UNIVERSIDADE DE SÃO PAULO

Programa de Pós-Graduação Interunidades em Bioinformática

ANA CAROLINA QUIRINO SIMÕES

**Planejamento, gerenciamento e análise de dados de
microarranjos de DNA para identificação de biomarcadores
de diagnóstico e prognóstico de cânceres humanos.**

São Paulo

04/02/2009

ANA CAROLINA QUIRINO SIMÕES¹

**Planejamento, gerenciamento e análise de dados de
microarranjos de DNA para identificação de biomarcadores
de diagnóstico e prognóstico de cânceres humanos.**

*Tese apresentada ao Programa de Pós-Graduação
Interunidades em Bioinformática da
Universidade de São Paulo para
obtenção do Título de Doutor em
Ciências (Bioinformática)*

Orientadores:

Profa. Dra. Aline Maria da Silva
(Instituto de Química – USP)

Prof. Dr. Eduardo Jordão Neves
(Instituto de Matemática e Estatística – USP)

São Paulo

04/02/2009

¹ Durante o desenvolvimento desta tese, a aluna foi bolsista da CAPES e FAPESP.

ANA CAROLINA QUIRINO SIMÕES

Planejamento, gerenciamento e análise de dados de microarranjos de DNA para identificação de biomarcadores de diagnóstico e prognóstico de cânceres humanos.

Tese apresentada ao Programa de Pós-Graduação Interunidades em Bioinformática da Universidade de São Paulo para obtenção do Título de Doutor em Ciências, área Bioinformática.

Aprovada em: _____

Banca Examinadora

Prof. Dr. _____

Instituição: _____

Assinatura: _____

Prof. Dr. _____

Instituição: _____

Assinatura: _____

Prof. Dr. _____

Instituição: _____

Assinatura: _____

Prof. Dr. _____

Instituição: _____

Assinatura: _____

Prof. Dr. _____

Instituição: _____

Assinatura: _____

Dedico este trabalho aos meus pais Edda e Plínio

AGRADECIMENTOS

- À professora Aline Maria da Silva pela orientação, dedicação e por todos os ensinamentos desde o início da minha vida acadêmica.
- Ao professor Eduardo Jordão Neves pela co-orientação e ensinamentos.
- Aos meus pais Edda e Plinio pelo eterno carinho e pelos bons exemplos.
- A minha família: Paula, Jamila, Nalini, Taj, Surya, Pedro, Olga (*in memorium*), Leda (*in memorium*), Fausto (*in memorium*), Ivete (*in memorium*), Luzia César e César.
- Aos grandes amigos Rodrigo Tamura, Paulo Medeiros, Joana Frias, Mirian Morale, Celina Takemura, Ana Beatriz Graciano, Beatriz Stransky, Jihan Zoghbi, Rogério Rizzo, Tatyana Fittipaldi, Felipe Fittipaldi, Helena Lima e Luiz Otávio Conceição por todo o apoio e carinho.
- Aos amigos do IME, David Pires, Jesus Mena, David Correia, Gustavo Esteves, Lucas Fahham, Michel Mozinho, Fernanda Almeida, Fabrício Lopes, Alex Paschoal, Ariane Machado e Kelly Braghetto pela ajuda, pelo companheirismo e pelas variadas sugestões ao longo do desenvolvimento deste trabalho.
- Aos amigos do IQ, Alexandre Sanchez, Daniela Gonzalez, Daniela Beton, Layla Farage, Fabiana Almeida, Paulo Zaini, Paulo Paiva, Thiago Venancio, Patrícia Pessoa e Renato Astolfi pela ajuda deste que me associei ao grupo da Profa. Aline no IQ-USP.
- Ao Dr. Luiz Fernando Reis, Isabela W. Cunha, Anna Coló, Nair Muto, Kátia Carvalho e Alex Fiorini do Hospital A. C. Camargo por compartilharem seus resultados e por sempre estarem dispostos a me ajudar.
- Ao professores Roberto Hirata Jr., Roberto Marcondes e Junior Barrera do *Computer Vision Research Group* no IME pelas sugestões no decorrer deste trabalho.
- Aos Drs. Antonio Galves e Pablo Ferrari do Núcleo de Apoio à Pesquisa em Modelagem Estocástica e Complexidade (NUMEC) da USP pelos comentários ao trabalho.
- Aos professores Sergio Verjovski-Almeida e Eduardo Moraes Rego Reis e a doutoranda Ângela Fachel, do IQ-USP, por compartilharem resultados e pelas sugestões em várias etapas deste trabalho.

- Ao Prof. Fabio Kon (IME-USP) por ter me apresentado a “filosofia” eXtreme Programming.
- A Patrícia Martorelli, secretária do Programa de Pós-Graduação Interunidades em Bioinformática, pela atenção e lembretes sempre necessários.
- A comunidade do IME-USP, em especial do NUMEC, por proporcionar condições adequadas para o desenvolvimento deste trabalho.
- À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa por seis meses no início do curso de Doutorado.
- À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pela concessão da bolsa de Doutorado por 44 meses e pelo apoio financeiro que possibilitou a realização desse trabalho.

Coming together is a beginning. Keeping together is progress. Working together is success.

Henry Ford

RESUMO

Simões, A.C.Q. **Planejamento, gerenciamento e análise de dados de microarranjos de DNA para identificação de biomarcadores de diagnóstico e prognóstico de cânceres humanos**. 2009. 137p. Tese de Doutorado - Programa de Pós-Graduação Interunidades em Bioinformática. Universidade de São Paulo, São Paulo.

Nesta tese, apresentamos nossas estratégias para desenvolver um ambiente matemático e computacional para análises em larga-escala de dados de expressão gênica obtidos pela tecnologia de microarranjos de DNA. As análises realizadas visaram principalmente à identificação de marcadores moleculares de diagnóstico e prognóstico de cânceres humanos. Apresentamos o resultado de diversas análises implementadas através do ambiente desenvolvido, as quais conduziram a implementação de uma ferramenta computacional para a anotação automática de plataformas de microarranjos de DNA e de outra ferramenta destinada ao rastreamento da análise de dados realizada em ambiente R. Programação eXtrema (eXtreme Programming, XP) foi utilizada como técnica de planejamento e gerenciamento dos projetos de análise dados de expressão gênica. Todos os conjuntos de dados foram obtidos por nossos colaboradores, utilizando-se duas diferentes plataformas de microarranjos de DNA: a primeira enriquecida em regiões não-codificantes do genoma humano, em particular regiões intrônicas, e a segunda representando regiões exônicas de genes humanos. A primeira plataforma foi utilizada para avaliação do perfil de expressão gênica em tumores de próstata e rim humanos, sendo que análises utilizando SAM (*Significance Analysis of Microarrays*) permitiram a proposição de um conjunto de 49 sequências como potenciais biomarcadores de prognóstico de tumores de próstata. A segunda plataforma foi utilizada para avaliação do perfil de transcritos expressos em sarcomas, carcinomas epidermóide e carcinomas epidermóides de cabeça e pescoço. As análises com sarcomas permitiram a identificação de um conjunto de 12 genes relacionados à agressividade local e metástase. As análises com carcinomas epidermóides de cabeça e pescoço permitiram a identificação de 7 genes relacionados à metástase linfonodal.

Palavras-chave: Análise de dados em larga escala, Marcadores Moleculares, Câncer, Microarranjos de DNA, Programação eXtrema, Classificadores

ABSTRACT

Simões, A.C.Q. **Planning, management and analysis of DNA *microarray* data aiming at discovery of biomarkers for diagnosis and prognosis of human cancers.** 2009. 137p. Tese de Doutorado - Programa de Pós-Graduação Interunidades em Bioinformática. Universidade de São Paulo, São Paulo.

In this PhD Thesis, we present our strategies to the development of a mathematical and computational environment aiming the analysis of large-scale microarray datasets. The analyses focused mainly on the identification of molecular markers for diagnosis and prognosis of human cancers. Here we show the results of several analyses implemented using this environment, which led to the development of a computational tool for automatic annotation of DNA microarray platforms and a tool for tracking the analysis within R environment. We also applied eXtreme Programming (XP) as a tool for planning and management of gene expression analyses projects. All data sets were obtained by our collaborators using two different microarray platforms. The first is enriched in non-coding human sequences, particularly intronic sequences. The second one represents exonic regions of human genes. Using the first platform, we evaluated gene expression profiles of prostate and kidney human tumors. Applying SAM to prostate tumor data revealed 49 potential molecular markers for prognosis of this disease. Gene expression in samples of sarcomas, epidermoid carcinomas and head and neck epidermoid carcinomas was investigated using the second platform. A set of 12 genes were identified as potential biomarkers for local aggressiveness and metastasis in sarcoma. In addition, the analyses of data obtained from head and neck epidermoid carcinomas allowed the identification of 7 potential biomarkers for lymph-nodal metastases.

Keywords: Large-scale data analysis, Molecular Markers, Cancer, DNA *microarrays*, eXtreme Programming, Classifiers

LISTA DE ABREVIATURAS E SIGLAS

ACCamargo3	Plataforma de dados de <i>microarrays</i> desenvolvida no Hospital A.C. Camargo.
AMPiR	<i>Annotate Microarray Platform in R</i>
BASE	<i>BioArray Software Environment</i>
Bioconductor	Projeto de desenvolvimento de <i>software</i> livre para análise de dados genômicos
bp	<i>Base pair</i>
cDNA	DNA complementar
CE	Carcinomas epidermóides
CECP	Carcinomas epidermóides de cabeça e pescoço
ESTs	<i>Expressed Sequence Tags</i>
E-utilities	<i>Entrez Programming Utilities</i>
GEO	<i>Gene Expression Omnibus</i>
gi	<i>GenInfo Identifier</i> (Identificador primário nas bases de dados do NCBI)
GO	<i>Gene Ontology</i>
KEGG	<i>Kyoto Encyclopedia of Genes and Genomes</i>
MAGE-ML	<i>MicroArray Gene Expression Markup Language</i>
MIAME	<i>Minimum Information About a Microarray Experiment</i>
MAIGES	<i>Mathematical Analysis of Interacting Gene Expression Systems</i>
MaigesPack	Pacote de análise de dados de <i>microarrays</i> desenvolvido pelo grupo MAIGES
ncRNA	RNA não codificante
NCBI	<i>National Center for Biotechnology Information</i>
PCR	Reação em Cadeia da Polimerase
PSA	Antígeno Prostático Específico
R	Plataforma de análise estatística
RCC	Carcinoma Renal de Célula Clara
SAGE	<i>Serial Analysis of Gene Expression</i>
SAM	<i>Significance Analysis of Microarrays</i>
SNP	Polimorfismo de nucleotídeo único
STKE	<i>The Signal Transduction Knowledge Environment</i>
TyPiR	<i>Track your Project in R</i>
XP	eXtreme Programming

SUMÁRIO

1. Considerações Iniciais e Objetivos	1
1.1. Organização do texto	2
2. Planejamento, gerenciamento e ambiente computacional	3
2.1. Introdução: Problemas em projeto de larga escala	3
2.1.1. Uma equipe de trabalho.....	3
2.1.2. As diversas estratégias.....	3
2.1.3. Ambiente de condução da análise	6
2.1.4. Ambiente de desenvolvimento	7
2.1.5. Interface de disponibilização dos dados e resultados	8
2.1.6. Rastreabilidade e reprodutibilidade.....	8
2.1.7. Anotação dos dados biológicos	9
2.2. Métodos.....	10
2.2.1. Instalação do BASE e implementação de <i>plugin</i>	10
2.2.2. Utilização do R + pacotes.....	12
2.2.3. Implementação do Moodle	12
2.2.4. Fundamentos de XP	12
2.2.5. Anotação dos dados.....	17
2.2.6. Rastreamento das análises	18
2.3. Resultados	18
2.3.1. Comparação das plataformas de condução de análise: BASE x R.....	18
2.3.1.1. Plugin SAM e interface	18
2.3.1.2. Persistência de Objetos: R + pacotes.....	19
2.3.1.3. Discussão	21
2.3.2. Um modelo para gerenciar análises de <i>microarrays</i> usando XP.....	22
2.3.2.1. O modelo.....	22
2.3.2.2. Escopo do modelo e discussões	27
2.3.3. A ferramenta AMPiR	28
2.3.3.1. Introdução	28
2.3.3.2. Especificação e Resultados: A anotação da plataforma ACCamargo3	29
2.3.3.3. Discussão	34
2.3.4. A ferramenta TyPiR	35
2.3.4.1. Introdução	35
2.3.4.2. Especificação e requisitos	36
2.3.4.3. Discussão	38
3. Busca por potenciais biomarcadores de diagnóstico e prognóstico de cânceres humanos	39
3.1. Introdução	39
3.1.1. Tecnologias da genômica para análise de expressão gênica no câncer	39
3.1.2. Marcadores moleculares no diagnóstico e prognóstico de câncer.....	41
3.1.3. O câncer de próstata e seus marcadores moleculares	43
3.1.4. Os Carcinomas Renais do tipo Célula Clara e seus marcadores moleculares	44
3.1.5. Os Sarcomas e seus marcadores moleculares.....	45
3.1.6. Os Carcinomas Epidermóides e seus marcadores moleculares	46
3.2. Métodos.....	47
3.2.1. A plataforma de microarranjos <i>non-coding</i>	48
3.2.2. Conjunto de dados: câncer de próstata	49
3.2.2.1. Filtragem e seleção com SAM para a análise de normais x tumorais	50
3.2.2.2. Filtragem e seleção com SAM para a análise de marcadores de prognóstico	51
3.2.2.3. <i>Clustering</i> hierárquico e validação.....	53
3.2.3. Conjunto de dados: Carcinoma renal do tipo célula clara	54

3.2.3.1.	Primeira estratégia de redução: genes diferencialmente expressos	55
3.2.3.2.	Segunda estratégia de redução: genes mais expressos	56
3.2.4.	A plataforma de <i>microarray</i> ACCamargo3 e análises de dados	57
3.2.5.	Conjunto de dados: Sarcoma.....	59
3.2.6.	Conjunto de dados: Carcinomas Epidermóides.....	63
3.2.7.	Conjunto de dados: Carcinomas Epidermóides de Cabeça e Pescoço	64
3.3.	Resultados	66
3.3.1.	Descoberta de potenciais marcadores para prognóstico do câncer de próstata	66
3.3.1.1.	Análise de normais e tumorais	66
3.3.1.2.	Análise usando somente as amostras tumorais.....	67
3.3.1.3.	Validação do clustering	72
3.3.1.4.	Discussão	74
3.3.2.	Descoberta de marcadores para Carcinomas Renais de Célula Clara	75
3.3.2.1.	Trios com genes diferencialmente expressos	75
3.3.2.2.	Trios com os genes mais expressos e também com o “busca e escolha”	77
3.3.2.3.	Discussão	81
3.3.3.	Descoberta de marcadores para Sarcoma	82
3.3.3.1.	A primeira etapa do projeto e os primeiros resultados	82
3.3.3.2.	A segunda etapa do projeto: XP + Fibromatoses como tema central	87
3.3.3.3.	Resultados para a comparação Fibromatoses x Benignos	91
3.3.3.4.	Resultados para a comparação Fibromatoses x Malignos	94
3.3.3.5.	Resultados para a comparação benignos x malignos	95
3.3.3.6.	Marcadores de agressividade local e de metástase.....	96
3.3.3.7.	Marcadores que diferenciam, dentro dos tumorais, entre alto grau e baixo grau	99
3.3.3.8.	Marcadores que usam a informação de grupos para separar as amostras	100
3.3.3.9.	Discussão das análises de Sarcoma	101
3.3.4.	Descoberta de marcadores para Carcinomas Epidermóides de Cabeça e Pescoço	101
3.3.4.1.	XP e as análises.....	101
3.3.4.2.	A exclusão de CNRP de muitas análises e tabelas de contingência	107
3.3.4.3.	Análises de módulos funcionais relacionados a metástase linfonodal	113
3.3.4.4.	Árvores de decisão para os genes em validação.....	115
3.3.4.5.	Discussão das análises de Carcinomas Epidermóides de Cabeça e Pescoço.....	117
3.3.5.	Descoberta de marcadores para Carcinomas Epidermóides.....	117
3.3.5.1.	XP e as análises	117
3.3.5.2.	A questão das topologias.....	118
3.3.5.3.	Discussão das análises de Carcinomas Epidermóides	121
4.	Discussão Final.....	121
5.	Bibliografia	125
	Lista de Anexos	138

1. Considerações Iniciais e Objetivos

Esta tese se insere no contexto de projetos de larga escala para análise de expressão gênica, em que um grande volume de dados precisa ser analisado, rastreado, atualizado, documentado e disponibilizado, com toda a variabilidade de parâmetros e técnicas inerentes. Nos últimos anos várias plataformas computacionais para armazenamento e/ou análise de dados de expressão gênica foram desenvolvidas e disponibilizadas (Baldi & Long, 2001; Colantuoni *et al.*, 2002; Saal *et al.*, 2002; Causton *et al.*, 2003; Dudoit *et al.*, 2003; Saeed *et al.*, 2003; Sarkans *et al.*, 2005; Troein *et al.*, 2006; Parkinson *et al.*, 2009). Entretanto, entendemos que gerenciar projetos desta natureza exige muito mais do que a inserção e análise de dados em uma plataforma computacional. É necessário primeiro decidir quais hipóteses biológicas serão investigadas e quais modelos estatísticos serão utilizados. Em seguida, quais as ferramentas e métodos de análise que serão empregados, quais os requisitos para que a análise ocorra e então avaliar os resultados e possivelmente reinseri-los na análise, para se aprofundar nas perguntas biológicas que parecem ter respostas mais promissoras. Isso tudo realizado em conjuntos de dados relativos a milhares de genes, anotações de bancos externos, com centenas de amostras biológicas e diversas variáveis clínicas, ou seja: um grande desafio. É preciso ainda que, a qualquer momento do projeto, todas as análises possam ser rastreadas bem como os parâmetros utilizados e os resultados obtidos. É desejável também que haja espaço para a incorporação de novas ferramentas à plataforma computacional.

Neste cenário, esta tese de Doutorado tem dois objetivos principais:

- a) apresentar soluções computacionais para resolver a questão do planejamento, gerenciamento e análise de dados expressão gênica obtidos através da técnica de microarranjos de DNA;
- b) apresentar contribuições destas soluções computacionais na busca por potenciais biomarcadores de diagnóstico e prognóstico de cânceres humanos.

Esta tese apresenta os resultados de nossa participação em cinco projetos de análise de dados de expressão gênica em cânceres humanos. O contato com estes projetos com objetivos tão variados foi bastante importante na especificação do ambiente de análise que implementamos. Os dados foram obtidos através da tecnologia

de microarranjos de DNA (também referenciada neste trabalho por *microarray*) por dois grupos de pesquisa com que estabelecemos colaborações: o grupo liderado pelo Dr. Luiz Fernando Lima Reis (Hospital A. C. Camargo) e o grupo liderado pelo Prof. Sergio Verjovski-Almeida (Instituto de Química, USP).

As principais contribuições computacionais que geramos neste trabalho são o modelo de aplicação de Programação eXtrema (XP) para condução de análises de dados de microarranjos de DNA, a ferramenta AMPiR (*Annotate Microarray Platform in R*) e a ferramenta TyPiR (*Track your Project in R*). Em relação à identificação de biomarcadores para diagnóstico e prognóstico do câncer humano, nós mostramos a aplicação do método SAM (*Significance Analysis of Microarrays*) como uma abordagem para selecionar candidatos a biomarcadores no câncer de próstata humano. Nós também identificamos novos candidatos a biomarcadores para o diagnóstico e prognóstico de sarcomas e carcinomas epidermóides como resultado da aplicação das soluções computacionais que desenvolvemos e que contemplam planejamento, gerenciamento e análise de dados expressão gênica obtidos através da técnica de microarranjos de DNA.

1.1. Organização do texto

As contribuições resultantes de nosso trabalho serão apresentadas em duas seções separadas. Primeiramente abordaremos as questões relativas ao planejamento gerenciamento e ambiente computacional para análise de dados de expressão gênica (Seção 2). As análises de dados de microarranjos de DNA, a que também referenciaremos neste trabalho como análises de dados de *microarrays*, visando à identificação de potenciais biomarcadores para os diferentes cânceres humanos estudados serão apresentadas na Seção 3. Cada uma dessas seções contempla uma breve introdução sobre o tópico, a contextualização dos problemas, a metodologia escolhida para resolver tais problemas, os resultados obtidos e sua discussão. A síntese das nossas contribuições e as perspectivas futuras deste trabalho estão reunidas na discussão final (Seção 4).

O detalhamento de alguns conceitos empregados nesta tese que julgamos ser úteis ao leitor bem como de informações sobre os manuscritos oriundos desta tese, súmula curricular e histórico escolar foram incluídos como anexos em CD-ROM suplementar a tese.

2. Planejamento, gerenciamento e ambiente computacional

2.1. Introdução: Problemas em projeto de larga escala

2.1.1. Uma equipe de trabalho

Projetos de larga escala na área da genômica exigem a interação e a contribuição de diversos profissionais organizados em grandes equipes. Os profissionais que compõem essas equipes são das áreas de biológicas (bioquímica, biologia molecular, microbiologia, farmácia, medicina, etc), de computação, de estatística e de bioinformática. Como exemplos de projetos de larga escala, citamos: *The Human Cancer Genome Project* (HCGP), *The Cancer Genome Anatomy Project* (CGAP), *ENCODE Project Consortium*, Projetos de sequenciamento do genoma humano, do genoma de *Dictyostelium discoideum*, de ESTs da cana-de-açúcar (SUCEST) e o do genoma da bactéria *Xylella fastidiosa* (Simpson *et al.*, 2000; Strausberg *et al.*, 2000; Camargo *et al.*, 2001; Lander *et al.*, 2001; Venter *et al.*, 2001; Vettore *et al.*, 2003; Eichinger *et al.*, 2005; Birney *et al.*, 2007). Em sua maioria são projetos inovadores, que costumam ter duas principais vertentes de desenvolvimento: de resultados biológicos e de resultados computacionais.

A necessidade de interação entre os tipos profissionais que compõem equipes multidisciplinares em projetos de genômica requer estratégias de planejamento, gerenciamento e comunicação para maximizar o sucesso do projeto. Nesse contexto, como será detalhado adiante, apresentamos um modelo de gerenciamento de análise de dados de microarranjos de DNA usando XP. O modelo propõe uma estratégia que compatibiliza o fato da equipe multidisciplinar trabalhar simultaneamente com dados e ferramentas de cunho computacional e biológico. O desenvolvimento das ferramentas AMPiR e TyPiR apresentadas nesta tese também considera uma equipe multidisciplinar, que tenha necessariamente um ou mais profissionais da área de bioinformática.

2.1.2. As diversas estratégias

Uma análise de dados de microarranjos de DNA poderia ser resumida de forma algo simplista a uma análise com milhares de genes envolvendo diversos passos, cada

um deles orientados por uma hipótese (biológica ou experimental) e com diversos parâmetros e suposições. A diversidade na análise de dados de microarranjos de DNA pode estar na plataforma de microarranjos eleita, no desenho experimental das amostras biológicas e das hibridizações, na plataforma computacional escolhida para a análise e também nos diversos métodos estatísticos e computacionais que podem ser aplicados para a investigação de uma mesma hipótese biológica. Justamente para suprir os desafios de rastreabilidade, confiabilidade e atualização que essa definição implica, surgiram diversas plataformas de análises de dados de *microarray*: BASE, ArrayExpress, SpotFire® e R como pacotes do Bioconductor (Saal *et al.*, 2002; Dudoit *et al.*, 2003; Saeed *et al.*, 2003; Gentleman *et al.*, 2004; Kaushal & Naeve, 2004; Sarkans *et al.*, 2005; Parkinson *et al.*, 2009). Elas diferem quanto a estratégias de modelagem dos dados, persistência dos dados, funcionalidades, capacidade de interagir com outras funcionalidades, flexibilidade e documentação. Algumas são *open source*, como o BASE, ArrayExpress, TM4, R e pacotes do Bioconductor; outras são *software* proprietários, como o SpotFire®. Algumas usam sistemas gerenciadores de bancos de dados, como o ArrayExpress, BASE e TM4.

A iniciativa do projeto Bioconductor é de agregar programas *open source* para a análise e compreensão de dados de genômica e bioinformática (<http://www.bioconductor.org>) (Gentleman *et al.*, 2004). A versão 2.2 do Bioconductor distribui no total 260 pacotes, incluindo o MaigesPack usado nesta tese (Esteves, 2007). Há também outras iniciativas de projetos de *software open source* para bioinformática, como o BioPerl (<http://BioPerl.org>), BioPhyton (<http://BioPhyton.org>) e BioJava (<http://BioJava.org>).

Diversos são os métodos estatísticos e computacionais que podem ser aplicados em conjuntos de dados de expressão: análise de genes diferencialmente expressos, *clustering* dos genes, busca por módulos de genes concordantes com uma classificação, análise da concordância de uma rede genética nos dados e busca por classificadores (Butte *et al.*, 2000; Quackenbush, 2001; Tusher *et al.*, 2001; Quackenbush, 2002; Causton *et al.*, 2003; Segal *et al.*, 2004). Para cada um desses diversos métodos, diversos parâmetros podem ser aplicados. Por exemplo, para análises de genes diferencialmente expressos podem-se aplicar métodos paramétricos, como o teste t ou não paramétricos, como o teste de Wilcoxon. Pode-se ainda aplicar o teste *Significance Analysis of Microarrays* (SAM), que é uma variante do teste t (Tusher *et al.*, 2001).

Naturalmente, cada método pressupõe um conjunto de suposições estatísticas cuja validade deve ser verificada. É importante notar também que a aplicação de cada um dos diversos métodos numa análise de dados de microarranjos é sempre guiada pela hipótese biológica que se está testando. Isso aponta para uma importante funcionalidade da interface. Porém, isso não é o que se encontra atualmente nas interfaces das plataformas computacionais de dados de microarranjos, pois elas enfatizam e explicitam a técnica aplicada (Saal *et al.*, 2002), como discutido na seção que trata da comparação das plataformas BASE e R (seção 2.3.1).

A diversidade também existe na montagem da equipe de profissionais que participará da análise. Pode ser que ela seja composta por profissionais das quatro áreas: bioinformática, biologia, computação e estatística (ordenados por ordem alfabética). Porém, dada a raridade de profissionais da área de bioinformática no mercado profissional e acadêmico e dada à competitividade do mercado de computação, nem sempre projetos de genômica possuem o apoio computacional e a estabilidade de profissionais ideal.

Como última fonte de variabilidade que apontamos, temos um dos grandes desafios da bioinformática e genômica: a questão de falta de padrões. Algumas propostas de padronização já foram feitas à comunidade de bioinformática e de *systems biology*, dentre as quais podemos citar MIAME (Brazma *et al.*, 2001), MAGE-ML (Spellman *et al.*, 2002; Ball & Brazma, 2006) e CELLML (Lloyd *et al.*, 2008). A padronização MIAME (*Minimum Information About a Microarray Experiment*) sugere um conjunto mínimo de informações de um experimento de microarranjo de DNA que deveria ser disponibilizado para que qualquer outro pesquisador consiga reproduzir o experimento. Já o padrão MAGE-ML (*MicroArray Gene Expression Markup Language*) refere-se à *markup language* para se representar dados de microarranjos que cumprem o MIAME. O padrão CELLML refere-se à *markup language* para representar modelos matemáticos de redes biológicas, isto é, da área de *sistems biology*. A adoção de padrões favorece a automatização das análises, comparação entre resultados encontrados entre diversos experimentos e análises de dados, assim como favoreceria a integração de plataformas bem como a integração do conhecimento biológico, por permitir que as várias iniciativas de anotação sejam integradas. É o que pode ser verificado com a base de dados GEO (*Gene Expression Omnibus*) do NCBI (Barrett *et al.*, 2009), que está interligada a todas as outras diversas bases de dados da mesma instituição através de *links* (ver seção 2.3.4).

2.1.3. Ambiente de condução da análise

Por ambiente de condução da análise subentendemos uma plataforma que permite a visualização das análises realizadas, bem como dos resultados obtidos. Idealmente ela deveria documentar os parâmetros usados em cada método, para que se possa reproduzir a análise caso necessário. Ele deve ser um ambiente amigável e de preferência intuitivo para o desenvolvedor da análise, tipicamente o bioinformata ou programador. Pode ser que a interface seja ao estilo de uma Shell, como o R acoplado com pacotes do Bioconductor (Dudoit *et al.*, 2003; Gentleman *et al.*, 2004) ou que tenha uma interface gráfica, como os encontrados para o BASE, TM4 e SpotFire® (Saal *et al.*, 2002; Saeed *et al.*, 2003; Kaushal & Naeve, 2004). Deve ser um ambiente também que faça uso de índices para organizar os dados, frente ao problema de se ter muitas análises em um projeto de análise de dados de *microarray*.

Quando o ambiente, como o ArrayExpress, BASE e TM4, usa um sistema gerenciador de bancos de dados, como o MySQL, PostgreSQL ou ORACLE, tem-se a persistência dos dados atrelada ao banco, por meio da modelagem dos dados. Isso pode vir a restringir a inserção de conjuntos de dados que saiam do padrão que foi modelado. Surge então o problema: Será que vou conseguir analisar meus dados usando essa plataforma? Tem como adaptá-los de alguma forma? O problema se agrava ainda mais quando a inadequação da modelagem é percebida com a análise em curso: Mudaremos de plataforma? Adaptações e mudanças nesse sentido implicam em perdas ao projeto, de tempo, pessoas e dinheiro, e devem ser minimizadas ao máximo.

A utilização de uma plataforma com menos restrições, como a opção por R acoplado com os diversos pacotes disponíveis de estatística e os voltados para genômica, por meio do Bioconductor, torna-se então interessante nesse contexto. Porém, essa opção demanda que se tenha um desenvolvedor envolvido na equipe do projeto que desenvolva programas em R e que, preferencialmente, entenda da questão biotecnológica, ou seja, um bioinformata. Ela também acarreta problemas de persistência dos dados, rastreamento e interface para a visualização dos resultados.

A utilização de índices para localizar análises realizadas também é de suma importância, uma vez que diversas análises são realizadas ao longo de um projeto de análises de dados de *microarray*. Logo, a eleição de índice eficiente torna-se um fator importante para o sucesso do ambiente de condução da análise.

2.1.4. Ambiente de desenvolvimento

Por estarmos tratando de projetos de larga escala e ser a área de bioinformática relativamente nova, bem como a área de desenvolvimento de *software* que apóia a genômica, é comum que concomitantemente ao desenvolvimento dos resultados biológicos do projeto, se tenha o desenvolvimento de ferramentas computacionais (Kent *et al.*, 2002; Paquola *et al.*, 2003a; Paquola *et al.*, 2003b; Chisholm *et al.*, 2006; Koide *et al.*, 2006). Portanto, quando falamos em ambiente de desenvolvimento, estamos nos referindo à plataforma de suporte computacional para desenvolvimento de novas funcionalidades e acoplamento das mesmas à plataforma de análise dos dados. Naturalmente, tem-se que quanto mais fácil for acoplar uma nova funcionalidade/ferramenta à plataforma de análise de dados, maior será o ganho de toda a equipe. Porém é importante ressaltar que a adoção de uma nova funcionalidade/ferramenta deve ser fácil. A questão de sua validade, no entanto, fica a cargo dos pesquisadores envolvidos na análise de dados.

A plataforma BASE (Saal *et al.*, 2002) resolveu parcialmente esse problema usando o conceito de *plugins*. Assim, podem-se adicionar algumas novas análises ao padrão da plataforma se elas tiverem sido transformadas em *plugins*. Na prática isso significa que a ferramenta receberá sempre um arquivo de certo tipo e então saberá o que fazer com ele. Mas e se a nova funcionalidade a ser desenvolvida precisa de outros campos que não são os padrões do formato exportado? Foi com o intuito de trazermos mais flexibilidade à plataforma que desenvolvemos uma interface que faz uma meta-análise do arquivo BaseFile disponibilizado pelo BASE e decide qual a entrada que será passada ao *plugin* (Simões & da Silva, 2006).

A flexibilidade do R existe através da possibilidade de utilização dos diversos pacotes disponibilizados no repositório CRAN do projeto “*The R Project for Statistical Computing*” ou da utilização dos pacotes do Bioconductor (Dudoit *et al.*, 2003; Gentleman *et al.*, 2004). O R ainda permite que se importe e se execute códigos/funções escritos em R, sem que esses estejam em pacotes fechados.

2.1.5. Interface de disponibilização dos dados e resultados

A interface de disponibilização dos dados e resultados a que estamos nos referindo é a interface gráfica que permite aos colaboradores da área de biológicas visualizarem os resultados obtidos. O ideal é que ela seja intuitiva ao usuário, leve ao mínimo de erros de manipulação e interpretação possíveis e ainda permita ao usuário inserir suas observações. Grandes avanços têm sido feitos na base de dados GEO do NCBI nesse sentido (Barrett *et al.*, 2009; Sayers *et al.*, 2009), em que se podem visualizar alguns resultados para conjuntos de dados específicos. Porém, aqui estamos tratando do rastreamento das análises já realizadas com certo conjunto de dados ainda não depositado em um repositório público, bem como da retroalimentação do fluxo da análise. Estamos tratando de interfaces que podem ser usadas com as plataformas computacionais de análise dos dados.

As plataformas computacionais ArrayExpress, BASE, SpotFire® e TM4 resolveram esse problema permitindo a visualização da árvore de análise de um conjunto de dados, como mostrado na seção que trata da comparação das plataformas BASE e R (seção 2.3.1). Ou seja, são plataformas que misturam o conceito de ambiente de condução da análise com interface de visualização dos resultados. É importante ressaltar que por se tratar de uma árvore, a retroalimentação da análise fica comprometida. Já no ambiente R, não há uma interface assim de pronto, apesar de haver iniciativas nesse sentido (Koide *et al.*, 2006). No transcorrer de uma análise no R, tem-se como resultado objetos que podem ser encadeados como o desenvolvedor decidir. Seja qual for a opção pela plataforma de análise, o problema de qual índice será usado para indexar os resultados persiste. O ideal é que o índice a ser escolhido seja o mais informativo para as pessoas que vão analisar os resultados obtidos.

Outro aspecto importante das interfaces de disponibilização dos resultados é o de permitir que usuários a distância do centro de análise do dado tenham acesso aos resultados. Nesse sentido interfaces via *Web* foram selecionadas por algumas plataformas, como o BASE. Nós propomos a adaptação do moodle (<http://moodle.org>), que é um software livre desenvolvido para o apoio à aprendizagem.

2.1.6. Rastreabilidade e reprodutibilidade

A necessidade de rastreabilidade da análise faz-se necessária devido ao grande número de variáveis e etapas envolvidas em um projeto de análise de dados de

microarray. Assim, para que se possa ter confiança nos resultados obtidos e se possa reproduzir a análise dos dados, o rastreamento de cada etapa realizada é necessário. As plataformas ArrayExpress, BASE, TM4 e Spotfire® (Saal *et al.*, 2002; Saeed *et al.*, 2003; Kaushal & Naeve, 2004) fazem a persistência dos resultados de forma automatizada. Já a plataforma R acoplada com pacotes não o faz, sendo necessário dar comandos específicos para que os objetos sejam salvos. Isso traz a flexibilidade de se registrar o que se quer, porém pode comprometer também a análise caso algum objeto não seja salvo. A utilização do R acoplado com o editor emacs e o pacote ESS (Rossini *et al.*, 2004) permite que se guarde o registro dos comandos digitados, exatamente como foram digitados, porém isso está longe de ser um registro como o apresentados por interfaces como o BASE (Saal *et al.*, 2002), uma vez que não faz uso de índices nem estruturas para organizar os dados.

2.1.7. Anotação dos dados biológicos

Atualmente existem diversas iniciativas de anotação de dados de genômica: Gene do NCBI, KEGG, Gene Ontology e UniProt (Consortium, 2000; Consortium, 2008; Kanehisa *et al.*, 2008; Sayers *et al.*, 2009). Porém é importante notar que elas não são necessariamente concordantes. Um exemplo interessante a ser notado é o campo de sinônimos presente no banco Gene do NCBI. Nele podemos ver os nomes alternativos, usados em outras bases de dados, para um mesmo gene; a citar a hemoglobina beta, que possui HBB como símbolo para o gene, e CD113t-C e beta-globin como sinônimos. Outro fator complicador na questão de anotação dos dados é o fato de que as anotações evoluem ao longo do tempo: ou porque foram encontrados erros de anotação ou pelo fato de que uma nova função ou similaridade ter sido encontrada para uma dada sequência gênica. Com isso, em projetos de larga escala como análise de dados de expressão, é importante saber com que versão da anotação se está trabalhando.

Essa questão da evolução da anotação da plataforma de microarranjos pode ser um entrave para as plataformas computacionais de condução da análise, pois se essa informação não for modelada no sistema, haverá o problema de que duas anotações para o dado não poderão coexistir.

As divergências quanto à padronização na anotação dos dados fica ainda mais evidente quando passamos para plataformas de vias metabólicas e redes de regulação,

como as catalogadas nas bases de dados KEGG, STKE e Reactome (Gough, 2002; Kanehisa *et al.*, 2008; Matthews *et al.*, 2009).

É importante notar que a frequência com que a anotação de um conjunto de dados será refeita deve ser planejada, tendo como medida o impacto que acarretará, uma vez que uma nova anotação dos dados pode levar a novas interpretações de resultados encontrados.

O projeto Bioconductor (Gentleman *et al.*, 2004) disponibiliza diversos pacotes que pode ser usados na anotação de dados de microarranjos de DNA: AnnBuilder, annotate, BioBase, annotationTools e AnnotationDbi. O pacote annotationTools (Kuhn *et al.*, 2008), por exemplo, disponibiliza a anotação da plataforma comercial Affymetrix. Já o pacote AnnBuilder (Zhang *et al.*, 2003) permite que se recupere a anotação de algumas bases de dados do NCBI, como recursos da base GoldenPath (<http://genome.ucsc.edu/goldenPath>), GeneOntology (Consortium, 2000) e KEGG (Kanehisa *et al.*, 2008). O pacote AnnotationDbi permite acesso a bancos de dados de anotação implementados localmente, ao invés de bases disponíveis via *Web*.

2.2. Métodos

2.2.1. Instalação do BASE e implementação de *plugin*

A plataforma BASE (Saal *et al.*, 2002), versão 1.2.16, foi instalada duas vezes, usando alternadamente os sistemas gerenciadores de banco de dados PostgreSQL e MySQL, em uma máquina Pentium 3 com 500MB de RAM. A interface *Web* do BASE foi desenvolvida em PHP. As interfaces que usamos para realizar perguntas (*queries*) aos bancos foram PhpMyAdmin e DBDesigner. De forma a simplificar a explicar o funcionamento do BASE, poderíamos considerá-lo como sendo constituído de dois “módulos”. O primeiro, módulo de dados brutos, contém as informações referentes às amostras, extratos, hibridizações, *reporters* e *design* do *array*. O segundo, módulo de análise, contém informações referentes às análises feitas sobre os dados brutos. Quando consideramos a implementação do BASE usando MySQL® temos que o primeiro módulo é representado pelo banco base, já o segundo é representado pelo banco baseDynamic. Ou seja, o BASE, quando instalado usando o sistema gerenciador de banco de dados MySQL® é na verdade dois bancos de dados “conversando” entre si. Uma das peculiaridades da instalação do BASE usando o sistema gerenciador de banco

de dados PostgreSQL é que os dois “módulos” (dados brutos e análise) passam a ser representados em um só banco.

A plataforma BASE fornece o armazenamento íntegro, o rastreamento das análises e uma interface de visualização. As diversas análises estatísticas a serem realizadas nos dados são viabilizadas sob a forma de adição de um *plugin* à plataforma. As linguagens em que os *plugins* são comumente escritos são C++, Java, Perl e R. Nós implementamos um *plugin* para o *Significance Analysis of Microarrays* – SAM (Tusher *et al.*, 2001) e conjuntamente implementamos uma interface que faz uma meta-análise do BaseFile fornecido pelo BASE ao *plugin*. Várias análises podem ser realizadas com a estratégia *Significance Analysis of Microarrays* (Tusher *et al.*, 2001), uma vez que diferentes tipos de respostas devem ser considerados:

- Uma classe de resposta (*One class*).
- Duas classes de resposta, de estudo com amostras não pareadas (*Two-class unpaired*).
- Duas classes de resposta, de estudo com amostras pareadas (*Two-class paired analysis*).
- Múltiplas classes de resposta (*Multiclass*).
- Dados de sobrevivência (*Survival data*).
- Resposta quantitativa (*Quantitative response*).

De forma geral o teste por *Significance Analysis of Microarrays* - SAM calcula a estatística d_i , em que i refere-se a cada um dos genes (transcritos), utilizando para tanto as intensidades oriundas dos experimentos de *microarray*, sendo que temos x_{ij} , em que $i = 1, 2, \dots, p$ genes (transcritos) e $j = 1, 2, \dots, n$ amostras, e o vetor resposta y_j , em que $j=1, 2, \dots, n$, tal que $r_i = |\bar{\xi}_i - \bar{\leq}_i|$, sendo $\bar{\xi}_i$ a média dos x_i e $\bar{\leq}_i$ a média dos y_i , sendo, conforme a equação 1:

$$d_i = \frac{r_i}{s_i + s_0}; i = 1, 2, \dots, p \quad \text{eq(1)}$$

Assim, temos que r_i é um *score*, s_i é o desvio padrão e s_0 é o *fudge factor*. Os parâmetros r_i e s_i são calculados de forma diferente para cada um dos tipos de análise possíveis: *quantitative response*, *two-class unpaired data*, *censored survival data*,

multiclass response, *paired data* e *one class data*. Para comportar a heterogeneidade das respostas, diferentes programas (*scripts*) foram escritos, sendo que todos chamam a interface referida anteriormente para obter os dados que são relevantes para tal *plugin*.

2.2.2. Utilização do R + pacotes

A plataforma R (Ihaka & Gentleman, 1996) foi instalada usando a versão 2.4.1 em uma máquina Linux, com os pacotes do CRAN e Bioconductor (Dudoit *et al.*, 2003) necessários para as análises. Em muitas das análises usamos o pacote MaigesPack (Esteves, 2007), disponibilizado pelo Bioconductor e desenvolvido pelo Dr. Gustavo Esteves. A diversidade de análises que esse pacote possui está detalhada na seção da plataforma de *microarray* ACCamargo3 (3.2.4). Os objetos resultantes de cada uma das etapas das análises foram salvos em um repositório compartilhado pelos desenvolvedores, assim como os comandos dados.

2.2.3. Implementação do Moodle

O Moodle (<http://moodle.org/>) foi instalado e configurado em um servidor Linux. O moodle é um *software* livre, tradicionalmente usado para apoio à aprendizagem. Ele possui uma interface *Web* com controle de níveis de acesso aos usuários. Ele foi incorporado aos nossos projetos como interface de disponibilização dos resultados. Para isso foram criadas as áreas para conjuntos de dados, que no moodle são representados por cursos. Então, dentro de cada conjunto de dados criamos diretórios, como os de resultados, dados brutos e bibliografia.

2.2.4. Fundamentos de XP

A abordagem de engenharia de *software* para desenvolvimento e gerenciamento de projetos chamada eXtreme Programming (XP) foi concebida por Kent Beck, culminando em um livro chamado “*Extreme Programming Explained*”, publicado em outubro de 1999 (Beck, 2000). É nessa primeira edição do livro que os fundamentos aqui apresentados foram baseados. Ela é uma das metodologias de desenvolvimento ágil de *software* e se encaixa bem no contexto de equipes de desenvolvedores interagindo

com clientes, a exemplo de fábricas de *software*. Outras questões sobre XP e seus fundamentos também são discutidos nos livros “*Questioning Extreme Programming*” de Pete McBreen (McBreen, 2003) e no livro “*Planning eXtreme Programming*” de Kent Beck e Martin Fowler (Beck & Fowler, 2001). Sua utilização neste projeto de doutorado faz-se pertinente dada a colaboração que temos efetuado com grupos de pesquisa nos diversos conjuntos de dados de expressão, como os apresentados aqui: Sarcoma, Carcinomas Epidermóides e Cabeça e Pescoço. Outros casos de utilização de XP em projetos diversos, assim como maiores detalhes sobre os fundamentos da técnica, podem ser vistos em (Silva, 2005; Franco, 2007; Sato, 2007; Bassi Filho, 2008).

XP possui quatro valores fundamentais: *Comunicação*, *Simplicidade*, *Retro-alimentação* e *Coragem*. A comunicação deve ser empregada para com o colaborador e internamente, dentro da equipe. A simplicidade deve ser exercida na escrita de código, sendo que o mesmo deve ser escrito de forma funcional e simples (fácil de ler, intuitivo, de forma que não esconda armadilhas). A retro-alimentação deve ser aplicada na constante avaliação das metas, pois “insucessos” devem ser incorporados no cronograma de trabalho. A coragem deve ser constantemente aplicada no projeto: coragem de se refatorar o código, sempre que necessário e de comunicar um problema, sempre que ele for apontado, o mais cedo possível.

XP também propoe princípios metodológicos fundamentais tais como: retro-alimentação rápida, assumir simplicidade, mudanças incrementais, abraçar as mudanças e trabalho de qualidade. A retro-alimentação rápida implica que a incorporação de mudanças no projeto, assim como a correção de erros, deve ser feita o mais rápido o possível. A postura de assumir simplicidade deve sempre ser empregada, ou seja, o código e/ou a funcionalidade deve ser desenvolvido da forma mais simples possível, para que quando a complexidade no sistema seja realmente necessária, haja tempo para se focar nela. Esse costuma ser um ponto difícil para os programadores, dado o conceito de planejamento e estruturação de um projeto. Já quanto às mudanças, tem-se que elas devem ser incrementais e pequenas, tanto no *design*, quanto no planejamento e na equipe. Além disso, as mudanças devem sempre ser esperadas e bem vindas. O princípio fundamental referente ao trabalho de qualidade parte da premissa de que ninguém se sente bem fazendo um trabalho de baixa qualidade. Logo, além da qualidade do trabalho não ser realmente uma variável, uma vez que a qualidade do sistema depende necessariamente dela, tem-se ainda que um trabalho de qualidade vá favorecer a dinâmica positiva da equipe e uma boa evolução do projeto.

Assim, os valores fundamentais aliados aos princípios metodológicos de XP permitem que se tenha um conjunto de posturas que deveriam ser aplicadas para bom andamento de um projeto. A Tabela 1 lista as doze práticas fundamentais que XP sugere sejam aplicadas em um projeto.

O jogo do planejamento é o que permite que se monte o cronograma do projeto e faz-se essencial para o bom andamento de XP (Beck, 2000; Beck & Fowler, 2001). Ele inclui ordenar as “histórias do cliente” por prioridades determinadas pelo próprio. Então, cada história do cliente é quebrada em tarefas, para as quais o tempo de implementação são estimados. Para cada “história do cliente” é criado um cartão, assim como para cada tarefa: cartão de história e cartão de tarefa, respectivamente. Também podem ser criados cartões cuja história não possui um cunho investigativo, mas sim técnico, como por exemplo: alterar a configuração da interface do repositório comum de dados ou implementar um algoritmo mais eficiente para o cálculo de um teste. Para cada cartão então a equipe estima quantas horas de trabalho serão necessárias para seu desenvolvimento. Então, dada uma semana de trabalho de 40 horas multiplicada pelo número de desenvolvedores, pode-se determinar do que consistirá a próxima iteração. É interessante notar que os cálculos de horas de trabalho que serão consideradas para o cálculo do cronograma são, na verdade, as horas ideais de programação que um desenvolvedor consegue disponibilizar, pois XP considera que um desenvolvedor não passa 100% do seu tempo programando, mas sim um percentual do seu tempo contribuindo com a equipe (tirando dúvidas dos outros, discutindo estratégias).

Tabela 1. Doze práticas fundamentais de XP

Prática	Detalhes
O jogo de planejamento	Define rapidamente o escopo do próximo <i>release</i> , pois define quais serão as próximas funcionalidades a serem implementadas levando em consideração a estimativa técnica de implementação (horas de trabalho por exemplo).
Pequenos releases	Colocar um sistema simples para funcionar o mais rápido possível. Então, novas funcionalidades serão agregadas, sempre na forma de pequenos releases em um curto período de tempo.
A metáfora	História simples do sistema que guia toda a equipe.
<i>Design</i> simples	O sistema deve ser o mais simples possível.
Testes	Devem existir sempre (de preferência de forma automatizada) e permitem a avaliação do sistema por completo.
Refatoração	O sistema deve ser reestruturado para se remover duplicações, simplificar, melhorar a comunicação e adicionar flexibilidade.
Programação pareada	Toda escrita de código deve ser feita por duas pessoas em um mesmo computador simultaneamente.
Propriedade coletiva	Qualquer um pode mudar qualquer código a qualquer momento.
Integração contínua	Toda vez que uma tarefa é cumprida, ela deve ser integrada ao sistema (testada individualmente e também o sistema como um todo e/ou a parte que essa tarefa poderia afetar).
Semana de trabalho de 40 horas	Sustentabilidade no desenvolvimento.
Cliente sempre presente	Para que se possa tirar dúvidas a qualquer momento.
Seguir os padrões de código	Os critérios de qualidade de <i>software</i> devem ser cumpridos, sempre.

Pode-se dizer que XP implementa a filosofia: “A fraqueza de um é coberta pela força de outros”. Essa postura fica em evidência com o exercício da comunicação, integração entre a equipe, uma metáfora de sistema, testes e programação pareada. Isso, de certa forma traz robustez ao andamento do projeto, pois a ausência de um membro da equipe passa a não prejudicar de forma crítica o desenvolvimento do projeto.

XP também descreve que como estratégia a ser incorporada em um projeto a etapa de planejamento, a etapa de *design*, a escrita de código e os testes. A forma como as 12 práticas se integram às estratégias não será detalhado nesta tese, porém maiores detalhes podem ser obtidos em (Beck, 2000; Beck & Fowler, 2001; McBreen, 2003; Silva, 2005; Franco, 2007; Sato, 2007; Bassi Filho, 2008).

Em uma equipe “típica” de XP há divisão de papéis e responsabilidades por toda a equipe. Os possíveis papéis estão representados na tabela 2.

Tabela 2. Papéis em uma equipe de XP

Papel	Função
Programador	Geralmente são vários; escrevem o código e testes do sistema, sempre primando por um código de alta qualidade
Cliente	Escreve as histórias e prioriza os cartões. Sua participação e envolvimento são cruciais para o sucesso do projeto, uma vez que é ele quem faz os testes funcionais e decide se uma funcionalidade foi satisfeita. É deste ponto que surge o “provérbio” de XP: Cliente sempre presente.
Testador	Ajuda o cliente a escrever os testes funcionais e executa-os, sendo que os resultados são passados para toda a equipe.
Monitor (em inglês, Tracker)	Possui um papel crucial na manutenção do prazo previsto para o próximo release, uma vez que compara o cumprimento dos cartões com as estimativas feitas para os mesmos. Caso ocorram discrepâncias, a equipe toda é informada.
Treinador (em inglês, Coach)	Deve ser uma pessoa com profundos conhecimentos e vivência em XP, pois é o responsável pelo projeto como um todo. Ou seja, possui um papel importantíssimo dentro da equipe e no sucesso do projeto.
Consultor	É um especialista que é consultado quanto a questões pontuais. Ele não fica presente como o Cliente.
Chefe	É responsável pelas tomadas de decisões no projeto como um todo.

2.2.5. Anotação dos dados

A anotação da plataforma de *microarray* usada nos projetos de rim e próstata já estava pronta e foi inserida no BASE (Saal *et al.*, 2002) para que pudesse ser usada nas análises. Colunas adicionais tiveram que ser inseridas ao banco para conseguirmos colocar a anotação de se as seqüências eram ou não codificantes. Porém, tal plataforma não comportava a questão de versões de anotação dos dados.

A anotação dos dados de Sarcoma e Carcinomas Epidermóides foi feita usando a ferramenta AMPiR (ver seção 2.3.4) uma vez que não tínhamos a anotação pronta por se tratar de uma plataforma não comercial. Para ela foram recuperadas as informações

de gi das bases de dados Nucleotide, Gene e UniGene. Também foram recuperadas as informações de geneSymbol das bases Gene e UniGene, bem como os outros nomes disponíveis para um mesmo gene. Uma comparação entre a qualidade de anotação usando geneSymbol ou usando também os outros nomes disponíveis (“*also known as*”) também será apresentada adiante na seção que descreve a ferramenta AMPiR.

2.2.6. Rastreamento das análises

O rastreamento das análises foi feito por meio de diferentes estratégias: usando a documentação que os cartões de XP fornecem; usando o repositório compartilhado onde os objetos são salvos; usando o código desenvolvido para as análises também depositado no repositório compartilhado e usando o registro dos resultados organizados pelas iterações de XP conforme a interface de disponibilização dos dados e resultados.

2.3. Resultados

2.3.1. Comparação das plataformas de condução de análise: BASE x R

2.3.1.1. Plugin SAM e interface

Como exemplo de resultado de análise usando a plataforma BASE (Saal *et al.*, 2002) e o *plugin* SAM (Tusher *et al.*, 2001) apresentamos a Figura 1, com as informações de tempo de execução, arquivos intermediários gerados e arquivos de resultados.

Figura 1: Visualização do resultado de uma análise por SAM *two-class unpaired* realizada via BASE.

Como mostra a Figura 1, o BASE armazena todos os arquivos e objetos que saem como resultado do *plugin*. Caso o *plugin* também retorne um BaseFile (arquivo de saída do BASE), esse arquivo é automaticamente lido pelo BASE e seu resultado, por exemplo uma lista de genes, será disponibilizado pela interface como um subconjunto.

A interface que desenvolvemos em Perl, por fazer uma meta-análise do BaseFile e extrair os campos necessários do mesmo conforme a chamada do *plugin*, permite que um grande número de pacotes e funcionalidades possam ser acoplados ao BASE, por exemplo *plugins* desenvolvidos em R.

2.3.1.2. Persistência de Objetos: R + pacotes

O ambiente R (Ihaka & Gentleman, 1996) não armazena de forma automatizada e estruturada os objetos gerados no transcorrer de uma análise. Com isso a visualização da árvore de análise, e respectivos objetos, fica dificultada. Um exemplo de árvore de

análise obtida com o BASE pode ser visto na Figura 2. Consideramos que além de se ter os objetos gerados, é preciso também ter os parâmetros usados para chegar a tal objeto armazenados. Porém, uma vez que o armazenamento do código e dos objetos é necessário, surge o problema de criação de uma estratégia de organização do repositório. Diversas poderiam ter sido as estratégias: organizar por data de desenvolvimento, organizar ordem de criação, organizar por pergunta biológica, etc.

Com o intuito de valorizar a pergunta biológica envolvida em cada análise e reforçar o conceito de iterações cercado um tema de análise, utilizamos a estratégia de armazenarmos em um repositório compartilhado todos os objetos gerados em uma análise, bem como o código em escrito R, todos organizados pelas iterações e respectivos cartões criados usando XP. Logo, as iterações, bem como os cartões de XP mostraram-se válidos como índices para organizar repositórios de análises em que se usa a plataforma de análise R acoplada com pacotes.

Hierarchical overview of BioAssaySet analyses			
Name	Date	Info	Functions
♦ todasAmostras	2005-08-22	Mean FG - Mean BG, 54 assays, 497664 spots, 3752 reporters	[Delete]
└─ selecionapMarkersKnownGenesNoMatch [F]	2005-09-05	1 spot criterion	[Delete] [Copy]
└─ └─ sohPMarkersKnownGenesNoMatch	2005-09-05	54 assays, 412128 spots, 3596 reporters	[Delete]
└─ └─ └─ Transformation: Minimum intensity [T]	2005-09-05	Transformation: Minimum intensity	[Delete] [Copy]
└─ └─ └─ └─ sohPMarkersKnownGenesNoMatch transf.	2005-09-05	54 assays, 412128 spots, 3596 reporters	[Delete]
└─ └─ └─ └─ └─ Normalization: Lowess [T]	2005-09-05	Normalization: Lowess	[Delete] [Copy]
└─ └─ └─ └─ └─ └─ sohPMarkersKnownGenesNoMatch transf.2	2005-09-05	54 assays, 412128 spots, 3596 reporters	[Delete]
└─ └─ └─ └─ └─ └─ └─ Analysis: ADG2 ver 2 [J]	2005-09-06	Analysis: ADG2 ver 2	[Delete] [Copy]
└─ └─ └─ └─ └─ └─ └─ └─ resultsADG2_normal_tumoral [F]	2005-09-06	1 spot criterion	[Delete] [Copy]
└─ └─ └─ └─ └─ └─ └─ └─ └─ sohResultsADG2_v0	2005-09-06	54 assays, 43956 spots, 396 reporters	[Delete]
└─ └─ └─ └─ └─ └─ └─ └─ └─ └─ Analysis: Hierarchical clustering [J]	2005-09-06	Analysis: Hierarchical clustering	[Delete] [Copy]
└─ └─ └─ └─ └─ └─ └─ └─ └─ └─ └─ Hypothesis Testing: B-statistics (Bioconductor) [J]	2005-09-08	Hypothesis Testing: B-statistics (Bioconductor)	[Delete] [Copy]
└─ └─ └─ └─ └─ └─ └─ └─ └─ └─ └─ └─ Analysis: MDS [J]	2005-09-08	Analysis: MDS	[Delete] [Copy]
└─ └─ └─ └─ └─ └─ └─ └─ └─ └─ └─ └─ Analysis: PCA [J]	2005-09-08	Analysis: PCA	[Delete] [Copy]
└─ └─ └─ └─ └─ └─ └─ └─ └─ └─ └─ └─ └─ Analysis: ADG2 p-value=0.01, gleason 5 e 6 x 9 e 10 [J]	2005-09-08	Analysis: ADG2 ver 2	[Delete] [Copy]
└─ └─ └─ └─ └─ └─ └─ └─ └─ └─ └─ └─ └─ └─ Analysis: ADG2 p-value=0.05, gleason 5 e 6 x 9 e 10 [J]	2005-09-08	Analysis: ADG2 ver 2	[Delete] [Copy]
└─ └─ └─ └─ └─ └─ └─ └─ └─ └─ └─ └─ └─ └─ └─ Transformation: Weak Spot Filter [J]	2005-09-08	Transformation: Weak Spot Filter	[Delete] [Copy]

Figura 2: Exemplo de árvore de análise da plataforma BASE.

2.3.1.3. Discussão

Enquanto no BASE (Saal *et al.*, 2002) as análises são criadas através da interação com a interface por meio do mouse, no R as análises são criadas por meio do desenvolvimento de código: pequenos programas. Essa diferença acarreta algumas desvantagens: Como primeiro ponto, podemos explicitar o fato de que somente os parâmetros que foram selecionados para serem registrados serão armazenados por sistemas como o BASE, o que é uma restrição na flexibilidade da análise e/ou no rastreamento da mesma. Um exemplo em que essa flexibilidade é importante é o de salvar a semente usada em algoritmos aleatórios, para que se possa reproduzir o resultado encontrado. Se esse critério não fazia parte do que é retornado à plataforma BASE, isso só poderá ser feito se o *plugin* for modificado, o que exigirá conhecimentos de programação. Ou seja, a plataforma BASE torna-se interessante para usuários que não possuem grande desenvoltura de programação, pois com um clique no *mouse* podem fazer análises padrão, sem entrar a fundo nos métodos usados. Outra forma de ver essa desvantagem é a de que somente uma pessoa que está familiarizada com as variáveis biológicas do conjunto de *microarray*, que sabe programar em R e que entende os métodos estatísticos implementados nos diversos pacotes do R terá condições de implementar análises para os dados usando a plataforma R aliada a pacotes.

Como segundo ponto, temos a questão de que o armazenamento explícito dos objetos e do código desenvolvido faz-se necessário para um rastreamento adequado quando é feita a eleição pela plataforma R + pacotes para análise de dados de *microarray*. Nesse ponto é necessário um adequado planejamento de como será o repositório dos objetos e o repositório do código escrito, sendo necessária a utilização de índices compreensíveis à equipe que acessa o repositório, pois o volume de etapas e análises feitas sobre conjuntos de dados de *microarray* é muito grande.

Entendemos que a utilização de eXtreme Programming (XP), para indexar e organizar o repositório de objetos e código das análises mostrou-se adequada, pois permitiu uma rápida localização do ponto a ser rastreado e recuperado.

2.3.2. Um modelo para gerenciar análises de *microarrays* usando XP

2.3.2.1. O modelo

É comum projetos de bioinformática e de genômica, por implementarem novas estratégias computacionais para superar os desafios que a genômica e a bioinformática possuem, acabarem por ter como resultado a análise dos dados bem como a implementação de uma ferramenta. A utilização de XP para planejar e gerenciar o desenvolvimento de *software* foi discutida anteriormente (Beck, 2000; Beck & Fowler, 2001; McBreen, 2003; Silva, 2005; Franco, 2007; Sato, 2007; Bassi Filho, 2008). Frente ao enorme volume de dados que os projetos de larga escala de *microarray* trazem, surge a necessidade de se planejar e gerenciar uma análise de *microarray* de forma mais organizada. Assim, propomos a aplicação de XP no planejamento e gerenciamento de projeto desta natureza.

Conforme explicado anteriormente (seção 2.2.4) vários valores, práticas, papéis e responsabilidades podem ser colocados em prática em um projeto que adota XP como metodologia de desenvolvimento de *software*. Aqui, propomos usar tais conceitos não só para desenvolver código, como se faz necessário para a plataforma R acoplada com pacotes, mas também para planejar e gerenciar um projeto de análise de expressão, organizar e direcionar as hipóteses biológicas.

No modelo proposto entendemos que há duas principais equipes de especialistas: especialistas da área de biológicas/biologia molecular e especialistas da área de computação e estatística. Entendemos também que há profissionais da área de bioinformática envolvidos no projeto, que devem dominar as duas áreas e com isso assumir o papel de planejadores e gerenciadores. A Tabela 3 mostra nossa proposta de alocação dos profissionais frente aos papéis e responsabilidades.

Tabela 3. Alocação dos profissionais conforme papel, responsabilidades e área de especialidade.

Papel	Responsabilidades	Área do profissional
Monitor (Tracker em inglês)	Organiza as reuniões, gráficos de acompanhamento, gerencia reuniões.	Bioinformática ou computação
Treinador (Coach em inglês)	Especialista em XP. Responsável pela equipe e sucesso de XP na mesma.	Bioinformática ou computação
Desenvolvedor	Escreve código, scripts. Especialista em computação.	Computação ou Bionformática
Colaborador	Especialista no problema biológico. É quem interpreta os resultados da análise e surge com novas idéias de análises.	Biológicas ou Bioinformática
Líder dos colaboradores	Especialista em biológicas com forte conhecimento do problema. Responsável pelo projeto na área de biológicas.	Biológicas ou Bioinformática
Líder dos desenvolvedores	pelo projeto/sistema na área de exatas (computação, estatística).	Bioinformática, Computação, Estatística ou áreas afins

Entendemos que todos os quatro valores básicos de XP são essenciais para que a metodologia funcione, logo propomos que todos sejam implementados: comunicação, simplicidade, retro-alimentação e coragem. E consideramos que os cinco princípios básicos (retro-alimentação rápida, assumir simplicidade, mudanças incrementais, abraçar as mudanças e trabalho de qualidade) também são essenciais para que XP funcione associado a um projeto de análise de dados de *microarray*.

Consideramos também que as doze práticas devem ser colocadas em prática, sempre que possível, ainda que talvez algumas sejam mais difíceis de funcionar (Beck, 2000; McBreen, 2003; Cusumano, 2007; Bassi Filho, 2008). Michael A. Cusumano, por exemplo, ao comparar XP com o método de desenvolvimento iterativo *Microsoft-style*, sugere que as práticas de *design* simples e propriedade coletiva de código não poderiam ser implementadas em grandes projetos da Microsoft® (Cusumano, 2007).

Um exemplo de como um cartão deveria ser criado neste modelo está mostrado na Figura 3. Nele fica explícito o foco na pergunta biológica, sendo ela colocada como a “história”, exatamente como exposta pelo colaborador.

Figura 3: Exemplo de cartão conforme modelo proposto.

O modelo que propomos (Figura 4) focaliza as questões biológicas, colocando-as em evidência.

Figura 4: Modelo proposto de planejamento e gerenciamento de análises de dados de microarranjos de DNA.

A aplicação do modelo na análise de dados de microarranjos de DNA faz-se pela criação do meta-objeto, que é a representação conceitual/cognitiva do conjunto de dados (Figura 5). O meta-objeto deve ser criado de forma que todos os componentes da equipe entendam o dado e suas reentrâncias: como as amostras e suas variáveis clínicas se relacionam; quais são os genes e as diversas anotações e como eles se relacionam entre si, em vias e grupos; e ainda qual é o estimador para a expressão gênica e como ele deve ser usado. O meta-objeto faz-se fundamental por toda a análise, pois ele que evitará erros de implementação e compreensão da análise e resultados.

Figura 5: Meta-objeto. O meta-objeto é a representação conceitual/cognitiva do projeto como um todo, de como as amostras e suas variáveis clínicas se relacionam; de quais são os genes e as diversas anotações e de como eles se relacionam entre si, em vias e grupos; e ainda de qual é o estimador para a expressão gênica e de como ele deve ser usado.

Após a fase de criação do meta-objeto tem-se a fase zero, em que a carga inicial dos dados é feita no ambiente computacional de escolha. Nesta mesma etapa tem-se a anotação dos dados, isto é, a anotação da plataforma de microarranjos de DNA. Nessa mesma iteração devem ser feitos testes de qualidade nos dados, para verificar se não houve erros na carga e se o dado possui qualidade mínima para ser analisado.

No início de cada uma das iterações subseqüentes novas perguntas biológicas passam a ser consideradas para a análise, sendo que elas são então priorizadas e cada uma tem seu tempo de implementação e execução estimados. Faz-se então o jogo de planejamento, que consiste em definir quando deverá ser a próxima entrega de resultados e quais cartões, isto é, perguntas biológicas, serão investigados. Há três possibilidades para as perguntas biológicas em análise: ou elas continuam sob investigação, ou não levaram a resultados relevantes ou os resultados podem ser considerados tão relevantes que serão validados experimentalmente.

É interessante notar que geralmente as primeiras perguntas a serem investigadas são as que instigaram a construção do experimento, mas que em iterações subseqüentes são aprofundadas ou mesmo incorporadas a outras perguntas biológicas.

2.3.2.2. Escopo do modelo e discussões

Entendemos que o escopo de aplicação do modelo são projetos de análise de dados de expressão, para os quais há diversas perguntas biológicas a serem investigadas. Entendemos também que faz parte desse escopo a existência de uma equipe multidisciplinar em que deve haver um profissional que transite bem entre o problema biológico e a parte computacional, profissional esse que consideramos ser idealmente da área de Bioinformática. Entendemos também que caberá a esse profissional fazer o papel de Coach e de Tracker, devido a sua capacidade crítica das áreas. Entendemos também que o modelo comporta desenvolvimento simultâneo às análises de software e ou funcionalidades, sendo para esses criados cartões como no modelo tradicional de XP, devendo eles ter seus tempos de implementação estimados e serem priorizados também pelo colaborador no jogo de planejamento.

Nossa avaliação é que a proposta do modelo faz-se necessária dado o grande volume de dados em análises de *microarray*. O modelo contribui ao sugerir como os profissionais de diversas áreas devem se posicionar nesse contexto e podem ser alocados para diferentes tarefas. Sugere também como as iterações devem ser dirigidas e como as perguntas biológicas devem ser colocadas de forma a serem valorizadas ao máximo e exauridas conforme a prioridade definida, que geralmente muda ao longo dos projetos. O modelo também comporta a questão híbrida de projetos de genômica de ter desenvolvimento de novas ferramentas computacionais simultaneamente ao desenvolvimento da análise.

O modelo ainda mostrou-se válido na utilização dos cartões como índices para os ambientes de condução de análise e interfaces de disponibilização dos dados. E trouxe ainda como vantagem, conforme discutido no capítulo de análise de dados, a vantagem de colaboradores de diferentes projetos conversarem de forma mais objetiva sobre seus projetos e com isso trocarem conhecimento e informações sobre possíveis análises e métodos.

2.3.3. A ferramenta AMPiR

2.3.3.1. Introdução

A ferramenta AMPiR (*Annotate Microarray Platform in R*) surgiu da necessidade de termos uma anotação confiável para as sequências (*probes*) representadas na plataforma de *microarray* ACCamargo3 (seção 3.2.4) e que pudesse ser cruzada à anotação de grupos gênicos para que pudéssemos estudar os dados de expressão também sob as óticas de grupos e vias (Butte *et al.*, 2000; Segal *et al.*, 2004; Esteves, 2007). A confiabilidade desejada baseia-se na priorização de relações entre bases de dados. A plataforma de *microarray* ACCamargo3, por ser uma plataforma não comercial, não possui de pronto uma anotação associada. Junto a isso havia o desejo de podermos atualizar a anotação da plataforma no momento que quiséssemos, uma vez que a anotação dos genes evolui temporalmente, como já discutido.

Várias são as bases de dados que possuem informações sobre genes, a citar NCBI (Gene, Unigene, RefSeq, Nucleotídeos, Proteínas), KEGG, BioCarta, STKE, GeneOntology, Reactome (Consortium, 2000; Kanehisa *et al.*, 2008; Matthews *et al.*, 2009). Porém, a transferência de nomenclaturas e anotação entre as mesmas não ocorre de forma simples, como apresentado na seção de resultados para o mapeamento de vias para a plataforma de *microarray* (seção 2.3.4). Logo, optamos por fazer o mapeamento das *probes* usando a base de dados Nucleotide do NCBI e então recuperar a informação de GeneOntology e KEGG pela recuperação da anotação dessa *probe* na base Gene do NCBI. Tal caminho entre as bases faz-se possível graças aos *links* entre as bases, que, na verdade, são relacionamentos entre as chaves primárias de cada banco, conforme detalhado a seguir, na descrição das E-utilities (Maglott *et al.*, 2007).

As E-utilities são interfaces que permitem ao usuário do banco de dados do NCBI (Maglott *et al.*, 2007; Sayers *et al.*, 2009) acessar seu conteúdo. Para tanto, foi desenvolvida a abordagem PowerScripting, que através da manipulação de URLs e variáveis de ambiente, armazenadas no servidor do NCBI, pode-se com uma linguagem de programação, por exemplo Perl ou Python, recuperar as entradas em formatos específicos e considerando *pipelines* de *queries*. O banco de dados que pode ser acessado pela interface do NCBI, é de fato composto de vários bancos de dados, como, por exemplo, o banco de sequências, o banco de dados de genes e o banco de dados de expressão chamado GEO (Sayers *et al.*, 2009), como mostra a Figura 6. Apesar de ser composto por vários bancos separados, seus bancos de dados estão relacionados, em

alguns casos por relacionamentos curados manualmente (exemplo: *link* direto entre seqüência de nucleotídeo e entidade gene presente na página web do NCBI) ou por relacionamentos (*links*), feitos de forma automatizada.

Figura 6: Relacionamentos entre os bancos de dados presentes no NCBI.

Extraído de <http://www.ncbi.nlm.nih.gov/Database/index.html>

A entidade gene, do banco de dados do NCBI, é muito importante para o mapeamento de vias e grupos de genes no trabalho que estamos desenvolvendo, pois tal entidade centraliza a anotação de genes, disponibilizando a informação de Gene Ontology, KEGG e interações entre o gene em questão e outros genes (Consortium, 2008; Kanehisa *et al.*, 2008).

2.3.3.2. Especificação e Resultados: A anotação da plataforma ACCamargo3

A plataforma ACCamargo3, utilizada na obtenção dos dados de *microarray* para os conjuntos de dados de sarcoma, carcinoma epidermóide e cabeça e pescoço, foi

anotada usando a ferramenta AMPiR, que desenvolvemos. Ela foi originalmente implementada em Perl e está sendo transferida para o R.

Cada lâmina possui 4800 spots, referentes a 4574 *probes* (sequências) únicas, sendo 8 de controles. Foram encontradas no total 3748 símbolos (geneSymbol) oriundos do banco Gene do NCBI para os genes representados na plataforma. Tais símbolos (geneSymbol) podem ser usados como chave para outras bases de dados de genômica, como o KEGG, Gene Ontology, Biocarta e STKE (Consortium, 2000; Gough, 2002; Kanehisa *et al.*, 2008). No total mapeamos 427 grupos funcionais para essa plataforma, que foram obtidos do KEGG, Gene Ontology, Biocarta, STKE e da literatura pertinente.

A abordagem escolhida para mapear especificamente esta plataforma de *microarray* está exemplificada na Figura 7. Para que se possa percorrer os relacionamentos entre os bancos de dados do NCBI, é preciso usar os **gi**, isto é, identificações únicas dentro de cada banco de dados. Uma vez com o **gi** de uma sequência de nucleotídeos, opta-se preferencialmente por usar o relacionamento entre o banco de dados nucleotídeos e o banco de dados de gene, uma vez que tal relacionamento foi criado de forma manual e criteriosa. Porém, como muitas sequências não possuem tal relacionamento, optamos por usar o relacionamento que sequências de nucleotídeos possuem com Unigene como um caminho intermediário para chegar ao banco de dados de gene. Para os casos em que havia mais de uma entrada em Unigene relacionada a uma sequência de nucleotídeos, representando possivelmente isoformas, optamos por não prosseguir com o mapeamento, ficando essa sequência sem anotação. Uma vez no banco de dados Unigene, percorre-se o relacionamento entre o banco Unigene e Gene, sendo, portanto possível recuperar as informações de Gene Ontology e Kegg associadas à entidade Gene.

Figura 7: Esquema dos relacionamentos usados entre os bancos para anotação das seqüências de DNA depositadas na plataforma de *microarray* ACCamargo3.

O esquema apresentado na Figura 7 mostra que o caminho preferencial de anotação é o demarcado pela seta dupla, pois é curado manualmente pelo próprio NCBI. Quando inexistente, opta-se pelo demarcado com seta fina simples, passando pelo Unigene, sendo assim possível chegar à entidade, base de dados, Gene. Como ponto de início para a anotação tínhamos uma lista de nomes de ESTs (*EST name*), sendo que em alguns casos o nome não estava completo. Portanto, começamos a anotação fazendo buscas no banco Nucleotide do NCBI para determinar qual era o nome correto da EST e com isso poder recuperar um **gi** único para essa seqüência. Com o **gi** do banco Nucleotide em mãos, tentamos então recuperar a anotação da base Genes do NCBI. Uma vez que se tem a representação de gene para cada um dos probes, pode-se então recuperar a informação de Gene Ontology e de KEGG.

A associação entre a anotação dos genes obtida com o banco Gene e a anotação de grupos ou vias obtidas com o KEGG, gene Ontology ou outras bases de dados não se faz de forma simples, pois não seguem a mesma padronização. Assim, quando usamos geneSymbol para cruzar com os símbolos disponibilizados pelo KEGG encontramos um certo conjunto identidade. Quando consideramos também os sinônimos dos nomes de genes (campo Other Aliases) para o cruzamento, temos mais resultados positivos. Porém, verificamos que muitos desses resultados positivos eram na verdade falsos positivos, isto é, apesar do nome do gene ser o mesmo, eles não tratavam do mesmo gene (Tabela 4).

Tabela 4. Mapeamento das *probes* da plataforma ACCamargo3 com as vias do KEGG. por geneSymbol e por Sinônimos.

Nome via do KEGG	por GeneSymbol	Por Sinônimos
adrenergic.tgf	1	5
Bcell_antigen_receptor.tgf	0	6
circadian.tgf	0	0
cytokinin.tgf	0	0
EGF_receptor.tgf	2	8
ERK1_ERK2_MAPK.tgf	0	6
estrogen_receptor.tgf	0	4
ethylene_signaling.tgf	1	0
fas_signaling.tgf	2	11
fibroblast_growth_factor_receptor.tgf	2	7
G_alpha_12.tgf	0	4
G_alpha_13.tgf	0	5
G_alpha_i.tgf	1	7
G_alpha_q.tgf	0	3
G_alpha_s.tgf	0	5
G_alpha_s.tgf	0	2
G_alpha_s.tgf	0	1
IL_13.tgf	1	0
IL_4.tgf	3	5
insulin.tgf	0	5
integrin.tgf	1	10
interferon_gamma.tgf	2	3
JAK_STAT.tgf	0	5
jasmonate_biochemical.tgf	0	0
jasmonate_signaling.tgf	1	0
JNK_MAPK.tgf	1	10
mitochondrial_apoptosis_antiapoptotic_BCL2Family.tgf	1	1
mitochondrial_apoptosis_BH3Only_BCL2Family.tgf	2	4
mitochondrial_apoptosis_caspases.tgf	1	7
mitochondrial_apoptosis_multidomain_BCL2Family.tgf	1	4
natural_killer_cell_receptor.tgf	1	6
P13K.tgf	0	4
p38_MAPK.tgf	0	7
PAC1_receptor.tgf	0	3
seven_transmembrane_receptor_beta_arrestin.tgf	0	2
STAT3.tgf	2	2
T_cell_signal_transduction.tgf	0	6
TGF_beta_development.tgf	0	3
TGF_beta.tgf	0	3
TNF.tgf	1	5
toll_like_receptor.tgf	2	8
type1_interferon.tgf	0	2
viaAKT_elier.tgf	4	3
wnt_beta_catenin.tgf	1	6
wnt_cyclic_GMP.tgf	0	1

Os resultados do mapeamento dos grupos e vias para a plataforma de *microarray* ACCamargo3 estão apresentados nas Figuras 8 e 9. Na Figura 8 é possível visualizar o mapeamento de vias do KEGG e na Figura 9 o mapeamento de vias do KEGG e categorias de Gene Ontology. Infelizmente, por serem muitas as vias mapeadas para tal plataforma, no eixo x não é possível visualizar o nome de todas as vias. No eixo y temos o número de elementos mapeados para cada uma das vias. A variabilidade quanto ao número de nós em cada grupo deve-se ao fato de podermos construir um grupo considerando diferentes alturas da árvore de ontologias ou vias, ou seja, do grau de resolução e compartimentalização que elegemos.

Figura 8: Vias do KEGG mapeadas para a plataforma ACCamargo3.

Figura 9: Vias do KEGG e categorias do Gene Ontology mapeadas para a plataforma ACCamargo3.

As informações de localização cromossômica mapeadas para essa mesma plataforma estão apresentadas na Figura 10.

Figura 10: Informação de quantidade de *probes* mapeadas por cromossomo

2.3.3.3. Discussão

O AMPiR visa suprir a necessidade de se anotar uma sequência e/ou uma plataforma de análise de dados de genômica, transcriptômica e /ou proteômica de forma modular, dando-se prioridade a caminhos, conforme exemplificado na seção anterior. A modularidade surgiu da utilização das E-utilities disponíveis no NCBI e sua incorporação como uma ferramenta de anotação.

Como para a plataforma ACCamargo3 tínhamos inicialmente os nomes dos ESTs, sendo que vários deles estavam incompletos, realizamos várias buscas para definir quais eram os verdadeiros nomes das ESTs. Esse problema foi possível de resolver, pois nos casos das dúvidas, as outras ESTs que apareciam como resultado da busca não haviam sido submetidas pelo grupo que gerou as seqüências depositadas na plataforma ACCamargo3. Essa possibilidade de fazer buscas de forma automatizada e flexível usando as E-utilities associadas ao R mostrou o quanto poderosa é essa estratégia. É importante ressaltar que a anotação da plataforma ACCamargo3 com o pacote AnnBuilder (Zhang *et al.*, 2003) não foi possível devido ao fato de termos

inicialmente nomes de ESTs no lugar de **gi** e também ao agravante de vários desses nomes de ESTs estarem incompletos.

Inicialmente a ferramenta AMPiR havia sido implementada em Perl. Atualmente ela está sendo transformada em um pacote escrito em R que será submetido ao Bioconductor. Nessa nova versão a ferramenta apresenta como dependências o R, do pacote XML e uma conexão com o servidor do NCBI. Ela está sendo desenvolvida de forma modular e que continue permitindo a eleição de caminhos prioritários para a anotação. Ela difere do pacote AnnBuilder na medida em que permite buscas para achar **gi**, enquanto o pacote AnnBuilder parte sempre da premissa que o **gi** já existe.

Esta ferramenta usa sempre que possível os campos `query_key` e `WebEnv` da ferramenta E-utilities, otimizando a requisição a cada novo acesso ao servidor dedicado do NCBI para tal queries. Porém o desempenho total da ferramenta depende de conexão com a internet e de certo processamento local no R.

2.3.4. A ferramenta TyPiR

2.3.4.1. Introdução

A ferramenta surgiu da necessidade de se rastrear e gerenciar as análises de dados de *microarray* feitas no ambiente R acoplado a pacotes. Em plataformas de análise computacionais como o BASE, SpotFire® e TM4 (Saal *et al.*, 2002; Dudoit *et al.*, 2003; Saeed *et al.*, 2003; Kaushal & Naeve, 2004; Sarkans *et al.*, 2005; Parkinson *et al.*, 2009) o rastreamento se faz de forma automatizada, sendo que a árvore de análise pode ser visualizada pela interface gráfica. Porém, na plataforma R acoplada a pacotes o mesmo não acontece. O ambiente R permite que se salve o histórico dos últimos comandos realizados, assim como os objetos gerados naquela sessão, porém, esses arquivos são sobrescritos na medida em que novas sessões são criadas. O pacote ESS (Rossini *et al.*, 2004) permite que se armazene os comandos fornecidos na sessão, por atrelar o uso do emacs ao ambiente R, mas a organização desses arquivos depende de cada usuário, o que passa a ser um problema em projetos em que há uma equipe de desenvolvedores. Entendemos que para que o rastreamento seja completo, é preciso armazenar os objetos obtidos com as análises assim como também os comandos dados, um atrelado ao outro e de forma indexada.

Ao considerarmos a organização das análises conforme o modelo para o planejamento e gerenciamento de análises de *microarray* que propomos (ver seção

2.3.2), tem-se ainda a opção de organizarmos os objetos resultados das análises associados aos cartões que os originaram, ou seja, associando os objetos e comandos fornecidos à pergunta biológica que instigou a análise.

2.3.4.2. Especificação e requisitos

A ferramenta está sendo implementada em R e possui como requisitos R (versão 2.4) e sistema operacional Linux (Siever *et al.*, 2005). Ela permite que se armazene objetos associados a comandos, de duas formas: colocando todos em um mesmo repositório ou armazenando-os relacionados aos cartões criados conforme o modelo de planejamento e gerenciamento de análises de *microarray*. Para comportar essas duas estruturas a criação de um projeto cria dois diretórios principais: Objects e Cards, que na Figura 11 aparecem em português como Objetos e Cartões.

Figura 11: Estrutura de diretórios criada pela ferramenta. O diretório Objetos armazena todos os objetos em um mesmo diretório. O Diretório Cartões armazena os objetos indexados pelos cartões criados conforme o modelo de planejamento e gerenciamento de análises de *microarray* proposto nesta tese.

As funções usadas para gerenciar tal estrutura de diretórios e viabilizar o gerenciamento de um projeto estão descritas na Figura 12. Os nomes dos diretórios e das funções estão em inglês, pois é assim que estão sendo implementadas.

Figura 12: Funções usadas para armazenar os objetos de uma análise de dados de *microarray*. No diretório Objects todos os objetos ficam armazenados conjuntamente. No diretório Cards ficam armazenados os objetos indexados pelos cartões criados conforme o modelo de planejamento e gerenciamento de análises de *microarray* proposto nesta tese

As funções criadas visam permitir o armazenamento dos objetos (saveObjects, saveObjectsAndCommand e saveObjectsCard), a listagem dos objetos presentes (listObjects, listObjectsAndCommand e listObjectsCard) e a remoção de objetos presentes (eraseObjects, eraseObjectsAndCommand e eraseObjectsCard). As funções que possuem “AndCommand” em seu nome são usadas quando se quer vincular a informação do comando que foi dado para gerar o objeto ao objeto armazenado pela ferramenta, em oposição as funções que estão vinculadas ao diretório Objects e não possuem esse padrão em seu nome. As funções createCard, eraseCard e listCards são funções que foram criadas para comportar a utilização do modelo proposto nesta tese para o planejamento e gerenciamento de análise de *microarrays*, bem como quando se faz a utilização de XP. Logo, as funções saveObjectCard, moveObjectCard, eraseObjectCard e listObjectCard só são usadas quando se fez opção de armazenar os objetos indexados pelos cartões.

2.3.4.3. Discussão

A ferramenta TyPiR visa resolver o problema de armazenamento de objetos no transcorrer de uma análise de dados de *microarray* quando se usa a plataforma computacional R acoplada a pacotes para a análise dos dados. Ela permite que se armazene objetos conjuntamente com a linha de comando que deu origem ao objeto. A ferramenta também possui a flexibilidade de se simplesmente armazenar objetos sem que a linha de comando que deu origem a cada um deles seja vinculada ao objeto.

Caso um projeto adote o modelo proposto nesta tese para o planejamento e gerenciamento de análises de dados de *microarray*, temos que a ferramenta permite que os objetos sejam armazenados indexados pelos cartões, isto é, indexados pela hipótese que instigou sua criação. A ferramenta XPlanner (<http://xplanner.org/>) permite que se gerencie um projeto de desenvolvimento de software usando XP. Nela pode-se visualizar a evolução do projeto, os cartões pendentes, quem está desenvolvendo o que e qual o desempenho de cada programador. Sua aplicação é muito diferente do intuito da ferramenta TyPiR. Enquanto a primeira visa resolver o problema do acompanhamento de um projeto que usa XP, o TyPiR visa armazenar os resultados de um projeto desenvolvido usando XP. Logo, de certa forma, ele poderia ser usado para se avaliar a evolução do projeto, uma vez que permite a visualização dos objetos gerados em durante a análise, mas esse não é seu objetivos principal.

3. Busca por potenciais biomarcadores de diagnóstico e prognóstico de cânceres humanos

3.1. Introdução

3.1.1. Tecnologias da genômica para análise de expressão gênica no câncer

Até recentemente o estudo do câncer em nível molecular dependia de abordagens metodológicas restritas a análise de um gene por vez. Entretanto, os avanços tecnológicos da genômica em concerto com o conhecimento da sequência do genoma humano trazem excelentes perspectivas para a elucidação dos processos e mecanismos que determinam a transição do crescimento normal para o neoplásico e para o desenvolvimento de novas estratégias terapêuticas. Além disso, espera-se que a classificação dos diferentes tipos e subtipos de tumores, essencialmente baseada em critérios morfológicos e histoquímicos, seja complementada por uma classificação molecular baseada em indicadores biológicos (biomarcadores) relevantes para diagnóstico e prognóstico da doença bem como para a definição de alternativas terapêuticas personalizadas (Sotiriou & Piccart, 2007; Sawyers, 2008; van' t Veer & Bernards, 2008).

Os dados já disponibilizados nos últimos anos asseguram que a genômica tem potencial para desvendar quais são os genes que estão alterados no câncer, quantos destes genes estão alterados em cada tipo de câncer e que mecanismos determinam estas alterações. Entre as abordagens genômicas, a comparação dos perfis de expressão gênica global de tecidos normais e tumorais tem sido amplamente utilizada na busca de respostas a estas questões em diversos tipos de neoplasias malignas (Weber, 2002). Entre as tecnologias da genômica empregadas com esta finalidade, as principais e mais rentáveis em termos de quantidade e qualidade de informação gerada, estão o sequenciamento em larga escala de bibliotecas de cDNA gerando a informação conhecida como ESTs (*Expressed Sequence Tags*), a tecnologia denominada SAGE (*Serial Analysis of Gene Expression*) e os *microarrays* ou *chips* de DNA (Adams *et al.*, 1991; DeRisi *et al.*, 1996; Schena *et al.*, 1998; Lal *et al.*, 1999; Velculescu *et al.*, 1999; Lockhart & Winzeler, 2000; Strausberg *et al.*, 2000; Schulze & Downward, 2001). Somam-se a estas abordagens, a crescente utilização de tecnologias de última geração

para o sequenciamento de DNA na rápida determinação de genomas e transcriptomas completos (Morozova & Marra, 2008).

Os microarranjos de DNA, em particular os *oligoarrays*, são ainda considerados uma excelente alternativa para análise quantitativa dos níveis de expressão de milhares de genes simultaneamente, permitindo com relativa facilidade a avaliação global da expressão gênica em organismos, tecidos ou células, em diferentes situações fisiológicas ou patológicas (Schena *et al.*, 1998; Schulze & Downward, 2001; Hoheisel, 2006). Inicialmente os microarranjos continham moléculas de DNA geradas através de PCR de clones de bibliotecas de cDNA ou oligonucleotídeos com tamanho variando de 20 a 70 nucleotídeos (*oligoarrays*), representando milhares ou dezenas de milhares de genes (Venkatasubbarao, 2004). Hoje além da possibilidade da confecção “doméstica” de microarranjos ou de *oligoarrays*, há diversas plataformas comercialmente disponíveis (ex: Agilent, Affymetrix, NimbleGen). Além disso, a tecnologia evoluiu para obtenção dos chamados *tiling (oligo)arrays* que representam cromossomos completos, permitindo também a avaliação da expressão de regiões não codificadoras de um genoma (Bertone *et al.*, 2004; Ishkanian *et al.*, 2004; Mockler *et al.*, 2005). Este tipo de plataforma evidenciou a importância biológica dos ncRNAs. Cabe aqui ressaltar o crescente potencial dos ncRNAs como marcadores para o câncer, bem como seu importante papel na diferenciação celular e tumorigênese (Reis *et al.*, 2005; Szymanski *et al.*, 2005; Calin & Croce, 2006; Lin *et al.*, 2006; Amaral & Mattick, 2008) .

Tanto a reconhecida heterogeneidade como a instabilidade genética do tecido tumoral é um fator que torna a comparação entre perfis de expressão gênica de tumores uma tarefa complexa. Mas os dados já disponíveis na literatura ilustram o potencial dos microarranjos de DNA no estudo de diferentes aspectos da biologia do câncer. Utilizando-se microarranjos de cDNA ou de oligonucleotídeos, representando um conjunto de genes humanos variando entre 4.000 a 25.000, o perfil de genes expressos já foi comparado em diversos tipos de câncer (Golub *et al.*, 1999; Alizadeh *et al.*, 2001; Bhattacharjee *et al.*, 2001; Garber *et al.*, 2001; Okabe *et al.*, 2001; Ramaswamy *et al.*, 2001; Sorlie *et al.*, 2001; Welsh *et al.*, 2001b; Chung *et al.*, 2002; Singh *et al.*, 2002; van't Veer *et al.*, 2002; van de Vijver *et al.*, 2002; Mackay *et al.*, 2003; Meireles *et al.*, 2004; Reis *et al.*, 2004; Jones *et al.*, 2005; Weigelt *et al.*, 2005; Hu *et al.*, 2006; Chen *et al.*, 2007). Os resultados destas comparações têm sido extraordinários, permitindo avanços significativos na identificação e classificação de subtipos tumorais, na

descoberta de biomarcadores para diagnóstico e prognóstico da doença, na identificação de alvos terapêuticos além de aprofundar o conhecimento sobre a biologia do câncer.

3.1.2. Marcadores moleculares no diagnóstico e prognóstico de câncer

Os biomarcadores para o câncer podem ser divididos em quatro categorias: de diagnóstico, prognóstico, preditor e farmacodinâmico (Sawyers, 2008; van' t Veer & Bernards, 2008). A primeira categoria é utilizada na detecção da doença e na definição de seus subtipos. A segunda classe engloba os marcadores usados como uma ferramenta para previsão da evolução da doença, no direcionamento da conduta clínica e também no acompanhamento da evolução da doença. Um marcador preditor tem como meta estimar a probabilidade de que um paciente venha a se beneficiar de um tratamento específico. Já o quarto tipo, marcadores farmacodinâmicos, possibilita uma avaliação de como o tumor reagirá ao tratamento com um fármaco/quimioterápico, permitindo-se o cálculo da dosagem necessária para a resposta pretendida.

Para que um gene ganhe *status* de marcador clínico, ele deverá passar pelos seguintes estágios: estudo, validação e qualificação (Sawyers, 2008; van' t Veer & Bernards, 2008). No primeiro estágio, uma enorme gama de técnicas pode ser aplicada para prospecção de potenciais marcadores. No estágio de validação geralmente se avalia o desempenho do marcador em amostras clínicas enquanto no estágio de qualificação a eficácia do marcador é avaliada na clínica. Nesse último estágio, frequentemente se encontra muita relutância por parte dos médicos, o que é compreensível na medida em que o novo marcador pode contrariar a conduta médica até então adotada e/ou a intuição do clínico.

A busca por marcadores moleculares evoluiu da utilização de uma molécula, como o PSA (antígeno prostático específico) (Savage & Waxman, 1996), que ainda é usada para o acompanhamento clínico do câncer de próstata, até a busca de mutações em genes, busca por SNPs e também por conjuntos de genes ou proteínas, os quais são denominados de classificadores e/ou assinatura de genes (Misra *et al.*, 2002; Chang *et al.*, 2005; Dai *et al.*, 2005; Weigelt *et al.*, 2005; Wessels *et al.*, 2005; Glas *et al.*, 2006; Quackenbush, 2006; Cai *et al.*, 2007; Sawyers, 2008; van' t Veer & Bernards, 2008).

Inúmeros são os trabalhos que mostram o sucesso de se utilizar a tecnologia de microarranjos de DNA na avaliação do perfil de expressão gênica diferencial em tumores humanos visando à identificação de potenciais biomarcadores (Golub *et al.*,

1999; Alizadeh *et al.*, 2001; Dhanasekaran *et al.*, 2001; Finlin *et al.*, 2001; Garber *et al.*, 2001; Pinheiro *et al.*, 2001; Ross & Perou, 2001; Sorlie *et al.*, 2001; Su *et al.*, 2001; Beer *et al.*, 2002; Chung *et al.*, 2002; Pomeroy *et al.*, 2002; Squire *et al.*, 2002; van' t Veer *et al.*, 2002; van de Vijver *et al.*, 2002; Mackay *et al.*, 2003; Meireles *et al.*, 2003; Meireles *et al.*, 2004; Reis *et al.*, 2004; Brentani *et al.*, 2005; Simões *et al.*, 2005; Weigelt *et al.*, 2005; Hu *et al.*, 2006; Chen *et al.*, 2007; Dupuy & Simon, 2007; Nakaya *et al.*, 2007; Brito *et al.*, 2008; Sawyers, 2008; van' t Veer & Bernards, 2008). Porém somente três conjuntos de assinaturas gênicas estão atualmente disponíveis no mercado para diagnóstico e/ou prognóstico do câncer de mama: MammaPrint (Agendia), Oncotype DX (Genomic Health) e TherosH/I/TherosMGI (AviaraDx/BioMerieux). O primeiro é composto de uma assinatura para tumores de mama composta por 70 genes; o segundo é composto por 16 genes e o terceiro avalia um total 7 genes. Estes testes permitem a predição do real benefício ou necessidade de quimioterapia para a paciente e também uma estimativa do risco de recidiva ou de metástase, através de uma avaliação da agressividade do tumor com base no perfil de expressão destes biomarcadores em amostras do tecido tumoral (van' t Veer *et al.*, 2002; Buyse *et al.*, 2006; Glas *et al.*, 2006; Ma *et al.*, 2008; van' t Veer & Bernards, 2008).

A dificuldade de introduzir na prática clínica testes baseados em assinaturas de genes oriundas da análise de dados de microarranjos de DNA deve-se, em parte, a discrepância entre dados resultante, entre outros aspectos, da diferença entre as plataformas de microarranjos utilizadas. Para avaliar esse problema foi realizado um projeto chamado MAQC (*Microarray Quality Control*), que testou com as mesmas amostras de RNA referência (disponíveis comercialmente) em seis plataformas comerciais de microarranjos de DNA e uma plataforma de *oligoarray* (microarranjo de oligonucleotídeos) preparada pelo *National Cancer Institute* (NCI). Os resultados mostraram razoável consistência de dados obtidos entre as plataformas testadas em termos dos genes identificados como diferencialmente expressos, validando, assim, esta tecnologia para análise de perfis de expressão gênica (Shi *et al.*, 2006). Entretanto, na prática, o problema de cruzamento de dados ainda persiste dado que diferentes seqüências de DNA depositadas no microarranjo podem representar um mesmo gene, porém essas diferentes seqüências podem amostrar distintos transcritos variantes de *splicing*-alternativo. Assim, a correta anotação e qualificação da seqüência de DNA depositada no microarranjo são de extrema importância, em particular quando se pretende comparar dados obtidos a partir de diferentes plataformas de microarranjos.

Associada aos aspectos relativos à qualidade e à confiabilidade das plataformas de microarranjos (Chung *et al.*, 2002; Draghici *et al.*, 2006; Quackenbush, 2006; Shi *et al.*, 2006; Dupuy & Simon, 2007), a utilização de métodos estatísticos e computacionais adequados para análise tem papel crítico na definição do conjunto de biomarcadores, uma vez que atribuem significância estatística ao resultado encontrado e permitem análises visando a diminuição do número de potenciais biomarcadores a ser posteriormente validado (Golub *et al.*, 1999; van't Veer *et al.*, 2002; van de Vijver *et al.*, 2002; Causton & Game, 2003; Causton *et al.*, 2003; Simões *et al.*, 2005; Esteves, 2007; Esteves *et al.*, 2007).

3.1.3. O câncer de próstata e seus marcadores moleculares

O câncer de próstata é a neoplasia maligna mais frequentemente diagnosticada em homens nos países ocidentais. Entre os fatores de risco associados ao câncer de próstata estão idade, raça e histórico familiar, sendo que apenas uma pequena fração (5-10%) dos casos de câncer de próstata são hereditários e diagnosticados em homens suscetíveis com menos de 50 anos (Elo & Visakorpi, 2001; Jemal *et al.*, 2008).

Os tumores de próstata são, em sua maioria, adenocarcinomas, sendo que seu crescimento é dependente de andrógenos e a eliminação cirúrgica ou química da produção deste hormônio é o tratamento de escolha nos casos de tumores em estado avançado. Contudo, o aparecimento de clones celulares independentes de andrógeno é um problema clínico freqüente (Elo & Visakorpi, 2001). As aberrações cromossômicas mais freqüentes observadas nos tumores de próstata incluem perda de heterozigose (LOH, *loss of heterozygosity*) em alguns cromossomos, embora ampliações também tenham sido observadas. Assim a relativa heterogeneidade de expressão gênica no câncer de próstata é esperada (Abate-Shen & Shen, 2000; Reynolds, 2008).

A medida dos níveis séricos do antígeno específico de próstata (PSA) é empregada na detecção precoce de casos de câncer de próstata bem como na avaliação do estágio do tumor e na monitoração do seu desenvolvimento. Este é um diagnóstico importante que associado à intervenção cirúrgica e à radioterapia tem reduzido significativamente o número de mortes, contudo não há ainda cura efetiva para tumores em estágio avançado. Por outro lado, níveis elevados de PSA podem ocorrer em uma variedade de situações não cancerosas levando a resultados falsos positivos. Ademais a

correlação entre níveis de PSA e o estágio do tumor não está precisamente estabelecida (Parekh *et al.*, 2007; Sardana *et al.*, 2008). Atualmente, o escore de Gleason é um dos indicadores mais importantes na determinação do estágio do câncer de próstata, e associado dos níveis séricos de PSA, é indicativo do tipo de tratamento. A graduação de Gleason leva em consideração a conhecida heterogeneidade do tumor e se caracteriza por um escore onde os 2 padrões predominantes, com graus variáveis de 1 a 5, são contabilizados para obtenção da soma no total de 2 a 10. Os tumores com graus de 2 a 4 são considerados bem diferenciados, 5 e 6, bem a moderadamente diferenciados, os de grau 7 moderadamente a pouco diferenciado e aqueles cuja soma é variável entre 8 e 10, são considerados pouco diferenciados. (Albertsen *et al.*, 1998).

Os problemas encontrados no diagnóstico de tumores de próstata, incluindo falso-positivos e falso-negativos, indicam a necessidade do desenvolvimento de novos marcadores tumorais, que tornem o diagnóstico mais preciso e sejam indicativos do prognóstico real. Assim, a busca por novos biomarcadores para diagnóstico, prognóstico e progressão do câncer de próstata tem sido intensa e, vários candidatos têm sido propostos e alguns estão em fase de validação (Dhanasekaran *et al.*, 2001; Elo & Visakorpi, 2001; Luo *et al.*, 2001; Srivastava *et al.*, 2001; Welsh *et al.*, 2001a; Reis *et al.*, 2004; Parekh *et al.*, 2007; Reynolds, 2008; Sardana *et al.*, 2008).

3.1.4. Os Carcinomas Renais do tipo Célula Clara e seus marcadores moleculares

O Carcinoma de Célula Renal (RCC) é o tipo de câncer urológico mais letal e representa 3% dos cânceres diagnosticados no mundo, sendo que um terço dos pacientes costuma apresentar metástase. A classificação do RCC baseia-se predominantemente em análises morfológicas e histológicas, que dividem este tipo de carcinoma em diferentes subtipos: de células claras (CCRc), papilífero, de células cromóforas, de ductos coletores e medular. Os tumores do tipo célula clara são os mais comuns (70 a 80% dos casos) e os mais agressivos (Reuter & Presti, 2000; Jones & Libermann, 2007; Jemal *et al.*, 2008). O CCRc ocorre frequentemente em indivíduos portadores da doença von Hippel-Lindau, uma síndrome autossômica dominante de suscetibilidade a tumor, causada pela perda nas células germinativas do gene supressor de tumor von Hippel-Lindau (Linehan *et al.*, 2005).

Análises do perfil de expressão gênica de RCC através de microarranjos de DNA possibilitaram a identificação de genes com potencial para classificação molecular, prognóstico e tratamento deste carcinoma, além de genes relacionados com a transformação maligna neste tecido (Boer *et al.*, 2001; Skubitz & Skubitz, 2002; Jones *et al.*, 2005; Schuetz *et al.*, 2005; Yao *et al.*, 2005; Hirota *et al.*, 2006; Kopper & Timar, 2006; Jones & Libermann, 2007; Brito *et al.*, 2008). Outros potenciais biomarcadores foram identificados através da análise do perfil de expressão de proteínas em amostra de tecido tumoral ou de soro de pacientes portadores de RCC (Seliger *et al.*, 2007).

3.1.5. Os Sarcomas e seus marcadores moleculares

Os sarcomas, também chamados de tumores originados de tecidos moles, são tumores relativamente raros, correspondendo a menos de 5% de todos os tipos de cânceres em adultos. Eles apresentam alta morbidade e mortalidade, sendo sua taxa de sobrevida global, em 5 anos, inferior a 40% para alguns subtipos histológicos (Jemal *et al.*, 2008). O estudo de marcadores moleculares para sarcomas é clinicamente importante, pois até o presente os tumores de tecidos moles são classificados com base em critérios histológicos, sendo que o tipo atribuído a que uma amostra é deve-se à semelhança com outros tecidos diferenciados. Desta forma, tecidos de diferenciação adipocítica são chamados de lipomas/lipossarcomas; tumores de diferenciação muscular lisa, de leiomiomas/leiomiosarcomas e assim por diante. Porém os sarcomas sinoviais, alveolares de partes moles e epitelióides não se assemelham a qualquer tecido humano. A imunohistoquímica trouxe uma grande evolução no diagnóstico dos tumores de partes moles, pois tornou possível definir histogeneticamente tumores pouco diferenciados. Porém ainda há tumores que não expressam nenhuma das proteínas disponíveis pela imunohistoquímica, sendo esses chamados de sarcomas pleomórficos. (Mackall *et al.*, 2002; Helman & Meltzer, 2003).

Um grupo de sarcomas, que inclui os sarcomas sinoviais, é caracterizado por translocações específicas, que inclusive servem como critério de diagnóstico. Outro grupo de sarcomas não apresenta alterações genéticas específicas mas apresenta um cariótipo desorganizado reflexo de alterações genéticas e instabilidade cromossômica, como por exemplo leiomiossarcomas, fibrossarcomas e sarcomas pleomórficos (Helman & Meltzer, 2003).

A maioria dos sarcomas apresenta alterações nos genes que codificam a proteína do retinoblastoma (RB) e p53, conhecidos supressores tumorais. Acredita-se também que alterações nas vias de reparo e em genes envolvidos na estabilidade cromossômica possuam um papel relevante no desenvolvimento do sarcoma. (Stratton *et al.*, 1990; Dei Tos *et al.*, 1996; Helman & Meltzer, 2003).

Já foram verificadas relações entre genes de vias de fatores de crescimento e sarcomas, como fator de crescimento do tipo insulina 1 (IGF1) em rabdosarcomas e leiomiosarcomas, *c-kit* em sarcomas sinoviais e rhabdomyosarcomas e *c-met* em GISTs (*GastroIntestinal Stromal Tumors*) (Hoppener *et al.*, 1988; Roholl *et al.*, 1990; Kalebic *et al.*, 1998; Wallenius *et al.*, 2000; Berman & O'Leary, 2001; Hornick & Fletcher, 2002; Mackall *et al.*, 2002).

As intervenções terapêuticas nos diferentes sarcomas utilizam estratégias para inibição específicas das vias de sinalização celular alteradas, tendo como diagnóstico prévio o tipo e o grau histológico, sendo este último determinado pela quantidade de necrose, diferenciação celular e índice mitótico (Mackall *et al.*, 2002).

Alguns grupos de pesquisa, usando dados de microarranjos de DNA, já descreveram conjuntos de genes capazes de separar os sarcomas entre si (Allander *et al.*, 2002; Nagayama *et al.*, 2002; Nielsen *et al.*, 2002; Ren *et al.*, 2003; Nielsen, 2006; Antonescu, 2008) confirmando o potencial desta estratégia na definição de potenciais marcadores clínicos para classificação de sarcomas.

3.1.6. Os Carcinomas Epidermóides e seus marcadores moleculares

Os tumores carcinomas epidermóides (CE) são oriundos da camada basal de qualquer epitélio estratificado escamoso e é o tipo histológico predominante entre os tumores de cabeça e pescoço, e freqüente em pulmão, esôfago e colo de útero (Kumar *et al.*, 2005).

O carcinoma epidermóide de cabeça e pescoço (CECP) representa 90% dos tumores dessa região (Sankaranarayanan *et al.*, 1998). Quando diagnosticados em estágios avançados, tais tumores apresentam pouca chance de cura. Muitos dos CE são associados a fatores externos, como o os CE de esôfago a hábitos alimentares, os CE de pulmão ao fumo, os CECP ao álcool e ao fumo e o CE de colo de útero à infecções

persistentes pelo HPV (*Human Papiloma Vírus*) de alto risco (Parkin *et al.*, 2001; Carvalho *et al.*, 2005; Jemal *et al.*, 2008; Termine *et al.*, 2008).

Atualmente a conduta de tratamento para os paciente com CECP é baseada no estadiamento dado pelo sistema TNM (onde T é o tamanho do tumor, N representa linfonodos regionais comprometidos pela doença e M representa as metástases à distância), sendo que o tratamento consiste de cirurgia seguida ou não por radioterapia e/ou quimioterapia. O prognóstico do paciente é estimado com base no tamanho do tumor, comprometimento de linfonodos ou metástases linfonodais, extravasamento extracapsular, padrão de invasão vascular e perineural e margem de ressecção cirúrgica (Genden *et al.*, 2003; Howe *et al.*, 2006; Haddad & Shin, 2008).

Algumas alterações moleculares em CECP já foram descritas, como expressão aumentada de ciclina D1 e EGFR (Receptor do fator de crescimento epidérmico), e a perda da expressão de p16, assim como mutações em p53 e a infecção e integração do genoma do HPV (Almadori *et al.*, 2008; Chang & Califano, 2008; Palka *et al.*, 2008). Genes relacionados à proliferação e sobrevivência celular, adesão e matriz extracelular, angiogênese e invasão têm sido freqüentemente apontados por análises de dados de microarranjos de DNA (Belbin *et al.*, 2002; Cromer *et al.*, 2004; Ginos *et al.*, 2004; Belbin *et al.*, 2005; Chung *et al.*, 2006; Giri *et al.*, 2006), permitindo em alguns casos a predição de recorrências e de metástase à distância. A busca por marcadores de comprometimento linfonodal levou à identificação de conjuntos de genes preditores capazes de diagnosticar metástases linfonodais em amostras independentes com acurácia de 86% e (precisão do diagnóstico clínico é de 68%), sendo de 100% para linfonodos livres de doença e 77% para linfonodos comprometidos (Roepman *et al.*, 2005; Roepman *et al.*, 2006). Entretanto, apesar do sucesso destas análises, estes potenciais marcadores ainda não foram incorporados à clínica.

3.2. Métodos

Os resultados discutidos nesta tese referem-se a análise de 5 conjuntos de dados obtidos a partir da avaliação do perfil de expressão gênica de amostras de diferentes tipos de tumores utilizando-se duas plataformas distintas de microarranjos de DNA, conforme esquematizado na Figura 13. A seguir detalhamos as características das

plataformas de microarranjos empregadas bem como dos conjuntos de amostras analisado.

Figura 13: Resumo das análises de dados de microarranjos apresentadas nesta tese.

3.2.1. A plataforma de microarranjos *non-coding*

A plataforma de microarranjos intitulada, nesta tese, de “non-coding” foi utilizada na obtenção dos dados de expressão gênica de tumor de próstata e de rim, os quais foram submetidos a nossa análise. Esta plataforma foi desenvolvida pelo grupo de pesquisa do Prof. Sérgio Verjovski-Almeida (IQ-USP) (Reis *et al.*, 2004). Os dados de expressão gênica foram também integralmente obtidos pelo grupo do Prof. Verjovski-Almeida e gentilmente disponibilizados para nossas análises. A plataforma *non-coding* é constituída de 4216 sequências depositadas em lâminas de vidro, representando aproximadamente 3700 ESTs humanas. Como entre os 4216 sequências estão incluídos *housekeeping* genes, controles negativos e controles positivos, pode-se considerar dados associados a genes, o total 3816 sequências ou pontos do microarranjo. 2292 sequências

referem-se a genes sabidamente relacionados a câncer, selecionados usando as bases de dados Entrez, OMIM e CGAP do NCBI. Os processos biológicos selecionados foram apoptose, tumorigênese, metástase, metabolismo de câncer, e progressão de câncer. As 2292 sequências relacionadas a genes conhecidos mapeiam em exons. O restante das sequências de DNA depositadas no microarranjo foi aleatoriamente selecionado dentre sequências consideradas *no-match*, das quais 822 representam ncRNAs (RNA não codificadores) que mapeiam em regiões intrônicas de genes anotados. Outras 241 sequências depositadas no microarranjo mapeiam em regiões não anotadas do genoma humano. A plataforma *non-coding* foi usada na avaliação do perfil de expressão gênica em tumores de próstata e carcinoma renal, possibilitando a demonstração do importante potencial dos transcritos intrônicos como biomarcadores (Reis *et al.*, 2004; Brito *et al.*, 2008).

3.2.2. Conjunto de dados: câncer de próstata

O conjunto de dados de microarranjos resultante da análise do perfil de expressão gênica de amostras de tumores de próstata humano previamente anotados e normalizados (Reis *et al.*, 2004) foi submetido a uma re-análise, onde utilizamos uma abordagem diferente (SAM) (Simões *et al.*, 2005)² daquela anteriormente aplicada (Correlação de Pearson). Dois tipos de análise foram realizados, sendo o primeiro a busca por marcadores de diagnóstico e o segundo a busca por marcadores de prognóstico.

O conjunto consiste de 48 amostras de RNA isolado de próstata humanas, das quais 23 são amostras de tecido normal e 25 de tecido tumoral. A amostra de referência usada nas hibridizações é um *pool* de RNA das amostras utilizadas. Como análise alternativa, decidimos por usar como critério de seleção dos genes o método SAM (*Significance Analysis of Microarrays*) (Tusher *et al.*, 2001), implementado como *plugin* para o Excel® (versão: 1.21). Com o intuito de descobrir potenciais marcadores, que possam distinguir entre tecido normal e tecido tumoral, pode-se usar o SAM considerando duas classes presentes nos dados (normal e tumoral) em uma análise denominada: *two-class unpaired analysis*. Optamos nessa fase por também fazer uma

² Simões, A. C. Q., da Silva, A. M., Verjovski-Almeida, S. & Reis, E. M. 2005. SAM Method as an Approach to Select Candidates for Human Prostate Cancer Markers. Lecture Notes in Computer Science 3594/2005:202-205. Consta dos anexos inclusos no CD-ROM suplementar a tese.

análise considerando três classes, para tanto executamos no SAM a análise: *multiclass response*.

Como entre os 4216 sequências estão incluídos *housekeeping genes*, controles negativos e controles positivos, foi necessário fazer a primeira filtragem nos dados resultando em 3816 sequências (*spots*). A partir desse conjunto é que fizemos os testes com os filtros seguintes.

3.2.2.1. Filtragem e seleção com SAM para a análise de normais x tumorais

Os filtros aplicados na tabela de intensidades dos transcritos foram escritos em Perl e tinham como meta:

- * **Filtro 1:** Retirar *spots* referentes a *housekeeping genes*, controles negativos, controles positivos e contaminantes.

- * **Filtro Global:** Excluir *spots* que possuem uma variação global muito pequena, considerando um número mínimo de amostras. Ou seja, deseja-se selecionar *spots* que apresentam intensidades fora de certa faixa em pelo menos certo número de amostras.

- * **Filtro Local:** Excluir *spots* que variam muito no conjunto de amostras normais, considerando um número mínimo de amostras. Ou seja, deseja-se selecionar *spots* que apresentam intensidades dentro de uma faixa definida em pelo menos certo número de amostras.

Para definir as faixas que foram consideradas duas abordagens foram implementadas para cada um dos filtros: usar a média e variância (V) das intensidades dos *spots* em questão ou usar o *fold change* (F) com relação às intensidades dos *spots* em questão. O número mínimo de amostras em questão é referido por N.

Após filtrar os transcritos realizamos a análise de *two-class unpaired data* com SAM, categorizando as amostras em duas classes: Normais (1) e Tumorais (2). Como parâmetros utilizamos: 1000 amostras de *bootstrap*, *K-Nearest Neighbors Imputer* como *Imputation Engine*, semente randômica e delta (tabela 5, seção 3.2.2.2).

3.2.2.2. Filtragem e seleção com SAM para a análise de marcadores de prognóstico

Seguindo o caminho indicado pelos resultados da busca por marcadores de diagnóstico, aplicamos o SAM na busca por potenciais marcadores de prognóstico. Para tal apenas dados derivados de amostras tumorais foram considerados nesta análise. O conjunto de dados é formado de 27 amostras tumorais de próstata, com Gleason Score (GS) variando entre 5 e 10. Vale lembrar que o Gleason Score é atualmente o parâmetro utilizado pelo patologista para classificação do tecido tumoral, cuja pontuação vai de 1 a 10 e deve-se a soma de duas pontuações, de 1 a 5, atribuídas por cada patologista conforme as características histológicas de malignidade do tecido (Albertsen *et al.*, 1998). Como já mencionado, a categorização destas amostras tumorais a partir do mesmo conjunto de dados de microarranjos já havia sido realizada (Reis *et al.*, 2004), sendo que o critério de seleção de genes candidatos utilizado foi o coeficiente de correlação de Pearson da expressão de um transcrito com o Gleason Score (GS) da amostra.

Inicialmente foi aplicado o filtro 1 descrito no item anterior que visa eliminar dados de *spots* contaminantes e controles. Um segundo filtro usando-se a média e a variância de cada um dos transcritos escrito em Perl foi aplicado aos dados, com o intuito de selecionar transcritos que variam bastante dentre as amostras tumorais, considerando um número mínimo de amostras (N). Mais especificamente, o N usado foi igual a 4 e o desvio padrão igual a 1,25. Porém, a busca por marcadores de prognóstico também foi realizada sem a aplicação desse segundo filtro.

Após os filtros, seguiu-se a etapa de seleção dos transcritos com o SAM. Tanto a análise de *two-class, unpaired data*, como a de *multiclass* foram realizadas. No primeiro caso somente as amostras com GS baixo (5 e 6) e alto (9 e 10) foram consideradas. Já no segundo caso os três níveis de GS foram considerados: baixo (5 e 6), intermediário (7 e 8) e alto (9 e 10). Para o conjunto de transcritos selecionado com a análise por SAM *two-class unpaired data* decidimos verificar como as amostras com GS intermediário (7 e 8) se separam. O fluxograma dessa análise está apresentado na Figura 14.

Como parâmetros para o SAM usamos 1000 amostras de *bootstrap* e os valores de delta. Os resultados obtidos estão apresentados na tabela 5. Os valores de *false*

discovery rate, FDR (Storey, 2002), em conjunto com o número de transcritos retornados indicam um conjunto de genes candidatos muito promissor. Com o intuito de verificarmos como estava a separação de tais amostras, considerando-se o conjunto de transcritos selecionados pelo SAM, optamos por realizar um agrupamento (*clustering*) hierárquico usando tais transcritos.

Figura 14: Fluxograma de análise para a busca por marcadores de prognóstico com SAM.

Tabela 5: Parâmetros e resultados para as análises com SAM na busca por marcadores de prognóstico.

Conjunto	Filtro de variação	# Spots	# Spots Two-class	delta	FDR %	# Spots Multi-class	delta	FDR %
Sem filtro	não	3816	67	1,45	0,91	19	0,42	3,87
Com filtro	sim	1604	49	1,45	1,21	12	0,42	5,80

3.2.2.3. *Clustering* hierárquico e validação

O *software* usado para fazer o *clustering* hierárquico foi o SpotFire DecisionSite™ versão 15.3.486 e para a normalização prévia ao *clustering* usamos Z-score. Utilizamos *Ward's method* como método de *clustering* e como função de ordenação, o valor médio.

Para avaliarmos a robustez do *clustering*, optamos por introduzir um ruído nos dados. Assim, uma vez com ruído poderíamos averiguar com que frequência o *clustering* se sustentava. Para tanto dez mil amostras foram geradas com ruído gaussiano e para cada uma, uma árvore foi feita. Por final uma árvore consenso foi calculada, em que a cada ramo está associado o número de vezes que tal ramo se sustentou, considerando as dez mil árvores. Os programas do ruído gaussiano foram escritos em R (versão 2.0.1 Patched) e a árvore consenso foi criada usando o Consensus Tree do programa Phylip (versão 3.61).

O gerador de ruído gaussiano foi escrito em R (Ihaka & Gentleman, 1996), usando as funções *rnorm*, *mean*, *std*. Ele usa uma curva gaussiana com média zero e desvio padrão obtido de uma curva obtida dos dados. Para obtermos a curva também usamos pacotes disponíveis no ambiente R, mais especificamente usamos a função *nlm* (*non linear minimization*), que minimiza uma função usando um algoritmo do tipo de Newton (Schnabel *et al.*, 1985; Dennis & Schnabel, 1987). Dado que *std* refere-se ao desvio padrão dos spots, *int* refere-se a média das intensidades e *p1* e *p2* aos parâmetros que se quer estimar, a função *f(x)* a ser minimizada, era, pela equação 2:

$$f(x) = \sum \{ [\text{std} - p1 \cdot e^{(-\text{int} \cdot p2)}]^2 \} \quad \text{eq(2)}$$

Assim, com a minimização pudemos encontrar os parâmetros *p1* e *p2* e, conseqüentemente a curva mostrada na Figura 15. Uma vez encontrados os parâmetros *p1* e *p2* da *f(x)*, podemos adicionar um ruído em cada uma das intensidades de forma randômica, porém usando uma distribuição para esse ruído (usando a função *R norm* do R). É importante ressaltar que a curva que fornece o valor de desvio padrão para o ruído gaussiano confere valores de desvio-padrão maiores para intensidades menores, uma

vez que por terem valores menores estão mais sujeitos à interferência do limite de detecção do *scanner* e variação experimental (erro experimental).

Figura 15: Curva ajustada ao gráfico por minimização não linear.

Os agrupamentos consensos foram feitos usando as dez mil árvores geradas a partir do dado com ruído. Para calcular a árvore consenso o programa Consensus Tree do Phylip foi usado.

3.2.3. Conjunto de dados: Carcinoma renal do tipo célula clara

As amostras de tecido de rim humano foram processadas pela aluna de Doutorado Angela Fachel, sob a orientação do Prof. Sergio Verjovski-Almeida (IQ-USP). O RNA total foi isolado de amostras de carcinoma renal (RCC) do tipo célula clara, sendo 20 amostras de tecido normal e 20 de tecido tumoral, sendo que 17 são pareadas. Como referência nas hibridizações foi usando um *pool* de RNA das próprias amostras do experimento. Os dados normalizados extraídos dos microarranjos nos (Fachel & Verjovski-Almeida, dados não publicados) foram submetidos a análise com a metodologia descrita por (Gomes *et al.*, 2005) com intuito de investigarmos trios de transcritos com potencial classificador. Esta metodologia baseia-se na análise por

discriminante linear de Fisher. Como medida de qualidade dos trios, optamos por usar *svd* (*singular value decomposition*, que leva em consideração a distância entre os grupos e a dispersão dos mesmos). Uma vez que a plataforma de *microarray* utilizada para a hibridização das amostras representa 4608 seqüências, se realizássemos o cálculo de todos os trios possíveis, teríamos da ordem de 10^{12} possíveis trios. Caso pudéssemos calcular todos os possíveis trios, teríamos certeza de ter escolhido o trio que permite a melhor separação entre as classes normal e tumoral. Porém, tal cálculo é proibitivo, dada a atual capacidade computacional que dispomos.

3.2.3.1. Primeira estratégia de redução: genes diferencialmente expressos

Com o intuito de reduzir o número de transcritos a serem usados nos trios, optamos por realizar as etapas esquematizadas na Figura 16 e descritas a seguir:

1. Os transcritos foram filtrados por critérios de qualidade e controle (exemplo: controles negativos, *housekeeping* genes).
2. Os transcritos foram separados nas categorias intrônicos e exônicos, e então foi feita a análise de genes diferencialmente expressos, considerando tal categorização.
3. Com o intuito de explorar todos os possíveis trios compostos por genes diferencialmente expressos selecionados por algum dos critérios descritos no item acima, optamos por fazer a união dos conjuntos do item anterior, que chamaremos de "conjunto união".

Como medida de qualidade dos trios optamos por usar *svd* (*singular value decomposition*, que na verdade leva em consideração a distância entre os grupos e a dispersão entre os grupos). Os *scripts* para tal análise foram escritos nas linguagens Perl e R.

Figura 16: Esquema da estratégia para seleção de transcritos a serem usados na busca por melhores trios para classificação das amostras entre normal e tumoral

3.2.3.2. Segunda estratégia de redução: genes mais expressos

A segunda estratégia de redução do número de genes a serem considerados para análise de trios consistiu em selecionarmos os genes mais expressos e então aplicar o método chamado “busca e escolha” implementado pelo aluno de doutorado Elier Broche Cristo (Meireles *et al.*, 2004; Gomes *et al.*, 2005). Tal método consiste em selecionar inicialmente os genes mais diferencialmente expressos pelo teste-t e então calcular o discriminante linear de Fisher para os pares de genes. O critério de qualidade é o *svd* (*singular value decomposition*), que é calculado usando a distância entre os grupos e a dispersão dos mesmos. Então, para os melhores pares, calcula-se os trios, também por discriminante linear de Fisher e ordenados por *svd*.

A questão de altos valores de expressão faz-se interessante caso o marcador seja colocado em rotina, pois estaria menos sujeito a erros experimentais por estar mais longe do limite inferior de detecção. Com isso criamos 4 grupos de genes, que diferiam no número de elementos, conforme Figura 16, uma vez que o tempo para calcular tais

trios é um fator limitante. Como podemos observar na Figura 17, não obtivemos resultados para a abordagem exaustiva, pois o objeto gerado não coube na memória.

Figura 17: Conjuntos de genes gerados para fazer a análise de trios.

3.2.4. A plataforma de *microarray* ACCamargo3 e análises de dados

A plataforma de microarranjos intitulada, nesta tese, de “ACCamargo3” foi utilizada na obtenção dos dados de expressão gênica de sarcomas e carcinomas epidermóides, os quais foram submetidos a nossa análise. Esta plataforma foi desenvolvida pelo na *Microarray Facility* do Laboratório de Inflamação do Instituto Ludwig de Pesquisa sobre o Câncer e do Hospital A.C. Camargo. As lâminas da plataforma “ACCamargo3” são constituídas de 4800 *spots* referentes a 4574 *probes* únicas, sendo 8 de controles. As sequências de DNA representadas na plataforma foram anotadas usando a ferramenta AMPiR, que desenvolvemos e que está descrita nesta tese, que encontrou no total 3748 símbolos (gene Symbol) oriundos do banco Gene do NCBI para os genes da plataforma. No total mapeamos 427 grupos funcionais para essa

plataforma, que foram obtidos do KEGG (<http://www.genome.jp/kegg/>), GO (www.geneontology.org/), Biocarta, STKE e publicações da literatura.

A abordagem que escolhemos para mapear uma plataforma de microarranjo está exemplificada na Figura 18. Para que se possamos percorrer os relacionamentos entre os bancos de dados do NCBI, é preciso usar os **gi**, isto é, identificações únicas dentro de cada banco de dados. Uma vez com o **gi** de uma sequência de nucleotídeos, opta-se preferencialmente por usar o relacionamento entre o banco de dados nucleotídeos e o banco de dados de gene, uma vez que tal relacionamento foi criado de forma manual e criteriosa. Porém, como muitos nucleotídeos não possuem tal relacionamento, optamos por usar o relacionamento que nucleotídeos possuem com Unigene como um caminho intermediário para chegar ao banco de dados de gene. Para os casos em que havia mais de uma entrada em Unigene relacionada a uma sequência de nucleotídeos, representando possivelmente isoformas, optamos por não prosseguir com o mapeamento, ficando essa sequência sem anotação. Uma vez no banco de dados Unigene, percorre-se o relacionamento entre o banco Unigene e Gene, sendo, portanto possível recuperar as informações de Gene Ontology e KEGG associadas a entidade Gene.

Figura 18: Esquema dos relacionamentos usados entre os bancos para anotação das sequências de DNA depositadas na plataforma de *microarray*. O caminho preferencial de anotação é o demarcado pela seta dupla), pois é curado manualmente. Quando inexistente, opta-se pelo demarcado com seta fina simples, passando pelo Unigene, sendo assim possível chegar na entidade Gene. Uma vez que se tem a representação de gene para cada um dos probes, pode-se então recuperar a informação de Gene Ontology e de KEGG.

Os conjuntos de dados gerados com a plataforma ACCamargo3 foram submetidos a etapa de normalização no pacote MaigesPack (Esteves, 2007). A

normalização dos dados é essencial uma vez que há uma diferença na incorporação dos fluoróforos Alexa 555 (verde) e Alexa 647 (vermelho) na etapa de síntese do cDNA como também entre a fluorescência dos mesmos (Quackenbush, 2002). Vale ressaltar que o pacote MaigesPack, além da normalização, permite diversas análises, como esquematizado na Figura 19.

Figura 19: Possíveis análises de serem realizadas com o pacote MAIGES.

As vias e módulos funcionais alterados foram identificados também através da metodologia descrita por (Segal *et al.*, 2004) com modificações (Gomes *et al.*, 2005). O método de módulos funcionais é baseado no conhecimento biológico dos genes e não se baseia em genes diferencialmente expressos. Os genes presentes na lâmina de *microarray* são classificados em funções ou vias das quais participam, conforme descrito em KEGG (<http://www.genome.jp/kegg/>) e GO (www.geneontology.org/).

3.2.5. Conjunto de dados: Sarcoma

O conjunto de dados intitulada Sarcoma consiste de 102 amostras de tumores de tecido mole, também chamados de sarcomas, classificadas como Benignos (BMT), Fibromatoses (DTF) e Sarcomas (MMT) (tabela 6). A extração do RNA total das amostras e os experimentos de hibridização dos microarranjos foram realizados pela

aluna Isabela Werneck da Cunha do Hospital A.C. Camargo, sob a orientação do Dr. Luis Fernando Reis.

A amostra de referência para hibridização dos microrranjos consiste de um *pool* de RNA total de 15 linhagens celulares de diversos tecidos epiteliais e mesenquimais, tumorais e normais (tabela 7). Cada lâmina da plataforma ACCamargo3 foi hibridizada com quantidades iguais de sondas de cDNA fluorescente, correspondentes a cada uma das amostras em estudo ou ao RNA referência. Utilizou-se a estratégia de *dye swap*, em que cada amostra é hibridizada duas vezes, sendo que na segunda hibridização há a troca dos corantes Alexa 555 (verde) e Alexa 647 (vermelho). Após a reação de hibridização foi feita a lavagem das lâminas e aquisição das imagens em um leitor a laser confocal (Scanarray Express, Packard). A leitura das imagens foi feita com o programa Scanarray.

Os sarcomas presentes nesse conjunto pertencem aos tipos: Sinovial (SS), Pleomórfico (SP), Adipocítico (TAM), Fibroblático (TFM), Muscular (TMM) e Neuronal (TNM). Já os tumores benignos são dos tipos: Muscular (TMB) e Neuronal (TNB). Como outro tipo de tumor temos as fibromatoses (TFB ou DTF), que possuem a peculiaridade de serem tumores agressivos localmente, porém não evoluem para tumores metastáticos. A tabela 6 resume as amostras deste conjunto de dados.

A peculiaridade das fibromatoses de serem agressivas localmente e não apresentarem metástase, quando comparada com os tumores benignos e malignos (tabela 8), permitiu a busca por marcadores moleculares para agressividade local e também por marcadores para metástase.

O Meta-objeto de sarcoma é a representação do conjunto de dados de microarranjos que foi usado como metáfora pela equipe de XP. De certa forma é a estrutura do objeto sumarizado do pacote MaigesPack (Esteves, 2007), que pode ser visualizada com o comando `str` do R. Porém, além da estrutura, o meta-objeto possui informações detalhadas sobre cada uma das variáveis e também sobre a anotação da plataforma de *microarray*. No total havia 26 variáveis biológicas conforme a tabela 9.

Tabela 6 – Amostras do conjunto de dados de sarcoma

Origem	Tipo	Sigla	Classificação	Diagnóstico	Amostras
Tumor Muscular	TMB	BMT	Benigno	Leiomioma	2
Tumor Neuronal	TNB	BMT	Benigno	Neurofibromas e Schwannomas	5
Tumor Fibroblástico	TFB	DTF	Fibromatose	Fibromatose tipo desdemóide	19
Sarcoma Sinovial	SS	MMT	Sarcomas	Sarcomas Sinoviais	19
Sarcoma Pleomórfico	SP	MMT	Sarcomas	Sarcomas Pleomórficos	17
Tumor Adipocítico	TAM	MMT	Sarcomas	Lipoblástico e Liposarcoma	5
Tumor Fibroblástico	TFM	MMT	Sarcomas	Fibrosarcoma	6
Tumor Muscular	TMM	MMT	Sarcomas	Leiomiosarcoma, Tumor de stroma gastrointestinal (GIST) e Sarcoma Alveolar	24
Tumor Neuronal	TNM	MMT	Sarcomas	Tumor de maligno de nervo periférico (MPNST)	5

Tabela 7 – Lista das linhagens celulares utilizadas para obtenção do RNA referência

Daudi	Linfoma de Burkitt
DLD-1	Adenocarcinoma de Cólon
DU 154	Carcinoma de Próstata
FaDu	Carcinoma Epidermóide de Faringe
GM 637	Fibroblasto
H 146	Carcinoma de Pulmão
H 1080	Fibrossarcoma
HB4 α	Célula Luminal de Mama
HEK 293	Rim Embrionário Humano
Jurkat	Leucemia Aguda de Células T
Saos-2	Osteossarcoma
SK-BR-3	Adenocarcinoma de Mama
SK-Mel-28	Melanoma
T24	Carcinoma de Bexiga
T98G	Glioblastoma

Tabela 8 – Características do conjunto de amostras de sarcoma

Tipo	Malignidade	Agressividade Local	Potencial Metastático
Benignos	Benigna	Pequena	Pequeno
Fibromatose	Benigna	Grande	Pequeno
Sarcomas	Maligna	Grande	Grande

Tabela 9– Variáveis biológicas do Meta-objeto

Identificação da Amostra	Código interno da amostra
microgramas de RNA extraído	Quantidade de RNA extraído do tumor
Arquivo	Nome do arquivo de carga com os dados brutos
Ref	Informação de com que corante a amostra referência está marcada
Grupo biológico	Tipo de sarcoma: TFB, SS, SP, TNB, TMM, TMB, TNM, TFM, TAM
Origem do tumor	Tumores fibroblásticos, sarcomas sinoviais, sarcomas pleomórficos, Tumores neurais, Tumores musculares, Tumores adipocíticos
Swap	nome do arquivo que contém os dados brutos de Dye Swap
Diagnóstico	Diagnóstico do tipo de tumor do paciente
Grau	Informação de grau do tumor. Exemplo: benigno, baixo, alto.
Localização	Localização do tumor
Tamanho	Tamanho do tumor em milímetros
Margens	Informação de se após a retirada do tumor ainda restou tumor no tecido. Se não restou: margens livres. Se restou: margens comprometidas.
IVL	Invasão vascular linfática
IVS	Invasão vascular sanguínea
IPN	Invasão perineural
Sexo	Gênero do paciente
Raça	Raça do paciente
Idade	Idade do paciente ao fazer a cirurgia
Data diagnóstico	data do diagnóstico
Quimio_terapia_neo	Informação de se fez quimioterapia antes da cirurgia
RXT_neo	Informação de se fez radioterapia antes da cirurgia
Quimio terapia adj	Informação de se fez quimioterapia após a cirurgia
RXT_adj	Informação de se fez radioterapia após a cirurgia
Recidiva local	Informação de se teve recidiva local
Metastase	Informação de se teve metástase
Local da metástase	Informação do local onde apareceu a metástase

Além das variáveis biológicas, devemos considerar também as variáveis para os genes representados no meta-objeto, resultantes da anotação da plataforma de *microarray* usando o AMPiR (tabela 10). E ainda, como parte do meta-objeto temos a informação dos 127 grupos gênicos e nove vias de sinalização anotados para a plataforma ACCamargo3. Os dados foram carregados no R usando o MaigesPack, (Esteves, 2007).

Tabela 10 – Variáveis da anotação da plataforma

Linha	Localização do spot na lâmina
Coluna	Localização do spot na lâmina
Nome do clone	Identificador interno da sequência
giNucl	Chave (gi) do banco de dados de Nucleotídeos do NCBI (obtido com o AMPiR)
accNucl	Chave de acesso (Accession) do banco de dados de Nucleotídeos do NCBI (obtido com o AMPiR)
giUniGene	Chave (gi) do banco de dados UniGene do NCBI (obtido com o AMPiR)
geneSymbolUniGene	Símbolo do Gene (Gene Symbol), obtido do banco de dados UniGene do NCBI (obtido com o AMPiR)
geneNameUniGene	Nome do Gene (Gene Name), obtido do banco de dados UniGene do NCBI (obtido com o AMPiR)
giGene	Identificador do gene, obtido do banco de dados Gene do NCBI (obtido com o AMPiR).
geneSymbolGene	Símbolo do Gene (Gene Symbol), obtido do banco de dados Gene do NCBI (obtido com o AMPiR)
geneNameGene	Nome do Gene (Gene Name), obtido do banco de dados Gene do NCBI (obtido com o AMPiR)
chromNum	Número do Cromossomo obtido do banco de dados Gene do NCBI (obtido com o AMPiR)
Lócus	Mapa do gene - locus, obtido do banco de dados Gene do NCBI (obtido com o AMPiR)

3.2.6. Conjunto de dados: Carcinomas Epidermóides

O conjunto de Carcinomas epidermóides é composto por tumores de carcinoma epidermóide de quatro topologias diferentes: cabeça e pescoço, pulmão, esôfago e colo uterino. Ele consiste de 171 amostras, sendo 121 de cabeça e pescoço, 9 de pulmão, 22 de esôfago e 19 de colo uterino. A extração do RNA total das amostras e os experimentos de hibridização dos microarranjos foram realizados pela aluna Anna Coló do Hospital A.C. Camargo, sob a orientação do Dr. Luis Fernando Reis. Os

procedimentos experimentais para obtenção dos dados de microarranjos destas amostras são idênticos aos descritos para as amostras de sarcoma.

No total havia 100 variáveis clínicas disponíveis para esse conjunto de dados, que estão resumidas na tabela 11.

Tabela 11 – Variáveis clínicas principais das amostras de carcinoma de cabeça e pescoço

Variável	O que descreve
Subregião_topologia	Topologia e detalhe da mesma
Topologia principal	Topologia Principal. Ex: BOCA, ORO...
Idade de nascimento	Idade de nascimento do paciente
Idade na cirurgia	Idade do paciente no momento em que foi retirada a amostra
Sexo	Gênero
Tabaco	Se havia exposição ao tabaco
Quantidade de tabaco	Quanto se expunha ao tabaco
Tempo de tabaco	Por quantos anos se expôs ao tabaco
Álcool	Se havia exposição ao álcool
Quantidade de álcool	Quanto se expunha ao álcool
Tempo de álcool	Por quantos anos se expôs ao álcool
Extensão extracapsular	Se houve extensão extracapsular
Esvaziamento	Se houve esvaziamento e quantos linfonodos deram positivos
Invasão Perineural	Se houve invasão perineural
Invasão Vascular	Se houve invasão vascular
Grau de diferenciação	Grau de diferenciação do tumor
Tamanho	Tamanho do tumor
Margens	Se foi possível retirar todo o tumor ou ainda ficou algum

3.2.7. Conjunto de dados: Carcinomas Epidermóides de Cabeça e Pescoço

Esse conjunto de dados pode ser entendido como um subconjunto do conjunto de dados de carcinomas epidermóides (seção anterior), uma vez que corresponde somente aos carcinomas epidermóides de cabeça e pescoço e que a normalização dos dados foi feita separadamente.

O conjunto de dados consiste de 121 amostras de carcinoma epidermóides oriundos das seguintes topologias: 55 de boca, 35 da cavidade oral, 7 da cavidade de seios para nasais, 18 de laringe e 6 de hipolaringe, conforme detalhado na tabela 12. Como mencionado anteriormente, a extração do RNA total das amostras e os experimentos de hibridização dos microarranjos foram realizados pela aluna Anna Coló do Hospital A.C. Camargo, sob a orientação do Dr. Luis Fernando Reis e segundo os procedimentos experimentais descritos para as amostras de sarcoma e carcinomas.

No total havia 104 variáveis clínicas disponíveis para esse conjunto de dados, que estão resumidas na tabela 11. As amostras possuem a informação de exposição a álcool, tabagismo e quimioterapia adjuvante ao tratamento; no entanto as amostras foram coletadas antes de qualquer tratamento de radioterapia ou quimioterapia. Entre as questões que podem ser abordadas a partir da análise destes dados, é se o álcool e tabaco têm influência na evolução do tumor. Outra se refere à discriminação de amostras conforme prognóstico ou pela sua topologia.

Tabela 12 – Amostras de carcinomas epidermóides de cabeça e pescoço

Topologia	Sigla	Amostras	Detalhes da topologia
Boca	BOCA	55	Língua, assoalho de boca, retromolar, gengiva, mucosa jugal, palato duro
Cavidade Oral	ORO	35	Base da língua, loja amigdaliana, palato mole, úvula, valecula
Cavidade dos Seios Paranasais	CNSP	7	Fossa nasal, seio maxilar
Laringe	LARI	18	Laringe supraglótica, glótica e subglótica
Hipolaringe	HIPO	6	Hipofaringe, seio piriforme

Além das variáveis clínicas listadas, temos também as informações de anotação gênica da plataforma ACCamargo3, porém com a informação de 527 grupos gênicos, construídos a partir de bases de dados como KEGG, gene Ontology ou artigos, somados a informação de nove vias de sinalização.

A criação do meta-objeto no caso da análise dos tumores de cabeça e pescoço foi crítica para que conseguíssemos analisar dependência entre variáveis no objeto. Como exemplo, temos a variável recidiva que possui como variável de confundimento a quimioterapia adjuvante. Isso deve-se ao fato de que a recidiva do tumor é afetada pelo fato de o paciente ter sido tratado com quimioterapia adjuvante, fazendo-se necessário portanto a segregação dos dados em dois conjuntos toda vez que formos analisar recidiva: os tratados com quimioterapia e os não tratados.

Inicialmente as perguntas biológicas eram:

- a) As topologias podem ser diferenciadas? Isto é, tumor de cabeça e pescoço de boca é diferente de tumor de orofaringe, que por sua vez é diferente de cns?
- b) Quais são os marcadores moleculares de prognóstico?
- c) Quais são as vias e processos biológicos responsivos a álcool e/ou tabaco

Para a análise de árvores de decisão apresentada na seção 3.3.4.4. desta tese, usamos a implementação CART do algoritmo de partição recursiva para árvores de decisão (Breiman *et al.*, 1984), disponível no pacote rpart do R. Esse algoritmo faz uma busca exaustiva sobre todas as possíveis partições maximizando a informação por nó.

3.3. Resultados

3.3.1. Descoberta de potenciais marcadores para prognóstico do câncer de próstata

3.3.1.1. Análise de normais e tumorais

Após filtrar os transcritos realizamos a análise de *two-class unpaired data* com SAM, categorizando as amostras em duas classes: Normais (1) e Tumorais (2). Como parâmetros utilizamos 1000 amostras de *bootstrap*, *K-Nearest Neighbors Imputer* como *Imputation Engine*, semente randômica e delta (tabela 13).

Tabela 13: Resultado da análise com SAM para as análises de *two-class unpaired data* e *multiclass data* conforme os parâmetros de cada um dos filtros. N refere-se ao número de amostras, V à variância e F ao Fold Change.

Conjunto	Filtro Global	Filtro Local	# Spots	# Spots Two-class	delta	FDR	# Spots Multi-class	delta	FDR
C1	-	-	3816	2	0,15	0,5	7	0,15	0,13
C2	V=0,8 N =11	-	2857	2	0,15	0,5	95	0,15	0,23
C18	V=0,8 N =11	V =0,7 N = 5	482	240	0,15	0,85	5	0,15	0,20
C34	-	V =0,7 N = 5	3354	2	0,15	0,5	70	0,15	0,30
C13	F =0,8 N =11	-	611	301	0,15	0,87	10	0,15	0,09
C24	F=0,8 N =11	F=0,3 N = 10	608	300	0,15	0,87	7	0,15	0,14
C39	-	F=0,3 N = 10	2307	2	0,15	0,5	1	0,15	0,92

Observando-se o índice de *False Discovery Rate* (FDR) mostrado na tabela 13 podemos ver que as análises de *two-class unpaired data* não levaram a bons resultados, não importando a presença do filtro. Observando a análise de *multiclass response*, em

que se considera três categorias (1: Amostras normais, 2: Amostras com Gleason Score entre 5 e 8, e 3: Amostras com Gleason 9-10) podemos ver o índice FDR apresenta-se menor e que o filtro por variância apresenta um melhor desempenho (leva a valores de FDR menores) que o filtro por *fold change*.

Os dados obtidos nos permitem concluir que as análises usando SAM considerando as duas classes, normal e tumoral não levaram a resultados significativos. Não podemos concluir, contudo, se tal resultado deve-se a aplicação do SAM para *feature selection* ou se deve-se ao fato das amostras consideradas normais estarem alteradas (isto é especulamos que o tecido considerado “normal” isolado da vizinhança do tecido tumoral não seja de fato equivalente a um tecido de fato normal).

Por ter a FDR diminuído na maioria das análises de *multiclass data* (considerando três classes: normal, Gleason 5-8 e Gleason 9-10) resolvemos realizar uma análise mais detalhada considerando somente as amostras tumorais.

3.3.1.2. Análise usando somente as amostras tumorais

O software usado para fazer o *clustering* hierárquico foi o SpotFire DecisionSite™ versão 15.3.486 e para a normalização prévia ao *clustering* usamos Z-score. Utilizamos *Ward's method* como método de *clustering* e como função de ordenação, o valor médio.

A Figura 20 mostra o *clustering* hierárquico obtido com os transcritos selecionados pelo SAM *two-class unpaired analysis* (Gleason 5 e 6 versus Gleason 9 e 10) para o conjunto de dados não filtrado. A Figura 21 mostra o *clustering* hierárquico obtido com os transcritos selecionados pelo SAM *multiclass* (Gleason 5 e 6, versus Gleason 7 e 8 versus Gleason 9 e 10), para o conjunto de dados não filtrado. Na Figura 22 vemos o *clustering* hierárquico obtido com os transcritos selecionados pelo SAM *two-class unpaired analysis* (Gleason 5 e 6 versus Gleason 9 e 10) para o conjunto de dados filtrado. Na Figura 23 está apresentado o *clustering* hierárquico obtido com os transcritos selecionados pelo SAM *multiclass* (Gleason 5 e 6, versus Gleason 7 e 8 versus Gleason 9 e 10), para o conjunto de dados filtrado. Em todos estes casos, observamos que uma separação visualmente interessante foi obtida.

A Figura 24 apresenta o *clustering* hierárquico obtido com os transcritos selecionados pelo SAM *two-class unpaired analysis* (Gleason 5 e 6 versus Gleason 9 e 10) para o conjunto de dados não filtrado, usando as amostras com Gleason 5, 6, 7, 8, 9

e 10. Podemos ver que algumas amostras com Gleason 7 distribuem-se entre as amostras com Gleason 9 e 10 e que outras entre as amostras com Gleason 5 e 6, o que pode ser um indicativo de pior ou melhor prognóstico, respectivamente.

Na Figura 25 podemos ver o cluster hierárquico obtido com os transcritos selecionados pelo SAM *two-class unpaired analysis* (Gleason 5 e 6 versus Gleason 9 e 10), para o conjunto de dados filtrado, usando as amostras com Gleason 5, 6, 7, 8, 9 e 10. Novamente podemos ver que algumas amostras com Gleason 7 distribuem-se entre as amostras com Gleason 9 e 10 e que outras entre as amostras com Gleason 5 e 6, o que pode ser um indicativo de pior ou melhor prognóstico, respectivamente.

Figura 20: *Clustering* hierárquico do conjunto de dados não filtrado para variação nas amostras tumorais usando os transcritos obtidos da análise com SAM *two-class unpaired data*, em que as categorias eram GS 5 e 6 (traço verde) x GS 9 e 10 (traço vermelho).

Figura 21: *Clustering* hierárquico do conjunto de dados não filtrado variação para nas amostras tumorais usando os transcritos obtidos da análise com SAM multiclass data, em que as categorias eram GS 5 e 6 (traço verde) x GS 7 e 8 (traço azul) x GS 9 e 10 (traço vermelho).

Figura 22: *Clustering* hierárquico do conjunto de dados filtrado para variação nas amostras tumorais usando os transcritos obtidos da análise com SAM two-class unpaired data, em que as categorias eram GS 5 e 6 (traço verde) x GS 9 e 10 (traço vermelho).

Figura 23: *Clustering* hierárquico do conjunto de dados filtrado para variação nas amostras tumorais usando os transcritos obtidos da análise com SAM multiclass data, em que as categorias eram GS 5 e 6 (traço verde) x GS 7 e 8 (traço azul) x GS 9 e 10 (traço vermelho).

Figura 24: *Clustering* hierárquico do conjunto de dados não filtrado para variação nas amostras tumorais usando os transcritos obtidos da análise com SAM two-class unpaired data e as amostras com GS 5 e 6 (traço verde); 7 e 8 (traço azul); 9 e 10 (traço vermelho).

Figura 25: *Clustering* hierárquico do conjunto de dados filtrado para variação nas amostras tumorais usando os transcritos obtidos da análise com SAM two-class unpaired data e as amostras com GS 5 e 6 (traço verde); 7 e 8 (traço azul); 9 e 10 (traço vermelho).

3.3.1.3. Validação do clustering

Exemplos de árvores consenso estão nas Figuras 26 e 27. Como mostrado na Figura 27, alguns ramos importantes, como o de 9.880, 6.206 e 5.466 se sustentam na maioria dos casos, o que reforça a hipótese de que tal separação das amostras tumorais não deve-se ao acaso, o que poderia indicar que algumas amostras com *Gleason Score* 7, por estarem no mesmo ramo que as amostras de *Gleason Score* 9 e 10, poderiam apresentar um pior prognóstico.

Figura 26: Árvore consenso para o conjunto de dados filtrado pela variação nas amostras tumorais usando os 49 transcritos apontados pelo SAM na análise de two-class unpaired data e as amostras com GS 5, 6, 9 e 10.

3.3.1.4. Discussão

A comparação do conjunto de 49 transcritos sugerido pela análise com SAM que descrevemos acima com o conjunto de 56 transcritos apontado em uma análise anterior (Reis *et al.*, 2004) , revelou uma coincidência de 10 transcritos (Figura 28). Ou seja 10 transcritos que selecionamos estão entre os 56 transcritos revelados na análise anterior. Além disso, desses 10, 4 foram validados experimentalmente (Reis *et al.*, 2004).

Figura 28: Intersecção entre os transcritos encontrados em análise prévia (Reis *et al.*, 2004) e transcritos encontrados pelo SAM *two-class unpaired data* para o conjunto filtrado para a variação dentre as amostras tumorais

Assim considerando a consistência de separação obtida entre os grupos de Gleason, que foi atestada pela validação do *clustering*, somada a intersecção dos dados com a análise anteriormente realizadas, podemos concluir o SAM parece ser um método promissor para encontrarmos potenciais marcadores de prognóstico.

É importante destacar, que entre os transcritos encontrados nesta análise bem como na análise anterior (Reis *et al.*, 2004) estão:

- PSA (*prostate-specific antigen*), atualmente já utilizado como marcador clínico no diagnóstico e prognóstico no câncer de próstata;
- RAB3B, um membro da família do oncogene *RAS*;
- RASSF1, um supressor tumoral.

Estas observações são um indicativo importante de que os transcritos apontados em nossa análise podem ser potenciais marcadores para o prognóstico do câncer de próstata. Contudo, a confirmação desta proposição irá requerer futura validação experimental e clínica.

3.3.2. Descoberta de marcadores para Carcinomas Renais de Célula Clara

3.3.2.1. Trios com genes diferencialmente expressos

A análise dos diferentes conjuntos possibilitou a identificação de alguns trios, sendo que apresentaremos alguns deles, nas Figuras 29 a 33.

Figura 29: Melhor trio considerando grupo híbrido de transcritos. Em vermelho temos os transcritos obtidos de amostras normais, em azul de tumorais.

Figura 30: Terceiro melhor trio considerando grupo híbrido de transcritos. Em vermelho temos os transcritos obtidos de amostras normais, em azul de tumorais.

Figura 31: Melhor trio considerando grupo de transcritos intrônicos. Em vermelho temos os transcritos obtidos de amostras normais, em azul de tumorais.

Figura 32: Segundo melhor trio considerando grupo de transcritos intrônicos. Em vermelho temos os transcritos obtidos de amostras normais, em azul de tumorais.

Figura 33: Pior trio considerando grupo híbrido de transcritos. Em vermelho temos os transcritos obtidos de amostras normais, em azul de tumorais.

Em alguns casos é possível visualizar um classificador linear que separe as amostras nas classes normais e tumorais. Porém, como podemos observar pelo pior trio exemplificado na Figura 33, nem sempre será possível usar um classificador linear para separar as classes.

Usando o critério de seleção descrito na Figura 16, calculamos da ordem de 10^9 trios. Apesar de termos obtido o melhor trio deste conjunto, sabemos que tal trio pode não vir a ser o melhor trio de todos, se considerarmos os 10^{12} possíveis trios de serem calculados, ou seja, é possível que o melhor trio que encontramos seja um classificador sub-ótimo, e não o ótimo, dado todo o espaço de classificadores possíveis de serem calculados considerando todos os transcritos da plataforma de *microarray* utilizada.

3.3.2.2. Trios com os genes mais expressos e também com o “busca e escolha”

Os valores de svd que obtivemos para cada um dos conjuntos estão dispostos na Figura abaixo. Na Figura 34 temos que os valores de svd para a A e B vão até 10, enquanto, em C vão até 12.8.

Figura 34: Valores de svd para todos os trios calculados. Eixo x: numero do trio. Eixo y: valor de svd do trio. Figura a) Valores de svd obtidos para os 50 genes mais expressos : maior svd 10. Figura B) Valores de svd obtidos para os 100 genes mais expressos : maior svd 10. Figura C) Valores de svd obtidos para os 500 genes mais expressos : maior svd 12.8.

Os melhores trios obtidos com cada uma das abordagens podem ser visualizados nas Figuras 35 a 37.

Figura 35: Melhor trio para os 50 mais expressos. Em vermelho estão as amostras normais, em azul as tumorais.

Figura 36: Melhor trio para os 100 mais expressos. Em vermelho estão as amostras normais e em azul as tumorais.

Figura 37: Melhor trio para os 500 mais expressos. Em vermelho estão as amostras normais, em azul as tumorais.

Optamos também por usar a implementação que o aluno Elier Broche Cristo fez para o pacote MaigesPack (Gomes et al., 2005; Esteves, 2007) também chamada de “busca e escolha”. Essa função parte da premissa que o melhor trio será composto por genes diferencialmente expressos. Então, após achar os genes mais diferencialmente expressos, calcula pares, medindo sua qualidade também por svd. Por último, tem-se a busca pelos melhores trios faz-se para os pares que apresentaram maiores valores de svd. Para tanto, precisamos utilizar os seguintes parâmetros: k e l . K corresponde ao número de trios que queremos visualizar no final da análise e l o número de genes diferencialmente expressos que serão usados para calcular o svd. Usando o parâmetro $k=100$ e $l=100$, obtivemos o trio presente na Figura 38 com valor de svd 13.936.

Figura 38: Visualização do melhor trio encontrado com a função “busca e escolha”. Em vermelho encontram-se as amostras normais. Em verde as tumorais.

Tais trios chegaram a valores de svd 13.936, o que sugere serem superiores aos encontrados apenas selecionando por intensidade média alta. Como medida de avaliação da robustez estamos usando CV (*cross-validation*). Assim, temos para os primeiros trios a tabela 14

Tabela 14: Valores de *Cross-validation* (CV) e svd para os melhores trios encontrados pelo “busca e escolha”.

Trio	CV	svd
1	100	13.936
2	100	13.93
3	100	13.795
4	100	13.634
5	100	13.525
6	100	13.466
7	100	13.422
8	100	13.238
9	100	13.237

Não conseguimos concluir a análise exaustiva usando o “busca e escolha” por questões de tempo hábil de execução.

3.3.2.3. Discussão

Nenhuma das abordagens que implementamos para a busca por trios para as amostras de Carcinoma Renal de Células Claras forneceu ótimos resultados. Essa dificuldade pode ter sido encontrada pela própria biologia das amostras, assim como não foi possível encontrar uma boa separação para as classes normais e tumorais em próstata. Ela também pode ser atribuída ao fato de, por limitações computacionais, termos que fazer uma seleção prévia do número de genes que serão usados no cálculo dos trios.

De qualquer forma, evidencia que nem sempre em uma análise de dados de *microarrays* conseguimos encontrar bons marcadores moleculares e/ou bons candidatos para validação.

3.3.3. Descoberta de marcadores para Sarcoma

3.3.3.1. A primeira etapa do projeto e os primeiros resultados

A colaboração para a análise dos dados de Sarcoma começou em maio de 2007. A carga dos dados foi feita usando o MaigesPack, então normalizada e sumarizada por amostras, dado que as amostras foram hibridizadas usando o desenho experimental de *dye swap*, conforme descrito na seção de procedimentos para a plataforma de *microarray* ACCamargo3 (seção 3.2.4). A análise dos dados de Sarcoma começou inicialmente sem seguir propriamente uma metodologia de gerenciamento. Inicialmente existiam as seguintes perguntas biológicas que iriam direcionar as análises:

Tabela 15. Perguntas biológicas iniciais

Perguntas Iniciais	Problema
P1: Quais os genes que distinguem entre os tipos de sarcoma?	Distinção entre sarcomas
P2: Quais os genes que podem dar alguma informação de grau?	Distinção de grau em sarcomas
P3: Quais são os genes preditores de metástase?	Predição de evolução - metástase
P4: Quais os genes envolvidos na distinção entre fibromatose e outros (pois fibromatose “nunca” metastiza)?	Predição de evolução - metástase
P5: É possível achar um conjunto de genes que avalie o tempo até a metástase x tempo de sobrevida?	Predição de evolução

Todos os dados resultados foram disponibilizados aos colaboradores através da interface do moodle. Os primeiros disponibilizados foram em junho de 2007. Assim, para essas buscas iniciais temos organização dos resultados conforme a Figura 39.

fibromatose_outros	2.4MB	25 June 2007, 02:35 PM
gfap	56KB	28 August 2007, 01:07 PM
maligno_X_benigno	1.7MB	25 June 2007, 03:14 PM
meta_naoMeta	1.7MB	25 June 2007, 01:42 PM
meta_naoMeta_semBenig	855KB	25 June 2007, 05:18 PM
porGrupos_grau	1.6MB	25 June 2007, 02:27 PM
porGrupos_grupos	1.8MB	25 June 2007, 02:28 PM
porGrupos_meta	1.7MB	25 June 2007, 02:28 PM
sinovial_outros	2.4MB	25 June 2007, 03:15 PM
tfb_v_tfm	2.4MB	25 June 2007, 03:16 PM
tmb_X_tmm	780KB	25 June 2007, 03:12 PM

Figura 39: Estrutura de diretórios do moodle para os resultados iniciais dos dados de Sarcoma

Dentro de cada diretório era então possível visualizar todos os resultados obtidos para a comparação em questão. A título de exemplo temos na Figura 40 os resultados das análises de genes diferencialmente expressos por Wilcoxon e teste-t, assim como o resultado do clustering hierárquico feito com os genes mais diferencialmente expressos usando SOM, como o resultado dos trios para essa comparação.

MAIGES ► Sarcoma ► Resources ► Results 25/06/2007 ► fibromatose_outros				Edit files	Update th
Name	Size	Modified
SOM-DE-fibroXoutros.png	10.6KB	25 June 2007, 01:12 PM
TFB_nao.jpg	60KB	25 June 2007, 02:15 PM
_DEfibroXoutros-name.html	444.9KB	25 June 2007, 01:18 PM
_DEfibroXoutros-semajuste-name.html	444.7KB	25 June 2007, 01:18 PM
_DEfibroXoutros-semajuste.html	393.1KB	25 June 2007, 01:18 PM
_DEfibroXoutros.html	393.3KB	25 June 2007, 01:18 PM
_Wilcox_fibromatose_tfb_Voutros_20070625_geneSymbol_comAjuste.html	308.6KB	25 June 2007, 01:16 PM
_Wilcox_fibromatose_tfb_Voutros_20070625_spotName_comAjuste.html	360.2KB	25 June 2007, 01:16 PM
fibromatose_tfb_Voutros_PorWilcox_KM_12Genes_nomeGene.png	13KB	25 June 2007, 01:13 PM
fibromatose_tfb_Voutros_PorWilcox_KM_12Genes_nomeNucl.png	28.7KB	25 June 2007, 01:13 PM
fibromatose_tfb_Voutros_PorWilcox_KM_12Genes_nomeUnigene.png	13.2KB	25 June 2007, 01:14 PM
genes_maisDifExpressos_tValor_usadoParaTrios.txt	1.5KB	25 June 2007, 02:13 PM
pares.txt	2.5KB	25 June 2007, 02:14 PM
trios.txt	3.1KB	25 June 2007, 02:14 PM			

Figura 40: Exemplo de resultados para uma análise disponibilizada no moodle. No caso trata-se da comparação TFB x outros.

Figura 42: Melhor trio obtido pelo método busca e escolha para as classes Fibromatoses x outras (*Cross Validation* = 99%, *svd*=15,697). Em verdes estão as fibromatoses. Em vermelho as outras amostras. Pode-se ver que uma amostra marcada em verde, como fibromatose, está bem misturada às vermelhas.

Figura 43: Melhor trio obtido pelo método busca e escolha para as classes Sinovial x outros (*Cross Validation* =100% e *svd*=23,293). Em verdes estão as sinoviais e em vermelho as outras amostras.

Após avaliar os resultados encontrados, a equipe então decidiu por focar nas perguntas P2, P3 e P4, que tratam das questões de se achar marcadores associados à metastização e agressividade local. Como iniciativa de resolver o problema da documentação das análises, documentação do código desenvolvido e organização dos resultados, propusemos a adaptação de eXtreme Programming (XP) ao projeto. Assim, temos que os resultados que serão apresentados a seguir foram gerados usando essa metodologia para gerenciar a análise.

3.3.3.2. A segunda etapa do projeto: XP + Fibromatoses como tema central

As iterações de XP desenvolvidas para a análise giraram em torno de questões conceituais, isso é, perguntas biológicas. A progressão das iterações pode ser visualizada na tabela 16. Uma nova anotação para a plataforma de *microarray* foi colocada assim como foi feita uma nova carga dos mesmos, uma vez que algumas correções nos dados foram necessárias. A classificação histológica de todas as amostras foi revista, sendo que duas mudaram de classe. As inconsistências nos dados referente a variáveis clínicas também foram resolvidas através da re Checagem dos prontuários dos pacientes. Novos testes de integridade e consistência foram feitos para os dados.

Tabela 16. Progressão das iterações em Sarcoma

Número Iteração	Tema
Zero	Anotação do dado, quanto as genes-plataforma de <i>microarray</i> e quanto as amostras. Carga dos dados no ambiente computacional e testes de consistência dos dados.
Um	Testes de qualidade dos dados de <i>microarray</i> . Normalização e sumarização dos dados.
Dois	Explorando as variáveis biológicas da etapa um do projeto: Fibromatose x outros e Sinovial x outros.
Três	Primeiras análises explorando a biologia das fibromatoses.
Quatro	Análises explorando fibromatoses, sarcomas e tumores benignos com agressividade, malignidade, grau e metástase. Análises explorando tetraspaninas.
Cinco	Análises mais profundas explorando fibromatoses, sarcomas e tumores benignos, com malignidade e agressividade. Análises explorando tetraspaninas.
Seis	Geração de elementos para o artigo submetido e artigos em preparação.

Além das perguntas girando em torno de Fibromatoses X Benignos X Malignos, tem-se também que os colaboradores da parte biológica tinham o interesse de se analisar Sarcomas sobre a perspectiva das tetraspaninas, que resultou em um artigo em preparação. Assim, essas análises foram incorporadas nas iterações quatro, cinco e seis.

No total foram criados 18 cartões referentes a perguntas biológicas, resumidos na tabela 17. Eles foram criados aos moldes do esquema apresentado na Figura 4 (Modelo usando XP). A título de exemplo de um cartão temos na Figura 44, o cartão 14 que junto com o cartão 15 (Figura 45) e cartões subsequentes levou ao resultados discutidos na seção 3.3.3.6, que exploram os marcadores de agressividade local.

Tabela 17. Cartões de XP para Sarcoma iterações 3 a 5

Número Cartão	Tema geral	Hipótese Biológica	Variáveis	Iteração
1		Por que fibromatose é agressiva localmente? O que a faz diferente dos outros benignos?	Origem: (TMB + TNB) x TFB	3
2		Quais os genes / módulos envolvidos na malignização?	Origem: TFB x TFM	3
3		O que faz com que as fibromatoses sejam agressivas localmente mas não malignas?	Origem: TFB x (SS + LPS + SP + TAM + TFM + TMM + TNM)	3
4		O que diferencia a agressividade local da metástase?	Origem + meta: TFB x ((SS + LPS + SP + TAM + TFM + TMM + TNM) & meta == sim)	3
5	Fibromatoses x Não - Metastáticos	O que diferencia a agressividade local de ficar maligno mas não metastizar?	Origem + meta: TFB x ((SS + LPS + SP + TAM + TFM + TMM + TNM) & meta == sim)	3
6	Agressividade local x Metastática	Quais os genes envolvidos em agressividade local mas não envolvidos na agressividade maligna-metastática?	Origem + metástase (vide cartão)	4
7	Agressividade local x Metastática	Quais os genes envolvidos em metástase mas não em agressividade local?	Origem + metástase (vide cartão)	4
8	Metástase	Quais os genes envolvidos com metástase?	Origem + metástase (vide cartão)	4
9	Grau x agressividade	Tumores que tem grau mais alto tendem a ser mais agressivos?	Origem + grau (vide cartão)	4
10	Técnica	Para as redes de relevância, quais os verdadeiros valores de p-valor?		4
11	Técnica	Visualização dos genes das redes de relevância com scatterplot para TFB x Fibromatose		4
12	Fibromatose x Malignos	Como são os boxplots dos genes dos módulos ativos para TFB x malignos	Origem	4
13	Fibromatose x	Para a comparação TFB x	Origem: TFB x TFM	4

	Malignos	TFM há 92 trios com CV = 100% e 92 pares com CV = 100% Pq em ambos há 92? Há um gene que está influenciando mais que outros? Ou seja, esses trios são compostos pelo mesmo par (só muda a 3ª componente?)?		
14	Agressividade local x Metastática	Quais os genes envolvidos em agressividade local mas não envolvidos na agressividade maligna-metastática?	Origem + metástase (vide cartão)	5
15	Agressividade local x Metastática	Quais os genes envolvidos em metástase mas não em agressividade local?	Origem + metástase (vide cartão)	5
16	Questão técnica	Quais os valores de fold e p-valor para os cartões 14 e 15?	Origem + metástase	5
17	Agressividade local x Metastática	Como se comportarão os marcadores encontrados nos cartões 14 e 15 para grau da amostra?	Origem, metástase e grau	5
18	Agressividade local x Metastática	Como se comportarão os marcadores encontrados nos cartões 14 e 15 para malignidade e metástase?	Origem, metástase	5

O resultado dessas análises compõem quatro manuscritos. O primeiro já submetido à publicação³ foca nos marcadores de agressividade local e marcadores de metástase para Fibromatoses, Sarcomas e Benignos, cujos resultados serão apresentados a seguir. Os demais manuscritos estão em fase de redação⁴. Um deles descreve GFAP (*Glial Fibrillary Acid Protein*) e sua utilização como marcador para Sarcomas. Outro discute o papel das Tetraspaninas em Sarcomas. Já o quarto manuscrito apresenta as

³ Cunha, I.W., Carvalho, K. C., Martins, W. K., Marques, S., Falzoni, R., Simões, A.C.Q., Fahham, L., Neves, E. J., Soares, F. A. & Reis, L. F. L. Identification of genes associated with local aggressiveness and metastatic behavior in soft tissue sarcomas. Submetido à publicação. Consta dos anexos inclusos no CD-ROM complementar a tese.

⁴ - Santos, G.C, Carvalho, K. C., Falzoni, R., Simões, A.C.Q., Reis, L. F. L., Soares, F. A. & Cunha, I. W. Glial Fibrillary Acidic Protein (GFAP) is an Intermediate Filament Present in Several Tumors Types with cartilaginous differentiation. Em preparação.

- Carvalho, K.C, Cunha, I.W., Maschietto, M., Simões A.C., Reis L.F.L. & Soares F.A. Tetraspanins expression and their possible involvement in malignant potential in soft- tissue tumors. Em preparação.
- Cunha I.W., Carvalho, KC, Martins, W.K., Marques, S.M., Fahham, L.V., Simões, A.C., Cristo, E.B., Neves, E.J., Soares, F.A & Reis, L.F.L. O perfil de expressão gênica separa os diversos tipos de sarcomas: análise de genes diferencialmente expressos em tumores mesenquimais de partes moles através da técnica de cDNA microarray. Em preparação.

Constam dos anexos inclusos no CD-ROM complementar a tese.

assinaturas moleculares de Sarcomas, cujos resultados também serão apresentados adiante.

Figura 44: Cartão 14 da análise de Sarcoma.

Figura 45: Cartão 15 da análise de Sarcoma.

3.3.3.3. Resultados para a comparação Fibromatoses x Benignos

Para explorarmos as diferenças entre fibromatoses e as amostras benignas, conforme proposto no cartão 1, foram feitos teste-t com bootstrap e teste de Wilcoxon. Também foi feita a análise de módulos ativos, busca por trios usando o de método busca e escolha e redes de relevância dentro dos grupos funcionais anotados, cujo resultado pode ser visualizado na Figura 46.

Figura 46: Rede de Relevância para as classes fibromatoses (tfb) e benignos (TnbTnb). Verde significa correlação negativa, vermelho positiva (para o grafo da esquerda e do centro) No grafo da direita temos a representação da significância da diferença de correlação entre os genes das duas classes.e benignos para o grupo “Extracellular region GO0005576”.

Na rede de relevância da Figura 46 podemos ver no grafo à esquerda as correlações entre as moléculas do grupo “Extracellular region GO0005576” para a classe de fibromatoses. Em vermelho estão as correlações positivas, em verdes as negativas. No grafo do meio temos as correlações entre as moléculas do grupo “Extracellular region GO0005576” para a classe dos benignos, sendo que o mesmo esquema de cores se aplica. Já no grafo da direita temos somente as diferenças de correlação com p-valor menor que 0,005 representadas pelas linhas, sendo que quanto mais grossa a linha mais significativa é a diferença de correlação encontrada entre as classes fibromatoses e benignos. Assim, com a rede de relevância pode-se visualizar a dinâmica das mudanças para moléculas do grupo “Extracellular region GO0005576” para as classes fibromatoses e benignas.

A análise de módulos ativos para as classes fibromatoses x benignos, Figura 47, considerou como significativa as mudanças para o módulo “assinatura de metástase”, sendo que o perfil apresentado é estar menos ativo para as amostras de fibromatose do que para os benignos.

Figura 47: Módulos Ativos para as classes fibromatoses x benignos.

3.3.3.4. Resultados para a comparação Fibromatoses x Malignos

A análise de módulos ativos agora considerando as classes fibromatoses e malignos, Figura 48, mostrou que o grupo “assinatura de metástase” aparece ativado para os malignos e inativo para as fibromatoses, resultado consistente com o encontrado para o mesmo tipo de análise com as classes fibromatoses e benignos.

Figura 48: Módulos ativos para as classes fibromatoses e malignos.

3.3.3.6. Marcadores de agressividade local e de metástase

A busca por marcadores de agressividade local e de metástase se fez usando as a separação de classes fibromatoses, benignos e sarcomas. Inicialmente considerando todos os genes da plataforma ACCamargo3 de *microarray* as amostras biológicas se separaram conforme o clustering da Figura 50. As amostras benignas estão demarcadas como BMT, as Fibromatoses como DTF e as malignas como MMT. Dado que as fibromatoses são muito agressivas localmente, mas raramente apresentam metástase e que os sarcomas possuem potencial metastático, tem-se que as comparações apresentadas nos cartões 14 e 15 (Figuras 44 e 45), permitem a definição de tais marcadores. Para ambos são feitas as mesmas comparações de genes diferencialmente expressos: Fibromatoses x benignos (comparação A), fibromatoses x malignos (comparação B) e malignos x benignos (comparação C). Mais especificamente, o cartão 14 busca marcadores de agressividade local e o cartão 15 marcadores de metástase. A tabela com os p-valores de tais comparações está disponível no endereço eletrônico: <http://www.maiges.org/sarcomaFibromatosis/>. Alguns resultados dessas mesmas comparações com outras técnicas foram apresentados nos itens anteriores desta tese.

Figura 50: Disposição das amostras do conjunto Sarcoma usando todos os genes da plataforma ACCamargo3. As amostras benignas estão demarcadas como BMT, as Fibromatoses como DTF e as malignas como MMT.

Figura 51: Comparações entre Fibromatoses, Benignos e Malignos para o conjunto Sarcoma, presentes nos cartões 14 e 15.

Como resultado da intersecção dos cartões 14 e 15 temos dois conjuntos de 6 genes, que aparecem como os mais diferencialmente expressos para a comparação C. Porém há uma dicotomia para esses genes: os genes que aparecem como diferencialmente expressos para a comparação A não aparecem como diferencialmente expressos para a comparação B, e vice versa, como pode ser visualizado na Figura 52. Assim, 6 genes desses genes foram definidos como marcadores de agressividade local. E os outros 6 genes como marcadores de metástase.

Figura 52: P-valores ($-\log_{10}$) das comparações A, B e C. Os genes diferenciados representam os marcadores encontrados. Os genes em vermelho são os encontrados para agressividade local. Os em verde para metástase.

3.3.3.8. Marcadores que usam a informação de grupos para separar as amostras

Esses resultados também fazem parte do manuscrito Cunha et al. (artigo em preparação) acima mencionado, que trata dos perfis gênicos e potenciais marcadores que encontramos para os sarcomas. Usando a informação de grupos gênicos (podendo este ser uma via, uma categoria funcional oriunda do Gene Ontology (Consortium, 2000), por exemplo, ou uma categoria de moléculas ortólogas, oriunda do KEGG (Kanehisa *et al.*, 2008), optamos por fazer análise de clustering. Ou seja, não selecionamos genes diferencialmente expressos para o cluster, mas sim, elementos do grupo. Logo, poderíamos considerar essa abordagem não supervisionada, pois não definimos que amostra é o que.

Como exemplo de resultados temos a Figura 54, em que não conseguimos separar as amostras usando os genes presentes no grupo de Fosforilação Oxidativa retirado do KEGG.

Figura 54: Amostras agrupadas por K-médias usando os genes do grupo de Fosforilação Oxidativa, retirado do KEGG. Em colorido, na base do *heatmap*, temos cada um dos 6 “tipos” de sarcoma representado.

3.3.3.9. Discussão das análises de Sarcoma

Diversas estratégias foram usadas para analisar o conjunto de dados Sarcoma: busca por genes diferencialmente expressos usando teste t, teste de Wilcoxon e teste t com bootstrap, análises de módulos ativos e análises de redes de relevância. As buscas por genes diferencialmente expressos e as relações entre essas buscas, como por exemplo a busca por marcadores que envolveram as comparações entre as amostras malignas, benignas e tumorais (seções 3.3.3.3 a 3.3.3.6), permitiram que encontrássemos potenciais marcadores individuais que parecem ser capaz de classificar as amostras quanto a agressividade local e metástase do tumor (Figura 52). As análises por módulos ativos também permitiram que tivéssemos informações quanto a módulos de genes envolvidos no processo de agressividade local e metástase dos sarcomas, tendo assim uma visão global dos processos biológicos envolvidos. Mais especificamente pudemos verificar o comportamento de módulos que estão sabidamente envolvidos em câncer (Hanahan & Weinberg, 2000), como matriz extracelular, adesão celular e comunicação celular nas comparações entre as amostras benignas, fibromatoses e malignas. Já as análises de redes de relevância feitas para “Extracellular region GO0005576” e para as tetraspaninas (resultado não mostrado) permitiram que tivéssemos uma visão da dinâmica da mudança das moléculas nas classes em questão.

3.3.4. Descoberta de marcadores para Carcinomas Epidermóides de Cabeça e Pescoço

3.3.4.1. XP e as análises

A análise dos dados de Cabeça e Pescoço foi a segunda análise em que XP foi usada. Logo, desde o início, cartões foram criados para representar a evolução do projeto. O projeto ainda está em andamento. No total, até janeiro de 2009, foram criadas 6 iterações, cujos temas centrais estão resumidos na tabela 18.

Tabela 18. Progressão das Iterações em Cabeça e Pescoço

Número Iteração	Tema principal
Zero	Carga dos dados, verificações de qualidade dos dados e dependência entre as classes de amostras. Criação do meta-objeto.
Um	Topologias e esvaziamento
Dois	Esvaziamento, recidivas e tamanho
Três	Tamanho e esvaziamento
Quatro	Diferenciação, invasão vascular, invasão perineural
Cinco	Tabaco, álcool, extensão extracapsular
Seis	Esvaziamento resumido, margens, recidiva, rádio adjuvante, topologia e invasão
Sete	Maiores detalhes sobre esvaziamento resumido, margens, recidiva, rádio adjuvante, topologia e invasão

Os cartões que foram criados para Cabeça e Pescoço usando XP estão resumidos na tabela 19.

Tabela 19. Cartões de Cabeça e Pescoço usando XP

Número Cartão	Hipótese Biológica	Variáveis	Iteração
1	Como os genes se comportam perante as 5 topologias?	Topologia	1
2	Quais são os marcadores de recidiva para linfonodo negativo (esvaziamento)?	Esvaziamento resumido e recidivas totais resumido. Amostras Boca, oro, hipo e lari	1
3	Quais os marcadores para linfonodo comprometido	Esvaziamento resumido. Amostras Boca, oro, hipo e lari	1
4	Quais são os marcadores de bom prognóstico p/ linfonodos comprometidos-positivos?	Esvaziamento resumido e recidivas totais resumido. Amostras Boca, oro, hipo e lari	1
5	Quais os marcadores de recidiva sem levar em consideração esvaziamento?	Recidivas totais resumido. Amostras Boca, oro, hipo e lari	1
6	Quais os marcadores de recidiva para pacientes esvaziados?	Esvaziamento resumido e recidivas totais resumido. Amostras Boca, oro, hipo e lari	1
7	Quais os marcadores de prognóstico p/ linfonodos negativos considerando radioterapia e margens?	Esvaziamento resumido e recidivas totais resumido, margens livres, radioterapia. Amostras Boca, oro, hipo e lari	2

8	Uma vez que BOCA e ORO são mais homogêneas entre si do que as outras topologias, quais marcadores de linfonodo negativos x linfonodos positivos só para BOCA e ORO?.	Esvaziamento resumido. Amostras Boca, oro.	2
9	Quais são os marcadores de recidiva considerando linfonodos positivos com rádio e margens livres?	Esvaziamento resumido e recidivas totais resumido, margens livres, radioterapia. Amostras Boca, oro, hipo e lari.	2
10	Uma vez que BOCA e ORO são mais homogêneas entre si do que as outras topologias, quais são os marcadores de recidiva considerando linfonodos positivos com rádio e margens livres?	Esvaziamento resumido e recidivas totais resumido, margens livres, radioterapia. Amostras Boca, oro.	2
11	Quais são os marcadores de recidiva considerando linfonodos esvaziados sem rádio com margens livres?	Esvaziamento resumido e recidivas totais resumido, margens livres, radioterapia. Amostras Boca, oro, hipo e lari.	2
12	Quais são os marcadores de recidiva considerando linfonodos esvaziados, com rádio e com margens livres?	Esvaziamento resumido e recidivas totais resumido, margens livres, radioterapia. Amostras Boca, oro, hipo e lari	2
13	Quais são os marcadores de recidiva considerando linfonodos esvaziados, com rádio e com margens livres e somente para amostras ORO e BOCA (mais homogêneas)?	Esvaziamento resumido e recidivas totais resumido, margens livres, radioterapia. Amostras Boca, oro.	2
14A	Avaliar a estabilidade das amostras frente a um conjunto de classificadores	Desenvolvimento de novo recurso.	2
14B	São muitos trios (por grupos). Tem como gerar um listão para cada cartão com todos os valores dos trios?	Desenvolvimento de novo recurso.	2
14C	Ma vez que a grossura das linhas não são comparáveis entre os grafos das redes de relevância, transformar os grafos em tabelas.	Desenvolvimento de novo recurso.	2
15	Como as topologias se dispersam considerando todos os genes (sem os controles)?	Por correlação. Por distância euclidiana (completa).	2
16	Análise de trios usando distribuição das amostras do cartão 8 e 9, fazendo trios por busca e escolha e exaustivo, usando lista de genes da planilha da Anna (união de genes que vieram das análises de módulos ativos)	Esvaziamento resumido e recidivas totais resumido, margens livres, radioterapia. Amostras Boca, oro, hipo e lari.	2
17	Há quatro classificações de tamanho para o tumor. Eu gostaria de ver os genes que vão crescendo (decrecendo) conforme o tumor cresce.	Tamanho do tumor. Amostras Boca, oro, hipo e lari.	3
18	Há quatro classificações de tamanho para o tumor. Eu gostaria de ver os genes que vão crescendo (decrecendo) conforme o tumor cresce só considerando BOCA e ORO.	Tamanho do tumor. Amostras Boca, oro.	3
19	Quais são os marcadores para tamanho?	Tamanho do tumor. Amostras Boca, oro, hipo e lari.	3

20	Quais são os marcadores para tamanho considerando só BOCA e ORO?	Tamanho do tumor. Amostras Boca, oro.	3
21	O tumor já nasce agressivo ou ele se desenvolve e ganha agressividade? A idéia é avaliar tumor em estágio inicial x tumor em estágio avançado. Eu consigo separar bem as amostras só pensando em [(T1 ou T2) + linfonodo negativo] x [(T3 ou T4) + linfonodo positivo]?	Tamanho do tumor e esvaziamento resumido. Amostras Boca, oro, hipo e lari.	3
22	O tumor já nasce agressivo ou ele se desenvolve e ganha agressividade? A idéia é avaliar tumor em estágio inicial x tumor em estágio avançado. Eu consigo separar bem as amostras só pensando em [(T1 ou T2) + linfonodo negativo] x [(T3 ou T4) + linfonodo positivo] usando só BOCA e ORO?	Tamanho do tumor e esvaziamento resumido. Amostras Boca, oro.	3
23	Quais são as diferenças biológicas por trás da diferenciação?	Diferenciação. Amostras Boca, oro, hipo e lari.	4
24	Quais são as diferenças biológicas por trás da invasão vascular?	Invasão vascular. Amostras Boca, oro, hipo e lari.	4
25	Quais são as diferenças biológicas por trás da invasão vascular só para BOCA e ORO?	Invasão vascular. Amostras Boca, oro.	4
26	Quais as diferenças biológicas por trás da invasão perineural?	Invasão perineural. Amostras Boca, oro, hipo e lari.	4
27	Quais as diferenças biológicas por trás da invasão perineural só para BOCA e ORO?	Invasão perineural. Amostras Boca, oro.	4
28	Considerando os resultado do clustering hierárquico encontrado para o conjunto Carcinomas Epidermóides, qual é a diferença responsável pela segmentação entre os grupos de cabeça e pescoço CP1 e CP2?	Topologia.	4
31	Será que teremos um ganho analítico se fizermos uma mistura de variáveis com pouca interferência clínica?	Invasão, Diferenciação, linfonodo e tamanho. Amostras Boca, oro, hipo e lari.	4
32	Quais são os marcadores para álcool?	Álcool e parou de beber. Amostras Boca, oro, hipo e lari.	5
33	Quais os marcadores para tabaco?	Tabaco e parou de fumar. Amostras Boca, oro, hipo e lari.	5
34	Quais são os marcadores para álcool e tabaco?	Álcool, parou de beber, tabaco e parou de fumar. Amostras Boca, oro, hipo e lari.	5
35	Quais são os marcadores para extravasamento capsular?	Extensão extracapsular resumido. Amostras Boca, oro, hipo e lari.	5
36	Quais são os genes envolvidos com recidiva local-regional mas não com metástase?	Recidiva local, recidiva regional, metástase à distância e radio adjuvante	6
37	Quais são as alterações responsáveis pela metástase à distância? Quero comparar uma evolução benigna da pior evolução.	Recidivas totais resumido e rádio adjuvante, metástase à distância.	6
38	Quais as alterações responsáveis pela existência do tumor fora do sítio primário?	Esvaziamento resumido, rádio adjuvante, recidivas totais	7

	Quero olhar a colonização de outros sítios.	resumido.	
39	Tentativa de separar os mais indolentes dos mais horríveis. Será possível traçar uma assinatura de morte?	Status resumido	6
40	Será possível encontrar genes associados a resistência a drogas? Drogas: cetuximab, cisplatina, Paclitaxel, Radioterapia	Radioterapia adjuvante, status resumido	6
41	Qual o valor dos folds para todos os genes diferencialmente expressos dos cartões 8, 9 e 12?	Esvaziamento resumido e recidivas totais resumido, margens livres, radioterapia.	6
42	Classificar os genes em “que variam muito” e “variavam menos” para as topologias.	Topologia	6
43	Fazer classificadores com 5 genes específicos (selecionados do cartão 24). (trios associados a invasão)	Invasão vascular. Amostras Boca, oro, hipo e lari	6
44	Considerando os genes validados por real time, como ficam as árvores de decisão?	Esvaziamento resumido, recidiva, rádio adjuvante. Amostras Boca, oro, hipo e lari	6
45	Como as variáveis selecionadas se dispõem considerando um clustering hierárquico com todos os genes (menos controles)?	Topografia, tamanho, esvaziamento resumido, diferenciação, recidivas totais resumido, álcool, parou de beber, tabaco, parou de beber.	7
46	Como as amostras se organizam considerando todos os genes (sem os controles)?	Topografia	6
47	Considerando somente os genes que não variam muito, encontrados no cartão 42, como as topologias se organizam?	Topologia	6
48	Como fica um heatmap com todas as amostras e todos os genes (sem os controles)?	Topologia. Amostras Boca, oro, hipo, lari e cnsp.	7
49	Análise que nem a do cartão 12 (recidiva), mas usando os genes do grupo: Rickman 2008 + Cromer 2004 + Giri 2006. Quais são os marcadores de recidiva considerando linfonodos esvaziados, com rádio e com margens livres?	Esvaziamento resumido e recidivas totais resumido, margens livres, radioterapia. Amostras Boca, oro, hipo e lari	7
50	Análise que nem a do cartão 12 (recidiva), mas usando os genes do grupo: Chung 2004. Quais são os marcadores de recidiva considerando linfonodos esvaziados, com rádio e com margens livres?	Esvaziamento resumido e recidivas totais resumido, margens livres, radioterapia. Amostras Boca, oro, hipo e lari	7
51	Análise que nem a do cartão 8 (recidiva), mas usando os genes do grupo: Nagata 2003, Schmalbach 2004, Waner 2004, Roepman 2005 e O'Donnel 2005. Uma vez que BOCA e ORO são mais homogêneas entre si do que as outras topologias, quais marcadores de linfonodo negativos x linfonodos positivos só para BOCA e ORO?.	Esvaziamento resumido. Amostras Boca, oro	.7

Os resultados das análises foram disponibilizados pelo moodle, conforme Figura 55.

Figura 55: Estrutura de diretórios do projeto Cabeça e Pescoço. As iterações de XP aparecem demarcadas.

Os resultados para uma iteração ficam organizados dentro do diretório respectivo, como pode ser visto para a iteração 2, na Figura 56.

Figura 56: Resultados da iteração 2 para Cabeça e Pescoço.

Como uma das questões importantes para esse conjunto de dados é a de que até que ponto as topologias são diferentes entre si e devem ser tratadas de formas diferentes, pode-se observar que houve uma seção no projeto quanto às análises/cartões, pois muitas perguntas biológicas eram feitas para as amostras boca + oro + hipo + lari e então repetida somente para boca + oro. Os resultados eram então comparados.

3.3.4.2. A exclusão de CNSP de muitas análises e tabelas de contingência

As análises de contingência por Fisher's Exact Test e Pearson's Chi-squared test, tabela 20, evidenciaram que certo cuidado deve ser tomado para as análises com as variáveis esvaziamentoscoring e GrauDiferenciaçãoscoring, se as cinco topologias forem levadas em consideração. Esse resultado era esperado para a variável esvaziamento, uma vez que há certa dificuldade clínica de se esvaziar a cadeia de linfonodos para as amostras de cavidade de seios paranasais. Porém, não era esperada para grau de diferenciação.

Tabela 20. Valores dos testes Fisher e Pearson

Variável	Fisher's Exact Test (p-valor)	Pearson's Chi-squared test (p-valor)
RecidivaLocal	0.9978	0.9904
recidivastotaisscore	0.5981	0.6605
tempo1arecscore	0.705	0.5538
MetastaseaDistância	0.4611	0.5879
GrauDiferenciacaoscore	0.003016	0.00794
invasaoperineuralscore	0.03229	0.08886
invasaovascularscore	0.08843	0.1954
extensaoextracapsularscore	0.3437	0.2578
esvaziamentoscore	9.967e-05	4.795e-07
idadenacirurgiascore	0.5665	0.4215
Alcoolscore	0.02881	0.03316
Tabaco	0.08507	0.0949

As tabelas de contingência para as variáveis esvaziamentoscore e GrauDiferenciacaoscore, estão apresentadas a seguir (tabelas 21 e 22).

Tabela 21. Tabela de contingência para esvaziamento

	Boca	CNSP	Hipo	Lari	Oro
Não esvaziado	8	6	0	1	1
Linfonodo negativo	13	0	1	9	7
Linfonodo positivo	34	1	5	8	27

Tabela 22. Tabela de contingência para Grau de diferenciação

	Boca	CNSP	Hipo	Lari	Oro
Não informado	1	0	0	1	1
Bem diferenciado	33	2	1	8	10
Moderadamente diferenciado	21	3	3	9	20
Pouco Diferenciado	0	2	2	0	4

A distribuição da topologia com relação a outras variáveis também pode ser vista nas Figuras de 57 a 63, em que se tem a distribuição de álcool, tabaco, esvaziamento com o valor positivo, extensão extracapsular positiva, grau de diferenciação pouco diferenciado, invasão perineural positiva, invasão vascular positiva, metástase a distância.

Figura 57: Porcentagem das amostras para cada topologia que bebe álcool ou fuma

Figura 58: Porcentagem de amostras de cada topologia que aparece com esvaziamento positivo.

Figura 59: Porcentagem de amostras de cada topologia que aparece com extensão extracapsular positiva.

Avaliação da relação grau de diferenciação e localização do tumor

Figura 60: Porcentagem de amostras de cada topologia que aparece com grau de diferenciação pouco diferenciado.

Avaliação da relação invasão perineural e localização do tumor

Figura 61: Porcentagem de amostras de cada topologia que aparece com invasão perineural positiva.

Avaliação da relação invasão vascular e localização do tumor

Figura 62: Porcentagem de amostras de cada topologia que aparece com invasão vascular positiva.

Avaliação da relação metástase a distância e localização do tumor

Figura 63: Porcentagem de amostras de cada topologia que aparece com metástase a distância.

Devido ao fato de CNSP aparecer sistematicamente com um perfil diferenciado das outras amostras para as variáveis de interesse, tem-se que tal amostra foi excluída de muitas análises por acreditar que traria mais ruído. Também pode-se concluir que deve tomar certo cuidado na interpretação dos resultados das análises que envolvem as variáveis esvaziamento, grau de diferenciação, álcool e tabaco.

3.3.4.3. Análises de módulos funcionais relacionados a metástase linfonodal

Diversas análises de módulos funcionais (Segal *et al.*, 2004; Gomes *et al.*, 2005) que envolveram a questão de metástase linfonodal foram feitas, conforme a lista de cartões na tabela 19, porém algumas revelaram sistematicamente o aparecimento de alguns genes como significativos, que estão sendo validados como potenciais marcadores. Os resultados dessas análises estão apresentados nas tabelas 23 a 26.

Tabela 23. Módulos funcionais alterados em amostras de CECP de pacientes que apresentam metástases linfonodais

Módulo	Gene	p-valor
Via de calicreínas	KLK13	4,0E-15
	KLK7	2,8E-5
	KLK6	1,5E-2
	SERPINA3	1,7E-2
Via de calicreínas	KLK13	4,0E-15
	KLK7	2,8E-5
	KLK6	1,5E-2
	SERPINA3	1,7E-2
Tabela periódica Whitfield 2 (ciclo celular)	PCNA	1,6E-2
<i>Protein binding</i> (GO5515)	FOSB	4,1E-31
	KLK6	5,7E-18
	AREG	1,3E-11
	CCL20	4,0E-8
	EFNB2	9,1E-8
	IL1B	7,5E-5
	MYC	2,0E-04
	MCM2	2,0E-4
	INHBA	7,0E-3
	EGFR	1,4E-2
	AKT1	2,1E-2
	SPHK1	9,0E-9
Regulação de apoptose (GO42981)	INHBA	9,0E-9
	AKT1	9,7E-3

Tabela 24. Módulos funcionais alterados em amostras de CECP pT1 ou pT2 que não apresentam metástases linfonodais.

Módulo	gene	p-valor
Angiogênese (GO1525)	SPHK1	2,3E-3
Migração do leucócito transendotelial (KEGG)	MMP9	5,8E-5
	ACTN1	1,0E-3
	PIK3R3	3,5E-2
Regulação de crescimento celular (GO1558)	SPHK1	2,3E-3
Via das calicreínas	KLK13	1,2E-10
	KLK7	6,2E-4
	KLK6	3,3E-2
Metabolismo de nicotinato e nicotinamida (KEGG)	NT5E	4,5E-4

Tabela 25. Módulos funcionais alterados em amostras de CECP de pacientes que apresentam recidivas após tratamento com cirurgia, esvaziamento cervical e radioterapia adjuvante.

Módulo	Gene	p-valor
Biossíntese de diterpenóides (KEGG)	EGLN3	1,9E-3
Sinalização célula-célula (GO7267)	IL1F9	2,4E-21
	INHBA	1,4E-15
	AREG	4,3E-15
	BST2	1,0E-10
	KLK6	6,1E-4
	CCL20	4,1E-3
Sinalização célula-célula (GO7267)	IL1F9	2,4E-21
	INHBA	1,4E-15
	AREG	4,3E-15
	BST2	1,0E-10
	KLK6	6,1E-4
	CCL20	4,1E-3

Tabela 25. Módulos funcionais alterados em amostras de CECP de pacientes que apresentam metástases linfonodais e recidivas após tratamento com cirurgia e radioterapia adjuvante.

Módulo	Gene	p-valor
Sinalização célula-célula (GO7267)	IL1F9	4,1E-20
	AREG	5,5E-15
	INHBA	1,5E-12
	BST2	7,6E-10
	CCL20	2,1E-3
	KLK6	8,9E-3
Sinalização célula-célula (GO7267)	IL1F9	4,1E-20
	AREG	5,5E-15
	INHBA	1,5E-12
	BST2	7,6E-10
	CCL20	2,1E-3
	KLK6	8,9E-3
Região extracelular (GO5576)	INHBA	6,6E-15
	POSTN	3,0E-9
	KLK13	1,9E-6
	AREG	2,4E-6
	MMP13	7,3E-4
	KLK6	1,6E-2
Assinatura de metástase	COL6A3	5,5E-6
	COL6A3	1,75E-3
	PLAU	1,9E-3
Via de sinalização de TGF- β (KEGG)	INHBA	9,1E-11

Diversos genes apareceram recorrentemente como significativos nas análises de módulos ativos para as análises de metástase linfonodal. Esses genes recorrentes forem selecionados para validação experimental. Foi interessante verificar que diversos processos já conhecidos como muito envolvidos em câncer apareceram como módulos ativos ou inativos nas comparações quanto à metástase linfonodal, a citar apoptose, angiogênese e sinalização célula – célula.

3.3.4.4. Árvores de decisão para os genes em validação

Realizamos uma análise com árvores de decisão usando o método CART (Breiman *et al.*, 1984) com o intuito de simular como os sete genes em validação particionam o espaço de classificação de amostras nas três categorias:

* linfonodo positivo, recidiva positiva, margens livres e rádio terapia adjuvante positiva

* linfonodo positivo, recidiva negativa, margens livres e rádio terapia adjuvante positiva

* sem informação.

A árvore de decisão usando o CART pode ser vista na Figura 64 abaixo.

Figura 64: Árvore de decisão montada com os 7 genes em validação. Podemos ver que 5 dos sete genes foram selecionados para particionar o espaço.

A Figura 64 mostra que 5 dos sete genes foram selecionados para particionar o espaço. Em cima de cada nó está o nome do gene e qual o critério de decisão que deveria ser tomado caso a árvore fosse usada, por exemplo, na clínica médica como classificador. Ou seja, se o gene EGLN3 apresentar valor de expressão (W) maior que 0,8904 deve-se então olhar o valor de W para o gene IL1F9. Se esse apresentar valor de expressão maior que 0,98, então deve-se classificar a amostra como não informativa, mas se o valor for menor ou igual a 0,98, então deve-se olhar o valor do gene KLK6. Se o último apresentar valor menor que -0,8392, então a amostra será classificada como linfonodo positiva, recidiva negativa, margens livres e radioterapia positiva. Caso contrário será classificada como não informativa. A interpretação do outro ramo da árvore faz-se de forma similar. É importante ressaltar o quanto dependente a separação

das classes é dos valores de expressão encontrados, sendo esse um dos pontos fracos dessa metodologia.

3.3.4.5. Discussão das análises de Carcinomas Epidermóides de Cabeça e Pescoço

Não foi possível definir assinaturas de genes para metástase linfonodal assim como Roepman fez (Roepman *et al.*, 2005; Roepman *et al.*, 2006), porém pudemos encontrar certa consistência no perfil dos módulos de genes envolvidos em metástase linfonodal.

As análises de cabeça e pescoço apontaram para potenciais marcadores de metástase linfonodal, que foram selecionados por meio de análises de módulos ativos. Essa abordagem fez-se diferente da normalmente utilizada de se eleger genes diferencialmente expressos para validação. Essa opção foi feita uma vez que muitas das comparações de genes diferencialmente expressos não mostraram resultados muito significativos, principalmente quando os p-valores eram corrigidos (resultados não mostrados) (Dudoit & Speed, 1999). Esses genes eleitos estão em processo de validação experimental. A análise com árvores de decisão levou a uma proposta de árvore para os 7 genes em validação experimental. Uma das críticas a essa metodologia é a de que esse método de classificação é muito sensível aos valores de expressão encontrados, ou seja, caso a árvore fosse usada em clínica, um paciente que apresentar valores limítrofes para um dos genes pode vir a ter a sua classificação errada, pois na verdade o valor encontrado pode estar errado devido ao erro experimental.

As análises de módulos ativos, além de terem apontado para o papel de módulos já sabidamente envolvidos em câncer como angiogênese, sinalização célula-célula e região extracelular, também apontaram o papel de outros módulos na metástase linfonodal, como por exemplo as calicreínas (Hanahan & Weinberg, 2000).

3.3.5. Descoberta de marcadores para Carcinomas Epidermóides

3.3.5.1. XP e as análises

Até o presente momento somente três iterações foram feitas para o conjunto de dados Carcinomas Epidermóides, conforme a tabela 27.

Tabela 27. Iterações de XP para Carcinomas Epidermóides

Número da Iteração	Tema Principal
Zero	Carga dos dados e verificação de qualidade
Um	Análises iniciais quanto a topologias
Dois	Comparação entre grupo 1 e 2

Os resultados apontados pela iteração 1 levaram a hipóteses que estão sendo mais profundamente exploradas na iteração 2. Os cartões desenvolvidos estão apresentados na tabela 28.

Tabela 28. Cartões para Carcinomas Epidermóides usando XP

Número Cartão	Pergunta Biológica	Variáveis	Iteração
1	Como as amostras se dispõem conforme a topologia? O que prevalece: a topologia ou o fato de serem todos carcinomas epidermóides?	Topologia	1
2	Os genes que são diferencialmente expressos entre Boca-Normal e Boca-Tumoral são os mesmos entre Boca-Normal e Oro-Tumoral e ainda entre Boca-Normal e Hipo-Tumoral e também entre Boca-Normal e Lari-Tumoral e por último entre Boca-Normal e CNSP-Tumoral?	Topologia	1
3	O que diferencia o Grupo 1 do Grupo 2	Topologia	2

3.3.5.2. A questão das topologias

Para o cartão 1 fizemos uma análise de *clustering* hierárquico (Quackenbush, 2002) que revelou que as amostras de útero, pulmão e esôfago não se misturavam com

as de cabeça e pescoço. Esse resultado foi recorrente com várias medidas (resultado não mostrado). Ele também mostrou uma divisão para as amostras de cabeça e pescoço, o que chamamos de grupo 1 e 2, conforme demarcado na Figura 65.

Figura 65: Clustering hierárquico com todos os genes para as quatro topologias. As amostras de cabeça e pescoço não se misturam com as de útero, pulmão e esôfago. E as amostras de cabeça e pescoço segmentam-se em dois grupos, chamados de grupo 1 e grupo 2.

Com o intuito de averiguar as possíveis diferenças entre esses dois grupos, fizemos análises de módulos ativos, Figura 66, que evidenciaram que o grupo 1 parece possuir um perfil mais agressivo quanto a metástase, fato esse que será averiguado nas próximas iterações.

Figura 66: Módulos ativos para as comparações entre os grupos 1 e 2 de amostras.

3.3.5.3. Discussão das análises de Carcinomas Epidermóides

De forma muito interessante os mesmos genes que apareceram como significativos com as análises de Carcinomas Epidermóides de Cabeça e Pescoço para metástase linfonodal estão aparecendo na separação entre os grupos 1 e 2. Maiores investigações estão sendo feitas para determinar a relevância dessa separação entre os grupos 1 e 2 das amostras. É interessante também verificar a separação entre carcinomas epidermóides de cabeça e pescoço com os outros tipos: pulmão, útero e esôfago, que de certa forma responde à pergunta de se carcinomas epidermóides eram todos iguais, não importando a topologia, devendo por tanto serem todos tratados da mesma forma clinicamente. Essa separação também se sustentou para a análise usando ANOVA (Hastie *et al.*, 2001) com quatro grupos (resultado não mostrado).

4. Discussão Final

Durante esse trabalho de doutorado tivemos a possibilidade de colaborar com vários projetos de análise de dados de expressão gênica gerados pela técnica de microarranjos de DNA. Além dos conjuntos de dados citados nesta tese também colaboramos na análise de dados de Varíola (Esteves *et al.*, 2007)⁶, Ciclo Celular de humanos (Simões *et al.*, 2007)⁷ e tumor de Willms (colaboração com a Dra. Dirce Carraro, Hospital A.C. Camargo). Tivemos a oportunidade de coordenar a análise de dados do projeto Sarcoma, Carcinomas Epidermóides e Carcinomas Epidermóides Cabeça e Pescoço, projetos nos quais pudemos propor a utilização de metodologias de análise e implementá-las, o que culminou na descoberta dos marcadores descritos no capítulo de resultados, nas ferramentas apresentadas e no modelo de utilização de XP para gerenciamento de análises de dados de transcriptômica. Já nos projetos de análise de dados de expressão gênica em câncer de rim e próstata, a colaboração se deu pela

⁶ Esteves, G. H., Simoes, A. C., Souza, E., Dias, R. A., Ospina, R. & Venancio, T. M. 2007. New insights about host response to smallpox using microarray data. BMC Syst Biol 1:38. Consta dos anexos inclusos no CD-ROM complementar a tese.

⁷ Simões, A. C. Q., Esteves, G. H., Cristo, E. B., Fahham, L. V., da Silva, A. M. & Neves, E. J. 2007. Active Modules and Relevance Networks during the cell cycle in HeLa cells. Presented at the The Eighth International Conference On Systems Biology, Long Beach, California, EUA. Consta dos anexos inclusos no CD-ROM complementar a tese.

proposição e implementação de diferentes estratégias, até então não usadas pelo grupo de pesquisa, para a busca por marcadores moleculares.

A bioinformática é uma área multidisciplinar, que integra conhecimentos de genômica, ciência da computação e estatística. A aplicação de XP em um projeto de bioinformática, conforme, por exemplo, o modelo proposto nesta tese, se insere na questão de como a área de bioinformática e os profissionais da mesma irão se inserir no mercado profissional e na área acadêmica. Por ter essa visão multidisciplinar, espera-se que os bioinformatas sejam capazes de participar de projetos de genômica atuando nos três níveis de *expertise* do projeto: na parte biológica, na parte computacional-estatística e na interface entre ambas. Assim, pode-se verificar que a aplicação de XP, como proposto neste trabalho, insere o profissional da bioinformática de forma natural no projeto, de forma que suas contribuições possam ser maximizadas.

XP envolve muito planejamento do que deve ser feito em curto prazo, e um nível menor de planejamento do que deve ser feito a médio e longo prazo, para que se possam ter os resultados prioritários sempre sendo trabalhados. Os casos apresentados nesta tese tornam-se interessantes sob a perspectiva de que se inserem em um ambiente acadêmico, que, por ter sua cultura própria, leva ao fato de que nem todas as regras se aplicam perfeitamente. Dois pontos foram considerados mais críticos quanto à adaptação de XP à academia: propriedade coletiva para o código e programação pareada, principalmente pela questão de propriedade intelectual e originalidade em projetos acadêmicos. Logo, ao se implementar as dinâmicas de programação pareada e propriedade coletiva de código, temos que a questão da contribuição e originalidade podem se tornar um problema. No entanto conseguimos implementar o planejamento, gerenciamento e análise de dados usando XP em projetos acadêmicos com propriedade coletiva de código e de resultados.

A ferramenta AMPiR, apresentada nesta tese, foi inicialmente desenvolvida para suprir a necessidade de anotação da plataforma não comercial ACCamargo3. O seu desenvolvimento mostrou o poder da integração de buscas dentro das bases do NCBI, sendo esse o principal da ferramenta desenvolvida com relação ao pacote AnnBuilder do Bioconductor (Zhang *et al.*, 2003). A sua transformação em um pacote R visa suprir as dificuldades de se anotar plataformas não comerciais com o ambiente R acoplado a pacotes e visa também fornecer ao usuário a possibilidade de priorizar caminhos para a anotação, o que entendemos que fornece uma maior robustez à anotação. É importante

ressaltar que essa ferramenta foi desenvolvida considerando que os usuários que a utilizarão serão profissionais da área de bioinformática.

A ferramenta TyPiR, também apresentada nesta tese, foi desenvolvida para suprir a dificuldade de rastreamento de análises realizadas no ambiente R associado a pacotes, uma vez que permite o armazenamento dos objetos resultantes das análises em um repositório que pode ser organizado por cartões, isto é, por perguntas biológicas, conforme o modelo proposto de planejamento e gerenciamento também apresentados nesta tese. Essa ferramenta foi desenvolvida considerando usuários que possuam familiaridade com a interface de programação R.

As análises para o conjunto de dados de Carcinomas Epidermóides de Cabeça e Pescoço quanto às diferenças moleculares para as cinco topologias levam a acreditar que não há grandes diferenças moleculares entre cinco topologias: boca, oro, cnsp, lari e hipo. Tal dado corrobora com a hipótese de que não é a topologia do carcinoma epidermóide que deve direcionar o tratamento da doença tumores de cabeça e pescoço. Porém, os dados de Carcinomas Epidermóides mostraram que as amostras de cabeça e pescoço são bem diferentes das de útero, pulmão e esôfago, quanto ao perfil global de expressão gênica, sugerindo que o tratamento de CECF deve ser diferente ao aplicado aos carcinomas de útero, pulmão e esôfago.

De modo geral, as análises com combinação de variáveis se mostraram muito mais ricas do que as com variáveis simples, o que evidencia a complexidade intrínseca dos tumores Sarcomas e Carcinomas Epidermóides e do câncer em geral. A utilização do meta-objeto, conforme o modelo proposto, foi crucial nessa etapa de combinação das variáveis, uma vez que permitiu a rápida identificação de dúvidas e inconsistências nas análises.

A busca por marcadores usando genes diferencialmente expressos bem como genes encontrados como significativos nas análises de módulos ativos e redes de relevância (Butte *et al.*, 2000; Segal *et al.*, 2004; Esteves, 2007), levaram a identificação de potenciais marcadores moleculares que estão em processo de validação experimental. Ademais, as análises por módulos e vias permitiu a inspeção de processos biológicos em separado e agrupados por contextos biológicos de interesse, o que acabou por enriquecer muito as análises e a busca por classificadores.

Apesar de nesta tese apresentarmos a identificação de potenciais marcadores para câncer, estamos cientes das dificuldades e desafios para que esses marcadores possam ser efetivamente validados e, se for o caso, incorporados na clínica médica

(Sawyers, 2008; van' t Veer & Bernards, 2008). Entendemos que os experimentos e resultados apresentados nesta tese podem ser aceitos como fase 1 das cinco etapas de desenvolvimento de um marcador clínico (Pepe *et al.*, 2001), uma vez que serviram para apontar potenciais marcadores que estão em validação usando o mesmo conjunto de amostras descrito aqui nesta tese. Porém, para que esses marcadores possam evoluir para a fase 2 de desenvolvimento, eles devem ser validados usando um novo, e de preferência grande, conjunto de amostras de tecidos tumorais e também de tecidos normais. Tal tarefa exigirá, além do acesso a amostras de tecidos de pacientes, mais recursos financeiros e humanos, o que por si só já é um desafio.

Uma forma de contornar, em parte, esse desafio da validação do marcador em novas amostras e em tecidos não tumorais é a integração e comparação desses experimentos apresentados nesse trabalho com outros, como é feito por exemplo na base de dados Oncomine (Rhodes *et al.*, 2007) ou na base de dados Gene Expression Omnibus (GEO Profiles) (Barrett *et al.*, 2005; Barrett *et al.*, 2009; Sayers *et al.*, 2009). Atualmente, a base de dados Oncomine armazena 18.000 perfis de expressão de genes em tumores curados manualmente pelo projeto. Ela permite que se avalie o perfil de um gene específico nas amostras de tumores presentes em sua base de dados, permitindo a avaliação desse gene em tumores do mesmo tipo para o qual estava sendo proposto como marcador e em outros tipos de tumores. Já a base de dados GEO Profiles, por ser um repositório público que possui experimentos de microarranjos de DNA de diversos tipos de tecidos e organismos, permite a avaliação do comportamento do marcador proposto em amostras de tecidos normais de vários órgãos e em vários tecidos tumorais. Portanto, propomos que na etapa 1 de desenvolvimento de marcadores também sejam usadas essas ferramentas computacionais antes da realização de experimentos de validação, para que o investimento seja feito nos marcadores mais promissores.

5. Bibliografia

1. **Abate-Shen, C. & Shen, M. M.** 2000. Molecular genetics of prostate cancer. *Genes Dev* 14:2410-34.
2. **Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F. & et al.** 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252:1651-6.
3. **Albertsen, P. C., Hanley, J. A., Gleason, D. F. & Barry, M. J.** 1998. Competing risk analysis of men aged 55 to 74 years at diagnosis managed conservatively for clinically localized prostate cancer. *Jama* 280:975-80.
4. **Alizadeh, A. A., Ross, D. T., Perou, C. M. & van de Rijn, M.** 2001. Towards a novel classification of human malignancies based on gene expression patterns. *J Pathol* 195:41-52.
5. **Allander, S. V., Illei, P. B., Chen, Y., Antonescu, C. R., Bittner, M., Ladanyi, M. & Meltzer, P. S.** 2002. Expression profiling of synovial sarcoma by cDNA microarrays: association of ERBB2, IGFBP2, and ELF3 with epithelial differentiation. *Am J Pathol* 161:1587-95.
6. **Almadori, G., Bussu, F. & Paludetti, G.** 2008. Should there be more molecular staging of head and neck cancer to improve the choice of treatments and thereby improve survival? *Curr Opin Otolaryngol Head Neck Surg* 16:117-26.
7. **Amaral, P. P. & Mattick, J. S.** 2008. Noncoding RNA in development. *Mamm Genome* 19:454-92.
8. **Antonescu, C. R.** 2008. Molecular profiling in the diagnosis and treatment of high grade sarcomas. *Ultrastruct Pathol* 32:37-42.
9. **Baldi, P. & Long, A. D.** 2001. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* 17:509-19.
10. **Ball, C. A. & Brazma, A.** 2006. MGED standards: work in progress. *Omics* 10:138-44.
11. **Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W. C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W. & Edgar, R.** 2005. NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic Acids Res* 33:D562-6.
12. **Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Muerlter, R. N. & Edgar, R.** 2009. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* 37:D885-90.
13. **Bassi Filho, D. L.** 2008. Experiências com desenvolvimento ágil. Instituto de Matemática e Estatística Universidade de São Paulo, São Paulo.154
14. **Beck, K.** 2000. eXtreme Programming: Embrace Change. Addison-Wesley.
15. **Beck, K. & Fowler, M.** 2001. Planning eXtreme Programming. Addison-Wesley.
16. **Beer, D. G., Kardia, S. L. R., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G. A., Gharib, T. G., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M. G., Iannettoni, M. D.,**

- Orringer, M. B. & Hanash, S.** 2002. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* 8:816-824.
17. **Belbin, T. J., Singh, B., Barber, I., Socci, N., Wenig, B., Smith, R., Prystowsky, M. B. & Childs, G.** 2002. Molecular classification of head and neck squamous cell carcinoma using cDNA microarrays. *Cancer Res* 62:1184-90.
18. **Belbin, T. J., Singh, B., Smith, R. V., Socci, N. D., Wreesmann, V. B., Sanchez-Carbayo, M., Masterson, J., Patel, S., Cordon-Cardo, C., Prystowsky, M. B. & Childs, G.** 2005. Molecular profiling of tumor progression in head and neck cancer. *Arch Otolaryngol Head Neck Surg* 131:10-8.
19. **Berman, J. & O'Leary, T. J.** 2001. Gastrointestinal stromal tumor workshop. *Hum Pathol* 32:578-82.
20. **Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M. & Snyder, M.** 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* 306:2242-6.
21. **Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J. & Meyerson, M.** 2001. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America* 98:13790-13795.
22. **Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R. *et al.*** 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447:799-816.
23. **Boer, J. M., Huber, W. K., Sultmann, H., Wilmer, F., von Heydebreck, A., Haas, S., Korn, B., Gunawan, B., Vente, A., Fuzesi, L., Vingron, M. & Poustka, A.** 2001. Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31,500-element cDNA array. *Genome Res* 11:1861-70.
24. **Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P. *et al.*** 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 29:365-71.
25. **Breiman, L., Friedman, J. H., Olsen, R. A. & Stone, C. J.** 1984. *Classification and Regression Trees*. Wadsworth & Brooks . Pacific grove, CA.
26. **Brentani, R. R., Carraro, D. M., Verjovski-Almeida, S., Reis, E. M., Neves, E. J., de Souza, S. J., Carvalho, A. F., Brentani, H. & Reis, L. F.** 2005. Gene expression arrays in cancer research: methods and applications. *Crit Rev Oncol Hematol* 54:95-105.
27. **Brito, G. C., Fachel, A. A., Vettore, A. L., Vignal, G. M., Gimba, E. R., Campos, F. S., Barcinski, M. A., Verjovski-Almeida, S. & Reis, E. M.** 2008. Identification of protein-coding and intronic noncoding RNAs down-regulated in clear cell renal carcinoma. *Mol Carcinog* 47:757-67.
28. **Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R. & Kohane, I. S.** 2000. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci U S A* 97:12182-6.

29. **Buyse, M., Loi, S., van't Veer, L., Viale, G., Delorenzi, M. *et al.*** 2006. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst* 98:1183-92.
30. **Cai, Z., Goebel, R., Salavatipour, M. R. & Lin, G.** 2007. Selecting dissimilar genes for multi-class classification, an application in cancer subtyping. *BMC Bioinformatics* 8:206.
31. **Calin, G. A. & Croce, C. M.** 2006. MicroRNA-cancer connection: the beginning of a new tale. *Cancer Res* 66:7390-4.
32. **Camargo, A. A., Samaia, H. P., Dias-Neto, E., Simao, D. F., Migotto, I. A. *et al.*** 2001. The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proc Natl Acad Sci U S A* 98:12103-8.
33. **Carvalho, A. L., Nishimoto, I. N., Califano, J. A. & Kowalski, L. P.** 2005. Trends in incidence and prognosis for head and neck cancer in the United States: a site-specific analysis of the SEER database. *Int J Cancer* 114:806-16.
34. **Causton, H. C. & Game, L.** 2003. MGED comes of age. *Genome Biol* 4:351.
35. **Causton, H. C., Quackenbush, J. F. & Brazma, A.** 2003. *Microarray/gene expression data analysis: a beginner's guide*. Blackwell Science Ltd, Oxford.
36. **Chang, H. Y., Nuyten, D. S., Sneddon, J. B., Hastie, T., Tibshirani, R., Sorlie, T., Dai, H., He, Y. D., van't Veer, L. J., Bartelink, H., van de Rijn, M., Brown, P. O. & van de Vijver, M. J.** 2005. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc Natl Acad Sci U S A* 102:3738-43.
37. **Chang, S. S. & Califano, J.** 2008. Current status of biomarkers in head and neck cancer. *J Surg Oncol* 97:640-3.
38. **Chen, H. Y., Yu, S. L., Chen, C. H., Chang, G. C., Chen, C. Y., Yuan, A., Cheng, C. L., Wang, C. H., Terng, H. J., Kao, S. F., Chan, W. K., Li, H. N., Liu, C. C., Singh, S., Chen, W. J., Chen, J. J. W. & Yang, P. C.** 2007. A five-gene signature and clinical outcome in non-small-cell lung cancer. *New England Journal of Medicine* 356:11-20.
39. **Chisholm, R. L., Gaudet, P., Just, E. M., Pilcher, K. E., Fey, P., Merchant, S. N. & Kibbe, W. A.** 2006. dictyBase, the model organism database for *Dictyostelium discoideum*. *Nucleic Acids Res* 34:D423-7.
40. **Chung, C. H., Bernard, P. S. & Perou, C. M.** 2002. Molecular portraits and the family tree of cancer. *Nat Genet* 32 Suppl:533-40.
41. **Chung, C. H., Parker, J. S., Ely, K., Carter, J., Yi, Y., Murphy, B. A., Ang, K. K., El-Naggar, A. K., Zanation, A. M., Cmelak, A. J., Levy, S., Slebos, R. J. & Yarbrough, W. G.** 2006. Gene expression profiles identify epithelial-to-mesenchymal transition and activation of nuclear factor-kappaB signaling as characteristics of a high-risk head and neck squamous cell carcinoma. *Cancer Res* 66:8210-8.
42. **Colantuoni, C., Henry, G., Zeger, S. & Pevsner, J.** 2002. SNOMAD (Standardization and Normalization of MicroArray Data): web-accessible gene expression data analysis. *Bioinformatics* 18:1540-1.
43. **Consortium, T. G. O.** 2000. Gene Ontology: tool for the unification of biology. *Nature Genet* 25:25-29.
44. **Consortium, T. U.** 2008. The Universal Protein Resource (UniProt). *Nucleic Acids Res* 36:D190-D195.
45. **Cromer, A., Carles, A., Millon, R., Ganguli, G., Chalmel, F., Lemaire, F., Young, J., Dembele, D., Thibault, C., Muller, D., Poch, O., Abecassis, J. & Wasylyk, B.** 2004. Identification of genes associated with tumorigenesis and

- metastatic potential of hypopharyngeal cancer by microarray analysis. *Oncogene* 23:2484-98.
46. **Cusumano, M. A.** 2007. Extreme programming compared with Microsoft-style iterative development. *Communications of the ACM* 50:15-18.
 47. **Dai, H., van't Veer, L., Lamb, J., He, Y. D., Mao, M., Fine, B. M., Bernards, R., van de Vijver, M., Deutsch, P., Sachs, A., Stoughton, R. & Friend, S.** 2005. A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. *Cancer Res* 65:4059-66.
 48. **Dei Tos, A. P., Maestro, R., Doglioni, C., Piccinin, S., Libera, D. D., Boiocchi, M. & Fletcher, C. D.** 1996. Tumor suppressor genes and related molecules in leiomyosarcoma. *Am J Pathol* 148:1037-45.
 49. **Dennis, J. E. & Schnabel, R. B.** 1987. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, NJ.
 50. **DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y. D., Su, Y. A. & Trent, J. M.** 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics* 14:457-460.
 51. **Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A. & Chinnaiyan, A. M.** 2001. Delineation of prognostic biomarkers in prostate cancer. *Nature* 412:822-826.
 52. **Draghici, S., Khatri, P., Eklund, A. C. & Szallasi, Z.** 2006. Reliability and reproducibility issues in DNA microarray measurements. *Trends in Genetics* 22:101-109.
 53. **Dudoit, S., Gentleman, R. C. & Quackenbush, J.** 2003. Open source software for the analysis of microarray data. *Biotechniques Suppl*:45-51.
 54. **Dudoit, S. & Speed, T. P.** 1999. A score test for linkage using identity by descent data from sibships. *Annals of Statistics* 27:943-986.
 55. **Dupuy, A. & Simon, R. M.** 2007. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *Journal of the National Cancer Institute* 99:147-157.
 56. **Eichinger, L., Pachebat, J. A., Glockner, G., Rajandream, M. A., Sucgang, R. et al.** 2005. The genome of the social amoeba *Dictyostelium discoideum*. *Nature* 435:43-57.
 57. **Elo, J. P. & Visakorpi, T.** 2001. Molecular genetics of prostate cancer. *Ann Med* 33:130-41.
 58. **Esteves, G. H.** 2007. Métodos estatísticos para a análise de dados de cDNA microarray em um ambiente computacional integrado. Tese de Doutorado. Bioinformática (IME/IFSC/ESALQ/IQ/IB/ICB/FMVZ/FCFRP), Universidade de São Paulo, São Paulo.174
 59. **Esteves, G. H., Simoes, A. C., Souza, E., Dias, R. A., Ospina, R. & Venancio, T. M.** 2007. New insights about host response to smallpox using microarray data. *BMC Syst Biol* 1:38.
 60. **Finlin, B. S., Gau, C. L., Murphy, G. A., Shao, H., Kimel, T., Seitz, R. S., Chiu, Y. F., Botstein, D., Brown, P. O., Der, C. J., Tamanoi, F., Andres, D. A. & Perou, C. M.** 2001. RERG is a novel ras-related, estrogen-regulated and growth-inhibitory gene in breast cancer. *J Biol Chem* 276:42259-67.
 61. **Franco, E. D.** 2007. Um modelo de gerenciamento de projeto baseado nas metodologias de desenvolvimento de software e nos princípios de produção enxuta. Escola Politécnica, Universidade de São Paulo, São Paulo.107
 62. **Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaessler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G. D., Perou, C. M.,**

- Whyte, R. I., Altman, R. B., Brown, P. O., Botstein, D. & Petersen, I.** 2001. Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci U S A* 98:13784-9.
63. **Genden, E. M., Ferlito, A., Bradley, P. J., Rinaldo, A. & Scully, C.** 2003. Neck disease and distant metastases. *Oral Oncol* 39:207-12.
 64. **Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M. *et al.*** 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5:R80.
 65. **Ginos, M. A., Page, G. P., Michalowicz, B. S., Patel, K. J., Volker, S. E., Pambuccian, S. E., Ondrey, F. G., Adams, G. L. & Gaffney, P. M.** 2004. Identification of a gene expression signature associated with recurrent disease in squamous cell carcinoma of the head and neck. *Cancer Res* 64:55-63.
 66. **Giri, U., Ashorn, C. L., Ramdas, L., Stivers, D. N., Coombes, K., El-Naggar, A. K., Ang, K. K. & Story, M. D.** 2006. Molecular signatures associated with clinical outcome in patients with high-risk head-and-neck squamous cell carcinoma treated by surgery and radiation. *Int J Radiat Oncol Biol Phys* 64:670-7.
 67. **Glas, A. M., Floore, A., Delahaye, L. J., Witteveen, A. T., Pover, R. C., Bakx, N., Lahti-Domenici, J. S., Bruinsma, T. J., Warmoes, M. O., Bernardis, R., Wessels, L. F. & Van't Veer, L. J.** 2006. Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics* 7:278.
 68. **Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. & Lander, E. S.** 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531-7.
 69. **Gomes, L. I., Esteves, G. H., Carvalho, A. F., Cristo, E. B., Hirata, R., Jr., Martins, W. K., Marques, S. M., Camargo, L. P., Brentani, H., Pelosof, A., Zitron, C., Sallum, R. A., Montagnini, A., Soares, F. A., Neves, E. J. & Reis, L. F.** 2005. Expression profile of malignant and nonmalignant lesions of esophagus and stomach: differential activity of functional modules related to inflammation and lipid metabolism. *Cancer Res* 65:7127-36.
 70. **Gough, N. R.** 2002. Science's signal transduction knowledge environment: the connections maps database. *Ann N Y Acad Sci* 971:585-7.
 71. **Haddad, R. I. & Shin, D. M.** 2008. Recent advances in head and neck cancer. *N Engl J Med* 359:1143-54.
 72. **Hanahan, D. & Weinberg, R. A.** 2000. The hallmarks of cancer. *Cell* 100:57-70.
 73. **Hastie, T., Tibshirani, R. & H., F. J.** 2001. *The Elements of Statistical Learning* (Hardcover). Springer, New York.
 74. **Helman, L. J. & Meltzer, P.** 2003. Mechanisms of sarcoma development. *Nat Rev Cancer* 3:685-94.
 75. **Hirota, E., Yan, L., Tsunoda, T., Ashida, S., Fujime, M., Shuin, T., Miki, T., Nakamura, Y. & Katagiri, T.** 2006. Genome-wide gene expression profiles of clear cell renal cell carcinoma: identification of molecular targets for treatment of renal cell carcinoma. *Int J Oncol* 29:799-827.
 76. **Hoheisel, J. D.** 2006. Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet* 7:200-10.

77. **Hoppener, J. W., Mosselman, S., Roholl, P. J., Lambrechts, C., Slebos, R. J., de Pagter-Holthuizen, P., Lips, C. J., Jansz, H. S. & Sussenbach, J. S.** 1988. Expression of insulin-like growth factor-I and -II genes in human smooth muscle tumours. *Embo J* 7:1379-85.
78. **Hornick, J. L. & Fletcher, C. D.** 2002. Immunohistochemical staining for KIT (CD117) in soft tissue sarcomas is very limited in distribution. *Am J Clin Pathol* 117:188-93.
79. **Howe, H. L., Wu, X., Ries, L. A., Cokkinides, V., Ahmed, F., Jemal, A., Miller, B., Williams, M., Ward, E., Wingo, P. A., Ramirez, A. & Edwards, B. K.** 2006. Annual report to the nation on the status of cancer, 1975-2003, featuring cancer among U.S. Hispanic/Latino populations. *Cancer* 107:1711-42.
80. **Hu, Z., Fan, C., Oh, D. S., Marron, J. S., He, X. *et al.*** 2006. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* 7:96.
81. **Ihaka, R. & Gentleman, R.** 1996. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* 5:299-314.
82. **Ishkanian, A. S., Malloff, C. A., Watson, S. K., deLeeuw, R. J., Chi, B., Coe, B. P., Snijders, A., Albertson, D. G., Pinkel, D., Marra, M. A., Ling, V., MacAulay, C. & Lam, W. L.** 2004. A tiling resolution DNA microarray with complete coverage of the human genome. *Nature Genetics* 36:299-303.
83. **Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., Murray, T. & Thun, M. J.** 2008. Cancer statistics, 2008. *CA Cancer J Clin* 58:71-96.
84. **Jones, J. & Libermann, T. A.** 2007. Genomics of renal cell cancer: the biology behind and the therapy ahead. *Clin Cancer Res* 13:685s-692s.
85. **Jones, J., Otu, H., Spentzos, D., Kolia, S., Inan, M., Beecken, W. D., Fellbaum, C., Gu, X., Joseph, M., Pantuck, A. J., Jonas, D. & Libermann, T. A.** 2005. Gene signatures of progression and metastasis in renal cell cancer. *Clin Cancer Res* 11:5730-9.
86. **Kalebic, T., Blakesley, V., Slade, C., Plasschaert, S., Leroith, D. & Helman, L. J.** 1998. Expression of a kinase-deficient IGF-I-R suppresses tumorigenicity of rhabdomyosarcoma cells constitutively expressing a wild type IGF-I-R. *Int J Cancer* 76:223-7.
87. **Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. & Yamanishi, Y.** 2008. KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 36:D480-4.
88. **Kaushal, D. & Naeve, C. W.** 2004. An overview of Spotfire for gene-expression studies. *Curr Protoc Bioinformatics Chapter 7:Unit 7 7*.
89. **Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. & Haussler, D.** 2002. The human genome browser at UCSC. *Genome Res* 12:996-1006.
90. **Koide, T., Salem-Izacc, S. M., Gomes, S. L. & Vencio, R. Z.** 2006. SpotWhatR: a user-friendly microarray data analysis system. *Genet Mol Res* 5:93-107.
91. **Kopper, L. & Timar, J.** 2006. Genomics of renal cell cancer-- does it provide breakthrough? *Pathol Oncol Res* 12:5-11.
92. **Kuhn, A., Luthi-Carter, R. & Delorenzi, M.** 2008. Cross-species and cross-platform gene expression studies with the Bioconductor-compliant R package 'annotationTools'. *BMC Bioinformatics* 9:26.

93. **Kumar, V., Abbas, A. K. & Fausto, N.** 2005. Robbins and Cotran Pathologic Basis of Disease., 7th edition ed. Elsevier Inc., Philadelphia, PA.
94. **Lal, A., Lash, A. E., Altschul, S. F., Velculescu, V., Zhang, L., McLendon, R. E., Marra, M. A., Prange, C., Morin, P. J., Polyak, K., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., Strausberg, R. L. & Riggins, G. J.** 1999. A public database for gene expression in human cancers. *Cancer Res* 59:5403-7.
95. **Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C. *et al.*** 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860-921.
96. **Lin, S. L., Miller, J. D. & Ying, S. Y.** 2006. Intronic MicroRNA (miRNA). *J Biomed Biotechnol* 2006:26818.
97. **Linehan, W. M., Grubb, R. L., Coleman, J. A., Zbar, B. & Walther, M. M.** 2005. The genetic basis of cancer of kidney cancer: implications for gene-specific clinical management. *BJU Int* 95 Suppl 2:2-7.
98. **Lloyd, C. M., Lawson, J. R., Hunter, P. J. & Nielsen, P. F.** 2008. The CellML Model Repository. *Bioinformatics* 24:2122-3.
99. **Lockhart, D. J. & Winzler, E. A.** 2000. Genomics, gene expression and DNA arrays. *Nature* 405:827-36.
100. **Luo, J., Duggan, D. J., Chen, Y., Sauvageot, J., Ewing, C. M., Bittner, M. L., Trent, J. M. & Isaacs, W. B.** 2001. Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res* 61:4683-8.
101. **Ma, X. J., Salunga, R., Dahiya, S., Wang, W., Carney, E., Durbecq, V., Harris, A., Goss, P., Sotiriou, C., Erlander, M. & Sgroi, D.** 2008. A five-gene molecular grade index and HOXB13:IL17BR are complementary prognostic factors in early stage breast cancer. *Clin Cancer Res* 14:2601-8.
102. **Mackall, C. L., Meltzer, P. S. & Helman, L. J.** 2002. Focus on sarcomas. *Cancer Cell* 2:175-8.
103. **Mackay, A., Jones, C., Dexter, T., Silva, R. L., Bulmer, K., Jones, A., Simpson, P., Harris, R. A., Jat, P. S., Neville, A. M., Reis, L. F., Lakhani, S. R. & O'Hare, M. J.** 2003. cDNA microarray analysis of genes associated with ERBB2 (HER2/neu) overexpression in human mammary luminal epithelial cells. *Oncogene* 22:2680-8.
104. **Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T.** 2007. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 35:D26-31.
105. **Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D. *et al.*** 2009. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* 37:D619-22.
106. **McBreen, P.** 2003. Questioning Extreme Programming. Pearson Education Inc., Boston.
107. **Meireles, S. I., Carvalho, A. F., Hirata, R., Montagnini, A. L., Martins, W. K., Runza, F. B., Stolf, B. S., Termini, L., Neto, C. E., Silva, R. L., Soares, F. A., Neves, E. J. & Reis, L. F.** 2003. Differentially expressed genes in gastric tumors identified by cDNA array. *Cancer Lett* 190:199-211.
108. **Meireles, S. I., Cristo, E. B., Carvalho, A. F., Hirata, R., Jr., Pelosof, A., Gomes, L. I., Martins, W. K., Begnami, M. D., Zitron, C., Montagnini, A. L., Soares, F. A., Neves, E. J. & Reis, L. F.** 2004. Molecular classifiers for gastric cancer and nonmalignant diseases of the gastric mucosa. *Cancer Res* 64:1255-65.

109. **Misra, J., Schmitt, W., Hwang, D., Hsiao, L. L., Gullans, S., Stephanopoulos, G. & Stephanopoulos, G.** 2002. Interactive exploration of microarray gene expression patterns in a reduced dimensional space. *Genome Res* 12:1112-20.
110. **Mockler, T. C., Chan, S., Sundaresan, A., Chen, H., Jacobsen, S. E. & Ecker, J. R.** 2005. Applications of DNA tiling arrays for whole-genome analysis. *Genomics* 85:1-15.
111. **Morozova, O. & Marra, M. A.** 2008. Applications of next-generation sequencing technologies in functional genomics. *Genomics* 92:255-64.
112. **Nagayama, S., Katagiri, T., Tsunoda, T., Hosaka, T., Nakashima, Y., Araki, N., Kusuzaki, K., Nakayama, T., Tsuboyama, T., Nakamura, T., Imamura, M., Nakamura, Y. & Toguchida, J.** 2002. Genome-wide analysis of gene expression in synovial sarcomas using a cDNA microarray. *Cancer Res* 62:5859-66.
113. **Nakaya, H. I., Amaral, P. P., Louro, R., Lopes, A., Fachel, A. A., Moreira, Y. B., El-Jundi, T. A., da Silva, A. M., Reis, E. M. & Verjovski-Almeida, S.** 2007. Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. *Genome Biol* 8:R43.
114. **Nielsen, T. O.** 2006. Microarray analysis of sarcomas. *Adv Anat Pathol* 13:166-73.
115. **Nielsen, T. O., West, R. B., Linn, S. C., Alter, O., Knowling, M. A., O'Connell, J. X., Zhu, S., Fero, M., Sherlock, G., Pollack, J. R., Brown, P. O., Botstein, D. & van de Rijn, M.** 2002. Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet* 359:1301-7.
116. **Okabe, H., Satoh, S., Kato, T., Kitahara, O., Yanagawa, R., Yamaoka, Y., Tsunoda, T., Furukawa, Y. & Nakamura, Y.** 2001. Genome-wide analysis of gene expression in human hepatocellular carcinomas using cDNA microarray: Identification of genes involved in viral carcinogenesis and tumor progression. *Cancer Research* 61:2129-2137.
117. **Palka, K. T., Slebos, R. J. & Chung, C. H.** 2008. Update on molecular diagnostic tests in head and neck cancer. *Semin Oncol* 35:198-210.
118. **Paquola, A. C., Machado, A. A., Reis, E. M., Da Silva, A. M. & Verjovski-Almeida, S.** 2003a. Zerg: a very fast BLAST parser library. *Bioinformatics* 19:1035-6.
119. **Paquola, A. C., Nishiyama, M. Y., Jr., Reis, E. M., da Silva, A. M. & Verjovski-Almeida, S.** 2003b. ESTWeb: bioinformatics services for EST sequencing projects. *Bioinformatics* 19:1587-8.
120. **Parekh, D. J., Ankerst, D. P., Troyer, D., Srivastava, S. & Thompson, I. M.** 2007. Biomarkers for prostate cancer detection. *Journal of Urology* 178:2252-2259.
121. **Parkin, D. M., Bray, F., Ferlay, J. & Pisani, P.** 2001. Estimating the world cancer burden: Globocan 2000. *Int J Cancer* 94:153-6.
122. **Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M. *et al.*** 2009. ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res* 37:D868-72.
123. **Pepe, M. S., Etzioni, R., Feng, Z., Potter, J. D., Thompson, M. L., Thornquist, M., Winget, M. & Yasui, Y.** 2001. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst* 93:1054-61.

124. **Pinheiro, N. A., Caballero, O. L., Soares, F., Reis, L. F. & Simpson, A. J.** 2001. Significant overexpression of oligophrenin-1 in colorectal tumors detected by cDNA microarray analysis. *Cancer Lett* 172:67-73.
125. **Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M. *et al.*** 2002. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415:436-442.
126. **Quackenbush, J.** 2001. Computational analysis of microarray data. *Nat Rev Genet* 2:418-27.
127. **Quackenbush, J.** 2002. Microarray data normalization and transformation. *Nat Genet* 32 Suppl:496-501.
128. **Quackenbush, J.** 2006. Microarray analysis and tumor classification. *N Engl J Med* 354:2463-72.
129. **Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C. H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., Poggio, T., Gerald, W., Loda, M., Lander, E. S. & Golub, T. R.** 2001. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* 98:15149-54.
130. **Reis, E. M., Louro, R., Nakaya, H. I. & Verjovski-Almeida, S.** 2005. As antisense RNA gets intronic. *Omics* 9:2-12.
131. **Reis, E. M., Nakaya, H. I., Louro, R., Canavez, F. C., Flatschart, A. V. *et al.*** 2004. Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. *Oncogene* 23:6684-92.
132. **Ren, B., Yu, Y. P., Jing, L., Liu, L., Michalopoulos, G. K., Luo, J. H. & Rao, U. N.** 2003. Gene expression analysis of human soft tissue leiomyosarcomas. *Hum Pathol* 34:549-58.
133. **Reuter, V. E. & Presti, J. C., Jr.** 2000. Contemporary approach to the classification of renal epithelial tumors. *Semin Oncol* 27:124-37.
134. **Reynolds, M. A.** 2008. Molecular alterations in prostate cancer. *Cancer Letters* 271:13-24.
135. **Rhodes, D. R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B. B., Barrette, T. R., Anstet, M. J., Kincead-Beal, C., Kulkarni, P., Varambally, S., Ghosh, D. & Chinnaiyan, A. M.** 2007. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* 9:166-80.
136. **Roepman, P., Kemmeren, P., Wessels, L. F., Slootweg, P. J. & Holstege, F. C.** 2006. Multiple robust signatures for detecting lymph node metastasis in head and neck cancer. *Cancer Res* 66:2361-6.
137. **Roepman, P., Wessels, L. F., Kettelarij, N., Kemmeren, P., Miles, A. J., Lijnzaad, P., Tilanus, M. G., Koole, R., Hordijk, G. J., van der Vliet, P. C., Reinders, M. J., Slootweg, P. J. & Holstege, F. C.** 2005. An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas. *Nat Genet* 37:182-6.
138. **Roholl, P. J., Skottner, A., Prinsen, I., Lips, C. J., Den Otter, W. & Van Unnik, J. A.** 1990. Expression of insulin-like growth factor 1 in sarcomas. *Histopathology* 16:455-60.
139. **Ross, D. T. & Perou, C. M.** 2001. A comparison of gene expression signatures from breast tumors and breast tissue derived cell lines. *Dis Markers* 17:99-109.
140. **Rossini, A., Heiberger, R., Sparapani, R., Maechler, M. & Hornik, K.** 2004. *Emacs Speaks Statistics: A Multiplatform, Multipackage Development*

- Environment for Statistical Analysis. Journal of Computational & Graphical Statistics 13:247-261.
141. **Saal, L. H., Troein, C., Vallon-Christersson, J., Gruvberger, S., Borg, A. & Peterson, C.** 2002. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol* 3:SOFTWARE0003.
 142. **Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W. *et al.*** 2003. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34:374-8.
 143. **Sankaranarayanan, R., Masuyer, E., Swaminathan, R., Ferlay, J. & Whelan, S.** 1998. Head and neck cancer: a global perspective on epidemiology and prognosis. *Anticancer Res* 18:4779-86.
 144. **Sardana, G., Dowell, B. & Diamandis, E. P.** 2008. Emerging Biomarkers for the Diagnosis and Prognosis of Prostate Cancer. *Clinical Chemistry* 54:1951-1960.
 145. **Sarkans, U., Parkinson, H., Lara, G. G., Oezcimen, A., Sharma, A., Abeygunawardena, N., Contrino, S., Holloway, E., Rocca-Serra, P., Mukherjee, G., Shojatalab, M., Kapushesky, M., Sansone, S. A., Farne, A., Rayner, T. & Brazma, A.** 2005. The ArrayExpress gene expression database: a software engineering and implementation perspective. *Bioinformatics* 21:1495-501.
 146. **Sato, D.** 2007. Uso eficaz de métricas em métodos ágeis de desenvolvimento de software. Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.139
 147. **Savage, P. & Waxman, J.** 1996. PSA and prostate cancer diagnosis. *Eur J Cancer* 32A:1097-9.
 148. **Sawyers, C. L.** 2008. The cancer biomarker problem. *Nature* 452:548-52.
 149. **Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K. *et al.*** 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 37:D5-15.
 150. **Schena, M., Heller, R. A., Thériault, T. P., Konrad, K., Lachenmeier, E. & Davis, R. W.** 1998. Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol* 16:301-6.
 151. **Schnabel, R. B., Koontz, J. E. & Weiss, B. E., .** 1985. A modular system of algorithms for unconstrained minimization. *ACM Transactions on Mathematical Software* 11:419-440.
 152. **Schuetz, A. N., Yin-Goen, Q., Amin, M. B., Moreno, C. S., Cohen, C., Hornsby, C. D., Yang, W. L., Petros, J. A., Issa, M. M., Pattaras, J. G., Ogan, K., Marshall, F. F. & Young, A. N.** 2005. Molecular classification of renal tumors by gene expression profiling. *J Mol Diagn* 7:206-18.
 153. **Schulze, A. & Downward, J.** 2001. Navigating gene expression using microarrays--a technology review. *Nat Cell Biol* 3:E190-5.
 154. **Segal, E., Friedman, N., Koller, D. & Regev, A.** 2004. A module map showing conditional activity of expression modules in cancer. *Nat Genet* 36:1090-8.
 155. **Seliger, B., Dressler, S. P., Lichtenfels, R. & Kellner, R.** 2007. Candidate biomarkers in renal cell carcinoma. *Proteomics* 7:4601-12.
 156. **Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A. *et al.*** 2006. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 24:1151-61.

157. **Siever, E., Weber, A., Figgins, S., Love, R. & Robbins, A.** 2005. Linux in a Nutshell, Fifth Edition ed. O'Reilly, Sebastopol.
158. **Silva, M. A.** 2005. Uma ferramenta Web colaborativa para apoiar a engenharia de requisitos em software livre Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.138
159. **Simões, A. C. Q. & da Silva, A. M.** 2006. SAM as plugin for BASE. Presented at the 4th Annual Internacional Conference On Intelligent Systems For Molecular Biology, ISMB2006, Fortaleza, Ceará, Brasil.
160. **Simões, A. C. Q., da Silva, A. M., Verjovski-Almeida, S. & Reis, E. M.** 2005. SAM Method as an Approach to Select Candidates for Human Prostate Cancer Markers. Lecture Notes in Computer Science 3594/2005:202-205.
161. **Simões, A. C. Q., Esteves, G. H., Cristo, E. B., Fahham, L. V., da Silva, A. M. & Neves, E. J.** 2007. Active Modules and Relevance Networks during the cell cycle in HeLa cells. Presented at the The Eighth International Conference On Systems Biology, Long Beach, California, EUA.
162. **Simpson, A. J., Reinach, F. C., Arruda, P., Abreu, F. A., Acencio, M. *et al.*** 2000. The genome sequence of the plant pathogen *Xylella fastidiosa*. The *Xylella fastidiosa* Consortium of the Organization for Nucleotide Sequencing and Analysis. Nature 406:151-9.
163. **Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R. & Sellers, W. R.** 2002. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1:203-9.
164. **Skubitz, K. M. & Skubitz, A. P.** 2002. Differential gene expression in renal-cell cancer. J Lab Clin Med 140:52-64.
165. **Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Eystein Lonning, P. & Borresen-Dale, A. L.** 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 98:10869-74.
166. **Sotiriou, C. & Piccart, M. J.** 2007. Opinion - Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? Nature Reviews Cancer 7:545-553.
167. **Spellman, P. T., Miller, M., Stewart, J., Troup, C., Sarkans, U. *et al.*** 2002. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol 3:RESEARCH0046.
168. **Squire, J. A., Bayani, J., Luk, C., Unwin, L., Tokunaga, J., MacMillan, C., Irish, J., Brown, D., Gullane, P. & Kamel-Reid, S.** 2002. Molecular cytogenetic analysis of head and neck squamous cell carcinoma: By comparative genomic hybridization, spectral karyotyping, and expression array analysis. Head Neck 24:874-87.
169. **Srivastava, M., Bubendorf, L., Srikantan, V., Fossom, L., Nolan, L., Glasman, M., Leighton, X., Fehrle, W., Pittaluga, S., Raffeld, M., Koivisto, P., Willi, N., Gasser, T. C., Kononen, J., Sauter, G., Kallioniemi, O. P., Srivastava, S. & Pollard, H. B.** 2001. ANX7, a candidate tumor suppressor gene for prostate cancer. Proc Natl Acad Sci U S A 98:4575-80.
170. **Stratton, M. R., Moss, S., Warren, W., Patterson, H., Clark, J., Fisher, C., Fletcher, C. D., Ball, A., Thomas, M., Gusterson, B. A. & *et al.*** 1990.

- Mutation of the p53 gene in human soft tissue sarcomas: association with abnormalities of the RB1 gene. *Oncogene* 5:1297-301.
171. **Strausberg, R. L., Buetow, K. H., Emmert-Buck, M. R. & Klausner, R. D.** 2000. The cancer genome anatomy project: building an annotated gene index. *Trends Genet* 16:103-6.
 172. **Su, A. I., Welsh, J. B., Sapinoso, L. M., Kern, S. G., Dimitrov, P., Lapp, H., Schultz, P. G., Powell, S. M., Moskaluk, C. A., Frierson, H. F., Jr. & Hampton, G. M.** 2001. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res* 61:7388-93.
 173. **Szymanski, M., Barciszewska, M. Z., Erdmann, V. A. & Barciszewski, J.** 2005. A new frontier for molecular medicine: noncoding RNAs. *Biochim Biophys Acta* 1756:65-75.
 174. **Termine, N., Panzarella, V., Falaschini, S., Russo, A., Matranga, D., Lo Muzio, L. & Campisi, G.** 2008. HPV in oral squamous cell carcinoma vs head and neck squamous cell carcinoma biopsies: a meta-analysis (1988-2007). *Ann Oncol* 19:1681-90.
 175. **Troein, C., Vallon-Christersson, J. & Saal, L. H.** 2006. An introduction to BioArray Software Environment. *Methods Enzymol* 411:99-119.
 176. **Tusher, V. G., Tibshirani, R. & Chu, G.** 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98:5116-21.
 177. **van 't Veer, L. J. & Bernards, R.** 2008. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature* 452:564-70.
 178. **van 't Veer, L. J., Dai, H. Y., van de Vijver, M. J., He, Y. D. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. & Friend, S. H.** 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415:530-536.
 179. **van de Vijver, M. J., He, Y. D., van 't Veer, L. J., Dai, H., Hart, A. A. M. *et al.*** 2002. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* 347:1999-2009.
 180. **Velculescu, V. E., Madden, S. L., Zhang, L., Lash, A. E., Yu, J. *et al.*** 1999. Analysis of human transcriptomes. *Nat Genet* 23:387-8.
 181. **Venkatasubbarao, S.** 2004. Microarrays--status and prospects. *Trends Biotechnol* 22:630-7.
 182. **Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J. *et al.*** 2001. The sequence of the human genome. *Science* 291:1304-51.
 183. **Vettore, A. L., da Silva, F. R., Kemper, E. L., Souza, G. M., da Silva, A. M. *et al.*** 2003. Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Res* 13:2725-35.
 184. **Wallenius, V., Hisaoka, M., Helou, K., Levan, G., Mandahl, N., Meis-Kindblom, J. M., Kindblom, L. G. & Jansson, J. O.** 2000. Overexpression of the hepatocyte growth factor (HGF) receptor (Met) and presence of a truncated and activated intracellular HGF receptor fragment in locally aggressive/malignant human musculoskeletal tumors. *Am J Pathol* 156:821-9.
 185. **Weber, B. L.** 2002. Cancer genomics. *Cancer Cell* 1:37-47.
 186. **Weigelt, B., Hu, Z., He, X., Livasy, C., Carey, L. A., Ewend, M. G., Glas, A. M., Perou, C. M. & Van't Veer, L. J.** 2005. Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. *Cancer Res* 65:9155-8.

187. **Welsh, J. B., Sapinoso, L. M., Su, A. I., Kern, S. G., Wang-Rodriguez, J., Moskaluk, C. A., Frierson, H. F., Jr. & Hampton, G. M.** 2001a. Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res* 61:5974-8.
188. **Welsh, J. B., Zarrinkar, P. P., Sapinoso, L. M., Kern, S. G., Behling, C. A., Monk, B. J., Lockhart, D. J., Burger, R. A. & Hampton, G. M.** 2001b. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proceedings of the National Academy of Sciences of the United States of America* 98:1176-1181.
189. **Wessels, L. F., Reinders, M. J., Hart, A. A., Veenman, C. J., Dai, H., He, Y. D. & van't Veer, L. J.** 2005. A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics* 21:3755-62.
190. **Yao, M., Tabuchi, H., Nagashima, Y., Baba, M., Nakaigawa, N., Ishiguro, H., Hamada, K., Inayama, Y., Kishida, T., Hattori, K., Yamada-Okabe, H. & Kubota, Y.** 2005. Gene expression analysis of renal carcinoma: adipose differentiation-related protein as a potential diagnostic and prognostic biomarker for clear-cell renal carcinoma. *J Pathol* 205:377-87.
191. **Zhang, J., Carey, V. & Gentleman, R.** 2003. An extensible application for assembling annotation for genomic data. *Bioinformatics* 19:155-6.

Lista de Anexos

(inclusos no CD-ROM complementar a tese)

Anexo 1 - Súmula Curricular

Anexo 2 - Histórico Escolar

Anexo 3 – Simões, A. C. Q., da Silva, A. M., Verjovski-Almeida, S. & Reis, E. M. 2005. SAM Method as an Approach to Select Candidates for Human Prostate Cancer Markers. Lecture Notes in Computer Science 3594/2005:202-205.

Anexo 4 - Simões, A. C. Q. & da Silva, A. M. 2006. SAM as plugin for BASE. Presented at the 4th Annual Internacional Conference On Intelligent Systems For Molecular Biology, ISMB2006, Fortaleza, Ceará, Brasil.

Anexo 5 - Simões, A. C. Q., Esteves, G. H., Cristo, E. B., Fahham, L. V., da Silva, A. M. & Neves, E. J. 2007. Active Modules and Relevance Networks during the cell cycle in HeLa cells. Presented at the The Eighth International Conference On Systems Biology, Long Beach, California, EUA.

Anexo 6 - Esteves, G. H., Simoes, A. C., Souza, E., Dias, R. A., Ospina, R. & Venancio, T. M. 2007. New insights about host response to smallpox using microarray data. BMC Syst Biol 1:38.

Anexo 7 - Cunha, I.W., Carvalho, K. C., Martins, W. K., Marques, S., Falzoni, R., Simões, A.C.Q., Fahham, L., Neves, E. J., Soares, F. A. & Reis, L. F. L. Identification of genes associated with local agressiveness and metastatic behavior in soft tissue sarcomas. Submetido à publicação.

Anexo 8 – Resumos dos Manuscritos em preparação