UNIVERSIDADE DE CAXIAS DO SUL
Centro de Ciências Exatas e Tecnologia
Curso de Bacharelado em Ciência da Computação

RENATA DE PARIS

DESENVOLVIMENTO DE WORKFLOW CIENTÍFICO PARA BIOINFORMÁTICA

Caxias do Sul
2008

Renata De Paris

DESENVOLVIMENTO DE WORKFLOW CIENTÍFICO PARA BIOINFORMÁTICA

Trabalho de Conclusão do Curso para obtenção do Grau de Bacharel em Ciência da Computação da Universidade de Caxias do Sul.

Helena Graziottin Ribeiro

Orientadora

Caxias do Sul

2008

AGRADECIMENTOS

A Deus por iluminar-me e abençoar-me sempre.

Ao meu pai Sérgio, por me passar os ensinamentos de vida, pelo apoio incondicional durante toda a minha vida acadêmica, pelo amor e carinho a mim dedicado. A minha mãe Maria, pelo amor e carinho, pela presença incansável em todos os momentos da minha vida apoiando-me, cuidando-me e aconselhando-me para a conclusão deste trabalho.

Ao meu namorado Mateus, pela paciência, ajuda e por ter suportado todas as minhas ausências, estando sempre disposto em motivar-me com suas palavras de conforto em todos os momentos.

A minha orientadora, professora Helena, por sua dedicação, disponibilidade e credibilidade depositada em mim, motivando-me do inicio ao fim do trabalho. Obrigado pelos ensinamentos e pelas brilhantes idéias, que contribuíram para o desenvolvimento deste trabalho. Agradeço também ao professor Daniel Notari pelo auxilio prestado, fornecendo bibliografias e materiais importantes para a realização deste trabalho.

Aos professores Ana e Sérgio do Laboratório de Biotecnologia Vegetal e Microbiologia Aplicada da Universidade de Caxias do Sul (UCS), que não mediram esforços em me apoiar e auxiliar durante as pesquisas e visitas realizadas. Agradeço pela paciência que tiveram comigo e pelos ensinamentos passados, ensinamentos estes que servirão não apenas para este trabalho, mas para toda a minha vida. Sem estes conhecimentos biológicos não teria sido possível a realização deste trabalho. Obrigado a Ana por sempre responder aos meus e-mails, atender aos meus contatos e receber-me no laboratório de forma prestativa.

Um agradecimento especial a minha amiga Franciele Leite, pelo companheirismo, amizade, apoio, enfim por todos os momentos que passamos juntas ao longo desta vida acadêmica.

A todos os amigos que fiz durante a minha caminhada de graduação.

Enfim, a todos amigos e familiares que contribuíram direta ou indiretamente na realização deste trabalho.

EPÍGRAFE

É melhor tentar e falhar, que preocupar-se
e ver a vida passar; é melhor tentar, ainda
que em vão, que sentar-se fazendo nada
até o final. Eu prefiro na chuva caminhar,
que em dias tristes em casa me esconder.
Prefiro ser feliz, embora louco, que em
conformidade viver... “

Martin Luther King

RESUMO

Workflows Científicos, além de utilizar serviços web garantindo a integração entre bancos de dados biológicos, oferecem uma infra-estrutura para automatização dos processos que permite projetar, reusar, anotar, validar, compartilhar e documentar experimentos científicos para a bioinformática com o propósito de facilitar o trabalho dos biólogos. O objetivo principal deste trabalho é utilizar a ferramenta *Taverna* Workbench para criar um workflow científico. Para tanto, realizou-se uma análise do funcionamento existente no laboratório de biotecnologia e modelou-se um workflow para a geração do arquivo em formato FASTA, contendo as seqüências genéticas, e do arquivo em formato PDB, contendo as coordenadas da imagem gráfica de uma proteína existente nos bancos de dados públicos de estruturas. Para realizar a geração destes arquivos é necessário executar a consulta dos dados e recuperar as informações biológicas, através de serviços que acessam e manipulam diferentes bancos de dados públicos. A integração destes bancos de dados, e a interoperabilidade entre as diferentes ferramentas para acesso são recursos das ferramentas de workflows científicos utilizados na bioinformática.

Palavras-Chave: Bioinformática, workflow científico, arquivo formato Fasta, proteína, banco de dados.

ABSTRACT

Scientific workflows use the internet services in order to integrate biological data base. Besides, they also offer process automation infra-structure which enables scientific experiments to be projected, reused, annotated, validated, shared, and documented aiming at easing biologists work. The main purpose of this paper is to use the tool called Taverna Workbench in order to create a scientific workflow. For such purpose, an analysis of the functioning in the biotechnology laboratory was made. A workflow was modeled in order to generate a FASTA file, which presents a genetic sequencing. Another file, PDB format, presenting the graphic image coordinates of a protein that is found in public data banks of structures was also modeled. In order to generate such files it is necessary to run data consulting and recover biological information through services that can access and manipulate different public data base. Integration of such data banks and their inter-operation among the different access tools are resources of scientific workflow tools used in bioinformatics.

Key-words: *Bioinformatics, scientific workflow, Fasta file, protein, data bank.*

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	Interfaces Programadas para Aplicações
BD	Banco de Dados
BLAST	*Basic Local Alignment Search Tool*
BLASTP	Alinhamento das seqüências de proteínas
BLASTX	Alinhamento das seqüências de nucleotídeos
DDBJ	*DNA Data Bank Japan*
DNA	Ácido Desoxirribonucléicos
EBI	*European Bioinformatics Institute*
EMBL	Laboratório de Biologia Molecular
GB	*GenBank*
GSDB	*Gene Sequence Database*
LSID	*Life Science Identifiers*
NCBI	*National Center for Biotechnology Information*
NIH	Institutos Nacionais de Saúde dos Estados Unidos
MoML	*Modeling Markup Language*
PDB	*Protein Data Bank*
PDBJ	*Protein Data Bank Japan*
PRF	*Protein Research Foundation*

PIR	*Protein Information Resource*
PSI-BLAST	Alinhamento de seqüências protéicas distantes
RCSB	*Research Collaboratory for Structural Bioinformatics*
RNA	Desoxirribose
SCUFL	*Simple Conceptual Unified Flow Language*
SGBD	Gerenciamento de Banco de Dados
SOAP	*Simple Object Access Protocol*
SP	*Swiss-Prot*
TBLASTN	Tradução de nucleotídeos através da seqüência de proteínas
TBLASTX	Tradução de nucleotídeos através da seqüência de nucleotídeos
UCS	Universidade de Caxias do Sul
URL	Localizador de Recursos Universal
XML	*Extensible Markup Language*
WFMS	Sistemas Gerenciadores de Workflow
WSDL	*Web Services Description Language*

SUMÁRIO

1 INTRODUÇÃO

Atualmente a tecnologia está avançando a passos largos em todas as áreas, necessitando para isso ferramentas eficientes e eficazes para um correto funcionamento de suas descobertas. A quantidade de processos informatizados aumenta em todos os segmentos, a utilização do computador torna-se indispensável para agilizar a execução destes processos. Dentre as inúmeras áreas existentes, este trabalho irá abordar o estudo de uma ferramenta computacional que auxilia na área da biologia, com foco principal na utilização de workflows científicos para apoiar atividades em bioinformática.

Segundo GIBAS (2001), bioinformática é a aplicação da tecnologia de informação ao gerenciamento de dados biológicos. Os especialistas em bioinformática (ou bioinformatas) são os criadores das ferramentas, e é fundamental que eles entendam os problemas biológicos tanto quanto as soluções computacionais para que produzam ferramentas úteis. Os algoritmos de bioinformática precisam incluir proposições científicas complexas para desenvolver ferramentas e modelar os dados de maneira exclusiva.

Os estudos nesta área fundida entre a computação e a biologia necessitam de biólogos, bioinformatas e cientistas da computação a fim de desenvolver os processos de maneira construtiva, pois de nada adianta o trabalho do biólogo, sem a intervenção do cientista, nem o trabalho do cientista sem a intervenção do biólogo, bem como, é necessário a visão do bioinformata para identificar necessidades imperceptíveis aos olhos do biólogo e do cientista. A união destes três especialistas é extremamente importante para desenvolver tarefas tais como: análise e interpretação de vários tipos de dados biológicos, criação de algoritmos eficientes, implementação e pesquisa de novas ferramentas para aperfeiçoar ainda mais os processos na área da biologia.

Atividades em bioinformática estão crescendo por todo o mundo, acompanhadas por uma proliferação de dados e ferramentas. Isto traz novos

desafios, por exemplo, como entender e organizar esses recursos, como compartilhar e re-usar experimentos bem sucedidos (ferramentas e dados), e como prover interoperabilidade entre dados e ferramentas de diferentes locais e utilizados por usuários com perfis distintos (DIGIAMPIETRI, 2007).

Atualmente os recursos disponíveis para a pesquisa de dados genéticos estão bastante dispersos, e a consulta muitas vezes apresenta informações desnecessárias que dificultam o trabalho do biólogo. Isso ocorre, porque existe um grande número de banco de dados genéticos publicados, onde cada um oferece muitas informações distintas e importantes, mas que implementam padrões específicos e diferenciados para o armazenamento destes dados. Ainda existem alguns casos em que os dados armazenados não possuem padronização, são apenas gravados e disponibilizados para acesso. A ocorrência de um ou de todos estes fatores causa um grande impacto para as pesquisas tanto na área da biologia, que carece de agilidade e dados confiáveis, bem como na área da computação, que encontra barreiras e busca soluções para realizar a consulta e integração das informações.

Para solucionar o problema de integrar diferentes bancos de dados, em diferentes ambientes de execução, existem ferramentas já disponíveis e outras que estão sendo criadas fornecendo suporte aos workflows científicos para bioinformática. Eles estão sendo cada vez mais utilizados como meios de especificar e coordenar a execução de experimentos que envolvem participantes em locais distintos, pois permitem a representação e execução de tarefas que usam dados e ferramentas heterogêneos. Como a bioinformática ainda é uma área nova, não há um consenso bem definido sobre como as tarefas devem ser executadas e como os resultados devem ser anotados, mas pode-se afirmar que uma das principais características destas ferramentas é o alto grau de intervenção humana na sua execução (DIGIAMPIETRI, 2007).

O projeto de workflows científicos em bioinformática é tipicamente manual e realizado por meio da especificação e composição de atividades, que podem ser definidas de várias formas, por exemplo usando linguagens *script* ou, mais recentemente, invocando serviços w*eb*. A ampla aceitação de tais serviços neste domínio depende da disposição da comunidade de biólogos em propor uma padronização para eles. Isto inclui a construção de ontologias, o desenvolvimento de

repositórios de ferramentas específicos para biologia e a definição das interfaces para serviços comuns (DIGIAMPIETRI, 2007).

O objetivo deste trabalho é o desenvolvimento de um workflow científico para a bioinformática, auxiliando na coordenação das tarefas e integração de informações necessárias para o estudo das proteínas. A atividade a ser modelada é a recuperação de dados e a geração de um workflow científico para a bioinformática, onde seu estudo foi realizado juntamente com o Laboratório de Biotecnologia Vegetal e Microbiologia Aplicada da Universidade de Caxias do Sul (UCS). Com base neste objetivo a proposta deste trabalho é utilizar o *software Taverna Workbench* para a criação deste workflow científico, que realize a conexão com os bancos de dados de proteínas disponíveis no portal *National Center for Biotechnology Information (*NCBI) e apresente como resultados:

- Lista de substâncias encontradas;
- Geração de arquivos em formato Fasta, apresentando a seqüência genética de proteínas;
- Geração de arquivos em formato *Protein Data Bank* (PDB), apresentando a imagem gráfica de determinada proteína.

Para explicar de forma detalhada como foi realizada as etapas envolvidas na criação deste workflow, o presente trabalho dividiu-se em 6 capítulos distintos.

O Capítulo 2 contém uma introdução sucinta de contextos biológicos e computacionais, necessária para a compreensão da proposta. Essa introdução apresenta uma visão geral sobre os conceitos biológicos de grande relevância para este trabalho, os diferentes bancos de dados públicos e seus meios de acessos para a realização de buscas e manipulação dos registros. O capítulo mostra também a importância dos arquivos em formato Fasta e PDB, detalhando suas estruturas e explicando a utilização através das ferramentas para bioinformática.

O Capítulo 3 define o conceito de workflow descrevendo as diferenças entre workflows científicos e workflows de negócios, direcionando o detalhamento em workflows científicos utilizados na bioinformática. O capítulo apresenta as características e funcionamento de duas ferramentas específicas para a criação e execução de workflows científicos, o *Taverna Workbench* e o *Kepler*, com a finalidade de justificar o motivo pelo qual o *Taverna Workbench* foi escolhido para o desenvolvimento da proposta deste trabalho.

O Capítulo 4 inicia o processo de desenvolvimento do workflow proposto. Em um primeiro momento é construído o cenário em que o laboratório encontra-se, em seguida são identificados os problemas relevantes a esta situação, permitindo descrever os requisitos necessários para desenvolver a modelagem do processo. Esse capítulo utiliza padrões da engenharia de *softwares* e da UML, com o objetivo de realizar uma análise embasada em conceitos técnicos e aplicada aos diagramas da UML.

O Capítulo 5 apresenta o workflow proposto, utilizando os ambientes de trabalho definidos no Capítulo 4 através da arquitetura dos diagramas UML. Além de exemplificar a utilização da ferramenta, o capítulo relata a validação através de exemplos, e a avaliação dos especialistas para aplicação no laboratório de biotecnologia.

Por fim, o Capítulo 6 resume o trabalho realizado, apresentando os resultados e contribuições para trabalhos futuros.

2 INTRODUÇÃO A BIOINFORMÁTICA

Neste capítulo será introduzido quais são os principais processos e as principais áreas de influência para o estudo da bioinformática, iniciando por um visão geral dos conceitos biológicos, em seguida estudando as diferentes estruturas de bancos de dados, e por fim apresentando as ferramentas de acesso e manipulação das informações biológicas no mundo atual. Desta forma, será mais fácil conduzir os próximos capítulos ao foco principal deste trabalho.

2.1 Conceitos de Biologia Molecular

Para estudar e compreender melhor os conceitos da biologia molecular e celular são utilizadas técnicas de aprendizado, auxiliando no entendimento dos assuntos a serem abordados. Atualmente existem muitas formas de criar essas técnicas, uma delas é utilizarmos a idéia principal, que define resumidamente o foco central dos estudos para um melhor entendimento. Foi nesta mesma linha de pensamento que os biólogos introduziram o Dogma Central e Periférico da Biologia Molecular estabelecendo que: O DNA atua como um modelo para se replicar, ele também é transcrito no RNA, e o RNA é convertido em proteína (GIBAS, 2001).

Para definir a Biologia Molecular de uma forma clara e precisa, será explicada cada uma das partes que compõem o Dogma Central. Estrutura, funcionamento e funcionalidades serão apresentadas nesta seção, a fim de familiarizar o leitor sobre os conceitos de proteínas, estrutura do DNA, funcionalidade dos genes para os seres vivos, e os processos de transição e tradução.

Com exceção dos vírus, todos os demais seres têm a sua estrutura baseada na célula. Muitos são apenas unicelulares; e outros são pluricelulares. Mas, a despeito de certas diferenças, a arquitetura fundamental da célula se repete em

todos os níveis de organização tornando-a, indiscutivelmente, a unidade da vida (SOARES, 1991).

Existem as células procarióticas que são desprovidas de núcleo e as células eucarióticas que possuem núcleo individualizado. As células eucarióticas em sua grande totalidade possuem funcionalidade e estrutura semelhantes, porém este grupo é representado pelas células animais e as células vegetais que possuem algumas diferenças notáveis, uma vez que a presença ou ausência de organismos em sua composição interfere diretamente no funcionamento celular. Por exemplo, as células vegetais além de possuir formato definido geograficamente, contêm a parede celular para proteção da célula, e os cloroplastos responsáveis pela fotossíntese da célula. Já as células animais possuem um formato arredondado e são desprovidas dos componentes citados pelas células vegetais (LOPES, 1997).

Tanto as células animais quanto as células vegetais possuem três componentes principais em sua estrutura: núcleo, membrana plasmática e citoplasma. O núcleo é responsável por coordenar e comandar todas as funções celulares. A membrana plasmática regula a entrada e saída de substâncias, levando em consideração a permeabilidade seletiva. Já o citoplasma através da sua composição realiza diversas funções importantes que garantem o correto funcionamento celular. Dentre as organelas citoplasmáticas[1] conhecidas podemos citar: retículo endoplasmático, complexo de golgi, centríolos, mitocôndria, lisossomos, cloroplastos e vacúolos. Como principais funcionalidades as organelas realizam a produção e sintetização de proteínas, o transporte intracelular e a respiração celular gerando energia necessária para a célula (SOARES, 1991). A Figura 2.1 representa a estrutura de uma célula vegetal e uma célula animal.

[1] As Organelas citoplasmáticas estão posicionadas entre a membrana plasmática e o núcleo. Delimitadas por uma membrana própria e com conteúdo aquoso, é usado para descrever várias estruturas com funções especializadas.

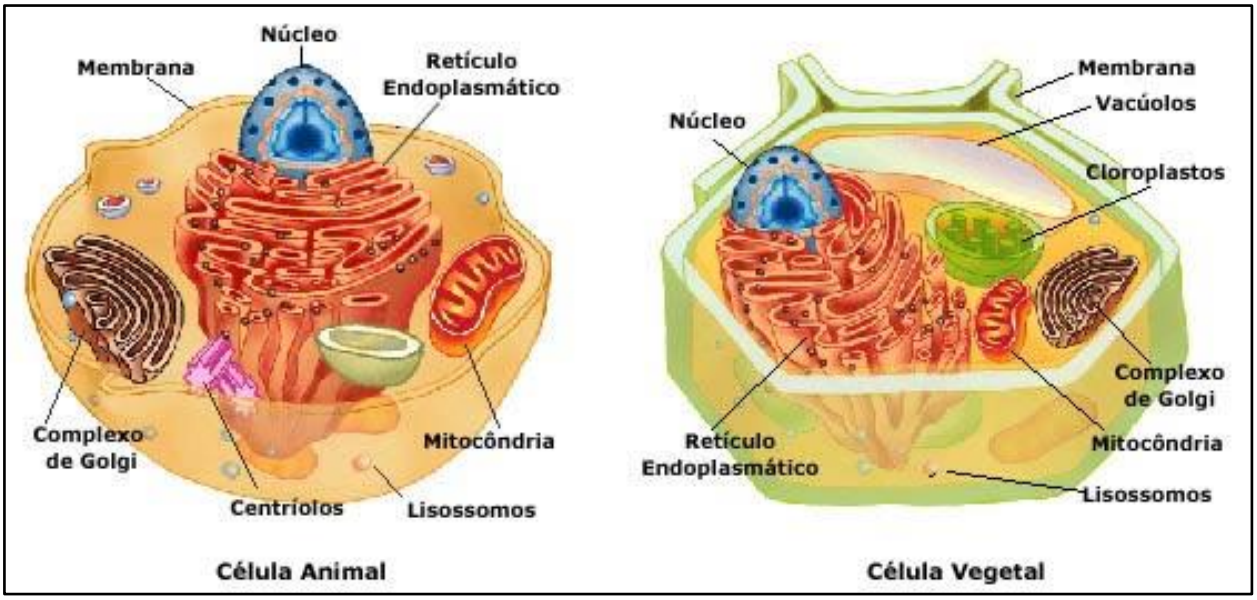

Figura 2.1: Células animal e vegetal.

As células são constituídas pelas seguintes substâncias: água (em torno de 70%), sais minerais (pequenas quantidades), carboidratos (açúcares), lipídios, vitaminas, protídeos (proteínas) e ácidos nucléicos. Esses componentes predominam em quantidade e variedade nas células, porém para esse trabalho será necessário entendermos um pouco mais sobre as duas últimas substâncias (SOARES, 1991).

Segundo LESK (2008), as proteínas evoluíram por meio de alterações estruturais, originadas por mutações nas seqüências de aminoácidos e rearranjos gênicos, que integram diferentes combinações de subunidades estruturais.

As proteínas desempenham funções muito importantes na célula, pois são responsáveis pelo crescimento, conservação, reconstrução, reprodução e funcionamento dos organismos, além disso, estão presentes em praticamente todas as partes de todas as células, e possuem um elevado peso molecular devido a sua composição polimerizada[2] de centenas de aminoácidos. Existem cerca de 20 tipos de aminoácidos que, ao serem interligados podem chegar a formar 43.000 estruturas protéticas diferentes (LESK, 2008). Devido a sua estruturação e classificação as proteínas possuem diferentes formatos, a estrutura primária é uma seqüência de aminoácidos em cadeia, já a estrutura secundária utiliza os arranjos alfa-hélice e folha-beta, na estrutura terciária ocorre o agrupamento e a interação

[2] Junção de duas ou mais moléculas de proteínas. É utilizado essa expressão para caracterizar a estrutura das proteínas, pois elas apresentam um formato de polímeros longos contendo milhares de átomos.

das hélices e folhas e a estrutura quaternária ocorre quando há o arranjo espacial resultante da união de duas ou mais estruturas vistas anteriormente. A Proteína Ribossomal, representada na Figura 2.2, é uma estrutura terciária com a interação dos arranjos alfa-hélice e folhas-beta.

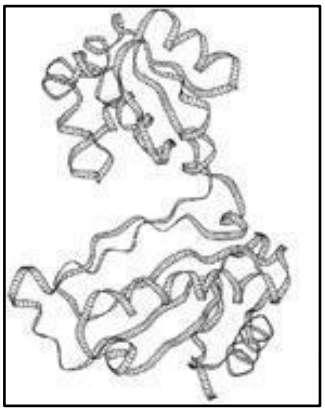

Figura 2.2: Proteína Ribossomal L1 de *methanococcus jannaschii* (LESK, 2008).

Dentre as diferentes proteínas que desempenham importantes papéis nos processos vitais, existem as enzimas que são de grande importância para a vida dos seres vivos, pois desempenham o papel de catalisar e estimular reações químicas através da energia de ativação, controlando a velocidade das reações evitando a elevação da temperatura. Outro conjunto de proteínas muito importante são os anticorpos, que atuam como substâncias estranhas para o organismo. Essas substâncias criam defesa ao organismo, pois estimula o organismo produzir anticorpos capazes de combater essas substâncias estranhas.

2.1.1 Síntese de Proteínas

Os seres vivos são formados por estruturas moleculares diferenciadas, ou seja, cada organismo vivo na terra possui características próprias que o identifica e o diferencia das demais espécies. Cada espécie carrega consigo um genótipo único herdado de seus ascendentes e descendentes no momento da sua criação, e um fenótipo caracterizando os organismos que sofrem a influência do seu ambiente e das limitações do genótipo. A hereditariedade dos seres vivos ocorre nos ácidos nucléicos através da síntese de proteínas.

Os ácidos nucléicos são substâncias orgânicas bastante complexas que se apresentam nas células com duas importantes funções: coordenar a síntese de

todas as proteínas celulares e transmitir as informações genéticas de ascendentes para descendentes, em todas as categorias de seres vivos. As unidades estruturais de um ácido nucléico são as mesmas tanto numa bactéria quanto num mamífero (SOARES, 1991).

Todos os ácidos nucléicos são formados por macromoléculas[3] de nucleotídeos. Os nucleotídeos são os componentes mais importantes do controle celular, pois possuem os genes responsáveis pela genética nos seres vivos representados por dois conjuntos: os ácidos ribonucléicos (RNA), e os ácidos desoxirribonucléicos (DNA). Segundo LOPES (1997), sua composição compreende-se por:

- Pentose, um açúcar formado por cinco carbonos. Utiliza a ribose para a composição química do RNA, e a desoxirribose para a constituição da DNA;
- Uma base nitrogenada, responsáveis por construírem as moléculas de DNA e RNA, e;
- Um ácido fosfórico, são ácidos que ligam as bases nitrogenadas.

Existem cinco tipos de bases que constituem as bases nitrogenadas: Adenina (A), Guanina (G), Timina (T), Citosina (C) e Uracila (U). As bases nitrogenadas A, C e G são encontradas tanto no RNA como no DNA, já T encontra-se apenas no DNA e U apenas no RNA (SOARES, 1991).

Cada molécula de DNA é formada por uma seqüência de genes, responsável por definir as características hereditárias dos seres vivos. Quando esses genes estão em ação eles são transcritos na molécula de RNA, sintetizando um único gene para uma determinada proteína. Cada gene tem uma seqüência de nucleotídeos que não se repete em nenhum outro gene, pois lhe é inteiramente pessoal. A “personalidade” de cada gene depende, pois, da seqüência dos nucleotídeos que formam o seu DNA. Essa seqüência é o que se convencionou chamar de Código Genético (SOARES,1991).

Tanto o RNA como o DNA sofrem processos de sintetização para criar o código genético dos seres vivos. No DNA existe o processo de duplicação que

[3] Moléculas orgânicas de elevada massa molecular relativa. São constituídas por proteínas, carboidratos, polisacarídeos e ácidos nucléicos.

realiza a replicação do DNA. No RNA ocorrem dois processos, o de transição realizando a síntese das proteínas e o de tradução criando uma nova proteína.

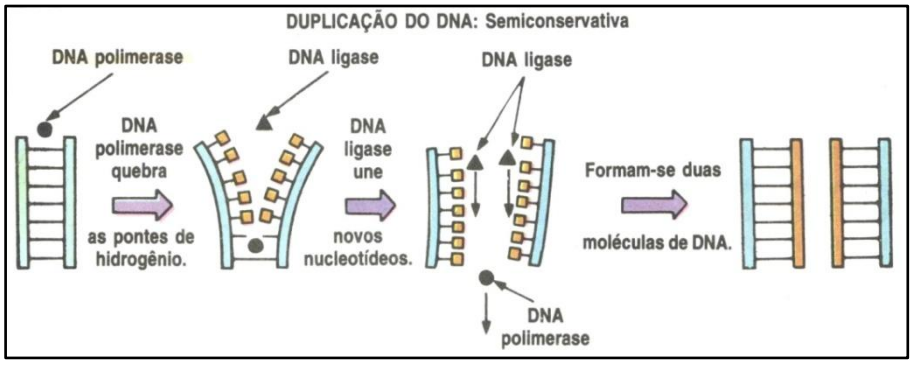

Figura 2.3: Processo de Duplicação do RNA (LOPES, 1997).

Na Figura 2.3 pode-se observar o processo de duplicação do DNA. Inicialmente é representada a estrutura de dupla hélice, em seguida a quebra das pontes de hidrogênio[4] através da enzima de DNA polimerase. Neste momento são criados novos nucleotídeos contendo outra enzima, chamada de DNA ligase que realiza a formação de duas novas moléculas de DNA.

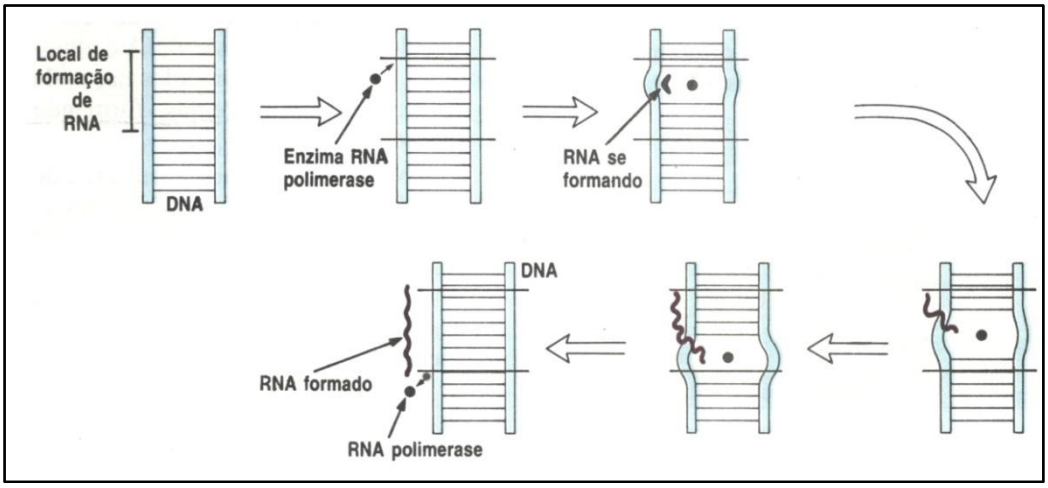

Figura 2.4: Processo de Transição do RNA (LOPES, 1997).

Na Figura 2.4 pode-se observar o processo da síntese de RNA. Partindo de uma molécula de DNA, é realizada a fragmentação de partes desse DNA através da enzima RNA polimerase, em seguida novos nucleotídeos surgem através de pareamentos[5] complementares do DNA, gerando uma molécula de RNA com

[4] Ligações de átomos de hidrogênio dentro de uma mesma molécula, sendo utilizado em Proteínas, no RNA e no DNA.

[5] Processo em que as bases nitrogenadas são trocadas para a criação de novas espécies. Realiza-se uma combinação entre bases nitrogenadas do DNA e bases nitrogenadas do RNA. Na biologia é utilizado o termo “pareia”, como por exemplo: o nucleotídeo que possui C pareia com o que possui G.

apenas uma hélice. Ao finalizar o processo de transição o RNA formado solta-se do DNA, e este por sua vez volta para sua estrutura original de dupla hélice.

Quando existe a formação de uma nova molécula de DNA ou de RNA é realizado o processo de pareamento de novos nucleotídeos, onde as bases nitrogenadas são trocadas para realizar a duplicação de uma nova molécula diferente da existente. O pareamento no DNA e no RNA é diferenciado, para o DNA segue a seguinte regra:

- Se a base é A, então pareia com a base T;
- Se a base é T, então pareia com a base A;
- Se a base é C, então pareia com a base G;
- Se a base é G, então pareia com a base C.

E para o RNA:

- Se a base é A, então pareia com a base U;
- Se a base é U, então pareia com a base A;
- Se a base é C, então pareia com a base G;
- Se a base é G, então pareia com a base C.

Na Figura 2.5 observar-se como é o funcionamento da parametrização de bases nitrogenadas na dupla hélice de um DNA, verifica-se cada lado da hélice ligando as bases através das pontes de hidrogênio.

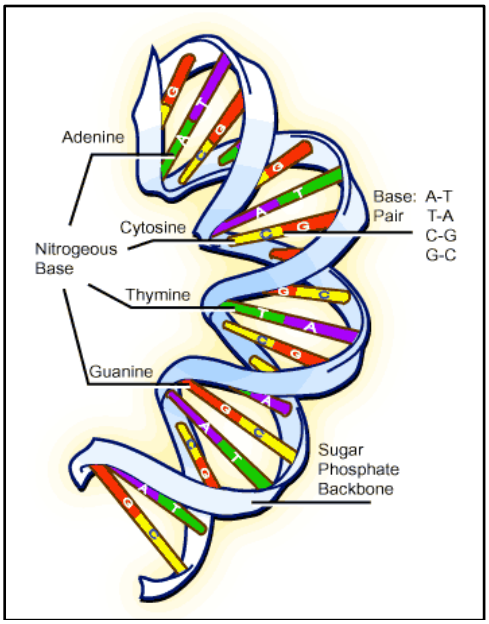

Figura 2.5: Dupla Hélice de DNA (SCIENCE CREATIVE QUARTELY, 2008).

Da mesma forma que o DNA possui o processo de replicação, o RNA também realiza essa mesma função, porém neste momento recebe o nome de tradução e origina uma nova proteína. Na maioria dos casos essa nova proteína é uma enzima responsável por catalisar reações químicas.

Segundo SOARES (1991): “a enzima catalisa reações químicas cujos produtos finais desencadeiam efeitos correspondentes à ação predeterminada pelo DNA que deu inicio a todo esse processo”. Durante muito tempo, prevaleceu a idéia de que a cada gene corresponderia sempre uma enzima. Isso caracterizou a teoria UM GENE – UMA ENZIMA. Hoje, sabemos que, em muitos casos, a proteína formada ao final da tradução do código genético já é a própria expressão do caráter genético.

A proteína realiza um papel de grande relevância para as células orgânicas e celulares de todos os seres vivos, pois é a partir dela que se descobre a seqüência do genoma diferenciado para as inúmeras espécies encontradas no nosso planeta. A Figura 2.6 representa de forma simplificada as 3 fases mais importantes para a formação do código genético, representados pela Duplicação do DNA para outro DNA, a Transição do DNA para RNA e por fim a Tradução do RNA em uma proteína.

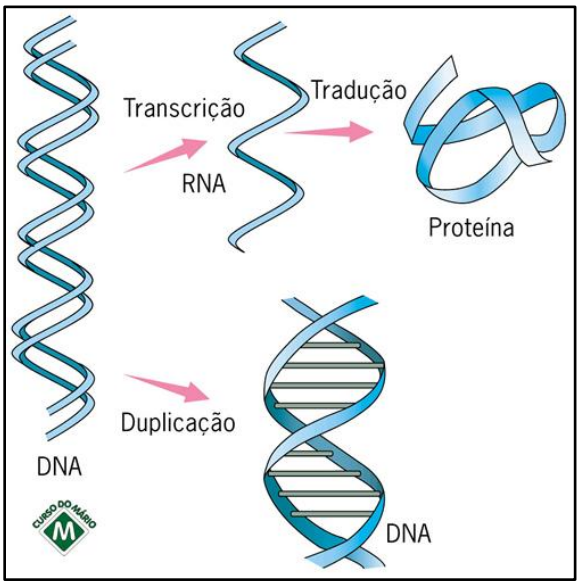

Figura 2.6: Processos realizados pelo DNA e RNA (DNA,2008).

Segundo LESK (2008): “as proteínas são as moléculas responsáveis pela maior parte da estrutura e atividade dos organismos”. Nossos cabelos, músculos, enzimas digestivas e anticorpos são todos proteínas. Tanto os ácidos nucléicos

como as proteínas são moléculas que se apresentam como cadeias longas e lineares. O código genético é de fato uma codificação: tripletos de letras sucessivas da seqüência de DNA especificam aminoácidos de proteínas. Tipicamente, proteínas são compostas de 200 a 400 aminoácidos, o que exige de 600 a 1.200 letras de mensagens de DNA expresso para especificá-las.

De forma geral todos os seres vivos possuem um genoma próprio e único que é o DNA completo de um organismo, incluindo os genes. Os genes possuem a informação para produzir todas as proteínas necessárias de todos os organismos. Estas proteínas determinam como um organismo parece, como pode se defender de infecções e doenças, e possivelmente sobre seu comportamento. Esta é o principal motivo pelo qual os cientistas estão muito interessados no seqüenciamento de genomas. Comparar DNA de outros organismos com o humano e entender a seqüência, irá ajudar os pesquisadores a aprender mais sobre evolução e a vida humana.

2.2 Bancos de Dados Biológicos

A estrutura de um banco de dados é formada quando cria-se um arquivo com informações sobre uma determinada área de conhecimento. Para a Bioinformática essas informações serão pesquisadas e armazenadas em arquivos por cientistas especializados, que realizam a coleta de dados de maneira empírica, extraindo inúmeras conclusões referentes ao estudo. Essas conclusões necessitam ser organizadas, arquivadas e publicadas em um domínio público de bancos de dados, contribuindo para a criação de novos organismos genéticos, e que serão reaproveitados por especialistas na área da biologia.

Dependendo do tipo de informação gerada (seqüências de proteínas, genes, DNA, etc.), são encontrados reservatórios específicos destinado para o armazenamento de informações categoricamente similares. Cada banco de dados possui uma classificação e uma estrutura já padronizada, para isso torna-se necessário publicar arquivos respeitando os formatos adequados de armazenamento de acordo com o banco de dados escolhido.

A maioria dos bancos de dados de biologia molecular são públicos e permite aos pesquisadores realizarem buscas nos dados, e armazenarem suas descobertas

em um desses bancos, sendo que está última deverá ser aprovada por uma série de requisitos classificatórios, pertinentes ao banco de dado público. A Tabela 2.1 apresenta uma relação dos bancos de dados (Fonte) utilizados ao longo deste trabalho, especificando à qual informação refere-se (assunto), e o endereço de acesso para a consulta virtual.

Tabela 2.1: Fontes de informações e dados biológicos

Assunto	Fonte	Link
Literatura biomédica	PubMed	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
Seqüência de ácido nucléico	Gen Bank	http://www.ncbi.nlm.nih.gov:80/entrez/query.fgi?db=Nucleotide
	SRS em EMBL/EBI	http://srs.ebi.ac.uk
	DNA Data Bank Japan (DDBJ)	http://www.ddbj.nig.ac.jp/
Seqüência de genoma	Entrez Genome	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Genome
Seqüencias de proteínas	Gen Bank	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Protein
	SWIS-PROT(SP) em ExPASy	http://www.expay.ch/spro/
	Protein Information Resource (PIR)	http://www-nbrf.georgetown.edu
Estrutura de Proteínas	Protein Data Bank (PDB)	http://www.rcsb.org/pdb/
	Protein Data Bank Japan (PDBJ)	http://www.pdbj.org/
Pesquisas sobre Proteínas Animal	Protein Research Foundation (PRF)	http://www.proteinresearch.net/
Entrez Structure DB Espectroscopia de massa de peptídeos e proteínas	PROWL	http://prowl.rockefeller.edu
Modificações pós-traducionais Informações bioquímicas e biofísicas	RESID	http://www-nbrf.georgetown.edu/pirwww/search/textresid.html
	ENZYME	http://www.expray.ch/enzyme/
	BIND	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Structure
Vias bioquímicas	PathDB	http://www.ncgr.org/software/pathdb/
	KEGG	http://www.genome.ad.jp/kegg/
2D-PAGE	SWISS-2DPAGE	http://www.expray.ch/ch2d/ch2d-top.html
Recursos na Web	Biocatálogo EBI	http://www.ebi.ac.uk/biocat/
	Arquivo IUBio	http://iubio.bio.indiana.edu

Dentre os bancos de dados mais antigos encontra-se o PDB, iniciado em 1971 em *New York*, atualmente é gerenciado pelo *Research Collaboratory for Structural Bioinformatics* (RCSB), uma organização distribuída em *New Jersey, Califórnia* e nos Estados Unidos. Segundo resultados pesquisados na data de edição deste trabalho, sua base contém em torno de 34.000 estruturas publicadas em arquivos. O PDB armazena estruturas em 3D de proteínas, ácidos nucléicos e uns poucos carboidratos, tendo como finalidade arquivar e anotar as estruturas para posterior distribuição dos conjuntos das coordenadas.

Em 1979 foi estabelecido o primeiro banco de dados (BD) de seqüências de DNA, com o nome de *Gene Sequence Database* (GSDB), atualmente é conhecido mundialmente por *GenBank* (GB). O GB disponibiliza dados de seqüências de DNA, RNA e proteínas através de três grandes institutos de pesquisa: Laboratório de Biologia Molecular Europeu (EMBL), Banco de Dados de DNA do Japão (DDBJ) e os Institutos Nacionais de Saúde dos Estados Unidos (NIH).

Devido às grandes dificuldades que os pesquisadores enfrentaram para descobrir o seqüenciamento do DNA, o GB evoluiu lentamente na sua primeira década de existência. A partir de 1995 com o lançamento do Projeto Genoma Humano, o número de publicações começou a aumentar rapidamente, impulsionando pesquisadores na descoberta de novas tecnologias para o seqüenciamento.

Nas Figuras 2.7 e 2.8 pode-se observar as estatísticas descritas acima, onde o crescimento nos repositórios de dados está aumentando constantemente nos últimos anos, e com o incentivo de novas ferramentas e tecnologias espera-se números cada vez maiores.

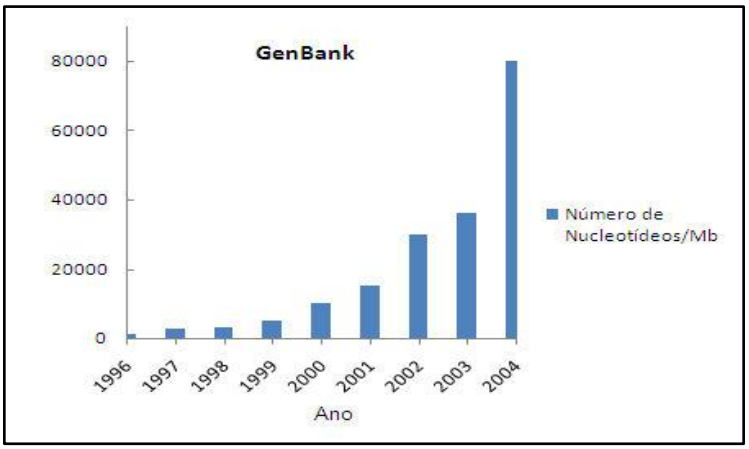

Figura 2.7: Gráfico Demonstrativo do Crescimento do GB (LESK, 2008).

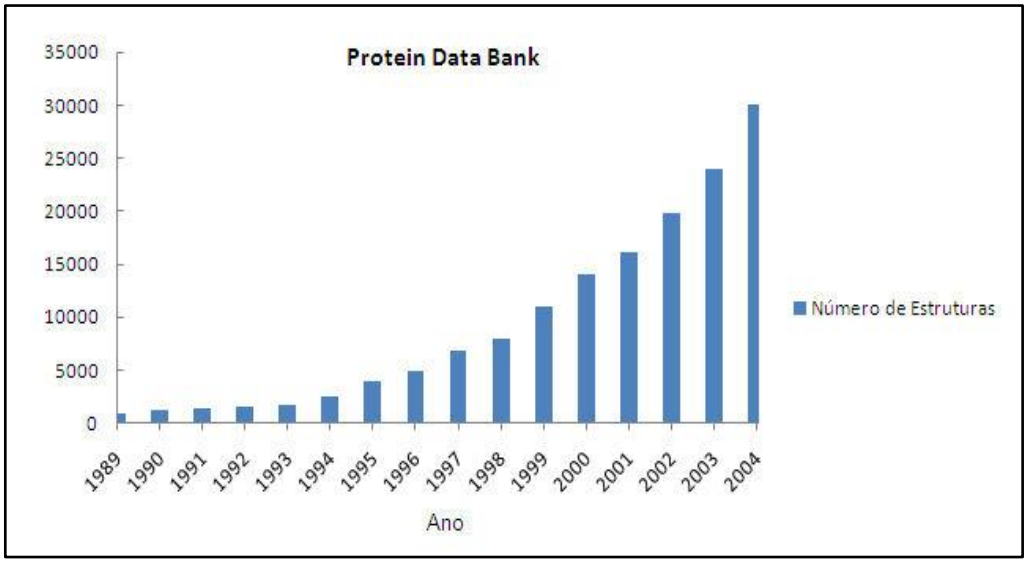

Figura 2.8: Gráfico Demonstrativo do Crescimento do PDB (LESK, 2008).

2.3 Acesso a Bancos de Dados Públicos

O grande número de bancos de dados públicos disponíveis na internet e o crescimento de estruturas genéticas publicadas sem identificação correta e muitas vezes duplicadas, conduziram a uma dificuldade em localizar genes específicos entre os especialistas. Através de portais de bancos de dados biológicos existentes é possível acessar múltiplos bancos de dados simultaneamente, realizando apenas uma pesquisa e encontrando uma variedade enorme de resultados.

Segundo BARNES (2003) as informações biológicas na Internet podem ser claramente divididas em duas amplas categorias, que são chamadas de “Internet Biológica” e “Informações biológicas na internet”. A Internet biológica tem a finalidade de construir ferramentas biológicas e os bancos de dados contendo informações biológicas detalhadas, bem como a seqüência do genoma humano, nucleotídeos e seqüência de proteínas, genótipo, polimorfismo[6] e toda a gama de informações biológicas. E as Informações biológicas na Internet, além de armazenarem dados biológicos, compreendem as informações arquivadas na web de forma informal, tais como endereços de *homepages* de laboratórios de pesquisa,

[6] São as variações de fenótipo que podem ser separada em classes distintas bem definidas.

resumos, anotações, ferramentas, bancos de dados especializados ou boutique[7] e qualquer outro dado que os cientistas considerem útil para arquivar na web.

Mesmo com toda essa gama de portais multiplicando-se constantemente para facilitar a vida dos biólogos, é importante definir as necessidades para depois determinar qual ferramenta ou banco de dados deverá ser utilizado. Neste trabalho optou-se pelo detalhamento do *National Center for Biotechnology Information* (NCBI) e o *European Bioinformatics Institute* (EBI), ambos utilizados no seqüenciamento e estruturação de proteínas.

O NCBI é o portal que possui o maior número de bancos de dados. Utilizando um dos maiores bancos de dados, o GB em cooperação com o EMBL e outras organizações internacionais, fornece as informações mais completas sobre dados de seqüencias de DNA disponíveis no mundo, possui um portal para informações de dados bibliográficos e acesso específico em pesquisas de estruturas e seqüenciamento do DNA. Atualmente mantido pela *United States National Library of Medicine, USA*, o portal utiliza o sistema ENTREZ, responsável por acessar arquivos com seqüências de DNA em diferentes divisões de bancos de dados. O ENTREZ realiza buscas em banco de dados de proteínas, peptídeo, nucleotídeo, gene, genoma, *popset*[8] e outros.

Por exemplo, ao solicitar uma pesquisa por proteínas através do NCBI, o Entrez acessa inúmeros bancos de dados que possui em seus arquivos o registro da proteína solicitada, dentre os principais bancos de dados estão: SwissProt, PIR, PRF, PDB, traduções de regiões codificantes e seqüências do GB. A partir desta definição criou-se uma estrutura genérica para representar os portais da web em bioinformática na Figura 2.9, apresentando os bancos de dados ligados nas diferentes ferramentas de acesso, que realizam a distribuição dos dados através dos Portais da Web, permitindo a visualização e extração destes dados.

[7] São bancos de dados que possuem em seu conteúdo informações diversificadas para as seqüências genéticas. Um exemplo de banco com essa característica é o *Protein Kinase Resource*, seus arquivos inclui seqüências, estruturas, informações funcionais, procedimentos laboratoriais, lista de cientistas interessados, ferramentas para análise, quadro de avisos e conexões.

[8] São informações sobre populações.

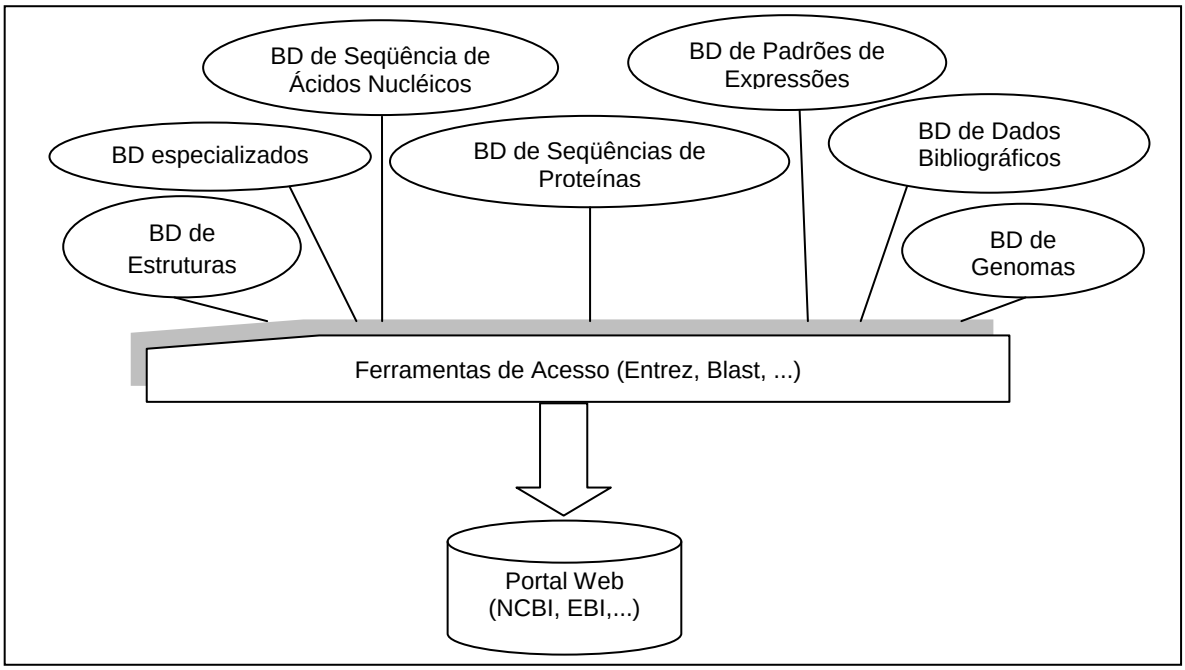

Figura 2.9: Acesso aos Bancos de Dados através dos Portais na Web.

Utilizado especificamente para acesso à bancos de dados, o NCBI além de armazenar seqüências genéticas, também possui um banco de dados de literaturas, disponibilizando serviços para consultas bibliográficas através do PubMed. Outra importante ferramenta disponível no portal é o *Basic Local Alignment Search Tool* (BLAST), um algoritmo para comparar seqüências genéticas que será estudado mais adiante, na seção sobre Ferramentas para Bioinformática.

O EBI coordenado pelo Instituto de Bioinformática da Europa, Inglaterra, através de pesquisa em Bioinformática desenvolve banco de dados biológicos e programas. Ao contrário do NCBI, o portal não possui uma padronização entre os registros contidos nos diferentes bancos de dados, pois o grande número de informação que o portal possui não utiliza ferramenta para acesso dos dados, apresentando as informações conforme a estrutura do banco de dados de origem.

De forma geral o EBI centraliza seu armazenamento com informações sobre proteínas e genoma humano, dentre os principais bancos de dados estão: EMBL, UniProt[9], Ensembl[10], InterPro[11] e PDB.

[9] Repositório central de seqüências e funções de proteínas criado pela junção das informações contidas no UniProtKB/Swiss-Prot, UniProtKB/TrEMBL e PIR.

[10] Fonte universal de informação sobre o genoma humano.

[11] Banco de Dados de proteínas que abrange famílias, domínios, repetições e regiões com características de proteínas conhecidas que podem ser aplicadas a novas seqüências de proteínas.

Na bioinformática existem sérios problemas para a integração dos bancos de dados existentes. Um dos principais motivos é a não existência na área de um padrão de armazenagem dos dados. Dessa forma, cada banco armazena os dados da forma que mais lhe convém, considerando que muitos desses bancos não foram projetados e construídos por profissionais de informática, mas sim por biólogos.

Como pode-se observar nos exemplos citados acima, alguns portais ainda enfrentam o problema da heterogeneidade, e outros estão buscando ferramentas para uma intercomunicação mais eficaz, aperfeiçoando a interoperabilidade para torná-la imperceptível ao usuário, e eliminando o limites de conexão para o BD's. Segundo LESK (2008) não está muito distante a existência de um único banco de dados de biologia molecular com inúmeras vias de acesso. Os cientistas poderão configurar o seu próprio acesso, selecionando partes da informação, criando "bancos de dados virtuais", adaptados às suas próprias necessidades.

2.4 Formato de Banco de Dados para Bioinformática

Até este momento estudou-se sobre o armazenamento e o acesso a diferentes seqüências genéticas. Nesta seção serão apresentados os formatos de recuperação de arquivos, bem como a sua utilização a partir de programas utilizados para analisar, integrar e retornar resultados importantes para a área de pesquisas biológicas.

A recuperação de seqüências genéticas nos bancos de dados é uma ferramenta disponibilizada pelos portais de banco de dados biológicos. Alguns disponibilizam mais de uma possibilidade de formato, outros apenas uma opção, e existem os portais que permitem o *dowloand* de todo o banco de dados para utilizá-lo em consultas locais. Devido ao aumento diário e constante do número de registros inseridos nos bancos de dados públicos, a utilização de bancos de dados locais não é uma opção aconselhada, uma vez que o armazenamento conduz à perda de informações atualizadas diariamente nos portais da *web*.

Alguns dos formatos mais utilizados pelos portais de bancos de dados biológicos são: Texto (txt), Fasta, XML, PDB, ANS.1[12], Gráfico. Neste trabalho

[12] *Abstract Syntax Notation One*, é uma especificação de dados genéricos, projetada para promover a interoperabilidade de bancos de dados, que é utilizada atualmente para armazenamento e recuperação de todos os tipos de dados no NCBI.

optou-se por utilizar o formato de arquivo PDB por possuir uma estrutura diferenciada para visualização gráfica de proteínas, e o formato de arquivo Fasta, devido a sua compatibilidade com muitos programas de análise de seqüências dentro da área de bioinformática.

O formato de arquivo PDB armazena informações sobre a proteína e as medidas que determinam sua estrutura geograficamente. Normalmente esses arquivos são interpretados por ferramentas específicas que realizam a leitura do conteúdo do arquivo e apresentam uma imagem gráfica da proteína, de acordo com as coordenadas existentes no arquivo.

Estes arquivos estão armazenados apenas no PDB, sua identificação é única e realizada através de um código PDB (PDB ID). O identificador possui uma codificação padrão, sendo composto por quatro caracteres, o primeiro caracter é sempre um número de 1 a 9 e os outros três caracteres por caracteres alfanuméricos. Um exemplo do arquivo PDB é apresentado na Figura 2.10, onde pode-se visualizar um fragmento do arquivo contendo as coordenadas da estrutura tioredoxina de *E.coli* , identificada pelo ID “2TRX”.

```
REMARK 525 SOLVENT
REMARK 525 THE FOLLOWING SOLVENT MOLECULES LIE FARTHER THAN EXPECTED
REMARK 525 FROM THE PROTEIN OR NUCLEIC ACID MOLECULE AND MAY BE
REMARK 525 ASSOCIATED WITH A SYMMETRY RELATED MOLECULE (M=MODEL
REMARK 525 NUMBER; RES=RESIDUE NAME; C=CHAIN IDENTIFIER; SSEQ=SEQUENCE
REMARK 525 NUMBER; I=INSERTION CODE):
REMARK 525
REMARK 525  M RES CSSEQI
REMARK 525    HOH   429        DISTANCE =  6.01 ANGSTROMS
REMARK 525    HOH   507        DISTANCE =  5.98 ANGSTROMS
DBREF  2TRX A    1   108  UNP    P0AA25   THIO_ECOLI       1    108
DBREF  2TRX B    1   108  UNP    P0AA25   THIO_ECOLI       1    108
SEQRES   1 A  108  SER ASP LYS ILE ILE HIS LEU THR ASP ASP SER PHE ASP
SEQRES   2 A  108  THR ASP VAL LEU LYS ALA ASP GLY ALA ILE LEU VAL ASP
SEQRES   3 A  108  PHE TRP ALA GLU TRP CYS GLY PRO CYS LYS MET ILE ALA
SEQRES   4 A  108  PRO ILE LEU ASP GLU ILE ALA ASP GLU TYR GLN GLY LYS
SEQRES   5 A  108  LEU THR VAL ALA LYS LEU ASN ILE ASP GLN ASN PRO GLY
SEQRES   6 A  108  THR ALA PRO LYS TYR GLY ILE ARG GLY ILE PRO THR LEU
SEQRES   7 A  108  LEU LEU PHE LYS ASN GLY GLU VAL ALA ALA THR LYS VAL
SEQRES   8 A  108  GLY ALA LEU SER LYS GLY GLN LEU LYS GLU PHE LEU ASP
SEQRES   9 A  108  ALA ASN LEU ALA
SEQRES   1 B  108  SER ASP LYS ILE ILE HIS LEU THR ASP ASP SER PHE ASP
SEQRES   2 B  108  THR ASP VAL LEU LYS ALA ASP GLY ALA ILE LEU VAL ASP
SEQRES   3 B  108  PHE TRP ALA GLU TRP CYS GLY PRO CYS LYS MET ILE ALA
SEQRES   4 B  108  PRO ILE LEU ASP GLU ILE ALA ASP GLU TYR GLN GLY LYS
SEQRES   5 B  108  LEU THR VAL ALA LYS LEU ASN ILE ASP GLN ASN PRO GLY
SEQRES   6 B  108  THR ALA PRO LYS TYR GLY ILE ARG GLY ILE PRO THR LEU
SEQRES   7 B  108  LEU LEU PHE LYS ASN GLY GLU VAL ALA ALA THR LYS VAL
SEQRES   8 B  108  GLY ALA LEU SER LYS GLY GLN LEU LYS GLU PHE LEU ASP
SEQRES   9 B  108  ALA ASN LEU ALA
FTNOTE   1 RESIDUES PRO A 76 AND PRO B 76 ARE CIS PROLINES.
FTNOTE   2 RESIDUES HIS A 6, LEU A 7, ILE A 23, ASP A 47, GLU A 48,
FTNOTE   2  LEU A 58, LEU A 80, HIS B 6, ASP B 47, LEU B 58, AND LEU B
```

Figura 2.10: Arquivo no formato PDB.

O formato do arquivo Fasta diferencia-se dos demais, pois além de possuir a seqüência genética como nos demais arquivos, identifica na primeira linha do arquivo as informações associadas ao gene e ao banco de dados de origem. Dependendo da espécie pesquisada o arquivo apresenta o nome do banco de dados e os códigos de identificação diferenciados, porém a estrutura do arquivo permanece igual. Abaixo, um exemplo do arquivo no formato Fasta para a glutationa peroxidase bovina (*bovine glutathione peroxidase*):

```
>gi|121664|sp|P00435.2|GPX1_BOVIN Glutathione peroxidase
MCAAQRSAAALAAAAPRTVYAFSARPLAGGEPFNLSSLRGKVLLIENVASLCGTTVRDYTQMNDLQRRLG
PRGLVVLGFPCNQFGHQENAKNEEILNCLKYVRPGGGFEPNFMLFEKCEVNGEKAHPLFAFLREVLPTPS
DDATALMTDPKFITWSPVCRNDVSWNFEKFLVGPDGVPVRRYSRRFLTIDIEPDIETLLSQGASA
```

Figura 2.11: Arquivo no Formato Fasta.

As informações do arquivo são as seguintes:

- Sinal > - obrigatório no inicio do arquivo.
- gi[13]|121664– Número do *GenInfo*, atribuído pelo NCBI para cada seqüência no seu banco de dados ENTREZ.
- SP – indica que o banco de dados fonte é o SWISS-PROT.
- P00435 – número de acesso da entrada no SWISS-PROT.
- GSHC _BOVIN– identificador da seqüência e da espécie no SWISS-PROT.
- Glutathione peroxidase – nome da molécula.
- As demais informações referem-se a seqüência da espécie.

O formato de arquivo Fasta é utilizado na maioria dos programas que realizam o pareamento de bases nitrogenadas através do DNA ou pareamento de proteínas, e encontram espécies relacionadas. Os programas para realizar o pareamento das espécies, normalmente interpretam a leitura para mais de um arquivo, realizando assim a combinação e transformação das espécies. Os arquivos podem ser salvos e lidos em um programa local, bem como pode-se interpretar através dos portais de acesso para a bioinformática, onde as informações são consultadas diretamente no banco de dados de origem, para realizar o processamento, ou seja, pareamento das seqüências de bases nitrogenadas, para no final retornar os resultados solicitados pelo usuário. O próximo capítulo apresenta

[13] O NCBI cria uma nova entrada com um novo número “gi” quando uma nova entrada de seqüência é disponibilizada, mas apenas cria um novo registro se as alterações afetarem apenas as informações que não concernem à seqüência, tal como referências em literatura.

algumas das ferramentas mais conhecidas e utilizadas pelos especialistas em biologia.

2.5 Ferramentas para Bioinformática

Antes de apresentar as ferramentas utilizadas pela bioinformática, é importante introduzir qual sua principal função e de que forma podem contribuir para a biologia.

Na biologia é realizado o alinhamento dos genes para organizar as seqüências primárias de DNA, RNA ou proteínas identificando regiões similares que possam ter conseqüência de relação funcional, estrutural ou evolucionária entre elas. Seqüências alinhadas de nucleotídeos ou resíduos de aminoácidos são representadas tipicamente como linhas de uma matriz. Colunas sucessivas de uma matriz são formadas por resíduos de caracteres semelhantes alinhados entre as diferentes linhas. Os resíduos de caracteres podem conter espaçamentos que são conhecidos por *gaps,* identificando falhas na seqüência. Um exemplo de alinhamento múltiplo[14] pode ser observado na Figura 2.12, onde para este fragmento o topo da matriz corresponde à identificação da coluna nas seqüências, a última linha da matriz apresenta o resultado dos caracteres conservados em todas as seqüências do alinhamento múltiplo, e os "-" no meio das seqüências são representados pelos *gaps.*

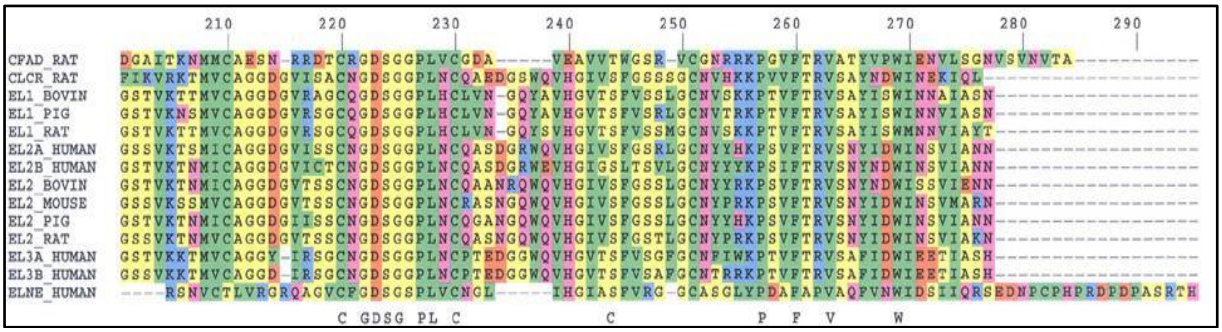

Figura 2.12: Alinhamento fragmentado do *Mammalian elasteses* (LESK, 2008).

Para realizar o alinhamento os programas utilizam diferentes processos, permitindo validar a similaridade entre os genes. Uma das formas conhecidas é o alinhamento local, onde dentro do arquivo localizam-se fragmentos de seqüências

[14] Alinhamento simultâneo de muitas seqüências em um mesmo processo.

semelhantes. A outra forma é o alinhamento global, realizando o alinhamento para toda a seqüência da fita genética. As duas formas são implementadas por algoritmos que buscam otimização e eficiência para encontrar uma solução ótima na solução do problema, as técnicas mais utilizadas são Programação Dinâmica e os programas baseados em Heurística (LESK, 2008).

Há 20 anos, quando a biologia ainda não trabalhava com a informática, realizar a similaridade entre seqüências de DNA era um trabalho muito demorado e cansativo, pois a comparação entre pares de bases nitrogenadas era uma a uma, combinando caractere a caractere até encontrar a combinação de um novo gene (GIBAS, 2001). Esse procedimento além de ser desgastante, necessitava de muita atenção do biólogo, para não equivocar-se e concluir resultados inválidos ou de espécies incorretas.

Em meados dos anos 80, a informática começou a desenvolver os primeiros programas que agilizaram e facilitaram a vida dos biólogos (LESK, 2008). Necessitando acompanhar o rápido crescimento de informações armazenadas nos bancos de dados biológicos, as ferramentas foram aumentando em quantidade e eficiência para integrar-se aos novos arquivos e serem utilizadas até os dias atuais como o principal recurso dos biólogos para a descoberta de novas espécies.

Muitas ferramentas amplamente disponibilizadas para a comunidade de biologia – incluindo desde alinhamento múltiplo, análise filogenética[15], identificação de motivos e software de modelagem por homologia, até serviços de pesquisa em bancos de dados na Web – contam com os algoritmos de comparação de pares de seqüências como o principal elemento da sua função (GIBAS, 2001).

As ferramentas e recursos na área da Bioinformática disponíveis para acesso remoto e local cresceram de forma descontrolada nos últimos anos, proporcionando o aperfeiçoamento e a inovação de tecnologias para obtenção de resultados mais precisos. Como o foco principal deste trabalho não é estudar o funcionamento destas ferramentas, mas introduzir de maneira geral suas funcionalidades, formas de acesso e diferentes possibilidades de utilização, será apresentado três modelos com estas características bastante diferenciadas. Optou-se por apresentar a ferramenta BLAST por possuir uma estrutura de consultas e acessos

[15] Estudo entre espécies que possuem antepassados em comum.

exclusivamente *on line* através do NCBI, e o ClustalX que é um *software* de utilização e processamento local.

O BLAST é um conjunto de ferramentas utilizadas para encontrar o melhor alinhamento de proteínas e DNA. O NCBI é o portal responsável pela manutenção e desenvolvimento destas ferramentas, compreendendo-se atualmente em cinco tipos de programas BLASTP, TBLASTN, PSI-BLAST contendo seqüências de aminoácidos, o BLASTX e o TBLASTX contem seqüências de nucleotídeos. Todos os programas realizam consultas em bancos de dados de proteínas e de nucleotídeos, porém buscam alinhar os genes de diferentes formas de acordo com a similaridade das seqüências. É possível detectar alinhamentos locais e globais, realizando comparações através de fragmentos de um ou mais arquivos de forma a encontrar similaridade entre essas regiões, bem como comparando toda a informação do arquivo com outros arquivos para então descobrir uma espécie.

O BLAST é uma ferramenta simples e eficiente, porém o tempo para processar a comparação dos pares de segmento é bastante lento. O algoritmo processa os dados utilizando a heurística comparativa por pares de segmentos, realizando testes exaustivos com os dados encontrados através do acesso aos diferentes bancos de dados públicos, tendo como finalidade encontrar o menor número de *gaps* (falhas) nos fragmentos resultantes. Dependendo do limite de falhas que o usuário estabelece previamente é possível determinar a quantidade de pares investigados em uma busca, bem como estimar o tempo necessário para processar as informações. Conseqüentemente existe uma chance pequena de que extensões adequadas não sejam selecionadas, mas na prática este compromisso é altamente aceitável.

Este algoritmo determina todos os pares de segmentos entre as seqüências do banco de dados e a seqüência em exame que apresentam escore acima de um valor de limiar pré-fixado pelo usuário. Esses pares de segmentos são apresentados como saída do programa, assim como seus valores estatísticos associados (TUTORIAL BLAST, 2008). A Figura 2.13 identifica com “1” os pares similares e com “0” os pares com *gaps*, calculando o *score* total da seqüência investigada.

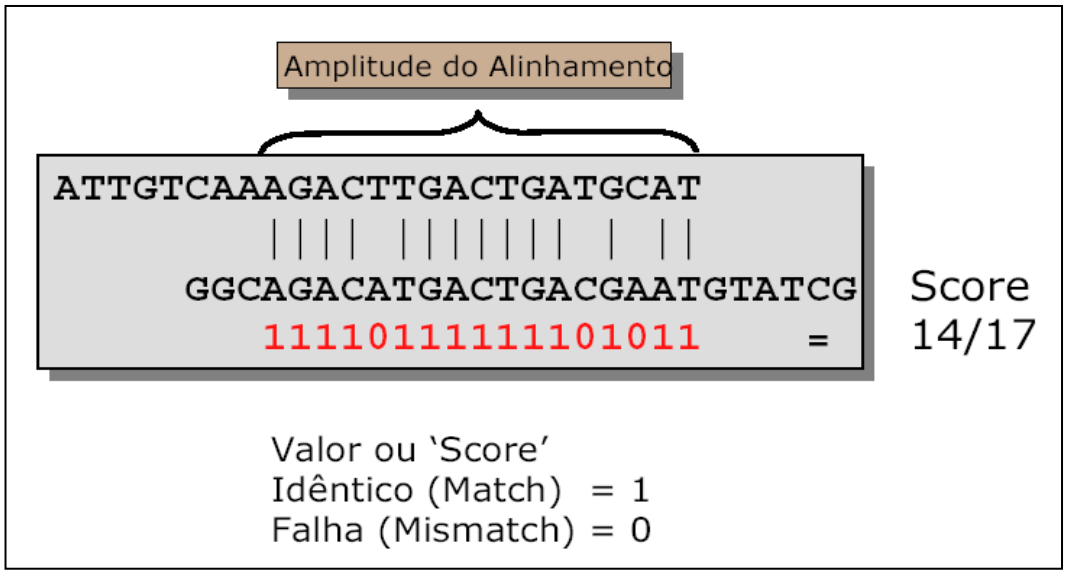

Figura 2.13: Alinhamento executado pelo BLAST (ALTSCHUL, 1990).

O ClustalX é uma ferramenta que realiza o alinhamento múltiplo de seqüências genéticas, e encontra-se disponível gratuitamente para ser instalado em todas as plataformas computacionais. Possui uma interface bastante iterativa, realizando a leitura de arquivos no formato FASTA, e gerando formatos diferentes para demonstrar resultados da análise dos alinhamentos realizados. Dentre os principais resultados apresentados estão o alinhamento de seqüências em cores e os diferentes formatos de árvores, facilitando o entendimento e a visualização dos resultados gerados pela ferramenta.

Caracterizando-se pelo rápido processamento de informações, sua plataforma realiza o alinhamento através da leitura de múltiplos arquivos no formato FASTA, selecionados pelos usuários, e utilizados como base de dados para o processamento da ferramenta. Ao contrário dos softwares *on-line*, como o BLAST, que possuem processamento mais lento por realizar constantes acessos nos bancos de dados públicos procurando por fragmentos de similaridade.

A Figura 2.14 mostra um exemplo do alinhamento em cores com os resultados obtidos após o alinhamento de várias seqüências, e na Figura 2.15 os diferentes formatos de árvores filogenéticas[16] que podem ser visualizadas.

[16] Representação em forma de árvore entre várias espécies que possuem antepassados com alguma semelhança. A representação de árvores filogenéticas compara-se a estrutura de um grafo, onde cada nodo é uma espécie que possuem descendentes interligados por arestas. As arestas ligam um nodo a outro, para as árvores filogenéticas os comprimentos das arestas significam uma medida de dissimilaridade entre duas espécies ou o tempo decorrido desde a sua separação.

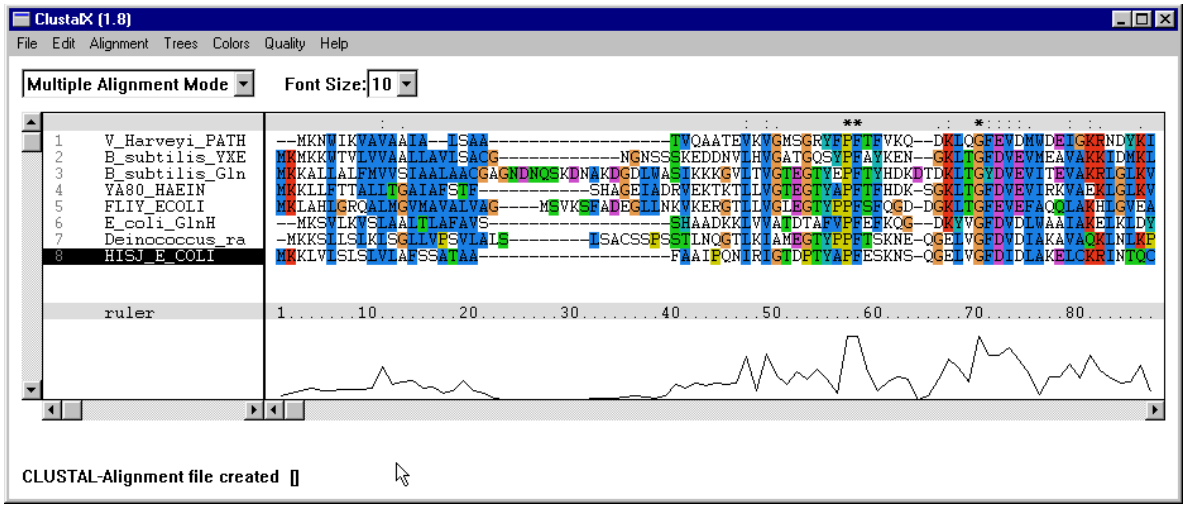

Figura 2.14: Alinhamento de seqüências pelo ClustalX (CLUSTALX, 2008).

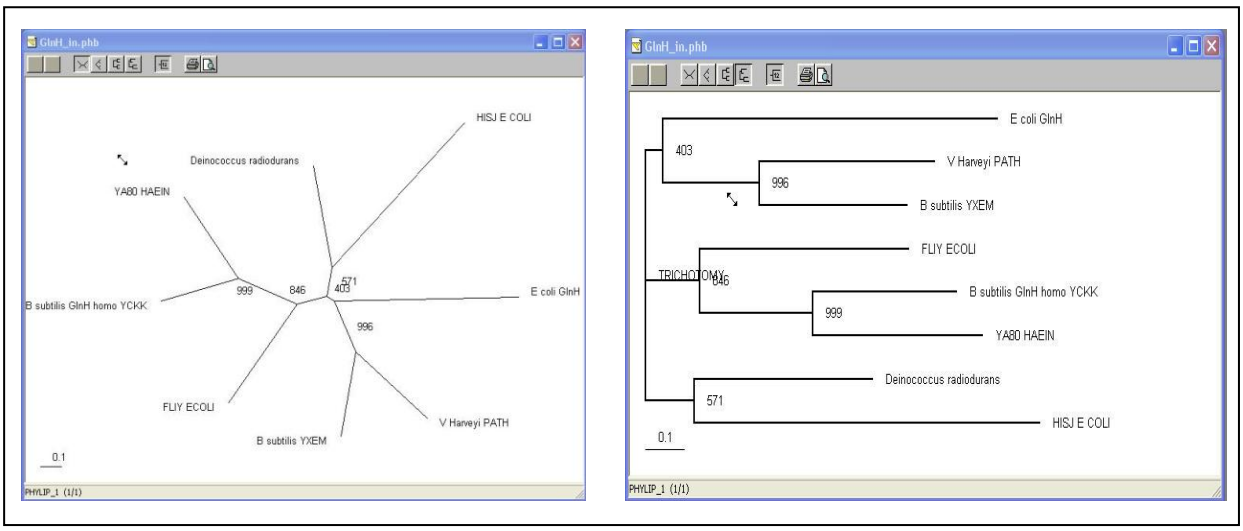

Figura 2.15: Árvores filogenéticas no ClustalX (CRUSTALX, 2008).

Baseado no funcionamento de alinhamento progressivo, esta ferramenta determina os alinhamentos para cada par de seqüências e constrói uma matriz de distâncias, onde é empregada na determinação de uma árvore filogenética guia, utilizada como peça chave para a determinação do alinhamento múltiplo. Seu código fonte está implementado na programação C, com arquitetura de Tree View.

3 WORKFLOW PARA BIOINFORMÁTICA

A automatização de processos utilizando novas tecnologias para a reorganização e aperfeiçoamento das atividades encontra-se em constante desenvolvimento principalmente nos ambientes empresariais e científicos. Devido a necessidade de não apenas documentar regras e procedimentos, mas criar ferramentas para manipular dados arquivados e executar processos existentes, é que os *softwares* de workflow surgiram, onde a realização de tarefas estruturadas proporcionou a eficiência e o aumento significativo na produtividade.

Para um melhor entendimento sobre workflow, este capítulo apresenta uma explicação sobre a estrutura de funcionamento padrão de workflow, definição sobre workflow de negócio e workflow científico e apresentação das ferramentas mais utilizadas atualmente em workflow científico.

3.1 Definição de Workflow

Um workflow diz respeito à automatização de procedimentos, onde documentos, informações ou tarefas são passadas entre os participantes de acordo com um conjunto pré-definido de regras, para se alcançar ou contribuir no objetivo global de um negócio. Apesar de um workflow permitir a organização manual, na prática a maioria dos workflows são organizados dentro de um contexto do sistema de informação para prover um apoio automatizado aos procedimentos (WfMC, 2004).

De acordo com Cruz (2000), um workflow é definido através de W(T,V,Sf,Cf), onde:

- T define um conjunto de tarefas W;
- V define um conjunto de variáveis de W definindo um fluxo de dados;
- Sf define uma função sucessora associada a cada tarefa t € T;
- Cf define uma função de condição associada a cada tarefa t € T.

A coleção de tarefas de um workflow é um conjunto de componentes de softwares independentes que implementam alguma funcionalidade (SILVA, 2007). Um workflow também define a ordem de execução dessas tarefas ou as condições em que serão executadas e sua eventual sincronização. As tarefas, em conjunto ou individualmente, podem produzir resultados parciais durante a execução do workflow. Esses resultados parciais chamados de dados de entrada e saída das tarefas (variáveis), são definidos como regras de um fluxo de informações podendo ocorrer de maneira condicional, onde existe mais de uma opção para a escolha da próxima etapa, ou apenas sucessor onde existe apenas um caminho que conduz para a próxima etapa.

Praticamente o processo de funcionamento em todos os workflows é semelhante e dividido em três importantes fases:

1- Entrada das informações (*input*);

2- Processamento destas informações;

3- Saída das informações (*output*).

A execução destas fases pode ser realizada tanto em nível de workflow onde a entrada de uma informação é o ponto de partida para a ativação de todo o processo, percorrendo cada uma das etapas de um workflow e gerando uma saída como resultado final, bem como em nível de processo específico, onde a entrada de um processo A pode depender da saída de um processo B que está antecedendo o processo A.

A necessidade de utilizar workflow surgiu, primeiramente, para reorganizar e automatizar processos em empresas, atuando tanto na área administrativa como na produtiva. A grande quantidade de papéis, regras e rotinas existentes em escritórios de forma passiva, conduziu para o desenvolvimento de uma ferramenta que além de organizar essas regras e rotinas arquivadas em diferentes lugares, também pudesse realizar a ativação desses processos (CRUZ, 2000). Ou seja, em um sistema de workflow cada tarefa contém inúmeros procedimentos, a partir desses procedimentos existem inúmeras atividades, e o conjunto destas atividades irão formar um processo que será executado de forma automatizada em um fluxo de processos mapeados seqüencialmente.

Os Workflows possuem, tradicionalmente, duas etapas conhecidas como: definição e execução (MENEZES, 2007). Na fase de definição é organizado o

funcionamento geral do Workflow, onde é definido o conjunto de tarefas, o conjunto de variáveis pertencentes ao fluxo de dados, as funções sucessoras e de condição, assim como os tipos de entrada e de saída. Já na execução é que são invocadas as regras definidas na fase anterior, produzindo resultados intermediários que serão utilizados de forma parcial ou total através de uma função sucessora ou de condição, a fim de acessar a tarefa seguinte até a obtenção do resultado final. A automatização destes processos é realizado pelas ferramentas de *software* para workflow.

A fim de oportunizar diferentes processos em um mesmo sistema de gerenciamento de atividades, onde cada qual responde por uma necessidade específica, realizando tarefas específicas e respondendo a resultados específicos dependendo da área de atuação, surgiram os diferentes modelos de workflow utilizados nas áreas administrativa, produtiva e científica.

Para padronizar e sustentar a criação dos diferentes tipos de workflows surgiram os Sistemas Gerenciadores de Workflow (WFMSs). Sua principal finalidade é viabilizar os projetos, estabelecendo e monitorando os workflows em ambientes heterogêneos e distribuídos (CITTON, 2004). Toda a estruturação de workflows deve atender as normas e exigências estabelecidas pelo WFMSs, bem como é possível criar novos padrões quando áreas específicas exigirem modificações relevantes às existentes, como é o caso dos workflows científicos, cujas suas características diferencem-se dos workflows existentes.

Workflows científicos diferem de workflows de negócio em diversos aspectos. Particularmente em bioinformática, eles caracterizam-se pelo alto grau de intervenção humana na sua execução (DIGIAMPETRI, 2007).

3.2 Workflow de Negócio

CRUZ (2000), descreve um sistema de workflow de negócio através de 3Rs, sendo formado por papéis, regras e rotas, em inglês *Roles, Roles and Routes.* Através destes componentes é possível construir workflows de negócio, produção,

ad hoc[17], orientados para objeto e baseado no conhecimento, utilizando caminhos (rotas) sequenciais, paralelos e condicionais.

Para a criação de um workflow é importante identificar quais são os papéis, regras e rotas a fim de criar uma ferramenta que realize da melhor forma possível a execução do processo. Um exemplo de workflow de negócio pode ser verificado na Figura 3.1, onde é apresentando um processo específico para organizar e automatizar a entrega de documentos internos solicitados em uma empresa. Para visualizar todos os objetos que fazem parte do workflow, foram divididos em três partes distintas:

1. Recursos Humanos e Computacionais; compreende as pessoas e os periféricos eletrônicos envolvidos.
2. Lógica do processo; representando tarefas e regras do processo.
3. Recursos de Informação; identificando os sistemas de aplicação e banco de dados.

O funcionamento do workflow é realizado da seguinte forma: Inicialmente um recurso humano solicita o documento (a), em seguida o documento passa pelo processo de avaliação (b), o passo seguinte é a notificação de entrega (c), para esse processo é utilizado uma aplicação do recurso de informação a fim de informar ao solicitando o prazo e a forma com que será entregue estes documento. A criação do documento (d) é inicializada após a notificação, utilizando os dados dos recursos de informação para encontrar as informações necessárias à serem incluídas no documento, a seqüência deste processo pode efetuar dois outros diferentes processos:

- Enviar por e-mail (e.1);
- Entregar Documento (e.2).

Esses dois diferentes processos identificam uma seqüência condicional, onde no final do processo ocorre apenas uma das duas condições.

[17] Um tipo de workflow criado para ser usado dinamicamente por grupos de trabalho cujos participantes necessitem executar procedimentos individualizados para cada documento processado dentro do fluxo de trabalho.

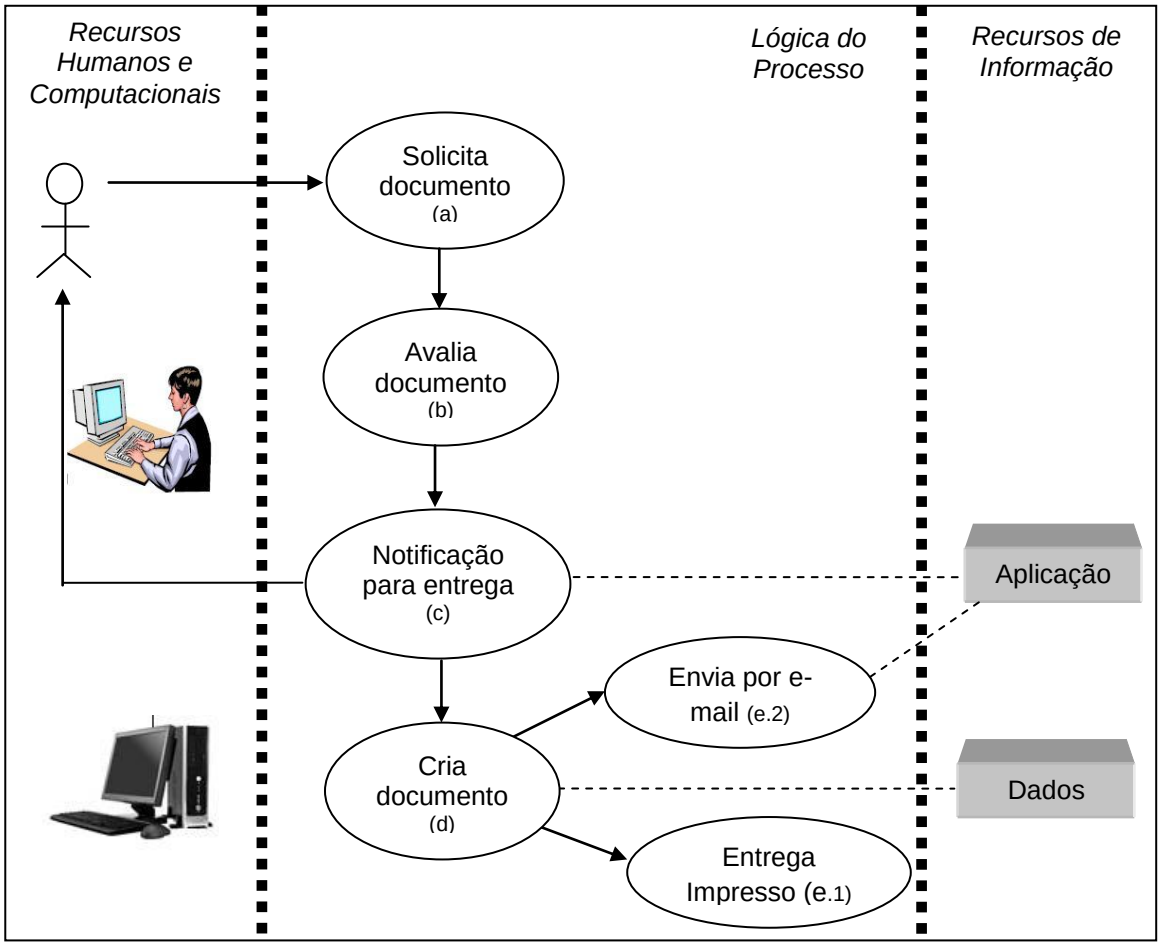

Figura 3.1: Exemplo workflow de negócio.

3.3 Workflow Científico

Workflows científicos são definidos como resolução de problemas científicos através de técnicas tradicionais de workflows. Ou seja, as idéias de execução de um conjunto de tarefas em uma determinada sequencia foram aproveitadas na área científica para a realização de experimentos e estudos. Nos workflows científicos os passos são na maioria das vezes compostos por programas computacionais que recebem, processam e geram um conjunto de dados científicos que podem ser repassados aos demais passos do workflow (SILVA, 2006).

A flexibilidade e interoperabilidade entre os workflows científicos estão aumentando o interesse nas mais diferentes áreas da *e-Science*. As aplicações em biodiversidade, química e física também estão utilizando essa nova tecnologia, porém foi na área da bioinformática que disseminaram os primeiros experimentos *in silico*, através da utilização de diversos programas. Os experimentos *in silico*

possibilitam desenvolver programas computacionais intensivos para a realização de experimentos científicos, uma vez que a execução de um determinado workflow pode não terminar ou não oferecer resultados conclusivos. Porém é importante manter registrado os experimentos com execução bem sucedida, bem como com execução defeituosas, pois o grande número de workflows para bioinformática disponíveis pode levar a criação de ferramentas já existentes, portanto antes de iniciar a criação de um novo workflow é importante pesquisar por ferramentas existentes e que permitem realizar o reuso no sentido de adaptá-las a necessidade existente.

A Figura 3.2 apresenta um exemplo de workflow científico, onde a entrada é o código identificador de uma espécie, podendo ser informada pelo usuário ou pré-determinada no workflow, em seguida são realizados acessos ao banco de dados e o processamento dos dados de saída. Ao final do processo são gerados dois tipos de saída, ou seja, um arquivo em formato normal e outro em formato gráfico obtido como resultado do procedimento software. Neste exemplo foi utilizado apenas um banco de dados e um *software*, normalmente o acesso aos bancos de dados é heterogêneo e o número de *softwares* também pode aumentar, dependendo dos resultados solicitados e dos acessos aos BD's.

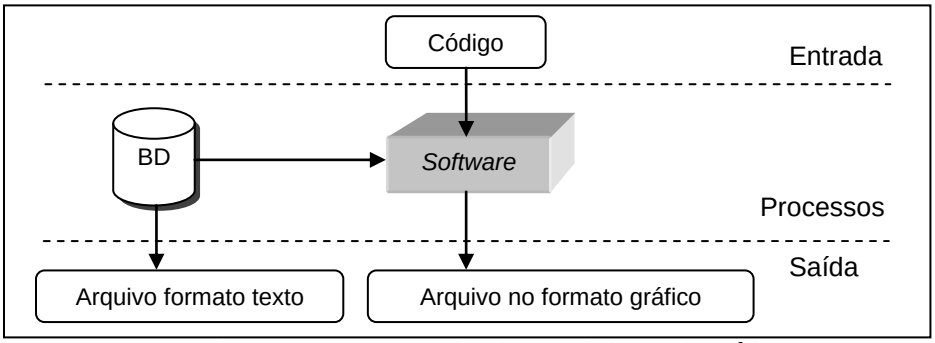

Figura 3.2: Exemplo de workflow científico.

As principais características dos workflows científicos segundo MENEZES (2007) são:

- Workflows científicos podem ter suas definições mudadas constantemente;
- Um experimento pode requerer múltiplas execuções de um mesmo workflow com parametrização diferente; gerando, conseqüentemente, resultados diferentes;

- Os resultados de workflows científicos são sempre importantes, independente de serem corretos ou não, pois os resultados refletem o andamento de um experimento;
- Há a necessidade de registrar como os dados foram gerados e manipulados pelas tarefas do workflow (proveniência de dados);
- Os programas que implementam as tarefas dos workflows podem estar em ambientes heterogêneos, dificultando a integração.

De forma geral os workflows científicos apresentam diferenças significativas em relação aos workflows de negócio. A comparação entre estas duas linhas de processos está apresentada na Tabela 3.1, onde pode-se observar como as características são representadas em cada workflow.

Tabela 3.1: Diferença entre Workflow Científico e de Negócio

Científico	Negócio
Definições alteradas constantemente.	Definições dificilmente são alteradas.
Execução em paralelo de um mesmo workflow (repetidas vezes).	Execução síncrona do workflow.
Resultado parcial e final: mesmo grau de importância.	Resultado final: mais importante.
Muita intervenção humana.	Pouca intervenção humana.
Procedência das informações e dados.	-

3.4 Ferramentas de Workflow Científicos

O avanço na tecnologia da informação vem aumentando e fornecendo ferramentas cada vez mais eficientes para a automatização de processos nas áreas empresariais e científicas. Muitas são as ferramentas de workflow disponíveis atualmente, antes de eleger um destes workflows para trabalhar é importante definir exatamente quais as regras em que as tarefas deverão ser processadas através do fluxo de informação, para que a escolha da ferramenta forneça recursos suficientes e resultados satisfatórios durante o inicio, meio e fim do processo.

Da mesma forma que workflows de negócios possuem características diferenciadas dos workflows científicos, também os *softwares* de workflow de negócios diferem-se dos *softwares* de workflow científicos, pois cada ferramenta

precisa desempenhar a função que lhe é competida de acordo com as características próprias definidas para cada workflow.

O foco principal deste trabalho é utilizar uma destas ferramentas de workflow disponíveis para o desenvolvimento de um workflow científico utilizado na área de bioinformática. Além disto, outros requisitos foram observados para a escolha, tais como confiabilidade, portabilidade, disponibilidade e isenção de custo para licenciar a utilização do *software*. As ferramentas que foram estudadas estão listadas abaixo, destas o Kepler e o Taverna serão explicados detalhadamente neste trabalho.

- Vistrails[18]; utilizado principalmente para a visualização e manipulação de resultados gráficos em workflows científicos, seu código fonte está escrito em linguagem Python.
- InServices[19]; é um sistema de gerenciamento e redefinição de workflows científicos com foco para serviços web, permite realizar o gerenciamento de workflow, os serviços de dados e o acesso para banco de dados e metadados.
- Kepler[20]; possui um ambiente que realiza a definição e execução de workflow para qualquer área científica, seu código fonte utiliza a linguagem Java.
- Taverna[21]; realiza o projeto e a execução de workflows específicos para a bioinformática, seu código fonte utiliza a linguagem Java.

Todas estas ferramentas de workflow desempenham atividades específicas que contribuem para a criação de workflows científicos, porém as ferramentas utilizadas para o desenvolvimento e execução de workflows em bioinformática são Kepler e Taverna. Além disso, estas duas ferramentas possuem características estruturais em comum, onde HOWINSON (2008) descreve como princípios básicos em comum entre Kepler e Taverna as seguintes citações:

- Etapas do fluxo de trabalho são realizadas pelos componentes que tem múltiplas portas de entrada e de saída.

[18] Fonte: http://www.vistrails.org

[19] Fonte: SILVA, Fabrício N. Da. *In Services*: Um sistema para gerenciamento de dados intermediários em workflows científicos na bioinformática. Instituto Militar de Engenharia, dissertação de mestrado, Rio de Janeiro, 2006.

[20] Fonte: http://www.kepler-project.org

[21] Fonte: http://taverna.sourceforge.net

- Componentes são conectados juntando uma porta de saída com outra porta de entrada.
- Um workflow é representado por um fluxograma e geralmente armazenado em um único arquivo no formato *XML*.

As ferramentas Kepler e Taverna por atenderem aos requisitos de regra e tarefas definidas para o desenvolvimento do workflow científico em bioinformática, serão estudadas mais detalhadamente nas próximas seções.

3.4.1 Kepler

Kepler é uma ferramenta utilizada para criação e execução de workflows científicos. Desenvolvida pelo projeto Ptolomeu II da Universidade de *Berkley* na Califórnia, utiliza um sistema baseado em Java e um conjunto de interfaces programadas para aplicações (APIs) em modelagem heterogênea hierarquizada (ALTINTAS, 2006). Tem como objetivo principal o uso de modelos computacionais que gerenciem a interação entre os componentes. Pode ser utilizada em diversas áreas, tais como ecologia, engenharia, astronomia e biologia. Os Workflows são compostos por um diretor e uma coleção de atores, o diretor controla a execução do workflow seguindo modelos computacionais específicos e os atores implementam as tarefas propriamente ditas, podendo ser simples ou compostas, o ator composto contém um conjunto de atores simples. O Kepler não possui uma estrutura formal de identificação dos dados. O armazenamento dos dados ocorre pela inserção de atores que realizam gravações em arquivo ou SGBD (MENEZES, 2007). Possui o código fonte implementado na linguagem Java, por isso é um *software* livre e pode ser utilizado em qualquer plataforma de sistema operacional.

Através da linguagem *Modeling Markup Language* (MoML) herdada pelo Ptolomeu II, os usuários podem construir workflows escolhendo atores com diferentes tarefas, e um ou mais diretores com diferentes estratégias de execução. Neste caso, os atores utilizam diferentes bibliotecas que podem estar em mais de um fluxo de trabalho, pois a aglomeração do MoML gráfico permite a utilização de um ou mais fluxos de trabalho em um mesmo workflow. A Figura 3.3 mostra um exemplo de workflow baseado na arquitetura MoML, neste exemplo existem quatro

atores constituídos por subworkflows, representados por: *PrepareInputFiles, PrepareExperiment, AddExperiment* e *ManageResearces* que são criados, implementados e testados separadamente, pois possuem em seu interior outro workflow com características e funcionalidades específicas entre os diretores e atores que constituem este subworkflow.

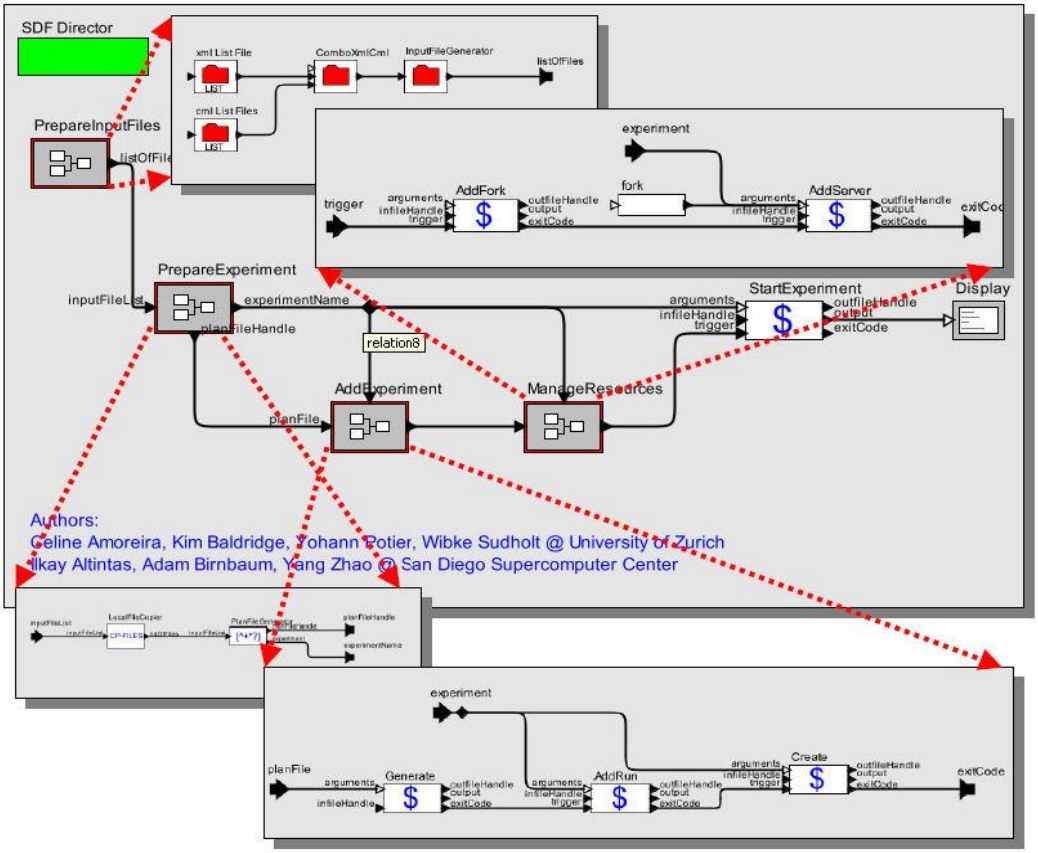

Figura 3.3: Exemplo Workflow baseado na arquitetura MoML (LUDÄSCHER, 2005).

Dentre as principais características apresentadas pelo Kepler estão:

- Possui uma interface de workflows desenhados no formato de circuitos eletrônicos, ligando-se por setas de entrada e saída que são conectados por objetos em formato de quadrados, conforme modelo da Figura 3.4.
- Trabalha na concepção de “diretórios” implementados a partir de modelos baseados na composição de componentes existentes chamados de agentes. A partir da observação do comportamento desses agentes, são executados simulações com diferentes semânticas computacionais dentro dos workflows.

- Destaca-se principalmente por desenvolver workflows na área da ecologia e geologia.

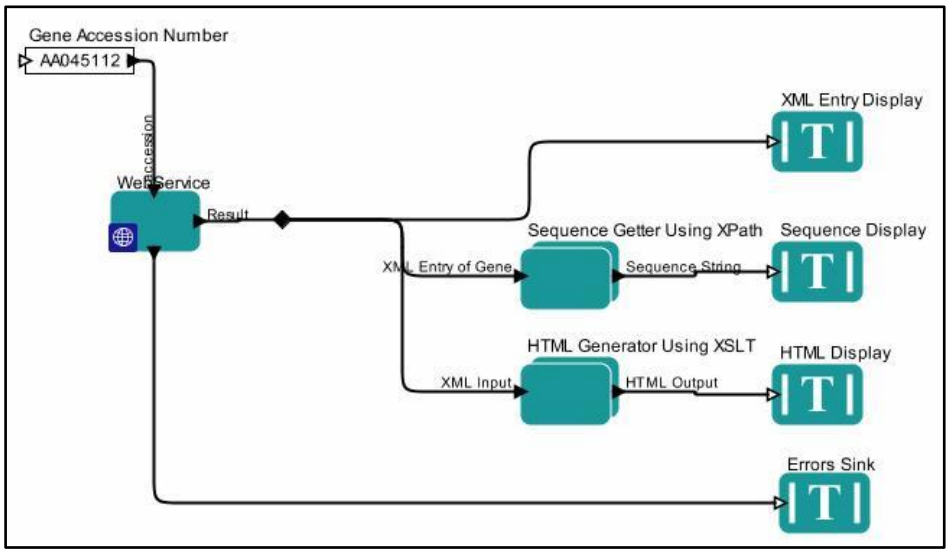

Figura 3.4: Exemplo do workflow construído no Kepler.

3.4.2 Taverna

Criado pelo projeto *MyGrid* e utilizado para o sistema de gerenciamento de workflow científico, o Taverna Workbench foi criado na Europa e atualmente é coordenado por Tom Oinn do Instituto de Bioinformática da Europa (EBI), juntamente com o *MyGrid development team* da Universidade de *Manchester*, a equipe conta com 10 desenvolvedores envolvidos para analisar, implementar testar e fornecer o suporte ao software. Taverna é uma ferramenta para desenvolver e executar workflows científicos, realizando a integração entre usuários de diferentes ferramentas de softwares, tais como EMBL-EBI (Cambridge, Reino Unido), NCBI (EUA), DDBJ e PDBJ (Japão). Da mesma forma que o Kepler, é implementado na linguagem Java é um *software* livre e pode ser executado em Windows, Linux, OSX e a maioria dos sistemas operacionais UNIX.

O projeto *MyGrid* possui componentes que auxiliam para uma melhor construção e execução do workflow. Cada componente fornece suporte para uma série de funções tais como, acesso das informações em bancos de dados públicos, manipulação de dados, armazenamento de dados intermediários e metadados, descoberta de novos recursos, comunicação entre diferentes ferramentas, utilização de *web services* para o acesso remoto e outros. Na Figura 3.5 pode-se observar a

forma com que os diferentes componentes interagem para o funcionamento do sistema.

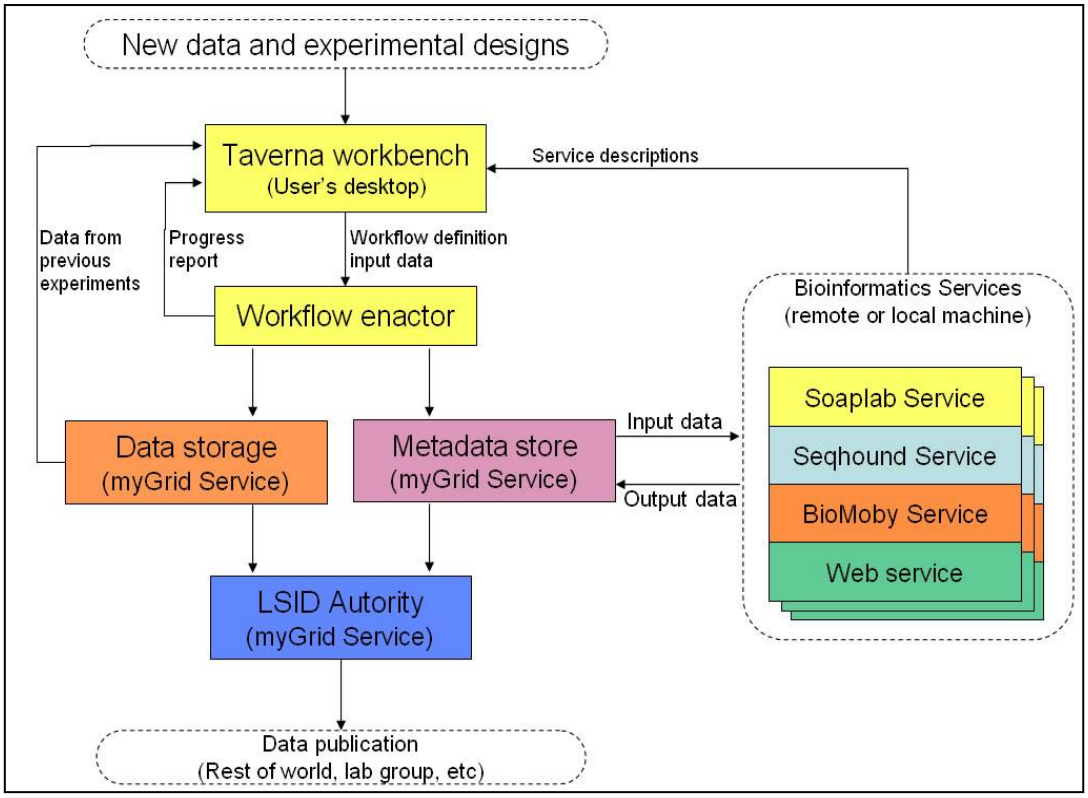

Figura 3.5: Componentes do *MyGrid* Toolkit (MYGRID PROJECT).

O Taverna Workbench é responsável por apresentar a interface ao usuário, fornecendo recursos para manipular as ferramentas disponíveis e para visualizar as etapas do projeto, permitindo a interação do usuário com o sistema. Além disso, esse componente realiza a junção de diferentes serviços, pois envia apenas uma solicitação que é distribuída pelos demais componentes, e o retorno é enviado pelo conjunto de diferentes serviços que fornecem os recursos para a utilização da ferramenta. Os três serviços de retorno existentes na arquitetura do Taverna Workbench são:

- Workflow *Enactor*; apresenta um relatório das etapas envolvidas durante a execução do workflow.
- *Bioinformatics Services*; captura os serviços remoto e/ou local, utilizados para manipular as informações.
- *Data Storage* e M*etadata Store*; captura os dados necessários para apresentar como resultado.

O Workflow *Enactor* responsável por controlar o fluxo de dados entre as etapas envolvidas na execução do workflow, realizando o gerenciamento de cada etapa do processo no momento da execução do workflow utilizando a linguagem de programação conhecida por *Simple Conceptual Unified Flow Language* (Scufl*)*. O Scufl é uma linguagem que permite aos usuários criar workflows sem ter de aprender a sintaxe da língua Scufl, pois manipula as ferramentas e desenvolve os workflow nos seguintes formatos: gráfico, XML e em estrutura de árvore. Além de ser uma linguagem simples para escrever, permite que outras linguagens o interpretem. Na Figura 3.6 pode-se observar o formato em XML no quadro A, o formato gráfico no quadro B e a estrutura em árvore no quadro C, representando os três formatos utilizados no Taverna pela linguagem Scufl.

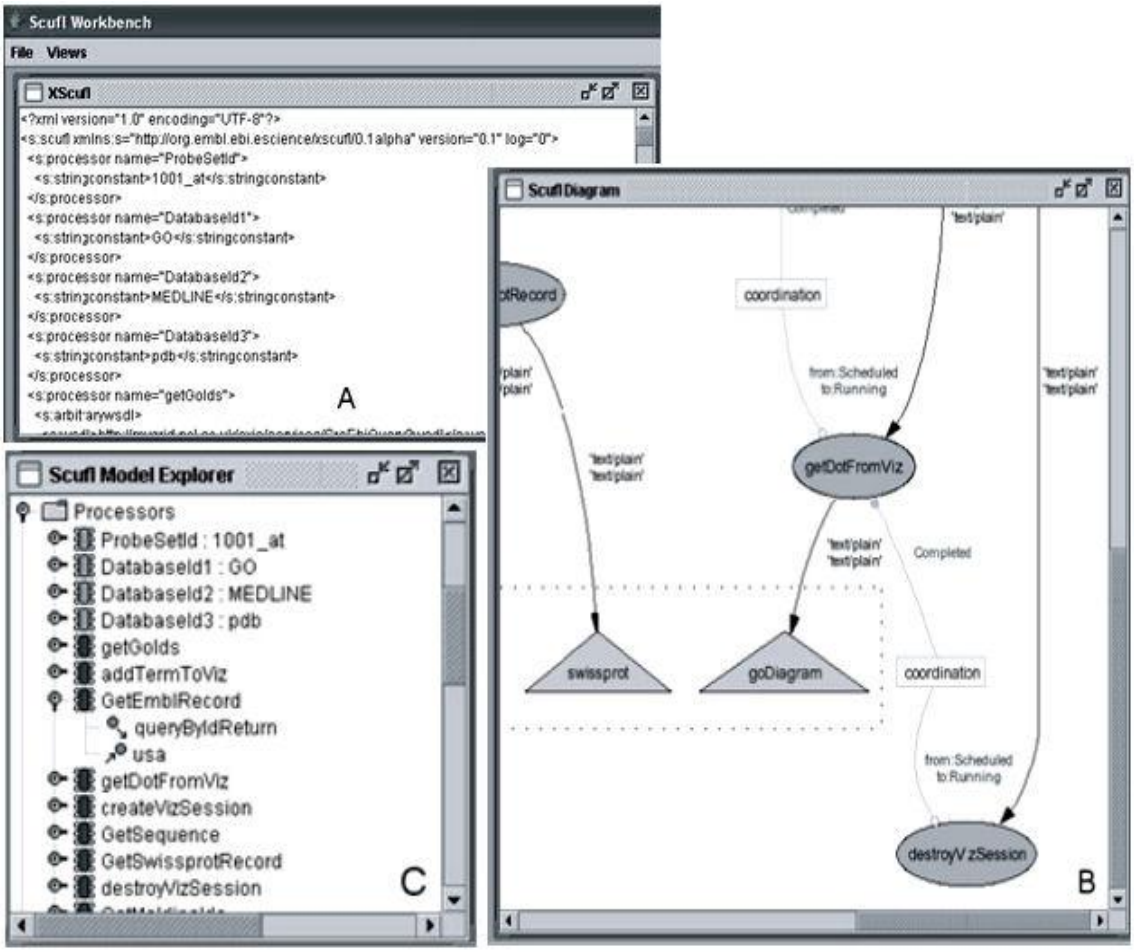

Figura 3.6: Formas de visualização da linguagem Scufl (OINN, 2004).

Para estruturar e organizar internamente os dados envolvidos durante a execução, o *MyGrid Toolkit* utiliza o Sistema de Gerenciamento de Banco de Dados (SGBD) MySQL, que armazena os dados resultantes de etapas intermediárias e

finais do processo através dos componentes *Data storage* e *Metadata store*. A interpretação de metadados necessita de API's que forneçam serviços para a bioinformática, dentre as principais utilizadas pelo Taverna estão: Soaplab, Seqhound e BioMoby. Taverna Workbench constantemente atualiza os acessos para esses serviços, permitindo interoperabilidade entre aos novos bancos de dados bem como para os novos serviços disponibilizados publicamente.

A identificação única e padronizada dos tipos de dados recebidos pelos componentes *Data storage* e *Metadata store*, é realizada pelo componente LSID *autority*. Sua funcionalidade é padronizar os tipos de dados oriundos de qualquer aplicação, sendo o responsável por permitir uma comunicação comum entre os dados através do uso do Life Science Identifiers (LSID).

Os componentes que estão inseridos nos serviços de bioinformática na ferramenta do Taverna Workbench fornecem os dados e as ferramentas necessárias para a realização dos processos, eles podem ser acessados de forma remota ou local. Os serviços remotos são compreendidos pela interação das URLs com *Web Services Description Language*[22] (WSDL), acessando os serviços *Simple Object Access Protocol* (SOAP). Os serviços locais utilizam componentes com uma linguagem de *script* simplificada em Java chamada de R_4, permitindo que o processamento dos workflows sejam acessados por arquivos, textos ou manipulação de listas.

A Figura 3.7 apresenta um modelo da interface gráfica utilizada pelo Taverna Workbench na construção de workflows científicos para bioinformática, onde cada etapa do fluxo pertence a um serviço específico. Os componentes que possuem em seu interior um triângulo com a seta voltada para cima e sem conectores de saída, representam as portas da saída de informações com a finalização do fluxo do workflow, os demais componentes apresentam uma conexão de entrada e uma conexão de saída contendo informações de saída e entrada parciais durante a execução do workflow.

[22] Linguagem baseado em XML utilizada para descrever *Web Services*. Contêm arquivos escritos em XML definindo os formatos de mensagem, os tipos de dados que o serviço manipula, os protocolos de transporte das mensagens de requisição aos serviços e de resposta do mesmo. As mensagens trocadas com os serviços Web seguem um padrão definido (SOAP) e podem ocorrer via o protocolo HTTP, podendo realizar a interoperabilidade com serviços Web mesmo que estejam implementadas em plataformas e/ou linguagens diferentes.

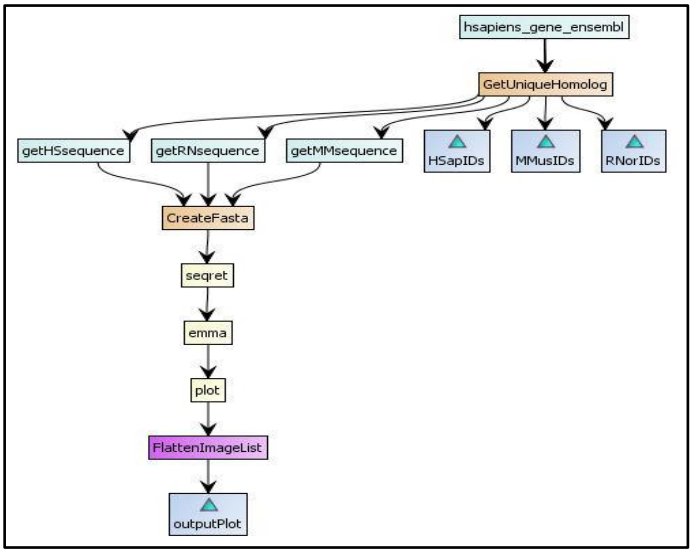

Figura 3.7: Exemplo da interface gráfica do Taverna Workbench.

3.5 Considerações Finais

Dentre as ferramentas de workflow científicos estudadas e/ou apresentadas neste trabalho, o Taverna Workbench foi escolhida como a ferramenta que será utilizada para a criação do workfow científico para bioinformática neste trabalho. Além de atender os requisitos do projeto, o Taverna Workbench possui as seguintes características que o diferencia das demais ferramentas para workflow científicos:

- Possui interface em um desenho de workflow no formato de árvores, sendo possível visualizar de forma gráfica os workflows ligando-se automaticamente entre si;
- Trabalha com um modelo de semântica computacional baseado no armazenamento dos dados em cada nível do fluxo, utilizando essas informações para inferir a proveniência de resultados intermediários e parciais, bem como validar a qualidade dos dados a fim de rastrear os passos em cada transformação*;*
- Seu foco principal é para a utilização na vida de cientistas nas áreas de biologia, química e para visualização de imagens médicas.

4 PROPOSTA DE MAPEAMENTO DE UM PROCESSO DA BIOLOGIA PARA WORKFOW CIENTÍFICO

Segundo GIBAS (2001) muitos pesquisadores, mesmo aqueles que fazem todo o seu trabalho no computador, mantêm livros de anotações de laboratório em papel, os quais ainda são o dispositivo de registro padrão para pesquisa científica, fornecendo um registro físico tangível das atividades de pesquisa. Entretanto, as anotações de laboratório podem ser uma fonte de informações inconvenientes, uma vez que ocorrem extravios e, em outras situações não recebem as devidas atualizações quando surgem novas descobertas científicas. Neste sentido os estudos em bioinformática estão tentando eliminar o trabalho manual dos biólogos, e introduzindo novas técnicas para o melhoramento dos processos.

A utilização de ferramentas para bioinformática, programas de computadores diversos, acesso em banco de dados, e as pesquisas em bancos públicos na web combinados com a interpretação humana, geram anotações referentes aos resultados encontrados, e muitas vezes referenciam as etapas do processo na tentativa de aperfeiçoar cada vez mais a execução dos procedimentos realizados durante a análise, através da manipulação de ferramentas e programas utilizados para a obtenção de resultados.

Para este capítulo serão apresentadas as etapas realizadas na criação do workflow proposto neste trabalho. Inicia-se pela descrição do cenário atual e levantamento de problemas encontrados, passa-se pela fase de levantamento de requisitos contextualizando as soluções dos problemas, para depois representar o fluxo do processo na biologia e especificar a automatização dos processos através do workflow.

4.1 Descrição do Processo Atual

Através de visitas realizadas ao Laboratório de Biotecnologia Vegetal e Microbiologia Aplicada da Universidade de Caxias do Sul (UCS), foi possível entender os processos de funcionamento das rotinas através de pesquisas, testes e esclarecimento de dúvidas com os especialistas. Nesse sentido, o estudo realizado para obtenção das informações necessárias ao desenvolvimento deste trabalho está baseado nas ontologias que possibilitam a especificação dos conceitos, ou seja, uma descrição concisa e não ambígua de quem são as entidades relevantes no domínio da aplicação e como elas se relacionam, a partir de experimentos empíricos da situação. A utilização de ontologias elimina incertezas e más-interpretações da semântica dos bancos de dados, programas e seus relacionamentos, facilitando a criação de sistemas.

Atualmente o Laboratório no qual foi desenvolvido esse trabalho não realiza processos totalmente automatizados, apenas utiliza ferramentas e programas de computadores separadamente para suprir uma determinada necessidade, e ainda mantêm ativo o processo de anotações muitas vezes automatizadas, mas sem uma procedência padronizada e organizada para o arquivamento destas informações.

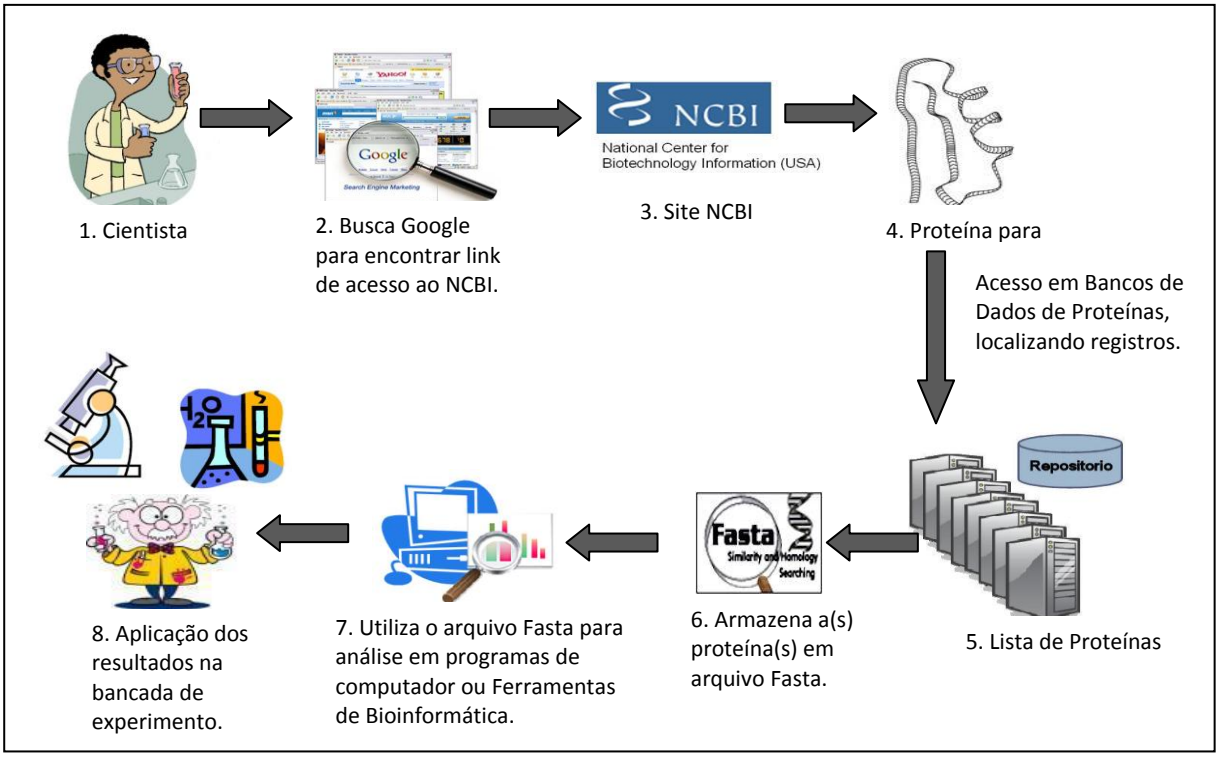

Figura 4.1: Cenário atual de experimentos no laboratório.

De acordo com a Figura 4.1, o processo inicia quando o cientista (1) utiliza uma ferramenta de busca na web (2) para encontrar o *link* de acesso ao portal do NCBI. Ao acessar o portal (3) são atribuídos parâmetros necessários para a localização de determinada proteína (4), o mecanismo de pesquisa do portal realiza o acesso nos bancos de dados disponíveis e encontra seqüências genéticas, que contenham nas entrelinhas dos arquivos armazenados a expressão da busca solicitada (5). Dentre a lista de resultados apresentados, o cientista realiza uma análise bastante criteriosa selecionando uma ou mais seqüências genéticas para a geração do arquivo em formato Fasta (6), recurso provido pelo próprio site do NCBI. Normalmente a gravação dos arquivos não tem local específico, eles são armazenados em um diretório qualquer para em seguida serem utilizados por um programa de computador que realiza a edição do arquivo, ou por uma ferramenta de bioinformática, como o Blast, ClustalX, que realiza o alinhamento das seqüências genéticas a partir dos registros do arquivo (7). O resultado obtido através da ferramenta que realiza o alinhamento é uma nova seqüência de código genético que passa da fase de análise para a fase experimental (8), onde esses resultados são aplicados nas bancadas de experimentos para gerar novas espécies, ou apenas gerando cópias de uma espécie geneticamente semelhante.

Dentro deste cenário procurou-se desenvolver uma ferramenta de workflow para ajudar o cientista, diminuindo o acesso em diversos programas de computadores, e ao mesmo tempo automatizando em parte o processo, utilizando-se das anotações e idéias dos especialistas da biologia. A seguir são apresentados os problemas relevantes do cenário atual, e em seguida descreve-se as soluções sugeridas para esses problemas.

4.1.1 Problemas do Cenário

Através da observação dos especialistas da biologia juntamente com os especialistas da computação, encontrou-se os seguintes problemas referente à utilização do processo atual:

1. Seqüência das ações está na “cabeça” do cientista.

A seqüência para execução dos programas e a necessidade de mover os dados de um ambiente para outro são ações que apenas o cientista responsável pelo experimento sabe como deve ser realizado e, mesmo que os processos sejam executados de forma diferente em cada situação o resultado final é obtido com sucesso. Porém a realização deste mesmo processo por outro cientista, ou por outra pessoa que não esteja familiarizada com o ambiente não proverá resultados satisfatórios.

2. Execução dos processos manualmente.

Tanto os processos manuais quanto as anotações são de grande importância e devem ser realizados de maneira precisa e não podem ser subestimadas, pois os erros de manipulação entre genes prejudicam a alta qualidade dos próprios dados de seqüências (LESK, 2008).

Mesmo que as ferramentas computacionais sejam utilizadas de forma independente, o acesso a elas e a forma com que os dados são manipulados ainda exigem alguns processos manuais ocorrendo em paralelo com a análise das espécies. Um exemplo de processo manual ocorre no momento em que é utilizado o bloco de notas do *windows* para editar determinado arquivo de formato Fasta, a fim de adequar o texto em um *layout* interpretável por outra ferramenta de bioinformática. Outro exemplo de processo manual é realizado ao armazenar e pesquisar arquivos Fasta. Neste momento o cientista necessita gravar as informações encontradas para em seguida realizar uma nova pesquisa, a fim de utilizar esse mesmo arquivo em outra ferramenta.

3. Falta de proveniência de dados e anotações.

Segundo MATOSO (2008), a procedência é útil para a compreensão do experimento e da cadeia de eventos usados na produção de um resultado, permitindo verificar se um experimento foi realizado de acordo com procedimentos científicos aceitáveis, identificando os dados de entrada de um experimento e sua origem e assegurando a qualidade dos dados.

O grande volume de dados públicos e a alta complexidade dos *softwares* para realizar o pareamento dos genes, muitas vezes não apresentam resultados satisfatórios ao final da análise. Ocorre a necessidade de realizar todo o processo

apresentado na Figura 4.1 inúmeras vezes, porém utilizando diferentes genes para uma mesma espécie, ou com espécies diferentes até encontrar um resultado aplicável para a biologia. Devido à falta de anotação ou identificação de proveniência dos dados a realização de uma nova busca torna-se bastante custosa, pois além do tempo gasto existe a possibilidade de reprodução do processo com espécies já utilizadas em outra análise.

Consultar dados de proveniência segundo MATTOSO (2008), é permitir o acesso de informações tais como, de onde vieram os dados, onde e quais as transformações que foram feitas neste dado e como cada resultado foi obtido é fundamental para a análise final.

4. Demora na busca de dados

O crescimento descontrolado dos bancos de dados públicos biológicos, e o grande número de novas fontes armazenadas interferem diretamente na capacidade que uma ferramenta terá para realizar a busca dos dados e mostrar os mesmos de forma interpretável para o usuário. Dependendo do tipo de informação solicitada e da ferramenta que estiver utilizando para realizar o acesso nos bancos de dados públicos, os resultados necessitam de mais tempo para serem interpretados e executados.

Outro fator importante é a instabilidade das redes de acesso utilizada nos laboratórios, pois o tempo de resposta das pesquisas depende da banda disponibilizada para o acesso a internet. Se o tráfego usual da rede for muito alto, o tempo de resposta das pesquisas tende a ser maior ainda, já se o tráfego usual da rede for mais baixo, o tempo de resposta das pesquisas normalmente é menor.

5. Acesso de várias ferramentas.

Nos últimos anos, diversas ferramentas para processamento de atividades em bioinformática foram disponibilizadas na Internet, sendo que muitas na forma de serviços web (DIGIAMPIETRI, 2007). Muitas destas ferramentas executam uma ótima análise, porém devido a dificuldades para manipulação e visualização de resultados não são utilizadas. A escolha pela ferramenta é realizada pelo próprio cientista que decide qual atende todos os requisitos e apresenta a melhor interface de trabalho. Por esse motivo no cenário apresentado pela Figura 4.1 não foram

identificadas as ferramentas utilizadas, apenas menciona-se as funcionalidades realizadas pelas ferramentas em cada etapa do processo.

6. Dificuldade em localizar as proteínas através de ferramentas web.

Muitos sites da *web* funcionam como portais entre dados armazenados em bancos de dados e as ferramentas computacionais disponíveis para a análise e busca dos dados. A recuperação de informações permite a seleção e a extração de dados a fim de fornecer os componentes de um projeto de pesquisa. O objetivo é fornecer a integração eficiente entre todas as etapas do processamento de dados necessárias para um projeto de pesquisa, por meio de uma conexão robusta entre as ferramentas para armazenamento, recuperação e análise de dados (LESK, 2008).

As ferramentas web nem sempre apresentam os resultados como os cientistas esperam receber, muitos portais utilizam mecanismos de pesquisas sofisticados e eficientes, porém devido a falta de padronização e informações dos bancos de dados de origem capturam também conteúdos irrelevantes e informações desnecessárias. Em outros casos são utilizadas expressões de busca desconhecidas e de alta complexidade para a interpretação de biólogos. Essa dificuldade encontrada pelos portais é verificada pelos cientistas da biologia que não conseguem localizar uma espécie específica, e pelos cientistas da computação que precisam desenvolver novas técnicas computacionais de busca, para a realização de pesquisas bem sucedidas entre os bancos de dados públicos.

4.2 Levantamento de Requisitos

Através dos padrões da engenharia de requisitos[23] que busca atender as capacidades e condições do sistema, o levantamento de requisitos realizou-se através da observação direta entre as partes envolvidas no projeto. Para isso foram utilizados o processamento e confirmação de resultados através de entrevistas com os especialistas, bem como a observação do desenvolvimento e das pessoas que desenvolvem o trabalho procurando identificar documentos que precisam ser coletados para posteriormente realizar o desenvolvimento do workflow.

[23] Fonte: PRESSMAN, Roger S. Engenharia de software. 6. Ed. São Paulo : McGraw-Hill, 2006.

Com a observação direta e com os problemas encontrados no cenário atual, descritos na seção anterior, identificou-se as principais necessidades que o workflow deve atender. Utilizando esse contexto avaliou-se as atividades realizadas manualmente pelo cientista da biologia, as ferramentas de software atualmente utilizadas e as rotinas realizadas neste processo e que necessitam de automatização a fim de agregar um melhor desempenho. Dentre as principais necessidades identificadas estão:

- ✓ Organização das atividades do processo;
- ✓ Necessidade de uma ferramenta única para a realização de todo o processo;
- ✓ Identificação da proveniência dos dados centralizada;
- ✓ Interoperabilidade entre diferentes softwares de bioinformática;
- ✓ Integração entre bancos de dados.

Para desenvolver o workflow com o objetivo de atender as necessidades descritas acima utiliza-se a metodologia de componentes reutilizáveis no contexto da engenharia de software (PRESSMAN, 2006), pois além de prover maior flexibilidade na estrutura do software produzido, o próprio ambiente de reuso fornece recursos para a utilização desta metodologia. O Taverna Workbench é considerado um ambiente de reuso, pois possui elementos de reutilização como acesso a bancos de dados e recuperação de componentes de software, proporcionando o suporte à integração de componentes para novos projetos, com capacidade em guardar e classificar as informações para depois recuperá-las.

O reuso é uma qualidade da engenharia de software baseada em componentes. Segundo PRESSMAN (2006), o desenvolvimento de softwares baseados em componentes elicita os requisitos do cliente, seleciona um estilo arquitetural apropriado para atender aos objetivos do sistema a ser construído e depois seleciona componentes potenciais para reuso, qualifica os componentes para se certificar de que se encaixam adequadamente na arquitetura do sistema, adapta os componentes se precisarem ser feitas modificações para integrá-los adequadamente e integra os componentes para formar subsistemas e toda a aplicação.

Atualmente a ferramenta Taverna oferece um *plugin* denominado *myExperiment*[24] que possibilita a busca e utilização de componentes adaptáveis aos workflows. O *myExperiment* publica aplicações desenvolvidas por estudantes e pesquisadores em bioinformática, permitindo a busca de componentes adaptáveis em aplicações criadas pelo Taverna Workbench e possibilitando adicionar e remover novos processos dinamicamente.

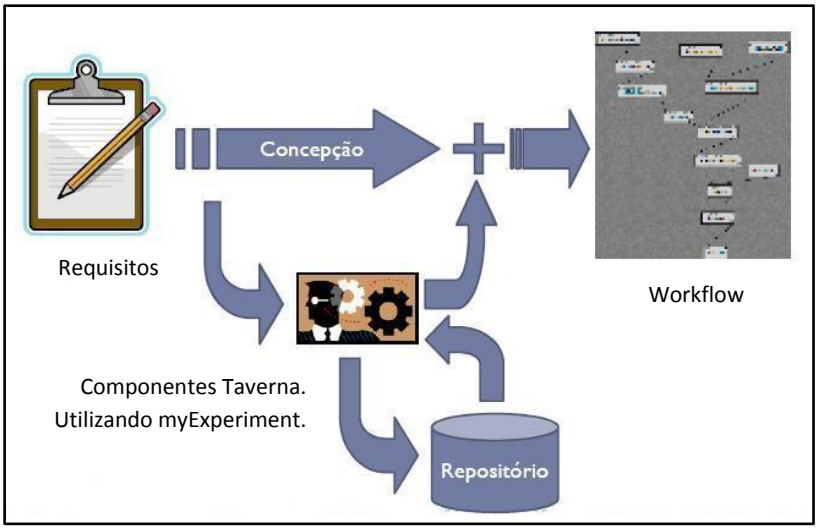

Figura 4.2: Processo de gerenciamento para o reuso (MATOSO, 2008).

Na Figura 4.2 observa-se o processo utilizado para a criação do workflow, utilizando o padrão de reuso. Primeiramente foi realizado a elicitação de uma lista de requisitos, identificando-se todos os problemas e necessidades encontrados no Laboratório de Biotecnologia Vegetal e Microbiologia Aplicada da Universidade de Caxias do Sul (UCS). Em uma segunda etapa realizou-se a busca por componentes reutilizáveis, através do portal *myExperiment,* e por fim a criação da ferramenta de workflow atendendo uma situação específica do laboratório. Para realizar a seleção de componentes foram observados os seguintes critérios:

- Componentes que pudessem ser adaptáveis para a situação;
- Componentes com acesso ao portal NCBI, pois possui o maior número de conexões para os vários bancos de dados biológicos; e
- Componentes com possibilidade de adição e remoção dinâmica de objetos.

[24] Fonte: http://www.myexperiment.org

Com objetivo de aplicar técnicas de engenharia de software existentes, o presente trabalho utiliza padrões de UML para representar o fluxo do processo proposto na biologia, e a representação funcional de cada etapa do processo através do workflow. Serão utilizados diagramas de UML aplicadas ao contexto de workflow científico, adaptando padrões específicos da UML com processo funcional de workflow, proporcionando o melhor entendimento do workflow proposto.

4.3 Apresentação da Proposta para o Processo na Biologia

Segundo LESK (2008), muitos *sites* da *web* funcionam como portais entre os arquivos em bancos de dados e as ferramentas computacionais disponíveis para a análise dos dados. A recuperação de informações permite a seleção e a extração de dados a fim de fornecer os componentes de um projeto de pesquisa. O objetivo é fornecer a integração eficiente entre todas as etapas do processamento de dados necessárias para um projeto de pesquisa, por meio de uma conexão robusta entre as ferramentas para armazenamento, recuperação e análise de dados.

A proposta para este trabalho proporciona a fusão e integração das fontes provedoras de dados em bioinformática, mapeando um entre os diversos processos que podem ser padronizados e gerenciados através das ferramentas de workflow. A presente aplicação desenvolvida executa as atividades de 1 até 6 representadas na Figura 4.1, onde o cientista realiza a busca de proteínas e solicita a geração do arquivo em formato fasta através do portal NCBI. Com o workflow proposto não será necessário realizar o acesso ao portal, pois o cientista executa todas as solicitações através do workflow que estabelece a conexão com os bancos de dados e portais necessários, fornecendo os resultados parciais e finais de acordo com os parâmetros informados e as execuções solicitadas no inicio do processo.

O objetivo inicial deste trabalho é a criação de um workflow para a geração de arquivo em formato Fasta de determinadas espécies genéticas, a fim de realizar esse processo de forma completa foi necessário adicionar um workflow para busca de proteínas que auxilia na geração deste arquivo. O workflow para a busca de proteínas fornece o número de Gi, conforme estudado no capítulo 2.4, utilizado como parâmetro de entrada do workflow de geração do arquivo Fasta. Também foi criado outro workflow com a finalidade de explorar os recursos existentes no Taverna Workbench, através da geração de imagens em 3D das proteínas.

Para realizar o mapeamento do fluxo deste processo proposto na biologia, utiliza-se a notação em UML do diagrama de atividades, pois além de possuir uma estrutura de fluxos de trabalho e processos do negócio, permite a visualização da seqüência ordenada das atividades nos diferentes processos. Através deste diagrama é possível descrever e apresentar como cada etapa do processo realiza as atividades, onde os estados são ações e a transição do evento é disparada automaticamente pelo término da ação (LARMAN, 2004).

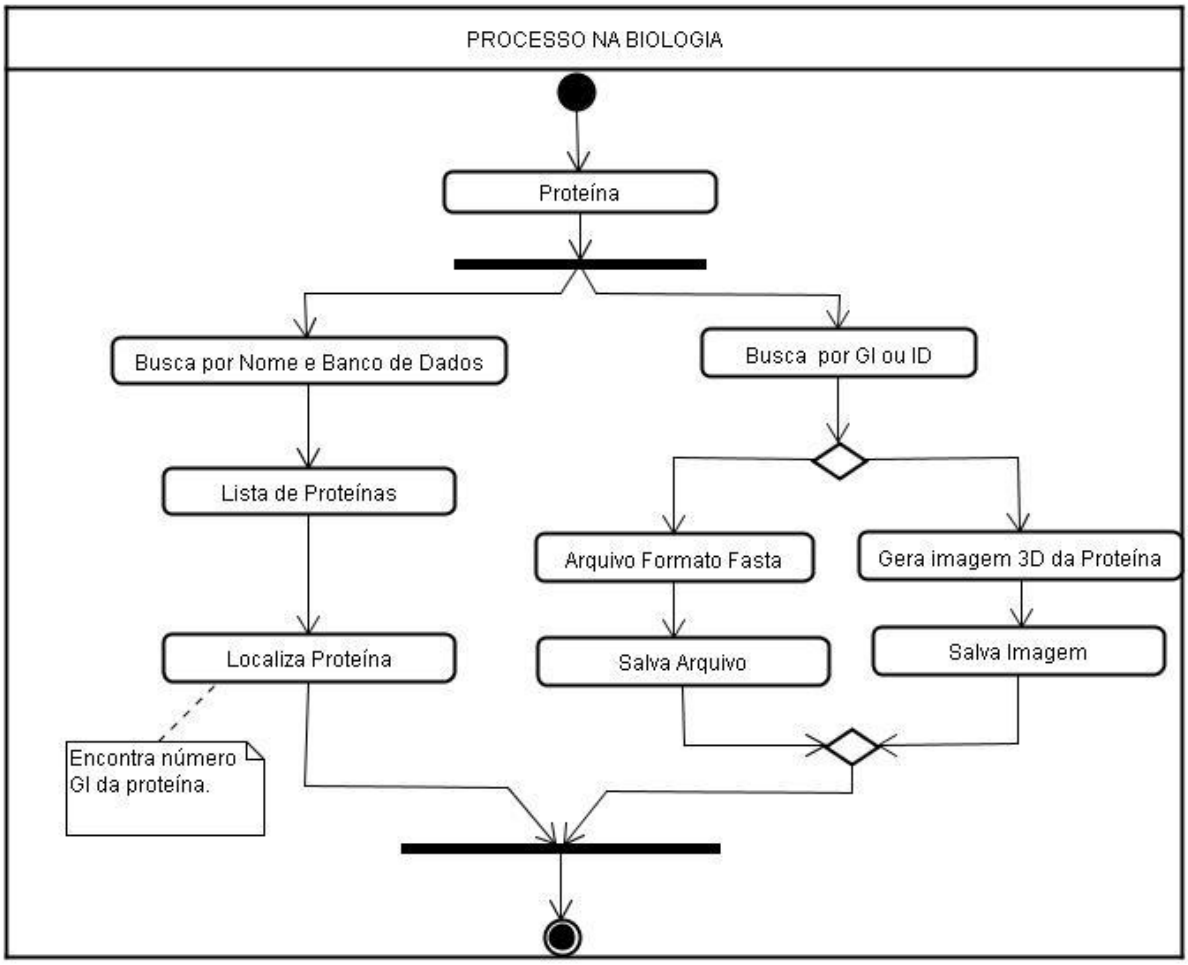

Figura 4.3: Fluxo do processo na biologia.

Na Figura 4.3 é possível visualizar o diagrama de atividade representando o fluxo proposto para o cenário atual, onde todas estas ações estão disponíveis do início ao fim na ferramenta de workflow proposta para bioinformática, e as atividades dependendo do tipo de solicitação executam um fluxo padrão de tarefas. Para este fluxo, a partir de uma determinada proteína é possível executar três diferentes situações:

- Através do nome e da localização (banco de dados) para acesso de uma proteína, gerar uma lista de proteínas encontradas com o número de Gi, ou;
- Realizar uma busca diretamente pelo número de Gi para a geração do arquivo em formato Fasta, e ainda;
- Gerar imagem para visualização em 3D de determinada proteína, através do banco de dados de estrutura de proteínas (PDB).

De acordo com o diagrama de atividades pode-se visualizar a separação inicial entre dois processos, e ao final a junção dos mesmos processos identificando a criação de dois diferentes workflows. Para tanto, utilizou-se um workflow utilizado na busca do número GI de várias proteínas, e outro a partir de uma proteína realiza a busca dos arquivos formato PDB e FASTA.

Alguns dos problemas levantados na seção 4.1.1 não puderem ser solucionados neste fluxo, devido aos fatores externos que influenciam diretamente a execução do workflow e estão limitados em alguns serviços, como por exemplo as restrições do Taverna Workbench, o acesso aos bancos de dados, a instabilidade de redes de acesso, entre outras. Na Tabela 4.1 é apresentada uma associação entre os problemas encontrados, identificando se foram solucionados pela proposta sugerida.

Tabela 4.1: Associação de Problemas X Soluções

	Descrição do Problema	**Solucionado**
1	Seqüência das ações está na cabeça do cientista	Sim
2	Execução dos processos manualmente	Sim
3	Falta de proveniência de dados e anotações	Sim
4	Demora na busca de dados	Não
5	Acesso em várias ferramentas	Sim
6	Dificuldade de localizar proteínas através de ferramentas web	Não

O problema número 4, referente à demora na busca dos dados não será solucionado na ferramenta proposta para este trabalho, pois depende de fatores externos tais como: a disponibilidade dos registros através dos bancos de dados públicos, a complexidade do algoritmo utilizado para realizar o processamento destas informações e a quantidade de tráfego na rede para acesso à Internet utilizado pela instituição.

O problema número 6, referente à dificuldade de localizar substâncias através de ferramentas *web* também não foi possível solucionar, pois o Taverna Workbench utiliza as mesmas conexões de acesso e busca à banco de dados do portal NCBI. Para a realização deste trabalho não foi encontrada nenhum objeto próprio do Taverna, que realizasse a busca de acesso aos bancos de dados de proteínas independente do mecanismo utilizado pelo NCBI.

4.4 Aplicação do Processo Proposto na Ferramenta de Workflow

O principal objetivo deste trabalho é desenvolver um workflow que realize a integração e geração do arquivo em formato fasta, através de uma ferramenta de workflow científico para bioinformática que suporte diferentes espécies de genes. Partindo deste desafio, desenvolveu-se um workflow específico baseado no fluxo apresentado pela Figura 4.3, utilizando componentes reutilizáveis e adaptando-os para atender as necessidades do cenário atual.

Para demonstrar a interação entre os objetos da ferramenta Taverna Workbench que interagem para executar os diversos passos de atividades, utiliza-se os modelos em UML de diagramas de colaboração[25], baseados na concepção de fluxos dos eventos representando cada processo em um diagrama. Os diagramas de colaboração são utilizados para expressar interações de mensagens similares, ilustrando os objetos em forma de grafo ou rede, na qual podem ser colocados em qualquer lugar do diagrama, sem precisar seguir uma seqüência (LARMAN, 2004).

De forma geral os diagramas de colaboração permitem mostrar uma organização espacial dos componentes e interações, mostrando uma interface organizada em torno dos objetos e seus vínculos entre si. Possui um número que determina a seqüência de mensagens e os segmentos simultâneos, onde uma situação específica é descrita por setas numeradas que mostram o movimento das mensagens durante o desenvolvimento de uma situação.

Nas Figuras 4.4, 4.5 e 4.6 podem ser observados os diagramas de colaboração representando os diferentes eventos existentes na ferramenta de workflow, para os processos de busca, geração do arquivo em formato Fasta e geração de imagem em formato PDB. Cada processo é executado de forma

[25] Fonte: LARMAN, Craig. Utilizando ULM e Padrões. 2.ed. Porto Alegre : Bookman, 2004.

independente, porém o Taverna permite realizar a execução de mais de um processo simultaneamente, de forma que cada atividade realiza os eventos de leitura do(s) dado(s), execução dos processos e apresentação do(s) resultado(s) de forma independente, possibilitando a abertura de uma ou mais interfaces para cada um destes processos.

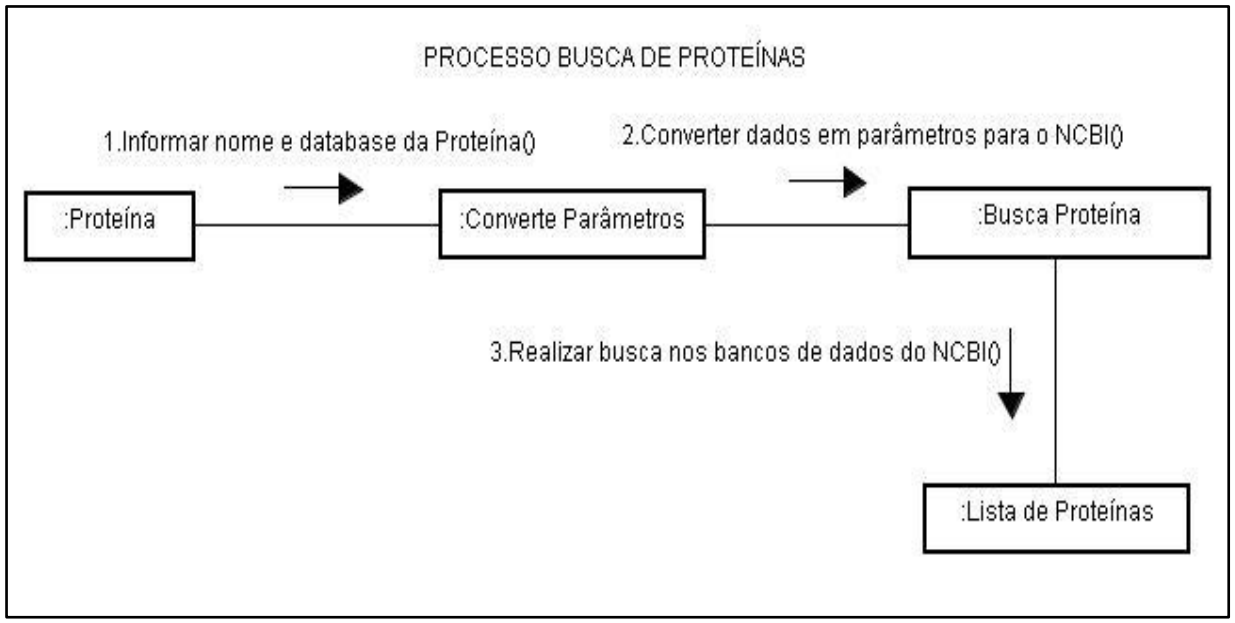

Figura 4.4: Processo para busca de proteínas.

O processo para buscar proteínas no portal do NCBI é representado pelo diagrama de colaboração da Figura 4.4. A execução deste processo é realizado separadamente dos outros dois eventos, pois esta busca torna-se necessária quando o usuário desconhece o número Gi, para solicitar um arquivo em formato Fasta de uma determinada proteína. Como entrada esse processo solicita dois parâmetros, um deles é uma expressão com o nome da proteína para a busca e o outro é uma expressão identificando em qual banco de dados está contida a proteína para ser localizada. Os eventos para converter parâmetros e busca de proteínas são executadas pelo workflow, que busca os componentes necessários para a realização destes processos. A saída de deste processo é um relatório em formato XML com o número de GI das proteínas encontradas a partir dos dados de entrada fornecidos.

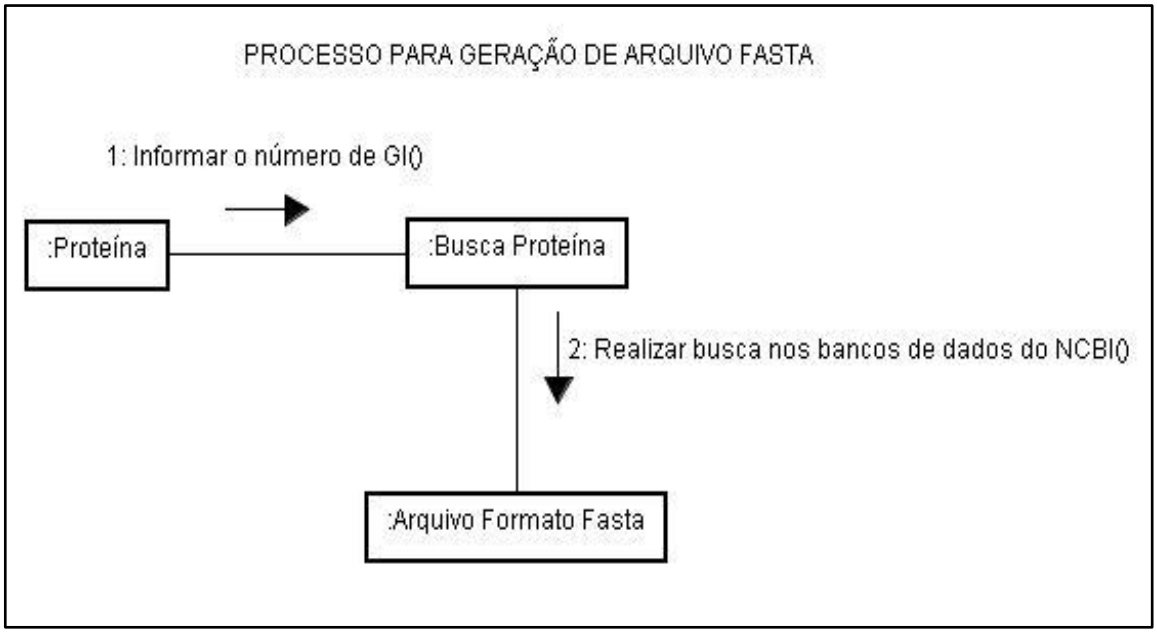

Figura 4.5: Processo para geração de arquivo Fasta.

Na Figura 4.5 é apresentado o processo para a geração do arquivo em Formato Fasta, onde o parâmetro de entrada é um número, ou vários números de Gi das proteínas que deseja-se obter o arquivo fasta. O evento de busca da proteína é executado pelo workflow que realiza a consulta no portal do NCBI, e retorna a seqüência da proteína em formato fasta, permitindo armazenar em arquivo texto.

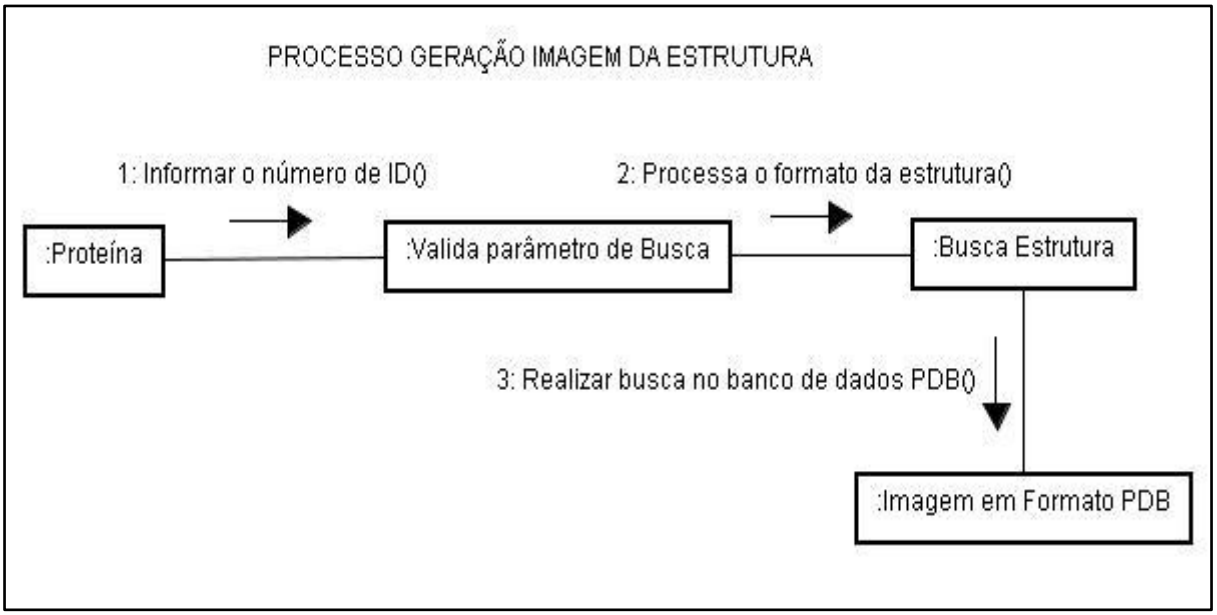

Figura 4.6: Processo para Geração de Imagem da Estrutura.

O processo que realiza a geração da imagem em 3D de determinada proteína está representado pelo diagrama de colaboração da Figura 4.6. Para a execução deste processo é necessário informar o ID, ou seja, o número que identifica a estrutura da proteína no PDB. Os processos para validação de parâmetros de busca e busca da estrutura é realizado internamente pelo Taverna Workbench que valida as informações para apresentar como saída a imagem 3D da proteína. O formato do arquivo gerado é PDB, pois é utilizado pelo banco de dados PDB, sua visualização pode ser realizada no próprio Taverna Workbench e também utilizando *softwares* específicos para a visualização e edição da imagem.

5 DESENVOLVIMENTO DO WORKFLOW PROPOSTO

A criação de workflows deve incluir processos, dados e recursos normalmente utilizados para oferecer mecanismos de extensibilidade a fim de acomodar novos processos, dados e recursos. Entre os processos estão os programas de análise de bioinformática, programas estes que filtram os resultados de outros programas e mecanismos de controle de execução do workflow (LEMOS, 2004).

Para entender o processo de desenvolvimento do workflow proposto, este capítulo apresenta uma definição detalhada dos recursos e dos serviços utilizados e disponíveis pelo Taverna Workbench. Em seguida explica-se os procedimentos necessários para ativar a execução do workflow, primeiramente descrevendo as diferentes etapas do processo e em seguida apresentando exemplos práticos com os workflows propostos. Por fim, um relato dos especialistas avaliando em diversos aspectos a utilização do workflow como um programa de bioinformática no laboratório de biotecnologia da UCS.

5.1 Criação do Workflow

Para a criação de um workflow através da ferramenta Taverna Workbench é importante saber o que pretende-se desenvolver e quais os serviços deseja-se utilizar, com o propósito de verificar se é possível desenvolver tal aplicação através dos serviços disponíveis nesta ferramenta. Após encontrar os serviços disponíveis na ferramenta para o desenvolvimento da aplicação, é necessário verificar se as entradas (*inputs*) disponíveis em cada serviço atendem as necessidades, bem como verificar se as saídas (*outputs*) disponíveis em cada serviço também atendem as necessidades de resultados esperados pelo especialista.

Em um mesmo workflow pode ser necessário utilizar um ou mais serviços com finalidades diferentes. Alguns serviços são configurados através de parametrização atribuindo valores possíveis e pré-definidos, e com a inserção de linhas de programação na linguagem *JavaScript*. Os serviços remotos têm a

finalidade de realizar a conexão através de *web services*, já os serviços locais utilizam de ferramentas locais herdadas da ferramenta Taverna Workbench.

Faz-se acesso a dois portais: o NCBI aplicando serviços locais, e o EBI utilizando os serviços do SOAPLAB[26], ambos os serviços disponíveis pela ferramenta Taverna Workbench. Para tanto foi necessário criar três algoritmos de workflow:

1- Realiza a busca dos números de GI de espécies, utilizando os serviços locais do NCBI;

2- Realiza a geração do arquivo em formato fasta, utilizando os serviços locais do NCBI para localizar as proteínas pelo número de GI, em seguida realizando o acesso ao banco de dados onde encontra-se a informação de saída;

3- Realiza a geração do arquivo em formato PDB, utilizando os serviços do SOAPLAB para acessar o banco de dados PDB, a fim de localizar as proteínas e obter a informação de saída.

As informações de entrada para os workflows propostos são:

- Expressão, permite informar qualquer palavra;
- Banco de Dados, permite informar os bancos de dados disponíveis para a busca da expressão. Os bancos de dados disponíveis são os mesmos utilizados pelo NCBI.
- Quantidade Registros, permite informar a quantidade máxima de números de GI que serão apresentados no relatório em XML.
- Número GI, número identificador da proteína no portal NCBI.
- Identificador ID, número identificador da proteína no banco de dados do PDB.

As informações de saída para os workflows propostos são:

- Lista em XML com o número de GI;
- Arquivo em formato Fasta, permitindo salvar em arquivo texto;
- Arquivo em formato PDB, permitindo salvar em arquivo PDB e no formato de imagem gráfica.

[26] É um *web service* disponível no Taverna Workbench, que fornece linhas de comandos para programas em bioinformática, incluindo bibliotecas e proporcionando suporte para a geração de aplicações web baseados nestas bibliotecas. Utilizando comandos baseados em XML, evitando linhas de comando de baixo nível para a programação de *web services*.

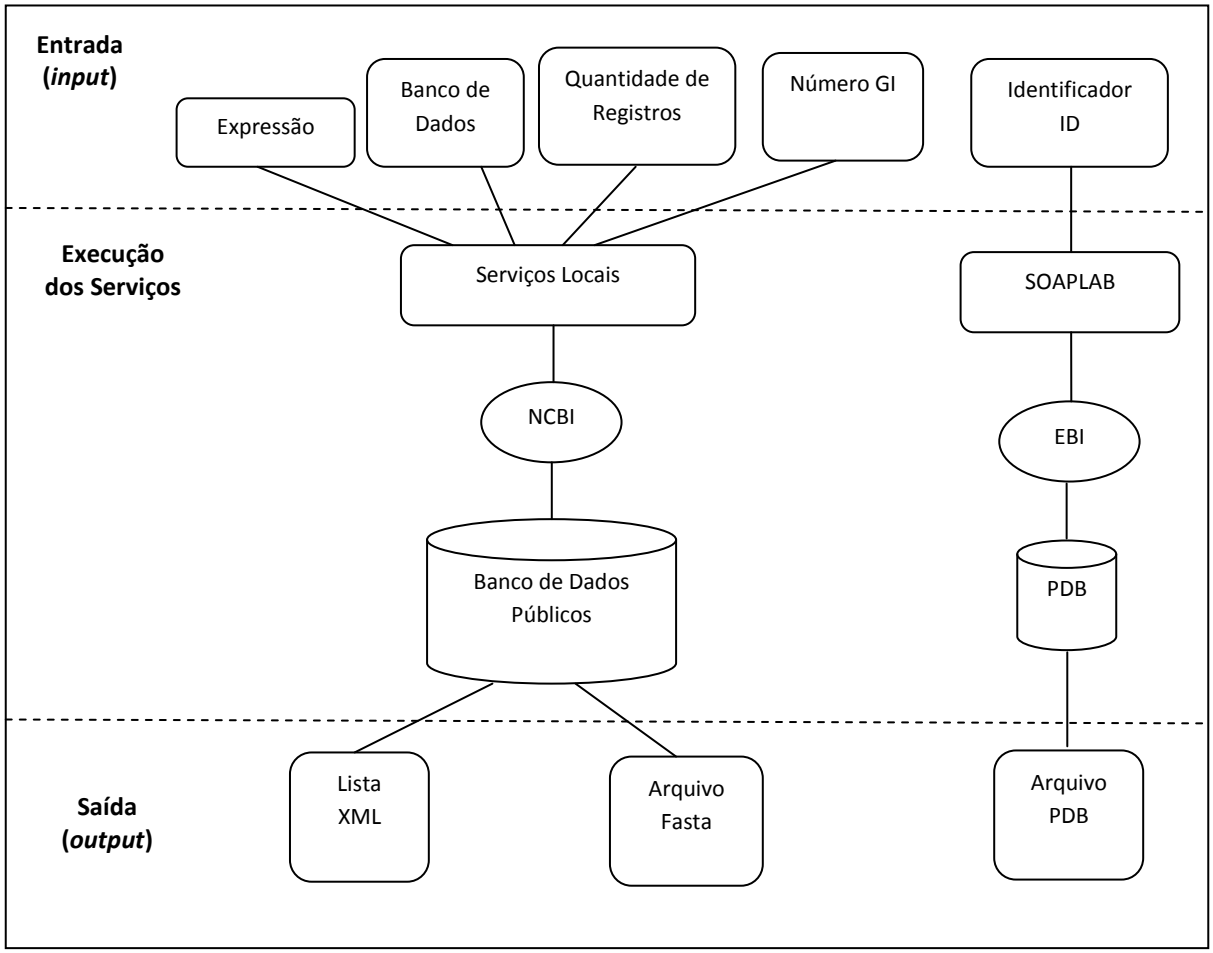

Figura 5.1: Fluxo das informações para a criação do Workflow.

De maneira geral os serviços acessados, as informações de entrada e as informações de saída do workflow proposto estão representadas na Figura 5.1, porém para executar os serviços de conexão aos bancos de dados e busca da informação correta utiliza-se diversos serviços, que interligados auxiliam para construir as informações de saída.

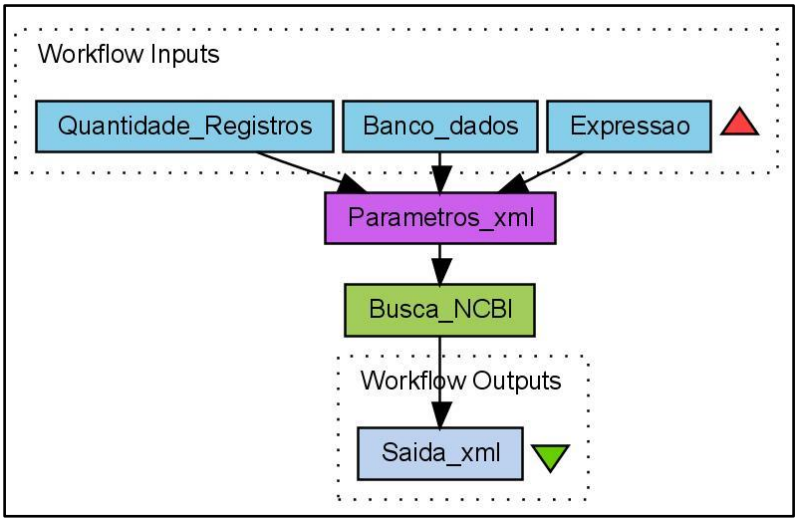

Figura 5.2: Workflow para busca do número de GI.

A Figura 5.2 apresenta o workflow criado para realizar a busca do número de GI, são utilizados os serviços Parametros_xml e Busca_NCBI. Em cada um destes serviços são utilizados recursos da ferramenta Taverna Workbench, o nome de cada processo e os serviços que são utilizados na aplicação, e uma breve explicação das funcionalidades está representado pela Tabela 5.1.

Tabela 5.1: Serviços Workflow Busca_NCBI

Nome	Processo	Serviço	Descrição
Parametros_ xml	*Get Protein GBSeq XML, do* ncbi.	Local	Converte a informação em formato texto para o formato em XML. Permite a entrada de várias informações, e retorna um texto em XML que pode ser utilizado como entrada para outro processo. No workflow da Figura 5.2 os dados de entrada compreendem-se em: expressão, banco de dados e quantidade de registros. Esses dados são transformados em formato XML por este componente, e utilizados como informação de entrada para o processo Busca_NCBI.
Busca_NCBI	*run_ eSearch*	WSDL do Ncbi	Esse processo utiliza a ferramenta ENTREZ do NCBI para executar a busca de informações. Recebe como entrada informações no formato XML, acessa o(s) banco(s) de dados solicitado(s) e busca registros que contenham a expressão de entrada, por fim retorna uma lista em XML com o número de GI das espécies pesquisadas.

A Figura 5.3 contém o workflow proposto para gerar os arquivos Fasta e PDB, utilizando os seguintes serviços: Teste_id_estrutura, Id, Estrutura, fetchData, Utiliza_estrutura e Busca_proteina_fasta.

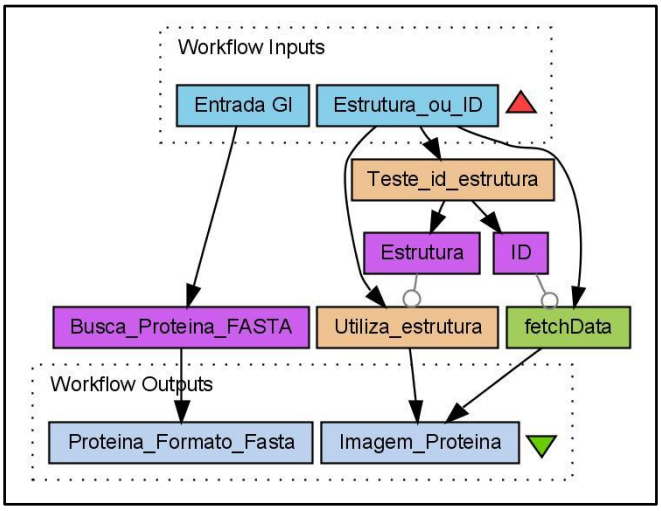

Figura 5.3: Workflow para geração de arquivo Fasta e PDB.

Na Tabela 5.2 representa-se os serviços utilizados na execução workflow gera arquivos em Formato Fasta e PDB. Para cada processo é apresentado a sua identificação através do nome de cada processo, o nome do processo na ferramenta Taverna Workbench, o tipo de serviço e uma breve descrição de sua funcionalidade.

Tabela 5.2: Serviços Workflow Gera Arquivos

Nome	Processo	Serviço	Descrição
Teste_id_ estrutura	*Beanshell scripting host*	Local	Permite inserir *scripts* na linguagem Java, com a finalidade de criar saídas adequadas para a entrada de um próximo processo. Esse componente possui um interpretador interno de objetos característicos do Java, permitindo utilizar sua sintaxe através de comandos, tipos de dados e outros recursos com extensões de uma linguagem em *JavaScript*. No workflow proposto utiliza-se um *script* atribuindo valores e realizando testes condicionais para a entrada dos dados, verificando se a entrada é uma estrutura ou se é apenas o ID da proteína existente no PDB, dependendo da saída será utilizado o processador para a interpretação de uma estrutura ou do ID.

Nome	Processo	Serviço	Descrição
Utiliza_estrutura	*Beanshell scripting host*	Local	Com a mesma funcionalidade do componente Teste_id_estrutura, porém para este caso o componente atribui na porta de saída o resultado da pesquisa realizada a partir da entrada de dados, no formato da estrutura bd:id.
Id	*Fail if true*	Local	Os serviços condicionais executam os processos através de ramificações, onde é possível quebrar a regra de que todos os serviços devem ter sidos executados para apresentar o resultado final do workflow. Permite realizar a execução dos serviços inferiores ao processo, desde que uma das partes esteja satisfeita pela condição. Neste workflow, os serviços inferiores ao processo ID somente serão executados se a entrada dos dados for apenas o número de ID do PDB, caso contrário executará o serviço Estrutura.
Estrutura	*Fail if true*	Local	Utilizando os mesmos serviços do componente Id, os processos ligados ao componente Estrutura somente serão executados se a entrada dos dados apresentar o formato de estrutura, com suas características devidamente estabelecidas, caso contrário irá executar o componente ID.
FetchData	*run_ eFetch*	WSDL SOAP do EBI	Esse serviço permite realizar o acesso aos bancos de dados através de identificadores de entrada ou número de adesão, retornando como resultado a saída de um arquivo em diferentes formatos. O formato da estrutura de entrada é bd:id, ou seja banco de dados:id do PDB. Para esse workflow são utilizados testes, validando se a informação de entrada apresenta a sintaxe solicitada, e como saída apresenta o arquivo em formato PDB.
Busca_proteina _fasta	*Get Protein* Fasta do NCBI		Recebe como entrada o número de Gi e retorna o arquivo em Formato Fasta.

Todos estes serviços estão implementados e estruturados em arquivos XML, gerados e interpretados pela própria ferramenta Taverna Workbench. Os arquivos possuem a sintaxe em XML e podem ser visualizados através de ferramentas para interpretação de texto. Os arquivos: “workflow_busca.xml”, utilizado para a busca do número de Gi e, “workflow_fasta_imagem.xml”, utilizado para geração dos arquivos no formato Fasta e PDB estão disponíveis no Anexo 1 deste trabalho.

5.2 Execução do Workflow

O Taverna Workbench fornece recursos para a execução do workflow através de características úteis, incluindo habilidade de interagir parcialmente com o usuário através dos fluxos de trabalho automatizados, que exigem intervenções humanas na execução de tarefas independentes ou em paralelo, na interrupção, na parada e no reinício da execução do workflow devido a alguma falha ocorrida durante o processo (CORNERY, 2005).

A ferramenta além de possuir um grande número de serviços para criação do workflow oferece também uma interface dinâmica e flexível para atender as necessidades do usuário. Permite executar o mesmo, ou workflows distintos uma ou mais vezes simultaneamente, apresentando a saída dos resultados separadamente, onde podem ser visualizados acessando os ícones no formato de paletas rotuladas com o nome dos workflows executados. Ao acessar a paleta do workflow, é possível visualizar informações como: o status, os resultados de saída e um relatório com os comandos em XML utilizados durante a execução, estes recursos também estão dispostos em formato de ícones em paleta para facilitar a visualização e a disposição das informações.

Para iniciar a execução de um workflow construído a partir dos recursos do Taverna Workbench é preciso abrir o arquivo em formato *XML,* e em seguida acessar a opção File -> “*Run* workflow...”, conforme a Figura 5.4 demonstra.

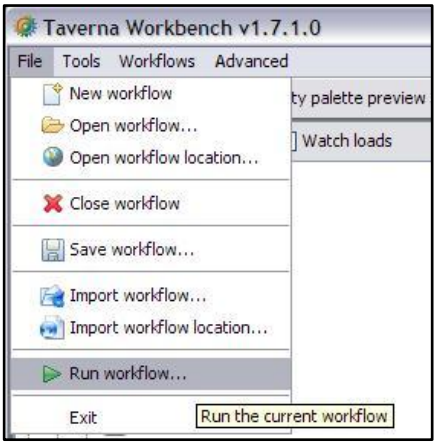

Figura 5.4: Opção para iniciar a execução do workflow.

A Figura 5.5 apresenta a interface aberta após selecionar a opção "*Run* workflow..". Nessa janela é necessário atribuir às informações de entrada necessárias para que o workflow possa interpretar, processar e retornar a saída dos resultados esperados. A parte superior desta tela apresenta o nome, a imagem e uma explicação do workflow que está sendo executado. A parte inferior oferece recursos para a inserção dos dados de entrada, permitindo localizar a partir de um arquivo ou descrevendo na própria interface os parâmetros de busca.

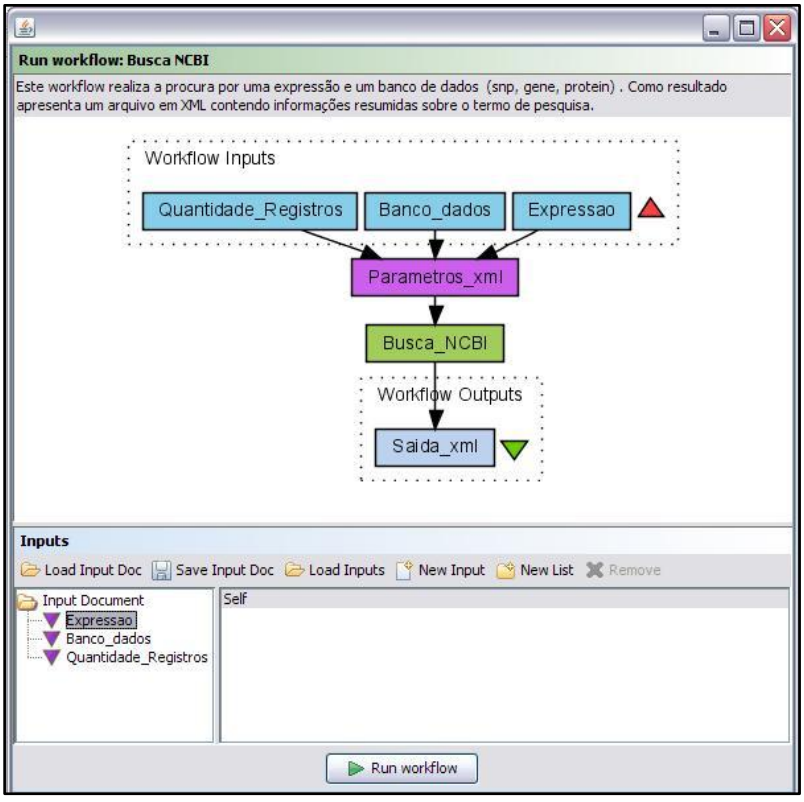

Figura 5.5: Interface inicial do workflow.

Ao selecionar a opção *Run* workflow da Figura 5.5, o Taverna ativa a opção para acesso de resultados, apresentando os processos de execução, invocação, chamadas e finalização dos processos, em seguida apresentando os resultados obtidos a partir dos dados de entrada fornecidos pelo usuário. Esse recurso não possui restrições, permite que sejam executados, através da opção *Run* workflow, mais de um workflow com parâmetros iguais ou diferentes, bem como permite abrir outras interfaces para workflows distintos, possibilitando a navegação rápida e flexível para o acesso das informações.

As Figuras 5.6 e 5.7 possuem a interface utilizada pelo Taverna enquanto o workflow é executado, apresentando os diferentes estados durante a realização de todos os processos, bem como a data e hora de início de cada processo e ainda um detalhamento dos eventos que estão sendo executados. Na Figura 5.6 um exemplo de workflow sendo executando com o status *Running* e no ícone de paleta *Status* os processos completos, invocados, chamados e com falhas. Já na Figura 5.7 o workflow está no processo finalizado, onde todos componentes apresentam status de processo finalizado, e contém outros dois ícones de paleta, um deles para acessar os resultados e outro para visualizar o algoritmo utilizado na execução do workflow. Além da possibilidade de visualizar as etapas do processo, a ferramenta permite a visualização gráfica do workflow, que pode ser observado na parte inferior destas Figuras, permitindo acessar as opções de proveniência dos dados, visualizar os dados de entrada e a construção em XML dos parâmetros utilizados para gerar as informações de saída.

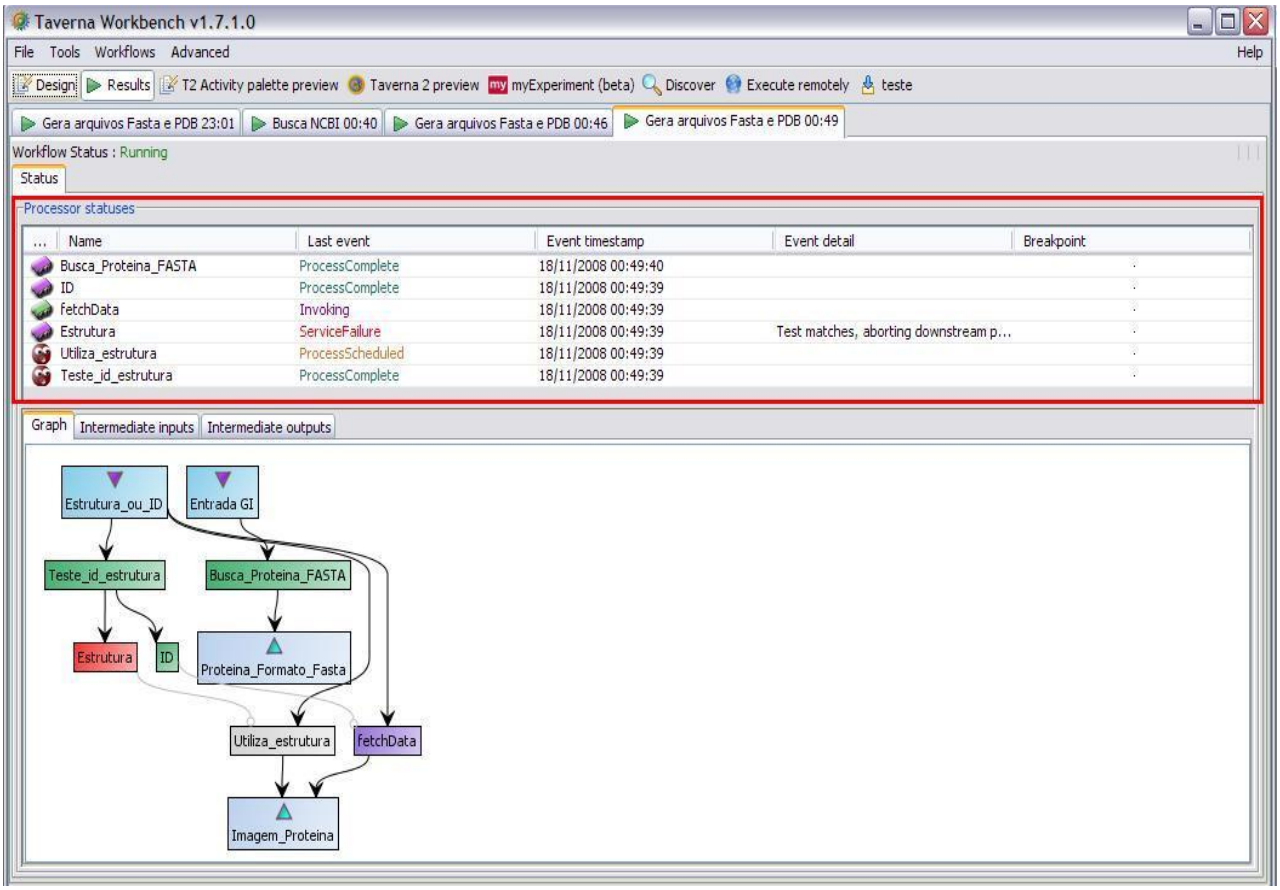

Figura 5.6: Interface do workflow em estágio de execução.

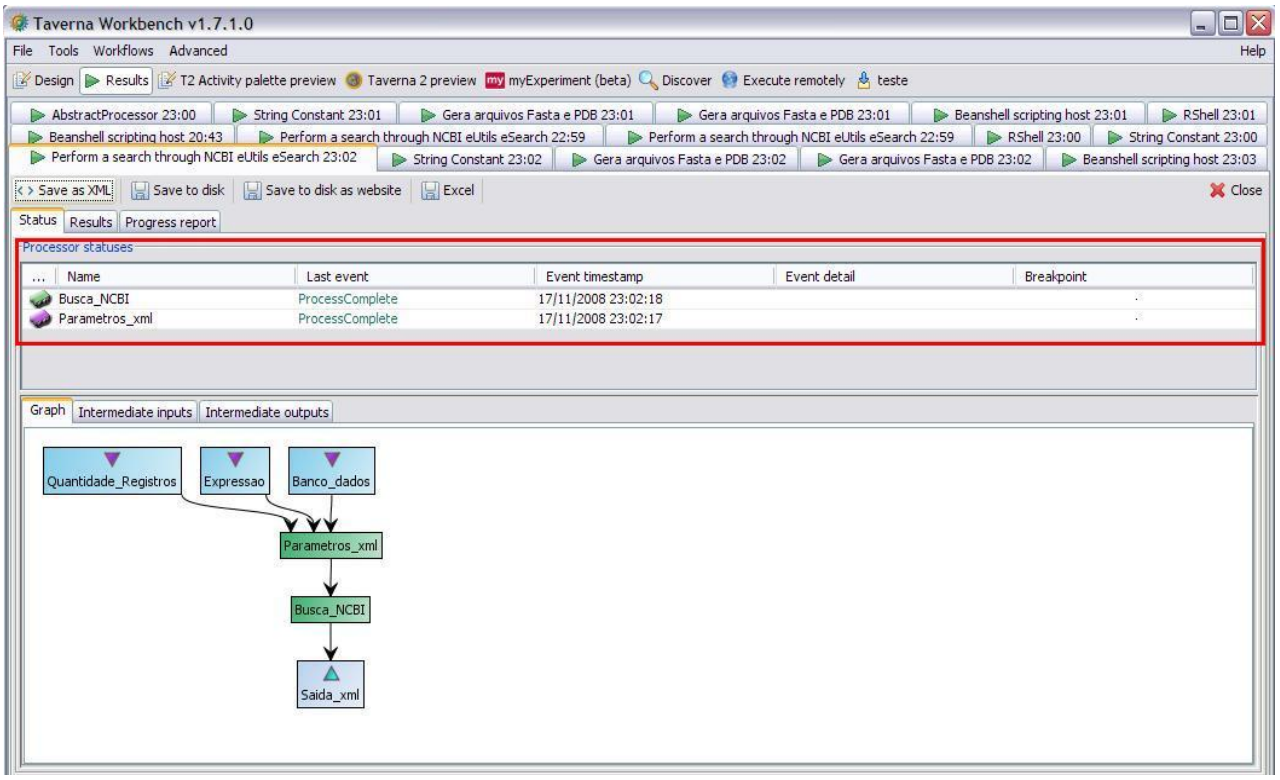

Figura 5.7: Interface dos resultados de saída dos workflows executados.

Na parte superior destas interfaces existem ícones em paletas possibilitando a navegação entre os workflows em execução e os já executados. Na Figura 5.8 pode-se visualizar um grande número de workflows executados, onde é identificado o nome e a hora de execução dos processos. Logo abaixo das paletas, existem as opções para salvar os resultados em arquivos, sendo possível salvar vários formatos dependendo de como os dados de saída estão dispostos.

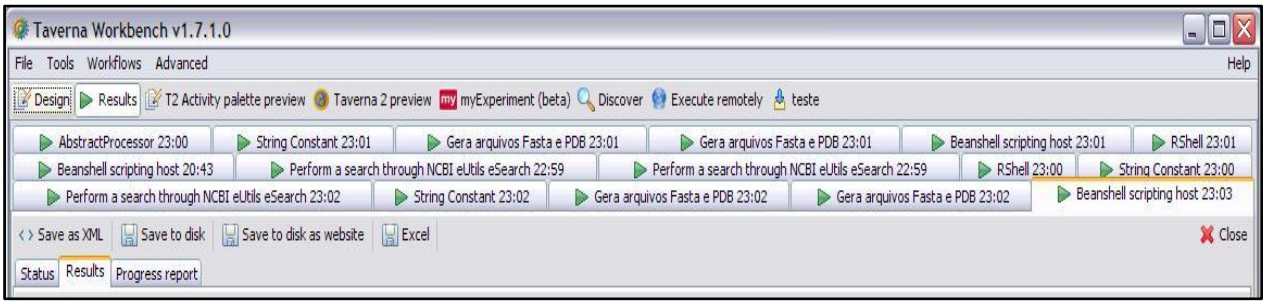

Figura 5.8: Acesso aos resultados dos workflows executados.

Para melhor entender as etapas realizadas na execução dos workflows, a seção 5.2 apresenta alguns exemplos práticos realizados ao longo do desenvolvimento do trabalho, e que utilizou-se como teste para validar as funcionalidades do workflow proposto.

5.3 Exemplos de Uso do Workflow

Nesta seção são demonstrados exemplos práticos utilizando os workflows propostos para o presente trabalho. Será apresentado uma situação específica em cada workflow, exemplificando o processo de execução, porém dependendo da necessidade do usuário e da disponibilidade dos serviços, a quantidade de entrada e a variedade de dados podem aumentar em um grande número.

O primeiro exemplo é do workflow para a busca do número de GI no NCBI, representado pela Figura 5.9. Nesta situação utilizou-se a opção *new input* inserindo informações para os parâmetros de entrada solicitados, neste caso é realizado uma busca pela expressão *glutathione*, nos bancos de dados de proteínas (*protein)* e atribuído o valor de 50, como sendo a quantidade máxima de registros listados no relatório em XML.

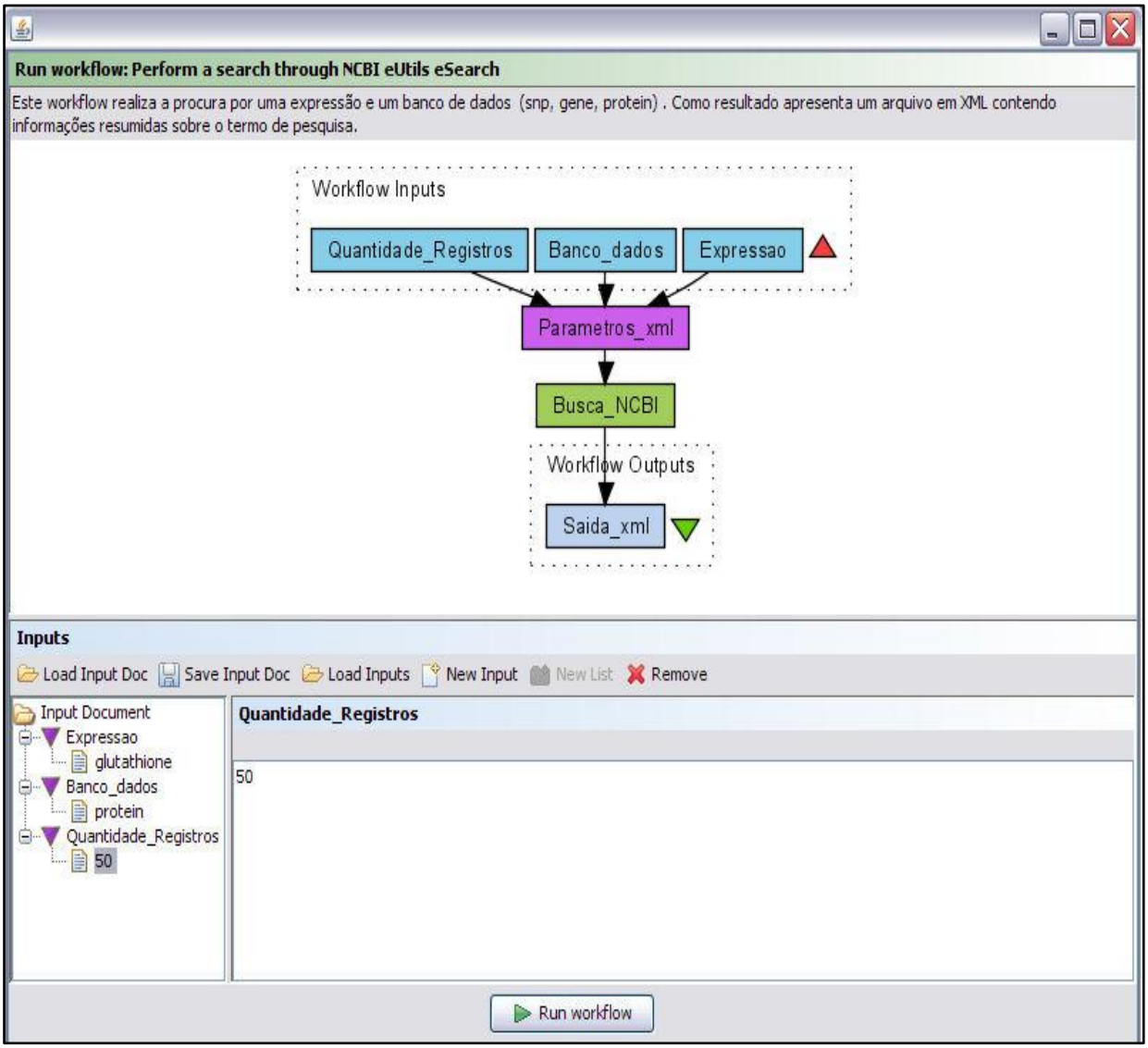

Figura 5.9: Parâmetros para executar o workflow “Busca NCBI”.

Após a execução do workflow utilizando os dados de entrada demonstrados na Figura 5.9, o resultado apresentado como saída está representado na Figura 5.10. Nesta interface é possível visualizar o número de GI de todos os registros pertencentes ao banco de dados de proteínas, e que contenham em seu interior a expressão “*glutathione*”.

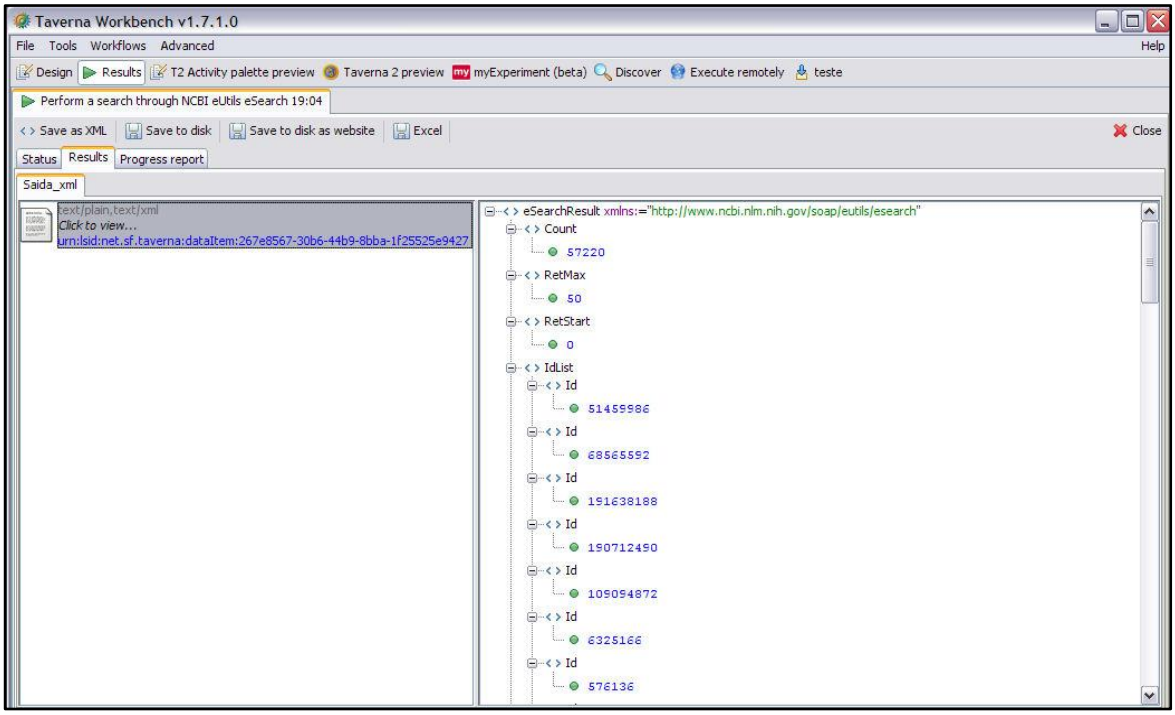

Figura 5.10: Resultado de saída para o workflow “Busca NCBI”.

O outro exemplo de workflow realiza a geração dos arquivos em formato Fasta e PDB. A interface parametrizando os dados de entrada, representado pela Figura 5.11, utiliza o número de GI “576136” e o número de ID do PDB “1glq”.

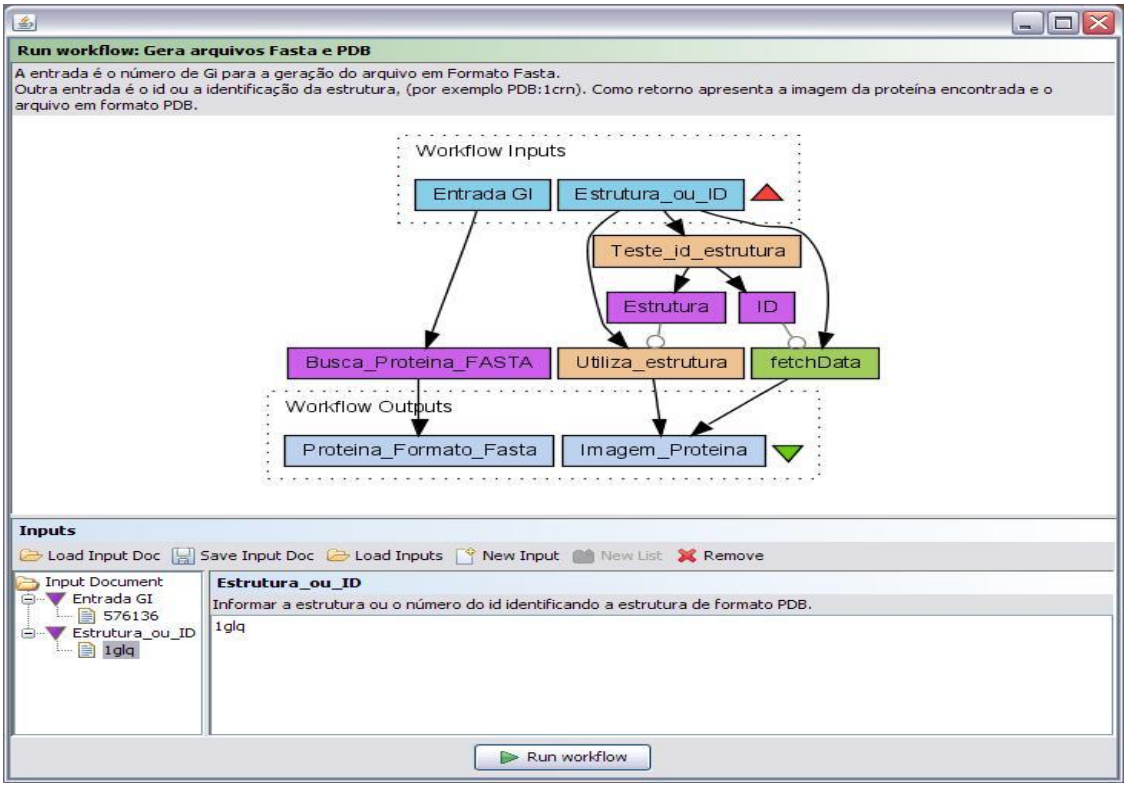

Figura 5.11: Parâmetros para executar o Workflow “Gera Arquivo Fasta PDB”.

Nas Figuras 5.12 e 5.13 pode-se visualizar os resultados de saída para o workflow “Gera Arquivo Fasta PDB”. A Figura 5.12 apresenta o arquivo em formato Fasta e a Figura 5.13 apresenta a imagem gráfica do arquivo no formato PDB.

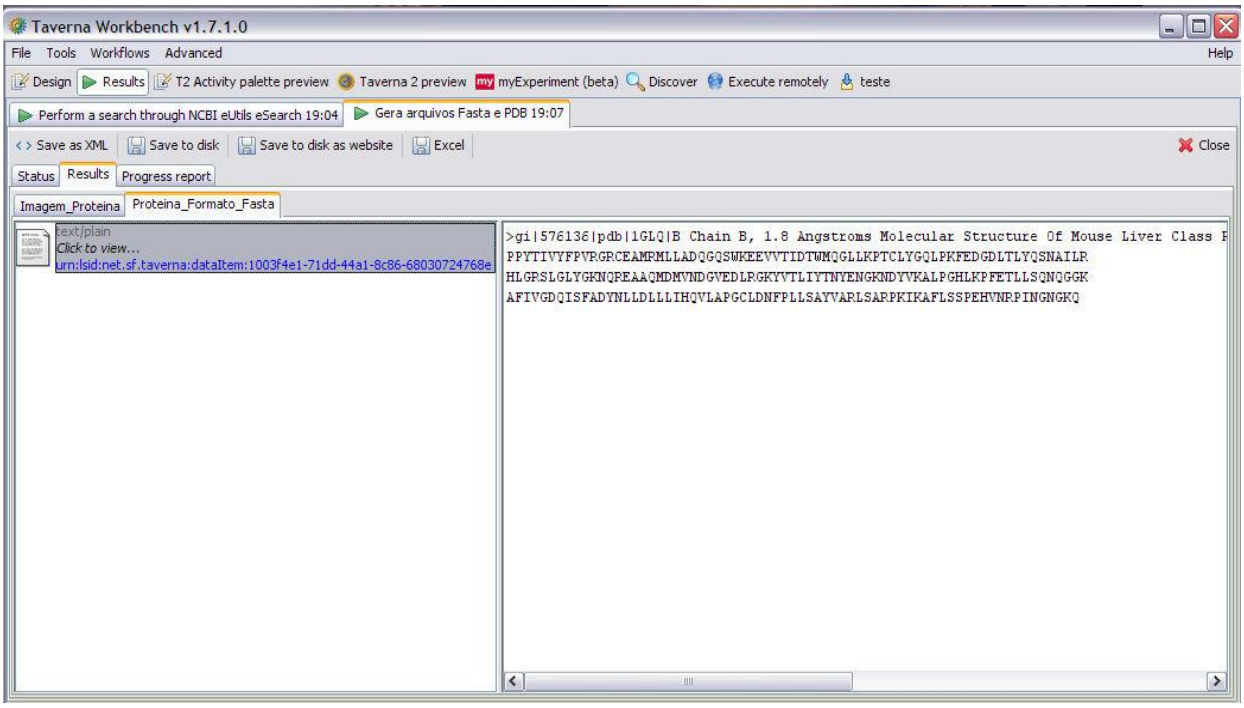

Figura 5.12: Resultado de saída Proteina_Formato_Fasta.

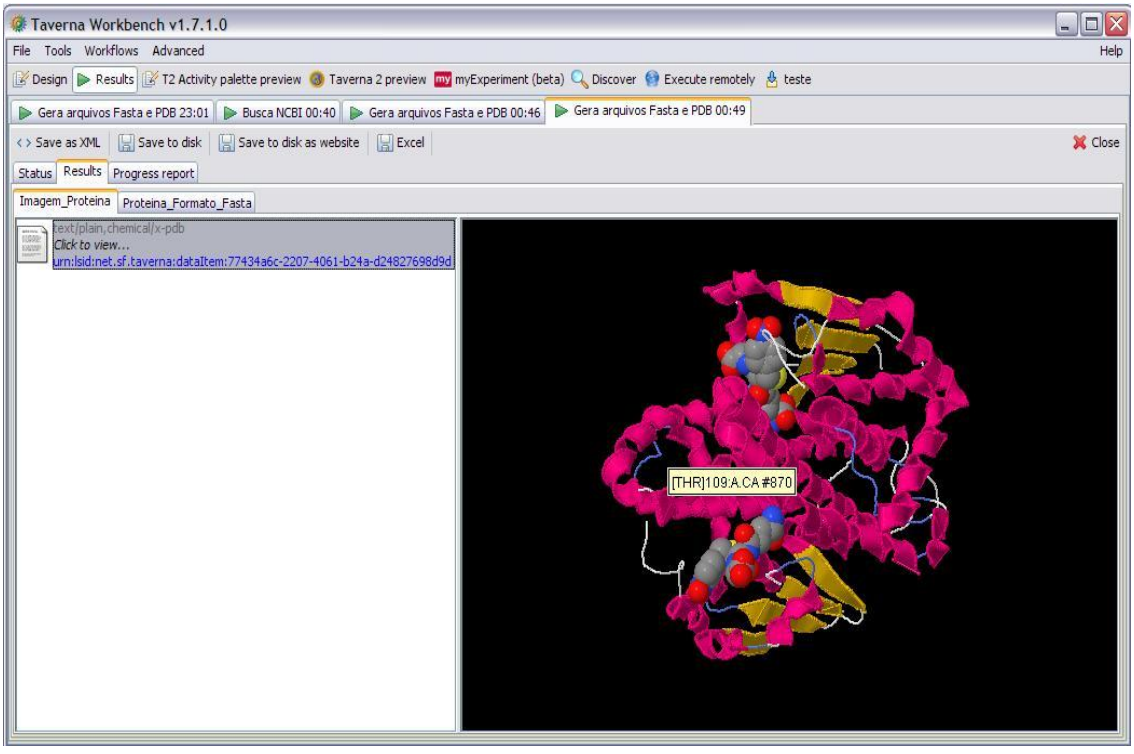

Figura 5.13: Resultado de saída Imagem_Proteina.

5.3 Avaliação do Especialista

Com o objetivo de aperfeiçoar o workflow proposto para contribuições futuras, bem como avaliar o grau de aceitação e satisfação dos especialistas, realizou-se uma avaliação com seis questões sobre a utilização do workflow para a bioinformática. O questionário permite descrever sugestões e comentários em todas as perguntas, três delas apresentam questões objetivas avaliando a contribuições para o laboratório, outras duas avaliam as vantagens e desvantangens em utilizar o workflow, e por fim uma questão descritiva solicitando sugestões de trabalhos futuros utilizando workflows científicos para a bioinformática.

A avaliação foi realizada em duas partes. Em um primeiro momento houve a apresentação do workflow no Laboratório de Biotecnologia Vegetal e Microbiologia Aplicada da Universidade de Caxias do Sul (UCS) com a presença de dois biólogos, e dos cientistas da computação, onde os cientistas apresentaram a ferramenta Taverna Workbench e o funcionamento dos workflows propostos para os biólogos. Para tanto houve questionamentos e sugestões de novos processos, que foram sendo identificados como rotina de trabalho por parte dos biólogos, e que poderiam ser utilizados através de workflows científicos.

Em uma segunda etapa solicitou-se a opinião dos biólogos através de um questionário com seis questões. A finalidade deste questionário é avaliar a contribuição do workflow para o laboratório, e validar o grau de satisfação por parte dos especialistas da biologia na utilização do workflow. O questionário compõe-se por cinco questões objetivas, com espaço para adicionar comentários, e uma questão discursiva permitindo sugerir idéias para o melhoramento do workflow proposto.

De forma geral o resultado da avaliação foi positivo, pois na maioria dos casos o workflow contribui muito para o laboratório e para os biólogos. Além, disso os resultados obtidos também contribuíram para o aumento de pesquisas e desenvolvimento de novos recursos para a bioinformática, com sugestões para o melhoramento e criação de novos processos para o workflow.

Os relatórios de avaliação com a respostas dos biólogos estão disponíveis no Anexo 2 deste trabalho.

6. CONSIDERAÇÕES FINAIS

O crescente volume de dados e processos utilizados na bioinformática torna cada vez mais fácil o desenvolvimento de novos recursos automatizados, provendo a eficiência na manipulação dos dados e auxiliando na descoberta de novas informações biológicas. Para tanto, os especialistas precisam cada vez mais da ajuda de sistemas que auxiliem os seus trabalhos, proporcionando uma melhor integração entre banco de dados e, uma eficiente interoperabilidade entre diferentes sistemas. O sistema formado por essas duas características denomina-se neste trabalho de workflow para bioinformática.

6.1 Conclusões

O workflow para bioinformática, construído com o auxilio da ferramenta Taverna Workbench, proposto neste trabalho teve por finalidade a geração de um arquivo em formato FASTA e outro arquivo em formato PDB, auxiliando nas pesquisas dos cientistas do laboratório de biotecnologia da UCS. Para a criação deste workflow houve a participação de dois especialistas que acompanharam todo o processo de análise, desenvolvimento, e por fim participaram da apresentação do workflow proposto. Após alguns testes no workflow os biólogos responderam a uma avaliação proposta pelo cientista da computação, tendo por finalidade contabilizar o grau de contribuição deste workflow no trabalho realizado pelos biólogos no laboratório de biotecnologia.

Com esta avaliação pode-se encontrar pontos positivos de grande contribuição para os biólogos, e pontos negativos que dificultam a utilização deste workflow e também de outras ferramentas utilizadas no laboratório.

Um dos pontos negativos, apontado como desvantagem por um dos biólogos, na utilização do workflow é a disponibilidade para acesso dos serviços utilizados pelo Taverna Workbech. Houveram algumas dificuldades encontradas pelos

especialistas quanto à demora na conexão dos serviços devido à velocidade da Internet atualmente existente no laboratório, e em algumas vezes ocorreu indisponibilidade dos serviços para acesso das ferramentas e dos bancos de dados. Essa última dificuldade depende exclusivamente dos bancos de dados públicos, uma vez que eles passam por processos de manutenções constantemente, bem como pode ocorrer falhas com interrupção dos serviços impossibilitando o acesso e utilização das informações.

Por outro lado o workflow proposto contribuiu muito em diversos processos realizados no laboratório, inclusive porque até o presente momento todas as atividades são realizadas de forma manual e com inúmeros sistemas, que são utilizados remotamente pela *web* ou por sistemas locais de acesso liberado e sem licença, além disso, a execução depende do conhecimento de quem realiza o processo. A criação deste sistema específico para o laboratório agradou muito os especialistas, que inclusive esperam a continuação deste trabalho adaptando outras funcionalidades a partir da utilização do arquivo em formato FASTA.

A geração do arquivo em formato FASTA, realizada por um dos workflows propostos neste trabalho, é de grande importância para os cientistas da biologia, pois além de conter seqüências genéticas de diferentes espécies, é reconhecido por inúmeros programas de bioinformática. Devido ao seu formato padronizado permite realizar o alinhamento de seqüências, a modelagem de estruturas e a descoberta de novas espécies genéticas. A figura 6.1 identifica alguns dos sistemas que interpretam o formato de arquivos FASTA.

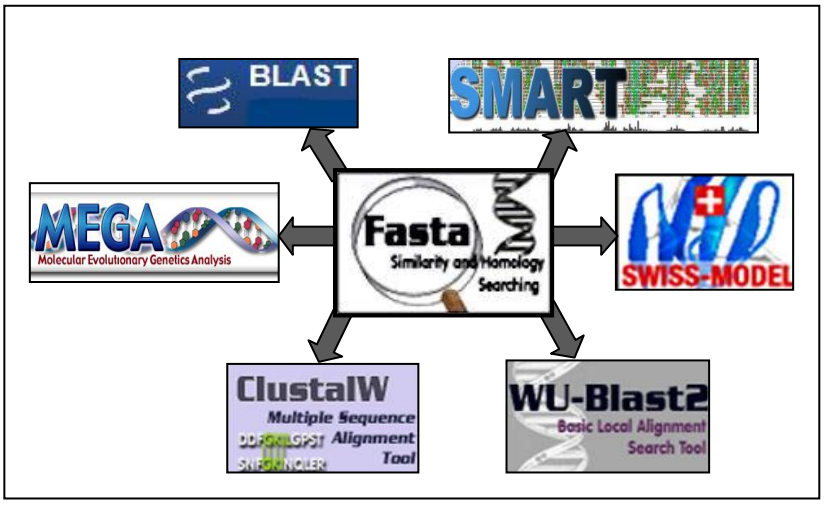

Figura 6.1 – Sistemas interpretadores do arquivo FASTA.

O presente trabalho contribui de forma significativa para auxiliar na execução dos processos na biologia, pois além de fornecer um formato de arquivo reconhecido mundialmente por todos os biólogos e programas de bioinformática, auxilia na construção de novos workflows científicos que realizam diferentes processos e contribuem de maneira eficiente no trabalho dos cientistas.

6.2 Trabalhos Futuros

Como trabalho futuro, pode-se apontar a utilização de outros serviços do Taverna Workbench para criar e adaptar novos workflows às atividades realizadas no laboratório. Alguns desses serviços poderiam auxiliar no armazenamento local dos resultados, possibilitando a inserção de anotações nestes arquivos e a posterior gravação em local padrão, pré configurado no workflow. Proporcionando desta forma a proveniência dos dados, realizando o acesso das informações de maneira instantânea, uma vez que o Taverna permite a leitura dos resultados deste workflow, e a visualização das anotações referentes às pesquisas realizadas e que estão armazenadas no arquivo.

Outro trabalho futuro seria o desenvolvimento de um workflow possibilitando a utilização da ferramenta BLAST do portal NCBI. Com esse recurso seria possível realizar a similaridade entre as seqüências genéticas de diferentes espécies, tanto de proteínas como de DNA. Além de executar a similaridade entre as espécies permitir a visualização de percentuais de semelhança, ou seja, os resultados de maior semelhança apresentam um número de maior percentual e os de menor semelhança um número de menor percentual. A criação deste workflow teria como entrada o número de GI ou o arquivo em formato Fasta de uma ou duas seqüências, e a identificação da espécie (proteína ou DNA). A execução do worklfow realizaria o processo de similaridade entre as espécies e a saída apresentaria os resultados em formato gráfico e texto, com o percentual de similaridade entre as espécies.

Por fim, outro trabalho futura poderia ser a utilização de serviços do *Swiss Model* para a criação de estrutura protéica tridimensionais. O *Swiss Model* acessa o banco de dados *Protein Data Bank* para realizar a consulta por similaridade de seqüências protéicas, utilizando a ferramenta BLASTP2 que realiza a modelagem por homologia automatizada de estruturas tridimensionais, e monta uma imagem a partir de um número pequeno de corpos rígidos, obtidos das estruturas de proteínas

alinhadas. O workflow criado para esse processo teria como entrada uma seqüência genética de proteínas no formato Fasta, a execução utilizaria os serviços do BLASTP2 e, realizaria a conexão com o banco de dados *ExPasy* do *Protein Data Bank.* Como resultado de saída apresentaria a imagem tridimensional da estrutura protéica, e o arquivo em formato PDB contendo as coordenadas para a criação da imagem.

7. REFERÊNCIAS

ALTINTAS, Ilkay; BARNEY, Oscar; JAEGER-FRANK, Efrat**. Provenance Collection Suport in the Kepler Scientific Workflow System**. In International Provenance and Annotation Workshop (IPAW), LNCS, Provenance and Annotation of Data, 4145: 118-132, 2006.

ALTSCHUL, S.F. et al. **Basic Local Alignment Search Tool – BLAST**. Journal Mol. Biol. 215: 403-440. 1990.

BARNES, Michael R.; GRAY, Ian C.. **Bioinformatics for Geneticists**. England : Wiley, 2003.

BERGERON, Bryan P. **Bioinformatics computing**. Upper Saddle River, NJ : Prentice Hall, c2003.

CITTON, Tais Adriana; STROGULSKI, Heitor. **Estudo sobre estrutura organizacional e modelagem de sistemas de Workflow**. Caxias do Sul, 2004. Monografia (Pós-Graduação) – Especialização em Novas Técnicas para o Desenvolvimento de Sistemas Computacionais. Universidade de Caxias do Sul, Monografia de Pós-Graduação, Caxias do Sul, 2004.

CORNERY, John S., CATCHEN, Julian M., LYNCH, Miahel. **Rule-based workflow management for bioinformatics.** The VLDB Journal, Spring, Verlag, 2005.

CRUZ, Tadeu. Workflow: **A tecnologia que vai revolucionar processos**. 2.ed. São Paulo : Atlas, 2000.

DIGIAMPIETRI, Luciano A. **Gerenciamento de workflows científicos em bioinformática**. Tese (Doutorado em Ciências da Computação) – UNICAMP, Campinas, 2007.

DNA: duplicação, tradução e transição. Disponível em: < http://turmadomario.com.br/cms/index.php/Conteudo/DNA.html>. Acessado em Agosto, 2008.

Ferramenta Blast, disponível em: <http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html>. Acessado em Setembro, 2008.

HOWISON, James; WIGGINS, Andrea; CROWSTON, Kevin. eResearch workflow for studying free and open source software development, Syracuse University School of Information Studies, Setembro de 2008.

KUNDE, Graziela F. **Evolução de Esquemas Conceituais em Workflow.** Dissertação (Mestrado em Ciências da Computação). Universidade Federal do Rio De Janeiro. Rio De Janeiro, Junho de 2000.

LARMAN, Craig. **Utilizando ULM e Padrões**. 2.ed. Porto Alegre : Bookman, 2004.

LEMOS, Melissa. **Workflow para Bioinformática**. Tese (Doutor em Informática). Pontifícia Universidade Católica (PUC). Rio De Janeiro, 2004.

LESK, Arthur M. **Introdução à Bioinformática**. 2.ed. Porto Alegre : Artmed, 2008.

LOPES, Sônia G. B. C. **Bio** – Volume Único. 6.ed. São Paulo : Saraiva, 1997.

LUDÄSCHER, Bertram; ALTINTAS, Ilkay; BERLEY, Chad; HIGGINS, Dan; JAEGER, Efrat; JONES, Matthew; LEE, Edward A.; TAO, Jing; ZHAO, Yang. **Scientific Workflow Management and the Kepler System.** Ed. Revisada., USA, 2005.

MATOSO, Marta; CRUZ, Sérgio M. S. da. **Gerência de workflows científicos: oportunidades e pesquisa em bancos de dados.** XXII Simpósio Brasileiro de Banco de Dados, Campinas, 13 a 17 de Outubro, 2008.

MENEZES, J. G. M. ; BAIÃO, Fernanda ; CAVALCANTI, M. C. **Uma Abordagem para Gerência de Dados Distribuídos em Workflows de Bioinformática.** In: XXII Simpósio Brasileiro de Banco de Dados e XXI Simpósio Brasileiro de Engenharia de Software, 2007, João Pessoa - Paraíba. VI WTDBD. João Pessoa : UFPB, 2007. v. 1.

MOTTA, Valter T., HESSELN, Ligia G., GIALDI, Silvestre. **Normas técnicas para apresentação de trabalhos científicos**. 3.ed.rev., atual. e ampl. Caxias do Sul : Educs, 2004.

OINN, Tom; ADDIS, Matthew; FERRIS, Justin; MARVIN, Darren; SENGER, Martin; GREENWOOD, Mark; CARVER, Tim; GLOVER, Kevin; POCOCK, Matthew R.;

WIPAT, Anil; LI, Peter. **Taverna: a tool for the composition and enactment of bioinformatics workflows**. Bioinformatics. Et. Al. Junho, 2004.

Portal EMBL, disponível em: <http://www.embl.org/aboutus/index.html>. Acessado em Setembro, 2008.

Portal NCBI, disponível em <http://www.ncbi.nlm.nih.gov/books/>. Acessado em Setembro, 2008.

PRESSMAN, Roger S. **Engenharia de software**. 6. Ed. São Paulo : McGraw-Hill, 2006.

Projeto Kepler. Disponível em: <http://kepler-project.org/>. Acessado em Setembro de 2008.

Projeto Ptolomeu. Disponível em <http://ptolemy.eecs.berkeley.edu/index.html>. Acessado em Outubro, 2008.

Projeto Taverna Workbench. Disponível em: <http://taverna.sourceforge.net/>. Acessado em Setembro de 2008.

Protein Data Bank, disponível em: <http://www.rcsb.org/pdb/>. Acessado em Setembro, 2008.

Science Creative Quarlety. Disponível em: <http://www.scq.ubc.ca/a-monks-flourishing-garden-the-basics-of-molecular-biology-explained/>. Acessado em Agosto, 2008.

Serviços BeanShell. Disponível em: < http://www.beanshell.org/intro.html>. Acessado em Novembro, 2008.

SILVA, Fabrício N. Da. **In Services: Um sistema para gerenciamento de dados intermediários em workflows científicos na bioinformática.** Dissertação (Mestrado em Sistemas e Computação). Instituto Militar de Engenharia. Rio de Janeiro, 2006.

SOARES, José L. **Biologia** – Volume Único. São Paulo : Scipione, 1991.

Tutorial Blast, disponível em: < http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/tut1.html>. Acessado em Outubro, 2008.

Tutorial ClustalX, disponível em: <http://virgil.ruc.dk/kurser/Sekvens/Treedraw.htm>. Acessado em Setembro, 2008.

Tutorial Diagramas de Colaboração. Disponível em <http://www.netbeans.org/>. Acessado em Outubro, 2008.

Tutorial Taverna. Disponível em <http://taverna.sourceforge.net/tutorial/TavernaTutorial.htm>. Acessado em Agosto, 2008.

Vistrails. Disponível em < http://www.vistrails.org/index.php/Main_Page>. Acessado em Outubro, 2008.

WfMC - Workflow Management Coalition, The Workflow Handbook 2004, Fischer,L.(ed.). Disponível em: <http://www.wfmc.org/information/handbook04.htm>. Acesso Setembro, 2008.

ANEXOS

ANEXO A – CÓDIGO FONTE DOS WORKFLOWS PROPOSTOS

ANEXO B – AVALIAÇÃO DOS ESPECIALISTAS

ANEXO A – CÓDIGO FONTE DOS WORKFLOWS PROPOSTOS

1. Worflow Busca NCBI

```
<?xml version="1.0" encoding="UTF-8"?>
<s:scufl xmlns:s="http://org.embl.ebi.escience/xscufl/0.1alpha" version="0.2" log="0">
  <s:workflowdescription lsid="urn:lsid:www.mygrid.org.uk:operation:QM8274VGE529" author="Renata De Paris"
title="Busca NCBI">Este workflow realiza a procura por uma expressão e um banco de dados  (snp, gene,
protein) . Como resultado apresenta um arquivo em XML contendo informações resumidas sobre o termo de
pesquisa.</s:workflowdescription>
  <s:processor name="Busca_NCBI">
    <s:arbitrarywsdl>
      <s:wsdl>http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/eutils.wsdl</s:wsdl>
      <s:operation>run_eSearch</s:operation>
    </s:arbitrarywsdl>
  </s:processor>
  <s:processor name="Parametros_xml">
    <s:local>
      org.embl.ebi.escience.scuflworkers.java.XMLInputSplitter
      <s:extensions>
        <s:complextype optional="false" unbounded="false" typename="eSearchRequest" name="parameters"
qname="eSearchRequest">
          <s:elements>
            <s:basetype optional="true" unbounded="false" typename="string" name="db" qname="string" />
            <s:basetype optional="true" unbounded="false" typename="string" name="term" qname="string" />
            <s:basetype optional="true" unbounded="false" typename="string" name="WebEnv" qname="string" />
            <s:basetype optional="true" unbounded="false" typename="string" name="QueryKey" qname="string" />
            <s:basetype optional="true" unbounded="false" typename="string" name="usehistory" qname="string" />
            <s:basetype optional="true" unbounded="false" typename="string" name="tool" qname="string" />
            <s:basetype optional="true" unbounded="false" typename="string" name="email" qname="string" />
            <s:basetype optional="true" unbounded="false" typename="string" name="field" qname="string" />
            <s:basetype optional="true" unbounded="false" typename="string" name="reldate" qname="string" />
            <s:basetype optional="true" unbounded="false" typename="string" name="mindate" qname="string" />
            <s:basetype optional="true" unbounded="false" typename="string" name="maxdate" qname="string" />
            <s:basetype optional="true" unbounded="false" typename="string" name="datetype" qname="string" />
            <s:basetype optional="true" unbounded="false" typename="string" name="RetStart" qname="string" />
            <s:basetype optional="true" unbounded="false" typename="string" name="RetMax" qname="string" />
            <s:basetype optional="true" unbounded="false" typename="string" name="rettype" qname="string" />
            <s:basetype optional="true" unbounded="false" typename="string" name="sort" qname="string" />
          </s:elements>
        </s:complextype>
      </s:extensions>
    </s:local>
  </s:processor>
  <s:link source="Banco_dados" sink="Parametros_xml:db" />
  <s:link source="Expressao" sink="Parametros_xml:term" />
  <s:link source="Quantidade_Registros" sink="Parametros_xml:RetMax" />
  <s:link source="Parametros_xml:output" sink="Busca_NCBI:parameters" />
  <s:link source="Busca_NCBI:parameters" sink="Saida_xml" />
  <s:source name="Expressao" />
  <s:source name="Banco_dados" />
  <s:source name="Quantidade_Registros" />
  <s:sink name="Saida_xml" />
</s:scufl>
```

2. Workflow Gera Arquivos Fasta e PDB

```
<?xml version="1.0" encoding="UTF-8"?>
<s:scufl xmlns:s="http://org.embl.ebi.escience/xscufl/0.1alpha" version="0.2" log="0">
  <s:workflowdescription lsid="urn:lsid:net.sf.taverna:wfDefinition:70d8a2a8-a369-4824-879d-dbfa36fcecc8"
author="Renata De Paris" title="Gera Arquivos Fasta e PDB">A entrada é o número de Gi para a geração do
arquivo em Formato Fasta. Outra entrada é o id ou a identificação da estrutura, (por exemplo PDB:1crn). Como
retorno apresenta a imagem da proteína encontrada e o arquivo em formato PDB.</s:workflowdescription>
  <s:processor name="Estrutura">
    <s:description>Fails if the workflow input is an identifier (i.e. is an actual structure).</s:description>
    <s:local>org.embl.ebi.escience.scuflworkers.java.FailIfFalse</s:local>
  </s:processor>
  <s:processor name="Busca_Proteina_FASTA">
    <s:local>net.sourceforge.taverna.scuflworkers.ncbi.ProteinFastaWorker</s:local>
  </s:processor>
  <s:processor name="ID">
    <s:description>Fails if the workflow input was a structure (i.e. is an identifer).</s:description>
    <s:local>org.embl.ebi.escience.scuflworkers.java.FailIfTrue</s:local>
  </s:processor>
  <s:processor name="Utiliza_estrutura">
    <s:description>Pass the input structure to the output.</s:description>
    <s:beanshell>
      <s:scriptvalue>out_structure = in_structure;</s:scriptvalue>
      <s:beanshellinputlist>
        <s:beanshellinput s:syntactictype="'text/plain'">in_structure</s:beanshellinput>
      </s:beanshellinputlist>
      <s:beanshelloutputlist>
        <s:beanshelloutput s:syntactictype="'text/plain'">out_structure</s:beanshelloutput>
      </s:beanshelloutputlist>
      <s:dependencies s:classloader="iteration" />
    </s:beanshell>
  </s:processor>
  <s:processor name="Teste_id_estrutura">
    <s:description>Return true if the input is a structure or false if the input is a structure identifer (e.g.
PDB:1crn).</s:description>
    <s:beanshell>
      <s:scriptvalue>lineLen = structure.indexOf("\n");
if(lineLen &amp;lt; 1) {
  lineLen = structure.length();
}
if(!structure.startsWith("ID ") &amp;amp;&amp;amp;
   structure.indexOf(":") &amp;gt; 0 &amp;amp;&amp;amp;
   structure.indexOf(":") &amp;lt; lineLen) {
  is_structure = "false";
} else {
  is_structure = "true";
}</s:scriptvalue>
      <s:beanshellinputlist>
        <s:beanshellinput s:syntactictype="'text/plain'">structure</s:beanshellinput>
      </s:beanshellinputlist>
      <s:beanshelloutputlist>
        <s:beanshelloutput s:syntactictype="'text/plain'">is_structure</s:beanshelloutput>
      </s:beanshelloutputlist>
      <s:dependencies s:classloader="iteration" />
    </s:beanshell>
  </s:processor>
  <s:processor name="fetchData">
    <s:description>Fetch the structure in PDB format from the identifer using EBI's WSDbfetch service (see
http://www.ebi.ac.uk/Tools/webservices/services/dbfetch).</s:description>
    <s:defaults>
      <s:default name="format">pdb</s:default>
      <s:default name="style">raw</s:default>
    </s:defaults>
    <s:arbitrarywsdl>
      <s:wsdl>http://www.ebi.ac.uk/Tools/webservices/wsdl/WSDbfetch.wsdl</s:wsdl>
      <s:operation>fetchData</s:operation>
    </s:arbitrarywsdl>
```

```
  </s:processor>
  <s:link source="Entrada GI" sink="Busca_Proteina_FASTA:id" />
  <s:link source="Estrutura_ou_ID" sink="Teste_id_estrutura:structure" />
  <s:link source="Estrutura_ou_ID" sink="Utiliza_estrutura:in_structure" />
  <s:link source="Estrutura_ou_ID" sink="fetchData:query" />
  <s:link source="Busca_Proteina_FASTA:outputText" sink="Proteina_Formato_Fasta" />
  <s:link source="Teste_id_estrutura:is_structure" sink="Estrutura:test" />
  <s:link source="Teste_id_estrutura:is_structure" sink="ID:test" />
  <s:link source="Utiliza_estrutura:out_structure" sink="Imagem_Proteina" />
  <s:link source="fetchData:fetchDataReturn" sink="Imagem_Proteina" />
  <s:source name="Entrada GI">
    <s:metadata>
      <s:mimeTypes>
        <s:mimeType>text/x-taverna-web-url</s:mimeType>
      </s:mimeTypes>
      <s:description>Informar o GI para a geração do arquivo em Formato Fasta.</s:description>
    </s:metadata>
  </s:source>
  <s:source name="Estrutura_ou_ID">
    <s:metadata>
      <s:description>Informar a estrutura ou o número do id identificando a estrutura de formato
PDB.</s:description>
    </s:metadata>
  </s:source>
  <s:sink name="Proteina_Formato_Fasta" />
  <s:sink name="Imagem_Proteina">
    <s:metadata>
      <s:mimeTypes>
        <s:mimeType>chemical/x-pdb</s:mimeType>
      </s:mimeTypes>
      <s:description>Estrutura em formato PDB.</s:description>
    </s:metadata>
  </s:sink>
  <s:coordination name="fetchData_BLOCKON_Fail_if_sequence">
    <s:condition>
      <s:state>Completed</s:state>
      <s:target>ID</s:target>
    </s:condition>
    <s:action>
      <s:target>fetchData</s:target>
      <s:statechange>
        <s:from>Scheduled</s:from>
        <s:to>Running</s:to>
      </s:statechange>
    </s:action>
  </s:coordination>
  <s:coordination name="Use_structure_BLOCKON_Fail_if_identifer">
    <s:condition>
      <s:state>Completed</s:state>
      <s:target>Estrutura</s:target>
    </s:condition>
    <s:action>
      <s:target>Utiliza_estrutura</s:target>
      <s:statechange>
        <s:from>Scheduled</s:from>
        <s:to>Running</s:to>
      </s:statechange>
    </s:action>
  </s:coordination>
</s:scufl>
```

ANEXO B – AVALIAÇÃO DOS ESPECIALISTAS

Nome: *Ana Paula Longaray Delamare*
Formação: *Bióloga/ microbiologia*
Atividade que realiza no Laboratório: *Pesquisadora*

1) Ao utilizar uma ferramenta de workflow o principal objetivo é otimizar e organizar processos, permitindo a integração de banco de dados e a interoperabilidade dos programas para bioinformática. O workflow proposto contribui para atingir esse objetivo?
   (X) Contribui muito
   ( ) Contribui parcialmente
   ( ) Não contribui

   Comentários: *Utilizando este workflow maximizamos o trabalho de análise de seqüência. Cabe lembrar que este foi o primeiro workflow utilizado, a partir deste poderemos criar novos programas, que serão utilizados no laboratório.*

2) O workflow proposto contribui de forma eficiente para a obtenção de arquivos em formato Fasta?
   (X) Contribui muito
   ( ) Contribui parcialmente
   ( ) Não contribui

   Comentários:

3) O workflow contribui para organizar a execução dos processos na obtenção do arquivo Fasta?
   (X) Contribui muito
   ( ) Contribui parcialmente
   ( ) Não contribui

   Comentários:

4) Foi encontrado vantagens em utilizar o workflow?
   (X) Sim. Quais? *Utilizar um programa remoto e não vários programas.*
   ( ) Não

5) Foi encontrado desvantagens em utilizar o workflow?
   (X) Sim. Quais? *Se a internet é lenta restringe a utilização do worKflow.*
   ( ) Não

6) Sugestões para o melhoramento do workflow proposto:

Ampliar a base de dados, no caso de estrutura protéica seria necessário desenhar o modelo tridimensional pelo SwissProt (escolher quantos modelos devem ser visualizados).

Nome: *Sergio Echeverrigaray*
Formação: *Doutor em Agronomia- Genética e melhoramento de Plantas*
Atividade que realiza no Laboratório: *Estudos genéticos e bioinformáticos em microrganismos e plantas.*

1) Ao utilizar uma ferramenta de workflow o principal objetivo é otimizar e organizar processos, permitindo a integração de banco de dados e a interoperabilidade dos programas para bioinformática. O workflow proposto contribui para atingir esse objetivo?
   (X) Contribui muito
   ( ) Contribui parcialmente
   ( ) Não contribui

   Comentários: *Um problema importante na área de bioinformática é a falta de interoperabilidade dos sistemas. Em geral, o trabalho é realizado de forma manual, mesmo quando se trata de uma seqüência de rotinas, mais ou menos estabelecidas. A integração de banco de dados, programas, e saída de resultados pode agilizar muito as análises*.

2) O workflow proposto contribui de forma eficiente para a obtenção de arquivos em formato Fasta?
   (X)Contribui muito
   ( ) Contribui parcialmente
   ( ) Não contribui

   Comentários: *Na sua versão atual o programa facilita a obtenção de arquivos FASTA. Eventualmente, algumas alterações são necessárias para facilitar ainda mais a procura, seleção e download destes arquivos.*

3) O workflow contribui para organizar a execução dos processos na obtenção do arquivo Fasta?
   ( ) Contribui muito
   (X) Contribui parcialmente
   ( ) Não contribui

   Comentários: *Eventualmente, algumas alterações são necessárias para facilitar ainda mais a a procura, seleção e download destes arquivos, especialmente quando se deseja obter diversos arquivos.*

4) Foi encontrado vantagens em utilizar o workflow?
   (X) Sim. Quais? *Facilidade de operação e organização dos processos.*
   ( ) Não

5) Foi encontrado desvantagens em utilizar o workflow?
   ( ) Sim. Quais?
   (X) Não

6) Sugestões para o melhoramento do workflow proposto:

Como dito anteriormente, as bases para utilização foram desenvolvidas. Atualmente é necessário realizar alguns ajustes específicos para download de arquivos FASTA a partir de analises de BLAST, e incorporação de novos programas de análises de seqüencias, tanto protéicas quanto de DNA. Acredito que com a base estabelecida será mais simples aprimorar o workflow.