A segunda abordagem envolve a construção de modelos qualitativos de interação gênica usando a quantidade limitada de dados de sinais de expressão que se encontram disponíveis geralmente. Os principais paradigmas para isso incluem os modelos de grafos orientados, redes Bayesianas [Friedman et al., 2000], redes Booleanas [Kauffman, 1969], redes lógicas generalizadas [Thomas et al., 1995] e redes gênicas probabilísticas que incluem o caso especial de redes Booleanas probabilísticas [Shmulevich et al., 2002].

Embora a seleção de características seja um dos passos principais na tentativa de descobrir relações biológicas verdadeiras, dificilmente um conjunto de características pode ser considerado relevante biologicamente sem validação posterior. O comportamento quase ótimo observado através de dados experimentais é apenas um ponto de partida para experimentação e validação biológica mais detalhada. Com isso, um procedimento de seleção de características propõe diversos conjuntos de características candidatos que podem ser usados posteriormente em estudos biológicos.

1.2 Objetivos

Esta tese analisa dois aspectos de seleção de características no problema de identificação de redes de regulação gênica a partir de sinais de expressão. Um primeiro aspecto abordado é o de modelagem de redes de regulação gênica a partir de sinais de expressão temporal provenientes de experimentos de *microarray*. O foco dessa parte do trabalho foi sobre a formalização de um modelo de redes gênicas probabilísticas (do inglês: *Probabilistic Genetic Networks* - PGN) em conjunto com um método de identificação de redes baseado nesse modelo.

Outro aspecto investigado em profundidade nesta tese é o fenômeno da predição intrinsecamente multivariada (do inglês: *intrinsically multivariate prediction* - IMP). Um conjunto IMP é constituído por um par de características preditores-alvo tal que o alvo

é predito de forma intrinsecamente multivariada por seus preditores. Predição intrinsecamente multivariada é aquela na qual os preditores em conjunto realizam uma ótima predição do alvo, mas quaisquer de seus subconjuntos propriamente contidos não realizam uma predição satisfatória do comportamento da característica alvo. Compreender como tal fenômeno ocorre pode ter implicações diretas no desenvolvimento de novos métodos de seleção de características. Além disso, sua compreensão pode ser fundamental na área de modelagem de redes de regulação gênica, pois existe a suspeita de que genes IMP têm potencial para regular diversas vias metabólicas vitais de um organismo (funções de canalização) [Kauffman, 1969].

1.3 Contribuições

As principais contribuições desta tese, além de uma revisão bibliográfica sobre modelagem de redes de regulação gênica encontrada no Capítulo 2, encontram-se a seguir. Os apêndices referentes aos artigos originados dessas contribuições podem ser encontrados em <http://www.vision.ime.usp.br/~davidjr/doutorado/tese/apendices>.

- Formalização de um modelo de construção de PGN através de seleção de características por entropia condicional média com penalização de instâncias raramente observadas. Esse trabalho iniciou-se ainda no mestrado [Martins-Jr., 2004] e contou com a colaboração de Carlos A. Bragança Pereira e Ricardo Z. N. Vêncio para realização de análise estatísticas, e da equipe de Hernando Del Portillo do Instituto de Ciências Biomédicas-USP para a análise e validação do modelo proposto em dados públicos de *Plasmodium falciparum*, um agente causador da malária [Barrera et al., 2004] [Barrera et al., 2005] [Barrera et al., 2006a] [Barrera et al., 2006b] (Capítulo 3 e Apêndices A, B, C e D).
- Investigação das condições para as quais o fenômeno de predição intrinsecamente multivariada ocorre. Dessa investigação, originou-se um estudo analítico para conjuntos de 2 e 3 preditores e outro estudo através de simulações computacionais de geração de redes com 3 ou mais nós. Essa pesquisa originou-se de uma colaboração com os professores Edward Russell Dougherty e Ulisses Braga-Neto durante o estágio de doutorado sanduíche realizado no Genomic Signal Processing Labora-

tory, Texas A&M University nos Estados Unidos. As análises envolvendo dados de *microarray* de melanoma contaram com a colaboração de Michael L. Bittner, pesquisador do Translational Genomics Research Institute (TGEN) nos Estados Unidos [Martins-Jr et al., 2008b, Martins-Jr et al., 2008a] (Capítulos 4 e 5 e Apêndices E e F).

- Generalização do critério de penalização de instâncias raramente observadas desenvolvido no mestrado [Martins-Jr., 2004] de forma a incluir um parâmetro de confiança β para a classe observada dada uma determinada instância que ocorreu apenas uma vez no conjunto de treinamento. Esse critério foi usado em [Barrera et al., 2004, Barrera et al., 2006a, Martins-Jr et al., 2006b, Lizier et al., 2008], sendo definido e discutido na Seção 2.3.2.

Além dessas, outras contribuições que envolvem seleção de características foram originadas de colaborações com outros pesquisadores durante o doutorado. Embora esses trabalhos não estejam enfocados nesta tese (alguns deles são apenas citados e brevemente comentados em seu decorrer), todos os artigos correspondentes encontram-se disponíveis em <http://www.vision.ime.usp.br/~davidjr/doutorado/tese/apendices>. A seguir, uma sinopse descritiva de cada uma dessas contribuições é apresentada:

- Idealização de uma nova abordagem para seleção de genes específicos que distinguem estados biológicos distintos baseando-se na análise de dados de SAGE (*Serial Analysis of Gene Expression*). Tal método aplica o erro *bolstered* para a identificação de genes fortes que separam os estados biológicos em um espaço de características definido pelas expressões gênicas de um conjunto de treinamento. Intervalos de credibilidade, definidos sobre um modelo probabilístico de medidas de SAGE, foram implementados com o intuito de identificar genes que distinguem os diferentes estados com maior confiabilidade dentre todos os grupos de genes selecionados pelo método de identificação de genes fortes descrito em [Silva et al., 2005]. Uma pontuação que leva em conta a credibilidade e o erro *bolstered* para classificar os grupos de genes considerados foi proposta. Os resultados obtidos a partir de dados de SAGE de gliomas (tipos de tumor cerebral) foram apresentados, corroborando a metodologia introduzida. Essa contribuição iniciou-se no mestrado (as idéias gerais desse método, bem como resultados preliminares, podem ser encontrados na

dissertação de mestrado [Martins-Jr., 2004]) e originou a publicação de um artigo durante o doutorado que formalizou as idéias do método. Esse trabalho envolveu uma colaboração com a pesquisadora Helena Brentani do Ludwig Institute for Cancer Research [Barrera et al., 2007] (Apêndice G).

- Formulação de um problema de otimização combinatória em que o espaço de busca é o conjunto potência de um conjunto finito estruturado como um reticulado Booleano. Esse reticulado tem uma propriedade em que a função custo de todos os intervalos que contêm o conjunto vazio e o conjunto completo de características formam uma “curva em U”. Tal formulação, aplicada a seleção de características, consiste de um algoritmo *branch-and-bound* que explora o espaço de busca sobre essa estrutura de reticulado. Diversos experimentos com dados públicos bem conhecidos indicam a superioridade do método proposto sobre a Busca Seqüencial Flutuante, uma heurística bastante utilizada por possuir um excelente custo-benefício em termos de custo computacional e qualidade dos resultados. Este trabalho surgiu de uma colaboração com Marcelo Ris, recém doutor em Bioinformática da USP [Ris et al., 2008] (Apêndice H).
- Implementação de um ambiente gráfico multiplataforma de código fonte aberto provendo algoritmos e funções critério de seleção de características, projeto de classificadores, métodos de estimação de erro e ferramentas de visualização gráfica. Esse sistema permite análises de dados utilizando diversos métodos e parâmetros. O principal foco desse software são as aplicações de bioinformática, embora seja útil para realizar tarefas de reconhecimento de padrões de um modo geral. Este projeto conta com a colaboração de Fabrício Martins Lopes, doutorando em Bioinformática do IME-USP, e sua versão atual está hospedada na Incubadora da FAPESP: <http://dimreduction.incubadora.fapesp.br> [Lopes et al., 2008b] (Apêndice I).
- Desenvolvimento de uma técnica que procura uma janela mínima para a estimação de um operador janela (W-operador) a partir dos dados de treinamento. A janela devolvida é aquela que maximiza a informação das classes observadas no conjunto de treinamento através da minimização da entropia condicional média estimada por penalização das instâncias não observadas (esse critério é definido e discutido na Seção 2.3.2). A técnica proposta foi testada com sucesso em duas aplicações de processamento de imagens e visão computacional: filtragem de imagens binárias e

reconhecimento de texturas em imagens com tons de cinza [Martins-Jr et al., 2005, Martins-Jr et al., 2006a] (Apêndices J e K).

- Proposta de um método de geração de malhas sobre texturas em imagens coloridas. Tal método combina classificação de texturas coloridas por operadores janela (W-operadores) e uma técnica de geração de malhas simples (*Imesh*) sobre imagens níveis de cinza. Uma extensão de W-operadores para tratamento de texturas coloridas é proposta, empregando uma combinação de canais RGB e HSV e o algoritmo de Busca Seqüencial Flutuante para Frente (SFFS) orientado pelo critério da entropia condicional média com penalização de instâncias raramente observadas. A idéia é selecionar um subconjunto de pontos da janela que contém a maior informação sobre as texturas em questão de acordo com um conjunto de treinamento. O W-operador é embutido na estimação de erro local usado pelo *Imesh* para a escolha dos vértices da malha. Posteriormente, o W-operador permite a atribuição de um rótulo aos triângulos durante a construção da malha, resultando em uma malha segmentada ao final do processo. Resultados mostram que a combinação de W-operadores com *Imesh* origina uma abordagem de geração de malhas triangulares baseada em classificação de texturas que possui um desempenho melhor que métodos pixel a pixel. Esta contribuição envolveu a colaboração de Mario Augusto de Souza Lizier, doutorando em Ciência da Computação e Matemática Computacional, e o Prof. Luiz Gustavo Nonato, ambos do Instituto de Ciências Matemáticas e de Computação - USP [Martins-Jr et al., 2006b, Lizier et al., 2008] (Apêndices L e M).

1.4 Organização do texto

O Capítulo 2 apresenta uma revisão sobre os conceitos necessários para o entendimento do restante da tese. A Seção 2.1 introduz o que são sinais de expressão gênica e discute a tecnologia de *microarray* de medição desses sinais. A Seção 2.2 apresenta tópicos de reconhecimento de padrões importantes em análise de expressão gênica, tais como o conceito de classificador Bayesiano e o problema da dimensionalidade. Na Seção 2.3, serão discutidos métodos de seleção de características com atenção especial aos algoritmos *wrapper* e funções critério, salientando a penalização de instâncias raramente ou nunca observadas, uma das contribuições desta tese. A Seção 2.4 apresenta um panorama da área de

modelagem e identificação de redes de regulação gênica.

No Capítulo 3 encontra-se a formulação do modelo para identificação de PGN por entropia condicional. Os resultados de simulações computacionais são apresentados na Seção 3.4 e os obtidos da análise de conjuntos de dados do *Plasmodium falciparum* podem ser encontrados na Seção 3.5.

O conceito de predição intrinsecamente multivariada (IMP) é introduzido no Capítulo 4, bem como a motivação para o estudo desse fenômeno. As condições para obter esse fenômeno em conjuntos de 2 e 3 preditores são estabelecidas e discutidas nas Seções 4.4 e 4.5. Os resultados obtidos através da análise de expressões gênicas de melanoma são apresentados na Seção 4.6, seguidos de uma seção de discussão (Seção 4.7).

O Capítulo 5 apresenta uma extensão do estudo de predição intrinsecamente multivariada para redes com 3 ou mais genes. A Seção 5.2 define o modelo de geração de redes por simulação bem como as hipóteses e restrições que ele assume. A descrição dos experimentos e dos resultados obtidos pode ser encontrada na Seção 5.3.

Finalmente, o Capítulo 6 encerra a tese com conclusões e perspectivas futuras para continuação dos trabalhos focados. A Figura 1.1 mostra um esquema de como os tópicos principais desta tese são relacionados entre si.

Figura 1.1: Esquema sobre a relação entre os tópicos principais desta tese.

Capítulo 2

Revisão e conceitos básicos

2.1 Sinais de expressão gênica

Dado que uma funcionalidade específica de uma célula eucariota é fortemente determinada pelos genes que ela expressa, e sendo a transcrição o primeiro passo no processo de converter a informação armazenada no DNA do organismo em proteínas, é de se esperar que esse processo seja regulado pela rede de controle que coordena a atividade celular [Dougherty et al., 2005]. Um meio primário de regulação da atividade celular é o controle de produção de proteína através da quantidade de RNA mensageiro (mRNA) expresso pelos genes. Mensagens de RNA passam pelos orifícios do núcleo celular chegando no ribossomo, no citoplasma, onde eles são traduzidos em seqüências de proteínas, construindo enzimas que catalisam reações metabólicas ou retornam ao núcleo para interagir com o DNA na regulação da síntese de RNA. A Figura 2.1 ilustra esse processo dentro da célula. Portanto, conjuntos de genes constituem redes de comunicação bastante complexas que controlam vias metabólicas celulares.

Existem diversas ferramentas baseadas em processos físico-químicos que permitem medir fenômenos moleculares tais como decodificação de DNA, descrição de estrutura de proteínas e estimação de concentração de mRNA e proteínas nas células. A existência de dados desse tipo permite a aplicação de técnicas matemáticas e computacionais para extrair informação e modelar fenômenos biológicos. Tendo em vista que o procedimento de medição dessa expressão não é uma tarefa fácil, existem diversas tecnologias que foram desenvolvidas para este fim: cDNA *microarrays*, RT-PCR (reação em cadeia da polymerase

Figura 2.1: Esquema simplificado da dinâmica celular.

de transcriptase reversa), SAGE (serial analysis of gene expression), entre outros.

Uma das tecnologias mais utilizadas para extração de sinais de transcrição é a tecnologia de *microarray* [Shalon et al., 1996]. Essa técnica consiste em um processo bioquímico que mede os níveis de expressão de milhares de genes simultaneamente a partir de amostras de mRNA de estados biológicos distintos. Durante a retrotranscrição, DNAs complementares (cDNA) sintetizados são marcados com fluorocromos de diferentes cores (Cy3: verde, Cy5: vermelho). Quantidades idênticas dos cDNAs marcados são misturados e colocados em contato com uma lâmina, sendo submetidos a um processo de hibridação competitiva entre as duas amostras (o cDNA é fixado em *spots* que possuem uma sequência de bases complementares). Após a hibridização, as lâminas de cDNA passam por um scanner a laser que produz uma imagem digital com duas bandas de cores (verde e vermelho) (Figura 2.2). A abundância relativa de mRNA nas duas amostras é representada por uma razão entre essas intensidades. Existem diversas técnicas computacionais que analisam e decodificam essas imagens em uma matriz de valores reais. Essa matriz serve como entrada para uma série de análises, desde classificação e caracterização de fenômenos biológicos até identificação de redes de regulação gênica.

Figura 2.2: Exemplo de imagem de *microarray*.

2.2 Reconhecimento de padrões em análise de sinais de expressão

Em análise de sinais genômicos, uma grande variedade de problemas envolve reconhecimento de padrões. Por exemplo, *microarrays* contêm medidas de expressão de milhares de genes e um dos principais objetivos é a classificação de padrões a partir desses perfis de expressão. Exige-se, portanto, o projeto de um classificador ψ que receba como entrada um vetor de níveis de expressão gênica $\mathbf{x} = (x_1, x_2, \dots, x_n)$ e devolva um rótulo que prediz a classe $y = \{0, 1, \dots, c - 1\}$ à qual o vetor considerado pertence. Um problema típico em análise de expressão gênica é a classificação de diferentes tipos de câncer ou diferentes estágios de desenvolvimento de um tumor [Porter et al., 2001]. Classificadores são projetados com base em um conjunto de amostras (vetores de expressão) que podem ser provenientes de tecidos diferentes ou de um mesmo tecido, que em geral é submetido a diversas condições ou em diferentes estágios de ciclo celular.

Um classificador baseado em expressão gênica produz uma lista de genes que podem indicar diferenças importantes do estado celular, de um único gene, ou de um grupo de genes. Em análise molecular de doenças, tais informações podem servir para diagnosticar a presença de uma anomalia e contribuir para a produção de um tratamento ou cura para tal a partir de intervenções gênicas por drogas [Dougherty et al., 2005].

Existem diversas abordagens para se realizar o reconhecimento de padrões. Dentre elas, este trabalho se encaixa na abordagem estatística em que cada padrão é represen-

tado por um vetor aleatório de n características $\mathbf{X} = (X_1, X_2, \dots, X_n)$ [Theodoridis and Koutroumbas, 1999, Duda et al., 2000]. Dado um conjunto de amostras de treinamento, o objetivo principal em reconhecimento estatístico de padrões é o de projetar um classificador que atribua um rótulo a um novo padrão com a menor margem de erro possível. Caso cada amostra desse conjunto de treinamento possua um rótulo Y conhecido *a priori*, trata-se de classificação supervisionada. Caso contrário, trata-se de classificação não-supervisionada [Watanabe, 1985]. Exemplos de classificação não-supervisionada são técnicas de agrupamento (*clustering*) [Jain et al., 1999]. Essas técnicas são bastante populares em estudos que utilizam expressões gênicas de *microarray*, já que muitos experimentos sugerem que genes distintos possuindo perfis de expressão similares podem ter funções similares na célula ou serem envolvidos numa mesma função biológica [Eisen et al., 1998]. Entretanto, esta tese concentra-se em classificação supervisionada.

Há três questões estatísticas importantes quando se trata de classificação baseada em expressões gênicas [Devroye et al., 1996]:

- Dado um conjunto de características, como projetar um classificador baseado em uma amostragem que realize uma boa classificação sobre a população geral?
- Como estimar o erro de um classificador nos casos em que há poucas amostras?
- Dado um grande conjunto de características (por exemplo, perfis de expressão gênica obtidos de *microarrays*), como selecionar um conjunto de características de qualidade para construir um bom classificador?

Em estudos envolvendo amostras de *microarrays*, existem milhares de características (genes) e algumas dezenas de amostras, o que dificulta sobremaneira o tratamento desses problemas [Dougherty, 2001]. Para evitar que o classificador seja super-ajustado ao conjunto de dados utilizado para treiná-lo (*overfitting*), deve-se projetar classificadores simples baseados em um pequeno conjunto de características que seja relevante de acordo com um determinado critério [Raudys and Jain, 1991, Devroye et al., 1996, Jain and Zongker, 1997, Vapnik, 1998]. Esta tese encaixa-se justamente nesse contexto.

2.2.1 Classificador Bayesiano

O conceito de classificador (ou preditor) Bayesiano é a base de todo o trabalho desta tese. Antes de introduzir esse conceito, é necessário introduzir noções básicas fundamentais de teoria da probabilidade e estatística. Como as contribuições desta tese se baseiam em dados discretos ou quantizados (não-contínuos), as definições a seguir são dadas apenas em termos de distribuições de probabilidades. Porém, cada uma dessas definições pode ser dada em termos de funções densidade de probabilidades (dados contínuos).

Probabilidade *a priori* de uma classe

A probabilidade *a priori* de uma classe $y \in Y$ ocorrer é denotada por $P(Y = y)$ (ou simplesmente $P(y)$), sendo definida como a probabilidade de y ocorrer na ausência de outras observações. Em um conjunto de c classes, $\sum_{y=0}^{c-1} P(Y = y) = 1$.

Probabilidade de um padrão

Seja $P(\mathbf{X} = \mathbf{x} | Y = y)$ (ou simplesmente $P(\mathbf{x} | y)$) a probabilidade de um padrão (ou instância) $\mathbf{x} \in \mathbf{X}$ ocorrer dado o conhecimento de que a classe y ocorreu. A probabilidade de ocorrência do padrão $\mathbf{x}$ é definida por:

$$P(\mathbf{x}) = \sum_{y=0}^{c-1} P(\mathbf{x} | y) P(y) \quad (2.1)$$

Probabilidade *a posteriori*

A probabilidade de uma classe ser y dado o conhecimento de um padrão $\mathbf{x}$ (probabilidade condicional ou *a posteriori*) é denotada por $P(y | \mathbf{x})$. Pela regra de Bayes, temos:

$$P(y | \mathbf{x}) = \frac{P(\mathbf{x} | y) P(y)}{P(\mathbf{x})} \quad (2.2)$$

Um classificador Bayesiano ψ recebe como entrada um novo padrão $\mathbf{x}$ e atribui um rótulo y^* tal que y^* seja o rótulo cuja probabilidade condicional $P(y^* | \mathbf{x})$ seja a maior

dentre as probabilidades condicionais de todas as classes ($P(y|\mathbf{x})$ para todo $y \in Y$). Formalmente ψ é definido como:

$$\psi_Y(\mathbf{x}) = y^* = \operatorname{argmax}_{y \in Y} P(y|\mathbf{x}) \quad (2.3)$$

Probabilidade conjunta

A probabilidade conjunta $P(\mathbf{x}, y)$, ou seja, a probabilidade de $\mathbf{x}$ ser associado à classe y , é obtida a partir do conhecimento da probabilidade condicional $P(y|\mathbf{x})$ e de $P(\mathbf{x})$:

$$P(\mathbf{x}, y) = P(y|\mathbf{x})P(\mathbf{x}) \quad (2.4)$$

A distribuição de probabilidades conjunta $P(\mathbf{X}, Y)$ oferece toda a informação de dependência entre $\mathbf{X}$ e Y . Um exemplo simples pode ser verificado na Tabela 2.1(a) que representa as probabilidades conjuntas de $\mathbf{X}$ e Y na qual $\mathbf{X} = \{X_1, X_2\} = \{0, 1\}^2$ e $Y = \{0, 1\}$. A tabela de probabilidades condicionais $P(Y|\mathbf{X})$ correspondente é mostrada na Tabela 2.1(b). Note que a tabela de probabilidades condicionais descreve apenas de forma parcial a relação de dependência entre $\mathbf{X}$ e Y , pois as probabilidades de ocorrência de cada possível padrão $P(\mathbf{x})$ não são explicitadas. Já a tabela de probabilidades conjuntas contém tanto as probabilidades condicionais como a probabilidade de ocorrência de cada padrão.

Taxa de erro

A probabilidade de erro de classificação ao associar $\mathbf{x}$ a uma classe y é dada por

$$\varepsilon_y(\mathbf{x}) = 1 - P(y|\mathbf{x}), y = 0, \dots, c - 1 \quad (2.5)$$

Portanto, o erro médio $\varepsilon_Y(\mathbf{X})$ de um classificador Bayesiano é a média ponderada de todos os erros $\varepsilon_{y^*}(\mathbf{x})$, $\forall \mathbf{x} \in \mathbf{X}$, ou seja:

$$\varepsilon_Y(\mathbf{X}) = \sum_{\mathbf{x} \in \mathbf{X}} P(\mathbf{x}) \varepsilon_{y^*}(\mathbf{x}), y^* = \operatorname{argmax}_{y \in Y} P(y|\mathbf{x}) \quad (2.6)$$

Tabela 2.1: Exemplo de tabelas de distribuição de probabilidades conjuntas e condicionais para $\mathbf{X} = \{X_1, X_2\} = \{0, 1\}^2$ e $Y = \{0, 1\}$. Note que, pela tabela de probabilidades conjuntas sabe-se que $P(X_1 = 0, X_2 = 0) = 0.4$, $P(X_1 = 0, X_2 = 1) = 0.25$, $P(X_1 = 1, X_2 = 0) = 0.1$ e $P(X_1 = 1, X_2 = 1) = 0.25$. Essas informações não estão presentes na tabela de probabilidades condicionais.

X_1	X_2	$P(X_1, X_2, Y = 0)$	$P(X_1, X_2, Y = 1)$
0	0	0.3	0.1
0	1	0.05	0.2
1	0	0.07	0.03
1	1	0.15	0.1

(a) probabilidades conjuntas

X_1	X_2	$P(Y = 0 X_1, X_2)$	$P(Y = 1 X_1, X_2)$
0	0	0.75	0.25
0	1	0.2	0.8
1	0	0.7	0.3
1	1	0.6	0.4

(b) probabilidades condicionais

O classificador Bayesiano que precisar inferir o valor de Y na ausência de qualquer outra observação deve atribuir o valor mais freqüentemente observado de Y , ou seja, $\psi_Y = y^* = \operatorname{argmax}_{y \in Y} P(y)$. Portanto, o erro cometido nessa situação (o erro *a priori*) é definido por:

$$\varepsilon_Y = 1 - \max_{y \in Y} P(y) \quad (2.7)$$

Como o classificador Bayesiano sempre atribui $\mathbf{x}$ à sua classe de maior probabilidade, tal classificador é estatisticamente ótimo. Porém, esse classificador só pode ser aplicado quando se conhece a distribuição de probabilidades conjuntas entre os padrões e as classes. Na prática, o projeto de um classificador se baseia em uma estimação dessas probabilidades a partir de um conjunto de dados de treinamento. Na maioria das situações, o conjunto de dados de treinamento disponível não é suficiente para obter uma boa estimação das probabilidades (em outras palavras, preencher a tabela de probabilidades conjuntas de forma satisfatória).

2.2.2 O problema da dimensionalidade

Um problema que existe ao usar o erro Bayesiano como erro de estimação é que essa medida de erro é monótona, ou seja, dados dois conjuntos de características $\mathbf{Z}$ e $\mathbf{X}$ tais

que $\mathbf{Z} \subseteq \mathbf{X}$, então $\varepsilon_Y(\mathbf{Z}) \geq \varepsilon_Y(\mathbf{X})$. Entretanto, se $\hat{\varepsilon}_Y(\mathbf{Z})$ e $\hat{\varepsilon}_Y(\mathbf{X})$ são os erros resultantes de classificadores projetados sobre um mesmo conjunto finito de amostras, então não é possível afirmar que $\hat{\varepsilon}_Y(\mathbf{Z}) \geq \hat{\varepsilon}_Y(\mathbf{X})$. De fato, o que se observa é que o erro do classificador em função do número de características que descrevem os padrões (dimensionalidade) forma uma “curva em U” (ver Figura 2.3) [Jain and Zongker, 1997]. Observando essa figura, constata-se que para as dimensões menores que d_1 , a adição de características implica em uma melhora no desempenho esperado do classificador. Entre as dimensões d_1 e d_2 , a inclusão de características passa a não causar qualquer impacto significativo em seu desempenho. O problema da dimensionalidade começa a ocorrer após o ponto d_2 em que novas características passam a afetar negativamente o desempenho esperado do classificador.

Figura 2.3: Gráfico das taxas de erro em função da dimensionalidade com número fixo de amostras ilustrando o problema da dimensionalidade. A curva do erro Bayesiano (erro do classificador ótimo) é dada por $\varepsilon_Y(\mathbf{X})$, enquanto a curva do erro esperado ao aplicar um classificador projetado a partir de um número finito de amostras é dada por $\hat{\varepsilon}_Y(\mathbf{X})$.

O número de amostras necessárias para que um classificador tenha um desempenho satisfatório é exponencial com relação à dimensão do vetor de características [Jain et al., 2000]. Isso pode ser verificado através do fato que a tabela de distribuição conjunta de probabilidades contém um número de linhas dado por $|R|^n$ em que $|R|$ é a cardinalidade do domínio R de valores das características e n é o número de características representantes dos padrões. Esse número de linhas representa a quantidade de configurações (instâncias) possíveis que os padrões podem possuir. Retomando o exemplo da Tabela 2.1, $|R| = 2$ e $n = 2$, resultando em $2^2 = 4$ instâncias possíveis. Caso seja acrescentada uma característica, o número de instâncias possíveis dobra ($2^3 = 8$). Para um conjunto de treinamento com um número fixo de amostras, cada característica adicionada torna a

tabela de frequências absolutas cada vez mais rarefeita, o que pode levar rapidamente a uma situação em que muitas instâncias são raramente ou nunca observadas. Essa situação configura o *overfitting*, ou seja, o classificador se torna excessivamente ajustado aos dados de treinamento.

Por conta do problema da dimensionalidade, existem inúmeras técnicas para reduzir a dimensão, ou seja, selecionar um subespaço do espaço de características que represente os padrões de forma satisfatória. Existem duas abordagens de redução de dimensionalidade: fusão (ou extração) de características e seleção de características. As técnicas de fusão criam novas características a partir de combinações das características originais, enquanto as de seleção devolvem um subconjunto dos atributos originais. O foco dos trabalhos originados desta tese é sobre seleção de características, como discutido a seguir.

2.3 Seleção de características

O problema de seleção de características consiste em selecionar um subconjunto de características que seja uma boa representação dos objetos em estudo. Os métodos de seleção de características consistem basicamente de duas partes principais: um algoritmo de otimização e uma função critério a ser otimizada. Esta seção apresenta uma visão geral das classes de métodos disponíveis para seleção de características e introduz dois métodos que embutem a estimação de erro diretamente em funções critério que se baseiam na estimação das probabilidades conjuntas entre os padrões e as classes (abordagem Bayesiana).

Os métodos de seleção de características podem ser divididos em dois grupos: filtros e *wrappers* [Kohavi and John, 1997]. Com as técnicas de filtro, as características são avaliadas de acordo com certas propriedades (correlação, informação mútua, poder discriminativo) sem levar em conta qual será o modelo de classificação [John et al., 1994, Xing et al., 2001]. A principal desvantagem das técnicas de filtro é que elas avaliam cada característica de forma independente com relação à predição da variável alvo (classe) [Dougherty et al., 2005]. Por exemplo, caso uma característica X_i seja um excelente preditor do comportamento da variável de classes, qualquer outra característica X_j que for bem correlacionada com X_i também fará uma boa predição dos rótulos. Características correlacionadas formam um conjunto redundante que, ao invés de melhorar a predição dos rótulos, pode fazer com que tal predição fique prejudicada. Outro exemplo, considerado

o oposto do anterior, é quando duas características X_i e X_j são irrelevantes isoladamente, mas que em conjunto realizam uma excelente predição. Tal conceito é caracterizado como predição intrinsecamente multivariada, sendo formalizado e intensamente discutido no Capítulo 4. As técnicas de filtro descartam tais características, que podem ser peças chave numa rede de regulação gênica [Martins-Jr et al., 2008b].

Por outro lado, com os algoritmos *wrappers*, a capacidade de predição de um classificador particular é investigada para todos os subconjuntos de genes ou para um subespaço do espaço total desses subconjuntos quando a busca exaustiva é computacionalmente inviável, o que freqüentemente ocorre. O subconjunto que obtiver a melhor avaliação de acordo com uma determinada função critério será a resposta do algoritmo *wrapper*.

2.3.1 Métodos *wrapper*

Um algoritmo do tipo *wrapper* considera o projeto de classificador como uma caixa preta e o utiliza para avaliar um determinado subconjunto de características. Em termos gerais, um método *wrapper* segue o fluxograma ilustrado na Figura 2.4. A cada passo o algoritmo seleciona um subconjunto de características, constrói um classificador com base nessas características e no conjunto de treinamento, e avalia o subconjunto com base no resultado da classificação do conjunto de teste, que pode ser o próprio conjunto de treinamento, um conjunto separado ou ambos (treinamento e teste). Essa avaliação depende da função critério empregada. No final do processo, caso não reste qualquer subconjunto a ser examinado, ou devido a algum outro critério de parada, o algoritmo devolve o melhor subconjunto dentre todos os examinados.

A busca pelo subconjunto ótimo é uma tarefa computacionalmente difícil em geral, já que para uma dimensão fixada d e n possíveis características, existem $\binom{n}{d}$ subconjuntos possíveis. Além disso, qualquer subconjunto de características pode ser um conjunto ótimo, ou seja, não há qualquer propriedade estrutural da solução ótima que ajude a restringir a busca [Cover and van Campenhout, 1977]. Conseqüentemente, apenas a busca exaustiva, que testa todas as possibilidades de subconjuntos, garante solução ótima. Em particular, o algoritmo “Melhores características individuais”, que devolve um subconjunto formado pelas melhores características individuais, não é apropriado, já que tal subconjunto pode estar relativamente distante da otimalidade. Como já discutido anteri-

Figura 2.4: Fluxograma de um algoritmo *wrapper* genérico. Os retângulos representam procedimentos que realizam uma certa atividade, os triângulos representam fontes de informação e as setas indicam fluxo de informação.

ormente nesta seção, duas características ruins podem dar origem a um par melhor do que um par formado por características boas (ver Capítulo 4 sobre predição intrinsecamente multivariada). Por causa disso e da ineficiência da busca exaustiva, existe uma repertório imenso de métodos de seleção de características propostos na literatura.

Métodos gulosos

Os métodos gulosos tentam adicionar as melhores características primeiro, porém a próxima característica a ser adicionada não é avaliada individualmente, mas sim em conjunto com as características já incluídas nos resultados intermediários [Jain and Zongker, 1997]. O método guloso mais simples é a Busca Sequencial para Frente (do inglês: *Sequential Forward Search* - SFS), que começa com um subconjunto resultado vazio e adiciona a melhor característica encontrada a esse subconjunto. Em seguida, adiciona uma segunda característica que, em conjunto com a primeira, forma o melhor par de características, e assim por diante. A Busca Sequencial para Trás (do inglês: *Sequential Backward Search* - SBS) é uma variante do SFS que começa com um conjunto completo e elimina sucessivamente as características menos relevantes de acordo com a função critério adotada.

Esses métodos são computacionalmente eficientes, mas os resultados obtidos podem

não ser satisfatórios, uma vez que uma característica é inserida (ou eliminada) não poderá mais ser eliminada do (inserida no) conjunto resposta. Dessa forma, pode ocorrer o *efeito nesting* em que uma característica já inserida (ou removida) do conjunto solução não pertence (ou pertence) ao conjunto ótimo. O fenômeno de predição intrinsecamente multivariada é uma das principais causas desse efeito (ver Capítulo 4).

Método Adiciona l - Remove r

Esse método (do inglês: *Plus l - Take Away r* , ou PTA) foi criado com o intuito de amenizar o *efeito nesting* [Somol et al., 1999]. A cada passo, esse algoritmo adiciona l elementos através do SFS e elimina r elementos através do SBS, sendo que l e r são parâmetros pré-definidos pelo usuário. O problema desse método é saber quais valores de l e r utilizar, sendo que o *efeito nesting* pode ser aliviado em maior ou menor grau dependendo dos parâmetros definidos.

Métodos de busca seqüencial generalizada

Esses algoritmos (do inglês: *Generalized Seqüencial Forward (Backward) Search* - GSFS e GSBS) inserem (no caso do GSFS) ou removem (no caso do GSBS) subconjuntos de características ao invés de características individuais [Somol et al., 1999]. Existe também a versão generalizada do PTA que acrescenta e remove subconjuntos de dimensão fixa. Apesar desses algoritmos reduzirem o *efeito nesting*, a seleção dos tamanhos dos subconjuntos a serem inseridos e removidos pode ser um fator que pode deixar o processo bastante lento. O problema é encontrar os parâmetros que promovem um equilíbrio entre desempenho e qualidade dos resultados.

Métodos flutuantes

Os métodos de busca seqüencial flutuante (do inglês: *Sequential Floating Forward Search* - SFFS e *Sequential Floating Backward Search* - SFBS) são generalizações do método PTA que determinam l e r dinamicamente [Pudil et al., 1994]. Esses métodos são computacionalmente eficientes e freqüentemente devolvem uma solução bastante próxima da ótima, sendo considerado um excelente algoritmo em termos de custo-benefício.

Existem também os métodos adaptativos (ASFFS e ASFBS) que adicionam e removem subconjuntos de tamanho determinado dinamicamente ao invés de manipular apenas características individuais como no caso do SFFS/SFBS [Somol et al., 1999]. Esses métodos devolvem resultados ainda melhores que o SFFS/SFBS, mas o custo computacional costuma ser bastante alto.

Métodos *branch-and-bound*

Os métodos *branch-and-bound* podem garantir solução ótima em termos de uma classe de funções critério monotônicas [Jain et al., 2000, Somol and Pudil, 2004]. Tal procedimento evita uma busca exaustiva através do uso de resultados intermediários, os quais são usados para obter limites nos valores mínimo e máximo da função critério adotada. Se a distribuição das probabilidades conjuntas entre os padrões e as classes for conhecida, dimensionalidades maiores implicam erros de classificação menores. Entretanto, como na prática essas distribuições são desconhecidas e devem ser estimadas, um problema com a adoção de algoritmos de *branch-and-bound* decorre justamente do fato de se basearem em funções critério monotônicas que não levam em conta os erros de estimação cometidos quando se considera padrões com alta dimensionalidade.

Na tentativa de contornar esse problema, foi proposto recentemente um novo algoritmo de *branch-and-bound* que se baseia em funções critério que embutem estimação de erro e formam “curvas em U”, refletindo o problema da dimensionalidade. Tal algoritmo, denominado *U-curve*, se baseia na exploração de uma estrutura de reticulado Booleano em que cada nó representa um subconjunto de características e possui um custo. Qualquer caminho que se percorre nesse reticulado a partir do conjunto vazio (nível 1) até o conjunto completo (nível n) dará origem a uma seqüência de custos que formará uma “curva em U” [Ris et al., 2008]. Esse método se originou de um produto de colaboração do autor desta tese com Marcelo Ris, recém doutor do curso de Doutorado de Bioinformática da USP. Resultados mostram que, para a maioria dos experimentos realizados que envolvem até 100 dimensões, o método “U-curve” atinge um resultado melhor que o algoritmo de Busca Sequencial Flutuante (SFFS) e em menor tempo. Tal trabalho abre caminho para uma série de desdobramentos futuros na tentativa de torná-lo mais eficiente, tanto em termos de consumo de memória como de tempo, para conjuntos de dimensionalidades maiores.

2.3.2 Funções critério

A escolha da função critério é muito importante para guiar os algoritmos na busca de um subconjunto de atributos que melhor descreva os padrões. Em seleção de características, o objetivo de uma função critério é o de minimizar o erro de classificação ou predição. Existem dois grupos de funções critério: as baseadas em distâncias entre as classes e as baseadas no desempenho do classificador.

Distâncias entre classes

As funções que se baseiam em distâncias entre classes são independentes do classificador adotado, levando em conta apenas a geometria dos pontos formada no espaço. Existem diversas medidas de distância, dentre as mais comuns são: entre os centróides das classes, entre vizinhos mais próximos, baseada em matrizes de espalhamento, de Mahalanobis, de Bhattacharyya e nebulosas [Theodoridis and Koutroumbas, 1999].

A distância é um operador entre dois elementos, portanto não podendo ser usado como medida de separabilidade entre três ou mais classes de uma maneira global. Uma alternativa para isso é usar critérios baseados em desempenho do classificador como discutidos a seguir.

Desempenho do classificador

As funções baseadas no desempenho do classificador construído sobre um subconjunto de características, diferentemente das baseadas em distância, podem avaliar a separabilidade entre duas ou mais classes de maneira global. O problema dessa abordagem é que, como as probabilidades conjuntas são estimadas a partir de um conjunto de treinamento, o subconjunto de características selecionado pode não ser satisfatório se o número de amostras for pequeno em relação à dimensionalidade (vide Seção 2.2.2). Na tentativa de contornar esse problema, propomos dois métodos que tentam embutir o erro de estimação diretamente na tabela de probabilidades conjuntas entre padrões e classes utilizada por funções critérios baseadas nessa tabela. Esses métodos atribuem uma penalidade às instâncias raramente ou nunca observadas. Embutimos esses métodos de penalização em duas funções critérios comumente utilizadas para seleção de características. A primeira (entropia condicional)

baseia-se na quantidade de informação que um subconjunto oferece com relação à variável de classes. A segunda (Coeficiente de Determinação não-linear) baseia-se na diminuição do erro do classificador ao utilizar um determinado subconjunto de características em relação ao erro *a priori*.

Entropia condicional média

A Teoria da Informação, originada pelos trabalhos de Shannon [Shannon and Weaver, 1963], pode ser empregada em problemas de seleção de características [Duda et al., 2000]. A entropia de Shannon de uma variável X é definida por:

$$H(X) = - \sum_{x \in X} P(x) \log P(x) \quad (2.8)$$

sendo adotado que $0 \cdot \log 0 = 0$. Essa definição se estende naturalmente considerando-se vetores aleatórios $\mathbf{X}$.

A entropia mede o grau de desordem de uma variável, ou seja, quanto maior a entropia de uma variável, mais caótico é o seu comportamento. A entropia condicional de uma variável Y dada uma instância $\mathbf{x} \in \mathbf{X}$ é definida de forma análoga:

$$H(Y|\mathbf{x}) = - \sum_{y \in Y} P(y|\mathbf{x}) \log P(y|\mathbf{x}) \quad (2.9)$$

A entropia condicional diz o quanto a variável $\mathbf{x}$ consegue prever o comportamento da variável Y . Quanto menor a entropia condicional de Y dado $\mathbf{x}$, melhor será a predição de Y através de $\mathbf{x}$. A Figura 2.5 ilustra dois histogramas, um para baixa entropia condicional e outro para alta entropia condicional de Y dado $\mathbf{x}$. O histograma da esquerda (baixa entropia condicional) configura o caso em que $\mathbf{x}$ realiza uma boa predição dos valores de Y , já que a distribuição de probabilidades condicionais apresenta massa altamente concentrada sobre $Y = 1$. Já o histograma da esquerda (alta entropia condicional) representa o caso em que $\mathbf{x}$ não é adequado para prever o comportamento de Y , pois a distribuição de probabilidades condicionais é próxima da uniforme.

A entropia condicional média é definida como a média ponderada das entropias condicionais de todas as possíveis instâncias $\mathbf{x} \in \mathbf{X}$. Sua equação é dada por:

Figura 2.5: O histograma da esquerda configura uma situação em que Y é bem predito por $\mathbf{x}$ porque a massa de probabilidades condicionais está bem concentrada em $Y = 1$ (entropia condicional baixa). Já para o histograma da direita, a massa de probabilidades está melhor distribuída ao longo das classes, o que faz com que o padrão $\mathbf{x}$ não seja um bom preditor de Y (entropia condicional alta).

$$H(Y|\mathbf{X}) = \sum_{\mathbf{x} \in \mathbf{X}} P(\mathbf{x}) H(Y|\mathbf{x}) \quad (2.10)$$

em que $H(Y|\mathbf{x})$ é a entropia condicional dada pela Equação 2.9. Quanto menor o valor de $H(Y|\mathbf{X})$, maior será o ganho de informação sobre Y através do conjunto de características $\mathbf{X}$. A informação mútua média é simplesmente a diferença da entropia *a priori* da entropia condicional média, ou seja:

$$IM(\mathbf{X}, Y) = H(Y) - H(Y|\mathbf{X}) \quad (2.11)$$

Coeficiente de Determinação

O Coeficiente de Determinação (do inglês: *Coefficient of Determination* - *CoD*) [Dougherty et al., 2000] é um critério útil para seleção de características [Hsing et al., 2005]. Sua equação é dada por:

$$CoD_Y(\mathbf{X}) = \frac{\varepsilon_Y - \varepsilon_Y(\mathbf{X})}{\varepsilon_Y} \quad (2.12)$$

em que ε_Y é o erro *a priori* definido na Equação 2.7 e $\varepsilon_Y(\mathbf{X})$ é o erro médio do classificador

ótimo (Bayesiano) ao prever Y tomando como entrada todas as instâncias possíveis de $\mathbf{X}$ (Equação 2.6). O CoD quantifica a diminuição do erro *a priori* na observação das instâncias de $\mathbf{X}$. Quanto maior o CoD, maior será essa diminuição ($CoD = 0$ significa que $\mathbf{X}$ não melhora o desempenho do classificador ótimo em relação ao erro *a priori*, enquanto $CoD = 1$ indica que Y é completamente determinado por $\mathbf{X}$).

Tanto a entropia condicional média como o CoD, tais como são definidas nas suas respectivas Equações 2.10 e 2.12, são funções monotônicas, ou seja, um seletor de características orientado por uma dessas duas funções sempre preferirá incluir características. Para evitar o problema da dimensionalidade, propomos os dois métodos de penalização a seguir.

Penalização das instâncias não observadas

Para embutir o erro de estimação devido a vetores de características de alta dimensão e número insuficiente de amostras, uma alternativa é envolver as instâncias não observadas no cálculo da função critério [Martins-Jr et al., 2006a, Lopes et al., 2008b]. Uma massa positiva de probabilidades é atribuída às instâncias não observadas, sendo que o valor da função critério correspondente a elas torna-se igual ao valor da função sem qualquer observação adicional (valor da função para o conhecimento *a priori* de Y).

No caso da entropia condicional média (Equação 2.10), as instâncias não observadas recebem a entropia da distribuição *a priori* de Y ($H(Y)$), enquanto para o caso do CoD elas recebem o valor do erro *a priori* (ε_Y). A massa de probabilidades para as instâncias não observadas é parametrizada por α . Este parâmetro é adicionado à frequência absoluta (número de observações) de todas as instâncias possíveis. Assim, a entropia condicional média com esse tipo de penalização é redefinida por:

$$H(Y|\mathbf{X}) = \frac{1}{\alpha M + s} \left[\alpha(M - N)H(Y) + \sum_{i=1}^N (f_i + \alpha)H(Y|\mathbf{X} = \mathbf{x}_i) \right] \quad (2.13)$$

em que M é o número de instâncias possíveis do vetor de características $\mathbf{X}$, N é o número de instâncias observadas (assim, o número de instâncias não-observadas é dada por $M - N$), f_i é a frequência absoluta (número de observações) de $\mathbf{x}_i$ e s é o número de amostras do conjunto de treinamento. As amostras $\mathbf{x}_i$ para $1 \leq i \leq N$ são aquelas observadas

(assim, as não observadas são $\mathbf{x}_j$ para $N + 1 \leq j \leq M$).

Já o CoD torna-se:

$$CoD_Y(\mathbf{X}) = \frac{\varepsilon_Y - \frac{1}{\alpha M + s} \left[\alpha(M - N)\varepsilon_Y + \sum_{i=1}^N (f_i + \alpha)(1 - \max_{y \in Y} P(y|\mathbf{x}_i)) \right]}{\varepsilon_Y} \quad (2.14)$$

A Tabela 2.2 mostra um exemplo contendo uma tabela de frequências absolutas dos valores de (X_1, X_2, X_3, Y) e suas respectivas entropias condicionais e erros *a posteriori*, com e sem penalização de instâncias não observadas, com $\alpha = 1$. A última linha da tabela (“Total”) fornece as respectivas entropias condicionais médias e os erros *a posteriori* médios.

Penalização de instâncias raramente observadas

Nesta penalização, as instâncias que não foram observadas não são levadas em conta. Tal penalização consiste em modificar as distribuições de probabilidades condicionais que têm apenas uma única observação [Barrera et al., 2006a]. Isso faz sentido porque se uma determinada instância $\mathbf{x}$ tiver apenas uma observação, o valor de Y é completamente determinado (ou seja, $H(Y|\mathbf{x}) = 0$ e $CoD_Y(\mathbf{x}) = 1$), mas a confiança de que essa é de fato a real distribuição de $P(Y|\mathbf{X} = \mathbf{x})$ é muito baixa. Um parâmetro β representa um valor de confiança. Dado que $Y = y$ é a classe observada para um padrão $\mathbf{x}$ observado uma vez, a confiança em y sendo o verdadeiro valor de Y para o padrão considerado pode ser modelada por um parâmetro β entre 0 e 1. A idéia é distribuir a massa dada por $1 - \beta$ igualmente para todo $\{Y \neq y|\mathbf{X} = \mathbf{x}\}$ e atribuir a massa β a $\{Y = y|\mathbf{X} = \mathbf{x}\}$. Em [Barrera et al., 2006a] (ver Capítulo 3), o valor β adotado foi de $\frac{1}{c}$ em que c é o número de classes (cardinalidade de Y), isto é, todas as instâncias observadas uma única vez tiveram distribuições de probabilidades condicionais uniformes (confiança nula no valor observado). Já em [Lizier et al., 2008], foi adotado $\beta = 0.9$, o que significa uma confiança de 90% no valor observado.

Adaptando essa penalização à Equação 2.10, a entropia condicional torna-se:

$$H(Y|\mathbf{x}) = \frac{M - N}{s} H((\Delta(0), \Delta(1), \dots, \Delta(c - 1))) + \sum_{\mathbf{x} \in \mathbf{X}: P(\mathbf{x}) > \frac{1}{s}} P(\mathbf{x}) H(Y|\mathbf{x}), \quad (2.15)$$

Tabela 2.2: Tabela de frequências absolutas de (X_1, X_2, X_3, Y) e suas respectivas entropias condicionais e erros de predição, com e sem penalização de instâncias raramente observadas para $\alpha = 1$. As frequências absolutas para $Y = 0$ e $Y = 1$ são indicadas respectivamente por $f(Y = 0)$ e $f(Y = 1)$. $f(\mathbf{x})$ e $f'(\mathbf{x})$ são as frequências absolutas da instância $\mathbf{X} = \mathbf{x}$ sem e com penalização, respectivamente. $H(Y|\mathbf{x})$ e $H'(Y|\mathbf{x})$ são as entropias condicionais sem e com penalização, respectivamente. $\varepsilon_Y(\mathbf{x})$ e $\varepsilon'_Y(\mathbf{x})$ são os erros *a posteriori* sem e com penalização, respectivamente. A última linha da tabela, denotada por “Total”, indica as frequências absolutas totais e as médias ponderadas das entropias condicionais e dos erros *a posteriori*. Para este exemplo, $H(Y) = 0.9932$, $\varepsilon_Y = 0.4516$ e os CoDs sem e com penalização valem 0.3577 e 0.2892, respectivamente.

X_1	X_2	X_3	$f(Y = 0)$	$f(Y = 1)$	$f(\mathbf{x})$	$f'(\mathbf{x})$	$H(Y \mathbf{x})$	$H'(Y \mathbf{x})$	$\varepsilon_Y(\mathbf{x})$	$\varepsilon'_Y(\mathbf{x})$
0	0	0	7	3	10	11	0.8813	0.8813	0.3	0.3
0	0	1	1	0	1	2	0	0	0	0
0	1	0	0	0	0	1	0	0.9932	0	0.4516
0	1	1	1	4	5	6	0.7219	0.7219	0.2	0.2
1	0	0	4	8	12	13	0.9183	0.9183	0.3333	0.3333
1	0	1	0	1	1	2	0	0	0	0
1	1	0	1	1	2	3	1	1	0.5	0.5
1	1	1	0	0	0	1	0	0.9932	0	0.4516
Total			14	17	31	39	0.8207	0.8843	0.2903	0.3210

em que $\Delta : \{0, 1, \dots, c - 1\} \rightarrow [0, 1]$ é a distribuição de probabilidades dada por

$$\Delta(i) = \begin{cases} \beta, & \text{para } i = y \\ \frac{1-\beta}{c-1}, & \text{para } i \neq y \end{cases},$$

s é o número total de amostras e N é o número de instâncias $\mathbf{x}$ tais que $P(\mathbf{x}) > \frac{1}{s}$ (mais de uma observação).

Como $\varepsilon_Y(\mathbf{x}) = 1 - \beta$ considerando $\beta \geq 1 - \beta$ (afinal, não faz sentido atribuir massas de probabilidade maiores para as classes não observadas do que para a classe observada) para $P(\mathbf{x}) = \frac{1}{s}$, o *CoD* com essa penalização é dado por:

$$CoD_Y(\mathbf{X}) = \frac{\varepsilon_Y - \left[\frac{(M-N)}{s}(1 - \beta) + \sum_{\mathbf{x} \in \mathbf{X}: P(\mathbf{x}) > \frac{1}{s}} P(\mathbf{x})(1 - \max_{y \in Y} P(y|\mathbf{x})) \right]}{\varepsilon_Y} \quad (2.16)$$

A Tabela 2.3 mostra um exemplo contendo a mesma tabela de freqüências absolutas dos valores de (X_1, X_2, X_3, Y) apresentada na Tabela 2.2 e suas respectivas entropias condicionais e erros *a posteriori*, com e sem penalização de instâncias observadas uma única vez, com $\beta = 0.8$. A última linha da tabela (“Total”) fornece as respectivas entropias condicionais médias e os erros *a posteriori* médios.

Tabela 2.3: Tabela de freqüências absolutas de (X_1, X_2, X_3, Y) e suas respectivas entropias condicionais e erros de predição, com e sem penalização de instâncias observadas uma única vez para $\beta = 0.8$. As freqüências absolutas para $Y = 0$ e $Y = 1$ e $\mathbf{X} = \mathbf{x}$ são indicadas respectivamente por $f(Y = 0)$, $f(Y = 1)$ e $f(\mathbf{x})$. $H(Y|\mathbf{x})$ e $H'(Y|\mathbf{x})$ são as entropias condicionais sem e com penalização, respectivamente. $\varepsilon_Y(\mathbf{x})$ e $\varepsilon'_Y(\mathbf{x})$ são os erros *a posteriori* sem e com penalização, respectivamente. A última linha da tabela, denotada por “Total”, indica as freqüências absolutas totais e as médias ponderadas das entropias condicionais e dos erros *a posteriori*. Para este exemplo, $\varepsilon_Y = 0.4516$ e os CoDs sem e com penalização valem 0.3517 e 0.3428, respectivamente.

X_1	X_2	X_3	$f(Y = 0)$	$f(Y = 1)$	$f(\mathbf{x})$	$H(Y \mathbf{x})$	$H'(Y \mathbf{x})$	$\varepsilon_Y(\mathbf{x})$	$\varepsilon'_Y(\mathbf{x})$
0	0	0	7	3	10	0.8813	0.8813	0.3	0.3
0	0	1	1	0	1	0	0.7219	0	0.1
0	1	0	0	0	0	0	0	0	0
0	1	1	1	4	5	0.7219	0.7219	0.2	0.2
1	0	0	4	8	12	0.9183	0.9183	0.3333	0.3333
1	0	1	0	1	1	0	0.7219	0	0.1
1	1	0	1	1	2	1	1	0.5	0.5
1	1	1	0	0	0	0	0	0	0
Total			14	17	31	0.8207	0.8673	0.2903	0.2968

2.4 Redes de regulação gênica

Grande parte dos processos celulares envolvem uma variedade de moléculas diferentes. O metabolismo de uma célula consiste de muitas reações interligadas. Produtos de uma reação podem servir de matéria prima para outras reações, formando uma rede metabólica. De forma similar, moléculas de sinalização são interligadas, formando uma rede de sinalização. E o mesmo ocorre para relações entre genes e seus produtos. Todos esses mecanismos são fortemente relacionados [Schlitt and Brazma, 2007]. O foco desta tese é sobre redes de transcrição de mRNA (expressão gênica), embora muitos princípios sejam válidos para outras redes.

As tecnologias atuais de medidas de expressão gênica, tais como as baseadas em *microarray*, permitem o estudo de aspectos de redes de regulação gênica de uma maneira global. Tais experimentos extraem os estados de um sistema regulatório de um determinado organismo submetido a condições de interesse, gerando um conjunto de dados contendo medidas de expressão. Esses dados são submetidos a comparações com predições geradas por modelos matemáticos construídos a partir de algum conhecimento biológico sobre o mecanismo regulatório. Esses modelos vão sendo constantemente corrigidos e aperfeiçoados para que as predições sejam cada vez mais fidedignas com a realidade representada pelos dados experimentais. A Figura 2.6 mostra um fluxograma que ilustra esse processo.

Figura 2.6: Fluxograma adaptado de [Jong, 2002] que ilustra o processo de análise de redes de regulação gênica. Os retângulos representam procedimentos que realizam uma certa atividade, os triângulos representam fontes de informação e as setas indicam fluxo de informação.

A modelagem de redes de regulação gênica é uma área em constante evolução, com novidades sendo publicadas quase diariamente [Kelemen et al., 2008]. Esta seção apresenta uma breve revisão baseada principalmente nos artigos *surveys* de Schlitt e Brazma [Schlitt and Brazma, 2007] e de De Jong [Jong, 2002] e nos livros de Dougherty *et al* [Dougherty et al., 2005] e Kelemen *et al* [Kelemen et al., 2008]. Tal revisão discute avanços recentes, limitações e desafios de modelagem. Embora a função desta seção seja a de mostrar um panorama do que se faz em pesquisas envolvendo redes de regulação gênica, esta tese se concentra em modelos de redes gênicas probabilísticas (PGN), incluindo redes Booleanas probabilísticas (PBN).

Diversas abordagens foram propostas para modelagem e simulação de redes gênicas. De acordo com [Schlitt and Brazma, 2007], essas abordagens podem ser agrupadas em quatro categorias principais ou níveis de organização: lista de partes, modelos topológicos, modelos de controle lógico e modelos dinâmicos.

2.4.1 Lista de partes

Uma lista de partes é uma coleção, descrição e automação dos elementos da rede de um organismo ou um sistema biológico [Schlitt and Brazma, 2007]. Tais elementos podem ser fatores de transcrição, promotores e sítios de ligação para fatores de transcrição. A montagem de uma lista de partes é a primeira etapa no desenvolvimento de um modelo complexo. Tais listas podem ser o resultado de um projeto de anotação e seqüenciamento genômico em que a seqüência de DNA de um organismo é determinada e grande parte das proteínas e dos genes são identificados. Essas listas também podem ser representadas como um banco de dados de elementos regulatórios ou funções desempenhadas por um conjunto de genes. Uma vez conhecidos os fatores de transcrição e seus sítios de ligação, é possível descrever a topologia de uma rede regulatória.

2.4.2 Modelos topológicos

Os modelos topológicos descrevem as conexões entre as partes. As topologias (ou arquiteturas) de redes regulatórias podem ser representadas por grafos cujos nós correspondem a genes, e cujos arcos (ou arestas) representam as interações regulatórias. Representações por grafos têm sido usados para vários conjuntos de dados biológicos, desde re-

des de interação entre proteínas até redes de expressão gênica [Albert and Barabási, 2002, da Silva and Stumpf, 2005].

A representação de uma rede regulatória pode ser generalizada de diversas maneiras [Jong, 2002]. As arestas e os nós podem ser rotulados de maneira a disponibilizar informação sobre a funcionalidade de cada gene e os tipos de interações existentes entre os genes. Um exemplo simples de rede de regulação gênica pode ser visto na Figura 2.7 em que existem 3 genes e os arcos indicam se a interação é de ativação ($\rightarrow$) ou de inibição ($\neg$).

Figura 2.7: Grafo orientado representando uma rede de regulação gênica. Os arcos indicados por $\rightarrow$ e $\neg$ representam interações de ativação e inibição, respectivamente.

Existe uma série de operações em grafos que podem ser realizadas para obter predições biologicamente relevantes sobre sistemas regulatórios [Jong, 2002]. Uma busca por caminhos entre dois genes, por exemplo, pode revelar interações regulatórias ausentes ou prover pistas sobre redundância na rede. Além disso, ciclos podem indicar relações de retroalimentação que são importantes para a homeostase e a diferenciação [Thomas, 1998]. Características de conectividade global de uma rede, tais como a média e a distribuição do número de reguladores por gene, podem indicar a complexidade da rede. Subgrafos fracamente conectados indicam módulos funcionais do sistema regulatório que podem ser analisados isoladamente. Análises de genes altamente conectados (*hubs*) mostram que tais genes fazem com que as redes tornem-se mais tolerantes a falhas aleatórias dos elementos da rede [Albert et al., 2001]. Nessa mesma linha, examina-se repetições de elementos estruturais locais (*motifs*) que ocorrem com frequência em dados reais, mas que não ocorrem em redes aleatórias, o que inclui ciclos de retroalimentação [Milo et al., 2002]. Pode-se ainda comparar redes regulatórias de organismos distintos e verificar características evo-

lucionárias que se mantêm conservadas em ambos [Jong, 2002].

No entanto, uma das principais razões para se estudar a topologia da rede é a de construir modelos mais detalhados de redes gênicas [Schlitt and Brazma, 2007]. Antes de qualquer modelo de lógica ou dinâmica de rede poder ser construído, é necessário saber quais genes interagem entre si e quais são mutualmente independentes. A existência de módulos (subredes que são relativamente isoladas do resto da rede) em sistemas biológicos é considerada uma premissa válida [Hartwell et al., 1999]. Entretanto, não existe uma definição precisa de módulo. Em uma representação por grafos, é natural definir um módulo como um componente “relativamente” isolado. Vários métodos foram propostos para identificar módulos como grupos de genes coexpressos sob condições específicas. Vias biologicamente significativas são usadas ocasionalmente para definir módulos e implementar uma abordagem reducionista, porém nem sempre é possível identificar as condições para as quais essas vias se interagem [Schlitt and Brazma, 2007].

Pode-se concluir que representações topológicas são úteis para fazer muitas observações interessantes a respeito das propriedades de redes, embora encontrar módulos bem definidos de tais redes no nível topológico seja um desafio considerável.

2.4.3 Modelos de lógicas de controle

Uma vez que a rede topológica é conhecida, o próximo passo é o estudo das regras de interação entre os elementos da rede. Por exemplo, se um promotor consiste de apenas um sítio de ligação para um fator de transcrição, podemos querer saber se tal fator é um ativador ou inibidor. Se vários fatores de transcrição se ligam a um promotor, saber o que cada fator faz e como eles se interagem pode ser importante. A Figura 2.8 ilustra um circuito lógico Booleano entre três genes de entrada que produz como saída o resultado do quarto gene. Estudos biológicos demonstram que alguns promotores apresentam comportamento combinatório aproximado por funções Booleanas (AND, OR, NOT e combinações dessas lógicas), mas em certos casos a interação é mais complexa [Louis and Becskei, 2002].

Funções Booleanas, árvores de decisão, redes de relevância e distribuições de probabilidades Bayesianas são freqüentemente usadas para descrever as lógicas da rede. Esses métodos podem ser distinguidos em discretos ou contínuos [Barrera et al., 2002, D’haeseleer, 2000]. Funções discretas são baseadas na hipótese de que um gene pode es-

Figura 2.8: Circuito lógico Booleano entre três genes de entrada produzindo como resultado o valor do quarto gene (Adaptado de [Schlitt and Brazma, 2007]). O quarto gene é ativado apenas quando os genes 1 e 2 estão ativos e o gene 3 está inativo, ou seja, $X_4 = [(X_1 \text{ AND } X_2) \text{ AND } \bar{X}_3]$.

tar em um número finito de estados. No caso mais simples, apenas dois estados descrevem o estado de um gene: expresso e não expresso. Dessa forma, adota-se funções Booleanas para descrever as interações entre genes. Essas funções são simplificações de redes gênicas, já que expressões gênicas e a relação entre eles assumem valores contínuos (por exemplo, concentração de moléculas varia de forma suave). Entretanto, redes Booleanas podem ser um bom ponto de partida para modelagem de redes gênicas [Kauffman, 1969].

Funções contínuas usam valores reais para representar a atividade do gene. Uma das abordagens mais utilizadas nessa categoria é a função linear, que considera um gene alvo como sendo uma combinação linear de seus preditores, cada um deles com um peso diferente e um sinal que pode ser negativo ou positivo. A topologia da rede determina quais desses preditores possuem peso 0 (genes não interconectados). Assim como as funções Booleanas, funções lineares são aproximações da realidade. Por exemplo, é impossível modelar uma situação em que o mesmo gene pode exercer uma função de ativador ou repressor de um outro gene, dependendo da presença ou ausência de outros genes (relações não-lineares). As relações não-lineares são melhor modeladas por funções Booleanas.

Árvores de decisão são modelos preditivos que permitem prever o comportamento de um gene particular (nó folha da árvore) baseados em dados de expressão de outros genes (nós internos da árvore) [Soinov et al., 2003]. Nesta abordagem, pode-se utilizar valores de expressão contínuos diretamente e as previsões com relação aos genes alvo (folhas) podem ser tanto discretas como contínuas.

As abordagens de redes de relevância consideram a similaridade entre pares de genes

no domínio contínuo [Butte and Kohane, 2000, Kelemen et al., 2008]. Para definir a topologia da rede, todos os pares de genes são comparados usando medidas de similaridade tais como coeficientes de correlação ou informação mútua. Em seguida, filtra-se as comparações para determinar conexões relevantes correspondendo a associações positivas ou negativas entre genes. Um método representativo dessa abordagem é o ARACNe (*Algorithm for the Reconstruction of Accurate Cellular Networks*), que se baseia em informação mútua para identificar genes relacionados [Margolin et al., 2006]. As desvantagens dessas abordagens é que elas, além de não inferirem causalidade, podem perder relações importantes entre múltiplos genes pelo fato de se basearem apenas em similaridade entre pares de sinais de expressão.

Redes Bayesianas provêm uma abordagem probabilística para modelagem de redes de regulação gênica [Friedman et al., 2000, Pe'er et al., 2001, Pournara and Wernisch, 2004]. Essas redes são representadas por um grafo orientado acíclico em que cada nó representa um gene e cada arco uma dependência. Para cada nível de expressão do gene i , denotado por X_i , existe uma distribuição condicional de probabilidades $P(X_i|\rho(X_i))$ em que $\rho(X_i)$ corresponde ao conjunto dos preditores diretos de i . Uma propriedade importante é que a distribuição de probabilidades conjuntas $P(\mathbf{X})$ de todos os nós de uma rede Bayesiana é dada pela multiplicação das probabilidades condicionais de cada gene dados os seus preditores, ou seja, $P(\mathbf{X}) = \prod_{i=1}^n P(X_i|\rho(X_i))$ (hipótese de Markov). A Figura 2.9 ilustra um exemplo de rede Bayesiana com todas as probabilidades condicionais associadas e a expressão que fornece a probabilidade conjunta para essa rede.

Figura 2.9: Rede Bayesiana com todas as probabilidades condicionais associadas e a correspondente expressão da probabilidade conjunta de $\mathbf{X} = [X_1, X_2, X_3, X_4, X_5]$.

Abordagens de rede Bayesiana para modelagem de redes de regulação gênica são atraentes por causa de sua sólida base estatística, o que permite lidar com aspectos estocásticos de expressão gênica, e medidas ruidosas, de maneira natural [Jong, 2002]. Além disso, redes Bayesianas podem ser empregadas quando existe apenas um conhecimento parcial do sistema. No entanto, esse modelo possui a desvantagem de não levar em conta os aspectos dinâmicos, embora essa desvantagem possa ser contornada através de generalizações que permitam relações de retroalimentação (ciclos) serem modeladas [Murphy and Mian, 1999]. Na próxima seção, serão discutidos modelos que levam em conta a dinâmica de redes.

2.4.4 Modelos dinâmicos

O conhecimento de lista de pares de uma rede, sua topologia e as lógicas de controle são requisitos necessários para expandir o modelo de forma a capturar mudanças dinâmicas ao longo do tempo [Schlitt and Brazma, 2007]. Tipicamente, esses modelos levam em conta um número relativamente pequeno de genes. O objetivo desses modelos é descrever e simular as mudanças dinâmicas do estado do sistema e prever a resposta da rede a vários estímulos e mudanças ambientais.

Existem diversos modelos dinâmicos propostos, sendo que Greller e Somogyi propuseram os seguintes atributos de divisão entre eles [Greller and Somogyi, 2002]:

- medidas quantitativas ou qualitativas (exemplo: Booleanas ou abundâncias);
- transições de estado determinísticas ou probabilísticas (exemplo: equações diferenciais ou cadeias de Markov);
- descrição geral do sistema determinístico ou estatístico (exemplo: campo vetorial ou distribuições de probabilidade de redes de crença Bayesianas);
- estado contínuo ou discreto;
- interações entre elementos e regras de atualização de estado não-lineares ou lineares;
- alta ou baixa dimensionalidade;
- forte ou fraca estocasticidade.

A seguir, são brevemente discutidos os principais modelos dinâmicos utilizados em redes gênicas.

Redes Booleanas

O modelo de redes Booleanas vem sendo empregado em redes de regulação gênica desde os anos 60 [Kauffman, 1969]. Redes Booleanas são baseadas na hipótese de que os valores ativo/inativo dos elementos podem descrever aspectos importantes da regulação gênica. Nesse modelo, o estado da rede é representado por uma sequência de n 0s e 1s indicando quais genes estão expressos ou não em um dado instante de tempo. A medida que o tempo prossegue, a rede transita pelo espaço de estados como mostrado na Figura 2.10. Para uma rede de n genes, existe um total de 2^n estados possíveis. Alguns estados podem nunca ser atingidos. Outros estados, denominados atratores, podem ser atingidos periodicamente, dependendo da condição inicial do sistema. No exemplo ilustrado na Figura 2.10(D) existem dois atratores: um deles é composto por um único estado (0, 0, 1) e o outro é composto por dois estados que se alternam: (1, 0, 1) e (0, 1, 0).

Figura 2.10: Exemplo de uma rede Booleana consistindo de 3 genes X_1 , X_2 , X_3 . (A) regras lógicas de transição de estado; (B) grafo de relações, dependências ou predições; (C) tabela de todos os possíveis estados (X_1 , X_2 , X_3) e as respectivas transições de estado para o instante de tempo posterior; (D) grafo representando as transições de estado. (Adaptado de [Schlitt and Brazma, 2007]).

O modelo de redes lógicas generalizadas (do inglês: *Generalized Logical Networks* - GLN) generaliza o modelo de redes Booleanas de forma a permitir não só que os elementos

assumam um número arbitrário de valores, mas também que as transições de estado ocorram de forma assíncrona [Thomas et al., 1995].

O modelo de redes Booleanas permite a introdução de estocasticidade, ou seja, os arcos das transições de estado podem ter probabilidades associadas. Dois modelos propostos nesse sentido são as redes ruidosas [Akutsu et al., 2000b] e as redes Booleanas probabilísticas (do inglês: *Probabilistic Genetic Networks* - PBN) [Shmulevich et al., 2002]. Existem evidências de que os modelos baseados em PBN podem revelar redes gênicas regulatórias biologicamente relevantes, podendo ser usadas para prever os efeitos da intervenção de um gene alvo [Dougherty et al., 2005]. O modelo de redes gênicas probabilísticas (PGN) é uma generalização do modelo PBN em que os elementos podem assumir dois ou mais valores [Barrera et al., 2006a]. Outra possível generalização de redes Booleanas é o modelo de redes de lógicas nebulosas (do inglês: *Fuzzy Logic Networks* - FLN), o qual é capaz de superar as restrições não realistas dos valores Booleanos (ligado/desligado) [Woolf and Wang, 2000, Cao et al., 2005]. Silvescu e Honavar descrevem uma generalização de redes Booleanas que lidam com dependências temporais entre genes por mais de uma unidade de tempo [Silvescu and Honavar, 1997].

Redes de Petri

Redes de Petri são uma extensão de modelos de grafos que permitem representações quantitativas simples de processos dinâmicos [Pinney et al., 2003, Hardy and Robillard, 2004, Moore et al., 2005]. Geralmente são grafos orientados consistindo de arcos e dois diferentes tipos de nós: nós de posição e de transição. Os arcos conectam apenas nós de posição a nós de transição e vice versa. O aspecto dinâmico é introduzido por *tokens* (bastões). Os nós de posição podem conter bastões. Cada arco tem um valor que indica o número de bastões necessários para realizar a transição através dele. E os nós de transição determinam a taxa de troca ao longo do caminho. Redes de Petri são particularmente úteis para modelagens de reações metabólicas, pois não necessitam de informação detalhada sobre as taxas das reações. Isso representa uma vantagem porque essas taxas são usualmente desconhecidas, embora sejam importantes para a descrição completa de certas redes metabólicas [Schlitt and Brazma, 2007].

Equações diferenciais

Redes Booleanas podem revelar importantes propriedades de rede, mas são insuficientes para capturar certos aspectos de sua dinâmica [Schlitt and Brazma, 2007]. Nesse sentido, modelos de equações diferenciais permitem descrições mais detalhadas da dinâmica da rede pela modelagem das alterações de concentração de moléculas ao longo do tempo [Chen et al., 1999, Maki et al., 2001]. Basicamente, a regulação gênica é modelada por equações da forma:

$$\frac{dx_i}{dt} = \mathcal{F}_i(\mathbf{x}), i \in \{1, \dots, n\}, \quad (2.17)$$

em que $\mathbf{x} = (x_1, x_2, \dots, x_n)$ é um vetor de expressões gênicas e $\mathcal{F}_i : \Re^n \rightarrow \Re$ uma função de controle lógico que pode ser linear ou não-linear (geralmente não-linear [Dougherty et al., 2005, Jong, 2002]).

Uma desvantagem em usar modelos de equações diferenciais no contexto de inferência de redes de regulação gênica é que o número de parâmetros que necessitam ser estimados é relativamente grande. Nesse caso, é necessário um conjunto de dados com alta resolução temporal (muitos instantes de tempo), o que dificilmente ocorre em problemas de bioinformática de um modo geral. Esta é a principal razão pela qual esses modelos são aplicados apenas a redes bem pequenas com poucos componentes. Outra razão reside na qualidade dos dados experimentais. Dados de *microarray*, por exemplo, geralmente contêm muito ruído e podem não ser apropriados para realização de modelagem quantitativa. Para isso, freqüentemente recorre-se à inclusão de conhecimento biológico *a priori* ou simplificações [Kelemen et al., 2008].

Modelos híbridos

Sistemas biológicos apresentam aspectos contínuos e discretos [Schlitt and Brazma, 2007]. Em geral, as concentrações são valores contínuos enquanto a ligação de um fator de transcrição ao DNA é expresso como um evento discreto (ligado ou não ligado). Entretanto, as fronteiras entre os aspectos contínuos e discretos dependem do nível de detalhe do modelo adotado. Modelos híbridos vêm sendo desenvolvidos na tentativa de descrever ambos os aspectos [Akutsu et al., 2000a].

Existem extensões do modelo de redes de Petri que permitem a inclusão de conhecimento sobre as dinâmicas de reações. Uma dessas extensões inclui atrasos de tempo estocásticos nas transições [Goss and Peccoud, 1998]. Nessas redes, os nós de transição disparam dependendo não só do número de bastões, como também de um componente estocástico. Outro tipo de extensão é o modelo de redes de Petri híbridas funcionais (do inglês: *Hybrid Functional Petri Nets* - HFPN) em que os nós de posição e de transição são contínuos [Matsuno et al., 2006]. O problema de usar HFPN é que a principal vantagem de usar redes de Petri sobre equações diferenciais (não depender da informação das taxas de reação) acaba sendo perdida. Por outro lado, HFPN permite uma mistura de nós contínuos e discretos, o que possibilita que sejam incluídas apenas informações sobre taxas de reação conhecidas [Schlitt and Brazma, 2007].

Modelo linear de estado finito

O modelo linear de estado finito (do inglês: *Finite State Linear Model* - FSLM) combina a simplicidade e o baixo custo computacional das redes Booleanas com a representação contínua dos valores dos elementos [Ruklisa et al., 2005]. Nesse modelo, a atividade dos genes é representada por dois estados (superexpresso e subexpresso), porém os níveis de expressão são representados por números reais. O tempo é contínuo e o estado da rede determina as taxas de mudança de concentração, enquanto o próprio estado é afetado por essas concentrações. Esse modelo pode ser empregado para construir modelos complexos, embora não seja possível construir tais modelos nos casos em que as redes de regulação são relativamente grandes por falta de informação biológica detalhada [Schlitt and Brazma, 2007].

2.4.5 Identificação de redes gênicas

A identificação de redes de regulação gênica, freqüentemente denominada “engenharia reversa”, é um processo que se inicia a partir de um conjunto de dados com o objetivo de projetar um modelo que se adapta aos dados de forma automática ou semi-automática, sem hipótese *a priori* sobre o sistema biológico [Schlitt and Brazma, 2007]. O modelo derivado dos dados é julgado por resultados de simulações. Por exemplo, pode-se usar um conjunto de dados de expressão gênica para construir um modelo de redes que seja

consistente com os dados.

Engenharia reversa somente é possível quando o modelo adotado pode descrever as propriedades do sistema real e a quantidade de dados que o descreve seja suficiente. E mesmo que essas condições sejam atendidas, resolver esse problema é um desafio e existem poucos métodos eficientes com esse objetivo. Esses métodos dependem fortemente da técnica de modelagem utilizada. Modelos quantitativos são mais exigentes tanto em termos de esforço computacional como em termos de quantidade necessária de dados. Modelos dinâmicos contêm diversos parâmetros, requerendo um conjunto de dados bastante detalhado para estimá-los satisfatoriamente [Schlitt and Brazma, 2007].

Somogyi *et al* propôs duas abordagens para identificação de redes gênicas [Somogyi et al., 1997]. A primeira abordagem (mecanicista-reducionista) é “Um gene para cada função e uma função para cada gene” e consiste das seguintes atividades principais:

- Redução total de organismos em genes;
- Determinação de estruturas de proteínas e atividades correspondentes;
- Mapeamento de interações moleculares dos produtos gênicos;
- Criação de um banco de dados de mecanismos moleculares;
- Criação de um modelo computacional para soma das partes.

A segunda abordagem (heurística, estratégia integrativa) é “Funções gênicas são distribuídas em redes de processamento paralelo” e consiste de:

- Identificação de elementos da rede gênica;
- Determinação dos estados da rede (padrões de expressão);
- Mapeamento de trajetórias e atratores;
- Busca por trajetórias paralelas (sugerem entradas compartilhadas);
- Determinação de relações temporais por sinais com formato ondulatório;

- Engenharia reversa da rede.

Existem diversos algoritmos propostos para inferência de redes Booleanas e Bayesianas [Liang et al., 1998, Akutsu et al., 2000a, Pe'er et al., 2001, Pournara and Wernisch, 2004, Friedman, 2004]. O Capítulo 3 apresenta um método de identificação de redes gênicas probabilísticas através de seleção de características por entropia condicional média, uma função critério similar à usada em [Liang et al., 1998], mas com um método de estimação de erro embutido.

Redes sintéticas

Uma abordagem poderosa para testar e validar métodos de identificação de redes de regulação gênica é a construção de redes sintéticas a partir de um modelo, sem se basear em nenhum conjunto de dados reais [Schlitt and Brazma, 2007]. Preditores de pequenos modelos foram testados experimentalmente com sucesso usando circuitos de controle específicos tais como laços de realimentação positiva (*feed forward loops*) e laços de retroalimentação (*feedback loops*) [Basu et al., 2004] [Becskei and Seraphin, 2000] [Elowitz and Leibler, 2000]. Mendes *et al* desenvolveram um modelo que gera redes de genes artificiais aleatórias de acordo com propriedades topológicas e cinéticas bem definidas para testar métodos de análise de dados de expressão gênica [Mendes et al., 2003].

Com o objetivo de avaliar as redes gênicas recuperadas por métodos computacionais, Lopes et al [Lopes et al., 2008a] desenvolveram um modelo que gera redes gênicas artificiais usando arquiteturas de redes complexas. Esse modelo define regras de interação entre os genes conectados com o objetivo de, a partir de um estímulo inicial, gerar séries temporais de expressão, as quais são submetidas a métodos de identificação de redes gênicas. Desse modo, as redes recuperadas podem ser validadas com base na rede geradora do sinal temporal.

Capítulo 3

Identificação de redes gênicas probabilísticas

3.1 Introdução

A vida de um organismo depende de diversas vias metabólicas que são reguladas por redes de expressão gênica e de sinalização. O mecanismo da regulação das vias envolve um sistema complexo com muitos sinais de naturezas distintas, algumas sendo inclusive de retroalimentação, produzindo ciclos na rede. Esses sinais podem ser de RNAs transcritos produzidos por expressão gênica e complexos proteicos produzidos pela interação de proteínas construídas pela tradução do RNA mensageiro.

Essa rede complexa de interações pode então ser modelada por um sistema dinâmico [Barrera et al., 2006a]. Em tal modelo, cada transcrito é representado por uma variável que toma seu valor de expressão. Todas essas variáveis tomadas em conjunto são os componentes que formam um vetor representando o estado do sistema. Cada componente do vetor de estados tem uma função associada que calcula seu próximo valor de expressão a partir de estados em instantes de tempo anteriores. Essas funções são os componentes de um vetor, chamado de função de transição, que define a transição de um estado ao próximo e representa os mecanismos de regulação gênica.

Seja R o conjunto de valores que cada componente do estado de um sistema dinâmico pode assumir. Por exemplo, $R = \{0, 1\}$ em sistemas binários e $R = \{-1, 0, 1\}$ em sistemas ternários. A função de transição Φ , para uma rede de n genes, é uma função de R^n em

R^n . Um sistema dinâmico finito é dado por:

$$\mathbf{x}[t+1] = \Phi(\mathbf{x}[t]) \quad (3.1)$$

em que $\mathbf{x}[t] \in R^n$, para todo $t \geq 0$. Um componente de $\mathbf{x}[t]$ é um valor $x_i[t] \in R$. Os sistemas definidos dessa forma são invariantes por translação no tempo, isto é, a função de transição é a mesma para todo tempo t . Quando Φ é uma função estocástica (ou seja, para cada estado $\mathbf{x}[t]$, o próximo estado $\Phi(\mathbf{x}[t])$ é uma realização de um vetor aleatório), o sistema dinâmico é um processo estocástico.

3.2 Redes gênicas probabilísticas

Um dos principais trabalhos originados desta tese refere-se à proposta de uma técnica de seleção de características baseada em entropia condicional para identificação de arquiteturas de redes gênicas probabilísticas a partir de dados de *microarray* obtidos do *Plasmodium falciparum*, um agente parasita causador da malária [Barrera et al., 2004, Barrera et al., 2006a]. Este trabalho foi desenvolvido em conjunto com outros pesquisadores do IME-USP e também em conjunto com a equipe do Prof. Dr. Hernando Del Portillo do Instituto de Ciências Biomédicas (ICB-USP).

As redes de expressões gênicas podem ser representadas por processos estocásticos. A função de transição estocástica é uma família particular de cadeias de Markov chamada de redes Booleanas probabilísticas (PBN) [Shmulevich et al., 2002]. As redes gênicas probabilísticas (PGN) são uma generalização das redes probabilísticas booleanas (PBN), ou seja, a PBN é uma PGN em que o domínio dos valores das expressões gênicas se restringe a apenas dois valores: 0 e 1 (domínio Booleano).

Considere uma seqüência de vetores aleatórios $\mathbf{X}[0], \mathbf{X}[1], \mathbf{X}[2], \dots$ assumindo valores em R^n e suas realizações denotadas respectivamente por $\mathbf{x}[0], \mathbf{x}[1], \mathbf{x}[2], \dots$. Uma seqüência de estados aleatórios $(\mathbf{X}[t])_{t=0}^{\infty}$ é chamada de cadeia de Markov se, para todo $t \geq 1$, $P(\mathbf{X}[t] = \mathbf{x}[t] | \mathbf{X}[0] = \mathbf{x}[0], \dots, \mathbf{X}[t-1] = \mathbf{x}[t-1]) = P(\mathbf{X}[t] = \mathbf{x}[t] | \mathbf{X}[t-1] = \mathbf{x}[t-1])$.

O conceito de cadeia de Markov reside no fato em que a probabilidade condicional de um evento futuro, dada a história passada, depende apenas do passado imediato e não do passado remoto. Uma cadeia de Markov é caracterizada por uma matriz de transição $\Pi_{\mathbf{Y}|\mathbf{X}}$ de probabilidades condicionais entre estados, cujos elementos são denotados por

$P(\mathbf{y}|\mathbf{x})$, e um vetor aleatório descrevendo a condição inicial $\mathbf{x}[0]$. Uma PGN é uma cadeia de Markov $(\Pi_{\mathbf{Y}|\mathbf{X}}, \mathbf{x}[0])$ tal que:

1. $\Pi_{\mathbf{Y}|\mathbf{X}}$ é homogênea, ou seja, $P(\mathbf{y}|\mathbf{x})$ não é uma função de t (invariante à translação no tempo);
2. $P(\mathbf{y}|\mathbf{x}) > 0$ para todo par de estados $\mathbf{x}, \mathbf{y} \in R^n$ (cadeia de Markov ergódica);
3. $\Pi_{\mathbf{Y}|\mathbf{X}}$ é condicionalmente independente, isto é, para todo par de estados $\mathbf{x}, \mathbf{y} \in R^n$, $P(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^n P(y_i|\mathbf{x})$;
4. $\Pi_{\mathbf{Y}|\mathbf{X}}$ é quase determinístico, isto é, para todo estado $\mathbf{x} \in R^n$, existe um único estado, $\mathbf{y} \in R^n$ tal que $P(\mathbf{y}|\mathbf{x}) \approx 1$.

As propriedades do modelo de PGN são inspiradas em fenômenos biológicos ou simplificações obrigatórias devido a uma usual falta de dados para a estimação do modelo. A primeira propriedade é uma restrição para simplificar o problema de estimação, mas poderia ser facilmente generalizado. A segunda impõe que todos os estados são atingíveis, ou seja, ruídos podem levar o sistema a qualquer estado. A terceira diz que a expressão de um gene em um dado instante de tempo t independe da expressão de outros genes no mesmo instante t . A quarta propriedade diz que o sistema tem uma dinâmica estrutural que está sujeita a pequenos ruídos.

É importante observar que o terceiro axioma pode não ser verificado dependendo da limitação da resolução temporal dos dados experimentais disponíveis. Entretanto, esse axioma foi adotado em nosso modelo para permitir alguma tratabilidade estatística.

3.3 Projeto de redes gênicas probabilísticas por entropia condicional

Uma PGN $\mathcal{G} = (\mathbf{X}, \Phi)$ de n genes é definida por um conjunto de n nós $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$, $X_i \in R$, e um vetor de n funções $\Phi = (\phi_1, \phi_2, \dots, \phi_n)$, $\phi_i : R^{k_i} \rightarrow R$ em que k_i é o número de genes dos quais o gene X_i depende (grau de entrada de X_i). Cada uma dessas funções é estocástica, ou seja, para cada uma das possíveis instâncias de $\mathbf{X}$, existe uma distribuição

de probabilidades no domínio de R que indica qual é a probabilidade do gene alvo X_i possuir um determinado valor em R dada a instância de $\mathbf{X}$.

Estimar a arquitetura de uma PGN a partir de algumas dezenas de amostras é um grande desafio. Neste trabalho, a arquitetura das PGNs é gerada a partir da adoção de alguns genes de interesse como sementes (genes alvo) e da obtenção de genes que melhor predizem esses genes alvo. Com isso, a idéia é extrair informações sobre as relações entre os níveis de expressão gênica. Após a descoberta de tais relações, pode-se aprender algo sobre os mecanismos responsáveis por gerar essas expressões observadas. As relações sugeridas pelo método estimador de arquiteturas de PGNs que forem potencialmente interessantes do ponto de vista biológico podem ser experimentalmente examinadas em bancada. Como esses experimentos possuem um alto custo, a idéia é oferecer ao biólogo especialista um número reduzido de relações que caracterize de forma satisfatória um determinado fenômeno de interesse.

Adotamos a informação mútua média como descrita no Capítulo 2 (Equação 2.11) como critério para estimação de arquiteturas de PGNs. A informação mútua depende da entropia da distribuição de probabilidades *a priori* do gene alvo ($H(Y)$) e da entropia da distribuição condicional do alvo dados os preditores ($H(Y|\mathbf{Z})$), $\mathbf{Z} \subseteq \mathbf{X}$. Como a ordem de classificação de um subconjunto $\mathbf{Z}$ não depende de Y , $H(Y)$ é sempre o mesmo para qualquer subconjunto considerado e, portanto, essa ordem depende apenas das entropias condicionais. O método de estimação da entropia condicional é baseado na penalização das instâncias observadas uma única vez (ver Capítulo 2, Equação 2.15). O parâmetro β dessa penalização utilizado nos experimentos deste capítulo foi $\frac{1}{c}$ em que c é o número de valores existentes no domínio de R (2 no caso binário ou 3 no ternário). A penalização através de $\beta = \frac{1}{c}$ deixa a distribuição de probabilidades condicionais uniforme (entropia máxima).

Para a estimação de arquiteturas de PGNs, considera-se um conjunto de genes alvo $\mathbf{Y} = \{y_1, y_2, \dots, y_n\} \subseteq \mathbf{X}$ e seus valores no instante de tempo posterior $\mathbf{Y}[t+1]$ como sendo predito por um subconjunto $\mathbf{Z} \subseteq \mathbf{X}$ de genes no instante de tempo t (i.e., $\mathbf{x}[t]$). Para cada gene alvo, aplica-se um algoritmo de seleção de características baseado no critério de informação mútua média para verificar quais subconjuntos de genes melhor predizem o comportamento do gene alvo. Esses subconjuntos são ligados ao gene alvo na rede resultante. Uma ilustração de como essa estimação é realizada para um único

gene alvo pode ser verificada na Figura 3.1. Ela consiste em uma matriz de expressões temporais discretas $(-1, 0, 1)$ na qual cada linha representa um perfil de expressão de um determinado gene e as colunas representam instantes de tempo (nesse exemplo, de 1 a 48). O procedimento representado por essa figura é o preenchimento da tabela das frequências absolutas de Y dado cada uma das instâncias de (X_1, X_2) que será transformada numa tabela de distribuição conjunta de probabilidades, sobre a qual a informação mútua média de Y dado X_1 e X_2 será calculada. Tal procedimento é repetido para todos os subconjuntos examinados pelo algoritmo de seleção de características adotado (caso seja o algoritmo de busca exaustiva nesse exemplo, todos os pares). Os melhores subconjuntos classificados de acordo com a informação mútua média são anexados como preditores do gene Y na rede.

Figura 3.1: Procedimento de obtenção das frequências absolutas do gene alvo Y dada cada uma das instâncias dos candidatos a preditores (X_1, X_2) a partir de uma matriz de expressões temporais ternárias composta por 48 instantes de tempo em que os genes $X_1, X_2, \dots, X_n$ são dispostos em linhas e os tempos $(t = 1, \dots, 48)$ em colunas. A figura da esquerda mostra o início do procedimento (coleta da primeira amostra $(x_1[1], x_2[1], y[2])$) e a da direita mostra o final do procedimento (coleta da última amostra $x_1[47], x_2[47], y[48]$). As cores verde, preto e vermelho representam, respectivamente, os valores -1 (subexpressão), 0 (normal) e 1 (superexpressão). Uma seqüência de slides que ilustra esse processo encontra-se em <http://www.vision.ime.usp.br/~davidjr/doutorado/tese/PGNdemo.pps>.

Após a aplicação do algoritmo de seleção de características considerado para todos os genes alvos, o resultado será uma rede estimada de dependências em torno dos genes alvo (ver ilustração na Figura 3.2).

Figura 3.2: Exemplo ilustrativo de arquitetura de PGN obtida utilizando-se alvos (vermelhos), cada um com o seu melhor par de preditores. Note que um alvo pode fazer parte do melhor subconjunto de preditores de um outro alvo.

3.4 Validação por meio de simulações

Para validar a técnica proposta de estimação de PGN, redes artificiais que satisfazem a definição de PGN foram criadas, simuladas e estimadas. Essas redes simuladas possuem 12 genes que podem ser preditos por um subconjunto de 1 a 5 genes ou não serem preditos por nenhum deles. Todos os genes da rede são ternários ($R = \{-1, 0, +1\}$) e a distribuição de probabilidades condicionais de um gene alvo dada qualquer instância de seus preditores $P(Y|\mathbf{Z} = \mathbf{z})$ possui 80% de sua massa concentrada em um único valor de Y . As simulações envolvem 48 iterações, que é o número de instantes de tempo com resolução de 1 hora presentes nos dados de *Plasmodium falciparum* utilizados pelos experimentos da próxima seção (Seção 3.5). Para cada gene alvo, todos os genes individuais, pares e triplas, tiveram suas informações mútuas calculadas, sendo que os cinco melhores de cada categoria foram considerados para calcular a *qualidade* de um gene preditor X_i com relação ao alvo Y . Definimos essa métrica (qualidade) como a soma das informações mútuas de todas as tuplas (individuais, pares ou triplas) preditoras de Y em que X_i está presente como

preditor (note que, para a categoria dos preditores individuais, a informação mútua do preditor em si equivale à sua qualidade). Finalmente, os preditores foram ordenados por esse critério de qualidade. A Figura 3.3 mostra a arquitetura de predição de uma rede que foi simulada por esse modelo. A Figura 3.4 mostra um esquema do procedimento dos cálculos das informações mútuas, qualidades e suas respectivas ordenações. A Tabela 3.1 contém os melhores preditores individuais e suas respectivas informações mútuas. As Tabelas 3.2 e 3.3 contém os 5 melhores genes de acordo com o critério de qualidade para os cinco melhores pares e as cinco melhores triplas, respectivamente. Maiores detalhes estão disponíveis em <http://www.vision.ime.usp.br/~davidjr/CAMDA2004/simulations>.

Figura 3.3: Rede artificial de 12 genes e seus respectivos preditores.

É importante notar que, para genes alvos preditos por 1 ou 2 genes, os preditores sempre foram recuperados corretamente através do ordenamento pelo critério de qualidade, mesmo com uma incerteza de 20% sobre o gene alvo a partir do conhecimento dos valores dos preditores. Já para genes alvo com 3 ou mais preditores, apesar do controle de qualidade devolver a maioria dos genes preditores corretamente, os erros de estimação começam a aparecer de forma bastante significativa. Isso mostra que, para um conjunto com apenas 48 amostras e com domínio ternário, três genes preditores formam o limite a partir do qual a qualidade de estimação dos preditores começa a ficar insatisfatória. Portanto, não é válido procurar subconjuntos com 4 ou mais preditores sob essas condições, pois a tabela de frequências absolutas começa a ficar extremamente rarefeita. Afinal, o número de linhas dessa tabela (número de instâncias possíveis) é dado pelo número de valores do domínio elevado ao número de preditores considerados (48 amostras são insuficientes para

Figura 3.4: Esquema do procedimento dos cálculos das informações mútuas, qualidades e suas respectivas ordenações.

observar todas as instâncias possíveis de um experimento ternário considerando 4 ou mais preditores, já que o número total de possíveis preditores é maior ou igual a $3^4 = 81$).

Tabela 3.1: Cinco melhores genes individuais obtidos de acordo com a informação mútua (IM , na tabela) para cada gene da rede disposto como alvo a partir de um conjunto de 48 amostras simuladas através da rede de predição descrita na Figura 3.3 com $P(Y|\mathbf{X} = \mathbf{x}) = 0.8$ para todo $\mathbf{x} \in \mathbf{X}$.

	Y (alvo)											
	0	1	2	3	4	5	6	7	8	9	10	11
1	2	11	1	9	1	6	8	4	7	6	2	2
IM	0.117	0.129	0.828	0.345	0.221	0.297	0.227	0.176	0.242	0.534	0.086	0.106
2	7	7	0	3	4	4	3	1	3	4	1	6
IM	0.097	0.068	0.073	0.083	0.205	0.102	0.073	0.170	0.146	0.194	0.078	0.091
3	11	5	10	2	6	8	9	2	4	2	3	3
IM	0.094	0.061	0.064	0.073	0.145	0.088	0.069	0.126	0.141	0.091	0.073	0.051
4	9	0	11	4	10	11	7	10	8	11	5	0
IM	0.087	0.054	0.042	0.047	0.137	0.072	0.056	0.102	0.090	0.059	0.067	0.044
5	3	3	7	11	7	3	4	5	10	3	8	7
IM	0.052	0.048	0.033	0.046	0.122	0.065	0.054	0.056	0.078	0.059	0.055	0.042

3.5 Resultados experimentais em dados de *Plasmodium falciparum*

O seqüenciamento do genoma do *Plasmodium falciparum* revelou que aproximadamente 60% do genoma anotado corresponde a proteínas hipotéticas e que muitos genes, cujas vias metabólicas ou produtos biológicos são conhecidos bioquimicamente, não foram preditos [Bozdech et al., 2003]. Através do uso da técnica de Transformada Discreta de Fourier (do inglês: *Discrete Fourier Transform* - DFT), foi sugerido que os parasitas seguem um programa de regulação de transcrição relativamente rígido ao longo de seus ciclos de vida [Bozdech et al., 2003]. Então, uma nova lista de genes codificantes com funções biológicas semelhantes aumentou de forma considerável a quantidade de novos alvos para vacinas e desenvolvimento de drogas. Nossa proposta é anotar genes sob uma diferente perspectiva: uma lista de propriedades funcionais é atribuída a redes de genes representando subsistemas do sistema regulatório de expressão do parasita

Tabela 3.2: Cinco melhores genes em termos de qualidade (Q , na tabela) para os cinco melhores pares obtidos de acordo com a informação mútua para cada gene alvo a partir de um conjunto de 48 amostras simuladas através da rede de predição descrita na Figura 3.3 com $P(Y|\mathbf{X} = \mathbf{x}) = 0.8$ para todo $\mathbf{x} \in \mathbf{X}$.

	Y (alvo)											
	0	1	2	3	4	5	6	7	8	9	10	11
1	2	11	1	9	1	6	8	1	1	6	2	2
Q	0.723	1.201	4.410	2.014	1.987	2.463	1.824	1.386	2.288	3.180	1.297	0.921
2	10	5	6	6	4	3	5	4	3	7	0	11
Q	0.689	0.578	0.891	0.420	0.993	0.565	0.401	0.978	0.492	0.660	0.687	0.679
3	9	2	11	10	6	1	9	2	2	4	7	6
Q	0.666	0.312	0.880	0.413	0.927	0.485	0.372	0.784	0.489	0.649	0.613	0.581
4	7	9	7	2	10	8	4	10	4	2	11	4
Q	0.632	0.305	0.880	0.408	0.483	0.481	0.353	0.743	0.448	0.643	0.300	0.350
5	11	7	10	11	7	11	10	8	8	1	9	8
Q	0.323	0.301	0.879	0.389	0.436	0.472	0.350	0.368	0.439	0.618	0.299	0.291

[Barrera et al., 2004, Barrera et al., 2006a]. Dado um conjunto de genes semente responsáveis por algum subsistema importante do mecanismos regulatório do parasita, e uma matriz de expressões gênicas representando a dinâmica de seus estados transcricionais ao longo do tempo, a idéia é construir uma rede em torno desses genes semente, de forma a facilitar a identificação das funcionalidades de alguns genes.

Para isso, adotamos o modelo proposto de redes gênicas probabilísticas estimadas por entropia condicional para construir redes que representem um subsistema da rede de expressão gênica do parasita.

3.5.1 Normalização e quantização

A quantização é uma tarefa de pré-processamento mandatória para a aplicação da técnica proposta de estimação de PGN. Os dados de expressão de *microarray* são valores reais resultantes da razão de duas expressões (Cy5/Cy3). As etapas de normalização e quan-

Tabela 3.3: Cinco melhores genes em termos de qualidade (Q , na tabela) para as cinco melhores triplas obtidas de acordo com a informação mútua para cada gene alvo a partir de um conjunto de 48 amostras simuladas através da rede de predição descrita na Figura 3.3 com $P(Y|\mathbf{X} = \mathbf{x}) = 0.8$ para todo $\mathbf{x} \in \mathbf{X}$.

	Y (alvo)											
	0	1	2	3	4	5	6	7	8	9	10	11
1	2	11	1	9	4	6	10	1	7	6	1	11
Q	2.579	2.432	4.485	2.290	3.631	2.882	1.721	3.651	1.919	3.442	2.517	2.001
2	9	2	11	1	1	8	8	4	4	7	2	1
Q	2.574	1.723	3.600	1.408	3.631	1.731	1.712	2.948	1.280	2.054	1.868	1.965
3	10	9	10	10	0	2	9	11	11	10	0	2
Q	1.924	1.274	1.797	1.338	0.758	1.724	1.677	1.426	1.277	1.387	1.861	1.965
4	0	1	9	2	6	0	1	9	10	2	11	4
Q	1.275	1.235	0.900	0.899	0.731	1.147	1.644	1.424	1.268	1.379	1.281	1.317
5	8	8	4	11	11	10	2	0	2	8	9	10
Q	0.656	0.615	0.894	0.482	0.721	0.587	1.131	0.754	1.268	0.696	0.653	1.317

tização envolvidas nos experimentos deste capítulo, na ordem em que são aplicadas, estão enumeradas a seguir:

1. Aplicação do logaritmo na base 2 de todas as expressões originais, resultando na matriz $\mathbf{G}$;
2. Os sinais de $\mathbf{G}$ foram normalizados por uma transformação normal dada por, para todo gene $g(t) \in G$, $\eta[g(t)] = \frac{g(t) - E[g(t)]}{\sigma[g(t)]}$, em que $E[g(t)]$ e $\sigma[g(t)]$ são, respectivamente, a esperança e o desvio padrão de $g(t)$;
3. Seja $g'(t) \in \eta[g(t)]$. A quantização de um gene g' é realizada por um mapeamento definido, para todo t por:

$$g''(t) = \begin{cases} -1, & \text{se } g'(t) < l \\ 0, & \text{se } l \leq g'(t) \leq h \\ 1 & \text{se } g'(t) > h \end{cases} \quad (3.2)$$

em que l é a média dos valores negativos de $g'(t)$ e h é a média dos valores positivos de $g'(t)$.

A transformação normal faz com que todos os sinais tenham média zero e desvio padrão 1, dispondo todos os sinais numa mesma escala de variação. Embora a transformação normal crie classes de equivalência que diminuem o erro de estimação, ela amplifica ruído em genes com perfis de expressão quase constantes (*housekeeping*). Para usar a transformação normal, esses genes devem ser filtrados previamente.

A quantização é baseada no fato da transformação normal fazer com que a média de cada sinal tenha valor zero. Dessa forma, o número de valores positivos e negativos serão aproximadamente o mesmo. Assim, fizemos com que os valores de subexpressão (-1) estivessem abaixo da média dos valores negativos e os valores de superexpressão (+1) acima da média dos valores positivos. Os valores próximos à média geral (acima da média dos negativos e ao mesmo tempo abaixo da média dos positivos) foram considerados como valores normais (0).

3.5.2 Dados de HB3

O transcriptoma do ciclo de desenvolvimento intraeritrocítico (do inglês: *Intraerythrocytic Developmental Cycle* - IDC) do *Plasmodium falciparum* foi gerado por medidas relativas dos níveis de abundância do mRNA (medidas transcricionais) de amostras coletadas de uma cepa chamada de HB3, que é bem caracterizada e originada de Honduras [Bozdech et al., 2003]. A partir do conjunto de dados de *microarray* original representando esse transcriptoma, foi construído um conjunto de dados, chamado USP-Dataset, resultante da aplicação de um filtro de pré-processamento que eliminou *spots* com sinal muito fraco em ambos os canais Cy3 e Cy5 de acordo com um determinado limiar de corte [Barrera et al., 2006a]. Como resultado desse pré-processamento, o conjunto USP-Dataset contém 6532 genes.

Glicólise

Durante os estágios assexuais do parasita (estágio em que se hospeda no sangue humano), ele depende inteiramente da glicólise para produção de ATP [Sherman, 1983]. Assim,

genes alvos que codificam para 10 enzimas pertencentes à via glicolítica foram escolhidos como sementes para testar a capacidade preditora do modelo de PGN proposto. Uma PGN foi gerada selecionando-se os 20 melhores preditores individuais (menores valores de entropia condicional) para cada alvo considerado (ver Figura 3.17 no final deste capítulo), interconectando 9 dos 10 genes alvo de forma bastante significativa. Além disso, uma análise dos 40 melhores preditores para cada alvo da glicólise (289 genes no total) revelou que a maior parte deles (96%) corresponde a genes hipotéticos relacionados com transcrição, tradução, síntese de DNA e RNA, motores de actina myosina e kinases (<http://www.vision.ime.usp.br/~davidjr/CAMDA2004/Table1S.html>). Os 4% restantes codificam antígenos de superfície, o que *a priori*, podem ser considerados falsos-positivos. Resultados similares foram obtidos de uma lista de 400 genes expressos em fase com a glicólise obtida dos dados de Bozdech *et al* [Bozdech et al., 2003]. Como esperado, nenhum gene do ciclo de ácido tricarboxílico (TCA) foi encontrado na PGN de glicólise, corroborando a falta de um ciclo de TCA funcional durante os estágios assexuais sangüíneos dos parasitas da malária [Sherman, 1983].

Além disso, é importante destacar que vários genes não levados em conta na análise de Bozdech *et al* [Bozdech et al., 2003] por possuírem sinais não-senoidais (cerca de 50% dos genes da malária possuem sinais senoidais na escala logarítmica), foram incluídos na PGN (Figura 3.17 no final do capítulo, podendo ser melhor visualizada em <http://www.vision.ime.usp.br/~davidjr/CAMDA2004/glycolysis.html>). Dois genes alvo da glicólise considerados na PGN não possuem sinal senoidal, sendo que um deles possui uma interconexão significativa com outros alvos.

Apicoplasto

As enzimas de uma organela conhecida como apicoplasto ou plastídeo estão se tornando novos alvos de drogas contra a malária, haja visto que não existem homólogos dessa organela nos humanos [Ralph et al., 2004, Wilson, 2002]. De acordo com a lista do laboratório de DeRisi (<http://malaria/ucsf.edu>), existem 27 genes com função anotada como sendo de apicoplasto. Esses genes foram adotados como sementes para a geração de uma PGN pelo nosso método. A Figura 3.18, encontrada no final deste capítulo (melhor visualizada em <http://www.vision.ime.usp.br/~davidjr/CAMDA2004/apicoplast.html>), indica que apenas 7 melhores preditores de cada gene alvo foram responsáveis por in-

terconectar 26 dos 27 alvos numa mesma componente. Uma análise mais detalhada dos resultados indica que nosso método é capaz de interconectar genes que têm sido experimentalmente demonstrados como parte do apicoplasto (proteínas acyl-carrier e ribossomal S9) [Mazumdar et al., 2006, Yung et al., 2003].

Controle negativo

Um experimento denominado de “controle negativo” foi idealizado para validar biologicamente os resultados encontrados pela metodologia de construção de PGNs. Tal experimento consiste em selecionar 8 genes alvos, sendo 4 alvos da glicólise e 4 alvos do apicoplasto, selecionados ao acaso para construir uma PGN baseada em predição individual e outra baseada em predição por pares de genes. A hipótese biológica por trás desse experimento é de que a rede deveria ter duas componentes conexas disjuntas, uma para os alvos da glicólise e outra para os alvos do apicoplasto, pois trata-se de dois sistemas biológicos distintos do parasita. Os resultados mostram claramente que as redes baseadas em predições individuais não foram inter-conectadas (ver Figura 3.5, sendo melhor visualizada em http://www.vision.ime.usp.br/~davidjr/CAMDA2004/ga_c.html). Já para a rede baseada em predição por pares, apenas dois genes fizeram a inter-conexão entre os dois conjuntos, sendo que o restante dos genes não foram inter-conectados (ver Figura 3.6, sendo melhor visualizada em http://www.vision.ime.usp.br/~davidjr/CAMDA2004/ga2_c.html). Note que para o caso de pares de preditores, existem 21.330.246 pares (combinação de 6532, 2 a 2), reforçando a importância desses resultados. Em conjunto, esses resultados demonstram o valor do modelo PGN em gerar redes significativas biologicamente e que inclui genes não tratados pela abordagem DFT [Barrera et al., 2004, Barrera et al., 2006a].

3.5.3 Dados de 3D7, Dd2 e HB3

Continuando os esforços de pesquisa em malária, DeRisi *et al* publicou dados com medidas de expressão gênica do IDC de outras duas cepas (3D7 e Dd2) junto com os dados da cepa HB3 usada em seu trabalho anterior [Llinás et al., 2006]. Essas cepas são originadas de diferentes locais (3D7: Holanda, HB3: Honduras, Dd2: Indochina) e possuem fenótipos distintos de sensibilidade a drogas.

Figura 3.5: Rede dos 50 melhores preditores individuais para cada um dos quatro alvos da glicólise (amarelo) e quatro alvos de apicoplasto (verde). No total, foram analisados 6532 genes.

Agora que existem três conjuntos de dados, um ponto chave é checar a possibilidade de unir os três conjuntos em um único com $48 \times 3 = 144$ instantes de tempo para melhorar a precisão da estimação da PGN. Com essa intenção, é importante que a maioria dos genes tenham perfis de expressão similares nos três conjuntos considerados. Existem diversas maneiras de medir a similaridade entre sinais. Em [Llinás et al., 2006], os autores aplicaram correlação de Pearson aos pares de sinais (3D7 com Dd2, 3D7 com HB3 e Dd2

com HB3) e concluíram que as diferenças entre as cepas em termos de expressão gênica são pequenas durante o IDC (70% das correlações de Pearson são maiores que 0.7 em uma escala de -1 a 1). Adicionalmente, após um pré-processamento dos dados (ver a próxima subseção), aplicamos duas outras medidas de similaridade para reforçar esse fato: correlação de Spearman (pares de sinais, escala de -1 a 1) e concordância de Kendall (três sinais simultâneos, escala de 0 a 1). A correlação de Spearman baseia-se no quadrado das distâncias entre os “postos” dos sinais g_1 e g_2 . Ou seja, dada a ordenação dos t valores de g_1 e g_2 , obtém-se a colocação (o posto) de cada um dos t elementos dos sinais originando os vetores de postos (A_1 e A_2). Quanto menor a soma dos quadrados das distâncias entre os postos, maior a correlação entre eles. Formalmente, a correlação de Spearman ρ é dada por [Kendall, 1962]:

$$\rho = 1 - \frac{6 \sum_{i=1}^t [A_1(i) - A_2(i)]^2}{t(t^2 - 1)} \quad (3.3)$$

A concordância de Kendall é uma extensão da correlação de Spearman para três sinais g_1 , g_2 e g_3 com seus respectivos vetores de postos A_1 , A_2 e A_3 . Formalmente, ela é definida por [Kendall, 1962]:

$$\rho_3 = \frac{12 \sum_{i=1}^t \left(\left[\frac{A_1(i) + A_2(i) + A_3(i) - 3(t+1)}{2} \right]^2 \right)}{9(t-1)t(t+1)} \quad (3.4)$$

Os histogramas contendo o número de genes para cada intervalo de correlação ou concordância estão dispostos na Figura 3.7. Todos os histogramas possuem um pico entre 0.8 e 0.9. Assim como foi constatado através do emprego do coeficiente de correlação de Pearson, no mínimo 70% da massa dos histogramas da correlação de Spearman e da concordância de Kendall está concentrada no intervalo de 0.7 a 1, o que indica que grande parte dos perfis de expressão estão bem correlacionados ao longo dos três conjuntos (note que todos os histogramas possuem aspectos bem parecidos). Portanto, pode-se concluir que os três conjuntos de dados podem ser unidos em um único conjunto com 144 (48×3) instantes de tempo. É importante ressaltar que caso todos os sinais fossem perfeitamente correlacionados (próximos de 1), o ganho de informação ao unir os 3 conjuntos seria praticamente nulo, o que não é o caso para os conjuntos de dados utilizados nos experimentos desta seção.

Dado que esses conjuntos são bem correlacionados entre as cepas, resolvemos aplicar

Figura 3.7: Histogramas de correlação de Pearson (pares de sinais), correlação de Spearman (pares de sinais) e concordância de Kendall (três sinais), contendo o número de genes para cada intervalo de correlação (ou concordância). Todos eles apresentam pelo menos 70% da massa concentrada no intervalo de 0.7 a 1, o que sugere que os três conjuntos podem ser unidos.

o esquema de projeto de PGN proposto usando os três conjuntos de dados em um único. Para isso, os três conjuntos disponibilizados por [Llinás et al., 2006] tratados por um controle de qualidade, foram tomados como entrada para todas as análises discutidas de agora em diante. Tal controle de qualidade consistiu em eliminar genes com mais de 40% de dados não observados nos respectivos perfis de expressão, transformação da matriz de expressões no logaritmo base 2 e suavização usando a função LOESS do pacote estatístico do R (<http://www.R-project.org>).

Esses conjuntos obtidos de 3D7, Dd2 e HB3 têm respectivamente 53, 50 e 48 instantes de tempo. Devido ao critério usado pelo controle de qualidade descrito anteriormente que descarta genes com mais de 40% de dados ausentes, o número de genes de cada conjunto também é diferente. Assim, existem 6287 perfis de expressão gênica para 3D7, 5294 para Dd2 e 6415 para HB3.

Pré-processamento

Considerando como entrada os conjuntos mencionados anteriormente, primeiramente os genes que não estão presentes nos três conjuntos foram filtrados, restando um total de 5185 genes. Nesses conjuntos filtrados, uma interpolação linear por vizinhos mais próximos seguida de uma reconstrução foi aplicada para todo sinal de forma a normalizar os três conjuntos em 48 instantes de tempo e preencher os dados faltantes. Finalmente, esses dados foram submetidos a procedimentos de transformação normal e quantização descritos na Seção 3.5.1.

Glicólise

Foram fixados os mesmos dez genes alvo da glicólise como sementes para criar uma PGN com um número mínimo de preditores necessário para formar uma única componente conexa. As PGNs com os melhores preditores individuais e melhores pares de preditores de acordo com a entropia condicional média foram consideradas. Para uma PGN glicolítica conexa através de preditores individuais, foram necessários 18 melhores preditores para cada alvo, enquanto 10 melhores pares foram necessários para conectar uma PGN da glicólise baseada em pares de preditores. As Figuras 3.8 e 3.9 apresentam as respectivas PGNs, que podem ser melhor visualizadas em http://www.vision.ime.usp.br/~davidjr/rede_malaria/glico_3D7_Dd2_HB3.html e http://www.vision.ime.usp.br/~davidjr/rede_malaria/glico_3D7_Dd2_HB3_duplas.html, respectivamente. Um resultado marcante conseguido através de ambas as PGNs projetadas é que dois outros genes com função de glicólise apareceram nas PGNs (indicados por setas vermelhas nas Figuras 3.8 e 3.9). Tal fato não ocorre para cada conjunto de dados em separado (no máximo um deles aparece). Essa é uma evidência do ganho de informação que se obtém por unir três conjuntos de dados para o projeto de PGNs mais precisas.

Figura 3.9: PGN dos 10 melhores pares de preditores dos alvos da glicólise formando uma única componente conexa (os nós amarelos representam os genes da glicólise). As setas vermelhas indicam dois outros genes da glicólise obtidos como preditores, fato que não ocorre quando se considera cada PGN de 3D7, Dd2 e HB3 separadamente.

Apicoplasto

Como feito para a glicólise, PGNs do apicoplasto com um número mínimo de preditores (simples e pares) para formar uma única componente conexa foram geradas. Dos 27 alvos do apicoplasto, 24 estão presentes nos três conjuntos de dados (3D7, Dd2 e HB3) e foram usados para criar as PGNs. As Figuras 3.10 e 3.11 mostram os resultados, que podem ser melhor visualizados em http://www.vision.ime.usp.br/~davidjr/rede_malaria/apico_3D7_Dd2_HB3.html e <http://www.vision.ime.usp.br/~davidjr>

/rede_malaria/apico_3D7_Dd2_HB3_duplas.html, respectivamente. De forma surpreendente, apenas 4 melhores preditores individuais foram suficientes para formar uma única componente conexa, enquanto apenas 2 melhores pares conectaram inteiramente a PGN baseada nos pares de preditores. Isso é um indício de que uma parte significativa dos genes presentes nessas PGNs podem ter função importante relacionada ao subsistema do apicoplasto, enquanto uma pequena parte deve ser composta por falsos positivos. Neste caso, o espaço de busca é pequeno, facilitando o trabalho do especialista em malária na tarefa de anotação e validação em bancada.

Figura 3.10: PGN dos 4 melhores preditores individuais dos alvos do apicoplasto formando uma única componente conexa (os nós verdes representam os genes do apicoplasto).

Figura 3.11: PGN dos 2 melhores pares de preditores dos alvos do apicoplasto formando uma única componente conexa (os nós verdes representam os genes do apicoplasto).

Controle negativo

Como feito para os dados de HB3 (Seção 3.5.2), foi realizado um experimento de controle negativo usando o conjunto de dados formado pela união dos dados de 3D7, Dd2 e HB3. A idéia novamente é verificar se a rede constituída pelos alvos da glicólise e do apicoplasto forma duas componentes desconexas ou fracamente conexas. Entretanto, uma diferença importante é que agora foram adotados todos os alvos de ambos os subsistemas (10 da glicólise e 24 do apicoplasto ao invés de 4 sorteados de cada subsistema) para validar o método com maior significância.

Um resultado bastante significativo desse experimento pode ser visto nas Figuras 3.12 e 3.13, que correspondem a uma PGN contendo o número máximo de preditores individu-

ais para o qual os subsistemas da glicólise e do apicoplasto permanecem desconexos (essas figuras podem ser melhor visualizadas em http://www.vision.ime.usp.br/~davidjr/rede_malaria/ga_3D7_Dd2_HB3.html). Essa rede tem um total de 552 genes, 297 preditores da glicólise e 255 do apicoplasto. A significância (p-valor) desse resultado, isto é, a probabilidade de intersecção nula entre dois subconjuntos de 297 e 255 elementos escolhidos aleatoriamente de um conjunto de 5185 elementos, é de 1.9812×10^{-7} .

Figura 3.12: 44 melhores preditores individuais dos 10 alvos da glicólise. 44 é o número máximo de preditores individuais em que os componentes formados pelos alvos da glicólise e do apicoplasto permanecem desconectados.

Outro resultado, talvez mais marcante que o anterior, é o fato de que o gene que faz

Figura 3.13: 44 melhores preditores individuais dos 24 alvos do apicoplasto. 44 é o número máximo de preditores individuais em que os componentes formados pelos alvos da glicólise e do apicoplasto permanecem desconectados.

a ponte entre os dois subsistemas para 45 melhores preditores permanece sendo o único gene fazendo tal ponte para PGNs com até 86 melhores preditores individuais (ver esquema na Figura 3.14). Tal PGN tem um total de 899 genes, 499 preditores da glicólise e 400 preditores do apicoplasto. A significância (p-valor) desse resultado, isto é, a probabilidade de intersecção entre dois subconjuntos de 499 e 400 elementos aleatoriamente escolhidos de um conjunto de 5185 elementos ter no máximo 1 elemento,

é de 2.2175×10^{-17} .

Figura 3.14: Esquema que mostra o gene OPFJ12793 (40S ribosomal protein, putative) fazendo a ponte de ligação entre os componentes formados por 86 melhores preditores individuais dos 10 genes alvo da glicólise e dos 24 genes alvo do apicoplasto. 86 é o número máximo de preditores individuais que conecta os dois componentes por no máximo 1 gene.

Intersecção entre PGNs de apicoplasto e genes apicoproteínas

Dos mais de 5000 genes do *Plasmodium falciparum*, cerca de 10% dos genes (550 genes) codificam proteínas associadas ao apicoplasto (também conhecidas como *apicoproteínas*). De um total de 289 genes expressos em fase com o genoma do apicoplasto (obtidos por DFT), 38 estão no conjunto das apicoproteínas (13.14%) [Bozdech et al., 2003]. O método de PGN foi aplicado para obter a intersecção entre genes de apicoplasto e de apicoproteínas e comparar com a intersecção obtida por [Bozdech et al., 2003].

Para aplicar o método de PGN, os mesmos 24 genes alvo do apicoplasto foram adotados como sementes. Então, o método foi aplicado para construção de PGNs com aproximadamente 280 genes para preditores individuais e para pares de preditores. Para a PGN baseada em preditores individuais, os 60 melhores preditores de cada alvo foram computados, produzindo 280 genes no total. Desses 280 genes, 40 estão presentes no conjunto das apicoproteínas (14.28%), uma porcentagem ligeiramente maior do que a produzida via DFT (13.14%). Por outro lado, para a PGN baseada em pares de preditores, 12 melhores pares para cada alvo foram computados, produzindo 273 genes no total. Desses 273 genes, 35 estão presentes no conjunto das apicoproteínas (12.82%).

Com o objetivo de medir a significância dessas porcentagens, 24 genes do conjunto das apicoproteínas foram selecionados aleatoriamente para gerar PGNs com 15 melhores preditores individuais, produzindo aproximadamente 280 genes em média. Esse procedimento foi aplicado 5000 vezes, sendo que a porcentagem média de genes recuperados foi de 10.32%. A Figura 3.15 ilustra esquematicamente esse procedimento e os procedimentos de PGN e DFT levando em conta os 24 alvos do apicoplasto, bem como cada um dos resultados obtidos.

O histograma de porcentagens de intersecção com o conjunto das apicoproteínas (excluindo os alvos escolhidos) após as 5 mil execuções do procedimento de geração de PGNs a partir de 24 apicoproteínas escolhidas aleatoriamente como alvo pode ser analisado na Figura 3.16. Esse histograma mostra que as porcentagens de intersecção com o conjunto das apicoproteínas para ambas as PGNs com alvos de apicoplasto (individuais e pares) e dos genes em fase com os alvos do apicoplasto são significativas, principalmente para a PGN baseada em preditores individuais, já que os alvos considerados (genes de apicoplasto) para gerar os três resultados (DFT, PGN com preditores individuais e PGN com pares de preditores) não estão presentes no conjunto das apicoproteínas.

3.6 Conclusão

Com o propósito de avançar o conhecimento biológico sobre o *Plasmodium falciparum*, redes gênicas probabilísticas (PGN), estimadas pela entropia condicional média, foram construídas a partir de sinais dinâmicos de expressão gênica do estado assexuado sangüíneo do parasita, publicados por Bozdech *et al* e Llinás *et al* [Bozdech et al., 2003, Llinás et al., 2006]. Diferentemente da abordagem DFT proposta para anotar genes que estão na mesma fase de alvos de um determinado subsistema, o método de PGN permite o uso de todos os elementos dos conjuntos de dados, independentemente das formas de seus sinais. Essa técnica foi validada através de simulações e aplicada a dados reais com genes alvo que codificam enzimas da via metabólica da glicólise e genes alvo do apicoplasto, com resultados potencialmente importantes do ponto de vista biológico [Barrera et al., 2004, Barrera et al., 2006a].

Esses resultados podem ser melhorados com a introdução de uma quinta propriedade que implica que cada gene seja caracterizado por uma combinação linear de outros genes

Figura 3.15: (A) Esquema ilustrando o procedimento de geração de PGNs a partir de 24 apicoproteínas selecionadas aleatoriamente dentre as 550 no total (15 melhores preditores individuais para cada alvo escolhido, obtendo 280 genes em média). Este procedimento foi executado 5000 vezes resultando numa média de 10.32% dos 280 genes como sendo de apicoproteínas. (B) Esquema que ilustra os experimentos de PGNs e DFT utilizando os 24 genes de apicoplasto como alvos. Dos 289 genes em fase com os alvos do apicoplasto (DFT), 13.14% foram de apicoproteínas. Já com relação aos genes resultantes da PGN com 60 melhores preditores individuais (280 genes), 14.28% deles foram de apicoproteínas. Finalmente, para PGN com 12 melhores pares (273 genes), as apicoproteínas compuserem 12.82% do total de genes resultantes.

[Barrera et al., 2006a]. Sua definição é a seguinte:

- Para todo gene j , existe um vetor $\mathbf{a}^j$ de números inteiros tal que para todo $\mathbf{x}, \mathbf{z} \in R^n$

Figura 3.16: Histograma de porcentagens da intersecção com o conjunto das apicoproteínas de 5000 execuções do método de PGN com 15 melhores preditores usando 24 apicoproteínas sorteadas como sementes. As setas indicam as barras às quais os experimentos que utilizam genes de apicoplasto pertenceriam se elas fossem uma das 5000 execuções. Note, portanto, que os genes de apicoplasto como alvos para os três métodos comparados levam a porcentagens significativamente acima da média. Tal média foi obtida justamente através das 5000 execuções usando 24 apicoproteínas como sementes, sendo que os genes de apicoplasto não fazem parte do conjunto das apicoproteínas.

e $y_j \in R$, se $\sum_{i=1}^n a_i^j x_i = \sum_{i=1}^n a_i^j z_i$ então $P(y_j|\mathbf{x}) = P(y_j|\mathbf{z})$.

Com esse propriedade, cada gene é caracterizado por um vetor de coeficientes $\mathbf{a}$ e um vetor de função estocástica ω_j do conjunto dos números inteiros em R . Se a_i^j é positivo, então o gene alvo j é ativado pelo gene i . Caso a_i^j seja negativo, então ele é inibido pelo gene i . Caso a_i^j seja zero, então j não é afetado por i . Um gene j é predito pelo gene i se a_i^j for diferente de zero. A componente j da função de transição estocástica ϕ , denotada ϕ_j , é construída pela composição de ω_j com a combinação linear de a^j e o estado anterior $\mathbf{x}[t]$, isto é, para todo $t \geq 1$,

$$\phi_j(\mathbf{x}[t]) = \omega_j\left(\sum_{i=1}^n a_i^j x_i[t]\right), \quad (3.5)$$

em que $\omega_j(\sum_{i=1}^n a_i^j x_i[t])$ é uma realização de uma variável aleatória em R , com distribuição $P(y|\sum_{i=1}^n a_i^j x_i[t])$.

A introdução dessa propriedade permitiria que os erros de estimação diminuam por atribuir algumas instâncias à mesma classe de equivalência de acordo com a combinação linear de suas entradas. Além disso, esse modelo permite distinguir entre sinais ativadores e inibidores, sendo relativamente consistente com a dinâmica de genes observada.

Figura 3.17: PGN construída a partir de 10 genes alvo da glicólise (nós amarelos) e 20 melhores preditores individuais de cada alvo. Quadrados representam genes com sinais de expressão senoidais enquanto círculos representam genes com sinais não senoidais. As setas indicam dois genes alvo da glicólise que não possuem sinal senoidal e que foram incluídas na PGN.

Figura 3.18: PGN construída a partir de 27 genes alvo do apicoplasto (nós verdes) e 7 melhores preditores individuais de cada alvo. Quadrados representam genes com sinais de expressão senoidais enquanto círculos representam genes com sinais não senoidais.

Capítulo 4

Predição intrinsecamente multivariada: derivação analítica

4.1 Introdução

Um conjunto de características preditoras é considerado de *predição intrinsecamente multivariada* (do inglês: “*intrinsically multivariate prediction*” - IMP) com relação a uma variável alvo caso o comportamento do alvo seja fortemente predito pelo conjunto de preditores, mas muito mal predito por qualquer de seus subconjuntos. A Figura 4.1 ilustra esse fenômeno a partir de um exemplo com 3 características preditoras (X_1, X_2, X_3) e uma característica alvo (Y). A Figura 4.1(a) mostra que qualquer preditor tomado isoladamente não serve para prever o comportamento do alvo. Além disso, qualquer subconjunto dos preditores tomado dois a dois também não realiza uma boa predição como mostrado na Figura 4.1(b). Porém, o conjunto completo dos três preditores faz uma predição muito boa da característica alvo (Figura 4.1(c)). Sob essas condições, diz-se que esse par preditores-alvo ($\mathbf{X}, Y$) é de predição intrinsecamente multivariada. No contexto de regulação gênica e controle, diz-se que o gene alvo é quem controla os preditores (o gene alvo é o mestre e seus preditores são escravos dele) [Dougherty et al., 2007]. Ou seja, supõe-se que o sentido do controle regulatório é inverso ao da predição.

Um exemplo numérico bem simples de predição intrinsecamente multivariada é a lógica Booleana ou-exclusivo (XOR: $Y = X_1 \oplus X_2$) como mostrada na Figura 4.1. Essa figura mostra que a saída dessa operação é totalmente indefinida dada uma das variáveis pre-

Figura 4.1: Ilustração do conceito de predição intrinsecamente multivariada para o caso de 3 preditores e 1 alvo: (a) Preditores tomados isoladamente não predizem o comportamento do alvo; (b) Preditores tomados aos pares também não realizam uma boa predição; (c) A tripla (X_1, X_2, X_3) é um excelente preditor de Y .

ditoras. Ou seja, observando $X_1 = 0$, o resultado da saída pode ser tanto 0 como 1, o mesmo ocorrendo para $X_1 = 1$, $X_2 = 0$ e $X_2 = 1$. Portanto, não há qualquer ganho de informação do comportamento da variável de saída a partir da observação de uma variável isoladamente. Porém, observando os dois preditores em conjunto, a predição é perfeita, ou seja, a saída torna-se completamente determinada pela observação do par (X_1, X_2) .

Tabela 4.1: Lógica ou-exclusivo (XOR) como exemplo de predição intrinsecamente multivariada.

X_1	X_2	saída
0	0	0
0	1	1
1	0	1
1	1	0

Neste capítulo será apresentada primeiramente a motivação para se estudar o fenômeno

de predição intrinsecamente multivariada em seleção de características de um modo geral, bem como no contexto de redes de regulação gênica (Seção 4.2). Em seguida, na Seção 4.3, será definido formalmente o conceito de variáveis de predição intrinsecamente multivariada com base no coeficiente de determinação não-linear (CoD) (Seção 4.3). Depois, são analisados os principais fatores que implicam no surgimento desse fenômeno, tais como lógicas de predição (Seção 4.4) e poder preditivo, incluindo os efeitos da covariância entre preditores e de suas distribuições de probabilidade marginais (*bias*) (Seção 4.5). Finalmente, na Seção 4.6 serão discutidas evidências da eficácia do modelo proposto de IMP em detectar canalização em dados reais de expressão em melanoma, encerrando este capítulo com uma discussão (Seção 4.7).

4.2 Motivação

Em seleção de características de um modo geral, o *efeito nesting* discutido no Capítulo 2 (Seção 2.3.1) está de certa forma relacionado com o fenômeno de predição intrinsecamente multivariada. Existem diversos algoritmos de seleção de características que tentam amenizar esse efeito de forma a conseguir um resultado próximo do ótimo de forma eficiente. Portanto, compreender como caracterizar a predição intrinsecamente multivariada e quais são as condições para originá-la é fundamental nessa área de pesquisa, tanto do ponto de vista teórico quanto das aplicações.

No contexto de sistemas biológicos, uma função muito importante que determinados genes desempenham chama-se *canalização* [Waddington, 1942]. Funções canalizadoras constituem um tipo especial de função Booleana em que pelo menos uma das variáveis de entrada é capaz de determinar o resultado da função [Kauffman, 1969] [Dougherty et al., 2005]. Genes canalizadores possuem um poder regulatório abrangente, podendo controlar diversos processos cujos genes afetados não são altamente correlacionados sob condições normais. Quando um gene controlador (alvo) está inativo (ou ativo), nenhum dos subconjuntos de genes preditores será um bom preditor, já que o comportamento deles se torna altamente variável nessa condição. Por outro lado, quando o gene alvo está ativo (ou inativo), seu comportamento é muito bem predito pelo conjunto de genes sob seu controle. Portanto, genes canalizadores podem ser investigados a partir do estudo de genes com predição intrinsecamente multivariada.

A lógica Booleana AND é um exemplo em que a variável de saída é canalizadora das variáveis de entrada, pois qualquer uma das variáveis que assumir o valor 0 determina completamente que o valor da variável de saída seja 0. Do ponto de vista da variável de saída, para o caso do AND, quando assume o valor 0, ela fica quase insensível aos valores das variáveis de entrada, já que estas podem assumir os valores (0, 0), (0, 1) e (1, 0) (ver Tabela 4.2). Apenas quando ambas as variáveis de entrada estão ativas, a variável de saída também se encontrará ativa. Ou seja, a variável de saída, quando ativa, “canaliza” os valores de entrada para (1, 1). Note que a variável de saída é relativamente estável, sendo que seu estado só se modifica para atender configurações bem específicas (apenas quando as variáveis de entrada estão ativas no caso da lógica AND). É exatamente isso que define um gene canalizador: ser relativamente robusto a perturbações, porém tendo uma maleabilidade suficiente para modificar seu estado atendendo a uma exigência específica do sistema regulatório.

Tabela 4.2: Lógica AND como exemplo de função canalizadora.

X_1	X_2	saída
0	0	0
0	1	0
1	0	0
1	1	1

A função de canalização é exigida por qualquer sistema complexo para proteção contra perturbações aleatórias ou erros operacionais. Genes canalizadores são freqüentemente encontrados em vias de sinalização que entregam informação de uma variedade de fontes à aparelhagem que ativa funções celulares centrais tais como ciclo celular, sobrevivência, apoptose (morte celular) e metabolismo [Waddington, 1942, Reichhardt and Bassler, 2007].

Investigamos a noção de predição intrinsecamente multivariada (IMP) em termos do Coeficiente de Determinação (CoD) [Martins-Jr et al., 2008b], embora existam abordagens baseadas em teoria da informação propostas recentemente [Anastassiou, 2007, Varadan et al., 2006], que também tentam quantificar a predição intrinsecamente multivariada, mas sem destacar as condições para que esse fenômeno ocorra. O CoD quantifica o incremento relativo no poder preditivo da variável alvo ao considerar a informação das

variáveis preditoras [Dougherty et al., 2000]. O CoD talvez tenha sido o primeiro paradigma preditivo utilizado no contexto de dados de *microarray* com o objetivo de prover uma medida de interação não-linear entre genes [Dougherty et al., 2000, Kim et al., 2000a, Kim et al., 2000b, Shmulevich et al., 2002].

Como poderia ser esperado ao analisar um sistema complexo e altamente integrado, o desenvolvimento de métodos para identificar genes canalizadores via predição intrinsecamente multivariada, a partir de expressões gênicas ao longo de um conjunto de amostras biológicas (conjunto de perfis de expressão), é um grande desafio.

4.3 Definição de predição intrinsecamente multivariada

Diz-se que um conjunto de características preditoras $\mathbf{X}$ é de predição intrinsecamente multivariada (IMP) para a característica alvo Y com respeito a λ e δ , para $0 \leq \lambda, \delta \leq 1$ e $\lambda < \delta$, se

$$\max_{\mathbf{Z} \subsetneq \mathbf{X}} \text{CoD}_Y(\mathbf{Z}) \leq \lambda \wedge \text{CoD}_Y(\mathbf{X}) \geq \delta \quad (4.1)$$

sendo CoD o coeficiente de determinação como definido no Capítulo 2, Equação 2.12. Ou seja, geralmente define-se um valor de λ baixo e um valor de δ alto para que um conjunto seja de predição intrinsecamente multivariada.

Para um par preditores-alvo $(\mathbf{X}, Y)$, o maior δ para o qual a predição é intrinsecamente multivariada é $\delta = \text{CoD}_Y(\mathbf{X})$. Dessa forma, pode-se definir uma pontuação de predição intrinsecamente multivariada (*IMP score*) através do máximo valor de $\delta - \lambda$. Assim, o *IMP score* é dado por

$$I_Y(\mathbf{X}) = \text{CoD}_Y(\mathbf{X}) - \max_{\mathbf{Z} \subsetneq \mathbf{X}} \text{CoD}_Y(\mathbf{Z}) \quad (4.2)$$

Neste capítulo, assume-se que as características são variáveis binárias, ou seja, $\mathbf{X} = \{x_1, x_2, \dots, x_n\} \in \{0, 1\}^n$ e $Y = \{0, 1\}$. Dessa forma, é possível analisar o fenômeno de predição intrinsecamente multivariada em termos de lógicas Booleanas de predição.

4.4 Lógicas de predição

Exemplos de lógicas de predição são o XOR e o AND, como ilustrado nas Tabelas 4.1 e 4.2. A variável de saída (alvo) Y é definida por uma lógica de predição em termos dos preditores $\mathbf{X}$ simplesmente se for uma função Booleana de $\mathbf{X}$, ou seja, $Y = \mathcal{F}(\mathbf{X})$ para $\mathcal{F} : \{0, 1\}^n \rightarrow \{0, 1\}$.

Esta seção supõe que a lógica de predição é o único fator relativo ao *IMP score*. Para tanto, assume-se que o poder preditivo de $\mathbf{X}$ com relação a Y seja o maior possível, ou seja, $P(Y = 1|\mathbf{x}) = 0$ ou $P(Y = 1|\mathbf{x}) = 1$, para todo $\mathbf{x} \in \mathbf{X}$. Isso é equivalente a dizer que o erro ao classificar a variável Y com base no conhecimento de $\mathbf{X}$ ($\varepsilon_Y(\mathbf{X})$) é nulo (predição determinística). Além disso, assumiremos que todas as instâncias de $\mathbf{X}$ são equiprováveis, isto é, $P(\mathbf{X} = \mathbf{x}) = 1/2^n$ para todo $\{x_1, x_2, \dots, x_n\} \in \mathbf{X}$. Neste caso, quando nos referimos a uma lógica de predição particular, por definição essa lógica provê a melhor predição dentre todos os preditores possíveis. Como $\varepsilon_Y(\mathbf{X}) = 0$, temos que $\text{CoD}_Y(\mathbf{X}) = 1$. Sob essas condições, pode-se definir três classes de lógicas com respeito ao *IMP score*:

1. **Lógica IMP:** uma lógica de predição é IMP se $\max_{\mathbf{Z} \subseteq \mathbf{X}} \text{CoD}_Y(\mathbf{Z}) = 0$;
2. **Lógica anti-IMP:** uma lógica de predição é anti-IMP se $\max_{\mathbf{Z} \subseteq \mathbf{X}} \text{CoD}_Y(\mathbf{Z}) = 1$;
3. **Lógica semi-IMP:** uma lógica de predição é semi-IMP se $0 < \max_{\mathbf{Z} \subseteq \mathbf{X}} \text{CoD}_Y(\mathbf{Z}) < 1$.

Para dois preditores binários, o número de lógicas possíveis é $2^4 = 16$ porque existem 4 instâncias (ou mintermos) possíveis, $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$, e para cada um desses mintermos, o alvo Y assume um valor binário. Definimos lógicas de k -mintermo como sendo as lógicas que possuem k mintermos em que o valor de saída é verdadeiro. Por exemplo, o AND é chamado de lógica de 1-mintermo enquanto que o XOR é chamado de lógica de 2-mintermo.

Cada lógica pode ser representada por uma palavra composta por 4 número binários na qual o primeiro número corresponde ao valor de Y quando $(0, 0)$ é observado, o segundo quando $(0, 1)$ é observado, o terceiro quando $(1, 0)$ é observado, e o último quando $(1, 1)$ é observado. Por exemplo, a lógica AND é expressa pela palavra 0001 e a lógica XOR pela palavra 0110. Para as lógicas constantes 0000 e 1111, o alvo é insensível à variabilidade

dos preditores. Isso corresponde ao caso degenerado no qual o erro ε_Y na ausência de observações já é nulo assim como o erro $\varepsilon_Y(\mathbf{Z})$ de qualquer conjunto de preditores $\mathbf{Z} \subsetneq \mathbf{X}$. Por definição, os CoDs são todos iguais a 1 nesse caso, implicando que as lógicas constantes são anti-IMP.

Das 14 lógicas restantes, 4 são de 1-mintermo. Essas lógicas estão listadas a seguir.

- $Y = X_1 \wedge X_2$ (0001) (AND);
- $Y = X_1 \wedge \bar{X}_2$ (0010);
- $Y = \bar{X}_1 \wedge X_2$ (0100);
- $Y = \bar{X}_1 \wedge \bar{X}_2$ (1000) (OR-negado ou NOR).

De forma análoga, existem 4 lógicas de 3-mintermo. Basta trocar os 1's por 0's e 0's por 1's nas palavras representativas das lógicas 1-mintermo para obter as lógicas 3-mintermo. Essas lógicas são listadas a seguir.

- $Y = \bar{X}_1 \vee \bar{X}_2$ (1110) (AND-negado ou NAND);
- $Y = \bar{X}_1 \vee X_2$ (1101);
- $Y = X_1 \vee \bar{X}_2$ (1011);
- $Y = X_1 \vee X_2$ (0111) (OR).

Repare que as lógicas 0010 e 0100 são redundantes com relação à ordem das variáveis, ou seja, elas são a mesma lógica mas com a ordem das variáveis trocada. Analogamente, 1101 e 1011 também são redundantes entre si.

Para uma lógica de predição qualquer, sua função preditora ótima (isto é, seguindo o princípio do classificador Bayesiano) $\psi_Y(\mathbf{X})$ pode ser descrita como:

$$\psi_Y(\mathbf{X}) = \begin{cases} 1, & \text{se } P(Y = 1|\mathbf{X}) > 0.5 \\ 0, & \text{c.c.} \end{cases} \quad (4.3)$$

A partir dessa definição, e levando em conta que todas as instâncias de (X_1, X_2) são equiprováveis (ou seja, cada instância ocorre com probabilidade 0.25), todas as lógicas

1-mintermo e 3-mintermo são IMP. As tabelas de probabilidades conjuntas contidas na Figura 4.2(a) ilustram a razão da lógica AND ser IMP. O $CoD_Y(X_1, X_2)$ é 1, já que o erro *a priori* ε_Y é 0.25, pois $P(Y = 0) = 0.75$ e $P(Y = 1) = 0.25$ e, além disso, $P(Y = 0|\mathbf{x}) = 0$ para todo $\mathbf{x} \in \mathbf{X}$ ($CoD_Y(X_1, X_2) = (0.25 - 0)/0.25 = 1$). Por outro lado, $CoD_Y(X_1) = 0$ pois $P(Y = 1|X_1 = 0) = 0$ e como $P(Y = 1|X_1 = 1) = P(Y = 0|X_1 = 1) = 0.25$, temos $CoD_Y(X_1) = (0.25 - 0.25)/0.25 = 0$. Pelo mesmo motivo, $CoD_Y(X_2) = 0$. Como o máximo dos CoDs individuais é 0 e o CoD do todo é 1, a lógica AND é IMP. As outras lógicas de 1-mintermo e 3-mintermo também são IMP por serem simétricas à lógica AND.

Como mencionado na Seção 4.2, a lógica AND é um exemplo de função de canalização. De fato, todas as lógicas de 1-mintermo e $2^n - 1$ -mintermo para $n \geq 2$, sendo n o número de preditores, são funções de canalização. Isso porque essas lógicas têm a particularidade em que apenas uma linha da tabela de probabilidades condicionais é diferente das demais (ou equivalentemente, um caractere é diferente dos demais na palavra binária que representa a lógica). Para que o conjunto de preditores esteja justamente nesse estado distinto, é necessário que cada um dos preditores possua um valor em específico. Caso pelo menos um desses preditores não atenda esse valor específico, o estado da variável alvo será completamente definido por tal variável (função de canalização). Da mesma forma, pode-se mostrar que essas mesmas lógicas também são de predição intrinsecamente multivariada, pois nenhum subconjunto próprio dos preditores que possuam esses valores específicos é capaz de determinar o estado da variável alvo sem a informação das variáveis restantes. Conclui-se então que toda função canalizadora é uma lógica IMP. Porém, a volta dessa afirmação nem sempre é verdadeira, ou seja, nem todas as lógicas IMP são funções canalizadoras (ver discussão das lógicas 2-mintermo a seguir).

Agora resta apenas analisar as lógicas 2-mintermo. Existem 6 possíveis (combinação de 4, 2 a 2):

- $Y = X_1$ (0011);
- $Y = X_2$ (0101);
- $Y = \bar{X}_1$ (1100);
- $Y = \bar{X}_2$ (1010);
- $Y = X_1 \oplus X_2$ (0110, XOR – OU-exclusivo);

- $Y = X_1 \oplus X_2$ (1001, NXOR – XOR negado).

Nas primeiras quatro lógicas 2-mintermo, Y é completamente determinado por um dos preditores, resultando em $\text{CoD}_Y(X_1) = 1$ ou $\text{CoD}_Y(X_2) = 1$, tornando-as anti-IMP. As últimas duas lógicas, XOR e NXOR, são IMP porque $\text{CoD}_Y(X_1) = \text{CoD}_Y(X_2) = 0$ em ambas. A Figura 4.2(b) contém tabelas de probabilidades conjuntas com seus respectivos CoDs que ilustram porque a lógica XOR é IMP.

Figura 4.2: Tabelas de probabilidades conjuntas para as lógicas AND e XOR, mostrando o porquê dessas lógicas serem IMP. CoD₁, CoD₂ e CoD₁₂ referem-se respectivamente a $\text{CoD}_Y(X_1)$, $\text{CoD}_Y(X_2)$ e $\text{CoD}_Y(X_1, X_2)$.

Note que as lógicas XOR e NXOR, apesar de pertencerem à categoria de lógicas IMP, não são funções canalizadoras, já que o estado da variável só é definido pelo conhecimento dos valores das duas variáveis.

Assim, no caso de dois preditores binários, existem apenas duas classes possíveis de lógicas: as anti-IMPs (lógicas constantes e de 2-mintermo, exceto XOR e NXOR) e as IMPs (1-mintermo, 3-mintermo, XOR e NXOR).

As lógicas semi-IMPs só aparecem para conjuntos com três ou mais preditores. De fato, para o caso de três preditores, das 256 (2^8) lógicas possíveis, 66 são IMP, 152 são semi-IMP e as restantes 38 são anti-IMP. A classe IMP é composta de todas as 8 lógicas 1-mintermo (todas as 8 7-mintermo por simetria), 16 lógicas 2-mintermo (16 lógicas 6-mintermo), 8 lógicas 3-mintermo (8 5-mintermo) e 2 lógicas 4-mintermo. A classe semi-IMP contém 48 lógicas 3-mintermo (48 lógicas 5-mintermo por simetria), todas com $\max_{\mathbf{Z} \subseteq \mathbf{X}} \text{CoD}_Y(\mathbf{Z}) = 0.67$, e 56 lógicas 4-mintermo, todas com $\max_{\mathbf{Z} \subseteq \mathbf{X}} \text{CoD}_Y(\mathbf{Z}) = 0.5$. Finalmente, a classe

anti-IMP é composta por lógicas constantes (00000000, 11111111), 12 lógicas 2-mintermo (12 6-mintermo, por simetria) e 12 lógicas 4-mintermo. Mais detalhes sobre essas lógicas, bem como seus respectivos CoDs, podem ser encontrados em (http://www.vision.ime.usp.br/~davidjr/imp_report/logicas.html).

No caso determinístico que está sendo considerado nesta seção, o valor do $\text{CoD}_Y(\mathbf{X})$ é 1 para todos os casos e a lógica de predição é o único fator que pode produzir características IMP. A situação é diferente na próxima seção (Seção 4.5), que considera o poder preditivo em lógicas estocásticas.

4.5 Poder preditivo

Aqui será considerado o caso em que a predição não é determinística. As lógicas corresponderão a tabelas verdade estocásticas nas quais todos os valores de saída são da forma p ou $1 - p$, sendo $0 \leq p \leq 1$. Os casos $p = 0$ e $p = 1$ produzem lógicas determinísticas, então o presente contexto é uma generalização estocástica do caso Booleano clássico discutido na seção anterior. Este modelo é relacionado, porém distinto, ao modelo “noisy-OR” bastante usado em redes de inferência Bayesiana [Murphy and Mian, 1999, Pearl, 1988]. Assumindo $p \geq 0.5$, temos $\varepsilon_Y(\mathbf{X}) = 1 - p$. Dessa forma, p mede o poder preditivo, isto é, a precisão do conjunto de preditores em prever o comportamento do alvo.

O impacto do poder preditivo p sobre as lógicas capazes de produzir IMP no caso de dois preditores será descrito analiticamente nesta seção (a Seção 4.5.2 fará uma análise superficial do caso de três preditores). Como visto na Seção 4.4, as lógicas 1-mintermo, 3-mintermo, XOR e NXOR são aquelas que possibilitam a geração de IMP. Para analisar o impacto do poder preditivo para dois preditores, daqui para frente, sem perda de generalidade, serão consideradas apenas as lógicas AND e XOR, já que a análise de IMP para outras lógicas pode ser reduzida à análise dessas duas lógicas.

Ao longo desta seção será adotada a seguinte notação simplificada: $\varepsilon_0 = \varepsilon_Y$, $\varepsilon_1 = \varepsilon_Y(X_1)$, $\varepsilon_2 = \varepsilon_Y(X_2)$, e $\varepsilon_{12} = \varepsilon_Y(X_1, X_2)$. Similarmente, $\text{CoD}_1 = \text{CoD}_Y(X_1)$, $\text{CoD}_2 = \text{CoD}_Y(X_2)$, e $\text{CoD}_{12} = \text{CoD}_Y(X_1, X_2)$.

O modelo de probabilidades para $X_1, X_2, Y \in \{0, 1\}$ pode ser descrito em termos de 8 parâmetros:

$$P(Y = 1|X_1 = i, X_2 = j) = \pi_{ij}, \quad i, j = 0, 1$$

$$P(X_1 = i, X_2 = j) = \alpha_{ij}, \quad i, j = 0, 1$$

Os parâmetros α_{ij} especificam a distribuição conjunta de probabilidades de X_1, X_2 (apenas três desses parâmetros são independentes já que um deles é 1 subtraído da soma dos demais), enquanto os parâmetros π_{ij} descrevem a tabela verdade estocástica da lógica de predição. Se todos os parâmetros π_{ij} são 0 ou 1, então este seria o caso de uma lógica determinística (Seção 4.4).

O erro do preditor ótimo de Y na ausência de preditores é dado por

$$\varepsilon_0 = \min\{P(Y = 1), P(Y = 0)\} = F[P(Y = 1)] = F\left(\sum_{ij} \alpha_{ij} \pi_{ij}\right)$$

em que $F : [0, 1] \rightarrow [0, 1]$ é dada por $F(x) = \min\{x, 1 - x\}$. A função F introduz uma não-linearidade na dependência do CoD nos parâmetros do modelo. Essa função tem uma importante propriedade de simetria: $F(p) = F(1 - p)$, para todo $0 \leq p \leq 1$ (ver Figura 4.3).

Figura 4.3: Gráfico de não-linearidade $F(x)$.

O erro ótimo para X_1 predizendo Y é

$$\varepsilon_1 = \sum_{i=0}^1 F[P(Y = 1|X_1 = i)]P(X_1 = i) = \sum_{i=0}^1 F\left(\frac{\pi_{i0}\alpha_{i0} + \pi_{i1}\alpha_{i1}}{\alpha_{i0} + \alpha_{i1}}\right)(\alpha_{i0} + \alpha_{i1}),$$

e, similarmente, o erro ótimo para X_2 predizendo Y é

$$\varepsilon_2 = \sum_{i=0}^1 F[P(Y = 1|X_2 = i)]P(X_2 = i) = \sum_{j=0}^1 F\left(\frac{\pi_{0j}\alpha_{0j} + \pi_{1j}\alpha_{1j}}{\alpha_{0j} + \alpha_{1j}}\right)(\alpha_{0j} + \alpha_{1j}).$$

Finalmente, o erro ótimo para X_1 e X_2 predizendo Y é

$$\varepsilon_{12} = E\{F[P(Y = 1|X_1, X_2)]\} = \sum_{ij} F(\pi_{ij})\alpha_{ij} \quad (4.4)$$

O modelo de probabilidade proposto considera tabelas parametrizadas por uma variável p , com $0 \leq p \leq 1$ tal que todas as entradas são $\pi_{ij} = p$ ou $\pi_{ij} = 1 - p$. Neste caso, aplicando a propriedade de simetria $F(p) = F(1-p)$, segue da Equação 4.4 que $\varepsilon_{12} = F(1-p)$. Assumindo $p \geq 0.5$, o preditor ótimo baseado no par (X_1, X_2) é $\varepsilon_{12} = 1 - p$, isto é, p é o poder preditivo de (X_1, X_2) com relação a Y .

Os CoDs baseados em X_1 , X_2 e (X_1, X_2) são, respectivamente,

$$\begin{aligned} \text{CoD}_1 &= 1 - \frac{\varepsilon_1}{\varepsilon_0} = 1 - \frac{\sum_{i=0}^1 F\left(\frac{\pi_{i0}\alpha_{i0} + \pi_{i1}\alpha_{i1}}{\alpha_{i0} + \alpha_{i1}}\right)(\alpha_{i0} + \alpha_{i1})}{F\left(\sum_{ij} \alpha_{ij}\pi_{ij}\right)}, \\ \text{CoD}_2 &= 1 - \frac{\varepsilon_2}{\varepsilon_0} = 1 - \frac{\sum_{j=0}^1 F\left(\frac{\pi_{0j}\alpha_{0j} + \pi_{1j}\alpha_{1j}}{\alpha_{0j} + \alpha_{1j}}\right)(\alpha_{0j} + \alpha_{1j})}{F\left(\sum_{ij} \alpha_{ij}\pi_{ij}\right)}, \\ \text{CoD}_{12} &= 1 - \frac{\varepsilon_{12}}{\varepsilon_0} = 1 - \frac{\sum_{ij} F(\pi_{ij})\alpha_{ij}}{F\left(\sum_{ij} \alpha_{ij}\pi_{ij}\right)} = 1 - \frac{F(1-p)}{F\left(\sum_{ij} \alpha_{ij}\pi_{ij}\right)} \end{aligned} \quad (4.5)$$

4.5.1 Preditores com distribuição uniforme

Agora, assumindo que a distribuição de (X_1, X_2) é uniforme, isto é, $\alpha_{00} = \alpha_{01} = \alpha_{10} = \alpha_{11} = \frac{1}{4}$, as equações anteriores se reduzem a:

$$\begin{aligned}
\text{CoD}_1 &= 1 - \frac{1}{2} \frac{F\left(\frac{\pi_{00}+\pi_{01}}{2}\right) + F\left(\frac{\pi_{10}+\pi_{11}}{2}\right)}{F\left(\frac{\pi_{00}+\pi_{01}+\pi_{10}+\pi_{11}}{4}\right)} \\
\text{CoD}_2 &= 1 - \frac{1}{2} \frac{F\left(\frac{\pi_{00}+\pi_{10}}{2}\right) + F\left(\frac{\pi_{01}+\pi_{11}}{2}\right)}{F\left(\frac{\pi_{00}+\pi_{01}+\pi_{10}+\pi_{11}}{4}\right)} \\
\text{CoD}_{12} &= 1 - \frac{1}{4} \frac{F(\pi_{00}) + F(\pi_{01}) + F(\pi_{10}) + F(\pi_{11})}{F\left(\frac{\pi_{00}+\pi_{01}+\pi_{10}+\pi_{11}}{4}\right)}
\end{aligned} \tag{4.6}$$

Nas duas próximas subseções, a análise será considerada para a lógica AND, que representa todas as lógicas 1- e 3-mintermo, e a lógica XOR, que representa as lógicas XOR e NXOR em termos de IMP.

Lógica AND

No caso da lógica AND, $\pi_{00} = \pi_{01} = \pi_{10} = 1 - p$ e $\pi_{11} = p$, com $0.5 \leq p \leq 1$. Substituindo estas igualdades na Equação 4.6, e usando a suposição $p \geq 0.5$, temos que:

$$\text{CoD}_1 = \text{CoD}_2 = 1 - \frac{1}{2} \frac{F(1-p) + F\left(\frac{1}{2}\right)}{F\left(\frac{3-2p}{4}\right)} = 1 - \frac{1}{2} \frac{1-p + \frac{1}{2}}{\frac{3-2p}{4}} = 0,$$

enquanto o CoD_{12} é dado por:

$$\text{CoD}_{12} = 1 - \frac{1}{4} \frac{3F(1-p) + F(p)}{F\left(\frac{3-2p}{4}\right)} = 1 - \frac{4(1-p)}{3-2p} = \frac{p - \frac{1}{2}}{1 - \left(p - \frac{1}{2}\right)}. \tag{4.7}$$

Assim, o CoD_{12} é não-linear como uma função do poder preditivo p , com CoD_{12} tendendo a 1 na medida que p tende a 1, e CoD_{12} tendendo a 0 para p tendendo a $\frac{1}{2}$ (ver Figura 4.4(a)). Já que $\text{CoD}_1 = \text{CoD}_2 = 0$, $\mathbf{X}$ é IMP para Y com respeito a $\lambda = 0$ e δ se e somente se

$$\text{CoD}_{12} = \frac{p - \frac{1}{2}}{1 - \left(p - \frac{1}{2}\right)} \geq \delta \Rightarrow p \geq \frac{1}{2} \frac{1 + 3\delta}{1 + \delta}.$$

Por exemplo, para $\delta = 0.8$, a condição IMP é $p \geq \frac{17}{18} = 0.944 \dots$

Lógica XOR

No caso da lógica XOR, $\pi_{00} = \pi_{11} = 1 - p$ e $\pi_{01} = \pi_{10} = p$, com $0.5 \leq p \leq 1$. Substituindo estas igualdades na Equação 4.6, e usando a suposição $p \geq 0.5$, temos:

$$\text{CoD}_1 = \text{CoD}_2 = 1 - \frac{1}{2} \frac{F\left(\frac{1}{2}\right) + F\left(\frac{1}{2}\right)}{F\left(\frac{1}{2}\right)} = 1 - \frac{1}{2} \frac{\frac{1}{2} + \frac{1}{2}}{\frac{1}{2}} = 0,$$

enquanto o CoD_{12} é dado por

$$\text{CoD}_{12} = 1 - \frac{1}{4} \frac{2F(p) + 2F(1-p)}{F\left(\frac{1}{2}\right)} = 1 - \frac{1}{4} \frac{4(1-p)}{\frac{1}{2}} = 2 \left(p - \frac{1}{2} \right). \quad (4.8)$$

O CoD_{12} neste caso é uma função linear do poder de predição p . Similarmente ao caso AND, CoD_{12} tende a 1 quando p tende a 1, e CoD_{12} tende a 0 para p tendendo a $\frac{1}{2}$ (ver Figura 4.4(b)). Novamente, os CoDs individuais são nulos, então $\mathbf{X}$ é IMP para Y com relação a $\lambda = 0$ e δ se e somente se:

$$\text{CoD}_{12} = 2 \left(p - \frac{1}{2} \right) \geq \delta \Rightarrow p \geq \frac{1}{2} + \frac{\delta}{2} \quad (4.9)$$

Por exemplo, com $\delta = 0.8$, a condição IMP é $p \geq 0.9$, que é menos restritiva do que no caso AND com o mesmo δ . Os gráficos dos valores críticos de p como uma função de δ , ilustrados na Figura 4.5, mostram que a condição IMP é sempre mais restritiva para a lógica AND que para a lógica XOR.

4.5.2 Preditores com distribuição não-uniforme

Nesta seção será considerada a situação em que a distribuição conjunta dos preditores não é uniforme. Veremos que o CoD pode ser fortemente afetado por essa distribuição. Os únicos casos em que o CoD_{12} não é afetado pela distribuição dos preditores é quando $p = 1$ (máximo poder preditivo) e $p = 0.5$ (poder preditivo nulo), com $\text{CoD}_{12} = 1$ e $\text{CoD}_{12} = 0$, respectivamente, independentemente da lógica e da distribuição dos preditores. Para $0.5 < p < 1$, as lógicas AND e XOR serão examinadas para observar o comportamento dos CoDs individuais e do CoD global. Primeiramente, o caso especial em que os preditores são independentes para conjuntos de dois e três preditores será discutido. Em seguida, o caso geral em que não se assume independência entre preditores será analisado em detalhes

Figura 4.4: CoD_{12} para as lógicas AND e XOR como função do poder preditivo p , com preditores uniformes.

Figura 4.5: Valor crítico de p para IMP como uma função de δ para as lógicas AND e XOR com preditores uniformes, mostrando que a condição IMP é sempre mais restritiva para a lógica AND do que para a lógica XOR.

apenas para conjuntos de dois preditores devido ao excesso de parâmetros inerente ao caso de três preditores.

Dois preditores independentes

Considerando o caso de dois preditores, seja $P_1 = P(X_1 = 1)$ e $P_2 = P(X_2 = 1)$. Dado o poder preditivo p para (X_1, X_2) em relação a Y , as tabelas de distribuição conjunta de probabilidades para as lógicas AND e XOR são mostradas nas Tabelas 4.3 e 4.4 respectivamente.

Tabela 4.3: Tabela de distribuição conjunta de probabilidades para a lógica AND assumindo independência entre X_1 e X_2 .

X_1	X_2	$P(Y = 0, X_1, X_2)$	$P(Y = 1, X_1, X_2)$
0	0	$(1 - P_1)(1 - P_2)p$	$(1 - P_1)(1 - P_2)(1 - p)$
0	1	$(1 - P_1)P_2p$	$(1 - P_1)P_2(1 - p)$
1	0	$P_1(1 - P_2)p$	$P_1(1 - P_2)(1 - p)$
1	1	$P_1P_2(1 - p)$	P_1P_2p

Tabela 4.4: Tabela de distribuição conjunta de probabilidades para a lógica XOR assumindo independência entre X_1 e X_2 .

X_1	X_2	$P(Y = 0, X_1, X_2)$	$P(Y = 1, X_1, X_2)$
0	0	$(1 - P_1)(1 - P_2)p$	$(1 - P_1)(1 - P_2)(1 - p)$
0	1	$(1 - P_1)P_2(1 - p)$	$(1 - P_1)P_2p$
1	0	$P_1(1 - P_2)(1 - p)$	$P_1(1 - P_2)p$
1	1	P_1P_2p	$P_1P_2(1 - p)$

Note que $\alpha_{00} = (1 - P_1)(1 - P_2)$, $\alpha_{01} = (1 - P_1)P_2$, $\alpha_{10} = P_1(1 - P_2)$, e $\alpha_{11} = P_1P_2$. Para a lógica AND, $\pi_{00} = \pi_{01} = \pi_{10} = 1 - p$ e $\pi_{11} = p$, com $0.5 \leq p \leq 1$. Substituindo esses valores na Equação 4.5, e assumindo $p \geq 0.5$, os CoDs para a lógica AND são dados por:

$$\begin{aligned}
\text{CoD}_1 &= 1 - \frac{(1 - P_1)(1 - p) + F[P_2(1 - 2p) + p]P_1}{F[P_1P_2(1 - 2p) + p]} \\
\text{CoD}_2 &= 1 - \frac{(1 - P_2)(1 - p) + F[P_1(1 - 2p) + p]P_2}{F[P_1P_2(1 - 2p) + p]} \\
\text{CoD}_{12} &= 1 - \frac{1 - p}{F[P_1P_2(1 - 2p) + p]}
\end{aligned} \tag{4.10}$$

Para a lógica XOR, temos que $\pi_{00} = \pi_{11} = 1 - p$ e $\pi_{01} = \pi_{10} = p$, com $0.5 \leq p \leq 1$. Assim, seus CoDs são dados por:

$$\begin{aligned}
\text{CoD}_1 &= 1 - \frac{F[P_2(1 - 2p) + p]}{F[P_1(1 - P_2) + P_2(1 - P_1) + (2P_1 - 1)(2P_2 - 1)p]} \\
\text{CoD}_2 &= 1 - \frac{F[P_1(1 - 2p) + p]}{F[P_1(1 - P_2) + P_2(1 - P_1) + (2P_1 - 1)(2P_2 - 1)p]} \\
\text{CoD}_{12} &= 1 - \frac{1 - p}{F[P_1(1 - P_2) + P_2(1 - P_1) + (2P_1 - 1)(2P_2 - 1)p]}
\end{aligned} \tag{4.11}$$

Note que, para $p = \frac{1}{2}$, o *IMP score* é zero para ambas as lógicas, não importando os valores de P_1 e P_2 (isto corresponde ao caso de poder preditivo nulo).

Com essas equações em mente, dados o poder preditivo e a lógica, pode-se gerar gráficos com os valores $0 < P_1 < 1$ no eixo X, $0 < P_2 < 1$ no eixo Y, e associando a cada ponto (P_1, P_2) uma intensidade em níveis de cinza correspondente a $\max\{\text{CoD}_1, \text{CoD}_2\}$ e a CoD_{12} , formando superfícies. Mais do que isso, pode-se juntar os dois gráficos em um único contendo dois canais de intensidade: vermelho para $1 - \max\{\text{CoD}_1, \text{CoD}_2\}$ e verde para CoD_{12} , sendo que o amarelo indica as regiões mais suscetíveis à geração de conjuntos IMPs (o amarelo indica alto CoD_{12} e baixo $\max\{\text{CoD}_1, \text{CoD}_2\}$ ao mesmo tempo). As Figuras 4.6 e 4.7 ilustram essas gráficos para as lógicas AND e XOR, repectivamente, para três valores de p . Note que para a lógica AND, a região mais favorável ao surgimento de conjuntos IMP é definida por $P_1 \leq 0.5$ e $P_2 \leq 0.5$ (quadrado inferior esquerdo), mas poderes preditivos menores encolhem esta região para $P_1 \approx 0.5$ and $P_2 \approx 0.5$. Para a lógica XOR, a região mais favorável é definida por $P_1 \approx P_2$ e $P_1 \approx 1 - P_2$ (diagonais) exceto para $P_1 \approx 0$ ou $P_1 \approx 1$.

As lógicas 1-mintermo e 3-mintermo são simétricas por rotação da lógica AND, como pode ser visto na Figura 4.8, que mostra quatro gráficos simétricos compostos por canais

Figura 4.6: Gráficos representando $\max\{\text{CoD}_1, \text{CoD}_2\}$ (primeira linha), CoD_{12} (segunda linha) e a junção de $1 - \max\{\text{CoD}_1, \text{CoD}_2\}$ com o CoD_{12} para a lógica AND e três valores distintos de p (0.90, 0.95 and 0.99) assumindo que X_1 e X_2 são independentes. As probabilidades P_1 e P_2 são dadas pelos eixos X e Y respectivamente. Nos gráficos das duas primeiras linhas, cores escuras correspondem a valores menores. Na última linha de gráficos, a cor vermelha indica que $\max\{\text{CoD}_1, \text{CoD}_2\}$ e CoD_{12} possuem valores baixos, a cor verde indica que esses dois valores são altos e, finalmente, quanto mais próximo de amarelo, menor será o valor do primeiro e maior será o valor do segundo. Assim, as regiões de interesse para geração de conjuntos IMP são compostas pela intersecção das regiões escuras nos gráficos da primeira linha com as claras nos da segunda linha. Essas regiões de intersecção são indicadas pela cor amarela nos gráficos da terceira linha.

verde e vermelho e as lógicas associadas a eles (em notação de palavra binária), para $p = 0.99$. Para as lógicas XOR e NXOR, os gráficos são exatamente os mesmos para todos os valores de p .

Há três observações importantes baseadas nos gráficos ilustrados nas figuras 4.6, 4.7 e

Figura 4.7: Gráficos representando $\max\{\text{CoD}_1, \text{CoD}_2\}$ (primeira linha), CoD_{12} (segunda linha) e a junção de $1 - \max\{\text{CoD}_1, \text{CoD}_2\}$ com o CoD_{12} para a lógica XOR e três valores distintos de p (0.90, 0.95 and 0.99) assumindo que X_1 e X_2 são independentes. As probabilidades P_1 e P_2 são dadas pelos eixos X e Y respectivamente. Nos gráficos das duas primeiras linhas, cores escuras correspondem a valores menores. Na última linha de gráficos, a cor vermelha indica que $\max\{\text{CoD}_1, \text{CoD}_2\}$ e CoD_{12} possuem valores baixos, a cor verde indica que esses dois valores são altos e, finalmente, quanto mais próximo de amarelo, maior será o valor do primeiro e menor será o valor do segundo. Assim, as regiões de interesse para geração de conjuntos IMP são compostas pela intersecção das regiões escuras nos gráficos da primeira linha com as claras nos da segunda linha. Essas regiões de intersecção são indicadas pela cor amarela nos gráficos da terceira linha.

4.8. A primeira delas é que, para todos os casos, as regiões centrais ($P_1 \approx 0.5$ and $P_2 \approx 0.5$) são aquelas com maior chance de produzir conjuntos IMP, indicando que uniformidade na distribuição de probabilidade dos preditores facilita a geração de conjuntos IMP. Segundo, os gráficos ilustram que o poder preditivo p exerce um efeito dramático na intensidade do CoD global. Mesmo para $p = 0.9$, que é um bom poder preditivo, a intensidade das regiões de cor amarela declina substancialmente. Uma última observação importante é que os maiores valores do CoD global obtidos para a lógica XOR são maiores que os obtidos pela lógica AND para o mesmo valor de p . Portanto, se o poder preditivo não

Figura 4.8: Lógicas de 1,3-mintermo e seus respectivos gráficos simétricos da junção dos valores de CoD_{12} e $1 - \max\{\text{CoD}_1, \text{CoD}_2\}$ em canais verde e vermelho. O poder preditivo é $p = 0.99$ e as lógicas correspondentes a cada gráfico estão referenciadas pela notação de palavra binária.

é tão forte ($p < 0.95$), então a lógica XOR (ou NXOR) é exigida para obter um *IMP score* substancial. Por outro lado, se o poder preditivo é bastante forte ($p \approx 1$), conjuntos IMP podem ser obtidos mais facilmente através de lógicas de 1,3-mintermo, pois uma considerável região do gráfico possui alta intensidade de amarelo (alto *IMP score*).

Três preditores independentes

A mesma análise feita para o caso de dois preditores independentes pode ser feita para três preditores, exceto que agora a análise não pode ser realizada através de visualização de superfícies, mas apenas por estatísticas sobre o espaço de parâmetros $(P(X_1 = 1), P(X_2 = 1), P(X_3 = 1)) = (P_1, P_2, P_3)$. Como discutido na Seção 4.4, para conjuntos de três preditores existem $2^8 = 256$ lógicas possíveis com 66 lógicas IMP, 152 semi-IMP e 38 anti-IMP. Felizmente, os conjuntos IMP e semi-IMP podem ser representados por poucas lógicas, com as lógicas restantes sendo obtidas por simetria e redundância pela ordem dos preditores. Um ponto crucial relacionado com independência entre dois preditores é que a uniformidade na distribuição de probabilidades conjuntas dos preditores oferece as melhores condições para geração de conjuntos IMP. Assim, é importante observar se essa uniformidade também contribui para a geração de IMPs no caso de três preditores. Como medida de uniformidade, será adotada a entropia de Shannon definida no Capítulo 2, Equação 2.8.

Para um dado poder preditivo p fixo, o interesse reside em calcular o λ mínimo e o δ máximo necessários para gerar IMPs para cada tripla considerada (P_1, P_2, P_3) com o

objetivo de verificar se a entropia da distribuição de probabilidades dos preditores está correlacionada com o $\delta - \lambda$, já que o *IMP score* aumenta com esta diferença. Considere a tripla de parâmetros (P_1, P_2, P_3) no intervalo $(0.01 \leq P_1, P_2, P_3 \leq 0.99)$ com granularidade 0.01 resultando num total de 970,299 pontos nesse espaço de parâmetros. Como cada ponto tem uma entropia associada e o mínimo λ e máximo δ para produzir IMPs, a idéia é gerar duas tabelas para cada lógica p considerada. A primeira tabela contém as contagens de pontos pertencentes a cada subconjunto determinado por $\lambda \leq l$ e $\delta \geq d$ para $0.1 \leq l \leq 1$ e $0 \leq d \leq 0.9$ com granularidade 0.1. A segunda tabela contém a entropia média das distribuições dos preditores para cada subconjunto.

Começando com as lógicas IMP, existem cinco subconjuntos de lógicas com os mesmos resultados em termos das tabelas descritas anteriormente, um para lógicas 1-mintermo e 7-mintermo, dois para 2-mintermo e 6-mintermo, um para 3-mintermo e 5-mintermo e um para 4-mintermo. A Tabela 4.5 mostra gráficos contendo quantidade de conjuntos IMP e entropia média das distribuições dos preditores para $p = \{0.9, 0.95, 0.99\}$ e para a lógica 00000001 representando lógicas 1,7-mintermo. Os resultados em formato de tabela numérica correspondentes a essa lógica e as outras quatro lógicas IMP podem ser vistos em http://www.vision.ime.usp.br/~davidjr/imp_report/imp_entropy_report.html. A partir desses mesmos resultados, é possível construir gráficos do tipo “*scatter plot*” da entropia H contra *IMP score* I , já que cada ponto (P_1, P_2, P_3) possui essas duas medidas associadas (o *IMP score* como sendo $\delta - \lambda$ com δ máximo e λ mínimo para os quais (P_1, P_2, P_3) gera conjuntos IMP). A Figura 4.9 mostra esses gráficos correspondentes aos resultados mostrados na Tabela 4.5. A partir desses resultados, pode-se observar que a entropia dos preditores e os *IMP scores* são bem correlacionados. Geralmente, quanto menor o poder preditivo, maior a correlação. Se um conjunto de preditores tem um poder preditivo que não é forte o suficiente ($p < 0.95$), então a entropia da distribuição dos preditores deve ser quase máxima para ser IMP. Uma última observação é que um menor poder preditivo leva a uma diminuição na quantidade de *IMP scores* na faixa $0.8 \leq I \leq 1$, já que o CoD global é afetado diretamente pelo poder preditivo.

As lógicas semi-IMP também possuem 5 subconjuntos de lógicas em termos dos resultados de número de conjuntos IMP, entropia média e correlação entre H e I , dois para lógicas 3-mintermo e 5-mintermo e três para lógicas 4-mintermo. O mesmo procedimento é aplicado a elas. A Tabela 4.6 mostra os gráficos contendo os resultados

Tabela 4.5: Número de conjuntos IMP e entropia média das distribuições dos preditores para cada subespaço do espaço de parâmetros (P_1, P_2, P_3) definido por λ e δ para $p = 0.90$, $p = 0.95$ e $p = 0.99$ e lógica 00000001. Os eixos X e Y correspondem aos valores de δ e λ respectivamente. As legendas estão dispostas na última coluna (o número de conjuntos IMPs é dado por potências de 2 e as entropias estão no intervalo de 0 a 1).

	$p = 0.90$	$p = 0.95$	$p = 0.99$	Legenda
Número de IMPs				
Entropia média

$p = 0.90$

$\text{Corr}(H, I) = 0.8245$

$p = 0.95$

$\text{Corr}(H, I) = 0.8256$

$p = 0.99$

$\text{Corr}(H, I) = 0.7245$

Figura 4.9: “Scatter plots” do IMP score (I) contra entropia (H) com 970.299 pontos do espaço de parâmetros (P_1, P_2, P_3) para $p = 0.90$, $p = 0.95$ e $p = 0.99$ e lógica 00000001, (lógica IMP). As correlações respectivas estão embaixo de cada um dos gráficos.

do número de conjuntos IMP e a entropia média para cada intervalo de valores definido por λ e δ para a lógica 00011001. Os resultados em formato de tabela numérica correspondentes a essa lógica e para as outras quatro lógicas semi-IMP representantes podem ser vistas em http://www.vision.ime.usp.br/~davidjr/imp_report/partly_

`imp_entropy_report.html`. Os gráficos do tipo “*scatter plot*” da entropia H contra IMP score I correspondentes estão ilustrados na Figura 4.10. Esses resultados mostram que as correlações da entropia com o IMP score para as lógicas semi-IMP são positivas mas significativamente menores que as obtidas pelas lógicas IMP. Outra diferença é que a altura do “*scatter plot*” é menor que $I = 0.7$, mesmo para $p \approx 1$, tal como $p = 0.99$. De fato, essa classe não apresenta um padrão bem definido para a altura do “*scatter plot*” já que existem lógicas para as quais a altura é sempre menor que $I = 0.6$, outras tendo alguns pontos maiores que $I = 0.6$ para $p = 0.9$ mas nenhum para $p = 0.99$, e ainda outras lógicas tendo alguns pontos mais altos que $I = 0.6$ para $p \approx 1$. O ponto importante é que o número de pontos maiores que $I = 0.6$ é drasticamente reduzido para todas as lógicas semi-IMP em relação às lógicas IMP. A única similaridade dessas duas classes de lógicas é o fato que um menor poder preditivo geralmente aumenta a correlação entre entropia e IMP score.

Tabela 4.6: Número de conjuntos IMP e entropia média das distribuições dos preditores para cada subespaço do espaço de parâmetros (P_1, P_2, P_3) definido por λ e δ para $p = 0.90$, $p = 0.95$ e $p = 0.99$ e lógica 00011001 (semi-IMP). Os eixos X e Y correspondem aos valores de δ e λ respectivamente. As legendas estão dispostas na última coluna (o número de conjuntos IMPs é dado por potências de 2 e as entropias estão no intervalo de 0 a 1).

Figura 4.10: “Scatter plots” do *IMP score* (I) contra entropia (H) com 970.299 pontos do espaço de parâmetros (P_1, P_2, P_3) para $p = 0.90$, $p = 0.95$ e $p = 0.99$ e lógica 00011001 (lógica semi-IMP). As correlações respectivas estão embaixo de cada um dos gráficos.

Preditores dependentes

Aqui é eliminada a hipótese de independência entre preditores. Ao fazer isto, a análise para três preditores acaba requisitando um número excessivo de parâmetros, tornando-se bastante complexa. Portanto, a análise será feita apenas considerando-se dois preditores. Para isto, é utilizada a covariância entre os preditores como medida de dependência. Note que, no contexto binário, os preditores são independentes se e somente se a covariância é nula. A covariância γ entre os preditores X_1 e X_2 é dada por $\gamma = E[X_1, X_2] - E[X_1]E[X_2]$, na qual $E[\cdot]$ é a função esperança. As Tabelas 4.7 e 4.8 mostram as tabelas de distribuição de probabilidades conjuntas adaptadas com a covariância γ .

Tabela 4.7: Tabela de distribuição de probabilidades conjuntas para a lógica AND adaptada com a covariância γ .

X_1	X_2	$P(Y = 0, X_1, X_2)$	$P(Y = 1, X_1, X_2)$
0	0	$((1 - P_1)(1 - P_2) + \gamma)p$	$((1 - P_1)(1 - P_2) + \gamma)(1 - p)$
0	1	$((1 - P_1)P_2 - \gamma)p$	$((1 - P_1)P_2 - \gamma)(1 - p)$
1	0	$(P_1(1 - P_2) - \gamma)p$	$(P_1(1 - P_2) - \gamma)(1 - p)$
1	1	$(P_1P_2 + \gamma)(1 - p)$	$(P_1P_2 + \gamma)p$

Tabela 4.8: Tabela de distribuição de probabilidades conjuntas para a lógica XOR adaptada com a covariância γ .

X_1	X_2	$P(Y = 0, X_1 = i, X_2 = j)$	$P(Y = 1, X_1 = i, X_2 = j)$
0	0	$((1 - P_1)(1 - P_2) + \gamma)p$	$((1 - P_1)(1 - P_2) + \gamma)(1 - p)$
0	1	$((1 - P_1)P_2 - \gamma)(1 - p)$	$((1 - P_1)P_2 - \gamma)p$
1	0	$(P_1(1 - P_2) - \gamma)(1 - p)$	$(P_1(1 - P_2) - \gamma)p$
1	1	$(P_1P_2 + \gamma)p$	$(P_1P_2 + \gamma)(1 - p)$

Primeiramente considere $P_1 = P_2 = \frac{1}{2}$. Baseando-se na Equação 4.5 e as tabelas de distribuição de probabilidades conjuntas 4.7 e 4.8, temos que $\alpha_{00} = (\frac{1}{4} + \gamma)$, $\alpha_{01} = (\frac{1}{4} - \gamma)$, $\alpha_{10} = (\frac{1}{4} - \gamma)$, e $\alpha_{11} = (\frac{1}{4} + \gamma)$ para todas as lógicas.

Os CoDs para a lógica AND, assumindo $P_1 = P_2 = \frac{1}{2}$ e $p \geq 0.5$, são dadas por

$$\begin{aligned} \text{CoD}_1 = \text{CoD}_2 &= 1 - \frac{1 - p + F\left(\frac{8p\gamma - 4\gamma + 1}{2}\right)}{2F\left(\frac{3 - 2p + 8p\gamma - 4\gamma}{4}\right)} \\ \text{CoD}_{12} &= 1 - \frac{1 - p}{F\left(\frac{3 - 2p + 8p\gamma - 4\gamma}{4}\right)} \end{aligned} \quad (4.12)$$

Após alguma manipulação algébrica, as equações podem ser simplificadas através do fato que $-\frac{1}{4} \leq \gamma \leq \frac{1}{4}$ (pois as probabilidades são não-negativas):

$$\begin{aligned} \text{CoD}_1 = \text{CoD}_2 &= \begin{cases} 0 & \text{se } -0.25 \leq \gamma \leq 0 \\ 1 - \frac{3 - 2p - 8p\gamma + 4\gamma}{3 - 2p + 8p\gamma - 4\gamma} & \text{se } 0 \leq \gamma \leq 0.25 \end{cases} \\ \text{CoD}_{12} &= 1 - \frac{4(1 - p)}{3 - 2p + 8p\gamma - 4\gamma} \end{aligned} \quad (4.13)$$

Os CoDs para a lógica XOR, assumindo $P_1 = P_2 = \frac{1}{2}$ e $p \geq 0.5$, são dados por

$$\begin{aligned} \text{CoD}_1 = \text{CoD}_2 &= 1 - \frac{F\left(\frac{-8p\gamma + 4\gamma + 1}{2}\right)}{F\left(\frac{-8p\gamma + 4\gamma + 1}{2}\right)} = 0 \\ \text{CoD}_{12} &= 1 - \frac{1 - p}{F\left(\frac{-8p\gamma + 4\gamma + 1}{2}\right)} \end{aligned} \quad (4.14)$$

A equação para CoD_{12} pode ser simplificada através da equação $-\frac{1}{4} \leq \gamma \leq \frac{1}{4}$ e assumindo $p > \frac{1}{2}$:

$$\text{CoD}_{12} = \begin{cases} 1 - \frac{2(1-p)}{8p\gamma - 4\gamma + 1} & \text{se } -0.25 \leq \gamma \leq 0 \\ 1 - \frac{2(1-p)}{-8p\gamma + 4\gamma + 1} & \text{se } 0 \leq \gamma \leq 0.25 \end{cases} \quad (4.15)$$

Note que, para $p = \frac{1}{2}$, o *IMP score* é zero para ambas lógicas, independente do valor de γ (isto corresponde ao caso em que o poder preditivo é nulo).

Outra observação importante é que, para qualquer valor do poder preditivo p , o *IMP score* máximo é obtido para $\gamma = 0$ para ambas lógicas AND e XOR (o caso de preditores independentes). Para a lógica XOR, como os *CoDs* individuais são zero, o máximo *IMP score* é obtido através da maximização do CoD_{12} , que corresponde à maximização do denominador na expressão para CoD_{12} na Equação 4.14. Mas, como visto na Figura 4.3, o máximo valor de F é alcançado no ponto $\frac{1}{2}$. Portanto, a equação resultante leva a $\gamma = 0$.

Para a lógica AND, a prova pode ser dividida em duas partes. Primeiramente, seja o caso $-0.25 \leq \gamma \leq 0$. Então, de acordo com a Equação 4.13, $\text{CoD}_1 = \text{CoD}_2 = 0$ e a maximização dos *IMP scores* corresponde a maximizar CoD_{12} , o qual é obtido através da maximização do termo $(3 - 2p + 8p\gamma - 4\gamma)$ na expressão para CoD_{12} . Já que $p \geq 0.5$, então o termo $8p\gamma - 4\gamma$ é igual a $4(2p - 1)\gamma$, o qual é crescente com os valores de γ . Portanto, quanto maior o γ no intervalo, maior será o valor do CoD_{12} , resultando em $\gamma = 0$ como ponto de máximo.

Agora, seja o caso $0 \leq \gamma \leq 0.25$. Da Equação 4.13, e após uma simples manipulação algébrica, o *IMP score* é dado por:

$$I = \frac{2p - 1 - (8p\gamma - 4\gamma)}{3 - 2p + 8p\gamma - 4\gamma}$$

Mais uma vez, já que $p \geq 0.5$, então o termo $8p\gamma - 4\gamma$ é igual a $4(2p - 1)\gamma$, o qual é crescente com os valores de γ . Portanto, quanto menor γ no intervalo, maior será o valor de CoD_{12} , resultando em $\gamma = 0$ como ponto de máximo. Unindo as duas partes da prova, temos que o maior *IMP score* é obtido para $\gamma = 0$.

A Figura 4.11 mostra gráficos de CoD_{12} , $\max\{\text{CoD}_1, \text{CoD}_2\}$ e $I = \text{CoD}_{12} - \max\{\text{CoD}_1, \text{CoD}_2\}$ como funções de p para alguns valores não-negativos de γ . A Figura 4.12 apresenta os mesmos gráficos para valores não-positivos de γ . Em todos os casos, o *CoD* conjunto, os *CoDs* individuais (exceto para os casos nos quais são nulos para todo intervalo), e o *IMP score* são crescentes com o poder preditivo. Além disso, em todos os casos o *IMP score* é crescente para valores decrescentes da covariância γ . Os gráficos são exatamente

os mesmos para as lógicas simétricas à lógica AND, tais como as lógicas 1000 (NAND), 0111 (OR) e 1110 (NOR). Para as lógicas 1,3-mintermo restantes (0100, 0010, 1011, 1101), os gráficos são idênticos àqueles para sinal inverso de γ considerando a lógica AND. Para a lógica NXOR, os gráficos são os mesmos da lógica XOR.

Figura 4.11: Gráficos dos CoDs e IMP score como funções de p para $P_1 = P_2 = 0.5$ e valores não-negativos de covariância ($\gamma \geq 0$).

Caso geral: preditores não-equiprováveis

Assumindo que P_1 e P_2 podem ter qualquer valor entre 0 e 1, baseando-se na Equação 4.5 e nas tabelas de distribuição conjunta de probabilidades (Tabelas 4.7 e 4.8), temos que $\alpha_{00} = (1 - P_1)(1 - P_2) + \gamma$, $\alpha_{01} = (1 - P_1)P_2 - \gamma$, $\alpha_{10} = P_1(1 - P_2) - \gamma$, e $\alpha_{11} = P_1P_2 + \gamma$ para todas as lógicas. Para a lógica AND (assumindo $p \geq 0.5$),

Figura 4.12: Gráficos dos CoDs e *IMP score* como funções de p para $P_1 = P_2 = 0.5$ e valores não-positivos de covariância ($\gamma \leq 0$).

$$\begin{aligned}
 \text{CoD}_1 &= 1 - \frac{(1 - P_1)(1 - p) + F \left[P_2 + \frac{\gamma - (2P_1P_2 + 2\gamma - P_1)p}{P_1} \right] P_1}{F [P_1P_2 + \gamma + (1 - 2P_1P_2 - 2\gamma)p]} \\
 \text{CoD}_2 &= 1 - \frac{(1 - P_2)(1 - p) + F \left[P_1 + \frac{\gamma - (2P_1P_2 + 2\gamma - P_2)p}{P_2} \right] P_2}{F [P_1P_2 + \gamma + (1 - 2P_1P_2 - 2\gamma)p]} \\
 \text{CoD}_{12} &= 1 - \frac{1 - p}{F [P_1P_2 + \gamma + (1 - 2P_1P_2 - 2\gamma)p]}
 \end{aligned} \tag{4.16}$$

Para a lógica XOR (assumindo $p \geq 0.5$),

$$\begin{aligned}
\text{CoD}_1 &= 1 - \frac{F \left[P_2 + \frac{(1-2P_2-P_1+2P_1P_2+2\gamma)p-\gamma}{1-P_1} \right] (1-P_1) + F \left[P_2 + \frac{(P_1-2P_1P_2-2\gamma)p+\gamma}{P_1} \right] P_1}{F [P_1 + P_2 - 2P_1P_2 - 2\gamma + (1-2P_1-2P_2+4P_1P_2+4\gamma)p]} \\
\text{CoD}_2 &= 1 - \frac{F \left[P_1 + \frac{(1-2P_1-P_2+2P_1P_2+2\gamma)p-\gamma}{1-P_2} \right] (1-P_2) + F \left[P_1 + \frac{(P_2-2P_1P_2-2\gamma)p+\gamma}{P_2} \right] P_2}{F [P_1 + P_2 - 2P_1P_2 - 2\gamma + (1-2P_1-2P_2+4P_1P_2+4\gamma)p]} \\
\text{CoD}_{12} &= 1 - \frac{1-p}{F [P_1 + P_2 - 2P_1P_2 - 2\gamma + (1-2P_1-2P_2+4P_1P_2+4\gamma)p]}
\end{aligned} \tag{4.17}$$

Note que, para $p = \frac{1}{2}$, o *IMP score* é zero para ambas as lógicas AND e XOR, independentemente dos valores de P_1 , P_2 e γ (isto corresponde ao caso de poder preditivo nulo).

Através dessas equações, é possível gerar gráficos de junção dos valores de CoD_{12} com $1 - \max(\text{CoD}_1, \text{CoD}_2)$ em canais verde e vermelho, assim como feito nas Figuras 4.6 e 4.7, para alguns valores de p e de γ . A Figura 4.13 ilustra esses gráficos para as lógicas AND e XOR. Para a lógica AND, é importante observar que covariância positiva suprime a região predisposta a gerar IMPs (quadrado inferior esquerdo) enquanto covariância negativa expande a região em volta do quadrado. Para a lógica XOR, tanto covariância positiva quanto a negativa expande a região central com maior *IMP score*. Além disso, a covariância positiva e a negativa expandem uma das diagonais. Portanto, uma pequena covariância (positiva ou negativa) aumenta a região de interesse para produção de IMPs (embora o *IMP score* em si diminua com a covariância), enquanto para a lógica AND, apenas uma pequena covariância negativa é benéfica.

4.6 Resultados em expressões gênicas de melanoma

Para testar o quanto as predições do modelo descrevem o comportamento de genes controladores, foi feito um exame de um conjunto de dados com amostras contendo versões ativas e inativas de um gene que atua como um controlador geral. O conjunto de dados investigado contém perfis de expressão de melanoma [Bittner et al., 2000]. Um gene conhecido por exercer amplo controle é o DUSP1 (*dual specificity phosphatase*) [Alessi et al., 1993]. DUSP1 pode servir como um antagonista efetivo a uma variedade de processos estimula-

Figura 4.13: Gráficos representando a junção dos valores de CoD_{12} e $1 - \max(\text{CoD}_1, \text{CoD}_2)$ nos canais verde e vermelho para as lógicas AND e XOR e para três valores distintos de p e cinco valores distintos de γ . As probabilidades P_1 e P_2 são dadas pelos eixos X e Y respectivamente. Regiões amarelas correspondem a regiões de interesse para surgimento de conjuntos IMP.

dos pelo MAPK ativado, incluindo proliferação e apoptose (morte celular) [Smalley, 2003], tendo sido reconhecido como antagonista funcional à replicação crônica dirigida por fatores de crescimento [Noguchi et al., 1993]. Em sua função como um guarda contra os altos níveis de proliferação freqüentemente associados ao câncer, DUSP1 atua como desfosforilador do MAPK1, uma proteína que serve como um ponto de integração para um conjunto diverso de processos celulares em adição à proliferação. Uma fração significativa das conseqüências da desfosforilação do MAPK1 deriva da inabilidade do MAPK1 desfosforilado em ativar fatores de transcrição por fosforilação.

O conjunto de dados em questão é composto de 31 amostras com 587 perfis de expressão gênica. Das 31 amostras, 19 são de tecidos normais e 12 são de tecidos com melanoma. As expressões são binarizadas para indicar mudança ou não-mudança relativa a um nível de expressão referência para cada gene individualmente. Uma mudança pode significar sub-expressão ou super-expressão. Tanto sub-expressão como super-expressão são rotuladas como 1, enquanto expressões sem mudança significativa com relação à referência são rotuladas como 0.

Tratando cada gene como um alvo, e dados λ and δ , o objetivo é obter uma lista de pares de preditores que são IMP com respeito a λ e δ para o gene alvo em questão. Alguns fatores que afetam a predição intrinsecamente multivariada foram analisados, tais como a probabilidade de cada preditor assumir o valor 1 (P_1 e P_2), a covariância γ entre os preditores, e o poder preditivo p . A simplificação assumida pelo modelo IMP de que todas as instâncias possuem o mesmo poder preditivo (ou seja, $\pi_{ij} = p$ ou $\pi_{ij} = 1 - p$) de forma a reduzir o número de parâmetros não se aplica exatamente aos dados reais. Porém, para uma lógica de predição específica, essa simplificação se aplica de forma aproximada, e os principais pontos levantados na análise teórica continuam valendo.

Considere os seguintes pares (λ, δ) : $(0.2, 0.7)$, $(0.3, 0.7)$, $(0.2, 0.8)$, $(0.3, 0.8)$. Esses pares são restritivos, mas admitem IMPs. As listas completas dos resultados podem ser encontradas em <http://www.vision.ime.usp.br/~davidjr/melanoma>. Aqui, a atenção será dada ao par $(0.2, 0.8)$, o par mais restritivo de todos os considerados.

Dos 587 genes, foram omitidos como alvos 37 genes tendo menos do que 6 zeros ou 6 uns. Tais genes apresentam muito pouca variabilidade, sendo que alguns deles levam a uma grande quantidade de conjuntos IMPs, mas devido ao pequeno número de amostras e ao pequeno número de variações, não é possível ter uma confiança razoável sobre esses

resultados. Dos 550 genes restantes, existem 101 alvos tendo no mínimo um par IMP. O gene alvo que possui o maior número de pares IMP é o DUSP1. Ele tem 19 pares IMP para $\lambda = 0.2$ e $\delta = 0.8$. Um desses 19 pares tem o maior *IMP score* possível (CoDs individuais iguais a 0 e CoD conjunto igual a 1). Os preditores ótimos para todos os 19 pares são determinados por lógicas de 1-mintermo: 14 lógicas AND (0001), 4 lógicas $Y = \bar{X}_1 \wedge X_2$ (0010 ou 0100), e 1 lógica NAND (1000). A Tabela 4.9 mostra a lista de pares IMP para DUSP1. Informações sobre lógica, P_1 , P_2 , e a covariância γ são apresentadas. Os gráficos contendo a junção dos valores de CoD_{12} e $1 - \max(\text{CoD}_1, \text{CoD}_2)$ nos canais verde e vermelho com as localizações (P_1, P_2) indicadas por um pequeno ponto azul são mostrados na Figura 4.14.

Figura 4.14: Gráficos com canais verde e vermelho representando os valores de CoD_{12} e $1 - \max(\text{CoD}_1, \text{CoD}_2)$ respectivamente para os conjuntos IMP de DUSP1 dados na Tabela 4.9. Cada gráfico foi gerado tomando como entrada a lógica, o poder preditivo p e a covariância γ obtidas. Os pontos azuis indicam as localizações (P_1, P_2) obtidas.

Considerando-se valores menos restritivos de λ e δ , DUSP1 também possui um número bastante significativo de conjuntos IMP. Para os seguintes pares $(\lambda, \delta) = \{(0.2, 0.7), (0.2, 0.8), (0.3, 0.7), (0.3, 0.8)\}$, DUSP1 é o gene alvo com o maior número de conjuntos IMP para três desses pares, tendo o terceiro maior número para $(\lambda, \delta) = (0.3, 0.8)$. A Ta-

Tabela 4.9: Lista de conjuntos IMP que têm DUSP1 como alvo. Os poderes preditivos e as covariâncias estão indicados nas colunas p e γ respectivamente. Os respectivos gráficos da junção dos valores de CoD_{12} e $1 - \max(\text{CoD}_1, \text{CoD}_2)$ nos canais verde e vermelho estão ilustrados na Figura 4.14. Esses gráficos foram gerados com base na lógica de predição, γ , p e o par (P_1, P_2) (o último indicado por um ponto azul nos gráficos).

#	X_1	X_2	Lógica	CoD_1	CoD_2	CoD_{12}	p	P_1	P_2	γ
1	RTN1	TEAD1	0001	0.0000	0.0000	1.0000	1.0000	0.5806	0.5161	-0.0763
2	CHN1	TOP1	0001	0.0000	0.1429	0.8571	0.9677	0.5484	0.4194	-0.0376
3	CASP3	STOM	0001	0.1429	0.0000	0.8571	0.9677	0.3548	0.4194	0.0462
4	EDG1	TEAD1	0001	0.0000	0.0000	0.8571	0.9677	0.3871	0.5161	-0.0065
5	MMP3	TEAD1	0001	0.0000	0.0000	0.8571	0.9677	0.4839	0.5161	0.0086
6	TGFBI	FOS	0001	0.0000	0.0000	0.8571	0.9677	0.4839	0.5161	-0.0581
7	UAP1	TOP1	0100	0.0000	0.1429	0.8571	0.9677	0.2581	0.4194	0.0548
8	TCF4	TEAD1	0001	0.0000	0.0000	0.8571	0.9677	0.5161	0.5161	-0.0753
9	TOP1	SERPINE1	0010	0.1429	0.0000	0.8571	0.9677	0.4194	0.4839	-0.0430
10	TOP1	TEAD1	0001	0.1429	0.0000	0.8571	0.9677	0.4194	0.5161	0.0430
11	TOP1	PLOD2	0010	0.1429	0.0000	0.8571	0.9677	0.4194	0.3548	0.0129
12	LAMA4	PCAF	0001	0.0000	0.0000	0.8571	0.9677	0.6452	0.4516	-0.0344
13	SERPINE1	PSFL	1000	0.0000	0.0000	0.8571	0.9677	0.4839	0.4516	-0.0258
14	IFIT1	TEAD1	0001	0.0000	0.0000	0.8571	0.9677	0.5484	0.5161	-0.0258
15	NR4A3	FOS	0001	0.0000	0.0000	0.8571	0.9677	0.6129	0.5161	-0.0602
16	CYP27A1	TEAD1	0001	0.0000	0.0000	0.8571	0.9677	0.4516	0.5161	0.0258
17	CYP27A1	ESTs	0010	0.0000	0.0000	0.8571	0.9677	0.4516	0.4839	0.0409
18	PCAF	FOS	0001	0.0000	0.0000	0.8571	0.9677	0.4516	0.5161	0.0258
19	ESTs	FOS	0001	0.0000	0.0000	0.8571	0.9677	0.6129	0.5161	-0.0602

bela 4.10 apresenta os três genes alvo com o maior número de IMPs para cada par (λ, δ) .

Esses resultados também corroboram a teoria de que se o poder preditivo p não é tão forte, apenas as lógicas XOR e NXOR lógicas podem produzir IMPs. Por outro lado, caso p seja forte, as lógicas de 1 e 3-mintermo possuem maior chance de produzir IMPs. Nos dados de melanoma, assumindo $\lambda = 0.2$ e $\delta = 0.8$, os conjuntos IMP assumiram um

Tabela 4.10: Genes alvo com o maior número de conjuntos IMP para os pares $(\lambda, \delta) = \{(0.2, 0.7), (0.2, 0.8), (0.3, 0.7), (0.3, 0.8)\}$.

	#	gene	#	gene
$\lambda = 0.2$	176	DUSP1	19	DUSP1
	103	SOCS1	16	ACY1
	87	MLANA	16	MLANA
$\lambda = 0.3$	350	DUSP1	49	MLANA
	225	MLANA	35	SOCS1
	215	SOCS1	33	DUSP1
	$\delta = 0.7$		$\delta = 0.8$	

dos quatro seguintes poderes preditivos: 1, 0.968, 0.936 e 0.903. A Tabela 4.11 mostra a distribuição de cada um desses valores para os conjuntos com lógicas de 1,3-mintermo e lógicas IMP de 2-mintermo (XOR e NXOR). É importante notar a correlação entre as lógicas e p e o fato de que nenhum dos conjuntos IMP com poder preditivo $p = 0.90$ é governado pelas lógicas de 1,3-mintermo e nenhum com poder preditivo $p = 1$ é governado por lógica IMP de 2-mintermo.

Tabela 4.11: Distribuição dos genes IMP classificados como lógicas de 1,3-mintermo e 2-mintermo (XOR e NXOR) para cada valor de poder preditivo apresentado nos resultados com $\lambda = 0.2$ e $\delta = 0.8$ ($p = 0.903, 0.936, 0.968, 1$).

	$p = 0.903$	$p = 0.936$	$p = 0.968$	$p = 1$
1,3-mintermo	0	92	102	3
2-mintermo	7	39	3	0

4.7 Discussão

Neste capítulo foi estudado o problema de predição binária usando o coeficiente de determinação (CoD). Fórmulas analíticas foram derivadas para o CoD ótimo como uma função da lógica de predição, poder preditivo, covariância entre preditores, e probabili-

dades marginais dos preditores. Esses fatores foram usados para determinar as condições sob as quais o fenômeno de predição intrinsecamente multivariada (IMP) ocorre, isto é, o CoD de cada subconjunto próprio dos preditores é baixo, mas o CoD conjunto de todos os preditores é alto. Em particular, isto implica que nenhum preditor individual determina o comportamento do gene alvo, enquanto a combinação desses preditores realiza uma excelente predição do alvo. Foi mostrado que IMP tende a ocorrer para alto poder preditivo e pequena covariância entre os preditores. Também foi demonstrado que IMP é favorecido por lógicas de predição específicas; por exemplo, com dois preditores, lógicas de 2-mintermo tais como XOR e NXOR são capazes de produzir IMPs com limiares mais restritivos do que os requeridos pelas lógicas de 1 e 3-mintermo tais como AND, NAND, NOR e OR.

Em nossa modelagem de regulação gênica Booleana, o CoD mede o quanto a expressão transcricional quantizada de um gene mestre é predito pelas expressões de seus genes escravos. Neste contexto, IMP corresponde ao caso em que o comportamento do mestre não pode ser confiavelmente predito por qualquer de seus genes escravos, mas combinações de genes escravos podem prever seu comportamento de forma quase determinística. As condições para o fenômeno IMP derivadas neste capítulo significam que IMP ocorre quando um gene mestre exerce uma forte regulação genômica (isto é, existe um alto poder de predição) de várias vias normalmente desconcorrelacionadas (isto é, há uma pequena covariância entre os preditores). Portanto, o modelo IMP captura a idéia de canalização de um sistema biológico com respeito a genes que podem exercer ações amplamente corretivas em processos celulares. Através do uso de dados de expressão gênica do estudo de melanoma, foi demonstrado que DUSP1, um gene conhecido por ser canalizador, exibe características IMP. Usando esses dados, foi mostrado empiricamente a constatação de que lógicas IMP de 2-mintermo produzem IMP para poder preditivo mais baixo do que lógicas de 1 e 3-mintermo.

Baseando-se nesses resultados, o critério de IMP pode ser promovido como uma ferramenta prática para a identificação de genes críticos de canalização em dados quantizados de expressão gênica. Mais estudos, ambos analíticos e experimentais, são necessários para obter um melhor entendimento do fenômeno IMP, e confirmar sua utilidade como ferramenta de descobrimento de genes canalizadores. No próximo capítulo será mostrado um estudo de quais características podem induzir o surgimento de genes IMP em um contexto

mais amplo: redes de regulação gênica simuladas com três ou mais genes.

Capítulo 5

Predição intrinsicamente multivariada: redes simuladas

5.1 Considerações preliminares

No Capítulo 4 foi introduzido o conceito de variáveis de predição intrinsicamente multivariada (IMP) e descritas suas propriedades analiticamente. Esse fenômeno foi minuciosamente analisado para o contexto de 2 e 3 variáveis preditoras com relação à variável alvo. Existem basicamente quatro propriedades que, em conjunto, podem originar conjuntos IMP: lógicas de predição, poder preditivo, covariância entre os preditores e distribuição de probabilidades de cada preditor tomado de forma isolada (probabilidades marginais). Tais propriedades foram empregadas para determinar as condições sob as quais conjuntos IMP surgem, isto é, todos os subconjuntos próprios de preditores são maus preditores enquanto o conjunto inteiro de preditores é um ótimo preditor do alvo considerado. O coeficiente de determinação não-linear (CoD), um critério que quantifica o incremento do poder preditivo por medição do erro de um conjunto preditor sobre apenas usar o erro *a priori*, foi o critério escolhido para realizar as análises.

Conjuntos IMP tendem a ocorrer para conjuntos de preditores com forte poder preditivo, baixa covariância e lógicas de predição específicas. Na modelagem proposta de regulação gênica Booleana, o CoD mede o quanto a expressão transcricional quantizada de um gene mestre é predita pelas expressões de seus genes escravos controlados. Nesse contexto, IMP corresponde ao caso em que o comportamento do mestre não pode ser con-

fiavelmente predito por qualquer de seus genes escravos isoladamente, mas combinações de genes escravos podem predizê-lo de maneira bastante confiável. O fenômeno IMP ocorre quando um gene mestre exerce forte regulação genômica de várias vias normalmente des-correlacionadas (baixa correlação entre preditores). O modelo IMP captura, portanto, a idéia de canalização de um sistema biológico [Waddington, 1942], ou seja, genes que podem exercer amplas ações corretivas em processos celulares.

No estudo do melanoma discutido na Seção 4.6, foi demonstrado que DUSP1, um gene conhecido por ser canalizador, possui diversos subconjuntos preditores com características IMP [Martins-Jr et al., 2008b]. Além disso, foi observado que lógicas XOR e XOR negado (NXOR) produzem IMP para poder preditivo menor que o necessário pelas lógicas de 1 e 3-mintermo (AND, OR, NOR, NAND, $X_1 \wedge \bar{X}_2$ e $X_1 \vee \bar{X}_2$).

Neste capítulo, é proposta uma extensão desse estudo em busca de características que são cruciais para produzir subconjuntos com alto *IMP score* em redes [Martins-Jr et al., 2008a]. Baseando-se nos resultados do Capítulo 4, uma hipótese plausível é que grandes proporções de lógicas de predição XOR ou NXOR nas redes podem levar a um aumento na chance de pelo menos um dos genes possuir alto *IMP score* com seus preditores. Mostra-se que existe uma alta correlação entre número de lógicas XOR/NXOR e o maior *IMP score* obtido numa rede considerada.

Uma outra característica importante a ser analisada é o tamanho do território de uma variável alvo. A idéia de analisar esta propriedade veio da observação de que as redes com apenas uma única variável fonte (variável de entrada) e que tiveram os *IMP scores* mais altos apresentam um território considerável. Dessa forma, obtém-se um indício de validação da hipótese obtida do Capítulo 4 [Martins-Jr et al., 2008b] de que subconjuntos com alto *IMP score* são mais suscetíveis a serem responsáveis pela regulação de várias vias metabólicas ou subsistemas, como observado nos experimentos com dados de melanoma. Mostramos ainda que isso ocorre devido ao fato de que a covariância dos preditores tem uma correlação negativa com o tamanho do território. Pelos resultados analíticos mostrados no Capítulo 4, uma alta covariância entre preditores diminui consideravelmente a possibilidade de obter alto *IMP score* com o alvo.

Todos os experimentos contidos neste capítulo envolvem redes construídas por simulações aleatórias descritas por um modelo. São consideradas apenas redes de estado estacionário (sem tempo envolvido) com uma única variável fonte, não permitindo ciclos

(grafos acíclicos orientados). Este capítulo é organizado assim: a Seção 5.2 descreve o modelo adotado para gerar redes simuladas; os resultados são ilustrados e discutidos na Seção 5.3; o capítulo é finalizado com algumas conclusões na Seção 5.4.

5.2 Modelo de redes

5.2.1 Hipóteses e restrições assumidas

Inicialmente, é necessário decidir quais tipos de conexão e restrições devem ser consideradas para definir o modelo de geração de redes. O modelo admitirá apenas conexões com graus de entrada 1, 2 ou 3, isto é, cada variável terá 1, 2 ou 3 preditores (exceto pela variável fonte que não possui preditores). Existem três razões para não considerar predição com grau maior do que 3 nas redes. A primeira razão é que, mesmo para predições de grau 3, o número de lógicas possíveis é bastante alto, o que pode fazer com que surja efeitos que confundam ou escondam algumas propriedades interessantes de redes. Outra razão é que dados biológicos normalmente possuem um pequeno número de amostras, o que impede uma correta estimação das distribuições de probabilidades. Além disso, não é muito comum observar predições com grau maior que 3 em sistemas dinâmicos ordenados, sendo que sistemas biológicos estão na fronteira entre comportamento ordenado e caótico (comportamento crítico) [Kauffman, 1993, Nykter et al., 2008].

Com relação às predições de grau de entrada 1, existem apenas duas lógicas de interesse (que resultam em CoD positivo): ativação e inibição. Dado um poder preditivo $p \geq 0.5$, as tabelas de distribuição de probabilidades condicionais dessas duas lógicas são mostradas na Tabela 5.1.

Tabela 5.1: Tabelas de distribuição de probabilidades condicionais para lógicas de predição de grau de entrada 1 (ativação e inibição) com poder preditivo p .

X	$P(Y = 0 X)$	$P(Y = 1 X)$
0	p	$1 - p$
1	$1 - p$	p

(a) ativação

X	$P(Y = 0 X)$	$P(Y = 1 X)$
0	$1 - p$	p
1	p	$1 - p$

(b) inibição

Como visto na Seção 4.4, existem 8 lógicas de grau 2 de predição capazes de produzir IMP: AND, OR, NAND, NOR, XOR, NXOR, $Y = X_1 \wedge \bar{X}_2$ e $Y = X_1 \vee \bar{X}_2$. No modelo de redes adotado aqui, todas essas lógicas são admitidas.

Considerando conexões de grau 3, lógicas parcialmente IMP não são consideradas para evitar efeitos de confusão (existem 152 delas, como mencionado na Seção 4.4). O modelo inclui apenas as lógicas IMP. Das 66 lógicas IMP, 36 são redundantes com relação à ordem das variáveis, restando 30 lógicas IMP com conexão de grau 3. Todas essas 30 lógicas estão incluídas no modelo.

5.2.2 Modelo proposto de redes

O modelo adotado aqui é o de grafos acíclicos orientados, ou seja, as redes não apresentam ciclos de dependência. Usando esse modelo, todo gene tem uma tabela de distribuição de probabilidades condicionais relacionada com os seus preditores, exceto para os genes fontes (genes que não possuem preditores). Através dessas tabelas de probabilidades condicionais, é possível construir uma tabela de probabilidades conjuntas com 2^n linhas na qual n é o número de genes da rede. A partir dessa tabela e das estruturas lógicas da rede, pode-se obter os CoDs e, por conseguinte, os *IMP scores* de todas as conexões de graus 2 e 3. Como os genes fontes não possuem preditores, a probabilidade marginal desses genes $P(X_n = 1)$ é escolhida de forma aleatória.

Sejam $\mathbf{X} = X_1, X_2, \dots, X_n$ os genes que compõem a rede. Os preditores de X_i compõem o subconjunto $\mathbf{Z}_i$ tal que se $X_j \in \mathbf{Z}_i$, então $i < j$. Dessa forma, o gene X_n é a fonte da rede (ausência de preditores) e sua distribuição marginal de probabilidades ($P(X_n = 1)$) é escolhida aleatoriamente com distribuição uniforme no intervalo entre 0.1 e 0.9 (valores menores que 0.1 e maiores que 0.9 não são permitidos para evitar que o gene fonte seja um “housekeeping”, isto é, quase constante).

O gene X_{n-1} é predito apenas pela fonte X_n com lógica de ativação ou inibição, cada uma sorteada com probabilidade 0.5 (ver Seção 5.2.1 para uma explicação dessas lógicas com grau de entrada 1). O gene X_{n-2} pode ter grau de predição 1 ou 2 de conexão (probabilidade 0.5 para ambos). Se predição de 1 grau for escolhida, um dos dois genes possíveis (X_{n-1} ou X_n) é escolhido aleatoriamente para assumir a função preditora (probabilidade 0.5 para ambos) com lógica de ativação ou inibição escolhidas aleatoriamente

com probabilidade 0.5 cada. Senão, caso o grau de entrada 2 tenha sido escolhido, cada uma das 8 lógicas IMP (ver Seção 4.4 para uma discussão sobre essas lógicas) pode ser escolhida através da distribuição uniforme de probabilidades ($\frac{1}{8}$ cada).

Para todo X_i com $i < n - 2$, o grau de conexão k ($k = \{1, 2, 3\}$) é aleatoriamente escolhido com probabilidade $\frac{1}{3}$ para cada grau. Após a escolha de k , um subconjunto de tamanho k é escolhido aleatoriamente de todos os subconjuntos possíveis de tamanho k no conjunto $\{X_{i+1}, X_{i+2}, \dots, X_n\}$. Lembrando, da Seção 4.4, que a ordem dos elementos no subconjunto importa para as lógicas $Y = \bar{X}_1 \wedge X_2$, $Y = \bar{X}_1 \vee X_2$ e algumas lógicas de grau 3. Portanto, temos que a ordem dos elementos no subconjunto escolhido também deve ser obtida de forma aleatória. Se $k = 1$, então a lógica é escolhida entre ativação e inibição com probabilidade $\frac{1}{2}$ cada. Se $k = 2$, então a lógica é escolhida dentre 8 lógicas IMP com probabilidade $\frac{1}{8}$. Finalmente, se $k = 3$, a lógica é escolhida dentre 30 lógicas IMP possíveis (Seção 4.4) com probabilidade $\frac{1}{30}$ cada.

Um fluxograma que ilustra esse mecanismo de construção de redes pode ser encontrado na Figura 5.1. Um exemplo de redes geradas por esse mecanismo é ilustrado na Figura 5.2.

5.3 Descrição dos experimentos

Foram realizados dois experimentos de geração de redes a partir de simulações através do modelo descrito na Seção 5.2 considerando os seguintes valores de poder preditivo: $p = 0.95$ (muito alto), $p = 0.9$ (alto), $p = 0.8$ (baixo) e $p = 0.6$ (muito baixo). Os resultados obtidos corroboram as análises descritas no Capítulo 4 [Martins-Jr et al., 2008b] e revelam uma propriedade marcante que mostra que os conjuntos IMP possuem uma tendência a controlar um grande número de genes, especialmente quando p é alto.

De cada rede gerada é possível obter informações importantes, tais como distribuições marginais de probabilidades ($P(X_i = 1)$), *IMP score* e estrutura (conexões e lógicas). O primeiro experimento consiste em analisar a correlação entre o número de categorias de lógicas (número de XOR/NXOR ou número de lógicas de 1-mintermo e 3-mintermo) e o máximo *IMP score* obtido nas redes. O segundo experimento tem como objetivo verificar alguma relação entre o tamanho do território de um nó (uma propriedade de teoria dos grafos que será explicada na Seção 5.3.2) com seu *IMP score*, de acordo com sua lógica (XOR/NXOR ou 1,3-mintermo).

Figura 5.1: Fluxograma ilustrando o mecanismo de construção de redes.

5.3.1 Número de lógicas XOR/NXOR

Nesse experimento, foram geradas 10.000 redes com 14 nós para cada p considerado (0.95, 0.9, 0.8 e 0.6). Cada rede tem um nó com *IMP score* máximo. A idéia é observar se existe alguma correlação entre o número de lógicas XOR/NXOR na rede com os máximos *IMP scores*, dado p .

De acordo com as afirmações feitas no Capítulo 4, as lógicas XOR/NXOR podem receber *IMP scores* maiores do que as lógicas de 1,3-mintermo considerando o mesmo p para ambas [Martins-Jr et al., 2008b]. Desta afirmação, seria esperado que, quanto maior o número de lógicas XOR/NXOR na rede, maior a possibilidade de se obter um alto *IMP score*. Em outras palavras, é esperada uma forte correlação positiva entre o número de

Figura 5.2: Exemplo de rede construída pelo modelo descrito com $n = 8$.

lógicas XOR/NXOR presentes na rede e seu *IMP score* mais alto. A Figura 5.3 mostra quatro diagramas, um para cada p considerado (0.95, 0.9, 0.8, 0.6). Esses diagramas representam a evolução da mediana dos máximos *IMP scores* ao longo dos conjuntos de pontos com número crescente de lógicas XOR/NXOR. As barras verticais indicam o intervalo em que 50% dos pontos estão concentrados (isto é, o fim do quartil inferior e o início do quartil superior). Os pontos dispostos acima das barras verticais indicam os *IMP scores* mais altos obtidos para cada conjunto considerado definido pelo número de lógicas XOR/NXOR. Essas figuras mostram que, de fato, existe uma correlação alta entre o número de lógicas XOR/NXOR e o *IMP score* máximo, como hipotetizado. Além disso, essa correlação aumenta para decrescentes valores de p , concordando com o que foi dito na Seção 4.5.2 em que, no caso de poder preditivo não tão alto, *IMP scores* altos devem ocorrer, em sua maioria, para predições com lógicas XOR/NXOR [Martins-Jr et al., 2008b].

Como as lógicas XOR/NXOR são as principais responsáveis pelos *IMP scores* mais altos, o número de lógicas de 1,3-mintermo ou conexões com grau 3 possuem uma correlação ligeiramente negativa com o *IMP score* máximo. As Figuras 5.4 e 5.5 apresentam esses resultados.

5.3.2 Tamanho de território

O território de um nó X é usualmente definido como o conjunto de genes $\mathbf{T}$ no qual existe pelo menos um caminho de X a qualquer nó $T \in \mathbf{T}$. Aqui, o conceito de território

Figura 5.3: Diagramas apresentando os valores das medianas dos *IMP scores* máximos para conjuntos representados pelo número de lógicas XOR/NXOR para $p = 0.95$, $p = 0.9$, $p = 0.8$ e $p = 0.6$. As barras verticais indicam a região com 50% de concentração de pontos para cada conjunto, enquanto as extremidades das barras indicam onde o quartil inferior termina e onde o quartil superior começa. As correlações entre o número de lógicas XOR/NXOR e o *IMP score* máximo são: (a) 0.4269, (b) 0.4806, (c) 0.5871 e (d) 0.5856.

será definido considerando os caminhos inversos, ou seja, $\mathbf{T}$ é o território do nó X se e somente se para todo $T \in \mathbf{T}$, existe pelo menos um caminho de T a X . Em particular, $X \in \mathbf{T}$, então o tamanho mínimo do território de um nó é 1. A Figura 5.6 ilustra o mesmo exemplo de rede dado na Figura 5.2 contendo o tamanho do território de cada um dos nós.

A idéia de analisar se o tamanho do território de um gene em uma rede é associado

Figura 5.4: Diagramas apresentando os valores das medianas dos *IMP scores* máximos para conjuntos representados pelo número de lógicas de 1,3-mintermo para $p = 0.95$, $p = 0.9$, $p = 0.8$ e $p = 0.6$. As barras verticais indicam a região com 50% de concentração de pontos para cada conjunto, enquanto as extremidades das barras indicam onde o quartil inferior termina e onde o quartil superior começa. As correlações entre o número de lógicas de 1,3-mintermo e *IMP score* máximo são: (a) -0.0086, (b) -0.0414, (c) -0.1295 e (d) -0.1647.

de alguma forma com seu *IMP score* surgiu da observação de que as redes melhor classificadas em termos do máximo *IMP score* obtidas no experimento anterior (Seção 5.3.1) apresentam o *IMP score* máximo localizado no alvo com um considerável tamanho de território (geralmente longe do gene fonte) para todo p considerado. As Figuras 5.7 e 5.8 apresentam as redes melhor classificadas para $p = \{0.95, 0.90, 0.80, 0.60\}$.

Figura 5.5: Diagramas apresentando os valores das medianas dos *IMP scores* máximos para conjuntos representados pelo número de lógicas de grau 3 para $p = 0.95$, $p = 0.9$, $p = 0.8$ e $p = 0.6$. As barras verticais indicam a região com 50% de concentração de pontos para cada conjunto, enquanto as extremidades das barras indicam onde o quartil inferior termina e onde o quartil superior começa. As correlações entre o número de lógicas de grau 3 e *IMP score* máximo são: (a) -0.1421, (b) -0.1563, (c) -0.1115 e (d) -0.0318.

Como as lógicas XOR/NXOR usualmente tendem a resultar em *IMP scores* mais altos, é necessário separar a análise em duas: a primeira é feita apenas para o conjunto de redes que tem lógicas XOR/NXOR como alvos com um dado tamanho de território; a segunda faz a mesma coisa considerando lógicas de 1 ou 3-mintermo. Para cada p considerado, foram geradas 120.000 redes, 10.000 para cada tamanho de território considerado (3, 4, ..., 14). Por construção, o nó 1 é aquele com o maior território (sempre que for dito que uma rede tem tamanho de território t , significa que o nó 1 dessa rede apresenta

Figura 5.6: Tamanho de território de cada nó na rede exemplo ilustrada na Figura 5.2.

tamanho de território t). 120.000 redes é um número suficiente de amostras tendo em vista que, mesmo para um quarto desse número (30.000 amostras), o comportamento geral do tamanho de território em função do *IMP score* é mantido para toda combinação de lógicas (XOR/NXOR ou 1,3-mintermo) e para todos os poderes preditivos considerados nos experimentos. Como um exemplo, a Tabela 5.2 apresenta as médias e os desvios padrões dos *IMP scores* para cada conjunto de amostras agrupadas por tamanho de território e para três números de amostras (30.000, 60.000 e 120.000) considerando lógicas XOR/NXOR para predição do nó 1 e poder preditivo $p = 0.95$. Observe que a variação entre os desvios padrões é mínima ao longo do número de amostras para todos os tamanhos de território considerados.

XOR/NXOR

Aqui foram observadas quantas amostras de cada tamanho de território considerado aparecem entre as redes possuindo os maiores 5% *IMP scores* (6.000 amostras de um total de 120.000), sendo tomada XOR/NXOR como lógica de predição para o nó 1. A Figura 5.9 apresenta histogramas mostrando as respectivas frequências absolutas de cada tamanho de território entre as 6.000 amostras para quatro valores de p . Para p alto (0.95 e 0.9), as barras são claramente crescentes com o tamanho do território, enquanto que para p

Figura 5.7: As melhores redes classificadas dentre as 10.000 redes simuladas para $p = 0.95$ e $p = 0.90$ no experimento apresentado na Seção 5.3.1. Os alvos com os mais altos *IMP scores* são indicados pelas setas vermelhas. Os tamanhos de território para esses dois conjuntos IMP são: (a) 13; (b) 11

baixo (0.80 e 0.60), os histogramas tornam-se mais uniformes. Uma possível explicação para esse fato pode ser encontrada logo mais na Seção 5.3.3.

Figura 5.8: As melhores redes classificadas dentre as 10.000 redes simuladas para $p = 0.80$ e $p = 0.60$ no experimento apresentado na Seção 5.3.1. Os alvos com os mais altos *IMP scores* são indicados pelas setas vermelhas. Os tamanhos de território para esses dois conjuntos IMP são: (a) 10; (b) 9.

1,3-mintermo

Agora, levando em conta os *IMP scores* obtidos para alvos com lógicas 1,3-mintermo de predição do nó 1 e seus tamanhos de territórios, para p alto (0.95, 0.80), os histogramas de frequências absolutas das amostras com um dado tamanho de território, pertencentes ao conjunto com os 5% maiores *IMP scores*, são crescentes ao longo de praticamente todo o domínio dos tamanhos de território considerados. Para p baixo (0.80, 0.60), diferentemente do caso XOR/NXOR, as barras ainda apresentam um comportamento crescente

Tabela 5.2: Médias (μ) e desvios padrões (σ) dos *IMP scores* para cada tamanho de território (t) e para três quantidades de amostras (30.000, 60.000, 120.000) considerando lógicas XOR/NXOR como preditoras do nó 1 e poder preditivo $p = 0.95$.

Amostras													
30.000	μ	0.4737	0.4952	0.4593	0.4480	0.4384	0.4318	0.4289	0.4378	0.4269	0.4359	0.4273	0.4324
	σ	0	0.1640	0.2105	0.2191	0.2216	0.2150	0.2191	0.2162	0.2103	0.2175	0.2129	0.2131
60.000	μ	0.4737	0.4976	0.4600	0.4400	0.4429	0.4451	0.4320	0.4255	0.4321	0.4344	0.4303	0.4314
	σ	0	0.1610	0.2130	0.2174	0.2211	0.2206	0.2160	0.2134	0.2130	0.2134	0.2141	0.2121
120.000	μ	0.4737	0.4960	0.4562	0.4438	0.4398	0.4368	0.4336	0.4341	0.4314	0.4302	0.4319	0.4353
	σ	0	0.1632	0.2135	0.21912	0.2206	0.2178	0.2152	0.2147	0.2152	0.2125	0.2124	0.2123
t		3	4	5	6	7	8	9	10	11	12	13	14

considerável até tamanho de território igual a 12 (ver Figura 5.10).

5.3.3 Discussão

A razão pela qual as redes melhor classificadas em termos dos *IMP scores* possuem um tamanho grande de território, especialmente para poder preditivo alto, pode ser explicada pela relação entre o tamanho do território e a covariância entre os preditores. O n -ésimo gene (fonte) possui alta covariância (em termos absolutos) com o $(n - 1)$ -ésimo gene já que este é unicamente predito pelo gene fonte. Assim, se o gene alvo considerado está localizado próximo à fonte (território pequeno), é esperado que seus preditores tendam a ser altamente covariados. Por outro lado, se o alvo posiciona-se distante do gene fonte (grande território), existe uma chance maior de que os preditores apresentem pequena covariância entre eles. Em outras palavras, a forte covariância entre o gene fonte e seu sucessor tende a se dissipar ao longo das vias da rede. Este resultado concorda com a análise apresentada na Seção 4.5.2 na qual pequenos valores de covariância podem beneficiar o *IMP score*, enquanto covariância alta o reduz de forma drástica [Martins-Jr et al., 2008b].

Para cada poder preditivo considerado ($p = \{0.95, 0.90, 0.80, 0.60\}$), foi visto anteriormente nesta seção que foram geradas 240.000 redes para cada valor de p (120.000 considerando o gene 1 predito por lógicas XOR/NXOR, e 120.000 predito por lógicas de

Figura 5.9: Histogramas de amostras com um dado tamanho de território e *IMP score* dentre os 5% maiores para lógicas XOR/NXOR de predição do nó 1 e $p = \{0.95, 0.9, 0.8, 0.6\}$.

1,3-mintermo), 20.000 redes para cada tamanho de território em questão (3,4. ..., 14). A covariância (em módulo) entre os preditores do gene 1 e sua correlação com o tamanho de seu território foram analisados. A Figura 5.11 apresenta quatro curvas, uma para cada p considerado, indicando a evolução da média das covariâncias ao longo dos tamanhos de território (cada média é calculada a partir de 20.000 amostras). As correlações entre as covariâncias e os tamanhos de território são -0.5105, -0.5919, -0.6510 e -0.5960 para $p = 0.95, 0.90, 0.80, 0.60$, respectivamente.

Além disso, as curvas exibidas na Figura 5.11 podem explicar o fato de que o comportamento crescente das barras dos histogramas tendem a desaparecer para poderes preditivos menores (esse desaparecimento se dá de forma mais rápida para lógicas XOR/NXOR). Na Seção 4.5.2 foi mostrado que uma pequena covariância em módulo ($|\gamma| \leq 0.08$) é de

Figura 5.10: Histogramas de amostras com um dado tamanho de território e *IMP score* dentre os 5% maiores para lógicas 1,3-mintermo de predição do nó 1 e $p = \{0.95, 0.9, 0.8, 0.6\}$.

fato benéfica para a chance do aparecimento de um par preditores-alvo que seja IMP se a predição for definida pela lógica XOR/NXOR [Martins-Jr et al., 2008b]. Para as lógicas de 1,3-mintermo, apenas covariância positiva ou negativa é benéfica. Se a covariância positiva é benéfica, covariância negativa é prejudicial, e vice-versa.

Analisando a Figura 5.11 com mais cuidado, é possível notar que as curvas para p alto (0.95 e 0.90) começam com alta média de covariâncias (maior que 0.15) e continua maior que 0.07 até um tamanho de território igual a 5, o que é prejudicial para todas as lógicas, induzindo uma tendência de crescimento notável das barras dos histogramas para ambos os casos. Isto porque é claramente mais fácil conseguir *IMP scores* altos quando a média das covariâncias não é tão alta (aproximadamente 0.05), o que acontece apenas

Figura 5.11: Evolução das médias das covariâncias (em módulo) entre os preditores do gene 1 ao longo dos tamanhos de território para $p = \{0.95, 0.90, 0.80, 0.60\}$. As respectivas correlações das covariâncias e dos tamanhos de território são -0.5105, -0.5919, -0.6510 e -0.5960.

para grandes territórios. Por outro lado, para p baixo (0.80 e 0.60), os trechos iniciais das curvas não são tão altos, então é perfeitamente possível obter *IMP scores* altos para tamanhos de territórios próximos a 6 ou 7 no caso das lógicas XOR/NXOR (a tendência crescente das barras dos histogramas desapareceram até mesmo para $p = 0.8$). Para as lógicas de 1,3-mintermo, a tendência crescente das barras dos histogramas ainda se conserva para pequeno valor de p , já que o valor absoluto da covariância pode ser tanto benéfico quanto prejudicial para obtenção de altos *IMP scores* neste caso.

A Figura 5.12 sumariza a relação do *IMP score* com as condições analisadas (lógicas, poder preditivo, tamanho de território e covariância entre preditores).

5.4 Conclusão

O fenômeno IMP foi analisado por meio de um estudo por simulações usando um grande número de redes simuladas. Duas propriedades de redes foram detectadas como favoráveis ao surgimento de *IMP scores*: número de lógicas XOR/NXOR e tamanho do território.

Figura 5.12: Relação do *IMP score* com as demais condições analisadas (lógicas, poder preditivo, tamanho de território e covariância entre preditores).

Os resultados confirmaram propriedades conhecidas e apresentadas no Capítulo 4, tal como o fato do número de lógicas XOR/NXOR serem correlacionados positivamente com o *IMP score* e, adicionalmente, uma nova propriedade em que o *IMP score* de uma rede com apenas um nó fonte é correlacionado positivamente com o tamanho do território. A explicação para isso reside basicamente no fato de que a covariância dos preditores tende a diminuir com o aumento do tamanho do território (covariâncias altas são prejudiciais para a obtenção de *IMP scores* altos). Isso é um indício de que genes IMP podem estar associados com o controle de diversas vias metabólicas, ou seja, um gene alvo com grande território tem mais chance de ser IMP com seus preditores e, portanto, ele pode ser um gene mestre que controla várias vias metabólicas independentes.

Capítulo 6

Conclusão

6.1 Considerações finais

Este trabalho apresentou resultados em duas frentes de pesquisa envolvendo seleção de características para tratar problemas de modelagem de redes de regulação gênica. Uma dessas frentes lidou com o problema de identificação de redes gênicas, visto como sendo uma das mais importantes etapas do longo e complexo processo de análise de dados genômicos. Nesse contexto, foi proposto um método de identificação de redes gênicas probabilísticas assumindo os dados temporais de expressão gênica obtidos de *microarray* como sendo um processo que segue uma cadeia de Markov de primeira ordem. Devido à falta de um número satisfatório de amostras, esse modelo considera que existe apenas uma função de transição de estados para todos os instantes de tempo considerados (invariância por translação), o que permite a aplicação de algoritmos de seleção de características para avaliar a dependência entre os genes. Para avaliar tal dependência, foi proposta uma função critério baseada em entropia condicional que penaliza as instâncias que são observadas uma única vez no conjunto de dados, embutindo a estimação do erro diretamente na função critério. Resultados experimentais obtidos através de simulações e análise de dados de *microarray* de *Plasmodium falciparum*, utilizando genes alvos de dois módulos vitais do parasita, demonstram a validade dessa técnica no objetivo de inferir novos conhecimentos biológicos. A análise dos dados biológicos contou com a colaboração da equipe do Prof. Hernando Del Portillo, na época pesquisador do Instituto de Ciências Biomédicas-USP especialista em parasitologia que, atualmente, faz parte do quadro de

pesquisadores do Centro de Pesquisa em Saúde Internacional de Barcelona (CRESIB). É importante ressaltar que tal colaboração foi iniciada em 2004, no final do mestrado, como uma iniciativa para disputar uma competição promovida pelo Critical Assessment of Microarray Data Analysis (CAMDA 2004) para a análise de dados públicos de *microarray* representando séries temporais de expressões gênicas do ciclo de vida do *Plasmodium falciparum*. Os resultados dessa iniciativa foram a publicação de um artigo que ganhou o prêmio de melhor apresentação do congresso e uma extensão desse artigo na forma de um capítulo pertencente ao livro *Microarray Data Analysis V*. Esse capítulo contém uma formulação detalhada da metodologia.

Outro principal foco desta tese foi sobre a questão da predição intrinsecamente multivariada, um fenômeno bastante importante que ocorre quando se analisa conjuntos de características para predição do comportamento de uma variável alvo (classe). Esse fenômeno é um dos principais responsáveis pelo fato de não existir, até o presente momento, um algoritmo de seleção de características que garanta otimalidade sem avaliar exaustivamente todos os conjuntos de características. Além disso, no contexto de redes de regulação gênica, genes IMP podem ser de vital importância como controladores de diversas vias metabólicas. Mostramos que, considerando genes como variáveis Booleanas, existem quatro fatores que determinam se um determinado conjunto de variáveis é IMP ou não: lógica de predição, poder preditivo, distribuição de probabilidades marginais (*bias*) e covariância entre preditores (entropia conjunta para o caso de 3 ou mais preditores). Outro fato importante mostrado é que qualquer conjunto de genes que tenha função canalizadora pode ser um conjunto IMP. Existem trabalhos na literatura evidenciando que genes canalizadores podem desempenhar papéis importantes no controle regulatório. Em particular, o gene DUSP1 em melanoma é conhecido por desempenhar funções de canalização. De fato, através de análises de dados de *microarray* de melanoma, ficou constatado que esse gene forma o maior número de conjuntos IMP como alvo da predição, sendo que todos eles são regidos por lógicas de 1-mintermo, que são canalizadoras. Posteriormente, uma outra evidência de que genes IMP são importantes no controle regulatório foi obtida através de simulações computacionais para geração de redes com 3 ou mais nós. Nesses experimentos, ficou constatado que o tamanho do território de um gene alvo pode influenciar em seu *IMP score* com seus preditores. Tal trabalho foi fruto de um estágio de doutorado com duração de 1 ano financiado pela CAPES (proc. BEX-3696-065) no Genomic Signal Processing Laboratory do Texas A&M University - EUA, sob orientação do Prof.

Edward Russell Dougherty, envolvendo a colaboração do Prof. Ulisses Braga-Neto. A análise dos resultados biológicos em melanoma foi executada pelo Prof. Michael Bittner do Translational Genomic Research Institute (TGEN) - EUA.

6.2 Perspectivas futuras

Com relação ao modelo de redes gênicas probabilísticas proposto neste trabalho, uma possibilidade imediata é considerar dependências lineares entre genes. Dessa forma, o número de linhas da tabela de probabilidades conjuntas (número de instâncias possíveis) dos preditores com relação ao alvo cresce linearmente ao invés de exponencialmente, o que pode reduzir sensivelmente o problema do erro de estimação e permitir que conjuntos com maior número de preditores sejam considerados. Porém, ao assumir essa característica no modelo, deve haver uma perda de informação significativa, já que a informação de dependência do alvo com relação aos preditores é comprimida na somatória correspondente à combinação linear entre os valores dos preditores. Mesmo assim, espera-se que redes com outros aspectos interessantes sejam obtidas. A análise dos resultados do método de PGN sobre as três cepas de *microarray* de *Plasmodium falciparum* continua em andamento, podendo ainda gerar novos conhecimentos biológicos futuramente.

A pesquisa sobre predição intrinsecamente multivariada em redes simuladas indica que a causa dos *IMP scores* mais altos serem atingidos para redes com grande território é o fato da covariância entre os preditores possuir uma tendência decrescente com o tamanho do território do alvo. Uma das próximas etapas dessa pesquisa é demonstrar essa conjectura. Além disso, com base em todos os resultados obtidos até então, abre-se novas linhas de pesquisa sobre genes IMP. Uma dessas linhas envolve o estudo de quais características ou conexões lógicas temporais em sistemas dinâmicos são responsáveis por produzir o fenômeno IMP no estado de equilíbrio, o que estenderia a pesquisa feita até o momento sobre redes estáticas a redes com evolução no tempo. Outra frente de pesquisa pode se concentrar em aplicar o conceito de predição intrinsecamente multivariada para identificação de redes regulatórias minimais, tendo em vista que um gene alvo sendo IMP com relação a seus preditores não pode ser predito por nenhum de seus subconjuntos próprios.

As funções critério discutidas nesta tese podem ser confrontadas. Em identificação de

redes gênicas probabilísticas, por exemplo, poderia ser utilizado o mesmo modelo com uma função critério diferente para seleção de características, tal como o CoD, e regerar as PGNs com o objetivo de compará-las. Já o conceito de predição intrinsecamente multivariada poderia ser redefinido utilizando entropia no lugar do erro do classificador Bayesiano na equação do CoD. Nesse caso, a predição intrinsecamente multivariada seria definida com base no *coeficiente de incerteza* que normaliza a informação mútua pela entropia *a priori* da variável alvo [Coombs et al., 1970]. O coeficiente de incerteza é definido por:

$$C_Y(\mathbf{X}) = \frac{H(Y) - H(Y|\mathbf{X})}{H(Y)} \quad (6.1)$$

Utilizando o coeficiente de incerteza, pode-se verificar certas mudanças nas condições dos fatores para a obtenção de características com alto *IMP score*. Porém, espera-se que, em linhas gerais, essas condições sejam similares. Ou seja, grande parte das conclusões estabelecidas no Capítulo 4 e 5 não devem sofrer alterações substanciais.

Referências Bibliográficas

- [Akutsu et al., 2000a] Akutsu, T., Miyano, S., and Kuhara, S. (2000a). Algorithms for inferring qualitative models of biological networks. In *Pacific Symposium on Biocomputing*, volume 5, pages 293–304.
- [Akutsu et al., 2000b] Akutsu, T., Miyano, S., and Kuhara, S. (2000b). Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, 16(8):727–734.
- [Albert and Barabási, 2002] Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(47).
- [Albert et al., 2001] Albert, R., Jeong, H., and Barabási, A.-L. (2001). Error and attack tolerance of complex networks. *Nature*, 409(6819):542.
- [Alessi et al., 1993] Alessi, D. R., Smythe, C., and Keyse, S. M. (1993). The human c1100 gene encodes a tyr/thr-protein phosphatase which potently and specifically inactivates map kinase and suppresses its activation by oncogenic ras in xenopus oocyte extracts. *Oncogene*, 8(7):2015–2020.
- [Anastassiou, 2007] Anastassiou, D. (2007). Computational analysis of the synergy among multiple interacting genes. *Molecular Systems Biology*, 3(83).
- [Barrera et al., 2002] Barrera, J., Cesar-Jr., R. M., Dantas, D. O., Martins-Jr., D. C., and Trepode, N. W. (2002). From microarray images to biological knowledge. In *Proceedings of the Second Brazilian Symposium on Mathematical and Computational Biology*.
- [Barrera et al., 2007] Barrera, J., Cesar-Jr, R. M., Humes-Jr, C., Martins-Jr, D. C., Patrão, D. D. C., Silva, P. J. S., and Brentani, H. (2007). A feature selection approach for identification of signature genes from sage data. *BMC Bioinformatics*, 8:169.
- [Barrera et al., 2004] Barrera, J., Cesar-Jr, R. M., Martins-Jr, D. C., Vencio, R. Z. N., Merino, E. F., Yamamoto, M. M., Leonardi, F. G., Pereira, C. A. B., and del Portillo, H. A. (2004). A

- new annotation tool for malaria based on inference of probabilistic genetic networks. In *Fifth International Conference for the Critical Assessment of Microarray Data Analysis (CAMDA)*, pages 36–40, Durham. (Winner of the Best Presentation Award).
- [Barrera et al., 2006a] Barrera, J., Cesar-Jr, R. M., Martins-Jr, D. C., Vencio, R. Z. N., Merino, E. F., Yamamoto, M. M., Leonardi, F. G., Pereira, C. A. B., and del Portillo, H. A. (2006a). Constructing probabilistic genetic networks of *Plasmodium falciparum* from dynamical expression signals of the intraerythrocytic development cycle. In *Methods of Microarray Data Analysis V*, chapter 2, pages 11–26. Springer.
- [Barrera et al., 2006b] Barrera, J., Cesar-Jr, R. M., Martins-Jr, D. C., Vêncio, R. Z. N., Becerra, C. F., Pereira, C. A. B., and del Portillo, H. A. (2006b). Probabilistic genetic networks analysis of three *Plasmodium falciparum* strains from dynamical expression signals. In *14th Annual International Conference On Intelligent Systems For Molecular Biology (ISMB)*.
- [Barrera et al., 2005] Barrera, J., Cesar-Jr, R. M., Martins-Jr, D. C., Vêncio, R. Z. N., Pereira, C. A. B., and Portillo, H. A. (2005). Estimation of probabilistic genetic networks of *Plasmodium falciparum* from dynamical expression signals. In *X-meeting*, page 58.
- [Basu et al., 2004] Basu, S., Mehreja, R., ThibThiberge, S., Chen, M. T., and Weiss, R. (2004). Spatiotemporal control of gene expression with pulse-generating networks. *Proceedings of National Academy Sciences USA*, 101(17):6355–6360.
- [Becskei and Seraphin, 2000] Becskei, A. and Seraphin, B. (2000). Engineering stability in gene networks by autoregulation. *Nature*, 405(6786):590–593.
- [Bittner et al., 2000] Bittner, M., Meltzer, P., Khan, J., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Gillanders, E., Leja, A., Dietrich, K., Beaudry, C., Berrens, M., Alberts, D., Sondak, V., Hayward, N., and Trent, J. M. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406(6795):536–540.
- [Bozdech et al., 2003] Bozdech, Z., Llinás, M., Pulliam, B. L., Wong, E. D., Zhu, J., and DeRisi, J. L. (2003). The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. *Plos Biology*, 1(1).
- [Butte and Kohane, 2000] Butte, A. and Kohane, I. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Pacific Symposium on Biocomputing*, pages 418–429.

- [Cao et al., 2005] Cao, Y., Wang, P. P., and Tokuta, A. (2005). *Gene Regulating Network Discovery*. Volume 5 of Studies in Computational Intelligence. Springer Verlag.
- [Chen et al., 1999] Chen, T., He, H. L., and Church, G. M. (1999). Modeling gene expression with differential equations. In *Pacific Symposium on Biocomputing*, pages 29–40.
- [Coombs et al., 1970] Coombs, C. H., Dawes, R. M., and Tversky, A. (1970). *Mathematical Psychology: An Elementary Introduction*. Prentice-Hall, Englewood Cliffs, NJ.
- [Cover and van Campenhout, 1977] Cover, T. M. and van Campenhout, J. M. (1977). On the possible orderings in the measurement selection problem. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-7(9):657–661.
- [da Silva and Stumpf, 2005] da Silva, E. and Stumpf, M. P. (2005). Complex networks and simple models in biology. *Journal of the Royal Society, Interface / the Royal Society*, 2(5):419–430.
- [Devroye et al., 1996] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag book series of Application Mathematics, vol. 31, New York.
- [D’haeseleer, 2000] D’haeseleer, P. (2000). *Reconstructing Gene Networks from Large Scale Gene Expression Data*. PhD thesis, University of New Mexico.
- [Dougherty, 2001] Dougherty, E. R. (2001). Small sample issues for microarray-based classification. *Comparative and Functional Genomics*, 2(1):28–34.
- [Dougherty et al., 2007] Dougherty, E. R., Brun, M., Trent, J., and Bittner, M. L. (2007). A conditioning-based model of contextual regulation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.
- [Dougherty et al., 2000] Dougherty, E. R., Kim, S., and Chen, Y. (2000). Coefficient of determination in nonlinear signal processing. *EURASIP Journal on Signal Processing*, 80(10):2219–2235.
- [Dougherty et al., 2005] Dougherty, E. R., Shmulevich, I., Chen, J., and Wang, Z. J. (2005). *Genomic Signal Processing and Statistics*. Eurasip Book Series on Signal Processing and Communications.
- [Duda et al., 2000] Duda, R. O., Hart, P. E., and Stork, D. (2000). *Pattern Classification*. John Wiley & Sons, NY.

- [Eisen et al., 1998] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Science USA*, 95:14863–14868.
- [Elowitz and Leibler, 2000] Elowitz, M. B. and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338.
- [Friedman, 2004] Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805.
- [Friedman et al., 2000] Friedman, N., Linial, M., Nachman, I., and Pe’er, D. (2000). Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620.
- [Goss and Peccoud, 1998] Goss, P. J. and Peccoud, J. (1998). Quantitative modeling of stochastic systems in molecular biology by using stochastic petri nets. *Proceedings of National Academy Science USA*, 95(12):6750–6755.
- [Greller and Somogyi, 2002] Greller, L. D. and Somogyi, R. (2002). Reverse engineers map the molecular switching yards. *Trends Biotechnology*, 20(11):445–447.
- [Hardy and Robillard, 2004] Hardy, S. and Robillard, P. N. (2004). Modeling and simulation of molecular biology systems using petri nets: modeling goals of various approaches. *Journal of Bioinformatics and Computational Biology*, 2(4):595–613.
- [Hartwell et al., 1999] Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–52.
- [Hsing et al., 2005] Hsing, T., Liu, L., Brun, M., and Dougherty, E. R. (2005). The coefficient of intrinsic dependence (feature selection using el cid). *Pattern Recognition*, 38(5):623–636.
- [Jain et al., 2000] Jain, A. K., Duin, R. P. W., and Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37.
- [Jain et al., 1999] Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.
- [Jain and Zongker, 1997] Jain, A. K. and Zongker, D. (1997). Feature-selection: Evaluation, application, and small sample performance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(2):152–157.

- [John et al., 1994] John, G., Kohavi, R., and Phleger, K. (1994). Irrelevant features and the feature subset problem. In *Proceedings of 11th International Conference on Machine Learning*, pages 121–129, San Francisco.
- [Jong, 2002] Jong, H. D. (2002). Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103.
- [Kauffman, 1969] Kauffman, S. A. (1969). Homeostasis and differentiation in random genetic control networks. *Nature*, 224(215):177–178.
- [Kauffman, 1993] Kauffman, S. A. (1993). *The Origins of Order*. Oxford University Press.
- [Kelemen et al., 2008] Kelemen, A., Abraham, A., and Chen, Y. (2008). *Computational Intelligence in Bioinformatics*. Springer.
- [Kendall, 1962] Kendall, M. G. (1962). *Rank correlation methods*. Griffin.
- [Kim et al., 2000a] Kim, S., Dougherty, E. R., Bittner, M. L., Chen, Y., Sivakumar, K., Meltzer, P., and Trent, J. M. (2000a). A general framework for the analysis of multivariate gene interaction via expression arrays. *Biomedical Optics*, 5(4):411–424.
- [Kim et al., 2000b] Kim, S., Dougherty, E. R., Chen, Y., Sivakumar, K., Meltzer, P., Trent, J. M., and Bittner, M. (2000b). Multivariate measurement of gene-expression relationships. *Genomics*, 67:201–209.
- [Kohavi and John, 1997] Kohavi, R. and John, G. (1997). Wrappers for feature selection. *Artificial Intelligence*, 97:273–324.
- [Liang et al., 1998] Liang, S., Furhman, S., and Somogyi, R. (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific Symposium on Biocomputing*, volume 3, pages 18–29.
- [Lizier et al., 2008] Lizier, M. A. S., Martins-Jr, D. C., Cuadros-Vargas, A. J., Cesar-Jr, R. M., and Nonato, L. G. (2008). Generating segmented meshes from textured color images. *Submitted*.
- [Llinás et al., 2006] Llinás, M., Bozdech, Z., Wong, E. D., Adai, A. T., and DeRisi, J. L. (2006). Comparative whole genome transcriptome analysis of three plasmodium falciparum strains. *Nucleic Acids Research*, 34(4):1166–1173.
- [Lopes et al., 2008a] Lopes, F. M., Cesar-Jr, R. M., and Costa, L. F. (2008a). AGN simulation and validation model. In *Advances in Bioinformatics and Computational Biology, Third*

- Brazilian Symposium on Bioinformatics*, volume 5167 of *Lecture Notes in Computer Science*, pages 169–173. Springer-Verlag.
- [Lopes et al., 2008b] Lopes, F. M., Martins-Jr, D. C., and Cesar-Jr, R. M. (2008b). Feature selection environment for genomic applications. *BMC Bioinformatics*, 9:451.
- [Louis and Becskei, 2002] Louis, M. and Becskei, A. (2002). Binary and graded responses in gene networks. *Science's Signal Transduction Knowledge Environment*, 2002(143):PE33.
- [Maki et al., 2001] Maki, Y., Tominaga, D., Okamoto, M., Watanabe, S., and Eguchi, Y. (2001). Development of a system for the inference of large scale genetic networks. In *Pacific Symposium on Biocomputing*, pages 446–458.
- [Margolin et al., 2006] Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla-Favera, R., and Califano, A. (2006). Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1).
- [Martins-Jr., 2004] Martins-Jr., D. C. (2004). Redução de dimensionalidade utilizando entropia condicional média aplicada a problemas de bioinformática e processamento de imagens. Master's thesis, Instituto de Matemática e Estatística - Universidade de São Paulo, Rua do Matão, 1010.
- [Martins-Jr et al., 2008a] Martins-Jr, D. C., Braga-Neto, U., Bittner, M. L., and Dougherty, E. R. (2008a). Network properties of intrinsically multivariate predictive genes. In *IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, Phoenix, AZ, USA.
- [Martins-Jr et al., 2008b] Martins-Jr, D. C., Braga-Neto, U., Hashimoto, R. F., Dougherty, E. R., and Bittner, M. L. (2008b). Intrinsically multivariate predictive genes. *IEEE Journal of Selected Topics in Signal Processing*, 2(3):424–439.
- [Martins-Jr et al., 2005] Martins-Jr, D. C., Cesar-Jr, R. M., and Barrera, J. (2005). Automatic window design for gray-scale image processing based on entropy minimization. In *X Iberoamerican Congress on Pattern Recognition*, Lecture Notes in Computer Science, pages 813–824. Springer-Verlag.
- [Martins-Jr et al., 2006a] Martins-Jr, D. C., Cesar-Jr, R. M., and Barrera, J. (2006a). W-operator window design by minimization of mean conditional entropy. *Pattern Analysis & Applications*, 9:139–153.

- [Martins-Jr et al., 2006b] Martins-Jr, D. C., Cesar-Jr, R. M., Barrera, J., Lizier, M. A. S., and Nonato, L. G. (2006b). W-operator window design for color texture classification. In *Proceedings of XIX Brazilian Symposium on Computer Graphics and Image Processing (SIB-GRAPI)*, Manaus/AM - Brazil. (abstract for poster presentation).
- [Matsuno et al., 2006] Matsuno, H., Inouye, S. T., Okitsu, Y., Fujii, Y., and Miyano, S. (2006). A new regulatory interaction suggested by simulations for circadian genetic control mechanism in mammals. *Journal of Bioinformatics and Computational Biology*, 4(1):139–153.
- [Mazumdar et al., 2006] Mazumdar, J., Wilson, E. H., Masek, K., Hunter, C. A., and Striepen, B. (2006). Apicoplast fatty acid synthesis is essential for organelle biogenesis and parasite survival in *Toxoplasma gondii*. *PNAS*, 103(35):13192–13197.
- [Mendes et al., 2003] Mendes, P., Sha, W., and Ye, K. (2003). Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, 19(Suppl 2):122–129.
- [Milo et al., 2002] Milo, R., Shann-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827.
- [Moore et al., 2005] Moore, J. H., Boczek, E. M., and Summar, M. L. (2005). Connecting the dots between genes, biochemistry, and disease susceptibility: systems biology modeling in human genetics. *Molecular Genetics and Metabolism*, 84(2):104–111.
- [Murphy and Mian, 1999] Murphy, K. and Mian, S. (1999). Modelling gene expression data using dynamic bayesian networks. Technical report, University of California, Berkeley, CA.
- [Noguchi et al., 1993] Noguchi, T., Metz, R., Chen, L., Mattéi, M. G., Carrasco, D., and Bravo, R. (1993). Structure, mapping, and expression of *erp*, a growth factor-inducible gene encoding a nontransmembrane protein tyrosine phosphatase, and effect of *erp* on cell growth. *Mol Cell Biol*, 13(9):5195–5205.
- [Nykter et al., 2008] Nykter, M., Price, N. D., Aldana, M., Ramsey, S. A., Kauffman, S. A., Hood, L. E., Yli-Harja, O., and Shmulevich, I. (2008). Gene expression dynamics in the macrophage exhibit criticality. *PNAS*, 105(6):1897–1900.
- [Pearl, 1988] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.
- [Pe’er et al., 2001] Pe’er, D., Regev, A., Elidan, G., and Friedman, N. (2001). Inferring sub-networks from perturbed expression profiles. *Bioinformatics*, 17(Suppl 1):S215–S224.

- [Pinney et al., 2003] Pinney, J. W., Westhead, D. R., and McConkey, G. A. (2003). Petri net representations in systems biology. *Biochemical Society Transactions*, 31(6):1513–1515.
- [Porter et al., 2001] Porter, D. A., Krop, I. E., Nasser, S., Sgroi, D., Kaelin, C. M., Marks, J. R., Riggins, G., and Polyak, K. (2001). A sage (serial analysis of gene expression) view of breast tumor progression. *Cancer Research*, 61(15):5697–5702.
- [Pournara and Wernisch, 2004] Pournara, I. and Wernisch, L. (2004). Reconstruction of gene networks using bayesian learning and manipulation experiments. *Bioinformatics*, 20(17):2934–2942.
- [Pudil et al., 1994] Pudil, P., Novovicová, J., and Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125.
- [Ralph et al., 2004] Ralph, S. A., Van-Dooren, G. G., Waller, R. F., Crawford, M. J., Fraunholz, M. J., Foth, B. J., Tonkin, C. J., Roos, D. S., and McFadden, G. I. (2004). Metabolic maps and functions of the *Plasmodium falciparum* apicoplast. *Nat. Rev. Microbiol.*, 2:203–216.
- [Raudys and Jain, 1991] Raudys, S. J. and Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):252–264.
- [Reichhardt and Bassler, 2007] Reichhardt, C. J. O. and Bassler, K. E. (2007). Canalization and symmetry in boolean models for genetic regulatory networks. *J. Phys. A: Math. Theor.*, 40:4339–4350.
- [Ris et al., 2008] Ris, M., Martins-Jr, D. C., and Barrera, J. (2008). A branch-and-bound optimization algorithm for u-shaped cost functions on boolean lattices. (*submitted*).
- [Ruklisa et al., 2005] Ruklisa, D., Brazma, A., and Viksna, J. (2005). Reconstruction of gene regulatory networks under the finite state linear model. *Genome Informatics*, 16(1):225–236.
- [Schlitt and Brazma, 2007] Schlitt, T. and Brazma, A. (2007). Current approaches to gene regulatory network modelling. *BMC Bioinformatics*, 8(Suppl 6):59.
- [Shalon et al., 1996] Shalon, D., Smith, S. J., and Brown, P. O. (1996). A dna microarray system for analyzing complex dna samples using two-color fluorescent probe hybridization. *Genome Res*, pages 639–45.
- [Shannon and Weaver, 1963] Shannon, C. E. and Weaver, W. (1963). *The mathematical theory of communication*. Univ. of Illinois Press.

- [Sherman, 1983] Sherman, I. W. (1983). Metabolism and surface transport of parasitized erythrocytes in malaria. In *Ciba Found Symp.*, pages 206–221.
- [Shmulevich et al., 2002] Shmulevich, I., Dougherty, E. R., Kim, S., and Zhang, W. (2002). Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18:261–274.
- [Silva et al., 2005] Silva, P. J. S., Hashimoto, R. F., Kim, S., Barrera, J., Brandão, L. O., Suh, E., and Dougherty, E. R. (2005). Feature selection algorithms to find strong genes. In *Pattern Recognition Letters*, volume 26, pages 1444–1453.
- [Silvescu and Honavar, 1997] Silvescu, A. and Honavar, V. (1997). Temporal boolean network models of genetic networks and their inference from gene expression time series. *Complex Systems* 13, 13(1):54–75.
- [Smalley, 2003] Smalley, K. S. (2003). A pivotal role for erk in the oncogenic behaviour of malignant melanoma? *Int J Cancer*, 104(5):527–532.
- [Soinov et al., 2003] Soinov, L. A., Krestyaninova, M. A., and Brazma, A. (2003). Towards reconstruction of gene networks from expression data by supervised learning. *Genome Biology*, 4(1):R6.
- [Somogyi et al., 1997] Somogyi, R., Fuhrman, S., Askenazi, M., and Wuensche, A. (1997). The gene expression matrix: towards the extraction of genetic network architectures. In *Nonlinear Analysis, Theory, Methods and Applications*, volume 30, pages 1815–1824.
- [Somol and Pudil, 2004] Somol, P. and Pudil, P. (2004). Fast branch & bound algorithms for optimal feature selection. *Pattern Analysis and Machine Intelligence*, 26(7):900–912.
- [Somol et al., 1999] Somol, P., Pudil, P., Novovicová, J., and Paclík, P. (1999). Adaptive floating search methods in feature selection. *Pattern Recognition Letters*, 20:1157–1163.
- [Theodoridis and Koutroumbas, 1999] Theodoridis, S. and Koutroumbas, K. (1999). *Pattern Recognition*. Academic Press, USA, 1st edition.
- [Thomas, 1998] Thomas, R. (1998). Laws for the dynamics of regulatory networks. *The International Journal of Developmental Biology*, 42(3):479–485.
- [Thomas et al., 1995] Thomas, R., Thieffry, D., and Kaufman, M. (1995). Dynamical behaviour of biological regulatory networks-i. biological role of feedback loops and practical use of the concept of the loop-characteristic states. *Bull Math Biol*, 57(2):247–276.

- [Vapnik, 1998] Vapnik, V. N. (1998). *Statistical Learning Theory*. John Wiley & Sons, New York.
- [Varadan et al., 2006] Varadan, V., III, D. M., and Anastassiou, D. (2006). Computational inference of the molecular logic for synaptic connectivity in *c. elegans*. *Bioinformatics*, 22(14):e497–e506.
- [Waddington, 1942] Waddington, C. H. (1942). Canalization of development and the inheritance of acquired characters. *Nature*, pages 563–565.
- [Watanabe, 1985] Watanabe, S. (1985). *Pattern Recognition: Human and Mechanical*. New York: Wiley.
- [Wilson, 2002] Wilson, R. J. M. (2002). Progress with parasite plastids. *Journal of Molecular Biology*, 319:257–274.
- [Woolf and Wang, 2000] Woolf, P. and Wang, Y. (2000). A fuzzy logic approach to analyzing gene expression data. *Physiological Genomics*, 3(1):9–15.
- [Xing et al., 2001] Xing, E. P., Jordan, M. I., and Karp, R. M. (2001). Feature selection for high-dimensional genomic microarray data. In *Proc. 18th International Conference on Machine Learning*, pages 1019–1022, San Mateo-CA.
- [Yung et al., 2003] Yung, S. C., Unnasch, T. R., and Lang-Unnasch, N. (2003). Cis and trans factors involved in apicoplast targeting in *toxoplasma gondii*. *Journal of Parasitology*, 89(4):767–776.