UM AMBIENTE PARA GERENCIAMENTO E EXECUÇÃO DE EXPERIMENTOS DE EXPRESSÃO GÊNICA COM MICROARRANJOS

Vinicius de Souza Von Held

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Aprovada por:

Profª. Marta Lima de Queirós Mattoso, D.Sc.

Prof. Geraldo Bonorino Xexéo, D. Sc.

Prof. Milton Ozório Moraes, D. Sc.

Prof. Alberto Martín Rivera Dávila, D. Sc.

RIO DE JANEIRO, RJ - BRASIL

JUNHO DE 2007

VON HELD, VINICIUS DE SOUZA

Um ambiente para gerenciamento e execução de experimentos de expressão gênica com microarranjos

[Rio de Janeiro] 2007

XII, 77 p. 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia de Sistemas e Computação, 2007)

Dissertação - Universidade Federal do Rio de Janeiro, COPPE

1. Workflows
2. Bioinformática
3. Microarranjos

I. COPPE/UFRJ II. Título (série)

Aos meus pais Álvaro e Cleide, pelo apoio incondicional

AGRADECIMENTOS

À minha orientadora Marta Mattoso, por aceitar me orientar nesse trabalho multidisciplinar, por me auxiliar sempre na elaboração de idéias e pela confiança na conclusão dessa dissertação.

Aos professores Milton Ozório e Alberto Dávila, por participarem da minha banca, por me receberem tão bem na FIOCRUZ e por serem tão solícitos quando lhes pedi auxílio.

Ao professor Xexéo, por participar da minha banca.

Às pesquisadoras da FIOCRUZ Anna Beatriz Ferreira e Luana Guerreiro pelas valiosas explicações que me deram sobre os experimentos com microarranjos, e por me ajudarem a 'fazer um *upgrade* no meu cérebro para que ele ficasse compatível com os microarranjos'.

À minha amiga e doutoranda de Biomedicina da UFRJ Leandra Baptista, por me ajudar a aumentar meus conhecimentos biológicos.

Aos meus pais Álvaro e Cleide, que sempre apoiaram minhas decisões, mesmo eu não tendo seguido os caminhos mais 'fáceis' e óbvios na minha vida profissional e acadêmica. E por me fornecerem metade de seus genes, que me transformaram em quem sou.

À minha namorada Isa, pela paciência e carinho nesses últimos meses quando estive um pouco ausente.

Aos meus amigos (e afilhados de casamento) Amanda Varella e Vinícios Pereira, pela amizade e companheirismo desde o início da faculdade. À Amanda especialmente pelo grande incentivo e orientação psicológica na fase final deste trabalho.

Ao meu amigo Cláudio Ferraz, colega de trabalho e parceiro de aventuras acadêmicas e turísticas, e também outro grande incentivador à conclusão desse trabalho.

À Camille Furtado e à Rógea Rocha, duas amizades que prezo muito. E também das quais fiz muito uso do 'ouvido amigo'.

A todos os demais amigos do mestrado, da faculdade e da COPPETEC com quem convivi nesses últimos anos.

Ao povo da montanha, os amigos dos grupos Montanhismo & Cia e Trilha & Cia, companheiros de aventuras e perrengues quase tão intensos quanto as atividades do mestrado. Especialmente aos amigos mais próximos: Júnior, Luizão, Xanda, Rose e Lu. Obrigado também por, a meu pedido, me excluírem de todos os 'eventos legais' dos últimos meses.

A Fundação COPPETEC pelas oportunidades profissionais e pelo apoio financeiro.

Ao governo brasileiro, pelo apoio financeiro através da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES.

A Deus, pela vida e pela determinação para conseguir chegar à conclusão deste trabalho.

Obrigado a todos que de alguma forma contribuíram para a realização desta conquista.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

UM AMBIENTE PARA GERENCIAMENTO E EXECUÇÃO DE EXPERIMENTOS DE EXPRESSÃO GÊNICA COM MICROARRANJOS

Vinicius de Souza Von Held

Junho/2007

Orientadora: Marta Lima de Queirós Mattoso

Programa: Engenharia de Sistemas e Computação

O entendimento das funções dos genes é fundamental para o desenvolvimento de novos diagnósticos e terapias baseadas em evidências genéticas. A técnica de microarranjos proporciona a análise de milhares de genes ao mesmo tempo, permitindo a comparação da expressão gênica de células em situações diversas (células saudáveis e tumorais, por exemplo).

Os experimentos com microarranjos são compostos por uma série de atividades em *wet lab* e computacionais (*in silico*). Apesar de existirem ferramentas computacionais e padrões para apoiar essas etapas, há uma carência de apoio na gerência do processo experimental, além da falta de padronização na anotação dos experimentos, tanto em relação ao formato quanto ao conteúdo. O objetivo desta dissertação é desenvolver um ambiente para apoiar os experimentos com microarranjos no sentido de facilitar a composição de processos e dados, o registro de atividades e monitoração dos experimentos, apoiando-se no uso de *workflows* científicos e padrões como o *MGED Ontology* e *Genomics Unified Schema*. Espera-se assim melhorar o grau de integração entre etapas de um experimento e entre experimentos diversos, o que é de extrema importância para uma análise mais ampla das funções dos genes.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

AN ENVIRONMENT FOR MANAGING AND EXECUTIONING GENE EXPRESSION EXPERIMENTS WITH MICROARRAYS

Vinicius de Souza Von Held

June/2007

Advisor: Marta Lima de Queirós Mattoso

Department: Systems and Computer Engineering

Understanding the gene functions is essential for developing new treatments and diagnosis based on genetics evidences. The microarray technique permits the simultaneous analysis of thousands of genes, allowing the comparison of cells' gene expression in distinct situations (e.g., healthy and cancerous cells).

Microarray experiments are composed by wet lab and computational (*in silico*) activities. Despite the existing tools and patterns which deal with these experiments, they lack a support for managing the experimental processes, and also a standard for annotation of experiments (format and content). This work aims at developing an environment for supporting microarray experiments to facilitate the composition of process and data, the registry of activities and the monitoring of experiments, based on scientific workflows and standards such as MGED Ontology and Genome Unified Schema. We expect to improve the integration level among the phases of one experiment and between distinct experiments, which is extremely important for a wider analysis of gene functions.

Índice

Índice de Figuras

Índice de Tabelas

Lista de Siglas

cDNA:	Complementary DNA
DNA (ADN):	Deoxyribonucleic Acid (Ácido Desoxirribonucléico)
DW:	Data Warehouse
GARSA:	Genomic Analysis Resources Sequence Annotation
GEO:	Gene Expression Omnibus
GMOD:	Generic Model Organism Database
GO:	Gene Ontology
GUS:	Genomics Unified Schema
MGED:	Microarray and Gene Expression Data
MIAME:	Minimum Information About a Microarray Experiment
MoML:	Modeling Markup Language in XML
mRNA:	Messenger RNA
OLAP:	Online Analytical Processing
PCR-RT:	Polymerase Chain Reaction-Reverse Transcription
RAD:	RNA Abundance Database
RAD-SA:	RAD Study-Annotator
RNA (ARN):	Ribonucleic Acid (Ácido Ribonucléico)
SGBD:	Sistema Gerenciador de Banco de Dados
SNOMED:	Systematized Nomenclature of Medicine
XML:	EXtensible Markup Language

Capítulo 1 – Introdução

"Com certeza eles não morreram pois são antigos mestres na arte da sobrevivência. (...) Mas não os procure flutuando livremente no mar. Eles abandonaram essa liberdade há muito tempo. Agora eles apinham-se em colônias imensas, vivendo com segurança dentro de robôs desajeitados gigantescos, murados do mundo exterior, comunicando-se com ele por meio de vias indiretas e tortuosas, manipulando-o por controle remoto. Eles estão em mim e em você. Eles nos criaram, corpo e mente. E sua preservação é a razão última de nossa existência. Transformaram-se, esses replicadores. Agora eles recebem o nome de genes e nós somos suas máquinas de sobrevivência."

Richard Dawkins, O gene egoísta

1.1 – Motivação

As pesquisas na área de genética têm crescido enormemente nos últimos dez anos (PROSDOCIMI, FILHO et al., 2003). Grande parte dessas pesquisas tem a finalidade de identificar genes e sua atuação no funcionamento dos organismos. Com o seqüenciamento genético de um número cada vez maior de espécies e o aprimoramento e criação de técnicas para análises primárias dos genes, a geração de dados nessa área de pesquisa vem crescendo de forma exponencial. No ano de 1999, o banco de dados EMBL (European Bioinformatics Institute) possuía menos de 6 milhões de seqüências genéticas em sua base de dados. Em 2003, possuía algo em torno de 34 milhões e hoje já são quase 100 milhões de seqüências armazenadas (EMBL, 2007c). Para acompanhar esse ritmo de crescimento, bancos e ferramentas para tratamento desses dados vêm se aperfeiçoando e se tornando mais robustos, rápidos e com mais funcionalidades. Resultados de pesquisas são disponibilizados cada vez mais rápido e podem ser utilizados para novas análises em diversos aplicativos desenvolvidos para fins de estudo biológico. E associado a esse crescimento, está o desenvolvimento da bioinformática, uma área multidisciplinar que reúne conhecimentos de outras áreas de conhecimento, como as ciências da computação e a estatística, além da biologia. O profissional de

bioinformática deve ser capaz de desenvolver soluções computacionais eficientes levando em conta detalhes e incertezas dos eventos biológicos.

Um esforço muito grande, bastante influenciado pela conclusão do seqüenciamento do genoma humano, tem sido concentrado na descoberta das funções dos genes, já que a mesma pode auxiliar no desenvolvimento de tratamentos preventivos e terapêuticos de doenças graves como o câncer. Os genes são as unidades mínimas de armazenamento de informação que regulam o funcionamento de um organismo, normalmente agindo em conjunto, ativando ou reprimindo uns aos outros. O conhecimento dessas etapas de "ligar" e "desligar", conhecido como regulação da expressão gênica pode, em última análise, definir os processos de desenvolvimento e diferenciação celular. Para adquirir tal conhecimento, uma técnica que tem sido bastante utilizada para análise de expressão gênica é a de microarranjos, que permite estudar a expressão (quais genes estão ativos/reprimidos) de milhares de genes de um tipo celular ou tecido ao mesmo tempo, ao invés do estudo individual de cada gene (MANDUCHI, GRANT et al., 2004; PIATETSKY-SHAPIRO e TAMAYO, 2003). Esta técnica permite comparar, por exemplo, o padrão de expressão gênica de uma célula saudável com o de uma tumoral.

Os experimentos com microarranjo, assim como outros experimentos biológicos, são compostos por uma série de atividades experimentais, tanto em bancada (*in vitro*) quanto em laboratórios computacionais (*in silico*). Há um processo experimental, caracterizado por um *workflow científico,* que muitas vezes não é explicitamente percebido pelo pesquisador. Usualmente cada etapa deste processo é realizada de maneira isolada e sem integração automatizada entre elas. Além disso, a falta de acompanhamento do processo não permite rastrear a ocorrência de sucesso ou eventuais erros e dificulta a re-execução de partes do processo, ou mesmo de todo ele. A falta de padronização na anotação de tais experimentos também consiste numa dificuldade para integração e compartilhamento de experimentos diversos.

São esses problemas citados acima que serão abordados nessa dissertação. Elaboraremos uma ferramenta para apoiar computacionalmente e integrar todo o processo experimental, incluindo as atividades laboratoriais de bancada e as computacionais. Essa ferramenta é especializada para os experimentos com microarranjos, pois vamos considerar particularidades deste tipo de experimento.

1.2 – Caracterização do problema

Com o crescente interesse pela tecnologia de microarranjos e o aumento do número de experimentos deste tipo realizados (JONATHAN KNIGHT, 2001), diversos pesquisadores começaram a tentar replicar experimentos uns dos outros. Mas por se tratar de uma tecnologia recente, não havia, a princípio, práticas comuns para documentar o experimento. Ou seja, um experimento podia conter uma vasta documentação, mas não conter dados necessários para sua replicação, ou mesmo não conter dados informados em outros experimentos, caso o pesquisador tivesse a intenção de realizar um estudo comparativo (NATURE, 2002). Dado esse problema, o grupo *Microarray Gene Expression Data Society* (MGED, 2006a) elaborou o *Minimum Information About a Microarray Experiment* (MIAME) (BRAZMA, HINGAMP et al., 2001), que relaciona as informações mínimas a respeito de um experimento com microarranjo, de forma que seja possível seu completo entendimento e replicação de seus resultados. Usualmente, tais informações são armazenadas em repositórios públicos de experimentos, como o GEO (GEO, 2006) e o *Array Express* (PARKINSON, SARKANS et al., 2005), e são requisitos para a publicação nas principais revistas científicas da área (BRAZMA, HINGAMP et al., 2001).

Entretanto, esses dados são normalmente armazenados de forma não estruturada (normalmente em textos livres) e sem padrões rigorosos para nomenclatura. Isso torna mais complicada a análise interexperimental, que nesse tipo de experimento é fundamental, já que diversos grupos realizam de forma independente estudos sobre os mesmo genes, que poderiam ser comparados a fim de se obter conclusões adicionais ou confirmação de resultados através, por exemplo, de meta-análise. Abordando essa problemática, o *Genomics Unified Schema* (GUS) (GUS, 2006) define um amplo esquema lógico relacional que, associado a uma série de aplicações, objetiva armazenar e manter diversos tipos de dados genômicos. O RNA *Abundance Database* (RAD) é um subesquema do GUS especializado em dados de experimentos com microarranjos, e que contempla as especificações do MIAME. Já o problema da falta de padrão de nomenclatura pode ser satisfatoriamente contornado com o uso de vocabulário controlado baseado em ontologias, que podem auxiliar no uso de uma descrição bem definida e não-ambígua do experimento. A *Gene Ontology* (GO) (GO, 2006) define uma ontologia especializada para a área de genética e o MGED *Ontology* (OWG, 2006) é especializado em anotações para experimentos com microarranjo. A abordagem

relacional do GUS, associada à aplicação de vocabulário controlado, utilizada, por exemplo no sistema RAD *Study Annotator* (MANDUCHI, GRANT et al., 2004), contribui para a estruturação e padronização de dados, que é um passo importante rumo a integração de informações provenientes de experimentos diversos.

Um outro grupo de ferramentas são aquelas utilizadas para gerenciar e integrar as diversas atividades que integram o processo experimental. O conjunto dessas atividades pode ser visto como um *workflow* científico, pois se compõe de tarefas repetitivas e interdependentes, e que representam uma unidade de trabalho. Além disso, trata-se de uma aplicação científica, que exige alto poder computacional, utilizada para validar descobertas na forma de algum modelo científico e aplicadas a um domínio de conhecimento específico (SANTOS, 2004). A gerência deste *workflow* viabiliza a integração de experimentos diversos uma vez que provê controle sobre os parâmetros de execução das etapas. Além disso, permite monitorar a proveniência dos dados, que é um indicativo da validade e qualidade dos dados (BUNEMAN, KHANNA et al., 2001). O reaproveitamento de dados, a re-execução de experimentos, ou elaboração de novos, têm um custo bem menor se apoiadas por um gerenciador automático de processos. O Kepler (KEPLER, 2006) e o *myGrid* (MYGRID, 2006) são ambientes que permitem a especificação e execução de *workflows* científicos. O GARSA (DÁVILA, LORENZINI et al., 2005) é uma ferramenta de gerência e execução de experimentos *in silico* de análises genômicas e anotações que integra fontes de dados e ferramentas diversas. Esses três sistemas, entretanto, cobrem somente as etapas *in silico* do processo experimental. Além disso não são aderentes aos padrões de metadados como o GUS, o MIAME e o MGED Ontology.

Completando as ferramentas computacionais utilizadas neste tipo de experimento, há aquelas utilizadas para realizar tarefas do processo experimental, como os sistemas que controlam equipamentos que processam objetos do experimento; ou aquelas usadas para realizar os experimentos *in silico*, utilizando abordagens estatísticas (MATHWORKS, 2006; R, 2006) ou outras, como *data mining* (PIATETSKY-SHAPIRO e TAMAYO, 2003) e data *warehousing* (KIRSTEN, DO et al., 2004) .

Vimos que há várias soluções computacionais aplicáveis aos experimentos com microarranjos. Entretanto, elas isoladamente não são atendem a diversos requisitos deste tipo de experimento. É desejável que um sistema gerencial de microarranjos seja capaz

de integrar todas as suas etapas, assim como os dados gerados por cada uma delas e também por elas utilizados.

1.3 – Objetivos

O objetivo desta dissertação é apoiar os experimentos com microarranjos no sentido de facilitar a composição de processos e dados, o registro de atividades e monitoração dos experimentos. Para tanto, visamos sistematizar as atividades que fazem parte do processo experimental com microarranjos e desenvolver um ambiente para gerência e execução de um workflow científico especializado nesse tipo de experimento.

A arquitetura deste ambiente visa integrar diversas ferramentas computacionais que atendem parcialmente os diversos requisitos exigidos para a gerência de experimentos com microarranjos. Essa integração vai gerar uma série de funcionalidades que atendem de forma satisfatória os principais requisitos para execução deste tipo de experimento: o sistema desenvolvido fará uso de padrões aceitos na comunidade científica, como o GUS/RAD, o MIAME e a MGED *Ontology*, facilitando, com a padronização de formato e de conteúdo, a troca de experiências de forma automatizada; as atividades *in vitro* e *in silico* do experimento serão contempladas; aplicativos diversos que realizam etapas do experimento serão integrados e será possível gerenciar a seqüência de execução dos mesmos, possibilitando o controle de fluxo das atividades; o registro de todos os passos do experimento e dos dados neles utilizados, permite o monitoramento do fluxo de dados, sendo assim possível verificar a qualidade dos mesmos através de seu histórico e rastrear a origem de eventuais erros.

O uso desse ambiente deve agilizar a execução dos experimentos, uma vez que promove a organização dos dados utilizados, facilita o acesso a eles e faz a manipulação automática dos mesmos para que possam ser utilizados por diferentes ferramentas. A interação do sistema com o usuário foi elaborada de modo que o pesquisador só precise registrar dados que não possam ser automaticamente capturados e deixando a manipulação de grandes volumes de dados como tarefas do sistema. Isso reduz a possibilidade de inserção de erros e inconsistências na base dados. É esperado também que o sistema facilite a elaboração de estudos adicionais, com a combinação de dados de diversos experimentos, já que a disponibilidade de metadados vai permitir que se

verifique que diferentes grupos de dados são comparáveis, ou fornecer meios para que se tornem comparáveis.

1.4 – Organização da dissertação

O restante desta dissertação está organizado da seguinte forma.

No capítulo 2 será feita, primeiramente, a revisão dos conceitos teóricos de biologia e informática relacionados ao tema desta dissertação. Falaremos sobre biologia molecular e genética, que envolvem conceitos básicos necessários ao entendimento dos experimentos com microarranjos. Em seguida, daremos uma visão geral sobre bioinformática, área da qual faz parte o tema dessa dissertação. A seguir falaremos sobre os experimentos com microarranjos, explicitando suas etapas. Depois falaremos sobre workflows científicos, modelo sobre o qual elaboramos a arquitetura de um sistema gerenciador de processos experimentais com microarranjos. Finalmente faremos uma discussão sobre os trabalhos relacionados a este tema.

No capítulo 3, apresentaremos uma modelagem do processo experimental com microarranjo através de um *workflow* científico. Caracterizaremos melhor o problema de gerência de processos experimentais e mostraremos como usualmente esse processo é executado, identificando os principais problemas que ocorrem na sua realização. Em seguida será proposta a arquitetura Wflex que visa atender os requisitos computacionais deste tipo de experimento.

No capítulo 4, explicaremos como foi feita a implementação de um protótipo da Wflex, e as tecnologias utilizadas. Em seguida falaremos sobre o estudo de caso através do ambiente real de experimentos onde o protótipo desenvolvido foi implantado e como se dará a utilização desse protótipo.

No capítulo 5, serão feitas as conclusões e discussões a respeito do desenvolvimento dessa dissertação, e também falaremos sobre possíveis trabalhos futuros.

Capítulo 2 – Experimentos com Microarranjos

O objetivo deste capítulo é apresentar conceitos teóricos básicos para o entendimento de experimentos biológicos, especialmente os de microarranjo; e discutir trabalhos, de informática e bioinformática, relacionados a este tipo de experimento. Nas seções 2.1, 2.2 e 2.3, apresentaremos conceitos relacionados à biologia molecular, bioinformática e experimentos com microarranjos, respectivamente, com o objetivo de fornecer informações básicas para o entendimento do trabalho proposto. Nas seções 2.4 e 2.5, mostraremos como se deu a evolução do apoio computacional aos experimentos com microarranjos, discutindo os principais trabalhos relacionados. Na seção 2.6, discutiremos a utilização de *workflows* científicos em experimentos com microarranjos. Na seção 2.7, faremos uma análise comparativa dos principais trabalhos relacionados ao tema dessa dissertação. Finalmente, na seção 2.8, faremos algumas considerações finais sobre o que foi discutido no capítulo.

2.1 – Biologia Molecular

Nessa seção explicaremos brevemente os processos biológicos que permitem aos genes reger todo o funcionamento de um organismo, além de apresentar as principais aplicações da área de pesquisa conhecida como genética.

2.1.1 – Código genético

Todos os seres vivos, desde os simples vírus até o complexo organismo dos seres humanos, possuem uma macromolécula que contém todas as instruções necessárias para a construção e funcionamento de todo organismo. No caso dos vírus e seres unicelulares, existe apenas um exemplar dessa estrutura. Mas nos demais seres, os chamados pluricelulares, existe uma cópia da mesma contida em cada célula do organismo. Essa molécula é chamada DNA (sigla em inglês para ácido desoxirribonucléico). Nos seres eucariotos, as células possuem um núcleo, separado do citoplasma por uma membrana, onde fica localizada essa molécula. Nos organismos cujas células não possuem núcleo, o DNA fica no citoplasma, juntamente com as demais substâncias intercelulares.

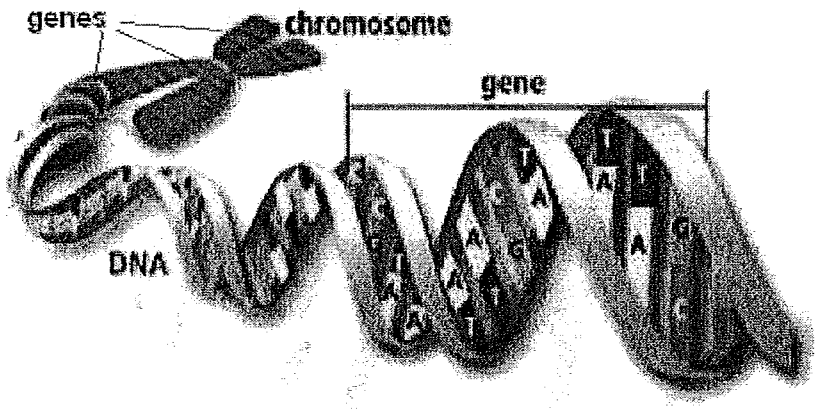

Figura 1: Macromolécula de DNA

O nucleotídeo é a unidade estrutural do DNA. Esta molécula é formada por duas longas cadeias complementares de nucleotídeos. Um nucleotídeo é formado por 3 componentes: um açúcar (a desoxirribose), um grupo fosfato (H3PO4) e uma base nitrogenada, que pode ser: Timina, Adenina, Guanina ou Citosina. Como é a base nitrogenada que diferencia cada unidade estrutural do DNA, a seqüência de nucleotídeos de uma cadeia é representada pelas iniciais dessas bases. Em geral, essa seqüência não se constitui em um único filamento, mas em diversos que se agrupam em estruturas chamadas cromossomos. O número de cromossomos em cada célula varia de acordo com a espécie. Como já foi dito, o DNA é formado por duas seqüências de nucleotídeos. Essas duas fitas se combinam da seguinte maneira: as bases timina (T) e adenosina(A) sempre se combinam entre si, o mesmo ocorrendo para as bases citosina(C) e guanina(G). Assim, diz-se que as duas seqüências de nucleotídeos são complementares e a partir de uma é possível se deduzir a seqüência da outra. Por isso a seqüência de um DNA é representada por uma única seqüência de bases nitrogenadas. É essa seqüência de bases (representada por letras) que chamamos de código genético.

O DNA é um modelo para a construção de proteínas, ou seja, ele possui o código a partir do qual são geradas as proteínas. As proteínas são estruturas essenciais presentes nas células, pois estas macromoléculas desempenham inúmeras funções, a maioria vital aos organismos vivos. Proteínas compõem estruturas de suporte a tecidos, são catalisadoras de reações bioquímicas e agentes do sistema imunológico, dentre outras tarefas que podem realizar.

O DNA não é traduzido diretamente para proteínas, e nem toda a extensão do mesmo é utilizada para codificar uma proteína. Na verdade apenas pequenos pedaços do mesmo (e nesses pedaços, apenas um dos dois filamentos) servirão de modelo à construção de uma proteína. Esses pedaços são chamados de genes. O conjunto de genes de um organismo é chamado de genoma. Os genes são primeiramente traduzidos para uma estrutura chamada RNA (sigla em inglês para ácido ribonucléico), que é semelhante ao DNA, porém possui apenas uma cadeia de nucleotídeos e o açúcar que compõe cada nucleotídeo é a ribose. Cada gene dá origem a um RNA específico. A transcrição é o processo que origina um RNA ao fazer a cópia de um gene para uma seqüência complementar ao mesmo. Ou seja, a seqüência origem (o gene) é copiada, trocando-se as bases A, G, T, C por U (Uracila), C, A e G, respectivamente. Como os filamentos do DNA são complementares, a seqüência de bases nitrogenadas da fita de RNA transcrita é idêntica ao outro filamento de DNA (aquele que não serviu de molde), excetuando-se pela base Timina, que no RNA é substituída pela Uracila. O RNA que é traduzido para proteína é do tipo mensageiro (mRNA). Existem outros tipos de DNA: o transportador e o ribossomial, que participam do processo de tradução do RNA mensageiro para uma proteína.

A tradução é o processo que gera uma proteína a partir de um RNA mensageiro e ocorre regida por uma estrutura celular chamada ribossomo. Outra estrutura que participa desse processo é o RNA transportador (tRNA), que possui uma trinca de bases nitrogenadas chamada anti-códon e também, ligada a ele, uma molécula chamada aminoácido. Cada RNA transportador possui um anti-códon e um aminoácido específico. O ribossomo faz com que cada trinca de bases nitrogenadas do RNA mensageiro (chamada códon) se ligue a um anti-códon com a seqüência complementar ao códon. Assim, os RNAs transportadores vão se ligando na seqüência de códons do RNA mensageiro e os aminoácidos vão se 'soltando' dos RNAs transportadores e se ligando entre si, através das chamadas ligações peptídicas. Ao final do processo, formam a proteína (com centenas ou milhares de aminoácidos), que toma uma conformação espacial que é determinante na função da mesma, já que muitas proteínas têm suas funções baseadas em encaixe com outras estruturas (outras proteínas inclusive).

Esse fluxo da informação contida no DNA (o código genético), até a síntese da proteína, passando pelo RNA mensageiro, é conhecido como Dogma Central da Biologia Molecular (CRICK, 1970).

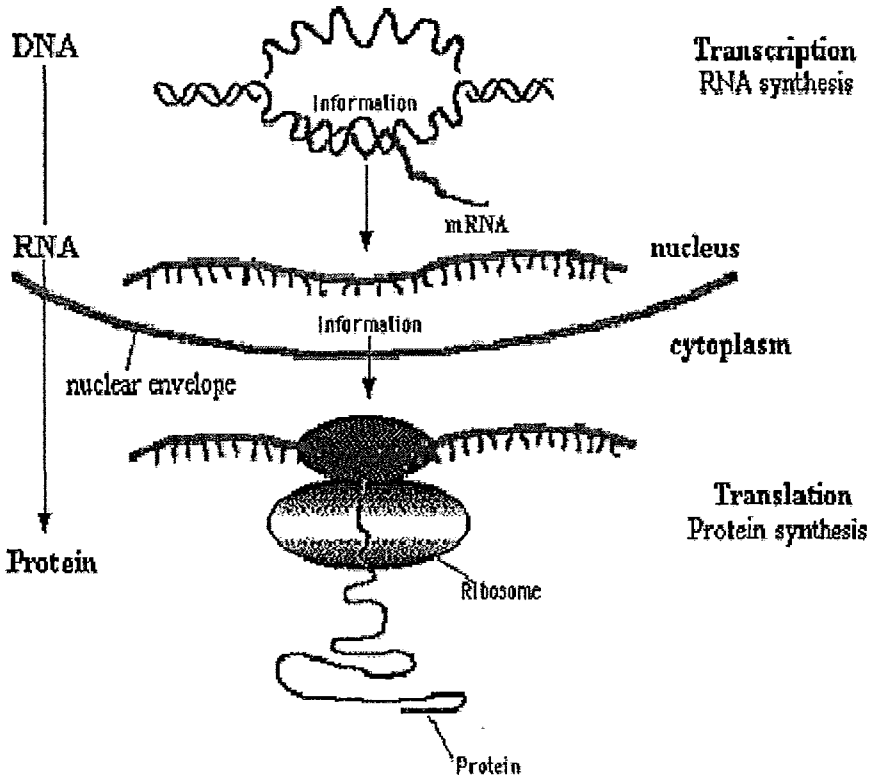

Figura 2: Dogma Central da Biologia Molecular

A função de um gene, na verdade, nada mais é que a função da proteína que este gene origina, e são elas as responsáveis por muitos dos processos e ciclos vitais que regem um organismo, como hormônios, enzimas, e anticorpos. Por sua vez, a função de uma proteína está intimamente ligada com a sua estrutura tridimensional. As enzimas, por exemplo, que funcionam como catalisadores de reações químicas, combinam-se apenas com outras estruturas (moléculas) nas quais possam se encaixar. Os anticorpos também são exemplos disto, já que os mesmos atacam apenas moléculas específicas (de bactérias, vírus, tecidos) baseado em suas estruturas espaciais. Portanto, a resolução do trinômio "Função do Gene-Função da Proteína-Estrutura da Proteína" se tornou um dos tópicos de pesquisa mais importantes da atualidade (HEAD-GORDON e WOOLEY, 2001).

2.1.2 – Genética

Genética é ciência dos genes. Ela se subdivide em algumas áreas, mas é com a genética molecular que o tema dessa dissertação mais se relaciona, já que ela se foca na estrutura e função dos genes ao nível molecular. Alguns temas que envolvem genética, como clonagem e alimentos transgênicos, se tornaram relativamente populares e a menção desses termos já não causam tanta estranheza ao público leigo. Entretanto, há outros temas não tão popularizados, mas que são alvo de diversos estudos nessa área. É o caso dos chamados tratamentos genéticos, que são métodos, preventivos ou terapêuticos, desenvolvidos baseados no entendimento das funções dos genes, ou seja, o papel que eles têm no desenvolvimento de um organismo. Grande parte das pesquisas em genética tem a finalidade de identificar genes e sua atuação no funcionamento dos organismos. Diversos outros benefícios podem ser alcançados como resultados das pesquisas na área de genética (HGP, 2006b).

Podemos ver no caso dos alimentos transgênicos como é utilizado um gene que tem sua função identificada. Um transgênico é, por definição, um organismo que possui genes de outras espécies, inseridos através de técnicas de recombinação de DNA. Em plantas podem ser acrescentados genes capazes de aumentar a resistência da mesma a uma determinada praga, aumentando assim a produtividade de alimentos e diminuindo o uso de agrotóxicos (TODABIOLOGIA, 2006).

Uma das grandes conquistas da genética foi a conclusão do Projeto Genoma Humano (HGP, 2006a), que seqüenciou 3 bilhões de nucleotídeos e identificou cerca de 25 mil genes, dos quais menos da metade têm funções conhecidas. Dentre os principais desafios pós-seqüenciamento encontra-se a definição dessas funções. Diversas outras espécies já tiveram seus genomas seqüenciados, como o do verme *Caenorhabditis elegans*, da mosca da fruta (*Drosophila melanogaster*) e do camundongo (*Mus musculus*) (ENSEMBL, 2006).

2.2 – Bioinformática

Em uma definição rígida, bioinformática é qualquer uso de computadores para lidar com informações biológicas. Mas na prática, como veremos nessa seção, a bioinformática se desenvolve no sentido de utilizar diferentes áreas do conhecimento, como a ciência da computação, a matemática, a estatística e a física, para analisar e

comparar seqüências genômicas ou protéicas com a intensa utilização das ferramentas da computação, tais quais bancos de dados, redes e inteligência artificial.

Com o seqüenciamento genético de um número cada vez maior de espécies e o aprimoramento e criação de técnicas para análises primárias dos genes, a geração de dados nessa área de pesquisa vem crescendo de forma exponencial. No ano de 1999, o banco de dados EMBL (EBI, 2006) possuía menos de 6 milhões de seqüências genéticas em sua base de dados. Em 2003, possuía algo em torno de 34 milhões e hoje já são quase 100 milhões de seqüências armazenadas. Além do armazenamento, ocorre paralelamente a necessidade de análise desses dados, o que torna indispensável a utilização de plataformas computacionais eficientes para a interpretação dos resultados obtidos. O surgimento e o desenvolvimento da bioinformática são resultados diretos dessas necessidades. Essa nova ciência envolve a união de diversas linhas de conhecimento, a engenharia de softwares, a matemática, a estatística, a ciência da computação e a biologia molecular (PROSDOCIMI, FILHO et al., 2003). A bioinformática é o instrumento que a comunidade de biologia molecular espera poder contar para transformar em informação útil (e valiosa) a imensa quantidade de dados produzidos pelos projetos de seqüenciamento de genoma.

Outro fator que levou ao desenvolvimento da bioinformática foi uma certa dificuldade de comunicação devido a abordagens distintas adotadas por profissionais de diferentes áreas: enquanto o biólogo procurava uma solução que levasse em consideração as incertezas e erros que ocorrem na prática, o cientista da computação procurava uma solução eficiente para um problema bem definido. Assim, surgiu a necessidade de um novo profissional, que entendesse bem ambas as áreas e fizesse a ponte entre elas. O bioinformata deve ter conhecimento suficiente sobre problemas biológicos reais para poder buscar soluções computacionais viáveis ao problema em questão (PROSDOCIMI, FILHO et al., 2003).

Algumas áreas da pesquisa em bioinformática merecem destaque devido a sua importância e a quantidade de estudos realizados. São elas: banco de dados de biologia molecular; comparação e análise de seqüências; e predição de estrutura de proteínas. .

Os bancos de dados de biologia molecular são grandes repositórios de dados (não necessariamente sistemas gerenciadores de banco de dados), normalmente estão

disponíveis nas *web* para consultas e são abertos a laboratórios do mundo todo para depósito de dados. Os principais tipos são os bancos de seqüências genéticas e de proteínas, como o GenBank (BENSON, KARSCH-MIZRACHI et al., 2007) e *Swiss-Prot* (BOECKMANN, BAIROCH et al., 2003) e os de estruturas tridimensionais de proteínas, cujo principal representante é o *Protein Data Bank* (WESTBROOK, FENG et al., 2002). Outros tipos de bancos armazenam dados de um tipo de experimento específico, como é o caso do *Gene Expression Omnibus* (GEO, 2006), que é um repositório de dados de experimentos de expressão gênica com uso de microarranjos.

A predição da conformação espacial de proteínas também é um problema corrente na bioinformática. O objetivo desses estudos é, dada uma seqüência de aminoácidos, (que pode ser obtida a partir da decodificação de uma seqüência genética), prever que formato espacial essa proteína vai tomar. Um método bastante utilizado para isto é a análise de outras proteínas, cujas seqüências de aminoácidos são semelhantes e estruturas espaciais já são conhecidas. Os critérios para se definir o quão semelhante uma estrutura é de outras são objetos de vários estudos, dentre eles: (BUHLER, 2003; CAMOGLU, KAHVECI et al., 2003; HUNT, ATKINSON et al., 2002). Os problemas citados, somados a grande quantidade de proteínas conhecidas, constituem uma forte motivação para o desenvolvimento de ferramentas integradas para armazenamento, indexação, consulta e análise de proteínas, tanto da estrutura primárias (seqüência de aminoácidos), quanto da espacial das mesmas.

A análise de seqüências ocorre principalmente através da comparação com outras seqüências já conhecidas, já que a semelhança entre seqüências pode ajudar na descoberta de relação funcional e evolucionária entre elas, assim como auxiliar no agrupamento das mesmas em famílias (ALTSCHUL, GISH et al., 1990). A comparação de seqüências se dá principalmente através de algoritmos que usam a idéia de alinhamento entre elas: uma seqüência a ser estudada é confrontada com milhões de outras seqüências armazenadas em bancos de dados, na tentativa de alinhá-la com um trecho de uma seqüência maior ou mesmo com toda a extensão da mesma. Como a quantidade de dados nesses bancos é imensa, esses algoritmos exigem um alto poder computacional para que sejam rodados. Uma das ferramentas de comparação de seqüências mais conhecidas e utilizadas é o Blast (ALTSCHUL, GISH et al., 1990).

Os experimentos com microarranjo fazem parte dos estudos de comparação e análise de seqüências, sendo, na verdade, um complemento a elas. Na próxima seção será explicado em detalhe esse tipo de experimento.

2.3 – Experimentos com microarranjos

Dentro do contexto de pesquisas sobre funções dos genes estão os experimentos com microarranjos. Embora todas as células de um organismo possuam o mesmo material genético, sua expressão gênica, ou seja, seu conjunto de genes ativos (que codificam proteínas), varia de acordo com o estágio de desenvolvimento, tecido, idade e condições do ambiente. O objetivo dos estudos com microarranjos é, portanto, determinar o padrão de expressão gênica sob uma determinada condição, comparando-o a expressão gênica da mesma célula sob condições consideradas normais.

Os genes são as unidades mínimas de armazenamento de informação que regulam a formação e o funcionamento de um organismo. Ao contrário do que pode parecer, eles não agem isoladamente, já que uma das diversas funções das proteínas produzidas a partir dos genes, é influenciar outros genes, acionando-os ou reprimindo-os (HUNTER, 1993). Assim, podemos dizer que os genes normalmente agem em conjunto, ativando ou reprimindo uns aos outros, determinando o surgimento de uma condição em um organismo, como as características físicas de uma pessoa e até mesmo o desenvolvimento de uma doença. O conhecimento dessas etapas de "ligar" e "desligar" genes, conhecido como regulação da expressão gênica pode, em última análise, definir os processos de desenvolvimento e diferenciação celular. A expressão gênica pode ser medida através da quantificação de RNA mensageiro presente na célula, já que esse tipo de RNA é exatamente o resultado da transcrição dos genes ativos da célula.

A utilidade desta técnica é, então, avaliar que genes estão ativos (ou deixaram de atuar) quando ocorrem mudanças no ambiente intercelular, como por exemplo, um câncer em desenvolvimento em um tecido ou uma célula posta em contato com uma substância tóxica, como em (MITCHELL, BROWN et al., 2004; YU, CHEN et al., 2004). Esse experimento pode sugestionar, então, que esses genes têm a função relacionada ao surgimento desta nova condição. O entendimento deste mecanismo

gênico pode levar a diagnósticos mais seguros, assim como auxiliar no desenvolvimento de tratamentos preventivos e terapêuticos.

2.3.1 – Etapas do experimento com microarranjos

Assim como a maioria dos experimentos biológicos, os experimentos com microarranjos são formados por várias etapas. Os experimentos em microarranjo são compostos por dois tipos de atividades: os experimentos *in vitro*, e as análises computacionais de dados, os chamados experimentos *in silico*. O ciclo de atividades deste tipo de experimento foi descrito em alto nível em (O'CONNELL, 2003). Uma descrição detalhada deste tipo de experimento é apresentada em (GE, 2002).

A primeira etapa do experimento consiste no planejamento do mesmo: são descritos os objetivos dos experimentos, os biomateriais (espécie, tecido, estágio de desenvolvimento, etc.) e os ambientes de cultura das amostras a serem comparadas. São definidas as plataformas do experimento e os protocolos serão utilizados em outras etapas, como a da hibridização e da extração de amostras. A definição quanto ao número de amostras (replicatas biológicas) e o número de vezes que cada uma será testada é uma decisão importante e que dependem do objetivo do experimento que esta sendo planejado (PAN, LIN et al., 2002; ZIEN, FLUCK et al., 2002). Os genes que serão estudados no experimento devem ser selecionados nesta etapa e, dentre os critérios de seleção, pode se incluir o resultado de análise de resultados de experimentos com microarranjos anteriores. Enfim, todos os detalhes do experimento devem ser definidos e documentados nessa etapa.

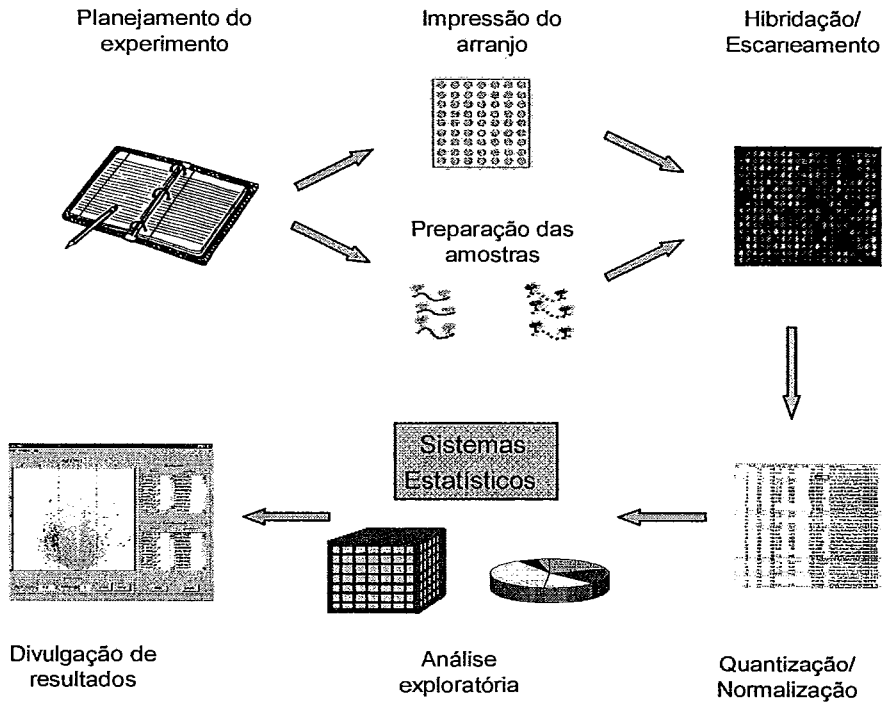

Figura 3: Etapas do experimento com microarranjos

O processo experimental prossegue com a **preparação do microarranjo**, com a impressão, em uma pequena lâmina retangular, das seqüências de DNA representativas dos genes selecionados para estudo (chamadas de sondas), dispostas na forma de uma matriz bidimensional. As sondas são seqüências de dois tipos principalmente. (WOO, AFFOURTIT et al., 2004). A primeira é uma seqüência de DNA complementar (cDNA), obtidos a partir da molécula de DNA através de técnicas como o PCR-RT (PROTOCOL ONLINE, 2007b). Os cDNAs possuem seqüências longas, de centenas ou milhares de bases nitrogenadas, e cobrem praticamente toda a extensão do gene que representa. O outro tipo de seqüência são os oligonucleotídeos, de poucas dezenas de bases, desenhados de forma a se combinarem a genes bem mais extensos que eles próprios, porém em regiões únicas do gene, de modo que eles sejam capazes de identificar um único gene. A impressão das sondas (oligonucleotídeos ou cDNA) nas lâminas é feita por um robô que mergulha pequenos pinos nas soluções, que estão armazenados em placas, divididas em dezenas ou centenas de pequenos poços que contém essas seqüências, e deposita pequenas quantidades das mesmas na lâmina. É possível também a utilização de arranjos comerciais, como o Affymetrix (AFFYMETRIX, 2006), que já são vendidas 'impressas', prontas para a utilização na próxima etapa.

Em paralelo a preparação do microarranjo, é feita a **preparação das amostras** de células. Normalmente elas são analisadas duas a duas, considerando-se amostras do mesmo tecido e de seres da mesma espécie, porém submetidas a diferentes condições. Pode-se, por exemplo, confrontar um tecido saudável com um doente; ou então confrontar amostras de células de um ser que vive em um ambiente limpo com as de outro que habita um local contaminado com uma determinada substância; até mesmo células que se desenvolveram sob as mesmas condições, porém de diferentes fases do desenvolvimento do organismo. Pode ocorrer também a comparação entre diferentes tecidos de uma mesma espécie ou até mesmo entre amostras de espécies distintas (ADJAYE, HERWIG et al., 2004). O importante é que haja variação significativa de apenas um aspecto, que é aquele sobre o qual se deseja avaliar a influência sobre a expressão gênica da célula. Deve-se manter, portanto, um controle sobre os demais aspectos para que os mesmos não modifiquem significativamente a ação dos genes nas células analisadas. As moléculas de RNA das amostras são extraídas (PROTOCOL ONLINE, 2007a) e, para diferenciar as moléculas de cada amostra, cada uma delas é marcada com substâncias fosforescentes de cores distintas, normalmente cores verde e vermelha. A partir deste ponto, podem-se juntar as amostras em um único recipiente, já que as moléculas de RNA de cada uma estão identificadas por uma cor.

A seguir, é feita a **hibridização** das amostras com a lâmina do microarranjo. Seguindo um determinado protocolo, a solução de amostras é mantida em contato com o microarranjo. Suponha o microarranjo representando os genes G1, G2, ..., Gn sendo hibridizado contra as amostras A e B, marcadas com as cores verde e vermelha, respectivamente. Cada molécula de RNA das amostras vai se combinar com a seqüência representativa do gene que a gerou e que está 'impressa', já que essas duas seqüências são necessariamente complementares. Após essa etapa, a lâmina será escaneada com laser e os fluoróforos emitidos em cada posição Gn do microarranjo indicará a quantidade da molécula de RNA através da intensidade de fluorescência detectada. Programas especializados fazem, então, a digitalização da lâmina hibridizada, e a cor que estiver presente em cada ponto da lâmina, indica a presença ou não do RNA, e conseqüentemente se o gene ali representado está sendo expresso em alguma amostra. A coloração eletrônica é meramente didática, mas quando Gn apresenta a cor verde ou vermelha, significa que o mesmo se expressou em apenas uma das amostras. Se Gn adquiriu a cor amarela (combinação de verde e vermelho), é um indício que o gene

estava expresso nas amostras A e B; mas se Gn manteve a cor original da lâmina (cor de fundo), fica indicada a inatividade do gene em ambas as amostras. Assim, os pontos verdes e vermelhos são os mais significativos para o experimento, pois indicam os genes que passaram a se expressar ou ficaram inativos de uma amostra para a outra. Ou seja, são os que demonstram a variação da expressão gênica influenciada pela variação da condição experimental.

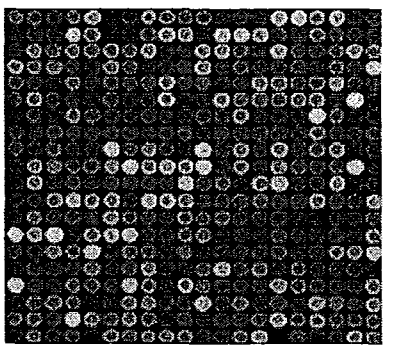

Figura 4: Ampliação de parte de um microarranjo hibridizado

É feita, a seguir, a **quantificação** de RNA em cada ponto Gn da lâmina, através da intensidade de cada ponto no arranjo. Apesar de misturadas, é possível quantificar separadamente o RNA de cada amostra, para cada Gn, devido a diferente freqüência dos fluoróforos utilizadas na marcação das amostras.

Nesse ponto do experimento, obtém-se uma tabela de genes com a quantidade dos respectivos RNAs em cada amostra. Um pré-processamento é feito sobre esses dados antes de se iniciar a análise exploratória sobre os mesmos (O'CONNELL, 2003). Basicamente, deve se normalizar os dados, filtrar dados de má-qualidade e finalmente unificar dados referentes a múltiplos pontos referentes ao mesmo gene.

A **normalização** (SMYTH e SPEED, 2003) é realizada para reajustar de dados, corrigindo variações decorrentes de efeitos não biológicos, com o objetivo de tornar dados de diferentes amostras comparáveis. Os principais efeitos que provocam essas distorções são: a cor de fundo da lâmina, que altera a intensidade real do ponto que contém RNA; e as diferentes cores das substâncias fosforescentes, que geram diferentes intensidades mesmo quando nas duas amostras há quantidade semelhante do RNA de um determinado gene. Nesse caso os dados das amostras devem ser normalizados entre si para comparação de amostras de um mesmo experimento. Uma técnica muito

utilizada é a de suavização Loess (BALLMAN, GRILL et al., 2004). Para permitir a comparação entre dados de diferentes experimentos, devem ser feitas normalizações que levem outros fatores em consideração, como os tipos de pinos utilizados na impressão do arranjo, ou os protocolos utilizados em outras etapas.

A filtragem de dados se constitui basicamente em eliminar das análises pontos de má-qualidade, como os pontos que possuem intensidade muito próxima a intensidade do fundo; ou aqueles cuja intensidade é muito diferente das suas réplicas.

Em arranjos onde há impressões do mesmo gene replicadas, ou mais de um ponto possuem seqüências que representem o mesmo gene (mesmo que seqüências diferentes) é necessária a junção dos dados relativos a todos esses pontos em um único ponto que represente o gene.

A etapa da análise de dados é quando se concentram os trabalhos de bioinformática relacionados a microarranjos (BAKEWELL e WIT, 2005; LU, LU et al., 2004; QIN e KERR, 2004; THYGESEN e ZWINDERMAN, 2004). Essa análise baseia-se principalmente em testes estatísticos para identificar grupos de genes que são diferencialmente regulados através de diferentes classes de amostras (GBG, 2006). Outras abordagens utilizadas nessas análises são a de aprendizado de máquina (CHO e WON, 2003; RUBINSTEIN, MCAULIFFE et al., 2003), *Data Warehousing* (GUÉRIN , MARQUET et al., 2007; KIRSTEN, DO et al., 2004) e *Data Mining* (BARRETT e EDGAR, 2006; PIATETSKY-SHAPIRO e TAMAYO, 2003). Pôde-se perceber após vários estudos que uma única análise não é suficiente para se tirar total proveito dos dados. São realizadas então várias análises sobre os mesmos dados, com o objetivo de se revelar diferentes aspectos sobre esses dados (LEUNG, 2002). Por isso essa etapa é mais conhecida como análise exploratória.

As análises dos dados se classificam basicamente em três grupos, relacionados aos objetivos da análise (PIATETSKY-SHAPIRO e TAMAYO, 2003):

- agrupamentos de genes em classes, utilizando o padrão de expressão gênica como critério (EISEN, SPELLMAN et al., 1998). Pode ser feita também a criação de novas classes ou mesmo o refinamento das já existentes;

- classificação da expressão gênica, baseada em outras já estudadas. Pode se descobrir, por exemplo, uma nova doença com uma assinatura digital semelhante a de outra já conhecida, dando uma idéia ao biólogo do tipo de tratamento que poderá ser aplicado (O'CONNELL, 2003);

- seleção de genes que se mostrem mais relacionados, devido a ação em conjunto dos mesmos (URIARTE e ANDRÉS, 2006). Os genes selecionados nessa etapa podem ser usados no *design* de um microarranjo mais eficiente, contendo um grupo de genes específicos para um determinado estudo (MANDUCHI, GRANT et al., 2004).

A última etapa é a elaboração das conclusões e disponibilização de resultados para a comunidade científica. Todo o experimento deve ficar disponível (dados, protocolos e anotações) de modo a permitir a comunidade científica a replicar o experimento de forma a retificar ou ratificar os resultados, ou mesmo usar parte dos dados para realizar novas pesquisas.

2.3.2 – Considerações sobre experimentos com microarranjos

O desenvolvimento dos experimentos com microarranjos deixou os biólogos com muitas expectativas devido à capacidade desta nova tecnologia revelar novos aspectos das funções gênicas. Mas também provocou algumas decepções após a publicação de alguns resultados incoerentes e a dificuldade de biólogos em realizar análises que exigem conhecimentos de áreas além das ciências biológicas, como a matemática e ciência da computação (LEUNG, LAM et al., 2001). Devemos considerar, entretanto, que essa tecnologia, relativamente nova, passa ainda por um processo de amadurecimento. Devido à inexperiência nesta técnica, erros básicos acabam ocorrendo e se propagando para todo o processo experimental, como, por exemplo, o etiquetamento errado de algum gene ou a contaminação dos mesmos devido a protocolos não aplicáveis a esta técnica (JONATHAN KNIGHT, 2001). E por envolver uma enorme quantidade de dados, os erros introduzidos tendem a se tornar tão grandes quanto o experimento.

Esses fatos mostram que, além da necessidade da colaboração de profissionais multidisciplinares, o experimento com microarranjos deve ter acompanhamento e

documentação rigorosos, e etapas bem definidas, para evitar a inserção de erros no processo e, mesmo que esses ocorram, possam ser rastreados e corrigidos.

2.4 – Evolução do suporte computacional aos experimentos com microarranjos

Nessa seção falaremos sobre os principais grupos de ferramentas de apoio computacional aos experimentos com microarranjos e como as mesmas vêm evoluindo para atender as peculiaridades deste tipo de experimento.

Os experimentos com microarranjo já eram tema de estudo em meados da década de 80 (AUGENLICHT, KOBRIN et al., 1984; AUGENLICHT, WAHRMAN et al., 1987), mas foi só a partir do final da década de 90 que o número de estudos publicados começou a crescer consideravelmente. De acordo com (JONATHAN KNIGHT, 2001), em 1999, foram publicados algumas dezenas de estudos que envolviam a tecnologia de microarranjos. Em 2000, esse número chegou próximo a 100 e atualmente são mais de 14 mil. Com o crescente interesse nessa tecnologia, diversos pesquisadores começaram a tentar replicar experimentos uns dos outros. Mas por se tratar de uma tecnologia recente, não havia práticas comuns para documentar o experimento. Ou seja, um experimento podia conter uma vasta documentação, mas não conter dados necessários para sua replicação, ou mesmo não conter dados informados em outros experimentos, caso o pesquisador tivesse a intenção de realizar um estudos comparativo (NATURE, 2002).

Visando solucionar esses problemas o grupo internacional *Microarray Gene Expression Data* (MGED, 2006a) propôs aos jornais científicos uma lista de requisitos para a publicação de estudos envolvendo microarranjos (BALL, BRAZMA et al., 2004), chamada MIAME – *Minimum Information about a Microarray Experiment* (BRAZMA, HINGAMP et al., 2001). Como o próprio nome sugere, o MIAME contém o mínimo de informações necessárias para o entendimento e replicação de um experimento, além de possibilitar o estudo comparativo entre diversos experimentos. Essas informações estão divididas em 5 grupos, que englobam as etapas de experimento até a obtenção dos dados: *Design* do experimento; design do arranjo; preparação e marcação das amostras; hibridização; leitura da imagem e obtenção de dados. As análises dos dados obtidos não

são contempladas nesses requisitos, já que o objetivo é permitir a análise independente dos dados obtidos através de um experimento bem definido e já realizado.

A proposta também sugere que os autores informem todos os dados citados pelo MIAME na publicação de seus estudos, além de depositarem tais dados em repositórios públicos de experimentos de microarranjo que atendem as resoluções do MIAME, como o GEO (EDGAR, DOMRACHEV et al., 2002), o Array Express (PARKINSON, SARKANS et al., 2005) e CIBEX (IKEO, ISHI-I et al., 2003). A lista dessas recomendações pode ser encontrada no apêndice A. Grandes veículos de publicação de trabalhos, como Nature group (NATURE, 2006), The Lancet (LANCET, 2006), Cell (CELL, 2006) e EMBO journal (EMBO, 2006), adotaram as recomendações do MIAME como requisito obrigatório para a publicação de trabalhos envolvendo a utilização de microarranjos (MGED, 2006b).

Os repositórios públicos de experimentos são serviços que, além do armazenamento dos dados experimentais, oferecem diversas ferramentas para submissão e acesso a tais dados. Além das funcionalidades básicas de submissão e pesquisa de dados, através de formulários eletrônicos, eles também provêm outros serviços mais avançados. O *Array Express*, por exemplo, disponibiliza um *data warehouse* com dados normalizados obtidos após a quantificação de RNA nas amostras dos experimentos. O GEO provê meios para se navegar entre os dados dos experimentos, além de diversos métodos de busca. O CIBEX possui uma estrutura ainda muito básica e possui poucas unidades de experimentos armazenados. As submissões a esses repositórios são revisadas por especialistas do grupo responsável pelo banco.

O MIAME e os repositórios públicos de experimentos foram grandes responsáveis pela plena disponibilidade, de forma útil, de estudos de microarranjos para a comunidade científica. Portanto, em relação ao conteúdo tornou-se possível a realização dos mais diversos tipos de estudos. Entretanto, com o constante crescimento do número de experimentos realizados, surge a necessidade de um tratamento cada vez mais automatizado dos experimentos. As recomendações do MIAME não especificam padrões para formatação ou descrição das informações submetidas, apenas o conteúdo. Observando os dados depositados nos principais repositórios públicos, percebe-se a grande quantidade de informação descrita em forma de texto livre, e sem padrões de

nomenclatura, ou seja, mais de um termo ou expressão se referindo a mesma coisa, como por exemplo: 'homem', 'humano' e 'homo sapiens'. Esses fatores dificultam o processamento automático dos experimentos, principalmente no que diz respeito á análise inter-experimentos. Deve-se levar em conta também que mesmo um único termo biológico pode ter mais de um sentido, e que o contexto deve ser levado em conta para se saber qual seu real significado (KENNEDY, KUKLA et al., 2005)

O *Genomics Unified Schema* (GUS, 2006) é um arcabouço de aplicativos, associado a um amplo esquema relacional, que objetiva armazenar, analisar e disponibilizar dados genômicos de diversos tipos, como genoma, proteoma e expressão gênica, provenientes de diversas fontes de dados. Associado ao GUS, existe um conjunto de ferramentas que permite instanciar o esquema em um banco de dados, carregar dados a partir de repositórios públicos existentes, inserir e acessar dados. Possui também um ambiente de desenvolvimento que facilita a criação de aplicativos para busca e navegação no banco instanciado. Diversos projetos de pesquisa utilizam o GUS (GUS, 2006). O sub-esquema do GUS denominado RAD - RNA *Abundance Database* (MANDUCHI, GRANT et al., 2004) permite armazenamento de todos os dados requisitados pelo MIAME, e também de resultados de análises. (MANDUCHI, GRANT et al., 2004). Além disso, no GUS também é possível armazenar ontologias.

O problema da falta de padronização de termos pode ser satisfatoriamente contornado com o uso de vocabulário controlado baseado em ontologias, que pode auxiliar no uso de uma descrição bem definida e não-ambígua do experimento (PASQUIER, GIRARDOT et al., 2004). O Gene *Ontology* (GO) (GO, 2006) define uma ontologia especializada para a área de genética e o MGED *Ontology* (OWG, 2006) é especializado em anotações para experimentos com microarranjo.

O RAD *study annotator* (MANDUCHI, GRANT et al., 2004) é uma interface gráfica que permite acesso ao banco de dados instanciado e que utiliza o GUS. A abordagem relacional do GUS, associada à aplicação de vocabulário controlado, baseado em ontologias, contribui para a estruturação e padronização de dados, que é um passo importante rumo à integração de informações provenientes de experimentos diversos.

Análogos ao MGED e ao GUS, existem respectivamente o *Generic Model Organism Database* (GMOD, 2007) e o CHADO. O GMOD é um projeto que visa desenvolver um conjunto de ferramentas para criação e administração de bancos de dados de organismos. Alguns componentes do projeto são: ferramentas para visualização e edição de genomas, ferramentas para administração de anotações em experimentos, uso de ontologias da área das ciências biológicas, um conjunto de procedimentos padrões para tratamento de dados biológicos, e um esquema relacional chamado CHADO, que possui um módulo que lida com dados de experimentos de expressão gênica com microarranjos, também chamado de RAD, e um módulo para armazenamento de ontologias. Projetos genômicos de algumas espécies estão associados ao GMOD, como é o caso do projeto da mosca de fruta e do camundongo.

A abordagem mais tradicional para a etapa de análise de dados é o uso de programas estatísticos, como o sistema R (R, 2006), o *Bioconductor* (BIOCONDUCTOR, 2006), ou mesmo rotinas mais simples desenvolvidos, por exemplo, no MatLab (MATHWORKS, 2006), na linguagem perl (PERL, 2003) ou numa forma mais sofisticada da mesma, o BioPerl (BIOPERL, 2006), que inclui pacotes de funções para aplicações biológicas. Entretanto, características como a enorme quantidade de dados a analisar, a existência de diversas variáveis a serem consideradas e a natureza comparativa (quantificação de uma substância em diferentes condições) sugerem a utilização de técnicas como *data mining* (PIATETSKY-SHAPIRO e TAMAYO, 2003), (BARRETT e EDGAR, 2006) e *data warehousing* (GUÉRIN , MARQUET et al., 2007; KIRSTEN, DO et al., 2004). A utilização de algoritmos de que envolvem aprendizado de máquina é outra abordagem comum em diferentes etapas do experimento, a partir do pré-processamento de dados (RUBINSTEIN, MCAULIFFE et al., 2003).

2.5 – Integração de dados experimentais

Um outro grupo de trabalhos se foca na integração de dados dos experimentos, até retomando alguns aspectos mencionados anteriormente. Um dos principais desafios dos sistemas gerenciais de experimentos com microarranjo é permitir a exploração, análise e interpretação dados de expressão gênica associados às anotações sobre os genes e as amostras (BRAZMA, HINGAMP et al., 2001). Dentro de um mesmo experimento, são 3 os principais grupos a serem tratados: anotações das amostras,

anotações dos genes e medidas de expressão gênica. Entretanto, considerando-se que dados oriundos de experimentos distintos, é necessário também o armazenamento de dados a respeito de execução de cada experimento (protocolos, algoritmos, etc) para possibilitar, de alguma maneira, que tais dados sejam normalizados e se torne possível uma análise comparativa entre eles.

(JARVINEN, HAUTANIEMI et al., 2004) apresenta algumas dificuldades para a integração de dados de experimentos diferentes com o objetivo de comparação entre eles. Dentre elas, foram destacadas: erros na obtenção dos oligonucleotídeos que representam os genes, gerando seqüências erradas, em arranjos customizados (impressos nos próprios laboratórios); dificuldade de encontrar correlação entre genes iguais em experimentos diferentes, devido ao uso de identificadores distintos; cDna e oligonucleotídeos hibridizam ao mRNA de maneiras diferentes (já que os oligonucleotídeos tem capacidade de identificar melhor seus alvos), o que pode gerar quantificação de mRNA diferente para a mesma amostra em arranjos de diferentes plataforma; outros fatores podem resultar em variações na quantificação, como diferenças no protocolo de hibridização ou mesmo diferenças físicas dos tipos de arranjos. Dadas essas dificuldades, conclui-se que a comparação entre experimentos é complicada e normalmente não confiável, se não forem realizados os devidos ajustes.

(MARKOWITZ, CAMPBELL et al., 2003), retomando alguns desse pontos mencionados, apresenta os principais desafios para se obter dados confiavelmente comparáveis e originados de experimentos distintos.

Primeiramente é destacada a importância do uso de padronizações de terminologias nas anotações dos experimentos. As anotações sobre genes e amostras representam grande dificuldade na integração inter-experimentos devido à falta de padrão de terminologia. O MIAME, o Gene Ontology e o SNOMED (SNOMED, 2006) são citados como esforços no sentido de facilitar essa integração.

Outro ponto abordado é o fato de a variação de alguns aspectos entre experimentos não permitir a adoção de um método eficiente para normalização de dados. Como exemplos são citados: o uso de versões diferentes de arranjos, com diferentes domínios de genes; utilização de conjuntos diferente de algoritmos, que geram dados em diferentes etapas, como escaneamento da imagem e quantificação de

mRNA. Nesse caso, a solução indicada é re-gerar os dados, a partir da re-execução dos procedimentos, biológicos e computacionais, utilizando as versões mais recentes dos arranjos ou dos algoritmos. Há outras fontes de variabilidade de dados, como diferenças entre amostras, processos para obter e armazenar amostras e ajustes de equipamentos, por exemplo. Esses dados podem ser normalizados através de métodos estatísticos, como mostrado em (HOAGLIN, MOSTELLER et al., 1983). Pode-se concluir então que a disponibilidade de dados e meta-dados do experimento é essencial que possam ser tratados e se tornarem comparáveis.

Finalmente, a abordagem de *data warehousing* é apontada como uma boa solução para sistemas como esse que precisam integrar dados originados de fontes diversas, que necessitam de validação e limpeza de dados e que dependem também de robustez e desempenho, devido a grande quantidade de dados tratados.

No próprio (MARKOWITZ, CAMPBELL et al., 2003), é apresentado o *Gene Express*, um sistema utilizado para gerenciar dados de expressão gênica gerados com a utilização da plataforma *Affymetrix GeneChip* (DAVID J.LOCKHART, HELIN DONG et al., 1996). Esse sistema integra dados de genes e amostras, e suas anotações, vindos de diferentes fontes de dados, tanto públicas quanto privadas, em um formato padrão, utilizando uma abordagem de *data warehousing*. Além disso, ele também utiliza o SNOMED como forma de padronizar e inter-relacionar os termos utilizados nas anotações. A adoção de uma plataforma principal, a padronização de formatação através da abordagem DW e o uso de vocabulário controlado, através do SNOMED foram os meios com os quais o *Gene Express* conta para facilitar a integração e comparação de dados inter-experimentos.

2.6 – Workflows científicos

Nesta seção, primeiramente vamos definir alguns conceitos relacionados a w*orkflows*. Em seguida, vamos apresentar as particularidades dos w*orkflows* científicos e mostrar alguns trabalhos nos quais essa abordagem é aplicada a experimentos biológicos.

Em empresas, há normalmente um conjunto de atividades realizadas numa determinada seqüência, repetidas vezes, para alcançar, ou contribuir com, o objetivo do negócio. Essas atividades compõem o chamado processo do negócio. O termo *workflow*

se refere à automação dessas atividades, onde documentos, informações e tarefas são passadas entre os participantes do processo, regidos por diretivas chamadas regras de negócio. *Workflows* podem ser manualmente organizados, mas na prática a maioria dessas atividades é apoiada ou automatizada por sistemas de informação (HOLLINGSWORTH, 1995). Um processo de negócio pode variar muito de complexidade, um ciclo completo pode durar de minutos a dias. Os sistemas utilizados para automatizar essas atividades podem ser implementados de diversas maneiras, utilizando diferentes tecnologias de computação e comunicação, operando em ambientes que podem estar contidos em um rede local, em uma rede entre empresas, que pode ser até mesmo a Internet, como ocorre no caso da utilização de serviços *Web* (W3C, 2006).

Em relação ao escopo computacional, podemos definir um *workflow* como uma coleção de atividades organizadas para acompanhar algum processo de negócio. As atividades ou tarefas são os componentes de software independentes que implementam alguma funcionalidade e são executadas por um ou mais sistemas de softwares. Exemplos de atividades incluem executar um programa, transformar um arquivo ou atualizar um banco de dados. Além disso, um *workflow* define a ordem de execução dessas atividades ou as condições em que essas atividades serão executadas e a sua eventual sincronização. Os dados de entrada e saída das atividades (variáveis) são definidos como o fluxo de dados do *workflow* (SANTOS, 2004).

Um sistema de gerência de *workflows* (WfMS - *Workflow Management Systems*) é aquele que provê automação do processo de negócio através do gerenciamento da seqüência de atividades e pela execução da mesma, ou seja, a invocação dos recursos humanos ou computacionais adequados associados a cada atividade do processo. Além disso, ele deve permitir a modelagem do *workflow* e ser capaz monitorar sua execução (HOLLINGSWORTH, 1995). Existem vários WfMSs disponíveis comercialmente, como o *Lotus Workflow* (IBM, 2006), o *MS Message Queuing* (MICROSOFT, 2006) e o *Oracle Workflow* (ORACLE, 2006).

Os experimentos científicos constituem-se de um conjunto de atividades experimentais repetitivas e interdependentes, executadas em uma ordem determinada e que representam uma unidade de trabalho. Assim, eles podem ser vistos como um *workflow*. Entretanto, por se tratar de uma aplicação científica, que exige alto poder

computacional, utilizada para validar descobertas na forma de algum modelo científico e aplicadas a um domínio de conhecimento específico, trata-se então de um *workflow* científico (SANTOS, 2004).

Os experimentos científicos são fortemente baseados em investigação experimental, acúmulos de evidências e assimilação de resultados. As atividades de análise não são, a princípio, determinadas. Elas vão sendo realizadas muitas vezes sob demanda, conforme vão sendo divulgados os resultados de etapas anteriores. Além disso, pode ser necessário se executar o *workflow*, ou parte dele, com o objetivo de se comparar resultados ou melhorá-los. Devido a essas características, pode-se destacar algumas particularidades dos *workflows* científicos (CAVALCANTI, 2003):

- Não são previsíveis, pois muitas vezes sua atividades serão definidas durante sua execução;
- são mutáveis, pois podem ter suas atividades alteradas conforme necessidades que podem surgir durante uma determinada execução do experimento;
- devem ser reusáveis, para permitir sua re-execução com mudança de parâmetros e dados de entrada;
- devem permitir re-execuções parciais, também no contexto da análise investigativa de dados.

Além disso, todas as execuções e re-execuções de *workflow*, mesmo as consideradas mal-sucedidas, devem ser registradas, inclusive parâmetros e dados de entrada, intermediários e resultados, pois constituem recursos científicos úteis. A chamada proveniência de dados, ou seja, os meios pelos quais os dados foram obtidos, como por exemplo, que programa foi utilizado e quais foram os dados de entrada e parâmetros, é um indicativo da validade e qualidade dos dados (BUNEMAN, KHANNA et al., 2001).

O *myGrid* (MYGRID, 2006) é um ambiente de grade computacional de alto nível para aplicações de bioinformática, que permite a integração de serviços *web* para a composição de *workflows* científicos. O *myGrid* utiliza ontologias para descrever, descobrir e compor serviços em um ambiente de experimentos científicos. O Taverna

(TAVERNA, 2006) é o componente do *myGrid* para criação e execução de *workflows* baseados no uso de serviços *web*. A definição do *workflow* é feita com o uso de uma linguagem chamada *Simplified Conceptual Workflow Language* (SCUFL), porém o sistema é capaz de gerar e exibir uma representação visual do *workflow*.

O Kepler (LUDÄSCHER, ALTINTAS et al., 2005) é um sistema que permite ao usuário a criação de modelos científicos executáveis, através da composição de outros aplicativos, utilizando uma representação visual do processo experimental, ou seja, do *workflow* científico. Ele se baseia no uso de componentes visuais para representar as atividades do *workflow* e o seqüenciamento das mesmas, no que diz respeito à parte de análise de dados do experimento. Os componentes que representam as atividades encapsulam chamadas a sistemas locais ou *web services,* acessam bancos de dados ou arquivos, remotos ou locais, ou ainda acessam repositórios públicos de dados científicos de áreas diversas como biologia e geociências.

O *Genomic Analysis Resources for Sequence Annotation* (GARSA) (DÁVILA, LORENZINI et al., 2005) é um ambiente *web* que objetiva facilitar a análise, integração e apresentação de informações genômicas através da instanciação de um *workflow*, ou *pipeline* como é chamado no meio científico das ciências biológicas. Esse *pipeline* é rodado através da execução integrada de vários sistemas de análise genômica, tratando, de forma transparente ao usuário, a conversão dos dados que transitam de um sistema para o outro. Um banco de dados relacional é utilizado para armazenamento de parâmetros e dados de entrada e saída. Os sistemas integrados no GARSA cobrem todo o experimento *in silico* de seqüências genômicas, incluindo etapas como a carga de dados de repositórios públicos, como o GenBank, e o alinhamento de seqüências utilizando-se o BLAST.

2.7 – Comparativo de trabalhos relacionados

Nessa seção faremos uma análise comparativa entre os trabalhos que consideramos mais representativos nas áreas em que se focam. O RAD-SA (MANDUCHI, GRANT et al., 2004) e o *Array Express* (MARKOWITZ, CAMPBELL et al., 2003) foram selecionados como sistemas especializados para gerenciar dados de experimentos com microarranjo. O *myGrid* (MYGRID, 2006) e o *Kepler* (KEPLER, 2006) são dois importantes sistemas de gerência de *workflows* focados em aplicações

científicas. O GARSA (DÁVILA, LORENZINI et al., 2005) representa a instância de um *workflow* para experimentos *in silico* de análise genômica, tema diretamente relacionado a esta dissertação, já que experimentos com microarranjos podem complementar as análises de dados genômicos.

Analisamos um conjunto de características que identificamos como essenciais (requisitos) a um sistema de gerência de experimentos com microarranjos. Dividimos essas características em dois grupos, um relacionado à manutenção de dados e metadados do experimento e outro que se relaciona ao acompanhamento de execução do processo experimental. A disponibilidade de metadados nesse tipo de experimento é bastante relevante pois através deles tem-se uma idéia da qualidade dos dados e também se provêm informações de como o experimento foi executado, permitindo a sua re-execução. O outro grupo de requisitos objetiva automatizar o acompanhamento e execução do experimento, reduzindo a inserção de erros no processo, facilitando a passagem da grande quantidade de dados entre as etapas e possibilitando a re-execução do experimento de forma padronizada. Os requisitos são listados a seguir, os quatro primeiros são os requisitos relacionados a suporte de dados e outros quatro são relacionados à gerência do processo:

- Atendimento às especificações do **MIAME**: verificamos se os sistemas permitem o armazenamento de todos os requisitos especificados pelo MIAME (MIAME, 2005). Esse critério é aplicável apenas aos sistemas próprios para experimentos com microarranjos;

- **Estruturação** de dados: verificamos se o sistema estrutura de alguma forma os dados e metadados dos experimentos, como, por exemplo, utilizando um esquema relacional;

- Uso de **Ontologia**: verificamos se alguma ontologia é utilizada em algum nível (dados ou metadados, por exemplo);

- Manutenção de dados de **experimentos *in vitro***: verificamos se os sistemas possuem funcionalidades para registrar ou analisar operações realizadas e dados gerados ou utilizados nas atividades *in vitro* do experimento. Informações sobre essas etapas são relevantes pois permitem a reprodução do experimento e podem ser usadas para se avaliar se dois experimentos são comparáveis.

- **Integração** de vários sistemas: verificamos se cada sistema funciona por si só, tentando abranger as várias etapas do experimento, ou se ele integra outros aplicativos, de modo a permitir a utilização de outros sistemas, mais especializados ou conceituados, de forma integrada, em etapas diversas do experimento;

- **Proveniência** de dados: verificamos se os sistemas registram informações sobre como os dados utilizados no sistema foram produzidos, como especificação de protocolos experimentais e parâmetros utilizados na execução de sistemas;

- **Execução e re-execução** de *workflow*: verificamos se os sistemas permitem, após a execução completa de um *workflow*, reusá-lo, ou seja, executá-lo novamente com alteração de parâmetros ou dados de entrada;

- **Execução parcial** de *workflow*: verificamos se é possível iniciar a execução de um *workflow* em uma etapa não inicial, utilizando dados intermediários armazenados, e a interrupção da execução deste *workflow* antes da etapa final;

Tabela 1: Comparação entre trabalhos relacionados

	RAD-SA	**Array Express**	myGrid	Kepler	GARSA
Suporte ao MIAME	Sim	Não	N/A	N/A	N/A
Estruturação de dados	GUS (RAD)	Proprietário	Proprietário	Não	baseado no GUS
Uso de Ontologia	MGED Ontology	SNOMED	Sim	Não	Não
Manutenção a dados de experimentos *in vitro*	Sim	Sim	Não	Não	Não
Integração de vários sistemas	Não	Não	Sim	Sim	Sim
Proveniência de dados	Sim	Não	Sim	Não	Não
Execução e re-execução de *workflow*	Não	Não	Sim	Sim	Sim
Execução parcial de *workflow*	Não	Não	Sim	Sim	Sim

Considerando-se as duas grandes categorias de projetos, ou seja, sistemas para experimentos com microarranjo e sistemas que utilizam a abordagem de *workflow*, podemos ver que os quesitos que caracterizam a segunda categoria são exatamente aqueles que se encontram entre as principais características dos sistemas de gerência de

workflow, ou seja, a possibilidade de executá-lo e re-executá-lo, parcial ou integralmente; e a integração de diferentes aplicativos. No *Kepler* e no *myGrid*, a re-execução parcial de *workflow* é feita de maneira indireta através do uso de *subworkflows*, ou seja, não é possível executar parcialmente um *workflow* contido em um diagrama, mas é possível encapsular partes do mesmo em outro *workflow* e executá-lo separadamente. No GARSA, a execução parcial é possível pois é o usuário que determina qual sistema e em que ordem ele será executados. Nos três sistemas, as atividades do *workflow* são executadas por outros aplicativos, cabendo ao sistema o gerenciamento da execução desses aplicativos. No grupo dos experimentos para microarranjos, essas características não estão presentes, mas outro quesito é exclusivo a ele, que é a manutenção de dados de experimentos *in vitro*, no caso do RAD-SA através do subesquema RAD do GUS. No caso do outro grupo, depende muito mais do usuário o armazenamento desses dados, através de adaptações do sistema, como por exemplo, inclusão de aplicativos que buscariam mais dados do experimento de alguma fonte remota. Também era de se esperar que o quesito de atendimento às especificações do MIAME fosse presente nos dois sistemas do primeiro grupo, porém o *RAD-Express* optou por selecionar um grupo de dados que considerou mais relevante ao experimento.

RAD-SA e o *Array Express* fazem uso de ontologia no nível de descrição de dados do experimento, mais especificamente nas anotações dos mesmos. O *myGrid* utiliza ontologia em vários níveis, inclusive na especificação do *workflow*, que pode ser usada, por exemplo, na busca por sistemas alternativos para execução de uma determinada atividade.

Em relação à estruturação de dados, podemos dizer que o *Kepler* não a possui na medida em que não armazena dados de forma embutida, mas sim provê acesso a fontes externas, como arquivos e bancos de dados, locais ou remotos. Ou seja, apesar de não possuir uma estrutura de armazenamento nativa, ele nos dá a flexibilidade de usar a estrutura que desejarmos. O GARSA utiliza em esquema relacional aberto e atualmente o grupo responsável por ele estuda a migração para o GUS. O RAD-SA, como já foi dito, utiliza o subesquema RAD do GUS e os demais utilizam um esquema proprietário de dados para armazená-los.

Dizemos que a proveniência de dados está presente no sistema quando o mesmo armazena os parâmetros de execução das atividades e/ou dados de especificação de

protocolos de algumas etapas do experimento. Usualmente é o esquema de armazenamento de dados que deve prover essa característica, como é o caso do GUS/RAD e da estrutura de armazenamento do *myGrid.*

2.8 – Considerações finais

Neste capítulo pudemos entender do que se trata o experimento com microarranjos e sua importância nas pesquisas de terapias e tratamentos genéticos. Vimos também que os desafios gerados a partir do desenvolvimento deste tipo de experimento são proporcionais às suas potencialidades.

Neste contexto, analisamos como vem evoluindo o apoio computacional a este tipo de experimento, mostrando os trabalhos que tratam os diversos elementos relativos a ele. A padronização do experimento foi tratada no aspecto do conteúdo através do MIAME. No aspecto da estruturação de dados, foram elaborados esquemas relacionais como o GUS e o CHADO. Em relação à terminologia, o uso de ontologias, como o MGED *Ontology*, busca a utilização de um vocabulário controlado. Os bancos públicos de experimentos foram responsáveis pela disponibilização de experimentos e resultados à comunidade científica, apesar de não necessariamente adotarem todos essas padronizações citadas. Em relação à análise de dados gerados, mostramos algumas abordagens utilizadas e verificamos também que mais de uma pode, e deve, ser utilizada para produzir resultados significativos.

Vimos também que a realização do experimento se compõe de diversas etapas, que podem ser realizadas utilizando-se diversos algoritmos, protocolos e aplicativos independentes. Atualmente, em grande parte dos casos, esta composição de programas é feita manualmente pelos biólogos, que executam um programa após o outro. Os dados gerados são analisados individualmente, e conforme o resultado da análise, o biólogo escolhe o próximo programa do fluxo de execução, copia e converte os dados anteriores para a nova execução (pois normalmente os formatos utilizados pelos programas não são compatíveis) e assim dá continuidade ao seu experimento. Essa abordagem, além de trabalhosa, é bastante suscetível a erros, devido à manipulação manual de dados, e tira o foco do biólogo no experimento, fazendo com que ele tenha que cuidar de detalhes de instalação e utilização de diversos aplicativos. Além disso, as informações sobre a execução do experimento não são registradas, perdendo-se assim também parte da

experiência que seria acumulada com a realização do experimento. Assim, a composição eficiente de atividades é um ponto crítico a ser tratado, não só em experimentos com microarranjos, mas em experimentos científicos em geral.

A questão da integração de dados também se torna crítica, já que a princípio, um experimento é modelado e executado de modo a se comparar os dados produzidos somente na sua execução. Mas a possibilidade de comparação de dados produzidos em experimentos distintos pode trazer resultados bem mais significativos, caso esses experimentos estejam estudando um grupo comum de genes. Assim, percebemos a real necessidade de armazenar e integrar não só os dados dos experimentos, mas também as informações sobre a execução dos mesmos, a fim de termos meios de tornar esses dados comparáveis.

Considerando as soluções já existentes e as pendências de apoio computacional à execução do experimento, vislumbramos a modelagem do experimento como a instanciação de um *workflow* científico para suprir parte significativa dessas necessidades. A necessidade de uma composição eficiente de aplicativos para tratar das atividades do experimento e de uma melhor proveniência de dados são os fatores que mais incentivaram a adoção da abordagem de um *workflow*. As funcionalidades dos sistemas que compõem o *workflow* podem vir a preencher os demais requisitos deste tipo de experimento, como a manutenção dos dados das atividades *in vitro*. O uso de padrões (MIAME e Gene *Ontology*, por exemplo), também é uma tendência nessa modelagem, já que os mesmos facilitam a colaboração entre grupos de pesquisas.

No próximo capítulo apresentaremos uma arquitetura de um *workflow* específico para experimentos com microarranjos, fazendo uso de padrões reconhecidos pela comunidade científica, como MIAME e o GUS; utilizando sistemas já consagrados neste tipo de experimento, como o sistema R, para desenvolvimento de rotinas de análise de dados; e experimentando o Kepler como sistema de execução do *workflow* deste experimento.

Capítulo 3 – Gerência do Processo Experimental com Microarranjos

No capítulo anterior, além de entendermos do que se tratam os experimentos com microarranjos e analisarmos as diversas tecnologias computacionais aplicadas sobre eles, discutimos também as principais carências de apoio computacional a este tipo de experimento. Vimos que a integração das etapas experimentais é um fator crítico para o sucesso dos experimentos com microarranjos e também que a modelagem deste experimento como a instanciação de um *workflow* científico pode tratar, além do problema da integração de etapas, diversos aspectos que são relevantes a este processo experimental.

Relembramos aqui que o objetivo desta dissertação é apoiar os experimentos com microarranjos no sentido de facilitar a composição de processos e dados, o registro de atividades e monitoração dos experimentos. Para tanto, visamos sistematizar as atividades que fazem parte do processo experimental e desenvolver um ambiente para gerência e execução de um *workflow* científico especializado nesse tipo de experimento. Assim, neste capítulo proporemos uma arquitetura que modela o desenvolvimento deste ambiente e que contempla todo o processo experimental com microarranjos, desde o planejamento do mesmo até o resultado das análises de dados a serem divulgados para a comunidade científica, além de permitir o gerenciamento sobre a execução das etapas dos processos.

Inicialmente, nas seções 3.1 e 3.2, modelaremos as principais atividades do experimento que são realizadas com a utilização de algum programa de computador, analisando os aspectos do processo experimental que demonstram a necessidade do gerenciamento integrado de suas etapas. Em seguida, na seção 3.3, vamos apresentar a arquitetura WFlex para um ambiente de integração das atividades computacionais do experimento, além de explicar os módulos que a compõem. Finalmente, na seção 3.4 faremos algumas considerações a respeito da arquitetura proposta.

3.1 – Atividades computacionais do processo experimental

Nesta seção, vamos mostrar quais as principais ferramentas computacionais utilizadas em um ambiente regular de experimentos biológicos, focando naquelas utilizadas especialmente em experimentos com microarranjos. Ou seja, faremos uma modelagem do processo experimental do ponto de vista dos sistemas computacionais utilizados.

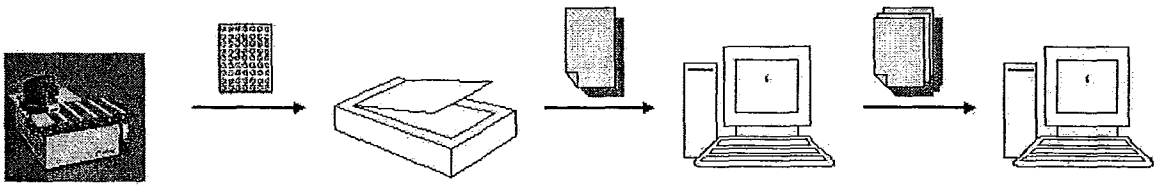

Figura 5: *Workflow* simplificado do processo experimental

A figura 5 dá uma visão geral das principais atividades computacionais realizadas nos experimentos com microarranjos, começando pela impressão do arranjo até as análises sobre os dados produzidos. Nas próximas subseções, detalharemos essas atividades, dentre outras que também fazem parte do processo, e mostrando os dados que são passados de uma para outra.

3.1.1 – Anotações experimentais

Primeiramente, os biólogos planejam o experimento, documentando o objetivo do mesmo, relacionando os genes a serem estudados e descrevendo as condições do ambiente onde as amostras a serem analisadas se desenvolveram. Em seguida, são realizadas as atividades experimentais em bancada de laboratório, como cultivo das amostras, extração do RNA total das mesmas e o preparo das soluções contendo as seqüências genéticas representativas dos genes a serem estudados. A observação da execução dessas atividades pode gerar mais anotações para o experimento. Até esse ponto, utilizam-se programas de computador de maneira acessória. Editores de texto e planilhas de dados são utilizados para fazer anotações sobre o procedimento experimental e relacionar listas de seqüências genéticas. É comum também que os biólogos realizem pesquisas através da Internet, em repositórios públicos de dados genéticos, por informações relacionadas aos objetos de estudo (genes e biomateriais especialmente). Todas essas anotações têm grande importância na interpretação dos resultados do experimento e por isso são requisitos do MIAME, adotados pelos

principais veículos de publicação de estudos científicos das áreas biomédicas. Por isso, de uma forma ou de outra, são anotadas para futuras referências.

3.1.2 – Sistemas para controle de equipamentos

As próximas atividades constituem-se da manipulação de programas que controlam o funcionamento de dois equipamentos usados especificamente nos experimentos com microarranjo. Nesta etapa, os programas de computador já não acessórios à realização do experimento, mas sim um intermediário na realização da tarefa. A interação com esses equipamentos é feita através de um sistema pois os parâmetros de entrada são relativamente complexos para serem feitos diretamente no equipamento, além de ser necessário por vezes, testar e ajustar tais parâmetros. Esses sistemas podem também gerar relatórios de execução da tarefa, que podem vir a fazer parte da documentação do experimento. A utilização desses sistemas é parte integrante do experimento propriamente dito e os dados utilizados por eles (e neles gerados) estão diretamente relacionados ao resultado do experimento.

O primeiro desses equipamentos é a impressora de microarranjos, que é um robô (*spotter*) que mergulha uma matriz configurável de agulhas nas soluções contendo os oligonucleotídeos ou cDNA representativos dos genes. O programa que controla esse equipamento permite ao usuário informar o desenho do microarranjo, ou seja, como a matriz de genes ficará organizada na lâmina. O usuário pode informar o número de agulhas que serão utilizadas e sua disposição espacial, se haverá repetições de pontos impressos e dados sobre a impressão propriamente dita, como tempo de contato das agulhas com a lâmina e duração de enxágüe das agulhas na lavagem das mesmas. Esses parâmetros influenciam na qualidade do arranjo que será obtido. O aplicativo gera um arquivo de registro informando alguns parâmetros utilizados, dados sobre a disposição das seqüências genéticas na placa de soluções e a coordenada de cada ponto onde tal seqüência foi impressa no arranjo.

O segundo equipamento é um scanner que faz uma leitura "colorida" do arranjo após o processo de hibridação. O *scanner* faz uma leitura a laser do microarranjo hibridizado e envia ao sistema que o controla duas imagens referentes à fluorescência emitidas por cada amostra, ou seja, por cada um dos canais do arranjo. As imagens têm cores diferentes para cada um desses canais, normalmente verde e vermelho. O biólogo pode alterar parâmetros da leitura, como intensidade do laser, ou da imagem já

digitalizada, alterando propriedades como o contraste, para obter imagens mais nítidas. Também é possível a combinação das duas imagens, obtendo-se uma terceira, que contém pontos nas duas cores originais e pontos na cor obtida com a combinação dessas duas cores.

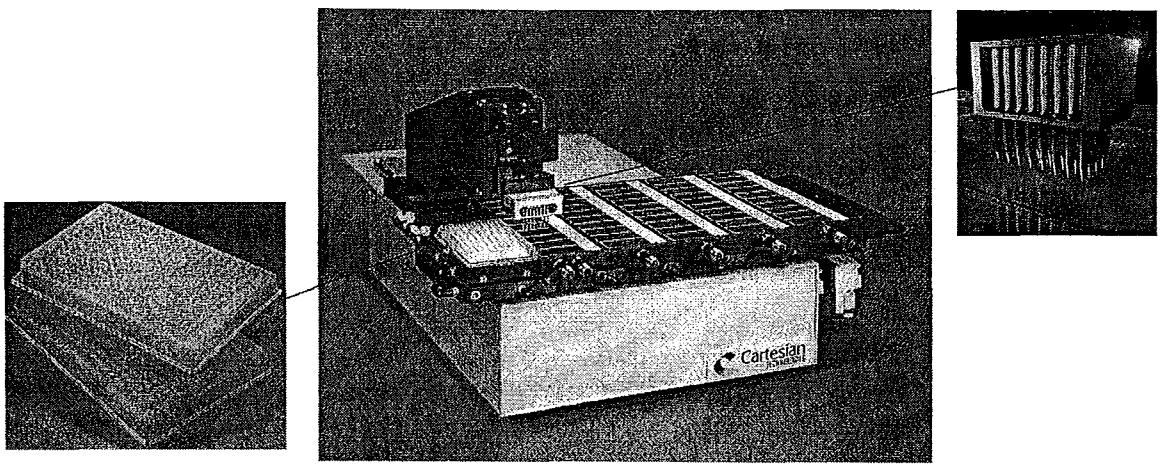

Figura 6: Impressora de Microarranjos

Na próxima fase é feita a utilização de um programa para analisar as imagens obtidas pelo *scanner* para quantificar o RNA existente nas amostras. Primeiramente o aplicativo deve identificar as áreas da figura que representam os pontos impressos e selecioná-las – todo o resto será considerado fundo. Esse passo normalmente exige interação com o usuário, que marca alguns pontos de referência para orientar o programa na localização dos demais. Após identificar a área de cada ponto, o sistema mede a intensidade dos *pixels* contidos nela. Lembrando que a intensidade do *pixel* representa uma quantidade de RNA, já que os mesmos foram marcados com substâncias fluorescentes, e quanto maior a quantidade de RNA, maior a fluorescência e mais intenso o ponto. Ou seja, é feita a quantificação de RNA das amostras analisadas. As principais medidas calculadas são a média, a mediana e o desvio padrão da intensidade dos *pixels* contidos na região de cada um dos pontos de cada canal do arranjo, além também de serem obtidas as mesmas medidas para a cor de fundo da lâmina do arranjo.

Entre a impressão do microarranjo e sua leitura a laser pelo *scanner*, há o procedimento laboratorial de hibridização das amostras com o microarranjo, que também deve ser anotado e documentado.

3.1.3 – Sistemas de análise de dados

A partir desse ponto, as tarefas passam a ser exclusivamente computacionais. Os dados obtidos na quantificação são a matéria-prima principal para toda a etapa de análise de dados. As ferramentas computacionais utilizadas são diversas, como mostrado na seção 2.4.

Neste ponto, em alguns casos onde há bioinformatas, é utilizada a abordagem de composição de chamadas de aplicativos através principalmente de rotinas em Perl (PERL, 2003) ou BioPerl (BIOPERL, 2006). Essa composição automática de aplicativos é viável para se realizar procedimentos básicos de análise, ou quando se há seqüências bem definidas de aplicações para se realizar uma determinada parte do estudo. Entretanto a análise exploratória exige composições dinâmicas de aplicações e o que acaba por ocorrer é que mais de uma rotina é criada para diferentes partes da análise ou, o mais comum, as aplicações são executadas isoladamente.

O resultado dessa etapa é a produção de diversas tabelas e gráficos, que são analisadas para que sejam elaboradas as conclusões do experimento, e possivelmente será utilizada como base em outros experimentos, com microarranjos e também de outros tipos.

3.1.4 – Publicação de resultados

Finalmente, na etapa de publicação de resultados, todas as anotações sobre os procedimentos experimentais e os dados obtidos até a etapa de quantificação devem ser compilados e formatados para envio a algum repositório público de experimentos com microarranjos. Normalmente o envio desses dados é feito através de uma página simples na Internet onde o biólogo faz a carga de um arquivo contendo os dados requisitados. Porém esse arquivo deve estar em um formato aceito pelo repositório e é de responsabilidade do pesquisador preparar tal arquivo. No sítio desses repositórios estão disponibilizados os modelos para esses arquivo, como em (EMBL, 2007b); ou aplicações, *web* ou locais, que permitem o cadastro dos dados para posterior envio diretamente ao repositório (EMBL, 2007a).

3.2 – Problemas correntes no processo experimental

Na seção anterior, vimos que diferentes ferramentas computacionais são utilizadas em diferentes níveis no processo do experimento, porém apoiando as etapas do mesmo de forma isolada. Nesta seção veremos que prejuízos têm o experimento quando não há uma forma automatizada de acompanhamento de experimentos complexos como aqueles que utilizam microarranjo.

Os experimentos com microarranjo, assim como outros experimentos biológicos, são compostos por uma série de atividades experimentais, tanto em *wet-lab* (*in vitro*) quanto em laboratórios computacionais (*in silico*). Há um processo experimental, caracterizado por um *workflow científico,* que muitas vezes não é explicitamente percebido pelo pesquisador. Essa falta de percepção, associada ao apoio computacional inadequado, ou mesmo ausente em determinados pontos, podem causar problemas que afetam o resultado do experimento, além de torná-lo mais lento e custoso. A utilização de ferramentas computacionais adequadas ao acompanhamento da execução do experimento deve torná-lo mais eficaz e menos suscetível a erros.

O problema mais usual na execução do processo experimental é a maneira como os aplicativos que executam algumas etapas do experimento são utilizados. Atualmente, em grande parte dos casos, a composição de programas que compõe as atividades *in silico* é feita manualmente pelos biólogos, que executam um programa após o outro. Os dados gerados são analisados individualmente, e conforme o resultado da análise, o biólogo escolhe o próximo programa do fluxo de execução, copia e converte os dados anteriores para a nova execução (pois normalmente os formatos utilizados pelos programas não são compatíveis) e assim dá continuidade ao seu experimento. Muitas vezes, os biólogos utilizam alguma linguagem de programação de *script* como *Perl* (PERL, 2003) para auxiliar na tarefa de composições de programas, cópias e conversão de dados, mas esta abordagem além de ser trabalhosa, possui grande dependência do ambiente (sistema operacional, localização dos programas e dados, direitos de execução sobre arquivos e diretórios), grande dificuldade de construção (criação e manutenção da definição do *workflow*, descoberta do programa correto a ser utilizado), além de grande dificuldade de utilização (definição dos parâmetros de uso, utilização simultânea por muitos usuários, grande quantidade de execuções, portabilidade para outros ambientes)

(SANTOS, 2004). Ou seja, cada programa é executado de maneira isolada e sem integração automatizada entre eles.

Associada a esta abordagem manual de composição de programas está a falta de acompanhamento do processo, que faz com que não seja possível rastrear a ocorrência de sucesso ou eventuais erros e dificulta a re-execução de partes do processo, ou mesmo de todo ele. Ou seja, a ausência de registro das execuções de programas leva a perda da experiência sobre os experimentos. Assim podemos dizer que a proveniência de dados fica severamente prejudicada. Além disso, a ocorrência de algum erro em alguma etapa (como o etiquetamento errado de algum gene ou alteração acidental de parâmetros de entrada de algum sistema) pode se propagar pelo resto do experimento, e será bastante complicado rastrear a origem desse erro de modo a corrigi-lo no ponto correto.

Outra atividade rotineira em na realização de experimentos biológicos são as anotações feitas a respeito dos procedimentos experimentais, como os protocolos utilizados nas diversas etapas e os biomateriais utilizados. Essas anotações são normalmente feitas em formato de texto livre, em arquivos isolados e sem seguir regras rígidas para estruturação do texto ou utilizar termos de forma padronizada para descrever procedimentos e biomateriais. Informações armazenadas dessa maneira são difíceis de serem associadas às outras etapas do experimento de forma manual e mesmo de forma automatizada. Comparações de resultados originados de experimentos distintos também têm grande dependência da análise dessas anotações. E se estas foram feitas de maneira muito distinta, pode se tornar inviável um estudo comparativo entre elas.

Deve-se notar que é da natureza desse tipo de experimento lidar com uma quantidade muito grande de dados (referentes normalmente a milhares de genes), o que torna ainda mais complexas as tarefas de manipulação e rastreamento de dados de forma manual. Também se torna complicado quando há necessidade de ligar esses dados às anotações feitas sobre eles, quando esses grupos de dados estão armazenados em locais (arquivos) isolados. Isso ocorre, por exemplo, quando se obtém uma lista de genes quantizados em um microarranjo e se deseja verificar anotações sobre algum gene. As informações sobre esse gene possivelmente estarão disponíveis na *Internet*, em algum arquivo da rede de computadores do laboratório ou mesmo na própria máquina que o biólogo está utilizando. Mas é dele a responsabilidade de saber onde se encontra a

informação e de recuperá-la. Para uma quantidade muito grande de genes, é um trabalho dispendioso e improdutivo.

Todas essas práticas citadas podem introduzir ruídos no experimento que podem prejudicar de maneira significativa os resultados do experimento. Também devemos considerar que o reaproveitamento de dados, a re-execução de experimentos, ou elaboração de novos, têm um custo bem menor se apoiadas por um gerenciador automático de processos experimentais. É desejável, então, a utilização de um sistema gerencial de experimentos com microarranjos que seja capaz de acompanhar a execução de tarefas do processo experimental, integrando dados oriundos de diferentes etapas de um mesmo experimento, e também de experimentos diversos. Apesar dos diversos ambientes já propostos para *workflows* científicos, e até para bioinformática, o apoio aos experimentos com microarranjos possui requisitos específicos. Um deles, por exemplo, é a necessidade de monitoração das atividades *in vitro*.

3.3 – Arquitetura WFlex

Nesta seção vamos propor a arquitetura WFlex (acrônimo para *Workflow* de Expressão Gênica) para desenvolvimento de um sistema de gerência de um *workflow* de experimentos com a utilização de microarranjos. Este sistema será responsável pela instanciação do *workflow*, ou seja, ele terá a função de realizar a execução das atividades computacionais que compõem este *workflow*. Esta arquitetura foi elaborada de modo a tratar os problemas levantados na seção anterior e atender aos requisitos relacionados na seção 2.7. Ela visa integrar, de forma automatizada, os programas utilizados em todo o processo, assim como os dados trafegados entre eles.

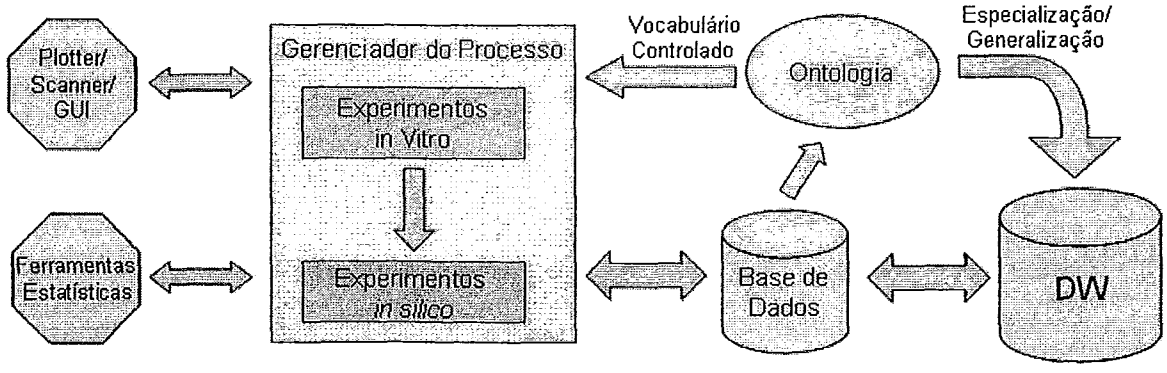

Figura 7: Arquitetura WFlex

O principal módulo da arquitetura é o gerenciador de atividades do processo experimental, responsável pelas chamadas dos aplicativos relativos as atividades *in vitro* e as *in silico* (composto usualmente por programas estatísticos), e pelo registro de execução dos mesmos. Esse gerenciador possui uma interface com os sistemas controladores da impressora e do equipamento leitor de microarranjos, além de uma interface gráfica com o usuário que permite entrada de anotações do experimento. O módulo de ontologia tem como finalidade padronizar termos e expressões utilizados nas anotações dos experimentos, além de possibilitar operações de generalização e especialização sobre os termos utilizados, baseado na hierarquia entre eles. Um banco de dados centralizado e único será utilizado para armazenar os diversos níveis de dados do experimento. O módulo de *data warehouse* vai permitir uma visão resumida dos resultados gerados pelo experimento com a utilização da arquitetura proposta. Esses módulos serão mais detalhados nas próximas subseções.

Serão utilizados, associados à arquitetura, diversos aplicativos especializados na execução das etapas do experimento, já que a intenção não é 'reinventar a roda' mas sim organizar e automatizar a execução de aplicativos que já realizam muito bem as tarefas que compõem o processo. Essa abordagem é bem ao que se refere o conceito de *workflow*.

3.3.1 – O gerenciador de atividades do processo

Este componente tem como função controlar a execução das atividades computacionais do processo. É responsável por registrar a execução dessas atividades e prover mecanismos para repetir seqüências de atividades pré-definidas ou já executadas, permitindo mudanças de parâmetros de entrada.

Figura 8: Modelo de classes para registro de execução de *workflow*

A figura 8 representa o modelo de classes sobre o qual será realizado o controle de execução das atividades do processo. Ele é baseado em um modelo de classes estendido ao do GUS, o GUS+, descrito em (SILVA, 2006). A classe 'Workflow' vai armazenar dados sobre uma instância do *workflow* ou parte dele, ou seja, um conjunto de atividades do processo realizadas, que pode ser o ciclo parcial ou completo do

experimento. 'Serviço' é qualquer aplicativo executado que represente uma atividade do processo experimental. 'Passo' indica a ordem de execução de 'Serviços' em uma instância do 'Workflow'. 'Parâmetro' representa a lista de dados utilizados como entrada de um 'Serviço'. Nela ficaram armazenados os nomes e os valores dos parâmetros utilizados.

Sempre que iniciada uma seqüência de atividades, o gerenciador deve registrar uma nova instância de 'Workflow'. Cada aplicativo executado deve provocar o registro de um 'Passo' do 'Workflow' e também o cadastro do serviço executado, assim como os parâmetros utilizados e seus valores. Para re-execução de *workflows*, o gerenciador consulta no modelo que passos devem ser executados, mas permite ao usuário a escolha de outro conjunto de dados como parâmetro de entrada.

3.3.2 – Interfaces com usuários, equipamentos e outros sistemas

Nos experimento *in vitro*, a interface com os equipamentos utilizados nos experimentos (impressora e scanner de microarranjos), é feita principalmente através de leitura de arquivos de registro de execução que esses equipamentos geram, e que normalmente informam tanto parâmetros de entrada quanto os dados produzidos como saída. O sistema que controla a impressora, por exemplo, gera um *log* que além de informar os dados do 'design' do arranjo, lista a correspondência das seqüências genéticas na placa de soluções com as coordenadas dos pontos onde essas seqüências foram impressas.

Interfaces com o usuário são através de formulários para que ele possa cadastrar dados e anotações dos experimentos. Elas existem para que o usuário cadastre somente informações que não podem ser capturadas diretamente pelo sistema através de leitura de arquivos ou consulta a dados no banco. Assim, diminui-se a probabilidade de inserção de inconsistências na base de dados.

A interação com os sistemas utilizados nas etapas de experimentação *in silico,* especialmente os estatísticos, são feitas diretamente do gerenciador de processos diretamente com tais sistemas. A partir da etapa de quantização, todos os dados necessários para análises se encontram registrados no banco, não havendo mais necessidade de leitura de arquivos. Assim, cabe ao usuário decidir que análises serão feitas e ao módulo gerenciador, executar os aplicativos que realizam essas análises,

cuidando da recuperação dos devidos dados no banco, além de registrar essa execução de aplicativos.

3.3.3 – Base de dados

Esse módulo se constitui de um banco de dados único e centralizado, estruturado para armazenar os diversos níveis de dados utilizados e gerados no experimento. Será instanciado seguindo o esquema relacional do GUS, sendo que nesta dissertação utilizaremos principalmente o sub-esquema RAD.

Os tipos de dados que o banco de dados será capaz de armazenar são os seguintes:

- dados obtidos no experimento, como a expressão os genes analisados e dados sobre os biomateriais utilizados;
- dados sobre a execução *in vitro*, como os protocolos de extração de RNA e hibridização dos genes com as amostras;
- dados sobre execução *in silico,* como os sistemas utilizados para análise dos dados obtidos, assim como parâmetros de entrada e dados de saída;
- ontologias, que incluem termos que representam diversas propriedades do experimento, de biomateriais e de protocolos utilizados, além dos domínios de valores para essas propriedades;
- e no mais alto nível, dados sobre a execução (e re-execução) do *workflow* (ou parte dele), utilizando uma extensão do esquema GUS baseado no GUS+ (SILVA, 2006).

Lembrando que os dados trafegados entre as etapas também serão armazenados no Banco, o que nos dará uma visão histórica dos mesmos dados.

3.3.4 – Módulo de Ontologia

Ontologia é uma especificação formal e declarativa de termos de um domínio e dos relacionamentos entre eles (GRUBER, 1993). Uma ontologia define um vocabulário comum para pesquisadores que precisam compartilhar informações de um domínio específico. As ontologias incluem um vocabulário padrão, o relacionamento desses

termos entre si, além da definição de cada termo do relacionamento existente entre eles. O uso de ontologia nos experimentos com microarranjo facilita a integração entre experimentos distintos uma vez que ela representa uma forma padronizada de comunicação de conhecimento a respeito de um domínio. A ontologia pode relacionar termos sinônimos e evitar o uso de termos diferentes para descrever a mesma coisa. E mesmo que termos em diferentes níveis de informação tenham sido utilizados para descrever algum aspecto em experimentos diferente, é possível adaptar uma comparação desse aspecto entre os experimento se a ontologia descrever o relacionamento entre tais termos. Por exemplo, se numa determinada análise foram utilizadas células de um anfíbio, e em outra, células de um rato, e a ontologia descrever que: rato é uma instância de mamífero; anfíbios e mamíferos são instâncias de classes de animais, pode-se visualizar uma comparação entres um anfíbio e um mamífero. Ou seja, os dados comparados foram elevados ao mesmo nível de informação.

Este módulo tem como objetivo disponibilizar uma ontologia de um domínio biológico para prover domínios de valores padronizados para diversas propriedades dos experimentos com microarranjo. Apesar de muitas anotações permitirem descrição em texto livre, várias delas possuem um conjunto determinado de valores possíveis. E esse conjunto pode estar contido na ontologia. Assim é assegurado que essas propriedades terão valores aplicáveis e expressos de forma padronizada. A ontologia também provê a descrição do significado tanto das propriedades quanto dos valores do seu domínio. Esses domínios estarão disponíveis ao usuário quando ele estiver fazendo as diversas anotações que devem ser feitas no decorrer do experimento.

Aproveitando o relacionamento entre termos e a descrição dos mesmos, essa ontologia também poderá ser utilizada em operações de agrupamento de valores de expressão gênica baseado em relacionamentos hierárquicos entre os termos, utilizando as chamadas operações de generalização e especialização de dados. Tais operações serão feitas pelo módulo de *data warehouse,* como explicaremos a seguir.

3.3.5 – Módulo de *data warehouse*

Nesse módulo, será aplicada a abordagem de data *warehousing* e OLAP sobre os dados obtidos nos experimentos. Os *data warehouses* são utilizados para se obter uma visão integrada e centralizada dos dados mais relevantes, em um ambiente onde eles estejam diretamente acessíveis e haja funcionalidades para análises dos mesmo a

um bom desempenho (KIRSTEN, DO et al., 2004). Em pesquisas com microarranjos, os dados mais relevantes são as medidas de expressão dos genes e os fatores que influenciam a variação da expressão gênica, como a idade do organismo, seu ambiente de desenvolvimento ou a presença de determinada doença ou substância.

A abordagem OLAP (On-line Analytical Processing) (JARKE, LENZERINI et al., 2003) refere-se ao conjunto de processos para criação, gerência e manipulação de dados multidimensionais para análise e visualização pelo usuário em busca de uma maior compreensão destes dados. A natureza analítica dos experimentos de expressão gênica com microarranjos sugere fortemente essa abordagem já que o objetivo é analisar a variação de uma medida, a de expressão gênica, sob diferentes aspectos. Neste tipo de experimento, a utilização de OLAP pode oferecer uma visão geral e resumida dos valores de expressão dos genes obtidos em experimentos diversos, onde as dimensões seriam esses fatores que acabamos de citar, capazes de alterar o valor dessa expressão. As navegações hierárquicas entre dimensões, típicas de sistemas OLAP, podem ser conseguidas através de consulta ao módulo de ontologia que fornece informações sobre o relacionamento entre esses fatores que serão utilizados como dimensão. Ou seja, podem-se aplicar operações de generalização e especialização sobre os dados para se obter diferentes níveis de abstração para análise dos mesmos.

A etapa de extração e limpeza de dados é facilitada pois eles se originarão de um repositório onde já foram tratados, tanto pela utilização de ontologias no cadastro de dados, quanto pela etapa de normalização das expressões quantizadas, fazendo com que eles estejam praticamente prontos para serem migrados para o repositório a parte que será criado para implantação do *data warehouse*.

3.4 – Considerações Finais

A arquitetura foi elaborada com o objetivo de cumprir a lista de requisitos identificados no capítulo anterior. Além disso, ela possui os componentes necessários para verificar se dados de diferentes experimentos são comparáveis ou para tornar tais dados comparáveis através de ajustes dos mesmos, ou a repetição de procedimentos, utilizando-se parâmetros diferentes dos originais.

O atendimento às especificações do MIAME é provido pelo esquema relacional escolhido para ser instanciado no banco de dados, o sud-esquema *RNA Abundance*

Database (RAD) do *Genomics Unified Schema* (GUS). Logicamente as interfaces com o usuário ou com outros sistemas permitem a captação de todos os dados relacionados na lista de requisitos do MIAME.

O uso de ontologia foi contemplado em duas partes da arquitetura: uso de vocabulário controlado no cadastro de dados do experimento; e para se realizar operações de especialização e generalização em etapas envolvendo análise de dados.

A estruturação de dados ocorre nos diversos níveis de dados e metadados, já que eles são armazenados de acordo com um esquema de dados relacional, portanto, estruturado. O armazenamento desses diferentes níveis de informação citados é o que faz com que a proveniência de dados também seja um aspecto contemplado nessa arquitetura.

A manutenção de dados de experimentos *in vitro* é possível uma vez que o esquema relacional utilizado contempla o armazenamento tanto de dados quanto de metadados dos processos laboratoriais.

O gerenciador de processos permite a execução e re-execução de *workflows* que representam parte do ciclo de atividades ou todo ele. Sendo assim, podemos dizer que os requisitos de Execução e re-execução de *workflow*, além da execução parcial são atendidos.

A integração automatizada de vários sistemas é, no final das contas, o grande objetivo do desenvolvimento dessa arquitetura. Esse aspecto será contemplado através principalmente da automatização na passagem de parâmetros entres os sistemas e na chamada de execução dos mesmos. Como é o módulo controlador de execução que trata essa integração de sistemas, ele que detém o controle sobre dados e parâmetros de entrada e saída dos aplicativos, facilitando assim o gerenciamento de proveniência de dados, e tirando do usuário a responsabilidade sobre o mesmo.

A arquitetura prevê bastante interação com o usuário, já que deve permitir o acompanhamento das atividades enquanto elas são realizadas, principalmente as etapas laboratoriais (carga de lista de genes, cadastro das informações sobre os protocolos experimentais), mas somente nos pontos necessário para evitar inserção de ruído no fluxo do experimento. E em relação à parte de análise de dados, deve haver interação

com o biólogo por se tratar de análise exploratória, onde o pesquisador deve poder alterar a seqüência de análises de acordo com os resultados que vão sendo obtidos durante a execução do experimento. Mas essa iteração deve ser restrita a escolha de aplicativos a serem utilizados, não deve ser para formatação de dados de entrada.

A arquitetura permite a integração de diferentes experimentos uma vez que procura padronizar tanto a estruturação quanto o conteúdo dos dados. A utilização de padrões como o MIAME e GUS, associada à possibilidade de re-execução de partes do *workflow* a ao armazenamento de metadados, facilitam o estabelecimento de comparações entre experimentos distintos, até mesmo realizados em diferentes tipos de pesquisa. Isto ocorre porque haverá informação suficiente para se verificar se os dados em questão são comparáveis, e se não forem, por terem sido obtidos, por exemplo, com a utilização de protocolos muito diferentes, haverá a possibilidade de se repetir procedimentos, tantos biológicos quanto computacionais, para transformá-los em dados comparáveis.

No próximo capítulo explicaremos como foi implementado um protótipo utilizando a arquitetura Wflex, assim como o ambiente experimental onde ele foi implantado.

Capítulo 4 – Aplicação em um ambiente real

Neste capítulo vamos apresentar um protótipo desenvolvido utilizando a arquitetura proposta no capítulo anterior. Na seção 4.1, vamos explicar como o protótipo foi implementado e as tecnologias utilizadas. Na seção 4.2, falaremos sobre o ambiente real de experimentos onde o protótipo foi implementado e como se dá a utilização desta ferramenta pelos biólogos. Finalmente, na seção 4.3, faremos algumas considerações sobre a construção e uso deste protótipo.

4.1 – Implementação do protótipo

A construção desse protótipo foi um trabalho de composição de vários aplicativos. O módulo gerenciador de processo foi construído como uma aplicação Java. O GUS/RAD foi instanciando em um Sistema Gerenciador de Banco de Dados (SGBD) PostgreSQL versão 8.2. A versão do GUS utilizada foi a 3.5, e foi obtida a partir de cópia de um banco de um laboratório da FIOCRUZ, onde ele vem sendo utilizado para estudo de anotações genômicas. O sistema utilizado para modelagem e execução de workflow foi o Kepler versão 1.0 beta3. O sistema R versão 2.4.0 foi utilizado para a construção de rotinas de análises estatísticas para a etapa *in silico* experimento.

Optamos pelo uso de interface *web* para acesso ao sistema. Assim, ele fica disponível, sem necessidade de instalação local, a partir dos diversos computadores utilizados durante a execução do experimento.

Optamos por utilizar o Kepler como a máquina de execução de workflows do sistema pois ele é um ambiente capaz de executar as análises de dados pertinentes aos experimentos com microarranjos, principalmente pela sua fácil integração com o sistema R. De fato ele foi utilizado para gerar os *sub-workflows* das possíveis análises que podem ser feitas sobre os dados obtidos a partir da quantização dos genes. Entretanto um aplicativo em Java foi desenvolvido para fazer o papel do módulo gerenciador de processo, ou seja, ele gerencia a execução desse workflow elaborado no Kepler (faz a chamada do workflow e registra sua execução). Esse módulo também é responsável, caso requisitado pelo usuário, por consultar na base de dados uma

determinada seqüência de atividades e executá-las na mesma ordem, fazendo as chamadas aos *sub-workflows* corretos, na ordem correta.

Diversos dados foram carregados a partir da MGED Ontology para o banco de dados. A utilização deles, como já foi dito, é para prover vocabulário controlado nas telas de cadastro, e para que seja possível realizar operações de generalização e especialização nas dimensões no módulo de *data warehouse*.

4.1.1 – Interação com o usuário

A tela principal do sistema é uma interface *web* e foi dividida em duas áreas principais. Do lado esquerdo, está a lista de atividades que fazem parte do experimento. Uma seta indica qual atividade está sendo realizada ou revisada no momento. Na área maior, do lado direito, estão os dados do atividade em questão. Nessa área, aparece ou uma tela de cadastro para os dados dessa atividade, ou somente os dados já armazenados relativos a esta atividade. Assim, esta segunda tela surge sempre após a conclusão de um cadastro ou quando o usuário clica em uma atividade já realizada.

Nas telas de cadastro deste protótipo, além dos campos livres de texto, há propriedades que possuem domínio de valores especificados na ontologia (MGED *Ontology*), sendo possível consultar a lista desses valores nas caixas de seleção, e escolher um deles. Além disso, as definições tanto das propriedades quanto dos valores do seu domínio ficam disponíveis para consulta, através de um botão de ajuda. E se o caso for aplicável, a tela pode exibir um campo para fazer carga de arquivos que contenham valores a serem cadastrados, como é o caso dos sistemas de controle do *spotter* e do *scanner* que geram arquivos de registros com dados de entrada e saída.

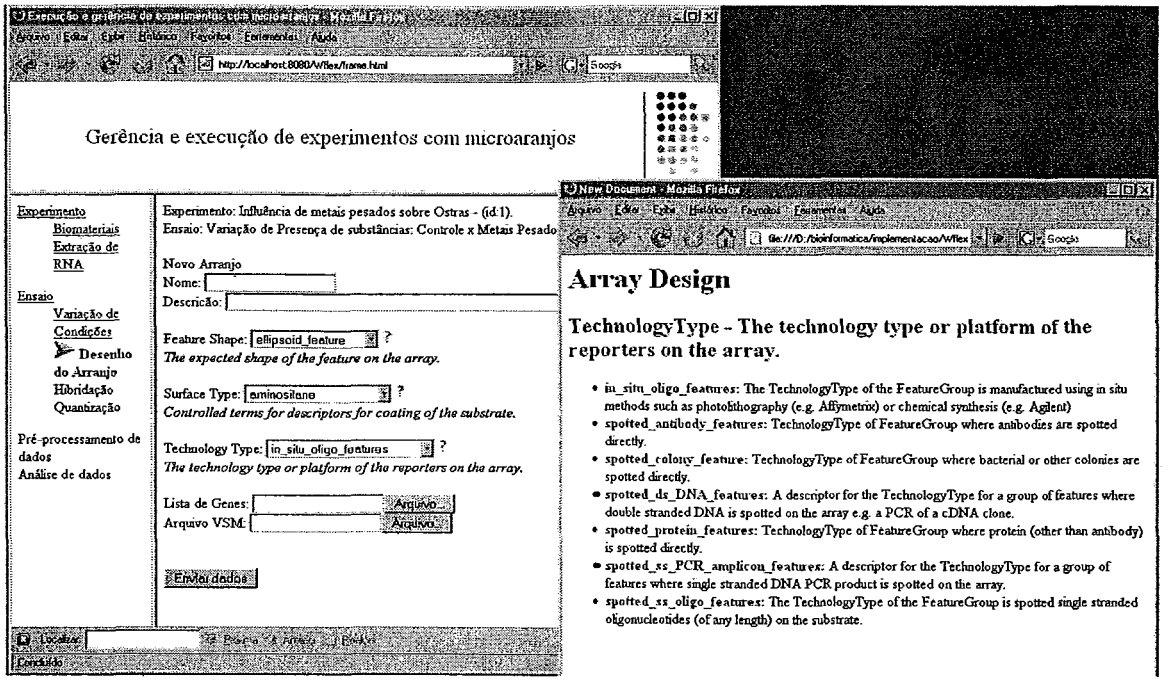

Figura 9: Cadastro de dados com apoio de Ontologia

Na Figura 9, podemos ver três propriedades do arranjo que tem lista de possíveis valores na MGED *Ontology*, além de uma tela com a lista de definições para os valores de uma das propriedades, e também a definição da própria propriedade. Também há campos para se fazer a cópia de dois arquivos locais para o servidor do sistema (*upload*). O primeiro desses arquivos é a lista de genes a serem impressos na lâmina; e no segundo estão contidas informações do desenho do arranjo na lâmina, gerado pelo sistema que controla a impressora de arranjos. Também podemos observar na área lateral da esquerda a lista das atividades do experimento. Uma seta indica que '*design* de Arranjos' é a tarefa atualmente sendo realizada. Os resultados produzidos nas tarefas anteriores à atual podem ser visualizados através dos *links* sobre os nomes das etapas. As etapas posteriores só podem ser executadas na ordem das atividades.

A Figura 10 mostra a tela exibida após a etapa da quantização de dados. Podemos notar que ela também exibe vários dados do design do arranjo, mas que são,de certa forma, relacionados à quantização. O nome dos genes possui um *link* para as anotações desse gene armazenadas no repositório público de seqüências genéticas GenBank (http://www.ncbi.nlm.nih.gov/).

Grid	SubGrid	Poco	Gene	Canal	Media	Fundo	Desvio Padrao	D.P.Fundo	Placa	Sequencia
1,1	1,1	P19	U15174	1	3186.60150375 94	2196.44864864865	631.529379825385	456.496513088089	OE0000437-AOP	AGCATGAGGAACACGAGCGTCATGAAGAAA
1,1	1,1	P19	U15174	2	3305.91729323308	2695.56216216216	674.831954429163	527.068126989801	OE0000437-AOP	AGCATGAGGAACACGAGCGTCATGAAGAAA
1,1	1,2	P09	AF022860	1	2666.92508143322	2117.11326860841	555.230404029182	413.660437929213	OE0000437-AOP	ATGGAACCCATCTCGGCTTTTGCAGTGGACA
1,1	1,2	P09	AF022860	2	3096.18241042345	2589.01618122977	670.656521053 87	507.296529869698	OE0000437-AOP	ATGGAACCCATCTCGGCTTTTGCAGTGGACA
1,1	1,3	L23	AF043339	2	3140.80110497238	2618.90421455939	687.227007779085	548.311185527413	OE0000437-AOP	GCTTTTGTTCAGGGCTGTGATCGGCCTGGGGA
1,1	1,3	L23	AF043339	1	2692.64640883978	2121.69731800766	571.567082532098	530.265568763291	OE0000437-AOP	GCTTTTGTTCAGGGCTGTGATCGGCCTGGGGA
1,1	1,4	L13	NM_002608	2	3797.06077348066	2732.44029850746	761.499771399585	604.82758878272	OE0000437-AOP	CCTTCCAAAACCTGCTTCCTTCAGTTTGTAAA
1,1	1,4	L13	NM_002608	1	3559.60220994475	2378.44402985075	820.426600952605	664.841149004 72	OE0000437-AOP	CCTTCCAAAACCTGCTTCCTTCAGTTTGTAAA
1,1	1,5	L03	M37763	2	2540.74468085106	2722.62941176471	553.316630882131	549.524829080056	OE0000437-AOP	TAAACTTCAGCAACCCTACAGTATATAAGCT
1,1	1,5	L03	M37763	1	1742.43971631206	1931.32352941176	445.521372670233	480.306147488183	OE0000437-AOP	TAAACTTCAGCAACCCTACAGTATATAAGCT
1,1	1,6	H17	X79981	2	2917	2520.78861788618	571.605765816923	566.190914509112	OE0000437-AOP	TTAGAGGAACCCAAGATGTGGCCTTTAGCAA
1,1	1,6	H17	X79981	1	2043.73529411765	1701.9593495935	421.909390590 08	419.892128972 84	OE0000437-AOP	TTAGAGGAACCCAAGATGTGGCCTTTAGCAA
1,1	1,7	H07	NM_001719	1	2633.14583333333	2107.30409356725	631.424966974256	410.747025245284	OE0000437-AOP	TCCCCTCCCTATCCCCAACTTTAAAGGTGTG
1,1	1,7	H07	NM_001719	2	3067.41666666667	2585.94736842105	528.611824492139	565.226610080447	OE0000437-AOP	TCCCCTCCCTATCCCCAACTTTAAAGGTGTG
1,1	1,8	D21	X03438	1	11346.6957831325	2817.22745098039	2691.5443316144	1166.59359899998	OE0000437-AOP	CCACTCACCAGTGTCCCCTCCACTGTCACAT
1,1	1,8	D21	X03438	2	8607.95180722892	3059.65490196078	1958.66680506399	917.36459583871	OE0000437-AOP	CCACTCACCAGTGTCCCCTCCACTGTCACAT
1,1	1,9	D11	NM_000639	2	2361.9620596206	2588.90189873418	538.2622135289 99	497.230565019925	OE0000437-AOP	AATATGGGTTGCATTTGGTCAAGATTTTGAAT
1,1	1,9	D11	NM_000639	1	1341.63956639566	1999.32911392405	314.479655898	437.812374730488	OE0000437-AOP	AATATGGGTTGCATTTGGTCAAGATTTTGAAT

Figura 10: Lista de genes e quantificação dos RNAs correspondentes

4.1.2 – O gerenciador de processo

O gerenciador de processo foi implementado de modo a registrar qualquer execução de sistemas durante o processo experimental. Ao se visualizar os dados de uma etapa, podemos também visualizar os dados de sua execução: quando a mesma ocorreu, em que ordem foi executada e se foi executada mais de uma vez. Isso possibilita o rastreamento das atividades do experimento e a identificação de práticas boas e ruins.

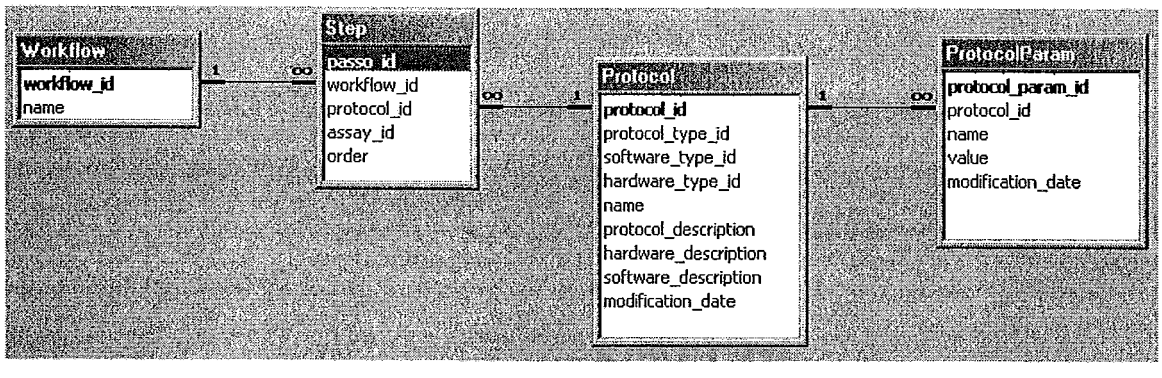

Figura 11: Esquema estendido para registro de execução de atividades

A figura 11 mostra como o modelo de classes da figura 8 foi mapeado para o esquema relacional utilizado. Foi necessária a criação de apenas duas tabelas novas no esquema do GUS, para representar uma execução de 'Workflow' e os passos ('Steps') dessa execução. Para registrar os sistemas utilizados (e seus parâmetros de entrada) nesses passos aproveitamos duas tabelas do sub-esquema RAD ('Protocol' e

'ProtocolParam'), já que a primeira possui atributos que permitem indicar que o protocolo na verdade foi realizado por um sistema, com alguns dados de entrada.

Para fazer a composição e orquestração dos sistemas que realizam as tarefas do *workflow*, utilizamos a interface gráfica do Kepler (KEPLER, 2006) para montar vários *sub-workflows* do processo experimental. O Kepler armazena a definição de *workflows* em um arquivo XML na linguagem MoML (LEE e NEUENDORFFER, 2000). Quando o usuário requisita a execução das tarefas representadas por este *workflow*, o gerenciador de processo faz a chamada para execução do mesmo sem utilizar a interface gráfica (o que não poderia ser feito de maneira automática), mas utiliza a máquina de execução de *Workflow* do *Ptolomy* (EKER, JANNECK et al., 2003) para executar o *workflow* definido em MoML. Essa máquina de execução é uma aplicação Java sem interface gráfica e é a mesma utilizada pela interface gráfica do Kepler para executar workflows. Antes da execução do workflow definido pelo XML, o gerenciador de processo abre esse arquivo para fornecer os parâmetros de entrada. O próprio *sub-workflow* é responsável por recuperar dados de entradas do banco e lá armazená-los. Mas é o próprio gerenciador do processo que registra a execução do *workflow*.

Mesmo utilizando nos *sub-workflows* chamadas a poucos aplicativos ou rotinas (as vezes a somente um), o uso do Kepler se justifica pelas facilidades que ele oferece para recuperação e manipulação de dados, para passagem de parâmetros, para chamadas a execução de aplicações e também pela representação visual do workflow. Isto facilitou bastante também os testes de execução de aplicativos em seqüência.

A figura 12 é representação gráfica de um *workflow* no Kepler que envolve atividades desde o desenho do microarranjo na lâmina até a etapa de quantização. Para a execução deste *workflow*, os dados básicos do *design* do arranjo já devem estar cadastrados. Esse *workflow* faz a leitura de três arquivos, um com a lista de genes a serem impressos e a posição dos mesmos nas placas com soluções; no segundo estão contidas informações do design do arranjo na lâmina, gerado pelo sistema que controla a impressora de arranjos. O terceiro arquivo contém dados da quantização dos dois canais da lâmina.

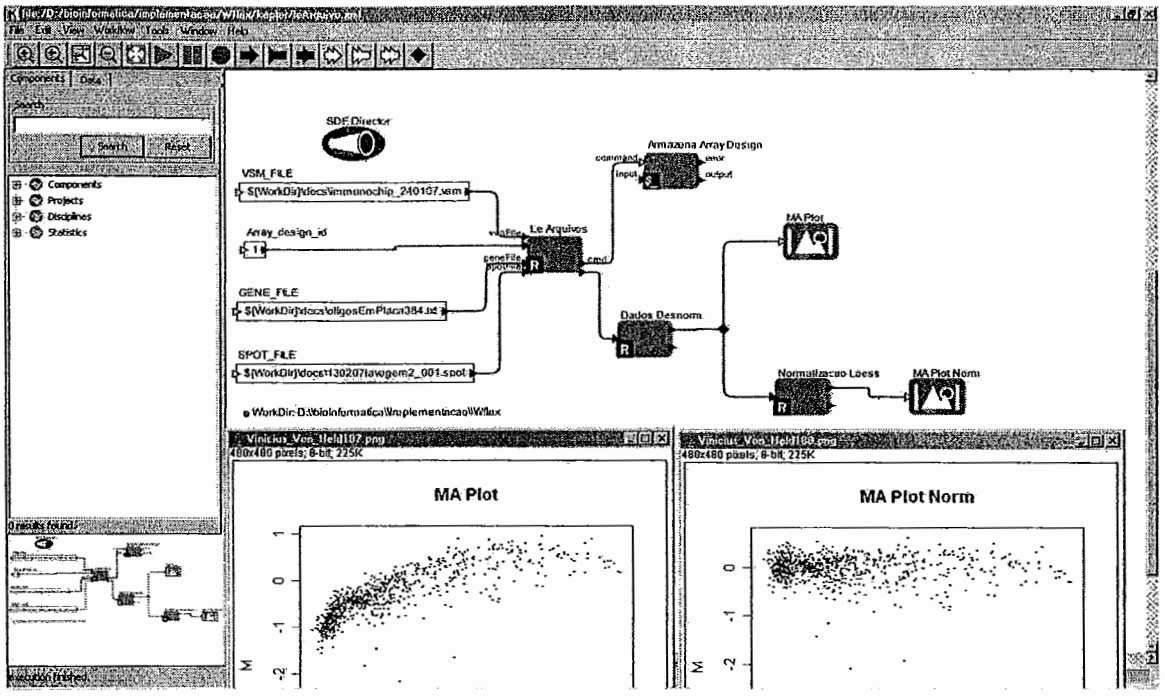

Figura 12: Representação visual de *workflow* no Kepler

4.1.3 – O esquema relacional

Utilizamos o sub-esquema do GUS chamado RNA *Abundance Database* (RAD) e também algumas tabelas de outros sub-esquemas que são usadas de forma compartilhada. Para dar uma idéia do mapeamento dos dados do experimento no esquema, vamos utilizar como exemplo um pedaço do esquema instanciado, mostrado a seguir.

A figura 13 mostra um pedaço do sub-esquema RAD. Essa parte, especificamente, contempla as principais tabelas responsáveis pelo armazenamento de dados e procedimentos de etapas bem características do experimento, que vão da preparação da lâmina até a quantização da imagem da lâmina hibridizada.

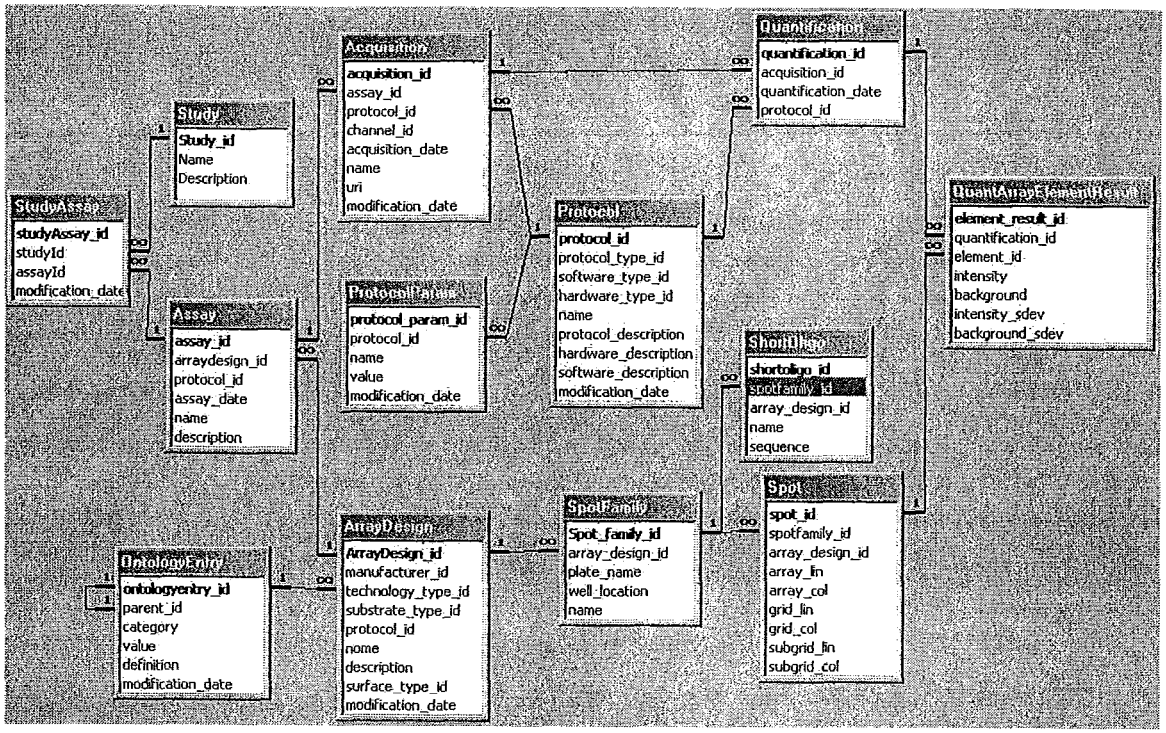

Figura 13: Parte do sub-esquema RAD

Segundo a MGED *Ontology* (WHETZEL, PARKINSON et al., 2006), um experimento é caracterizado por uma série de ensaios experimentais realizados com um propósito comum. No caso dos microarranjos, um ensaio é representado por uma hibridização, ou seja, um experimento com microarranjo é caracterizado por uma série de hibridizações com o propósito de determinar a influência de uma determinada condição sobre a expressão gênica da célula. No esquema da Figura 13, a tabela 'Study' contempla dados de experimentos enquanto a tabela 'Assay', os dados de um ensaio experimental, incluindo o protocolo de hibridização da lâmina, já que é este o evento que caracteriza o ensaio. A tabela 'StudyAssay' relaciona os ensaios experimentais de um experimento.

Acompanhando a ordem cronológica dos eventos, temos a tabela 'ArrayDesign' com os dados do desenho do arranjo na lâmina. Um conjunto de três tabelas contempla os dados dos pontos impressos na lâmina e das seqüências genéticas contidas na placas de soluções: um registro de 'SpotFamily' representa o conjunto de pontos impressos a partir do mesmo poço de uma das placas de soluções, ou seja, é o grupo de replicatas de um ponto impresso. Por esse motivo, ele contempla os dados de um poço, assim como sua localização na placa de soluções de genes. Assim vários 'Spots' podem fazer parte de uma mesma 'SpotFamily', porém, cada um deles tem localização distinta na lâmina.

Um 'ShortOligo' representa a seqüência genética contida em cada poço de cada placa, há uma para cada 'SpotFamily'.

Um outro grupo de tabelas contempla os dados da digitalização da imagem do arranjo hibridizado com as amostras e os dados da quantização desta imagem. Uma aquisição de imagem (tabela 'Acquisition') só pode pertencer a um ensaio, entretanto um ensaio gera mais de uma aquisição, já que uma imagem é gerada para cada canal (RNAS referentes a cada amostra) da lâmina. Cada aquisição então é quantificada ('Quantification') e o valor associado a cada 'Spot' da lâmina é registrado em 'QuantArrayElementResult'.

Podemos notar também a presença das tabelas 'Protocol' e 'ProtocolParam', que nessa parte do esquema, registra os protocolos de preparo do ensaio, desenho do arranjo, hibridização, aquisição de imagem e quantificação. De maneira geral, essas duas tabelas servem para armazenar de maneira estruturada metadados dos procedimentos que não possuem locais específicos para armazenamento no restante do esquema. Essas tabelas possuem campos para armazenar tanto o nome da propriedade quanto seu valor, assim como o tipo do dado.

Na figura 13, podemos ver também como as ontologias são armazenadas neste esquema de dados. Chaves estrangeiras de várias tabelas se referem a registros de 'OntologyEntry' (como é o caso de 'technology_type_id' em 'ArrayDesign' e 'protocol_type_id' em 'Protocol'). Os valores possíveis para esses campos são registros encontrados na MGED *Ontology*. Como os termos dessa ontologia têm relacionamentos onde um é subclasse ou instância do outro, o esquema de armazenamento no banco se dá simplesmente através de auto-relacionamento dos termos da ontologia, onde um termo é pai do outro.

4.1.4 – Uso de Ontologias

O MGED *Ontology* descreve uma ontologia especializada para termos utilizados em anotações de experimentos com microarranjos. A organização da ontologia é baseada em uma hierarquia de classes, e no mais baixo nível dessa hierarquia, algumas classes possuem instâncias de valores. É esse nível de instâncias que fornece domínios de valores para algumas propriedades do experimento. Todos os níveis da hierarquia possuem definição para os termos.

Grid	SubGrid	Poco	Gene	Canal	Media	Fundo	Desvio Padrao	D.P.Fundo	Placa	Sequencia
1,1	1,1	P19	U15174	1	3186.6015037594	2196.44864864865	631.529379825385	456.496513088089	OE0000437-AOP	AGCATGAGGAACACGAGCGTCATGAAGAAA
1,1	1,1	P19	U15174	2	3305.91729323308	2695.56216216216	674.831954429163	527.068126989801	OE0000437-AOP	AGCATGAGGAACACGAGCGTCATGAAGAAA
1,1	1,2	P09	AF022860	1	2666.92508143322	2117.11326860841	555.230404029182	413.660437929213	OE0000437-AOP	ATGGAACCCATCTCGGCTTTTGCAGTGGACA
1,1	1,2	P09	AF022860	2	3096.18241042345	2589.01618122977	670.65652105387	507.296529869698	OE0000437-AOP	ATGGAACCCATCTCGGCTTTTGCAGTGGACA
1,1	1,3	L23	AF043359	2	3140.80110497238	2618.90421455939	687.227007779085	548.311185527413	OE0000437-AOP	GCTTTTGTTCAGGGCTGTGATCGGCCTGGGG
1,1	1,3	L23	AF043359	1	2692.64640883978	2121.69731800766	571.567082532098	530.265568763291	OE0000437-AOP	GCTTTTGTTCAGGGCTGTGATCGGCCTGGGG
1,1	1,4	L13	NM_002608	2	3797.06077348066	2732.44029850746	761.499771399585	604.82758878272	OE0000437-AOP	CCTTCCAAAACCTGCTTCCTTCAGTTTGTAAA
1,1	1,4	L13	NM_002608	1	3559.60220994475	2378.44402985075	820.426600952605	664.84114900472	OE0000437-AOP	CCTTCCAAAACCTGCTTCCTTCAGTTTGTAAA
1,1	1,5	L03	M37763	2	2540.74468085106	2722.62941176471	553.316630882131	549.524829080056	OE0000437-AOP	TAAACTTCAGCAACCCTACAGTATATAAGCT
1,1	1,5	L03	M37763	1	1742.43971631206	1931.32352941176	445.521372670233	480.306147488183	OE0000437-AOP	TAAACTTCAGCAACCCTACAGTATATAAGCT
1,1	1,6	H17	X79981	2	2917	2520.78861788618	571.605765816923	566.190914509112	OE0000437-AOP	TTAGAGGAACCCAAGATGTGGCCTTTAGCAA
1,1	1,6	H17	X79981	1	2043.73529411765	1701.9593495935	421.90939059008	419.89212897284	OE0000437-AOP	TTAGAGGAACCCAAGATGTGGCCTTTAGCAA
1,1	1,7	H07	NM_001719	1	2633.14583333333	2107.30409356725	631.424966974256	410.747025245284	OE0000437-AOP	TCCCCTCCCTATCCCCAACTTTAAAGGTGTG
1,1	1,7	H07	NM_001719	2	3067.41666666667	2585.94736842105	528.611824492139	565.226610080447	OE0000437-AOP	TCCCCTCCCTATCCCCAACTTTAAAGGTGTG
1,1	1,8	D21	X03438	1	11346.6957831325	2817.22745098039	2691.5443316144	1166.59359899998	OE0000437-AOP	CCACTCACCAGTGTCCCCTCCACTGTCACAT
1,1	1,8	D21	X03438	2	8607.95180722892	3059.65490196078	1958.66680506399	917.36459583871	OE0000437-AOP	CCACTCACCAGTGTCCCCTCCACTGTCACAT
1,1	1,9	D11	NM_000639	2	2361.9620596206	2588.90189873418	538.262213528999	497.230565019925	OE0000437-AOP	AATATGGGTTGCATTTGGTCAAGATTTTGAAT
1,1	1,9	D11	NM_000639	1	1341.63956639566	1999.32911392405	314.479655898	437.812374730488	OE0000437-AOP	AATATGGGTTGCATTTGGTCAAGATTTTGAAT

Figura 10: Lista de genes e quantificação dos RNAs correspondentes

4.1.2 – O gerenciador de processo

O gerenciador de processo foi implementado de modo a registrar qualquer execução de sistemas durante o processo experimental. Ao se visualizar os dados de uma etapa, podemos também visualizar os dados de sua execução: quando a mesma ocorreu, em que ordem foi executada e se foi executada mais de uma vez. Isso possibilita o rastreamento das atividades do experimento e a identificação de práticas boas e ruins.

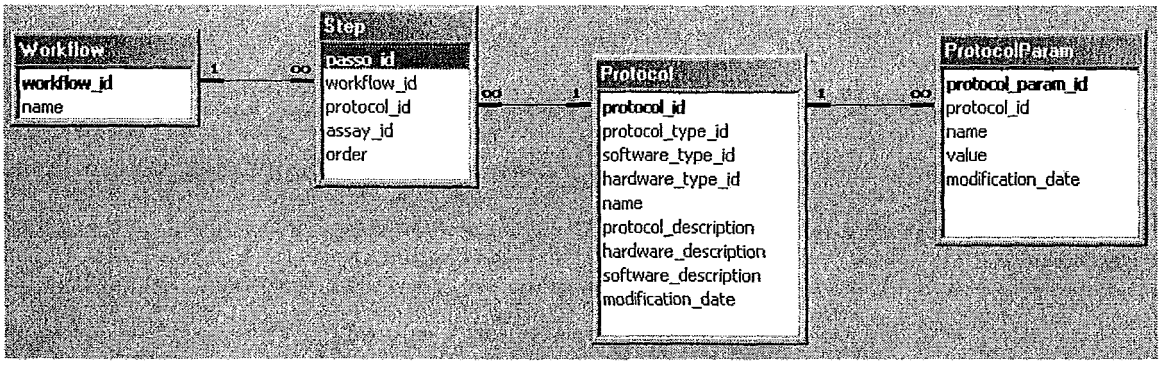

Figura 11: Esquema estendido para registro de execução de atividades

A figura 11 mostra como o modelo de classes da figura 8 foi mapeado para o esquema relacional utilizado. Foi necessária a criação de apenas duas tabelas novas no esquema do GUS, para representar uma execução de 'Workflow' e os passos ('Steps') dessa execução. Para registrar os sistemas utilizados (e seus parâmetros de entrada) nesses passos aproveitamos duas tabelas do sub-esquema RAD ('Protocol' e

'ProtocolParam'), já que a primeira possui atributos que permitem indicar que o protocolo na verdade foi realizado por um sistema, com alguns dados de entrada.

Para fazer a composição e orquestração dos sistemas que realizam as tarefas do *workflow*, utilizamos a interface gráfica do Kepler (KEPLER, 2006) para montar vários *sub-workflows* do processo experimental. O Kepler armazena a definição de *workflows* em um arquivo XML na linguagem MoML (LEE e NEUENDORFFER, 2000). Quando o usuário requisita a execução das tarefas representadas por este *workflow*, o gerenciador de processo faz a chamada para execução do mesmo sem utilizar a interface gráfica (o que não poderia ser feito de maneira automática), mas utiliza a máquina de execução de *Workflow* do *Ptolomy* (EKER, JANNECK et al., 2003) para executar o *workflow* definido em MoML. Essa máquina de execução é uma aplicação Java sem interface gráfica e é a mesma utilizada pela interface gráfica do Kepler para executar workflows. Antes da execução do workflow definido pelo XML, o gerenciador de processo abre esse arquivo para fornecer os parâmetros de entrada. O próprio *sub-workflow* é responsável por recuperar dados de entradas do banco e lá armazená-los. Mas é o próprio gerenciador do processo que registra a execução do *workflow*.

Mesmo utilizando nos *sub-workflows* chamadas a poucos aplicativos ou rotinas (as vezes a somente um), o uso do Kepler se justifica pelas facilidades que ele oferece para recuperação e manipulação de dados, para passagem de parâmetros, para chamadas a execução de aplicações e também pela representação visual do workflow. Isto facilitou bastante também os testes de execução de aplicativos em seqüência.

A figura 12 é representação gráfica de um *workflow* no Kepler que envolve atividades desde o desenho do microarranjo na lâmina até a etapa de quantização. Para a execução deste *workflow*, os dados básicos do *design* do arranjo já devem estar cadastrados. Esse *workflow* faz a leitura de três arquivos, um com a lista de genes a serem impressos e a posição dos mesmos nas placas com soluções; no segundo estão contidas informações do design do arranjo na lâmina, gerado pelo sistema que controla a impressora de arranjos. O terceiro arquivo contém dados da quantização dos dois canais da lâmina.

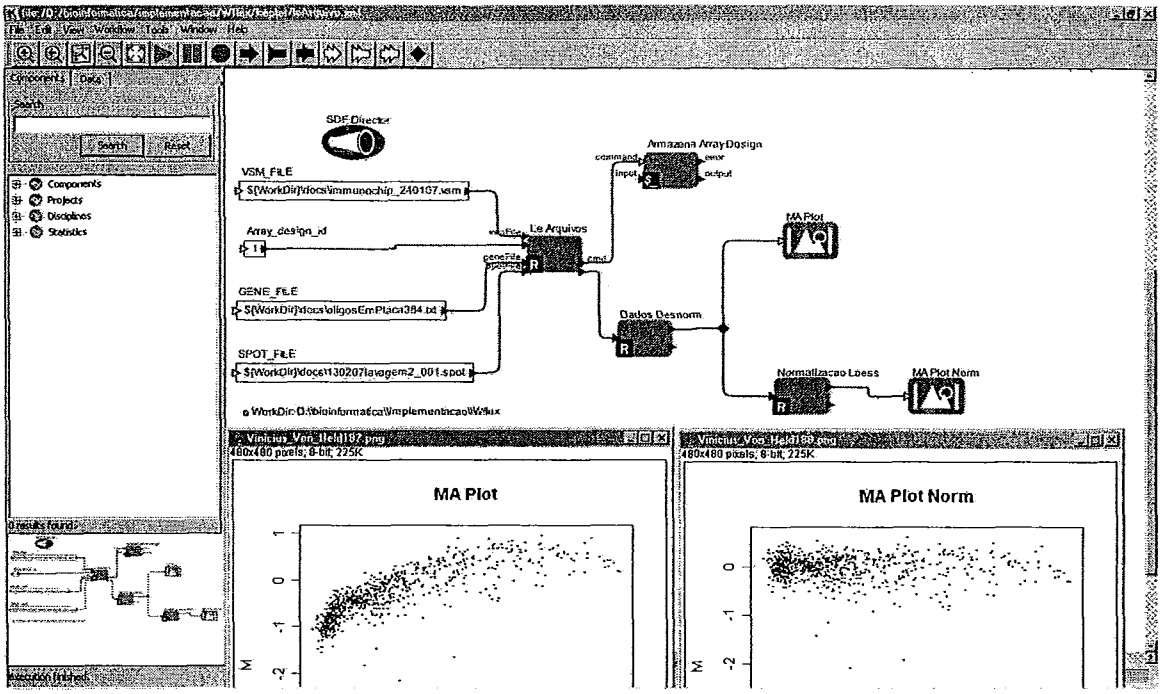

Figura 12: Representação visual de *workflow* no Kepler

4.1.3 – O esquema relacional

Utilizamos o sub-esquema do GUS chamado RNA *Abundance Database* (RAD) e também algumas tabelas de outros sub-esquemas que são usadas de forma compartilhada. Para dar uma idéia do mapeamento dos dados do experimento no esquema, vamos utilizar como exemplo um pedaço do esquema instanciado, mostrado a seguir.

A figura 13 mostra um pedaço do sub-esquema RAD. Essa parte, especificamente, contempla as principais tabelas responsáveis pelo armazenamento de dados e procedimentos de etapas bem características do experimento, que vão da preparação da lâmina até a quantização da imagem da lâmina hibridizada.

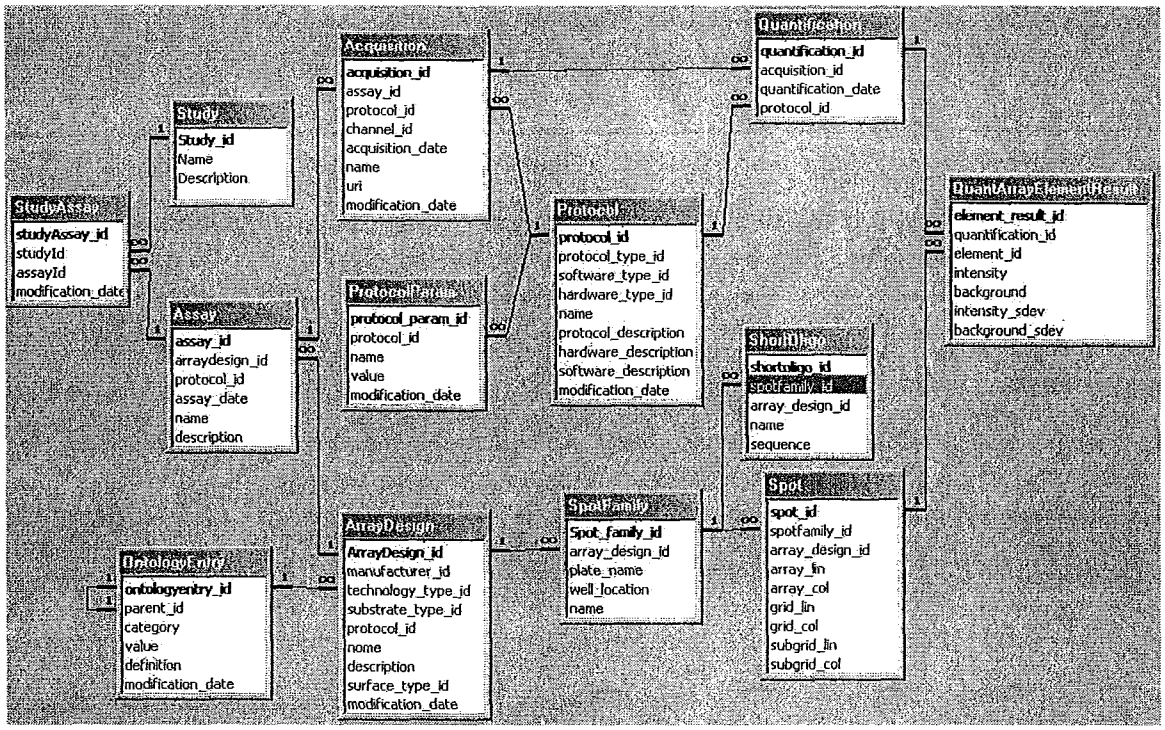

Figura 13: Parte do sub-esquema RAD

Segundo a MGED *Ontology* (WHETZEL, PARKINSON et al., 2006), um experimento é caracterizado por uma série de ensaios experimentais realizados com um propósito comum. No caso dos microarranjos, um ensaio é representado por uma hibridização, ou seja, um experimento com microarranjo é caracterizado por uma série de hibridizações com o propósito de determinar a influência de uma determinada condição sobre a expressão gênica da célula. No esquema da Figura 13, a tabela 'Study' contempla dados de experimentos enquanto a tabela 'Assay', os dados de um ensaio experimental, incluindo o protocolo de hibridização da lâmina, já que é este o evento que caracteriza o ensaio. A tabela 'StudyAssay' relaciona os ensaios experimentais de um experimento.

Acompanhando a ordem cronológica dos eventos, temos a tabela 'ArrayDesign' com os dados do desenho do arranjo na lâmina. Um conjunto de três tabelas contempla os dados dos pontos impressos na lâmina e das seqüências genéticas contidas na placas de soluções: um registro de 'SpotFamily' representa o conjunto de pontos impressos a partir do mesmo poço de uma das placas de soluções, ou seja, é o grupo de replicatas de um ponto impresso. Por esse motivo, ele contempla os dados de um poço, assim como sua localização na placa de soluções de genes. Assim vários 'Spots' podem fazer parte de uma mesma 'SpotFamily', porém, cada um deles tem localização distinta na lâmina.

Um 'ShortOligo' representa a seqüência genética contida em cada poço de cada placa, há uma para cada 'SpotFamily'.

Um outro grupo de tabelas contempla os dados da digitalização da imagem do arranjo hibridizado com as amostras e os dados da quantização desta imagem. Uma aquisição de imagem (tabela 'Acquisition') só pode pertencer a um ensaio, entretanto um ensaio gera mais de uma aquisição, já que uma imagem é gerada para cada canal (RNAS referentes a cada amostra) da lâmina. Cada aquisição então é quantificada ('Quantification') e o valor associado a cada 'Spot' da lâmina é registrado em 'QuantArrayElementResult'.

Podemos notar também a presença das tabelas 'Protocol' e 'ProtocolParam', que nessa parte do esquema, registra os protocolos de preparo do ensaio, desenho do arranjo, hibridização, aquisição de imagem e quantificação. De maneira geral, essas duas tabelas servem para armazenar de maneira estruturada metadados dos procedimentos que não possuem locais específicos para armazenamento no restante do esquema. Essas tabelas possuem campos para armazenar tanto o nome da propriedade quanto seu valor, assim como o tipo do dado.

Na figura 13, podemos ver também como as ontologias são armazenadas neste esquema de dados. Chaves estrangeiras de várias tabelas se referem a registros de 'OntologyEntry' (como é o caso de 'technology_type_id' em 'ArrayDesign' e 'protocol_type_id' em 'Protocol'). Os valores possíveis para esses campos são registros encontrados na MGED *Ontology*. Como os termos dessa ontologia têm relacionamentos onde um é subclasse ou instância do outro, o esquema de armazenamento no banco se dá simplesmente através de auto-relacionamento dos termos da ontologia, onde um termo é pai do outro.

4.1.4 – Uso de Ontologias

O MGED *Ontology* descreve uma ontologia especializada para termos utilizados em anotações de experimentos com microarranjos. A organização da ontologia é baseada em uma hierarquia de classes, e no mais baixo nível dessa hierarquia, algumas classes possuem instâncias de valores. É esse nível de instâncias que fornece domínios de valores para algumas propriedades do experimento. Todos os níveis da hierarquia possuem definição para os termos.

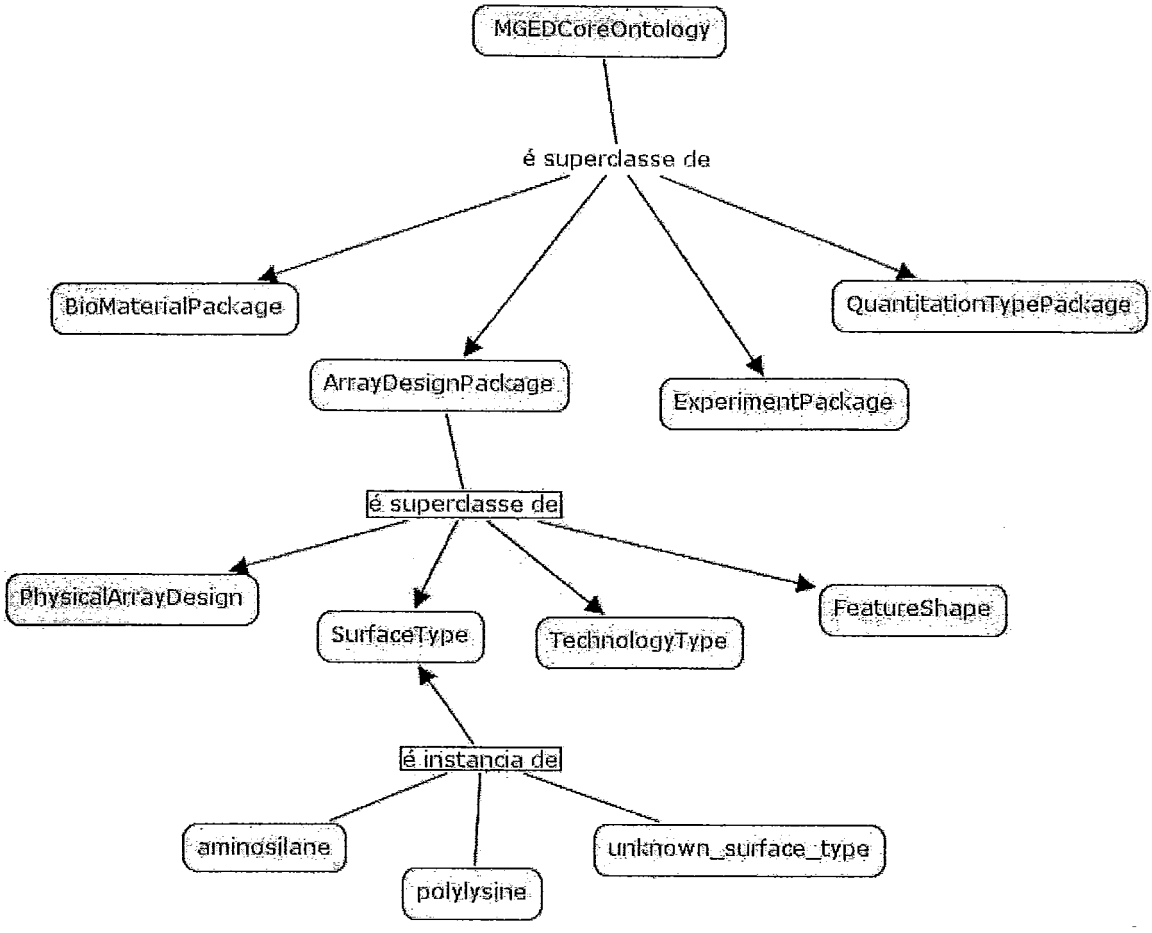

Figura 14: Parte do MGED Ontology

As ontologias utilizadas pelo protótipo devem ser previamente carregadas para o banco de dados a partir dos dados capturados no sítio do MGED *Ontology*. Se valores de propriedades que não constam na ontologia forem utilizados pelo usuário, tais valores serão adicionados à ontologia pelo sistema. Entretanto o sistema não prevê a inserção de novas propriedades. Caso isto seja necessário, elas podem ser cadastradas como parâmetros de protocolos.

Na seção anterior foi mostrado como as ontologias foram mapeadas no subesquema RAD. A Figura 14 mostra o esquema de uma parte da MGED *Ontology*, detalhando a parte de descrição da lâmina do microarranjo.

4.1.5 – Esquema estrela do *data warehouse*

Elaboramos um esquema estrela para o módulo de *dataware house* para que pudéssemos utilizar a abordagem OLAP na análise dos dados quantificados dos genes. Como já foi explicado, a hierarquia de classes fornecida pela MGED *Ontology* pode ser aproveitada nesse módulo para operações de generalização e especialização, que nesse contexto são chamadas de *drill-up* e *drill-down*. A MGED *Ontology* ainda não possui classes que contemplam todas as dimensões escolhidas neste esquema, portanto tais operações só podem ser utilizadas de forma bem restrita. A figura 15 ilustra este esquema.

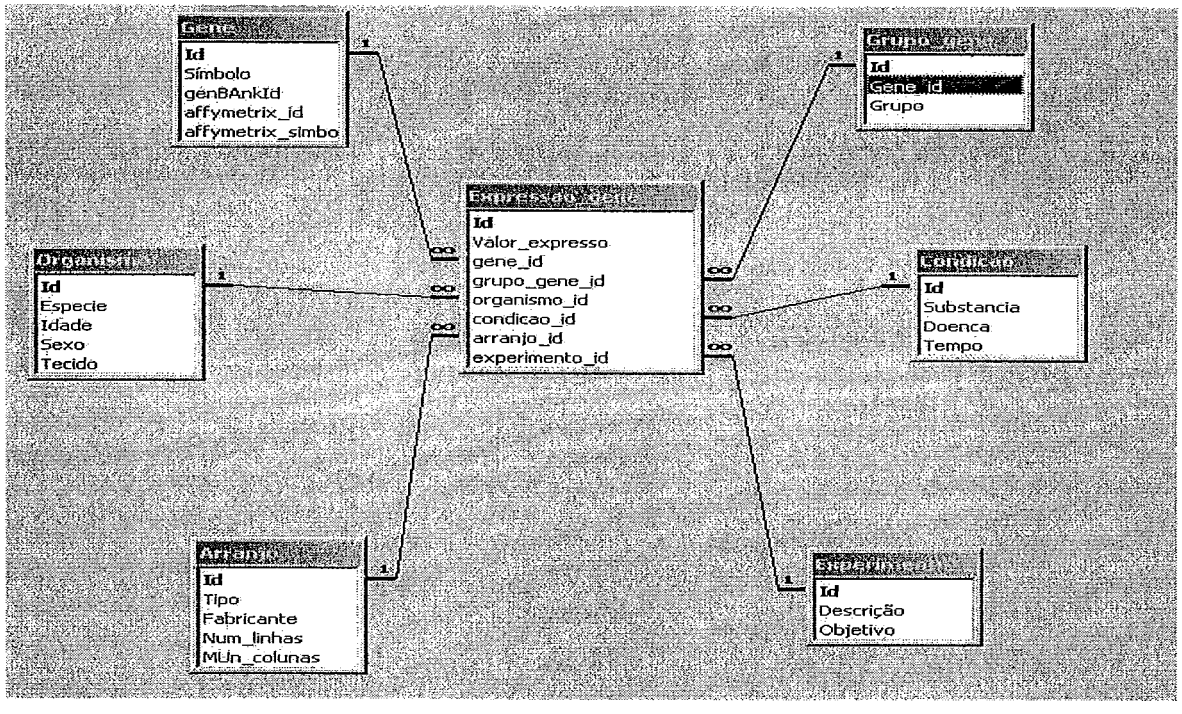

Figura 15: Esquema estrela do DW

4.2 – O ambiente de produção

O ambiente real de experimentos com microarranjos onde foi implantando o protótipo desenvolvido nessa dissertação é um dos laboratório do departamento de Micobacteriose da Fundação Instituto Oswaldo Cruz (FIOCRUZ) do Rio de Janeiro. Esse laboratório foi criado para comportar a aquisição da impressora e do *scanner* de microarranjos e é utilizado especialmente para a realização deste tipo de experimento. A FIOCRUZ é a única instituição de pesquisa no Rio de Janeiro que possui estes equipamentos. Alguns de seus pesquisadores já fizeram intercâmbio em outras

instituições de ensino, no Brasil e no exterior, para adquirir experiência com uso de microarranjos. Outros estão tendo os primeiros contatos com este tipo de experimento.

Este laboratório é dividido em três ambientes. O primeiro deles é onde se localiza o equipamento responsável pela hibridização das lâminas com microarranjos com as amostras de RNA. Neles também é possível realizar algumas atividades experimentais de bancada (*wet lab*). Em um outro ambiente, fica o *scanner* de microarranjos e um computador com o sistema que o controla. No terceiro ambiente, mais isolado, por necessitar de controle de umidade, fica a impressora (*spotter*) de microarranjos, além de um computador com o sistema que o controla. Algumas atividades que não são específicas para o experimento com microarranjos, como cultivo de amostrar e extração de RNA, são feitas em outros laboratórios.

A realização deste tipo de experimento nesse local é relativamente recente. Durante um grande período a maior parte do trabalho realizado foi no sentido de calibrar os equipamentos e ajustar protocolos para se começassem a obter lâminas de microarranjos de qualidades. Até então, o processamento dos dados obtidos era feito de forma bastante manual. Os únicos sistemas utilizados eram aqueles destinados ao controle do *spotter* e do *scanner* de microarranjos, além daquele para quantificar a imagem gerada pelo *scanner*. Os dados gerados nesses sistemas eram visualizados através de planilhas de dados e eventualmente a junção entre dois ou mais arquivos, como, por exemplo, a associação de lista de genes com os locais onde foram impressos na lâmina, era feita manualmente. Foram feitas algumas análises básica com o uso de planilhas de dados, porém devido à imensa quantidade de dados, era muito complicada a realização de análises mais complexas. Em relação às etapas *in vitro*, os protocolos vinham sendo anotados e arquivados em documentos de texto comuns, e tiveram (e ainda têm) muita utilidade para se rastrear os melhores parâmetros para execução dos procedimentos laboratoriais, ou seja, para se fazer ajustes desses parâmetros de forma a se obter melhores produtos experimentais.

O servidor *Web* e o SGBD foram instalados em uma das máquinas do laboratório de microarranjo. A implantação deste protótipo está sendo feita em etapas, se adaptando as necessidades dos pesquisadores da FIOCRUZ, já que mesmo para eles o processo completo do experimento ainda não está totalmente estabelecido. A parte de análise de dados, por exemplo, ainda está em fase inicial de elaboração.

4.3 – Considerações sobre a utilização do protótipo

A implantação de protótipo proporcionou algumas contribuições básicas para o ambiente experimental. O primeiro grande impacto da implantação desse protótipo foi a eliminação de processamento manual de uma grande quantidade de dados, que é um trabalho muito dispendiosos e sujeito a erros. Foi possível, assim, a visualização mais rápida dos dados obtidos a partir dos microarranjos e a verificação, também mais rápida, da consistência dos mesmos. Outra mudança estrutural básica foi a introdução de uma estrutura de organização dos dados, que previne situações, não muito difíceis de ocorrerem, de dificuldade de localização de dados importantes para o experimento, como lista de genes que foram impressos no arranjo, devido ao fato de nem sempre haver local padrão para depósito desses arquivos.

Contribuições mais avançadas ocorrerão com o aumento do número de experimentos realizados. Será possível perceber melhor as potencialidades da arquitetura proposta, pois será mais fácil a realização de estudos comparativos entre experimentos diferentes, seja pela maior facilidade para se resgatar dados de diferentes experimentos, e em meio a um grande número de dados, ou seja pelos procedimentos padronizados e metadados disponíveis, que torna possível a comparação desses dados de alguma maneira.

O protótipo propicia uma certa flexibilidade para os biólogos executarem *workflows* externamente ao sistema, através da interface gráfica do Kepler, utilizando os próprios *workflows* elaborados para o sistema, ou estendendo as funcionalidades dele. O sistema R, integrado ao Kepler, agregar várias funcionalidades, como acesso ao banco de dados (que eventualmente pode ser útil), funções para manipulação de dados e arquivos, geração de gráficos e exibição de imagens. Os dados podem ser recuperados do banco de dados ou dos arquivos originais desses dados, que também foram armazenados. Perde-se a funcionalidade de registro de execução, mas se ganha na a flexibilidade de usar uma interface gráfica de modelagem e execução de workflows para elaborar e executar seqüências de análises não previstas pelo sistema.

Houve algumas dificuldades na implementação desse protótipo. A primeira delas foi com o uso de um componente do Kepler que faz acesso aos dados no SGBD. O componente não funcionou devido ao (imenso) tamanho do esquema do GUS. Isso

compromete o funcionamento do sistema já que o sistema precisa fazer acesso ao banco de dados em todas as etapas. Essa questão foi resolvida com acesso ao banco de dados através de rotinas elaboradas no Sistema R, usando *driver* ODBC. A arquitetura do protótipo foi um pouco prejudicada pois o acesso ao SGBD ficou embutido em rotinas que realizam outras atividades do *workflow*.

Também houve dificuldade para mapear os dados nas tabelas pertinentes no sub-esquema RAD do GUS, apesar da documentação existente no GUS *Schema Browser* (http://www.gusdb.org/documentation.php), houve incerteza no uso de alguns campos e até mesmo de algumas tabelas. Faz parte do GUS todo um ferramental para carregar dados de repositórios públicos para o seu esquema relacional, e são essas ferramentas constituem a principal funcionalidade do GUS. Talvez por isso não tenha sido dado o devido tratamento a esta questão do mapeamento manual de dados nas tabelas do esquema.

Capítulo 5 – Conclusões

As pesquisas na área de genética têm crescido enormemente nos últimos 10 anos e a quantidade de dados gerados a partir dela tem aumentado exponencialmente. Dentro dessa área, há um especial interesse nos estudos a respeito das funções dos genes. A técnica de microarranjos permite o estudo de centenas ou milhares de genes ao mesmo tempo, sendo possível verificar a variação da expressão gênica das células dada uma mudança no ambiente celular.

Os experimentos com microarranjos são compostos, assim como outros experimentos biológicos, por uma série de atividades experimentais. Entretanto, a composição dessas atividades nem sempre é vista como um processo que deve ser tratado de forma integrada. Usualmente são realizadas de maneira isolada e sem integração automatizada entre elas. Assim, dificulta-se a monitoração e acompanhamento deste processo. Outro problema é a falta de padronização, tanto de conteúdo quanto de estrutura, nas anotações do experimento, o que dificulta bastante a realização de estudos comparativos entre experimentos diversos.

Tem havido uma evolução de apoio computacional que para este tipo de experimento. Vimos propostas de padrões para conteúdo de dados, através do uso de ontologias, e para estruturação dos mesmos, com o uso de um esquema relacional padrão difundido na comunidade científica. Analisamos também sistemas para definição e execução de *workflow* científico capazes de modelar muito bem experimentos *in silico*; e também sistemas especializados para gerenciar dados de experimentos com microarranjos. Entretanto nenhum deles representava uma solução completa para gerenciar todo o processo experimental com microarranjos, com a modelagem do mesmo como uma instância de um *workflow* científico.

Nesta dissertação visamos sistematizar as atividades que fazem parte do processo experimental com microarranjos e desenvolver um ambiente para gerência e execução de um *workflow* científico especializado nesse tipo de experimento. Para isso, propomos uma arquitetura para desenvolvimento de um sistema de gerência de experimentos com microarranjos e desenvolvemos um protótipo baseado nessa

arquitetura. Esse protótipo foi implantado em um ambiente real de experimentos, na FIOCRUZ do Rio de Janeiro.

Nossa solução atende a uma série de requisitos para gerenciamento de experimentos com microarranjos. O principal deles é de ter a funcionalidade de máquina de execução de *workflow* para o processo de experimento com microarranjo, que faz a composição de diferentes aplicativos e que também realiza o registro de execução dos mesmos. Outros requisitos atendidos são a padronização e estruturação dos dados relativos ao experimento, obtidos com o uso do esquema GUS/RAD, que atende as especificações do MIAME para experimentos com microarranjos, associado ao uso de uma ontologia especializada para anotações deste tipo de experimento. Por fim, o registro das atividades executadas permite ao pesquisador monitorar o fluxo dos dados e rastrear, por exemplo, a origem de uma inconsistência caso a mesma só tenham sido identificada em uma etapa mais avançada do experimento. O requisito de proveniência de dados é atendido através da combinação desse monitoramento de dados com o registro de metadados do experimento, também possibilitado pelo nosso sistema.

Podemos dizer que nossa arquitetura tem uma abordagem de composição de outras soluções existentes que atendem parcialmente aos requisitos do experimento. Associamos uma máquina de execução de *workflows* ao uso de ontologia e de um esquema relacional padrão para atender o conjunto completo de tais requisitos. Não foi criado nenhum aplicativo para realizar alguma etapa do experimento, mas o uso desses aplicativos foi, de alguma forma, intermediado pelo nosso sistema.

A principal vantagem desta arquitetura em relação a outras abordagens de *workflow* científico é que ela é direcionada especificamente a microarranjos, o que permite um tratamento dirigido especificamente a este tipo de experimento. Lidamos, por exemplo, com uma ontologia e um esquema relacional especializados para experimentos com microarranjos.

Houve também uma preocupação na utilização de padrões abertos aceitos na comunidade científica, para que os dados gerenciados pelo nosso sistema fossem mais facilmente compartilhados e comparados com dados de outros experimentos.

O protótipo implementado introduziu funcionalidades básicas no ambiente de produção, onde a interação entre as etapas do experimento era feita de forma bem

manual (não automatizada). A introdução de uma estrutura de organização e a automatização da manipulação de dados que são passados de um aplicativo para outro são funcionalidades simples mas que agilizam o processo experimental e diminuem a incidência de erros.

Funcionalidades mais avançadas serão percebidas quando o número de experimentos registrados for grande o suficiente para se realizar estudos interessantes com cruzamento de dados de experimentos distintos do ambiente de produção da FIOCRUZ, que ainda começa a explorar as potencialidades deste tipo de experimento. Porém, de imediato, já se percebeu a possibilidade de utilização dos *workflows* elaborados para serem executados pelo sistema, diretamente na interface gráfica do Kepler. Perde-se a funcionalidade do registro de execução, mas se ganha na flexibilidade do uso do Kepler, associado a um sistema estatístico, para estender ou altera o escopo do workflow e realizar análises não previstas inicialmente pelo sistema.

O Kepler se mostrou um ambiente de definição e execução de workflows prático e eficiente. O uso de uma interface gráfica, com funcionalidades integradas para manipulação de arquivos, acesso a banco de dados e criação e visualização de gráficos, além da representação visual da cadeia de tarefas, facilitou enormemente a elaboração de *workflows*. Além disso, o fato de a definição do *workflow* poder ser destacada da interface gráfica foi de grande utilidade, já que precisamos executar esse *workflow* de forma automática pelo gerenciador de processos, sem o uso de interface gráfica.

Dentre as contribuições desta dissertação, podemos destacar a elaboração de *workflows* que realizam análises de dados obtidos no experimento, e que podem ser utilizados (e estendidos) na interface gráfica do Kepler de forma independente do protótipo construído.

Outra contribuição foi a definição de uma série de requisitos para sistemas de gerenciamento de experimentos com microarranjos, identificados a partir das necessidades e objetivos do experimento. Analisando diversos estudos que envolvem esta tecnologia e percebemos a forte de necessidade de integração de experimentos diversos de forma a se possibilitar estudos comparativos, e também devido à natureza comparativa deste tipo de experimento. A necessidade de integração entre etapas em um mesmo experimento também se mostrou evidente, devido principalmente a grande

quantidade de dados trafegados entre elas. A maioria dos requisitos é motivada por essas necessidades de integração.

Também contribuímos com a elaboração de um esquema estrela para uma visão multidimensional da expressão gênica, onde algumas das dimensões são os fatores que determinam a variação de expressão gênica. Essa visão resumida das expressões gênicas, agrupada de acordo com os fatores que influenciaram seus valores, atende bem aos objetivos de estudos comparativos dos experimentos com microarranjo.

Entretanto, a principal contribuição deste trabalho é promover aumento significativo no grau de integração entre experimentos diversos, permitindo análises e comparações entre experimentos realizados de forma independente, mas que possuem interseções nos objetos de estudo (normalmente os genes analisados). Essa integração a um ambiente maior, ao adotar um modelo de dados único, no caso o GUS, permite que os experimentos em microarranjo não fiquem isolados das demais atividades de análise genômica e anotações. Resultados adicionais obtidos através desse tipo de análise cruzada são bastante promissores, pois dão uma visão ainda mais ampla sobre o comportamento genético.

Esse desenvolvimento faz parte de um ambiente maior, o BioWebDB (www.biowebdb.org), que visa gerenciar dados e workflows científicos em bioinformática para promover interoperabilidade entre diferentes fontes de dados genômicos e ferramentas de análises de tais dados, baseado no uso de ontologias, computação distribuída e algoritmos de inteligência artificial.

Um possível trabalho futuro seria a possibilidade de se criar novos *workflows* no Kepler e torná-los acessível, sem grandes esforços de programação, pelo sistema gerenciador do experimento. Como existe a possibilidade de o pesquisador utilizar isoladamente os *workflows* do sistema na interface gráfica do Kepler, ele pode estender as análises feitas pelo mesmo e acabar por criar um novo *workflow*, com um novo padrão de análises, que deseje incluir no sistema para que seja executado e gerenciado por ele. Porém, na estrutura atual do sistema, é exigido um certo esforço de programação para alterar o código-fonte do sistema. A nova funcionalidade poderia se constituir de uma interface para cadastrar o novo *workflow* e definir a origem dos dados de entrada e o destino dos dados produzidos para o após a execução do *workflow*.

Outro trabalho complementar a este seria o desenvolvimento de componentes no Kepler especializados em armazenamento e recuperação de dados do esquema relacional GUS, uma vez que os sistemas de estatísticas normalmente não lêem dados vindos de uma base de dados, mas sim de um arquivo. Assim, o acesso ao banco de dados ficaria encapsulado em um componente, e não misturado a outros procedimentos como está sendo feito atualmente.

Referências Bibliográficas

ADJAYE, J., HERWIG, R., ERRMANN, D., et al, 2004, "Cross-species hybridisation of human and bovine orthologous genes on high density cDNA microarrays", *BMC Genomics*, v. 5, n. 1, pp. 83-

AFFYMETRIX, 2006, "Affymetrix Genechip Arrays". Em: http://www.affymetrix.com/products/arrays/index.affx.

ALTSCHUL, S. F., GISH, W., MILLER, W., et al, 1990, "Basic local alignment search tool", *Journal of Molecular Biology*, v. 215, pp. 403-410

AUGENLICHT, L. H., KOBRIN, D., PAVLOVEC, A., et al, 1984, "Elevated expression of an endogenous retroviral long terminal repeat in a mouse colon tumor", *Journal of Biological Chemistry*, v. 259, n. 3, pp. 1842-1847

AUGENLICHT, L. H., WAHRMAN, M. Z., HALSEY, H., et al, 1987, "Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate in vitro.", *Cancer Research*, v. 47, n. 22, pp. 6017-6021

BAKEWELL, D. J.,WIT, E., 2005, "Weighted analysis of microarray gene expression using maximum-likelihood Source", *Bioinformatics*, v. 21, n. 6, pp. 723-729

BALL, C. A., BRAZMA, A., CAUSTON, H., et al, 2004, "Submission of Microarray Data to Public Repositories", *PLoS Biology*, v. 2, n. 9, pp. 1276-1277

BALLMAN, K. V., GRILL, D. E., OBERG, A. L., et al, 2004, "Faster cyclic loess: normalizing RNA arrays via linear models", *Bioinformatics*, v. 20, n. 16, pp. 2778-2786

BARRETT, T.,EDGAR, R., 2006, "Mining Microarray Data at NCBI's Gene Expression Omnibus (GEO)", *Methods in Molecular Biology*, v. 338, pp. 175-190

BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J., et al, 2007, "GenBank:update", *Nucleic Acids Research*, v. 35, pp. D21-D25

BIOCONDUCTOR, 2006, "Bioconductor Project". Em: http://www.bioconductor.org/.

BIOPERL, 2006, "BioPerl Project". Em: http://www.bioperl.org/.

BOECKMANN, B., BAIROCH, A., APWEILER, R., et al, 2003, "The SWISS-PROT protein knowledgebase and its supplement", *Nucleic Acids Research*, v. 31, n. 1, pp. 365-370

BRAZMA, A., HINGAMP, P., QUACKENBUSH, J., et al, 2001, "Minimum information about a microarray experiment (MIAME) - toward standards for microarray data", *Nature Genetics*, v. 29, pp. 365-371

BUHLER, J., 2003, "Provably Sensitive Indexing Strategies for Biosequence Similarity Search", *Journal of Computational Biology*, v. 10, n. 3-4, pp. 399-417

BUNEMAN, P., KHANNA, S., TAN, W.-C., 2001, "Why and Where: Characterization of Data Provenance", *Lecture Notes in Computer Science*, v. 1973, pp. 316-330

CAMOGLU, O., KAHVECI, T., SINGH, A. K., 2003, "PSI: Indexing Protein Structures for Fast Similarity Search", *Bioinformatics*, v. 19, n. Suppl 1, pp. i81-i83

CAVALCANTI, M., 2003, *Scientific Resources Management: Towards an In Silico Laboratory*, Tese de Doutorado, COPPE/UFRJ, RJ, Brasil

CELL, 2006, "Cell". Em: http://www.cell.com.

CHO, S.-B.,WON, H.-H., 2003, "Machine learning in DNA microarray analysis for cancer classification", *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003*, v. 19, pp. 189-198

CRICK, F., 1970, "Central Dogma of Molecular Biology", *Nature*, v. 227, pp. 561-563

DAVID J.LOCKHART, HELIN DONG, MICHAEL C.BYRNE, et al, 1996, "Expression monitoring by hybridization to high-density oligonucleotide arrays", *Nature Biotechnology*, v. 14, pp. 1675-1680

DÁVILA, A. M. R., LORENZINI, D. M., MENDES, P. N., et al, 2005, "GARSA: genomic analysis resources for sequence annotation", *Bioinformatics*, v. 21, n. 23, pp. 4302-4303

EBI, 2006, "EMBL-EBI". Em: http://www3.ebi.ac.uk/Services/DBStats/.

EDGAR, R., DOMRACHEV, M., LASH, A. E., 2002, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository", *Nucleic Acids Research*, v. 30, n. 1, pp. 207-210

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O., et al, 1998, "Cluster analysis and display of genome-wide expression patterns", *Proceedings of the National Academy of Sciences USA*, v. 95, n. 25, pp. 14863-14868

EKER, J., JANNECK, J. W., LEE, E. A., et al, 2003, "Taming Heterogeneity---the Ptolemy Approach", *Proceedings of the IEEE*

EMBL, 2007a, "MIAMExpress". Em: http://www.ebi.ac.uk/miamexpress/login.html.

EMBL, 2007b, "Tab2MAGE ArrayExpress submissions". Em: http://www.ebi.ac.uk/cgi-bin/microarray/tab2mage.cgi.

EMBL, 2007c, "The EMBL Nucleotide Sequence Database, statistics". Em: http://www3.ebi.ac.uk/Services/DBStats.

EMBO, 2006, "The EMBO Journal ". Em: http://www.emboj.org/.

ENSEMBL, 2006, "Ensembl". Em: http://www.ensembl.org/.

GBG, 2006, "Microarray Data Analysis", *Genomics and Bioinformatics Group*.http://discover.nci.nih.gov/microarrayAnalysis/Statistical.Tests.jsp

GE, 2002, "Microarray Handbook". *GE Healthcare*. Em http://www6.gelifesciences.com/APTRIX/upp00919.nsf/Content/WD:Microarray+Hand(219987857-B500)

GEO, 2006, "Gene Expression Omnibus". Em: http://www.ncbi.nlm.nih.gov/geo/.

GMOD, 2007, "Generic Model Organism Database". Em: http:// www.gmod.org.

GO, 2006, "The Gene Ontology". Em: http://www.geneontology.org/.

GRUBER, T., 1993, "A Translation Approach to Portable Ontology Specifications", *Knowledge Acquisition*, pp. 199-220

GUÉRIN , E., MARQUET, G., BURGUN, A., et al, 2007, "Integrating and Warehousing Liver Gene Expression Data and Related Biomedical Resources in GEDAW", *DILS*, v. 3615, pp. 158-174

GUS, 2006, "GUS: The Genomics Unified Schema". Em: http://www.gusdb.org/.

HEAD-GORDON, T.,WOOLEY, J. C., 2001, "Computational challenges in structural and functional genomics", *IBM Systems Journal*, v. 40, n. 2, pp. 265-296

HGP, 2006a, "Human Genome Project Information". Em: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml.

HGP, 2006b, "Potential Benefits of Human Genome Project Research". *Human Genome Project*. Em http://www.ornl.gov/sci/techresources/Human_Genome/project/benefits.shtml

HOAGLIN, D. C., MOSTELLER, F., TUKEY, J. W., 1983, *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, Inc

HOLLINGSWORTH, D., 1995, "The Workflow Reference Model", *The Workflow Management Coalition*, v. TC00-1003, n. 1.1

HUNT, E., ATKINSON, M. P., IRVING, R. W., 2002, "Database indexing for large DNA and protein sequence collections", *The VLDB Journal*, v. 11, n. 3, pp. 256-271

HUNTER, L., 1993, "Molecular Biology for Computer Scientists". In The MIT Press, *Artificial Intelligence & Molecular Biology*, chapter 1

IBM, 2006, "Lotus Workflow". Em: http://www.lotus.com/workflow.

IKEO, K., ISHI-I, J., TAMURA, T., et al, 2003, "CIBEX: center for information biology gene expression database.", *Comptes rendus biologies*, v. 326, n. 10, pp. 1079-1082

JARKE, M., LENZERINI, M., VASSILIOU, Y., et al, 2003, *Fundamentals of Data Warehouses*. 2nd Ed., Springer-Verlag

JARVINEN, A. K., HAUTANIEMI, S., EDGREN, H., et al, 2004, "Are data from different gene expression microarray platforms comparable?", *Genomics*, v. 83, n. 6, pp. 1164-1168

JONATHAN KNIGHT, 2001, "When the chips are down", *Nature*, v. 410, n. Nature, pp. 860-861

KENNEDY, J. B., KUKLA, R., PATERSON, T., 2005, "Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration", *DILS*, v. LNBI 3615, pp. 80-95

KEPLER, 2006, "Kepler Project". Em: http://kepler-project.org/.

KIRSTEN, T., DO, H.-H., RAHM, E., 2004, *A Data Warehouse for Multidimensional Gene Expression Analysis*. Interdisciplinary Centre for Bioinformatics, Universität Leipzig, DE

LANCET, 2006, "The Lancet". Em: http://www.thelancet.com/.

LEE, E. A.,NEUENDORFFER, S., 2000, *MoML - A Modeling Markup Language in XML — Version 0.4*

LEUNG, Y. F., 2002, "Unravelling the mystery of microarray data analysis", *Trends in Biotechnology*, v. 20, n. 9, pp. 366-368

LEUNG, Y. F., LAM, D. S. C., PANG, C. P., 2001, "The miracle of microarray data analysis", *Genome Biology*, v. 2, n. 9, pp. reports4021.1-reports4021.2

LU, Y., LU, S., FOTOUHI, F., et al, 2004, "Incremental genetic K-means algorithm and its application in gene expression data analysis", *BMC Bioinformatics*, v. 5, n. 172

LUDÄSCHER, B., ALTINTAS, I., BERKLEY, C., et al, 2005, "Scientific Workflow Management and the Kepler System", *Concurrency and Computation: Practice & Experience*, v. 18, n. 10, pp. 1039-1065

MANDUCHI, E., GRANT, G. R., HE, H., et al, 2004, "RAD and the RAD Study-Annotator: an approach to collection, organization and exchange of all relevant information for high-throughput gene expression studies", *Bioinformatics*, v. 20, n. 4, pp. 452-459

MARKOWITZ, V. M., CAMPBELL, J., CHEN, I.-M. A., et al, 2003, "Integration Challenges in Gene Expression Data Management". In LaCroix, Z. and Critchlow, T., *Bioinformatics: Managing Scientific Data*, chapter 10

MATHWORKS, 2006, "MATLAB". Em: http://www.mathworks.com/products/matlab/.

MGED, 2006a, "MGED - Microarray Gene Expression Data Society". Em: http://www.mged.org/.

MGED, 2006b, "MGED Mission". Em: http://www.mged.org/Mission/index.html.

MIAME, 2005, "Minimum information about a microarray experiment". Em: http://www.mged.org/Workgroups/MIAME/miame.html.

MICROSOFT, 2006, "Microsoft Message Queuing". Em: http://www.microsoft.com/windowsserver2003/technologies/msmq/default.msp x.

MITCHELL, S. A., BROWN, K. M., HENRY, M. M., et al, 2004, "Inter-platform comparability of microarrays in acute lymphoblastic leukemia", *BMC Genomics*, v. 5, n. 1, pp. 71-

MYGRID, 2006, "MyGrid Project". Em: http://www.mygrid.org.uk/.

NATURE, 2002, "Microarray standards at last", *Nature*, v. 419, n. 6905, pp. 323-

NATURE, 2006, "Nature Group". Em: http://www.nature.com/.

O'CONNELL, M., 2003, "Differential Expression, Class Discovery and Class Prediction using S-PLUS and S+ArrayAnalyzer", *SIGKDD Explorations Newsletter*, v. 5, n. 2, pp. 38-47

ORACLE, 2006, "Oracle Workflow". Em: http://www.oracle.com/technology/products/applications/workflow/index.html.

OWG, 2006, "MGED Ontology". Em: http://mged.sourceforge.net/ontologies/index.php.

PAN, W., LIN, J., LE, C. T., 2002, "How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach", *Genome Biology*, v. 3, n. 5, pp. research0022.1-research0022.10

PARKINSON, H., SARKANS, U., SHOJATALAB, M., et al, 2005, "ArrayExpress—a public repository for microarray gene expression data at the EBI", *Nucleic Acids Research*, v. 33, pp. D553-D555

PASQUIER, C., GIRARDOT, F., FOMBELLE, J. F., et al, 2004, "THEA: ontology-driven analysis of microarray data", *Bioinformatics*, v. 20, n. 16, pp. 2636-2643

PERL, 2003, "Perl Mongers". Em: http://www.perl.org/.

PIATETSKY-SHAPIRO, G.,TAMAYO, P., 2003, "Microarray data mining: facing the challenges", *SIGKDD Explorations Newsletter*, v. 5, n. 2, pp. 1-5

PROSDOCIMI, F., FILHO, F. C., CERQUEIRA, G. C., et al, 2003, "Bioinformática: Manual do Usuário", *Biotecnologia Ciência & Desenvolvimento*, n. 29, pp. 18-31

PROTOCOL ONLINE, 2007a, "mRNA Isolation". Em: http://www.protocol-online.org/prot/Molecular_Biology/RNA/RNA_Extraction/mRNA_Isolation/.

PROTOCOL ONLINE, 2007b, "RT-PCR". Em: http://www.protocol-online.org/prot/Molecular_Biology/PCR/RT-PCR/.

QIN, L. X.,KERR, K. F., 2004, "Empirical evaluation of data transformations and ranking statistics for microarray analysis", *Nucleic Acids Research*, v. 32, n. 18, pp. 5471-5479

R, 2006, "The R Project for Statistical Computing". Em: http://www.r-project.org/.

RUBINSTEIN, B. I. P., MCAULIFFE, J., CAWLEY, S., et al, 2003, "Machine learning in low-level microarray analysis", *SIGKDD Explorations Newsletter*, v. 5, n. 2, pp. 130-139

SANTOS, R. T., 2004, *O ambiente 10+C para a definição e execução de workflows in silico através de serviços web.*, Tese de Mestrado, COPPE, UFRJ, Rio de Janeiro, RJ

SILVA, F. N., 2006, *In Services: Um Sistema para Gerenciamento de Dados Intermediários em Workflows Científicos na Bioinformática*, IME/EB , Rio de Janeiro, RJ

SMYTH, G. K.,SPEED, T. P., 2003, "Normalization of cDNA Microarray Data", *Methods*, v. 31, pp. 265-273

SNOMED, 2006, "SNOMED - Systematized Nomenclature for Medicine". Em: http://www.snomed.org/.

TAVERNA, 2006, "Taverna Project". Em: http://taverna.sourceforge.net/index.php.

THYGESEN, H. H.,ZWINDERMAN, A. H., 2004, "Comparing transformation methods for DNA microarray data", *BMC Bioinformatics*, v. 5, n. 77

TODABIOLOGIA, 2006, "Alimentos Transgênicos". Em: http://www.todabiologia.com/genetica/transgenicos.htm.

URIARTE, R. D.,ANDRÉS, S. A., 2006, "Gene selection and classification of microarray data using random forest", *BMC Bioinformatics*, v. 7, n. 3

W3C, 2006, "Web Services Architecture". Em: http://www.w3.org/TR/ws-arch/.

WESTBROOK, J., FENG, Z., JAIN, S., et al, 2002, "The Protein Data Bank: unifying the archive", *Nucleic Acids Research*, v. 30, n. 1, pp. 245-248

WHETZEL, P. L., PARKINSON, H., CAUSTON, H. C., et al, 2006, "The MGED Ontology; a resource for semantics-based description of microarray experiments", *Bioinformatics*

WOO, Y., AFFOURTIT, J., DAIGLE, S., et al, 2004, "A Comparison of cDNA, Oligonucleotide, and Affymetrix GeneChip Gene Expression Microarray Platforms", *Journal of Biomolecular Techniques*, v. 15, pp. 276-284

YU, S. L., CHEN, H. W., YANG, P. C., et al, 2004, "Differential gene expression in gram-negative and gram-positive sepsis", *American journal of respiratory and critical care medicine*, v. 169, n. 10, pp. 1135-1143

ZIEN, A., FLUCK, J., ZIMMER, R., et al, 2002, "Microarrays: How Many Do You Need?", *Proceedings of the Sixth Annual Conference on Research in Computational Molecular Biology*, pp. 321-330

Anexo A – The MIAME Checklist

The purpose of this checklist is to guide authors, journal editors and referees in helping them to ensure that the data supporting published results based on microarray experiments are made publicly available in a format that enables unambiguous interpretation of the data and potential verification of the conclusions (see [1]). For more detail regarding the rationale of MIAME see [2]. MGED strongly recommends that the data is made publicly available through one of the public repositories for microarray data (see [3]).

Experiment Design:

- The goal of the experiment – one line maximum (e.g., the title from the related publication)
- A brief description of the experiment (e.g., the abstract from the related publication)
- Keywords, for example, *time course, cell type comparison, array CGH* (the use of MGED ontology terms is recommended).
- Experimental factors - the parameters or conditions tested, such as *time, dose, or genetic variation* (the use of MGED ontology terms is recommended).
- Experimental design - relationships between samples, treatments, extracts, labeling, and arrays (e.g., a diagram or table).
- Quality control steps taken (e.g., replicates or dye swaps).
- Links to the publication, any supplemental websites or database accession numbers.

Samples used, extract preparation and labelling:

- The origin of each biological sample (e.g., name of the organism, the provider of the sample) and its characteristics (e.g., gender, age, developmental stage, strain, or disease state).
- Manipulation of biological samples and protocols used (e.g., growth conditions, treatments, separation techniques).
- Experimental factor value for each experimental factor, for each sample (e.g., '*time* = 30 min' for a sample in a time course experiment).
- Technical protocols for preparing the hybridization extract (e.g., the RNA or DNA extraction and purification protocol), and labeling.
- External controls (spikes), if used.

Hybridization procedures and parameters:

- The protocol and conditions used for hybridization, blocking and washing, including any post-processing steps such as staining

Measurement data and specifications:

- Data

- The raw data, i.e. scanner or imager and feature extraction output (providing the images is optional). The data should be related to the respective array designs (typically each row of the imager output should be related to a feature on the array – see Array Designs).
- The normalized and summarized data, i.e., set of quantifications from several arrays upon which the authors base their conclusions (for gene expression experiments also known as gene expression data matrix and may consist of averaged normalized log ratios). The data should be related to the respective array designs (typically each row of the summarized data will be related to one biological annotation, such as a gene name).

- Data extraction and processing protocols,
  - Image scanning hardware and software, and processing procedures and parameters.
  - Normalization, transformation and data selection procedures and parameters.

Array Design:

- General array design, including the platform type (whether the array is a spotted glass array, an *in situ* synthesized array, etc.); surface and coating specifications and spotting protocols used (for custom made arrays), or product identifiers (the name or make, catalogue reference numbers) for commercially available arrays.
- Array feature and reporter annotation, normally represented as a table (for instance see Tables 1, 2 below), including
  - For each feature (spot) on the array, its location on the array (e.g., metacolumn, metarow, column, row) and the reporter present in the location (note that the same reporter may be present on several features).
  - For each reporter unambiguous characteristics of the reporter molecule, including
    - Reporter role – control or measurement
    - The sequence for oligonucleotide based reporters
    - The source, preparation and database accession number for long (e.g., cDNA or PCR product based) reporters
    - Primers for PCR product based reporters
  - Appropriate biological annotation for each reporter, for instance a gene identifier or name (note that different reporters can have the same biological annotation)
- Principal array organism(s)

Table 1. Oligonucleotide array description file example:

Feature				Reporter						Biological annotation			
Coordinates on Array				Reporter ID (user defined) Oligo ID	Biosequence Type	Sequence	DDBJ/ EMBL/ Genbank	Reporter Usage	Control Type	ID	Designation	Related Gene Symbol, if appropriate	Database Entry
Meta Col	Meta Row	Col	Row										
1	1	1	1	Cy3Cy5	Oligo	AAAAAAAAAAAAAAAAAA	–	Control	Positive	C001_01	Labeled. oligo	–	–
1	1	2	1	M00868_01	Oligo	ACCAGCAGATACCTCCTTG	D83002	Experimental	–	C002_01	Gene	ALK	LocusID 11682
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4	6	10	8	M00264_01	Oligo	ATGTCCGTTGAATTGG	D83002	Experimental	–	C002_01	Gene	ALK	LocusID 11682
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4	6	11	8	M02404_01	Oligo	AGTGGCGAGGAGGAGGAC	L11065	Experimental	–	C449_01	Gene	OPRK1	LocusID 18387
4	6	12	8	M03172_01	Oligo	CCACCACCAAGACCTACTCC	U34891	Experimental	–	C450_01	Gene	KLRA9	LocusID 16640

Table 2. cDNA array description file example:

Feature				Reporter						Biological annotation			
Coordinates on Array				Reporter ID (user defined) HGMP Ref	Biosequence Type	Clone ID	DDBJ/ EMBL/ Genbank	Reporter Usage	Control Type	ID	Designation	Related Gene Symbol	Database Entry
Meta Col	Meta Row	Col	Row										
1	1	1	1	370503	cDNA clone	IMAGE 32017	R17905	Experimental	–	C1	Gene	FNTA	LocusID2339
1	1	2	1	370504	cDNA clone	IMAGE 2962831	BC005866	Experimental	–	C2	Gene	MLH1	LocusID 4292
1	1	3	1	370505	Genomic clone	Cosmid 9H11	L40416	Control	Positive	–	–	–	–
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮			⋮	⋮
4	8	24	12	380696	cDNA clone	IMAGE 5214483	BC028215	Experimental	–	C285	Gene	PTEN	LocusID 5728