MINISTÉRIO DA DEFESA
EXÉRCITO BRASILEIRO
DEPARTAMENTO DE CIÊNCIA E TECNOLOGIA
INSTITUTO MILITAR DE ENGENHARIA
CURSO DE MESTRADO EM SISTEMAS E COMPUTAÇÃO

KELE TEIXEIRA BELLOZE

UMA EXTENSÃO DO PROCESSO DE ANOTAÇÃO GENÔMICA PARA AMPLIAR O USO E A EVOLUÇÃO COLABORATIVA DE ONTOLOGIAS NO DOMÍNIO DA BIOLOGIA MOLECULAR

Rio de Janeiro
2007

INSTITUTO MILITAR DE ENGENHARIA

KELE TEIXEIRA BELLOZE

UMA EXTENSÃO DO PROCESSO DE ANOTAÇÃO GENÔMICA PARA AMPLIAR O USO E A EVOLUÇÃO COLABORATIVA DE ONTOLOGIAS NO DOMÍNIO DA BIOLOGIA MOLECULAR

Dissertação de Mestrado apresentada ao Curso de Mestrado em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de Mestre em Sistemas e Computação.

Orientadora: Profª Maria Claudia R. Cavalcanti – D.Sc.
Co-orientadora: Profª Renata Mendes Araujo – D.Sc.

Rio de Janeiro
2007

C2007

INSTITUTO MILITAR DE ENGENHARIA
Praça General Tibúrcio, 80 – Praia Vermelha
Rio de Janeiro - RJ CEP: 22290-270

004.33 Belloze, Kele Teixeira
B446 Uma Extensão do Processo de Anotação Genômica para Ampliar o Uso e a Evolução Colaborativa de Ontologias no Domínio da Biologia Molecular, Kele Teixeira Belloze.
- Rio de Janeiro: Instituto Militar de Engenharia, 2007.
144p.: il., graf., tab.

Dissertação: (mestrado) - Instituto Militar de Engenharia, Rio de Janeiro, 2007.

1. Bioinformática - anotação genômica.
2. Sistemas colaborativos - ontologia.
I. Instituto Militar de Engenharia. II. Título.

CDD 004.33

INSTITUTO MILITAR DE ENGENHARIA

KELE TEIXEIRA BELLOZE

UMA EXTENSÃO DO PROCESSO DE ANOTAÇÃO GENÔMICA PARA AMPLIAR O USO E A EVOLUÇÃO COLABORATIVA DE ONTOLOGIAS NO DOMÍNIO DA BIOLOGIA MOLECULAR

Dissertação de Mestrado apresentada ao Curso de Mestrado em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de Mestre em Sistemas e Computação.

Orientadora: Profª Maria Claudia Reis Cavalcanti – D.Sc.

Co-orientadora: Profª Renata Mendes Araujo – D.Sc.

Aprovada em 29 de junho de 2007 pela seguinte Banca Examinadora:

_____

Profª. Maria Claudia Reis Cavalcanti – D.Sc. do IME – Presidente

_____

Profª. Renata Mendes Araujo – D.Sc. da UNIRIO

_____

Prof. Alberto Martín Rivera Dávila – D.Sc. da FIOCRUZ

_____

Profª. Claudia Marcela Justel – D.Sc. do IME

_____

Profª. Maria Luiza Machado Campos – Ph.D. da UFRJ

Rio de Janeiro

2007

Dedico este trabalho à minha mãe.

AGRADECIMENTOS

À minha mãe Ilda Teixeira Belloze por todo o desprendimento quanto a minha ida para o Rio de Janeiro e toda a ajuda nos momentos em que o tempo era curto pra resolver questões rotineiras. Por todo o carinho dedicado a mim.

À minha irmã Vanessa Teixeira Belloze pela torcida. A ela, juntamente com seu marido Devanilson Jardel Visoná e a nossa "irmã de coração" Roseli Aparecida de Oliveira por todo apoio à minha mãe nos momentos de minha ausência.

À minha sobrinha, Tamirys de Oliveira Visoná pelos abraços e sorrisos mais sinceros quando das minhas voltas a Juiz de Fora.

À minha orientadora, Maria Claudia Reis Cavalcanti por conduzir de forma tão cuidadosa as orientações sobre este trabalho. Por todos os ensinamentos, ajudas, cobranças, paciência e amizade.

À minha co-orientadora, Renata Mendes Araujo, pelas palavras e decisões corretas nos momentos certos. Pela atenção dedicada a este trabalho, assim como os ensinamentos e amizade.

Aos professores Alberto Martín Rivera Dávila, Claudia Marcela Justel e Maria Luiza Machado Campos, por terem aceitado tão prontamente a fazerem parte de minha banca.

A todos os professores do curso de Mestrado em Sistemas e Computação do Instituto Militar de Engenharia pelo conhecimento repassado. Aos funcionários, por toda ajuda fornecida.

Ao professor Alberto Martín Rivera Dávila, pela paciência nas diversas explicações sobre biologia e disponibilidade para todas as reuniões, inclusive as marcadas repentinamente.

À Priscila, Glauber e Henrique, que conheci na FIOCRUZ, pelo apoio, disponibilidade para reuniões, diversas ajudas sobre conteúdos do sistema GARSA, dúvidas relativas a área de biologia e amizade conquistada.

Ao Felipe Albrecht por todo o apoio no desenvolvimento do protótipo deste trabalho.

A todos os amigos do mestrado. Especialmente Cardoso, Flavio, Francis e Rafael pelo grupo de estudo, apoio em todos os momentos, conversas na madrugada e os "chopps". Ao Thiago por sempre estar ao meu lado, e que junto com Marcelo tornaram o ano de disciplinas menos pesado, com todas as brincadeiras e momentos de risadas.

À amiga Viviane por todo o apoio e amizade desde a época da graduação e que juntamente com Luciana e Saulo tornaram minha ida para o Rio de Janeiro facilitada.

Às amigas de morada no Rio de Janeiro, Lourdes e Dila, pela compreensão nos momentos da necessidade de silêncio e concentração. Por todos os almoços, cachorros-quentes, pipocas, conversas. Pela grande amizade conquistada.

Ao amigo Dáves por sempre encararmos juntos os mesmos desafios. E a ele, Ana, Bruna e Fabiano por todo o carinho, mesmo nos momentos em que fui ausente.

Aos amigos Renato, Rodrigo e Thiago por entrarem na minha vida num momento em que eu precisava de alegria e descontração.

A todos os amigos da graduação que sempre torceram por mim.

Aos amigos, conhecidos e parentes que apoiaram e também torceram por esta conquista.

Aos professores do Departamento de Ciência da Computação da Universidade Federal de Juiz de Fora por todo o apoio para a conquista deste título. Em especial Regina Maria Maciel Braga, Raul Fonseca Neto e Rubens de Oliveira pelo incentivo e amizade.

À professora Maria Clicia Stelling de Castro por sempre acreditar no meu trabalho.

Ao professor Marco Antônio Pereira Araújo por ser o grande e primeiro incentivador para eu ingressar no mestrado.

A todos que de alguma forma contribuíram para a realização deste trabalho.

À CAPES pelo apoio financeiro concedido.

SUMARIO

LISTA DE ILUSTRAÇÕES 10

LISTA DE TABELAS 12

LISTA DE SIGLAS 13

1 INTRODUÇÃO 17

1.1 Motivação 18

1.2 Caracterização do Problema 19

1.3 Visão Geral da Proposta 20

1.4 Contribuições 21

1.5 Organização do Trabalho 22

2 ANOTAÇÃO GENÔMICA 24

2.1 Bancos de Dados Genômicos 25

2.1.1 GenBank 28

2.1.2 Pfam 30

2.2 Anotação 32

2.3 Sistemas de Anotação 33

2.3.1 Apollo 34

2.3.2 Artemis 36

2.3.3 BioNotes 37

2.3.4 DAS 39

2.3.5 GARSA 40

2.3.6 Comparação entre os Sistemas de Anotações 40

2.4 Ontologias para Bioinformática 43

2.4.1 Gene Ontology 43

2.4.2 Sequence Ontology 46

2.5 Considerações 48

3 O PROCESSO DE ESTUDO DE SEQÜÊNCIAS E ANOTAÇÕES 49

3.1 Workflows para Estudo de Pesquisas Genômicas 50

3.1.1 Workflow MHOLLINE 51

3.1.2 Workflow do Projeto GENOPAR 52

3.1.3 Workflow SABIA 53
3.1.4 Workflow GARSA 54
3.2 Modelagem do Processo de Estudo de Seqüências Genômicas 56
3.2.1 Modelagem de Processos 56
3.2.2 Papéis 57
3.2.3 Bancada Molhada (*In-Vitro*) 58
3.2.4 Bancada de Experimentos *In Silico* 58
3.2.5 A Ontologia como Recurso no Processo de Anotação Genômica 67
3.3 Anotação no GARSA 68
3.3.1 Recebimento das Seqüências e Pré-análise 69
3.3.2 Análise das Seqüências 69
3.3.3 Anotação 74
3.4 Considerações 78

4 EXPANDINDO O USO DA ONTOLOGIA NO PROCESSO DE ANOTAÇÃO GENÔMICA 80
4.1 Colaboração na Bioinformática 81
4.1.1 MyGrid 82
4.1.2 TCruziDB 82
4.1.3 Consórcio Gene Ontology 83
4.1.4 Consórcio BioWebDB 84
4.2 Cenários de Colaboração e Papéis 84
4.3 Extensão do Processo de Anotação Genômica com Foco no Uso da Ontologia 85
4.4 Considerações 95

5 GARSA NOTES 97
5.1 Visão Geral do Protótipo 97
5.2 Funcionalidades do GARSA Notes 101
5.2.1 Termo Encontrado 101
5.2.2 Termo Não Encontrado 108
5.2.3 Geração de Relatórios 110
5.3 Modelo de Dados 113
5.4 Ambiente de Desenvolvimento 116
5.5 Considerações Referentes ao GARSA e GARSA Notes 117

6 OBSERVAÇÃO SOBRE O USO DO GARSA NOTES 119

6.1 Exemplo – Projeto *P.serpens* 119
6.1.1 Exemplo 1 120
6.1.2 Exemplo 2 123
6.2 Considerações 125
7 CONCLUSÃO 127
7.1 Contribuições 129
7.2 Melhorias e Trabalhos Futuros 130
7.2.1 Melhorias na Proposta 130
7.2.2 Melhorias no GARSA Notes 131
7.2.3 Trabalhos Futuros 131
8 REFERÊNCIAS BIBLIOGRÁFICAS 133
9 ANEXOS 143

LISTA DE ILUSTRAÇÕES

FIG. 2.1 Crescimento *GenBank* I (GENBANK, 2005) 29
FIG. 2.2 Crescimento *GenBank* II (GENBANK, 2005) 29
FIG. 2.3 Crescimento PfamA e PfamB (PFAM, 2006a) 31
FIG. 2.4 Tabela de descrição de *feature* (APOLLO USER GUIDE, 2005) 35
FIG. 2.5 Resultado da análise de programas no Artemis (ARTEMIS EXAMPLES, 2005) 36
FIG. 2.6 Tela para inserir uma anotação manual no BioNotes (BIONOTES, 2005) 38
FIG. 2.7 Arquitetura básica do DAS (DOWELL et. al., 2001) 39
FIG. 2.8 Anotação no sistema GARSA (GARSA, 2007) 40
FIG. 2.9 Exemplo da Gene Ontology (THE GENE ONTOLOGY CONSORTIUM, 2000) 45
FIG. 2.10 Relacionamentos entre termos da SO (EILBECK et al, 2005) 47
FIG. 3.1 Seqüências do *Wokflow MholLIne* (SANTOS, 2004) 51
FIG. 3.2 Workflow do Projeto GENOPAR (GENOPAR, 2005) 53
FIG. 3.3 Visão geral do workflow SABIA 54
FIG. 3.4 Workflow do GARSA (GARSA, 2005) 55
FIG. 3.5 Visão geral dos processos para estudo de seqüências genômicas 59
FIG. 3.6 Atividades da pré-análise 61
FIG. 3.7 Atividades da análise das seqüências 62
FIG. 3.8 Atividades da anotação 64
FIG. 3.9 Atividades da consolidação da anotação 67
FIG. 3.10 Predição de genes 70
FIG. 3.11 Característica do gene predito 71
FIG. 3.12 Resultados do programa BLAST 72
FIG. 3.13 Resultado da análise de similaridade cluster TGEG101003D05.g 73
FIG. 3.14 Resultado da anotação pelos termos da GO 75
FIG. 3.15 Interface de anotação das seqüências 76
FIG. 3.16 Dados estatísticos 77
FIG. 4.1 Processo de anotação genômica com foco no uso da ontologia 94
FIG. 4.2 Ciclo produtivo de anotação 96
FIG 5.1 Buscas de seqüências (GARSA, 2007) 98

FIG. 5.2 Edição das anotações..... 99
FIG. 5.3 Módulos do GARSA Notes..... 100
FIG. 5.4 Análise e classificação do termo encontrado 101
FIG. 5.5 Consulta pesquisador – perguntas..... 102
FIG. 5.6 E-mail enviado ao pesquisador especialista..... 103
FIG. 5.7 Consulta pesquisador – respostas..... 103
FIG. 5.8 Busca na literatura..... 104
FIG. .5.9 Cadastro de links de interesse 105
FIG. 5.10 Pesquisa a bancos de dados específicos..... 105
FIG. 5.11 Cadastro de bancos de dados 106
FIG. 5.12 Revisão de Similaridade..... 106
FIG. 5.13 Discussão entre pesquisadores 107
FIG. 5.14 Referência a outros itens 108
FIG. 5.15 Funcionalidades para termo não encontrado..... 109
FIG. 5.16 Opções para geração de relatório para termo encontrado 110
FIG. 5.17 Relatório gerado para termo encontrado..... 111
FIG. 5.18 Opções para geração de relatório para termo não encontrado 112
FIG. 5.19 Relatório gerado para termo não encontrado 112
FIG. 5.20 Opção de relatórios gerais..... 113
FIG. 6.1 Edição da anotação do cluster PSADSAU001A12.g (GARSA,2007)..... 120
FIG. 6.2 Revisão de similaridade para o termo ‘phosphatidylinositol 3-kinase activity’ 121
FIG. 6.3 Pesquisa a bancos de dados para o termo ‘response to nutrient’ 122
FIG. 6.4 Informações sobre o termo não encontrado 122
FIG. 6.5 Edição da anotação do cluster PSADEST001H06.b (GARSA,2007)..... 123
FIG. 6.6 Busca na literatura para o termo ‘ATP binding’ 124
FIG. 6.7 Busca na literatura para o termo ‘protein folding’ 124
FIG. 6.8 Informações sobre o termo encontrado..... 125

LISTA DE TABELAS

TAB. 2.1 Comparação entre sistemas de anotações 42
TAB. 2.2 Conteúdo da GO (THE GENE ONTOLOGY CONSORTIUM, 2006) 45

LISTA DE SIGLAS

ACT	Artemis Comparison Tool
BBOP	Berkeley Bioinformatics and Ontologies Project
BDGP	Berkeley Drosophila Genome Project
BLAST	Basic Local Alignment Search Tool
CDS	Coding Sequences
CGI	Common Gateway Interface
CNPq	Conselho Nacional de Desenvolvimento da Pesquisa
DAS	Distributed Annotation System
DDBJ	DNA Data Bank of Japan
DNA	DeoxyriboNucleic Acid
DTD	Document Type Definition
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
EPSRC	Engineering and Physical Sciences Research Council
EST	Expressed Sequence Tags
Fiocruz	Fundação Oswaldo Cruz
FPT	File Transfer Protocol
FT	Feature Table
GAME	Genome Annotation Markup Elements
GARSA	Genomic Analysis Resources for Sequence Annotation
GFF	General Feature Format
GMOD	Generic Model Organism Database
GUS	Genomics Unified Schema
IOC	Instituto Oswaldo Cruz
mRNA	messenger RiboNucleic Acid
NCBI	National Center for Biotechnology Information
NCBO	National Center for Biomedical Ontology
ORF	Open Reading Frame
ORG	Ontology Research Group
PERL	Practical Expert Report Language
RNA	RiboNucleic Acid
SABIA	System for Automated Bacterial Integrated Annotation
SGBD	Sistema Gerenciador de Banco de Dados

TIGR	The Institute for Genomic Research
tRNA	transfer RiboNucleic Acid
URL	Uniform Resource Locator
XML	eXtensible Markup Language

RESUMO

A anotação genômica é uma das tarefas mais importantes na área de pesquisa genômica. Uma anotação é o registro do significado biológico dos genes identificados nas seqüências genômicas. Contudo, em geral a anotação é feita através de um vocabulário do próprio anotador ou do grupo de pesquisa. Isto pode dificultar a troca de informações entre pesquisadores do mesmo grupo, entre projetos parceiros e pesquisadores interessados na pesquisa em questão, e consequentemente prejudicar a evolução da pesquisa. Para resolver este problema, alguns grupos de pesquisa genômica têm feito o uso da anotação baseada em ontologia. Embora o uso de ontologias seja crescente na comunidade de Bioinformática, estas ainda não acompanham a contento as descobertas genômicas. Tanto as falhas quanto os sucessos no uso destas ontologias são identificados no contexto dos projetos de pesquisa genômica. Entretanto, falhas e sucessos raramente são reportados aos desenvolvedores/curadores das ontologias.

Esta dissertação propõe uma extensão do processo de anotação genômica para capturar o raciocínio feito sobre a anotação baseada em ontologia. Por exemplo, as ações envolvidas na confirmação dos termos das ontologias, revelando as razões pela qual um termo foi identificado como adequado ou não, ou os problemas existentes, ou ainda possíveis sugestões para quando um termo não foi encontrado. O registro deste raciocínio contribui para a ampliação do uso e evolução da ontologia, e consequentemente, ampliando-se o uso de ontologias, facilita-se as colaborações inter e intra projetos. Um protótipo denominado GARSA Notes, o qual possui as funcionalidades necessárias para apoiar a proposta de extensão foi desenvolvido e anexado ao GARSA, um sistema de anotação em uso por um grupo de pesquisa da FIOCRUZ.

ABSTRACT

Genomic annotation is one of the main concerns of genomic research. An annotation is the registry of the biological meaning of a genoma sequence region. However, free textual annotations do not help results share among researchers and co-projects, and consequently, impairs research evolution. To address this problem, several genomic research groups have invested on ontology-based annotation. Although the use of ontologies has become popular within the Bioinformatics community, these ontologies evolve very slowly, and do not cope with the pace of the biological findings. Most of the ontology failures and successes are identified in the context of genomic research projects. However, they are rarely reported to the ontology designers/curators.

This work proposes a mechanism to capture the ontology-based annotation rationale, i.e., the actions involved in the ontology term choice, revealing the reasons why a term was discarded or successfully used, or existent problems or even suggestions when a term is not found. The registry of the rationale contributes to the use and evolution of the ontology, and consequently it facilitates “inter and intra” projects collaboration. A prototype named GARSA Notes, which provides the functionalities to support the proposed process extension was developed and integrated into GARSA, an annotation system in use for some years by a FIOCRUZ research group.

1 INTRODUÇÃO

Os projetos genoma possuem como objetivos a descoberta e a descrição de genes. Os principais referem-se aos genomas microbianos, de plantas e humano (VASCONCELOS, 2003). Para alcançar estes objetivos, além da realização de diversos experimentos biológicos, é feito o uso de computadores para apoiar estas tarefas de descoberta e descrição de genes. Assim, o desenvolvimento de projetos genoma aliado à tecnologia dos computadores, fez surgir uma área conceituada como Bioinformática, que atualmente é uma das áreas que mais evoluem.

De acordo com WESTHEAD et al. (2002), os computadores são importantes na Bioinformática por duas razões. Primeiro, porque muitos problemas na Bioinformática requerem que a mesma tarefa seja realizada diversas vezes, por exemplo, comparar uma nova seqüência genômica com outras seqüências de mesmo tipo armazenadas em banco de dados a fim de descobrir similaridades. Nestes casos, a funcionalidade dos computadores para processar as informações é indispensável. Segundo, porque os computadores através de *softwares* específicos são requisitados para resolver problemas como inferir o que cada seqüência de nucleotídeos ou aminoácidos representa e interpretar biologicamente esses dados.

A experimentação apoiada pelo uso dos computadores é denominada na comunidade científica por experimentos *in silico*. Nos projetos genoma, uma das tarefas mais importantes é a interpretação dos dados experimentais gerados, a partir de experimentos *in silico*, para se obter o conhecimento biológico e assim identificar o que representa cada uma das seqüências que compõem o genoma de um organismo. O processo de identificação dessas seqüências é denominado de anotação genômica.

A anotação genômica é um processo que depende não somente dos resultados gerados pelos experimentos *in silico*, mas também da experiência dos pesquisadores em observar os resultados e inferir o que neles pode conter. Os centros de pesquisa no Brasil e no mundo que possuem como foco o estudo do genoma de determinados organismos como – Instituto Oswaldo Cruz da Fundação Oswaldo Cruz (IOC/FIOCRUZ, 2007), Laboratório de Bioinformática do Laboratório Nacional de Computação Científica (LNCC) (LABINFO, 2007), National Center for Biotechnology Information (NCBI, 2007), European Bioinformatics Institute (EBI, 2007), envolvem diversos pesquisadores, denominados bioinformatas, os quais em geral trabalham no contexto de um grupo de pesquisa. Existem também centros de pesquisas como (NIG,

2007), EBI e NCBI, que são parceiros de outros e assim compartilham trabalhos e resultados.

Após a publicação em eventos ou revistas específicos da área de genômica sobre as descobertas dos genes das seqüências estudadas, os grupos e centros de pesquisa disponibilizam suas descobertas juntamente com as respectivas anotações em bancos de dados públicos, tornando-as acessíveis para toda a comunidade interessada na descoberta em questão.

1.1 MOTIVAÇÃO

Em geral, a anotação genômica é feita de acordo com um vocabulário que o próprio grupo de pesquisa ou o bioinformata sozinho estão acostumados. Isto se torna um problema, pois dificulta o entendimento daqueles que venham a acessar esta anotação, como: bioinformatas que colaboram entre si na realização de seus trabalhos, grupos de pesquisas parceiros, centros de pesquisas mantenedores de bancos de dados públicos e a comunidade de Bioinformática interessada nos resultados anotados.

Para contornar este problema, além da anotação feita através do vocabulário próprio do grupo de pesquisa, passou-se a fazer uso de ontologias para uniformizar o vocabulário para anotar. As ontologias constituem-se de modelos de dados, que representam determinados conceitos ou conjunto de termos de um domínio e o relacionamento entre eles (GRUBER, 1993). Estes conceitos formam um vocabulário controlado e comum a todos que o utilizarem. Na Bioinformática, a ontologia mais disseminada é a Gene Ontology (GO) (THE GENE ONTOLOGY CONSORTIUM, 2000).

Utilizar esta ou outras ontologias para fazer a anotação genômica possibilita um maior entendimento entre bioinformatas e o seu grupo de pesquisa, permite uma divulgação mais abrangente dos resultados em periódicos específicos da área, facilita a troca de informações entre projetos de pesquisa de todo o mundo e permite o desenvolvimento de trabalhos colaborativos, já que o vocabulário torna-se comum entre todos.

Sendo observada essa necessidade de manter além da anotação usual do grupo, também a anotação através de ontologias, os centros de pesquisas já têm incluído nos fluxos de processos (*workflows*) de seus projetos de pesquisa genômica, suporte à anotação através dos termos das ontologias. Sistemas como ARTEMIS

(RUTHERFORD et. al., 2000), GARSA (DÁVILA et. al., 2005) e SABIA (ALMEIDA, et. al., 2004), fazem o uso da Gene Ontology.

Uma das maneiras de promover e estabelecer a colaboração entre os membros de um grupo ou pesquisadores externos é usar um vocabulário controlado, como as ontologias. Porém o processo de anotação genômica não engloba uma verificação mais detalhada sobre a anotação baseada em ontologia.

1.2 CARACTERIZAÇÃO DO PROBLEMA

Observa-se que a anotação através dos termos das ontologias ainda não se mostra como sendo a principal, devido a ocorrência de problemas que envolvem as ontologias. Verificam-se na literatura diversos problemas encontrados pela comunidade de Bioinfomática no uso das mesmas, como os relacionados à Gene Ontology (DAVID, et al., 2002), (SMITH et al., 2003). Alguns destes problemas se referem à definição de termos (SMITH et al., 2004). Se os termos possuem problemas eles podem não representar a seqüência genômica de forma ideal e assim acarretar um baixo número de seqüências que conseguem ser anotadas desta forma.

No contexto de sistemas de anotação genômica, há duas formas principais de reportar problemas para os mantenedores das ontologias. A primeira e mais comum é através de canais de comunicação, que os próprios mantenedores oferecem. Por exemplo, as contribuições feitas à GO acontecem principalmente através de listas de discussões ou do *Curator Requests Tracker* [1] disponibilizado pelo Consórcio.

A segunda forma de contribuição é através de colaborações de centros reconhecidos e oficialmente comprometidos com a evolução da Ontologia, como os associados à GO. Uma das funções principais destes centros é prover a GO com novas anotações para o seu banco de dados.

Nestas duas abordagens a proposta de inclusão/atualização de termos é dissociada do processo de anotação. Desta forma, qualquer problema referente aos termos, assim como esclarecimento de dúvidas e sugestões devem ser tratadas à parte, junto aos consórcios mantenedores das ontologias. Logo, o repasse dos problemas ou sugestões fica condicionado ao interesse e disponibilidade dos bioinformatas, podendo então ser realizado somente em momentos futuros, o que frequentemente ocasiona uma perda de

[1] http://geneontology.sourceforge.net

conteúdo ou detalhamento, causando assim dificuldades na colaboração do grupo de pesquisa e na colaboração à própria ontologia.

Nos sistemas Artemis, GARSA e SABIA, a discussão e registro dos termos encontrados da Ontologia não ocorrem. Qualquer tipo de colaboração no sentido de reportar informações aos curadores das ontologias fica sujeita ao interesse e disponibilidade dos bioinformatas, sendo necessário que estes se preocupem em registrar tais informações à parte.

1.3 VISÃO GERAL DA PROPOSTA

A hipótese deste trabalho diz que se ficar bem caracterizado o processo de anotação genômica e a colaboração inerente a ele, é possível sugerir soluções de apoio a colaboração para ampliar o uso e a evolução de ontologias deste domínio, compartilhando conhecimento, estimulando o processo de anotação baseado em ontologia e melhorando a qualidade das anotações.

Logo, neste trabalho, é proposta uma extensão do processo de anotação genômica com vistas a ampliar o uso das ontologias. Inicialmente, foi necessário caracterizar o ambiente de anotação genômica, fazendo o reconhecimento de todas as atividades e papéis envolvidos no processo, a identificação das necessidades de colaboração e a forma como as ontologias são utilizadas.

Os experimentos *in silico* para o estudo das seqüências genômicas e consequentemente para a anotação genômica podem ser vistos como a instanciação de um processo que envolve diversas atividades e que devem ser realizadas levando-se em consideração uma ordem, independentemente do foco principal de cada centro de pesquisa.

Representar estas atividades através de um modelo permite visualizar e entender de forma mais ampla todo o processo. Além do modelo do processo, a descrição detalhada de cada atividade favorece também o entendimento do processo, e juntos possibilitam identificar a colaboração inerente ao processo. Através do trabalho colaborativo, é possível compartilhar experiências e conhecimentos entre os membros do grupo, permitindo ampliar as suas habilidades em produzir anotações de melhor qualidade, e consequentemente, obterem melhores resultados.

Com o levantamento das informações obtidas através de pesquisas sobre a forma de trabalho de alguns centros de pesquisas (SANTOS, 2004), (GENOPAR, 2005)

(ALMEIDA et al, 2004), (DÁVILA et. al., 2005) e de reuniões e observações feitas com um grupo de pesquisa genômica específico, tornou-se viável a construção de modelos de processos os quais demonstram todas as atividades desempenhadas para o estudo das seqüências genômicas e consequentemente da anotação genômica. Este modelo representa os processos de forma genérica e a partir dele foi possível propor extensões a este processo que ampliem a colaboração e o uso de ontologias em sua execução.

A extensão proposta permitiu identificar os problemas e registrar o raciocínio desempenhado pelos pesquisadores em relação à anotação baseada em ontologia, no momento em que eles de fato ocorrem, evitando a perda desta valiosa informação. Desta forma, é possível reportar tais problemas e colaborar com os curadores das ontologias, de modo a agilizar a evolução destas, e consequentemente, ampliar o seu uso. O registro do raciocínio também proporciona anotações futuras de forma mais rápida e fácil, e auxilia os novos pesquisadores em suas anotações.

A partir da especificação proposta foi desenvolvido um protótipo, o GARSA Notes, que além de permitir a monitoração da anotação baseada em ontologia, facilita também a geração de documentos contendo as informações sobre esta monitoração e sugestões caso existam. O protótipo compreende uma extensão ao sistema GARSA para a análise e anotação genômica, de forma a avaliar a viabilidade de seu suporte ao registro do raciocínio no uso da ontologia durante a execução do processo de anotação. O GARSA foi utilizado, devido ao centro de pesquisa desenvolvedor deste sistema ser o ambiente de observação e apoio para a realização deste estudo.

1.4 CONTRIBUIÇÕES

Dentre as contribuições deste trabalho podemos citar:

- Caracterização através de modelagens e descrições do processo de estudos de seqüências genômicas e consequentemente do processo de anotação genômica de forma genérica;
- Sugestão de formas para ampliação da colaboração no processo de anotação genômica;
- Identificação dos problemas relacionados à anotação baseada em ontologia;
- Proposta de extensão do processo de anotação genômica, a qual engloba um tratamento sobre a anotação baseada em ontologia;
- Desenvolvimento de um protótipo referente a extensão proposta;
- Observação de uso do protótipo no contexto de um grupo de pesquisa; e

- Disponibilização do protótipo funcional integrado ao sistema GARSA.

A observação de uso do protótipo mostrou que com o tratamento sobre a anotação baseada em ontologia e o registro de todo o raciocínio realizado neste, obtém-se evidências de que o conhecimento sobre os termos das ontologias pode ser ampliado e que este tipo de anotação pode vir a ser mais estimulada. Mostrou também que as informações relevantes ao processo podem ser armazenadas permitindo resgatar o histórico do que foi feito durante o uso da ontologia para anotação por parte dos pesquisadores.

A partir deste histórico, a colaboração com os mantenedores das ontologias pode ser ampliada uma vez que os problemas ocorridos ou sugestões de atualização e inserção estão bem caracterizadas, contribuindo assim para a evolução das ontologias.

Cabe ressaltar que, a proposta de manter um registro das ações dos pesquisadores no uso de ontologias, pode ser adotada para outros domínios, e assim contribuir com o desenvolvimento do trabalho dos pesquisadores desse domínio e com a evolução das ontologias utilizadas.

1.5 ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado da seguinte forma: o capítulo 2 faz uma visão geral sobre anotação genômica, descrevendo todos os conceitos que envolvem este tema, como bancos de dados genômicos e sistemas de anotação. Faz uma comparação sobre estes sistemas e por fim descreve o uso de ontologias em aplicações da área de Bioinformática.

O capítulo 3 fornece uma descrição de alguns *workflows* de apoio a pesquisa genômica, e apresenta a modelagem do processo de estudo de seqüências genômicas de forma genérica, apresentando os papéis envolvidos neste processo. Descreve também uma exemplificação do processo de acordo com o sistema GARSA.

O capítulo 4 apresenta a proposta de extensão do processo de anotação genômica, de forma a cobrir o tratamento sobre a anotação baseada em ontologia proposto neste trabalho.

O capítulo 5 descreve o protótipo funcional GARSA Notes. O capítulo 6 mostra um exemplo do uso do protótipo, e por fim, o capítulo 7 apresenta as conclusões do trabalho, as contribuições obtidas e os possíveis trabalhos futuros. Informações

adicionais sobre diversos termos da área de Bioinformática citados durante os capítulos encontram-se no Glossário.

2 ANOTAÇÃO GENÔMICA

No começo da revolução genômica, um interesse da Bioinformática era a criação e manutenção de bancos de dados para armazenar informações biológicas, tais como seqüências de nucleotídeos e aminoácidos.

O desenvolvimento deste tipo de banco de dados envolve não somente os projetos, mas o desenvolvimento de interfaces para que os pesquisadores possam acessar dados existentes tão bem quanto submeter dados novos ou revisados.

Deste modo, o campo da Bioinfomática evoluiu, e surgiram novas tarefas que envolvem a análise e interpretação de vários tipos de dados, incluindo seqüências de nucleotídeos e aminoácidos, domínios e estruturas das proteínas.

Essencialmente, a Bioinfomática possui três componentes (WESTHEAD, 2002), (NCBI, 2005a):

- A criação de bancos de dados, permitindo o armazenamento e gerenciamento de grandes conjuntos de dados biológicos;
- O desenvolvimento de ferramentas que permitem o acesso eficiente para uso e gerenciamento de vários tipos de informação;
- O desenvolvimento de novos algoritmos e estatísticas para acessar relacionamentos entre membros de grandes conjuntos de dados, tais como métodos para localizar um gene numa seqüência, verificar estruturas de proteínas ou sua função e outros.

As pesquisas genômicas beneficiam diversos segmentos como Saúde, Agricultura e Pecuária. Neste contexto, o Brasil, através dos ministérios da Saúde e da Defesa, já vem investindo nas pesquisas sobre agentes biológicos. Alguns destes agentes são vistos como armas biológicas e podem ser utilizados em possíveis guerras ou ataques, não só contra a população humana, mas também contra rebanhos ou plantações, o que afetaria significativamente a economia do país.

Conhecendo as propriedades químicas e físicas, tais como o DNA, o RNA de uma planta ou qualquer outro organismo, é possível prevenir doenças, desenvolver formas de tratamento, contribuir para a melhoria das espécies, entre outros. Por exemplo, para os humanos ou animais, o estudo de suas propriedades poderia prevenir algumas doenças, bem como facilitaria no desenvolvimento de vacinas e formas de tratamento; para plantações de gênero alimentício, o estudo das propriedades do gênero plantado traria

um aumento na produção, pois produtos específicos para o seu crescimento, bem como aqueles para o combate de “pragas” seriam melhor desenvolvidos.

Assim, tipicamente, a descoberta e descrição dos organismos a serem estudados envolvem o uso de *softwares* específicos de Bioinfomática. Estes *softwares* geram dados bastante volumosos, o que requer que sejam armazenados em bancos de dados para posterior análise e consulta. Os bancos de dados possibilitam também que pesquisadores possam compartilhar as seqüências genéticas e trocar informações com outros laboratórios de qualquer parte do mundo.

O armazenamento dos dados genômicos – seqüências e anotações – é feito inicialmente em bancos de dados privados de cada centro de pesquisa; são os bancos de projetos. Após a publicação destes dados em revistas ou eventos especializados da área, estes são então armazenados também em bancos de dados públicos, de acordo com a espécie pesquisada. Alguns destes bancos de dados são apresentados na seção 2.1.

Uma vez que estes dados encontram-se armazenados em bancos de dados públicos, podem ser acessados por todos aqueles que têm interesse sobre o estudo realizado. O entendimento dos resultados se dá por meio da leitura/observação das anotações feitas para cada seqüência pesquisada. As anotações constituem-se na identificação de cada gene contido nas seqüências genômicas armazenadas nos bancos de dados, que em geral possuem formatos específicos de armazenamento nestes bancos. O conceito de anotação é apresentado de forma detalhada na seção 2.2. As anotações são feitas através de sistemas que apóiam esta atividade, logo, a seção 2.3 descreve alguns sistemas de anotação, comentando suas principais funcionalidades, os tipos de anotações que suportam e também apresenta uma comparação entre eles.

Uma forma de realizar a anotação é através do uso de ontologias. A ontologia cria um vocabulário comum em relação ao conteúdo anotado e assim a compreensão da anotação torna-se facilitada para todos aqueles que tenham acesso a ela. A seção 2.4 apresenta conceitos e exemplos sobre ontologias específicas para o domínio da Bioinformática.

2.1 BANCOS DE DADOS GENÔMICOS

Um banco de dados genômico está geralmente associado a um *software* projetado para atualizar, consultar e recuperar componentes dos dados genômicos armazenados no sistema. Um banco de dados pode ser simples, sendo composto somente por um único arquivo contendo muitos registros, cada qual incluindo o mesmo conjunto de

informação. Por exemplo, um registro associado a um banco de dados de seqüência de nucleotídeos contém tipicamente a informação de qual o nome de contato, a seqüência de entrada com uma descrição do tipo da molécula, o nome científico do organismo, e geralmente, citações da literatura associadas com a seqüência.

A maioria dos bancos de dados genômicos são acessíveis através de sistemas na *Web*, os quais possuem recursos como a busca e filtragem da informação.

Os bancos de dados genômicos podem ser classificados em públicos ou privados. Os bancos de dados públicos (BERGERON, 2002) são criados e mantidos por laboratórios ou centros de pesquisas sem fins lucrativos, e possibilitam que pesquisadores possam compartilhar seqüências genéticas e trocar informações com outros laboratórios de qualquer parte do mundo. Como exemplo temos: *GenBank* (BENSON et.al., 2005), *Refseq* (PRUITT, 2005), *Pfam* (FINN et.al., 2006), *SWISSPROT* (BAIROCH, et. al, 2000), *CDD* (MARCHLER-BAUER, 2005), entre outros.

Já os bancos de dados privados são criados e mantidos por laboratórios e companhias que visam lucro. Geralmente esses laboratórios são associados com instituições acadêmicas. Como exemplo de bancos de dados privados pode-se citar o *LifeSeq* (INCYTE, 2007) que contém seqüências de genes humanos e ratos (BERGERON, 2002).

Além dos bancos públicos e privados, tem surgido na comunidade de Bioinformática uma terceira categoria, que são os bancos de dados de projeto. Estes bancos são próprios de grupos de pesquisa genômica, e possuem uma parte pública e outra restrita. A parte pública destes bancos disponibiliza os dados do grupo de pesquisa que já foram publicados em revistas especializadas ou outros veículos de comunicação da área. Enquanto que os dados ainda em fase de estudo e aqueles que ainda não tiveram sua respectiva publicação ficam numa parte restrita do sistema, onde somente usuários autorizados podem ter acesso. São descritos a seguir exemplos de banco de dados públicos.

Com relação aos bancos de dados públicos, existem basicamente duas classificações, os bancos de dados primários e os bancos de dados secundários:

- Os bancos de dados primários são aqueles derivados diretamente dos dados obtidos a partir do sequenciamento de aminoácidos ou proteínas. Apresentam resultados desses dados de forma experimental e que são publicados com alguma interpretação, onde não há uma análise cuidadosa dos dados, podendo haver redundâncias. Normalmente, as seqüências identificadas pelos laboratórios são

submetidas aos bancos de dados primários antes da publicação do respectivo artigo sobre a descoberta da seqüência, já que a maioria das revistas considera isso um pré-requisito para publicação. Contudo, a disponibilização dos dados é feita somente após a publicação ou a um intervalo de tempo que chega até um ano. Estas bases de dados agem primeiramente como o armazenamento de arquivos de informações da seqüência. Conseqüentemente, cada entrada é possuída pelo provedor da seqüência e a integração das informações anotadas é quase impossível.

- Os bancos de dados secundários, também denominados de bancos curados, são aqueles onde há uma compilação e interpretação dos dados de entrada de forma que podem ser obtidos dados mais representativos. Cada seqüência é curada manualmente e este tipo de banco de dados não possui redundâncias. Desta forma é garantida a qualidade e a confiabilidade das informações armazenadas. São bancos de dados desenvolvidos por comunidades particulares. Este tipo de banco é encontrado em quantidade menor se comparado à quantidade de bancos de dados primários.

Os bancos de dados genômicos podem ser também classificados de acordo com o conteúdo que armazenam, entre:

- Seqüências e anotações: seqüências de nucleotídeos e suas respectivas anotações;
- Proteínas e informações sobre as respectivas funções;
- Estrutura de moléculas de proteínas: secundárias, representadas em um plano, ou terciárias, representada em três dimensões;
- Taxonomia: classificação dos organismos vivos;
- Bibliografia: artigos, jornais, periódicos e outros na área de biologia molecular.

Os bancos de dados em biologia molecular são importantes principalmente para proporcionar à comunidade científica (e a quem mais interessar) uma forma de tornar os dados produzidos em todo o mundo acessíveis de forma mais fácil, rápida e inteligente (NCBI, 2005a). Com esse intuito, o governo americano lançou em 1988 o NCBI (*National Center for Biotechnology Information*) (NCBI, 2005a) para reunir bancos de dados públicos contendo seqüências de DNA.

O NCBI disponibiliza um grande número de ferramentas de informática e recursos para auxiliar o cientista na pesquisa genética, constituindo-se assim como o maior órgão de referência de Bioinformática. É lá que é mantido o maior banco de dados genômico público, o *GenBank*.

Além de tornar a acessibilidade aos estudos genômicos mais facilitada, os bancos de dados servem como apoio para verificar se uma determinada seqüência já está cadastrada e também para buscar grandes conjuntos de dados que servirão para análises de similaridade e consequentemente para a anotação.

Para melhor compreensão do conteúdo armazenado nestes bancos, em especial as anotações, as seções 2.1.1 e 2.1.2 apresentam a descrição de dois bancos de dados bastante conhecidos na comunidade de Bioinfomática e que diferem nas suas classificações. São eles: *GenBank* e *Pfam*, respectivamente.

Um item que não é tratado neste trabalho, mas que é importante para a comunidade de Bioinformática e desta forma relevante ser comentado, refere-se às iniciativas de esquemas de bancos de dados genéricos como o CHADO (CHADO, 2005) e o GUS (GUS, 2005) para auxiliar o desenvolvimento dos projetos de pesquisa genômica. Entre os projetos que utilizam o CHADO estão WormBase (CHEN, 2004) e Flybase (CROSBY et al., 2007) e quanto ao GUS podemos citar os projetos GeneDB (HERTZ-FOWLER et al., 2004) e TCruziDB (AGÜERO et al., 2006). O objetivo destes esquemas é estabelecer uma padronização dos bancos de dados genômicos. Outros detalhes referentes a estes esquemas podem ser encontrados em (CHADO, 2005), (GUS, 2005), (ALMEIDA, 2006).

2.1.1 GENBANK

O *GenBank* (BENSON et.al., 2005) é um banco de dados de seqüências de nucleotídeos produzido pelo NCBI. É usado como referência no sentido de verificar se uma dada seqüência já está catalogada. O histórico do volume de seqüências armazenadas no *GenBank* demonstra que, a cada ano, o número de seqüências armazenadas, bem como o número de bases, cresce cerca de 70% ao ano, como ilustra a FIG. 2.1. A cada ano novas versões do banco são disseminadas (GENBANK, 2005). O *GenBank* é um banco de dados dito primário e seus colaboradores recebem seqüências produzidas em laboratórios ao redor do mundo.

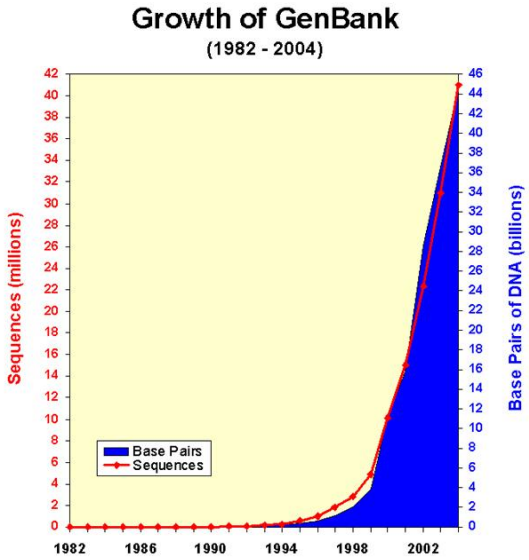


FIG. 2.1 Crescimento *GenBank* I (GENBANK, 2005)

Apesar de existirem outros bancos de dados que seguem a mesma linha do *GenBank*, ou seja, bancos do tipo primário, a comunidade científica utiliza praticamente apenas o *GenBank*. Isto pode ser percebido na FIG. 2.2 retirada do *release* sobre o *GenBank*, lançado em agosto de 2005 pelo NCBI, onde é mostrado um quadro comparativo do crescimento deste banco de dados e os outros da mesma linha.

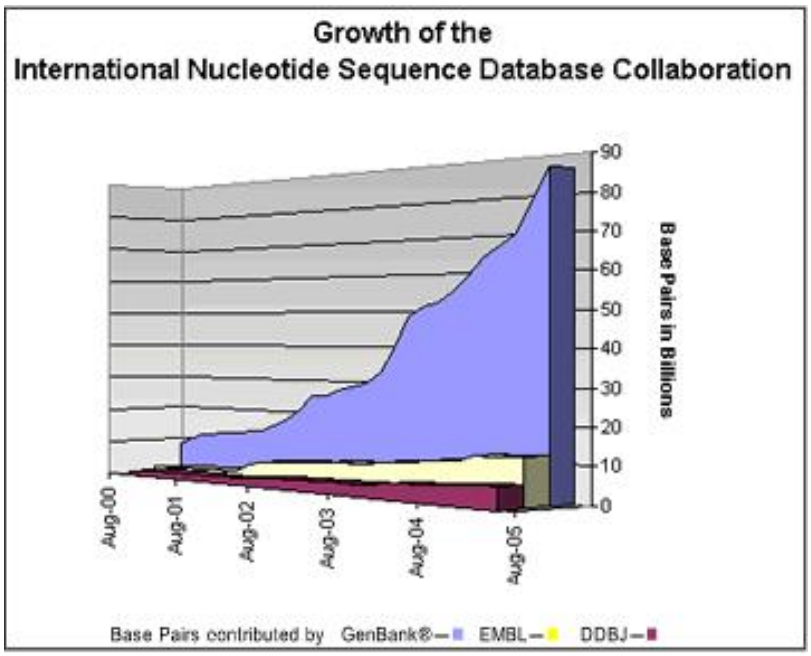


FIG. 2.2 Crescimento *GenBank* II (GENBANK, 2005)

Os bancos de dados EMBL (KULIKOVA et. al., 2007) e DDBJ (TATENO et. al., 1998) que aparecem da FIG. 2.2 são repositórios de seqüências de DNA. O *GenBank*

troca dados com os bancos EMBL e DDBJ de forma a manter o repositório de seqüências o mais completo possível.

Uma seqüência submetida ao *GenBank* inclui diversas informações, tais como: a descrição concisa da seqüência; um número de identificação (*access number)*, o qual é único para cada seqüência; palavras chaves associadas ao gene da seqüência; citações a artigos onde a seqüência foi publicada; características da seqüência; entre outras. Estas informações são submetidas de acordo com um formato para anotação genômica denominado de *Feature Table.* Maiores detalhes sobre este formato podem ser encontrados em (DDBJ/EMBL/GENBANK, 2007).

As características das seqüências são as anotações referentes a cada região codificante da seqüência. Sempre que as anotações são alteradas, os dados devem ser atualizados, sendo que a seqüência recebe um novo identificador (número de acesso). O identificador anterior é armazenado de forma a não se perder a referência anterior e desta forma poder ser visualizada a evolução do estudo sobre a seqüência em questão.

2.1.2 PFAM

O *Pfam* (FINN et.al., 2006) é um banco de dados de família de proteínas desenvolvido pelo *The Sanger Institute* (THE SANGER INSTITUTE, 2005), um instituto de pesquisa genômica do Reino Unido.

Neste banco de dados, para cada família de proteínas, pode-se:

- Ver os alinhamentos múltiplos;
- Visualizar as arquiteturas de domínios das proteínas;
- Examinar a distribuição das espécies;
- Seguir os *links* para outras bases de dados;
- Ver estrutura de proteínas conhecidas.

O *Pfam* é um banco de dados subdivido em duas partes, *Pfam-A* e *Pfam-B*, equivalentes a um banco de dados secundário e outro primário respectivamente. A primeira parte, *Pfam*-A, é a que contém dados classificados como secundários, - aqueles em que há uma verificação dos dados antes de serem armazenados, tornando-os mais confiáveis e representativos. Possui quase 9000 famílias de proteínas (dados de novembro de 2006). E para dar ao *Pfam* uma cobertura mais compreensiva de proteínas conhecidas, automaticamente é gerado um suplemento chamado *Pfam*-B que contém dados classificados como primários – dados que não sofreram uma validação para serem

armazenados. Este contém um grande número de famílias pequenas que não foram compiladas e interpretadas ainda de forma suficientemente cuidadosa, a ponto de merecerem fazer parte do *Pfam*-A. Embora a qualidade das famílias do *Pfam*-B seja considerada inferior, elas podem e costumam ser usadas quando nenhuma família no *Pfam*-A é encontrada.

A FIG. 2.3 mostra a cobertura das duas partes do *Pfam*, evidenciando que a cobertura do *Pfam-A* é bem maior, apesar da grande dificuldade que existe para manter um banco de dados secundário.

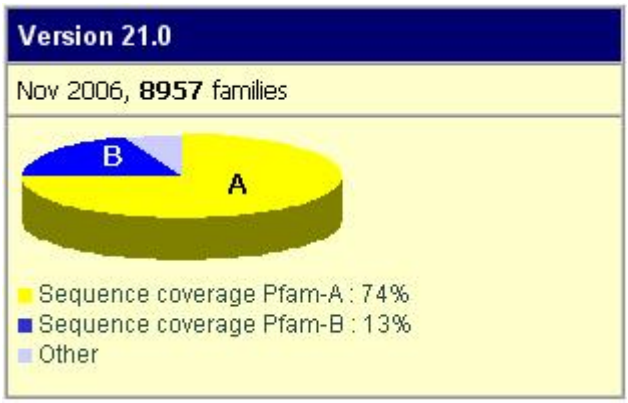


FIG. 2.3 Crescimento PfamA e PfamB (PFAM, 2006a)

Os dados armazenados no banco de dados *Pfam* seguem um formato que se subdivide em quatro seções: *header*, *reference*, *comment*, *alignment* (PFAM, 2006b).

***Header*:** possui campos que incluem informações específicas do *Pfam* tais como números de identificação (*access number*), um nome curto e reprentativo assim como uma pequena descrição sobre a família.

Reference: contém principalmente referências a outros bancos de dados e literatura.

Comment: contém informação funcional sobre a família como as anotações e outras informações. A anotação não possui um formato estrito como no formato *Feature Table*, utilizado pelo *Genbank*.

***Alignment*:** possui informações sobre o alinhamento da família de proteínas.

Todos os dados do *Pfam* são disponíveis para *download* em formatos *flatfile* via FTP a partir do *sites*[2] disponibilizados pelo *Pfam* e também via conjunto de dados de um SGBD relacional MySql (PFAM, 2006b).

[2] http://www.sanger.ac.uk/Software/Pfam/ftp.shtml

2.2 ANOTAÇÃO

Uma das tarefas mais importantes dos projetos genoma é a interpretação dos experimentos *in-silico* para se obter o conhecimento biológico destes dados e assim identificar o que representa cada uma das regiões de cada seqüência estudada. A anotação consiste simplesmente no processo de identificação e descrição das regiões codificantes das seqüências.

Para alcançar este objetivo, os pesquisadores tipicamente executam essa tarefa, extraindo e fragmentando o DNA do organismo estudado, consultando fontes de dados externas, executando programas de análise nas seqüências, e gerando anotações manuais de acordo com suas interpretações na bancada de experimentos biológicos e *in silico*.

Na bancada de experimentos *in silico*, um desafio da Bioinfomática é ter ferramentas efetivas que ajudem os pesquisadores a identificar grandes conjuntos de seqüências (de nucleotídeos ou proteínas). As ferramentas devem ajudar um pesquisador a explorar e rapidamente testar suas hipóteses, acessar anotações armazenadas em fontes de dados públicas para compará-las com seus resultados, executar programas de análise e obter as anotações automaticamente, analisar anotações existentes até então com a ajuda de uma interface apropriada, e gerar automaticamente ou manualmente novas anotações (LEMOS, 2005).

Assim, podemos classificar anotação como:

Manual: diretamente criadas pelo pesquisador;

Automática: gerada por programas de análise ou importada de fontes de dados públicas.

Geralmente o processo de anotação de genomas é feito em três etapas (LEMOS, 2003):

Na primeira etapa é feito o uso das ferramentas de Bioinfomática e são importados dados referentes às seqüências das fontes de dados públicas. Assim, as seqüências obtidas passam por diversos programas, os quais ajudam os anotadores a identificá-las, representando assim a anotação automática.

A segunda etapa necessita de especialistas que observem os dados obtidos na primeira etapa pelas ferramentas automáticas e pela importação das fontes de dados públicas; e identifiquem as seqüências de acordo com critérios pré-definidos.

Após a identificação dos genes, na terceira etapa é feita a anotação manual que geralmente é feita somente pelo anotador. Contudo, o ideal seria promover a interação

entre vários anotadores, bioinformatas e biólogos especialistas em diferentes áreas e no organismo estudado para discutir como as informações obtidas nas etapas anteriores podem estar relacionadas com a biologia do organismo em questão.

Em resumo, anotação é a escrita da identificação das regiões codificantes das seqüências, inicialmente feita de acordo com um vocabulário próprio do bioinformata ou do grupo de pesquisa.. O termo anotação é utilizado para descrever a etapa final de todo um processo de análise realizado sobre as seqüências.

Além das etapas citadas anteriormente que englobam as anotações automática e manual, outro tipo de anotação que vem sendo utilizada e constitui mais uma etapa, é a anotação baseada em ontologia, onde os genes são identificados através de termos pertencentes ao domínio da biologia molecular. A seção 2.4 descreve com mais detalhes as ontologias.

O processo de anotação genômica faz parte de um processo maior que é o estudo das sequencias genômicas (experimentos *in-silico*). Para gerenciar todo o processo de anotação, os pesquisadores contam com sistemas desenvolvidos para esta finalidade que empacotam diversos programas tanto para análise quanto para anotação das seqüências.

A maioria destes sistemas suportam ambas anotações automática e manual. Já a anotação baseada em ontologia ainda se apresenta em um número menor de sistema. Exemplos destes sistemas são descritos na seção 2.3.

2.3 SISTEMAS DE ANOTAÇÃO

As ferramentas que ajudam os pesquisadores a criar, recuperar e analisar as anotações que identificam e caracterizam as seqüências são chamadas de sistemas de anotação. Muitos destes sistemas, como os citados adiante, são sistemas de propósito geral, que podem ser utilizados pelos centros de pesquisas, e por diversos órgãos mantenedores de bancos de dados genômicos, como *The Sanger Institute* (THE SANGER INSTITUTE, 2005) e TIGR (TIGR, 2007).

Existem ainda os sistemas desenvolvidos especialmente para atender à necessidade de alguns centros de pesquisa, como o Sistema GARSA (DÁVILA et al., 2005), que além da parte de análise das seqüências, captura e registra todas as anotações realizadas.

Existem bancos de dados genômicos que não se apóiam nestes sistemas (ou *softwares*) de anotação, seja de propósito geral ou específico. Eles apresentam estruturas, como tabelas que irão ser populadas com as anotações (também chamada de *features*) das seqüências, e deste modo oferecem formatos os quais os usuários devem

seguir para anotar e fazer a submissão de suas seqüências. O formato *Feature Table* (DDBJ/EMBL/GENBANK, 2005) utilizado para submissão de anotação de seqüências de bancos de dados como *GenBank*, EMBL e DDBJ pode ser citada como exemplo.

Os sistemas de anotação diferem entre si com relação a aspectos como a tecnologia utilizada para desenvolvimento, a portabilidade, recursos oferecidos entre outros. Alguns conseguem importar e ler os padrões pré-definidos como os formatos usados pelos bancos de dados genômicos: FT (comentado anteriormente), GFF e outro menos comum, o GAME.

O GFF (DURBIN et.al., 2005) é um formato simples para transferência de anotações genômicas. Descreve genes e outras características associadas com o DNA, o RNA e seqüências de proteínas.

O GAME (XML COVER PAGES, 2002) provê um DTD - XML[3] para a troca de anotação genômica. Ele permite que pesquisadores compartilhem estas descrições no formato XML (O'REILLY XML, 2000).

Os diversos sistemas, para facilitar a anotação da seqüência em estudo se aproveitam de anotações de seqüências similares já anotadas. Desta forma eles possuem uma funcionalidade que é a leitura de dados provenientes de bancos de dados como *GenBank*, EMBL, DDBJ, *Pfam*, *Ensembl* (HUBBARD et. al., 2007), Swiss-Prot (BAIROCH, et. al, 2000) e outros. De forma a entender como funcionam os sistemas de anotação e levantar os tipos de anotações adotados: automática, manual e baseada em ontologia, a seguir são descritos alguns sistemas, a saber: *Apollo, Artemis*, *BioNotes,* DAS e GARSA e ao final destas descrições é apresentada uma tabela comparativa.

2.3.1 APOLLO

Apollo Genome Annotation Curation Tool (Apollo) (LEWIS, et. at., 2002) é um sistema de anotações de código aberto (*open-source*) que gera anotações automáticas e também permite que os pesquisadores explorem anotações em vários níveis de detalhe e criem anotações manuais, tudo em um ambiente gráfico, contudo o sistema não suporta anotação baseada em ontologia

Desenvolvido a partir de uma colaboração entre *Berkeley Drosophila Genome Project* (BERKELEY DROSOPHILA GENOME, 2005) e *The Sanger Institute,* o

[3] DTD XML: o DTD é utilizado na linguagem XML para especificação de documentos. Indica elementos, atributos, valores e os relacionamentos entre os elementos. Pode também definir entidades.

sistema Apollo está sendo usado pelos biólogos do *FlyBase* (CROSBY, et. al., 2007) para fazer as últimas anotações do genoma da *Drosophila* (BDGP, 2007), e também será utilizado para compartilhar estas anotações na comunidade.

O sistema foi projetado para ser flexível e extensível de forma que possa atender diferentes organismos. O projeto GMOD (GMOD, 2005) adotou o modelo do Apollo para o módulo de anotação.

O sistema Apollo também disponibiliza para execução (embutidas no sistema) diversas ferramentas de análise de dados genômicos como GENSCAN (BURGE et. al, 1998) e BLAST (busca por similaridade entre as seqüências) (ALTSCHUL et. al., 1990).

É uma aplicação escrita em linguagem de programação Java e que pode ser feito o *download* e executado em sistemas *Windows*, *Mac OS X*, ou qualquer sistema tipo *Unix* (incluindo *Linux*).

Type	Name	Range	Score
Primate	Q9GZT9	1092503 - 1093355	225.0
Primate	CAC42509	1092503 - 1093355	225.0
Rodent	Mm#S2171818	1092503 - 1093370	220.0
Rodent	Rn#S197560	1092503 - 1093370	212.0
Worm	Q9U4H6	1092479 - 1093361	198.0
Worm	O45918	1092479 - 1093361	198.0

FIG. 2.4 Tabela de descrição de *feature* (APOLLO USER GUIDE, 2005)

A FIG. 2.4 mostra a tabela de descrição de *feature*. Cada linha descreve uma *feature* única. As colunas descrevem o tipo da *feature* (*Type*), na figura designado pelo nome do organismo; o nome (*Name*), de acesso ao alinhamento; a posição genômica do alinhamento (*Range*), e o *score* (*Score*) do programa *Blastx* (modalidade do programa Blast que converte uma seqüência de nucleotídeos em proteínas). Se quiser informações detalhadas de uma *feature* nesta tabela, como o nome da seqüência alinhada, seu tamanho e descrição completos, basta selecionar a linha da *feature* e as informações aparecerão em uma segunda tabela (não mostrada aqui).

2.3.2 ARTEMIS

O *Artemis* (RUTHERFORD et. al., 2000) é uma ferramenta de anotação que permite a visualização das anotações da seqüência e os resultados de análise com o contexto da seqüência. Desenvolvido pelo *The Sanger Institute*, o *Artemis* é um *software* livre que está disponível para qualquer sistema operacional em uma versão desenvolvida em Java, porém tem maior suporte para ambientes *UNIX* e *GNU/Linux*.

O *Artemis* lê seqüências completas no formato FASTA[4] dos bancos de dados *GenBank* e EMBL. As anotações das seqüências podem estar no formato FT ou GFF. A principal anotação gerada pelo sistema é a automática, contudo, informações complementares podem ser adicionadas (anotação manual). O sistema suporta anotação baseada em ontologia.

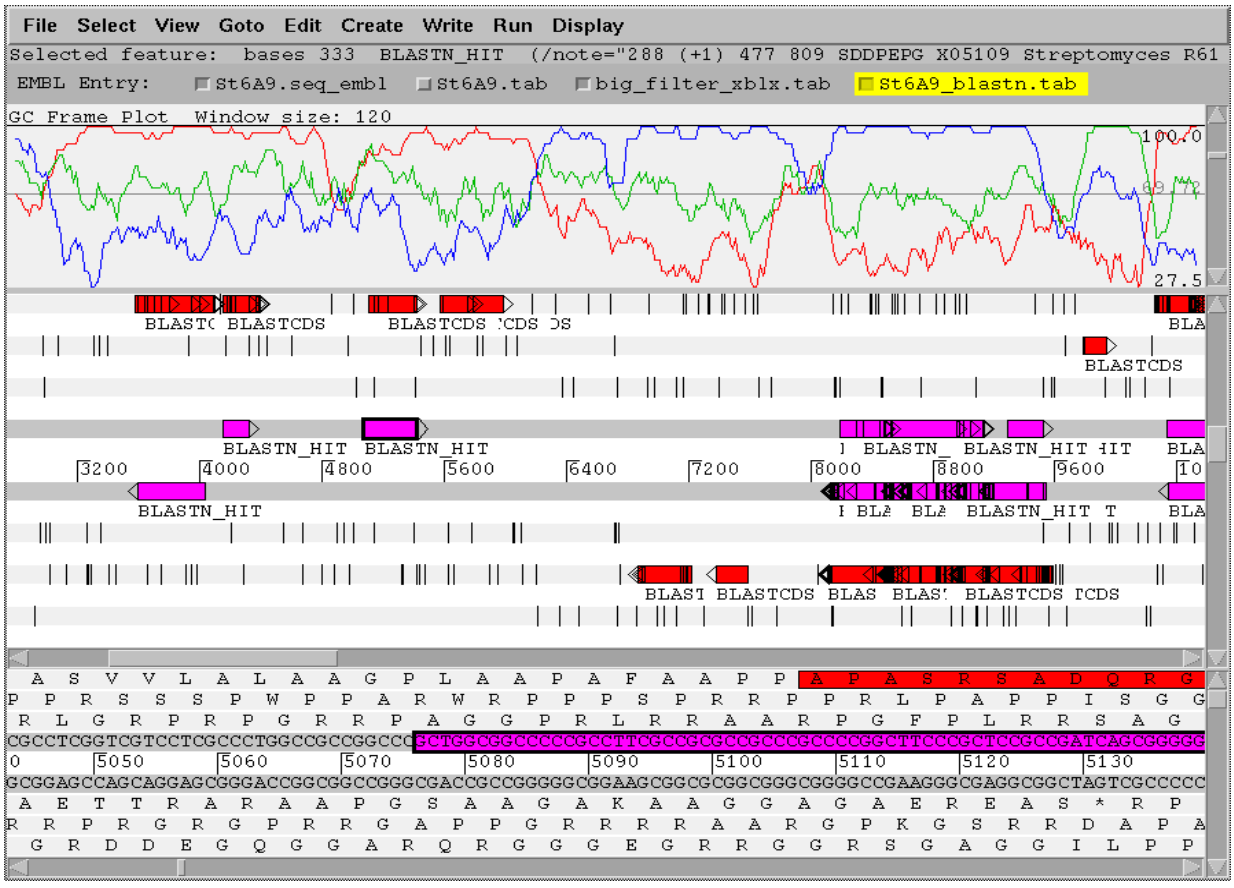


FIG. 2.5 Resultado da análise de programas no Artemis (ARTEMIS EXAMPLES, 2005)

O *Artemis* é a principal ferramenta de anotação usada para análise de genoma microbial do *The Sanger Institute*. Além disso, o sistema mostra o resultado de vários programas de análise da seqüência, como predições de *CDs*, *Blastn*, *Blastx*, e outros.

[4] FASTA: formato para representação de seqüências genéticas em arquivos digitais para que elas possam ser usadas por qualquer *software* de análise de seqüências.

Estes programas rodam externamente, o que significa que os resultados de qualquer análise podem ser facilmente incorporados sem modificação do *Artemis*. A FIG. 2.5 mostra os resultados dos programas *Blastn, Blastx* e *FramePlot* para o organismo *Streptomyces coelicolor* (micróbio que vive no solo).

Um outro programa que faz parte da distribuição do *Artemis* é o ACT (CARVER et. al., 2005). Este programa é um visualizador de comparação de seqüências de DNA que usa os componentes do *Artemis* para exibir as seqüências e assim herda boas ferramentas de busca e análise. Permite uma visualização interativa de comparações entre seqüências genômicas completas e anotações associadas. Além das leituras de seqüências dos bancos *GenBank* e *EMBL*, pode ler seqüências no formato FASTA ou até no formato inicial da seqüência. Anotações de seqüências extras podem estar nos formatos FT e GFF. A comparação de seqüências visualizadas pelo ACT é geralmente o resultado da execução de uma busca com *Blastn*.

2.3.3 BIONOTES

O *BioNotes* (LEMOS, et. al., 2003), desenvolvido pelo Departamento de Informática da Pontifícia Universidade Católica do Rio de Janeiro (INF PUC-RIO, 2005), é um sistema de anotação de seqüências que ajuda os pesquisadores a buscarem fontes de dados externas, executar programas de análise, analisar anotações e adicionar novas anotações para registrar sua interpretação do dado.

Um usuário pode atualmente buscar anotações importadas de fontes públicas externas, executar e armazenar anotações automaticamente geradas por algum programa de análise, e manualmente adicionar, apagar e atualizar anotações, para o qual elas tornam-se disponíveis para sua comunidade. O sistema não suporta a anotação baseada em ontologia.

A arquitetura do *BioNotes* e seu modelo de dados tem o objetivo de ser flexível o bastante para permitir que o sistema seja usado para anotar qualquer genoma e EST ou projetos genoma completo.

O sistema *BioNotes* possui uma arquitetura cliente-servidor que permite às comunidades compartilharem suas anotações através da *Web*, o que torna mais fácil o acesso às anotações e assim uma descoberta mais rápida de novas informações biológicas. É um sistema multi-usuário, instalado em um servidor, permitindo acesso

distribuído por uma grande comunidade de usuários, que de forma intensa facilita o compartilhamento de anotações.

Para manter a segurança no sistema, o *BioNotes* oferece diferentes perfis de usuários para facilitar o controle de quais usuários podem executar quais comandos. Além disso, o sistema suporta o conceito de uma fonte de dados privada, que é, um conjunto de seqüências e anotações armazenadas pelo sistema que são privados e acessíveis somente por uma grupo de usuários. Um exemplo deste tipo de grupo é o que tem acesso à *TCruzi*, uma fonte de dados privada que armazena anotações e seqüências relacionadas com o organismo do *Trypanosoma Cruzi*; e GLUCONA, que armazena anotações e seqüências ligadas ao organismo *Gluconacetobacter diazotrophicus*.

O *BioNotes* é escrito em linguagem de programação Java para facilitar portabilidade, e oferece interface *Web*.

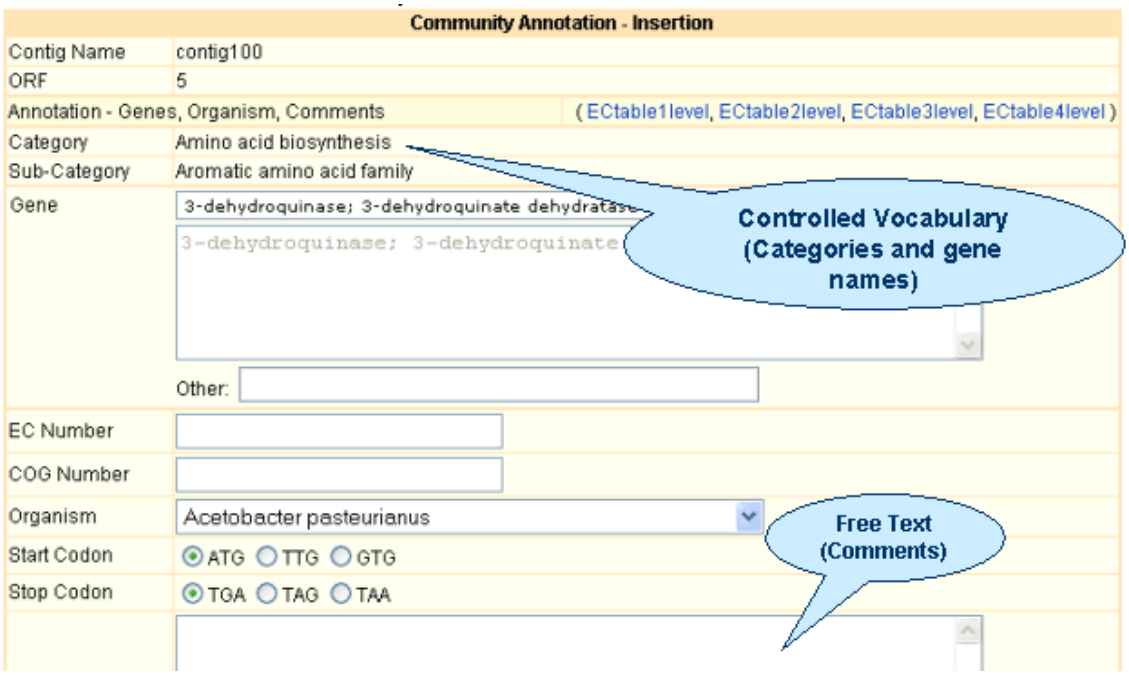


FIG. 2.6 Tela para inserir uma anotação manual no BioNotes (BIONOTES, 2005)

A FIG. 2.6 mostra um exemplo de uma anotação manual que foi determinada para uma ORF. Alguns campos como categorias e nomes de genes possuem opções pré-definidas, constituindo um vocabulário controlado para a anotação ser feita. Contudo, existe ainda um campo de texto livre para possíveis comentários.

2.3.4 DAS

O DAS (DOWELL et. al., 2001) foi projetado como um sistema leve para integração de dados de um número de bancos de dados distribuídos heterogêneos. É um sistema em que o usuário (cliente) seleciona um único servidor de referência (servidor genômico) e um número qualquer de servidores de anotação. O cliente DAS combina a informação retornada dos servidores em uma única tela gráfica.

A FIG. 2.7 mostra um exemplo de uma arquitetura básica do DAS. Um servidor, no caso o *Washington University Genome Sequencing Center* é projetado para ser o servidor de referência. Outros servidores de anotação como *Ensembl*, *Whitehead* e *Sean Eddy Laboratory* provêem informações das seqüências que estão no servidor de referência. O cliente, no caso o *Cold Spring Harbor Laboratory*, obtém os dados dos vários servidores de anotação e gera automaticamente uma visão integrada destes dados e os representa graficamente em uma única tela.

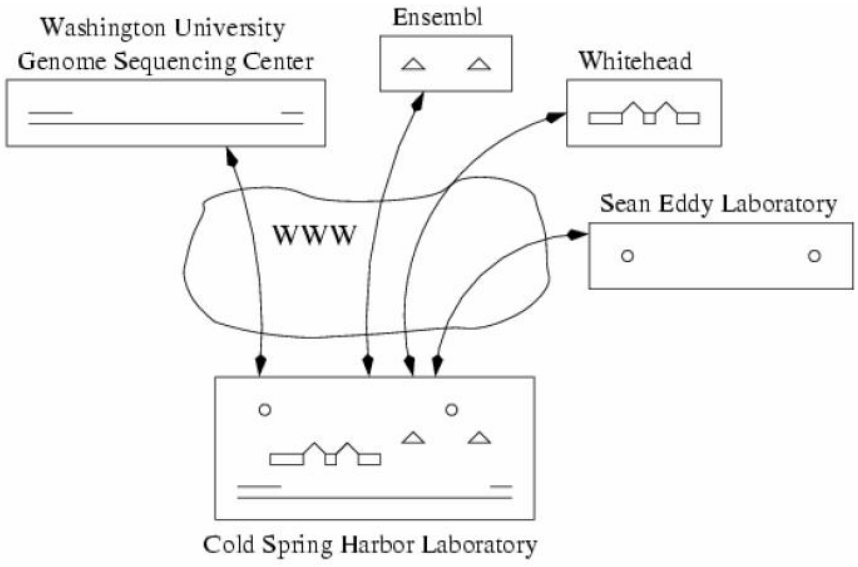


FIG. 2.7 Arquitetura básica do DAS (DOWELL et. al., 2001)

As anotações no DAS estão associadas às posições (marcadas por início e fim) das seqüências de acordo com o organismo. Em alguns projetos as seqüências podem ser de cromossomos inteiros, enquanto em outros elas podem ser *contigs* ou *reads*. As anotações seguem o formato FT e possuem uma descrição de como a anotação foi obtida automaticamente (que pode ser uma referência para um programa de análise). Caso o usuário queira anotar algo diferente do definido pela FT é permitido adicionar novos tipos de anotações sem restrição de vocabulário (anotação manual). O sistema

pode também fazer o uso de ontologias padrões da área, como a *Gene Ontology* (citada na seção 2.4)

2.3.5 GARSA

O GARSA (DÁVILA et al., 2005) é um dos projetos que fazem parte do Consórcio *BioWebDB* (BIOWEBDB, 2007), no qual os cientistas são diretamente envolvidos com genômica comparativa e bancos de dados genômicos.. É um ambiente baseado na *Web*, que tem como objetivo facilitar a análise, apresentação integrada da informação genômica, concatenação de algumas ferramentas de bioinformática e bancos de dados de seqüências, usando uma abordável flexível e amigável para o usuário. A FIG. 2.8 mostra um exemplo das anotações do GARSA.

Validate and Annotate CDS

Annotations

Cluster	Start	Stop	Description	Molecular Functional	Biological Process	Cellular Component
TGEG101001A01.g.1	65	568	trans-sialidase	exo-alpha-sialidase activity	pathogenesis	
TGEG101001B01.g.1	1	465	trans-sialidase	exo-alpha-sialidase activity	pathogenesis	
TGEG101001C01.g.1	452	51	Hypothetical Protein			
TGEG101001C08.g.1	192	79	Hypothetical Protein			
TGEG101001D01.g.1	452	33	mitochondrial ATP-dependent zinc metallopeptidase	metallopeptidase activity	proteolysis	membrane
TGEG101001D06.g.1	573	64	Hypothetical Protein			

FIG. 2.8 Anotação no sistema GARSA (GARSA, 2007)

A FIG. 2.8 (apresentada parcialmente) mostra as seqüências estudadas representadas através de Clusters (*Cluster*), a posição inicial e final da região codificante (*Start e Stop*), e as respectivas anotações. O GARSA permite a anotação automática e esta é confirmada através da anotação manual, representada pelo campo *Description* na figura. Os campos *Molecular Functional*, *Biological Processes* e *Cellular Component* correspondem à anotação baseada em ontologia. Maiores detalhes sobre o GARSA são encontrados no próximo capítulo.

2.3.6 COMPARAÇÃO ENTRE OS SISTEMAS DE ANOTAÇÕES

Um estudo de comparação entre sistemas de anotação foi realizado por LEMOS (2005), cujos aspectos utilizados para a pesquisa e a comparação dos sistemas foram: o

modelo de dados, a linguagem de programação utilizada, a interface adotada, o sistema de gerência de bancos de dados utilizado, a forma de armazenamento das anotações e das fontes de dados externas, processamento dos programas de análise dos dados e o controle de versão das anotações e tratamento das anotações manuais.

Para o contexto deste trabalho, procurou-se apresentar as definições sobre anotação genômica e dar exemplos de sistemas de anotação de propósito geral. Para estes exemplos houve a preocupação de observar itens que não foram contemplados na comparação feita por LEMOS (2005), tal como o uso de ontologias. Diante destes e de outros itens, construiu-se uma outra tabela comparativa.

Devido ao pouco tempo e à quantidade considerável de sistemas de anotação, não foram analisados todos os sistemas contidos em (LEMOS, 2005). A análise apresentada aqui limitou-se aos cinco sistemas que foram investigados e que foram descritos anteriormente: Apollo, Artemis, BioNotes, DAS e GARSA.

Os seguintes itens foram observados:

Versão gratuita: verifica se o sistema possui versão gratuita ou não;

Plataforma: em qual sistema operacional o sistema pode ser executado;

Visualização gráfica: se o sistema possui uma visualização que é “amigável” para os usuários;

Acesso multiusuário: verifica se o sistema permite sua utilização com mais de um usuário por vez;

Anotação baseada em ontologia: se o sistema trabalha com ontologias para anotação genômica e sendo assim, se possui alguma ferramenta para a escolha dos termos.

Lê outros formatos de anotações: verifica se o sistema suporta a importação e a leitura de outras fontes de anotações, provenientes de fontes de dados externas, como banco de dados públicos. Os formatos em geral são o FT, GFF, GAME e o formato usual para seqüências, o FASTA.

Incorpora resultados da seqüência: os sistemas em sua maioria permitem a execução de ferramentas, como programas de análise e buscas por similaridade, controlando a execução destas e assim incorporam seus resultados para a anotação da seqüência. Outros sistemas apenas importam resultados.

Quem usa: determinados sistemas de anotação têm sido tomados como padrão por alguns centros de pesquisa da comunidade de Bioinfomática.

TAB. 2.1 Comparação entre sistemas de anotações

Sistema de Anotação	Versão gratuita	Plataforma	Visualização gráfica	Acesso multiusuário	Anotação baseada em ontologia	Lê outros formatos de anotações	Incorpora resultados das seqüências	Quem usa
Apollo	S	2	S	-	N	S / (FT, GFF, GAME)	S	GMOD FlyBase
Artemis	S	2	S	-	S	S / (FT, FASTA)	S	Sanger
BioNotes	N	-	S	S	N	S / (-)	S	Riogene
DAS	-	-	S	-	S	S / (FT)	S	WormBase, FlyBase, Emsembl, TIGR, UCSC
GARSA	-	2	S	-	S	S/ (FASTA)	S	Fiocruz

Legenda

- Versão gratuita S: Sim, N: Não, '-': informação não encontrada.
- Visualização Gráfica / Acesso Multiusuário / Anotação Baseada em Ontologia / Permite a execução de outras ferramentas: S: Sim, N: Não.
- Plataforma: 1: Windows e Linux, 2: Qualquer sistema, '-': informação não encontrada.
- Lê outros formatos de anotação: S: Sim / Formato (FT, GFF, GAME, FASTA), '-': informação não encontrada.

Os sistemas na sua maioria são de distribuição gratuita, mas o contrário pode ocorrer como é o caso do BioNotes. Sobre a questão de plataformas, os sistemas em geral podem ser executados em todos os diferentes sistemas operacionais e todos possuem uma boa visualizarão gráfica. O acesso multiusuário foi identificado somente no BioNotes. Já o uso de ontologias é feito somente por três destes sistemas, Artemis, DAS e GARSA. Todos os sistemas lêem outros formatos de anotações e permitem a execução de ferramentas de análise da seqüência, em geral sendo executadas externamente e seus resultados anexados ao sistema. Os diversos centros de pesquisas que utilizam estes sistemas de propósito gerais são também listados na tabela.

2.4 ONTOLOGIAS PARA BIOINFORMÁTICA

Uma ontologia é uma descrição de conceitos dentro de um domínio e do relacionamento entre eles. De acordo com GRUBER (1993), uma ontologia é definida como “uma especificação formal de uma conceitualização”. Basicamente, uma ontologia consiste de instâncias, classes, conceitos e relacionamentos.

Na área de ciência da computação as ontologias são aplicadas em diversos ramos como: web-semântica, engenharia de requisitos, inteligência artificial, banco de dados e outros. Uma área interdisciplinar, na qual vem sendo crescente o uso de ontologias é a área Biomédica e consequentemente a Bioinformática.

Na Bioinformática algumas ontologias já foram desenvolvidas e possuem finalidades distintas: Gene Ontology (THE GENE ONTOLOGY CONSORTIUM, 2000), Sequence Ontology (EILBECK et. al., 2005), TAMBIS Ontology (BAKER et.al., 1998), RiboWeb (ALTMAN et.al., 1999), EcoCyc Ontology (KARP, 2000), The Ontology for Molecular Biology (MBO) (SCHULZE-KREMER, 1998). Uma revisão destas ontologias, exceto a Sequence Ontology, pode ser encontrada em (STEVENS et. al., 2000).

A OBO (*Open Biomedical Ontologies*) é um esforço colaborativo, em que participam vários grupos (NCBO, 2007), (GO, 2007), (ORG, 2007), (BBOP, 2007), no sentido de produzir vocabulários bem estruturados para uso entre diversos domínios médicos e biológicos. Desta forma, a OBO inclui diversas ontologias, onde algumas são aplicadas para a descrição de quaisquer organismos e outras para espécies determinadas, como as plantas. Podemos citar como ontologias mais disseminadas da OBO, a Gene Ontology e a Sequence Ontology.

2.4.1 GENE ONTOLOGY

O projeto da Gene Ontology (GO) é um Consórcio criado para suprir a necessidade de descrições consistentes para produtos de genes em diferentes bancos de dados. O que é observado em cenários onde não é feito o uso da ontologias, é um grande gasto de tempo e esforço na busca de informações nas bases de dados genômicas, uma vez que existe uma ampla variação na terminologia dos dados armazenados.

Percebe-se que a anotação feita com os termos da GO é bastante útil, pois, uma vez que as seqüências são disponibilizadas nos bancos de dados públicos, torna-se mais clara a identificação das mesmas se ela for anotada por termos da GO, isto porque estes

fazem parte de um vocabulário controlado que todos os centros de pesquisa podem ter acesso.

A GO possui três estruturas, denominadas vocabulários controlados (ontologias), que descrevem produtos de genes em termos de acordo com suas associações. São elas:

Componente celular: refere-se ao componente de uma célula, mas com o detalhe que pode ser parte de algum objeto maior, que pode ser uma estrutura anatômica;

Processo biológico: série reconhecida de eventos ou funções moleculares;

Função molecular: descreve atividades que ocorrem no nível molecular.

A existência de três ontologias se deve aos seguintes aspectos: o desenvolvimento e manutenção das próprias ontologias; a anotação de produtos de genes, os quais fazem associações entre as ontologias e os genes e produtos de genes na colaboração entre bancos de dados; e o desenvolvimento de ferramentas que facilitam a criação, manutenção e uso das ontologias.

Cada termo da GO possui um identificador numérico na forma GO:nnnnnnn, um nome, e uma associação para uma das três ontologias. A maioria dos termos ainda possui uma descrição textual e alguns termos possuem sinônimos. Os termos são ligados por dois tipos de relacionamentos: “é_um” e “parte_de”. As ontologias são ligadas através de um grafo direcionado acíclico, que se difere de outras hierarquias no sentido em que um “filho” pode ter vários “pais”.

A FIG. 2.9 ilustra um exemplo da Gene Ontology para a ontologia processo biológico, o qual descreve o metabolismo do DNA (*DNA Metabolism*). Note que um nó pode ter mais que um pai, como por exemplo, *mitochondrial DNA-dependent DNA replicaton* tem como pais *mitochondrial genome maintenance e DNA-dependent DNA replicaton*.

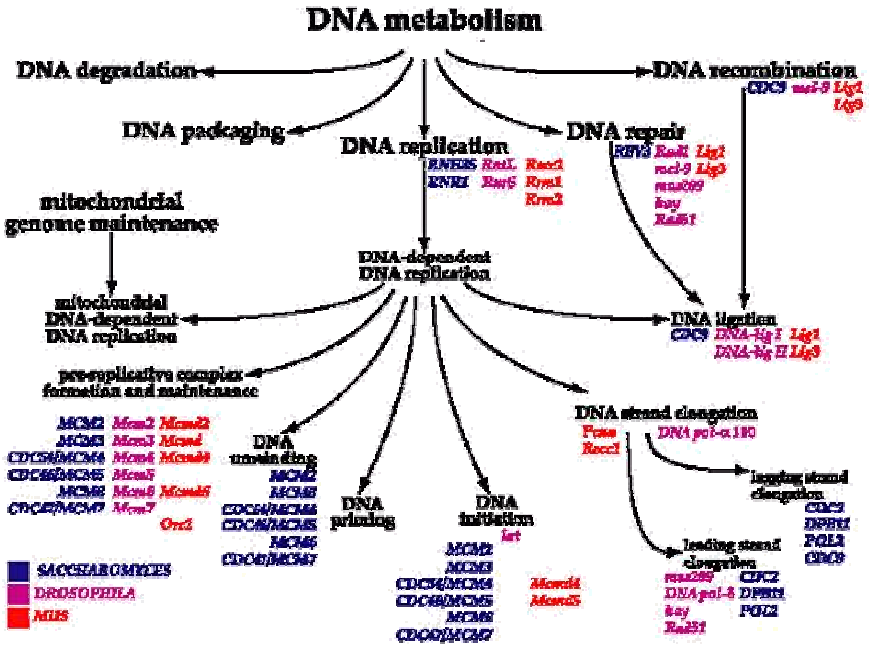


FIG. 2.9 Exemplo da Gene Ontology (THE GENE ONTOLOGY CONSORTIUM, 2000)

O *Web-Site* AmiGO (AMIGO, 2005) é o visualizador oficial da GO, onde pode-se pesquisar por termos ou genes ou proteínas e verificar sua localização no grafo direcionado acíclico. A TAB. 2 resume o conteúdo da ontologia até o ano de 2005, de acordo com (THE GENE ONTOLOGY CONSORTIUM, 2006):

TAB. 2.2 Conteúdo da GO (THE GENE ONTOLOGY CONSORTIUM, 2006)

Biological process terms	9805
Molecular function terms	7076
Cellular component terms	1574
Sequence Ontology terms	963
Genomes with annotation[a]	30
Annotated gene products	
Total	1 618 739
Electronic only	1 460 632
Manually curated	158 107

[a]Excludes annotations from UniProt, which represent 261 annotated proteomes.

Pode-se observar que a quantidade de produtos de gene anotados com a GO passa de um milhão. E este número tende a crescer, assim como o número de termos das ontologias da GO (função molecular, processo biológico e componente celular). Isto porque no decorrer do tempo e com o avanço e crescimento das pesquisas genômicas, surgem novas descobertas de genes e suas funções. Assim, o Consórcio da GO se atualiza em relação às novas descobertas para posteriormente disponibilizar os novos termos para comunidade de Bioinfomática.

Todas as mudanças nas ontologias (incluindo as atualizações) são coordenadas através do *GO Editorial Office*. As mudanças são propostas pelos curadores da GO, anotadores de organismos modelos e pesquisadores em geral envolvidos com esta área. Os curadores da GO, através de um sistema *online*, provido pelo *SourceForge* (GO SOURCE FORGE, 2006) recebem as solicitações de mudanças e inserção de novos termos nas ontologias. Dados também do ano de 2005, apontam que cerca de 2800 requisições foram postadas e destas, cerca de 2100 foram atendidas. É de total responsabilidade dos pesquisadores terem a iniciativa de solicitar mudanças e contribuir para a evolução da GO. Se a maioria das requisições feitas é atendida, para garantir o sucesso de uso da GO é importante que os anotadores se preocupem com esta tarefa e passem a contribuir mais.

2.4.2 SEQUENCE ONTOLOGY

O Projeto da Sequence Ontology (SO) é uma junção de esforços de centros de anotação genômica como *WormBase*, *FlyBase*, *Mouse Genome Informatics* e *Sanger Institute*, cujo objetivo é desenvolver uma ontologia satisfatória para a descrição de seqüências biológicas.

Assim, a Sequence Ontology descreve através de um conjunto de termos as características e atributos de seqüências biológicas, tais como *hits* de similaridade de nucleotídeos e interpretações de genes. Ela oferece recursos tais como:

- Prover um vocabulário controlado estruturado para a descrição de DNA e RNA e para a descrição de mutações das seqüências e;
- Prover uma representação estruturada dessas anotações nos bancos de dados.

Desta maneira, a SO facilita a troca, a análise e o gerenciamento de dados genômicos. Duas versões da Sequence Ontology englobam:

Versão completa: chamada SO e possui mais de cem termos, os quais são em geral utilizados em projetos de anotação genômica onde as seqüências são fortemente curadas;

Versão SOFA: termos que podem ser diretamente localizados na seqüência biológica, é a versão mais utilizada da SO;

A SO, assim como a GO é uma ontologia que recebe sugestões de atualizações e mudanças. Novos termos e definições são propostos, debatidos e aprovados ou rejeitados por um grupo aberto de indivíduos via lista de e-mail.

Uma versão bastante disseminada é a SOFA (*Sequence Ontology Feature Annotation*), a qual inclui somente características localizáveis das seqüências e é projetada para uso onde as saídas são no formato GFF3. Por este motivo que todos os termos da SOFA são "*unix-friendly*", por exemplo: não contêm espaço em branco, nunca iniciam com um número e não incluem caracteres como aspas ou hífens.

SOFA é um subconjunto da SO e todos os termos da SOFA que estão também na SO são marcados com uma *tag* "SOFA". Esta versão tende a ser mais estável que a versão completa da SO.

Atualmente a SO usa três tipos básicos de relacionamentos entre seus termos:

"tipo_de": especifica o que alguma coisa "é";

"deriva_de": utilizado para denotar relacionamentos de processos entre dois termos;

"parte_de": denota relacionamentos onde um termo faz parte de outro.

A FIG. 2.10 ilustra os vários relacionamentos entre alguns de seus termos:

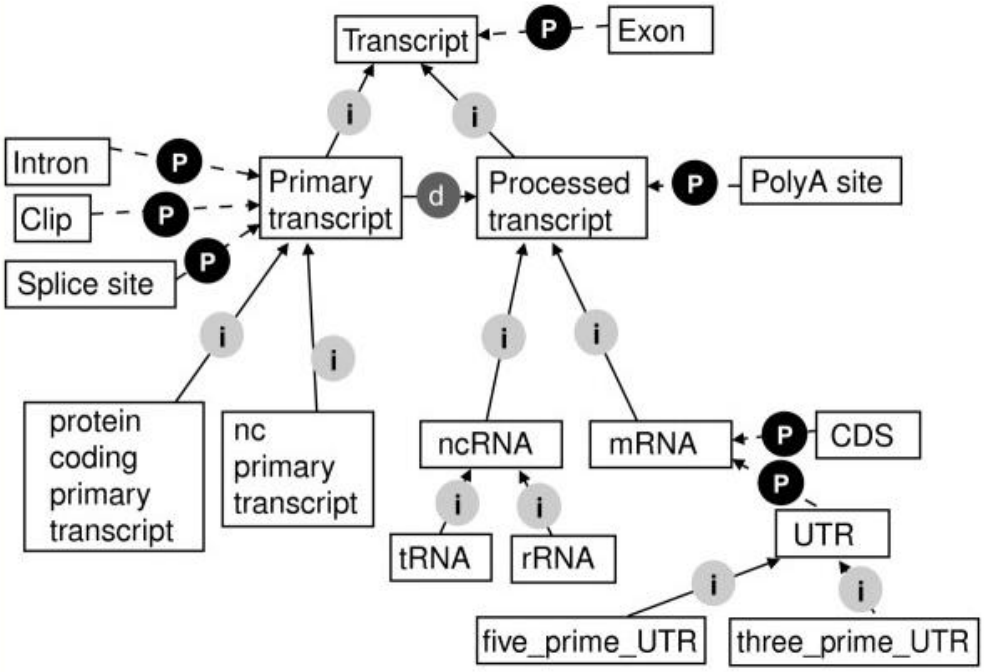


FIG. 2.10 Relacionamentos entre termos da SO (EILBECK et al, 2005)

A FIG. 2.10 mostra como os termos e relacionamentos são usados juntos para descrever conhecimento entre seqüências. Os relacionamentos são mostrados através das flechas, onde as que contêm a letra "i" são relacionamentos "tipo_de", a letra "P", os relacionamentos "parte_de" e a letra "d" os relacionamentos "deriva_de". Através

das setas tracejadas que conectam os termos, diferentes inferências lógicas podem ser feitas considerando-se o que o termo representa.

Uma das formas de se visualizar a Sequence Ontology é através do *Web-site* miSO (MISO, 2007), que provê um meio gráfico para buscar os termos da ontologia. A visualização é feita através de uma abordagem pai-filho. A apresentação dos termos é feita em detalhes mostrando os tipos de relacionamentos, a definição, os sinônimos e as referências.

2.5 CONSIDERAÇÕES

O objetivo principal deste capítulo foi definir anotação genômica, como capturá-las e representá-las. Para isto, tornou-se necessário caracterizar os bancos de dados genômicos; e como as anotações são representadas nestes bancos de acordo com os formatos pré-definidos. Foram levantados também exemplos de sistemas de anotação.

Como dito anteriormente, para que haja colaboração no processo de anotação genômica, é ideal que haja um vocabulário comum entre todos que participam. Percebemos que os sistemas de anotação tendem a não fazer uso de ontologias e compartilhar o conhecimento sobre o uso dos termos, o que também dificulta a colaboração.

Foram então descritos exemplos de ontologias já em uso pela comunidade de Bioinformática, no intuito de apresentar como estas são utilizadas, favorecendo assim um entendimento sobre as mesmas para propor sugestões que ampliem o uso das ontologias bem como proporcionem a sua evolução.

3 O PROCESSO DE ESTUDO DE SEQÜÊNCIAS E ANOTAÇÕES

Para caracterizar o processo de anotação genômica e a colaboração inerente a ele, é preciso resgatar o contexto onde este processo está inserido. O processo de anotação genômica faz parte de um processo mais amplo que é o de estudo de seqüências genômicas. Assim sendo, neste capítulo procuramos caracterizar o processo de estudo de seqüências, através de seu estudo e modelagem.

A proposta da caracterização e descrição do processo é ser genérica, de modo a representar o processo independente do foco principal de estudo de cada grupo, capturando as atividades comuns e mais relevantes. Apoiamos este trabalho através da literatura e descrições existentes sobre os processos utilizados por projetos de pesquisa (seção 3.1) e na observação específica do trabalho desenvolvido pelo grupo de pesquisa da Fiocruz que utiliza o sistema GARSA.

O grupo de pesquisa atua no Laboratório de Biologia Molecular de Tripanosomatídeos do DDBM/IOC/FIOCRUZ, um centro de pesquisa representante de estudos da área de biologia molecular. Um fator importante pelo qual o estudo apoiou-se no GARSA é por este reunir grande parte das funcionalidades de anotação, como anotações automáticas, manual e baseada em ontologia. Outro fator importante se refere ao centro de pesquisa possuir parceria com pesquisadores do Instituto Militar de Engenharia, o que tornou facilitado o acesso ao grupo, onde observações dos trabalhos desenvolvidos por eles foram possíveis, bem como reuniões e também o acesso ao sistema GARSA. Este acesso veio suprir a necessidade de se ter um ambiente para testar o apoio à colaboração, validar a hipótese deste trabalho e desenvolver o protótipo proposto neste.

A partir do estudo das atividades desempenhadas pelos grupos de pesquisa estudados e da constatação de que estes desempenham atividades semelhantes, tornou-se viável a construção de um modelo representativo do processo do estudo de seqüências genômicas, como apresentado na seção 3.2. No caso particular da colaboração, sugere-se que o modelo de um processo possa ser analisado e projetado com vias a buscar oportunidades de ampliar a colaboração existente, bem como identificar requisitos específicos de apoio computacional à colaboração. Para a construção deste modelo foi importante identificar quais são as características e/ou funções das pessoas que desenvolvem estas atividades e quais são suas

responsabilidades dentro do grupo de pesquisa, bem como definir a modelagem de processos. A seção 3.2 também apresenta estes itens.

O processo modelado do estudo de seqüências genômicas cobre desde o recebimento dos arquivos que contém as seqüências vindas da bancada de experimentos biológicos, até a sua disponibilização (juntamente com as respectivas anotações) através de bancos de dados públicos. Em um primeiro momento, todas as atividades deste processo são descritas de forma breve e geral. Em um segundo momento, essas mesmas atividades são detalhadas, vistas como sub-processos, contendo suas próprias atividades. Nesta modelagem, foi dada maior ênfase no processo de anotação genômica, onde as atividades são explicadas de forma mais detalhada.

Após a modelagem dos processos, é citada também na seção 3.2 a ontologia como um recurso computacional para apoiar o processo de anotação genômica e para finalizar, de forma a complementar o entendimento do processo de estudos de sequência genômica e o uso das ferramentas que o apóiam, na seção 3.3 é mostrado em detalhes como este processo é “executado” no contexto do grupo da Fiocruz. Isto é feito através do uso do Sistema GARSA e a partir de exemplos reais de um projeto de pesquisa.

3.1 WORKFLOWS PARA ESTUDO DE PESQUISAS GENÔMICAS

A pesquisa de organismos para descoberta do genoma é desempenhada por centros de pesquisas, que podem ser laboratórios ou universidades, com ou sem fins lucrativos e que são constituídos por grupos de pesquisas. O principal centro de pesquisa genômica é o NCBI (comentado no capítulo 2). No Brasil, existem diversos centros de pesquisas que podem se diferenciar de acordo com o organismo que estudam, mas todos são voltados para as pesquisas na área de biologia molecular, como o estudo de seqüências de nucleotídeos e proteínas.

Os centros de pesquisa, em geral, possuem definidos um processo ou *workflow* para o estudo de seqüências genômicas e, por conseguinte para a anotação genômica. Apesar de cada centro de pesquisa enfocar o que lhe é mais importante, as atividades desempenhadas para estes estudos são semelhantes, o que possibilita uma modelagem genérica dos processos. A seguir são citados alguns dos *workflows* para o estudo de seqüências genômicas desenvolvidos e em uso por alguns centros de pesquisa no Brasil.

3.1.1 WORKFLOW MHOLLINE

O Instituto de Biofísica Carlos Chagas Filho da Universidade Federal do Rio de Janeiro (IBCCF/UFRJ) é um centro de pesquisa que desenvolve pesquisas em Bioinformática. O *workflow* científico, denominado *MholLine* que foi desenvolvido pelo IBCCF tem como objetivo a predição e construção de modelos tridimensionais de proteínas a partir de uma seqüência de aminoácidos. Este tipo de experimento tem grande importância na descoberta das funções dos genes dos organismos. O *workflow MholLine* através da combinação de programas, como o BLAST, visa determinar a estrutura mais confiável para uma seqüência de moléculas utilizando como modelos, estruturas de proteínas relacionadas (SANTOS, 2004). A combinação de programas deste *workflow* pode ser verificada na FIG. 3.1:

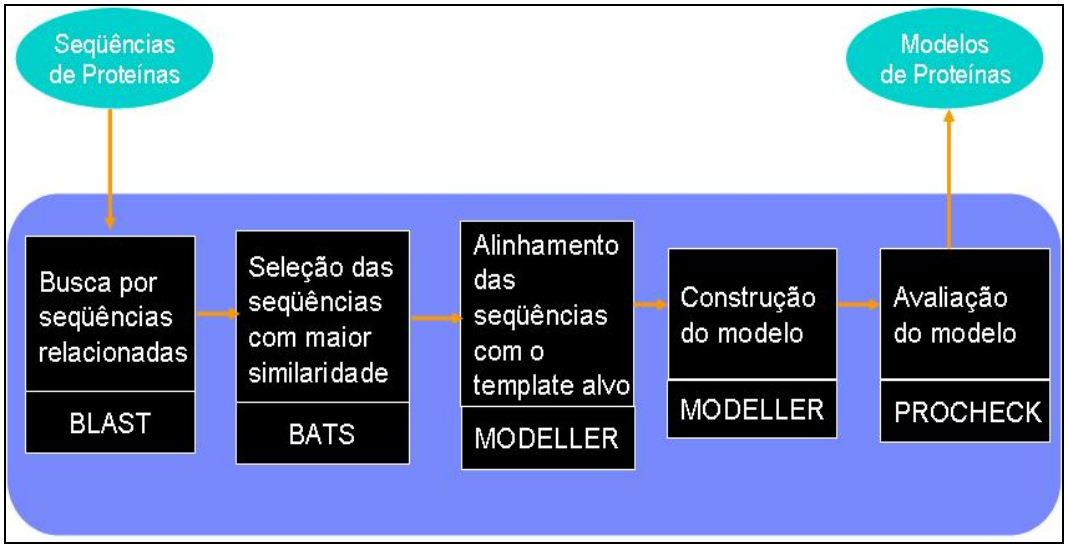


FIG. 3.1 Seqüências do *Wokflow MholLIne* (SANTOS, 2004)

Para iniciar o processo realizado pelo *workflow MholLine* as seqüências devem ser traduzidas de nucleotídeos para aminoácidos e assim tem-se as proteínas codificadas. Então, estas seqüências de proteínas são submetidas ao primeiro passo do *workflow* que irá executar o programa BLAST para buscar seqüências que estejam relacionadas por similaridade com as seqüências de entrada. Estas seqüências similares servem como base para a execução do programa seguinte, o BATS (RÖSSLE, 2004), que através de outros parâmetros busca dentre aquele conjunto de seqüências de proteínas similares, quais possuem um melhor padrão de similaridade, fazendo assim um refinamento das seqüências inicialmente selecionadas. As seqüências selecionadas pelo BATS são então

alinhadas de acordo com o modelo (*template*) alvo através do programa MODELLER (SALI, 2001). Este mesmo programa irá então construir o modelo (tridimensional) para tal proteína. Este modelo entra como dado do próximo programa, PROCHECK (LASKOWSKI et al., 1994), que irá avaliá-lo, para verificar se o modelo é uma estrutura possível de existir na natureza. Assim, a saída do *workflow* são modelos de proteínas tridimensionais.

3.1.2 WORKFLOW DO PROJETO GENOPAR

O Projeto GENOPAR (Projeto Genoma do Paraná) reúne pesquisadores e recursos dos laboratórios do Paraná, de Santa Catarina e do Rio Grande do Sul. Essas entidades juntas puderam implantar uma rede de laboratórios com equipamentos de última geração para estudos de genômica e biologia molecular. Faz parte dos estudos do Projeto GENOPAR o seqüenciamento da bactéria fixadora de nitrogênio *Herbaspirillum seropedicae* (GENOPAR, 2005).

Para o seqüenciamento de procariotos (bactérias), o Projeto GENOPAR tem usado, uma estratégia de seqüenciamento misto, onde a maior parte da seqüência do genoma é obtida por exaustivo seqüenciamento aleatório do genoma inteiro. Entretanto, na prática, essa estratégia não se mostra satisfatória para o completo fechamento da seqüência de um genoma, mesmo de procariotos. Por esta razão, ela deve ser complementada com a estratégia de seqüenciamento aleatório de clones, que corrige possíveis falhas deixadas pelo seqüenciamento aleatório do genoma. Um esquema geral das etapas de um projeto de seqüenciamento genômico deste tipo é mostrado na FIG. 3.2 (GENOPAR, 2005).

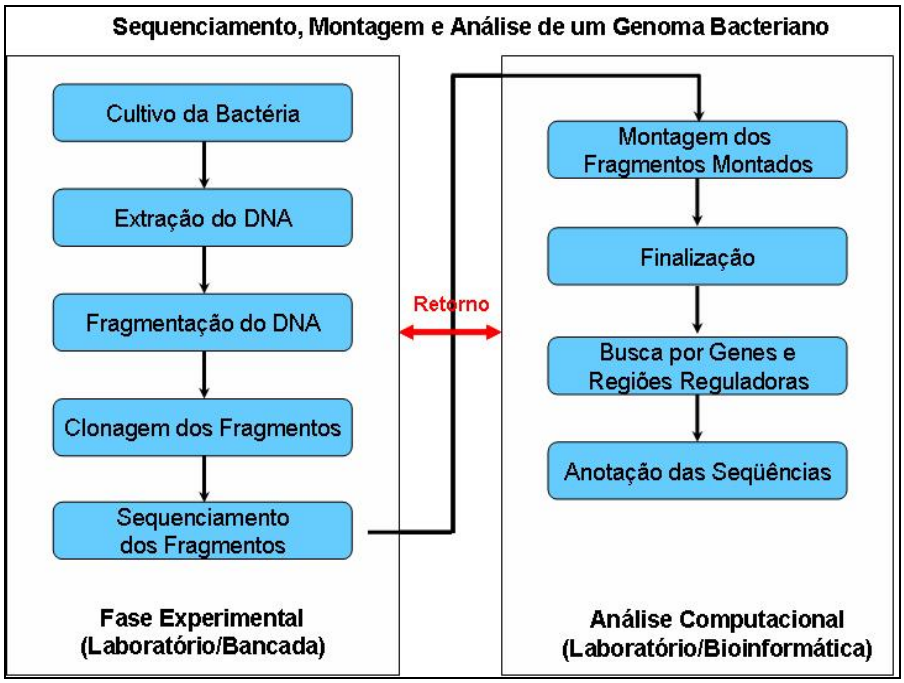


FIG. 3.2 Workflow do Projeto GENOPAR (GENOPAR, 2005)

Na fase experimental, após o cultivo da bactéria e a extração de seu DNA, o genoma a ser seqüenciado é fragmentado, os fragmentos são clonados e os clones seqüenciados. As seqüências geradas na fase experimental são analisadas na fase computacional, com o objetivo de remontar o genoma a partir dos fragmentos seqüenciados e, posteriormente, analisar este genoma quanto a presença de genes, regiões promotoras, seqüências repetitivas, RNAs estáveis, etc (GENOPAR, 2005). A análise do genoma tem seu término na anotação das seqüências. Os passos da fase computacional são realizados através do uso de programas de Bioinformática.

3.1.3 WORKFLOW SABIA

O SABIA (ALMEIDA et al, 2004) foi desenvolvido pelo Laboratório de Bioinformática (LABINFO, 2005) que faz parte do Laboratório Nacional de Computação Científica (LNCC) e possui como objetivo a montagem e anotação automática de genomas procariotas. Ele executa tarefas automáticas de análise para montagem, identificação de ORFs e regiões extragênicas. A FIG. 3.3 dá uma visão geral do *workflow*.

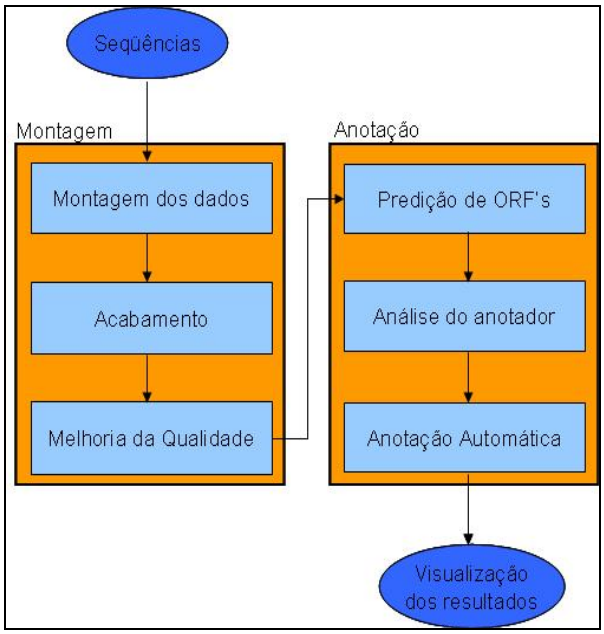


FIG. 3.3 Visão geral do workflow SABIA

O *workflow* SABIA é dividido em duas partes principais: montagem e anotação. Na parte de montagem, as seqüências são submetidas ao sistema e este utiliza o pacote PHRED/PHRAP/CONSED (PHRAP, 2007) para montar os dados e assim obter os *contigs*. A tarefa mais importante na montagem é o acabamento, o qual envolve fechamento de lacunas e melhoria da qualidade da seqüência.

Na parte de anotação todas as possíveis ORFs dos *contigs* gerados na parte de montagem são preditos. Cada ORF é submetida a vários bancos de dados para comparação e os resultados são apresentados para então serem avaliados pelo usuário/anotador. Na etapa de anotação automática são usados programas para busca de similaridade como o BLAST; identificação de proteínas como o Interpro (MULDER, et al., 2005), classificação funcional utilizando os bancos de dados KEGG (KEGG, 2005) e COG (TATUSOV et al., 2000) e de busca do termo correspondente usando a *Gene Ontology*.

3.1.4 WORKFLOW GARSA

O sistema constitui-se de um *workflow* composto de pacotes selecionados de *softwares* de Bioinformática (DÁVILA et al, 2005), como pode ser visto na FIG. 3.4.

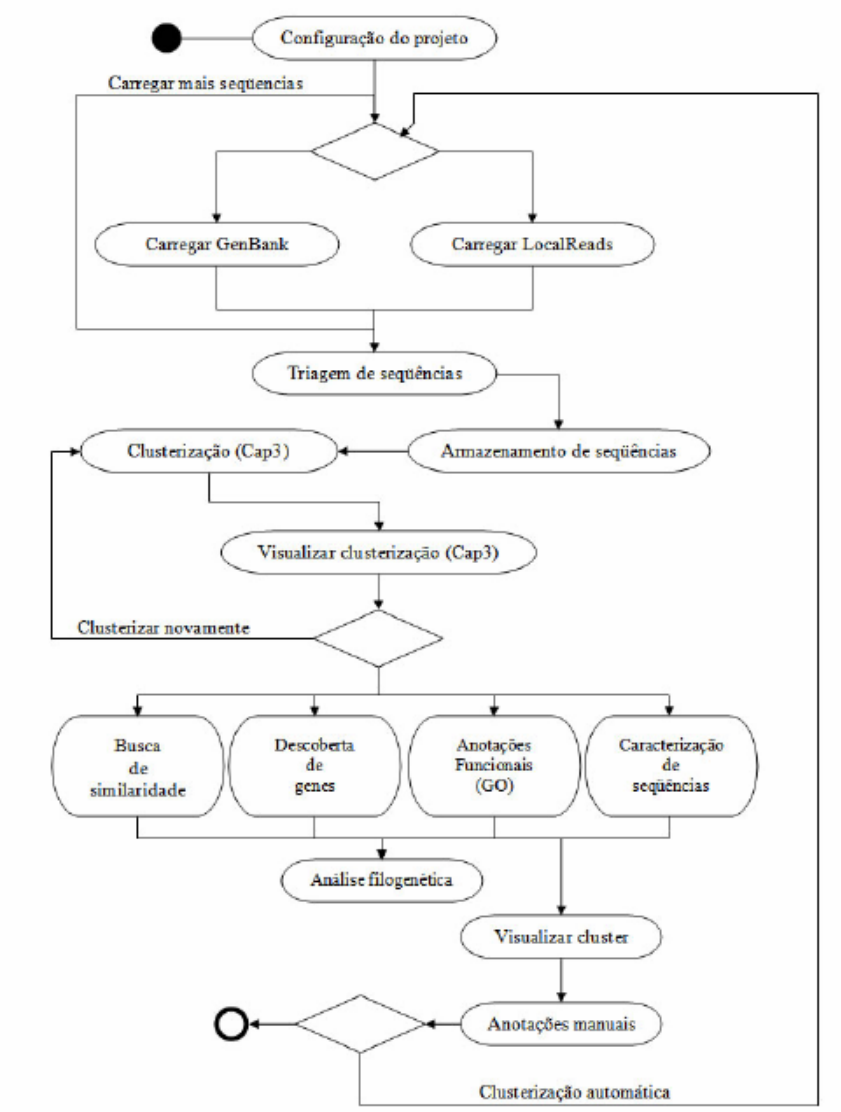


FIG. 3.4 Workflow do GARSA (GARSA, 2005)

O *workflow* do GARSA parte da configuração do projeto, como escolha de nome e quais os pesquisadores envolvidos. As seqüências são carregadas através do *GenBank* ou de *reads* oriundos de um sequenciamento. Estas seqüências passam então por uma limpeza (triagem) e são armazenadas. A seguir, as seqüências passam por um processo de clusterização, quando se gera uma sequência consenso. Logo após, são realizadas buscas de similaridades com seqüências de outros bancos de dados, execução de programas para descobrir regiões codificantes, anotação baseada em ontologia através da GO e descobertas das características das seqüências (anotação automática). Após a busca de similaridade e a descoberta de genes terem sido concluídas, a análise filogenética pode ser realizada. Por fim, após a visualização de todos os resultados

obtidos para um cluster (*Visualize cluster*), uma anotação manual é realizada para complementar as informações obtidas.

3.2 MODELAGEM DO PROCESSO DE ESTUDO DE SEQÜÊNCIAS GENÔMICAS

A modelagem do estudo de seqüências genômicas foi realizada para facilitar a visualização das atividades de cada processo, a identificação dos papéis que participam das atividades, e posteriormente a identificação de ferramentas que apóiam o processo em questão, permitindo desta forma uma caracterização bem definida do ambiente. As seções seguintes descrevem, portanto, sobre a caracterização de papéis e a modelagem de processos.

3.2.1 MODELAGEM DE PROCESSOS

O termo modelagem de processos tem sido associado com um número de idéias, todas preocupadas com o comportamento dinâmico das organizações. Basicamente, tais comportamentos podem ser tratados como um número de processos que são inter-relacionados (SNOWDON, 2006).

Do ponto de vista de metodologias para a realização da modelagem de processos de negócio, existem diversas abordagens. Cada uma delas utiliza uma determinada notação e linguagem, mas em geral, abrangem o mesmo conjunto de passos/atividades, a saber (SHARP E MCDERMOTT, 2001): 1) definir o escopo da modelagem; 2) documentar a missão, estratégia, metas e objetivos da organização; 3) compreensão do processo/organização “como está” (“as-is”); 4) avaliação do modelo obtido; 5) decidir quanto a: abandonar, contratar, manter como está, melhorar ou redesenhar os processos; 6) discutir e gerar idéias; 7) determinar características do processo desejado; 8) projetar os processos desejados (“to-be”).

Nas seções a seguir é explorado o uso da modelagem de processos no entendimento de um contexto de trabalho e a identificação de oportunidades de apoio à colaboração.

O processo de estudo das seqüências genômicas modelado cobre desde o tratamento das seqüências digitalizadas até a finalização e publicação das anotações, constituindo assim somente a parte dos experimentos *in-silico*. Contudo é preciso entender um pouco as atividades que são realizadas na bancada “molhada” (*in-vitro*), cujas atividades não são modeladas, devido a essa parte do trabalho não corresponder ao nosso alvo de estudo.

3.2.2 PAPÉIS

No ambiente dos projetos de pesquisa genômica, vários bioinformatas podem estar envolvidos, trabalhando separadamente ou em grupo, com base em um *workflow* definido (como os exemplos descritos anteriormente). Dessa forma, para entender o processo do estudo de seqüências genômicas há a necessidade de identificar os papéis de quem os realiza. Nesse contexto, são especificados dois papéis principais: anotador e curador.

Anotador: O anotador é um pesquisador que estuda o seqüenciamento de genomas, e que faz a anotação das seqüências de acordo com o seu grupo de pesquisa e de acordo com os padrões pré-definidos da comunidade de Bioinformática (para posterior submissão das seqüências para bancos de dados públicos). Profissionais de diferenrtes áreas, mas que estejam envolvidos no âmbito da pesquisa genômica e que tenham conhecimento suficiente, podem realizar estudos de seqüências, e fazer anotação, como por exemplo, médicos ou outros profissionais da saúde. Contudo, na maior parte das vezes, o anotador é um biólogo, que é responsável pela maior parte das tarefas, seja relacionada aos experimentos biológicos ou à atividades do *workflow* do projeto, como a exeução dos programas de análise das seqüências e anotações.

Curador: O curador é um pesquisador mais especializado, com experiência reconhecida na área, possuindo desta forma um maior conhecimento do assunto ou do organismo em estudo. É responsável por supervisionar os resultados das atividades do *workflow* e avaliar as anotações realizadas. O processo realizado por ele é conhecido por “curagem”. Outras responsabilidades incluem coordenar a anotação individual ou do grupo de pesquisa, verificar as informações anotadas fazendo as devidas observações, críticas, sugestões, e se for o caso, corrigir, até que a anotação esteja devidamente correta. É ideal que toda seqüência anotada passe por um curador antes de ser publicada.

3.2.3 BANCADA MOLHADA (*IN-VITRO*)

Na bancada de experimentos biológicos são estudadas cepas do organismo que se deseja pesquisar, de onde é retirado o DNA para ser analisado. O DNA é cortado em pontos específicos, e assim têm-se fragmentos de DNA de diversos tamanhos. Esses fragmentos são levados para um aparelho chamado seqüenciador.

O seqüenciador gera como resultados diversos arquivos binários, de acordo com a quantidade de amostras fragmentadas. Os arquivos binários são então analisados por um *software* específico que irá inferir a seqüência de nucleotídeos, correspondentes ao código genético, gerando cromatogramas ou *reads* correspondente a cada arquivo. Desta forma o DNA passa a ser representado por uma cadeia de caracteres que representam as bases ou nucleotídeos, são elas: A (adenina), C (citosina), G (Guanina), T(timina).

O momento em que é feita a conversão dos arquivos binários em cadeias de caracteres, representando os nucleotídeos, podemos considerar como o início dos processos de trabalhos na bancada de experimentos computacionais ou experimentos *in-silico*, pois começam a fazer parte do processo, os programas de Bioinformática.

3.2.4 BANCADA DE EXPERIMENTOS IN SILICO

O primeiro processo *in-silico* modelado se refere ao estudo das seqüências genômicas de forma geral. O objetivo é modelar o contexto em que o processo de anotação genômica está inserido, sendo visto os processos que são realizados antes e depois deste.

A **Erro! Fonte de referência não encontrada.** mostra uma visão geral do processo, o qual é composto das atividades: registrar recebimento, pré-análise, análise e anotação, e consolidação da anotação. As atividades de análise e anotação podem ser realizadas em paralelo por serem complementares: à medida que se faz análise se realiza anotação de acordo com os resultados obtidos na análise, e à medida que a anotação não é satisfatória ou suficiente, novas análises são realizadas.

Observe na figura que os papéis de anotador e curador aparecem na realização de todo o processo. Porém o anotador é responsável por quase todas as atividades dos processos, enquanto que o curador participa das atividades em que envolvem tomadas de decisões ou validações, que serão evidenciadas nos subprocessos, como detalhado mais adiante.

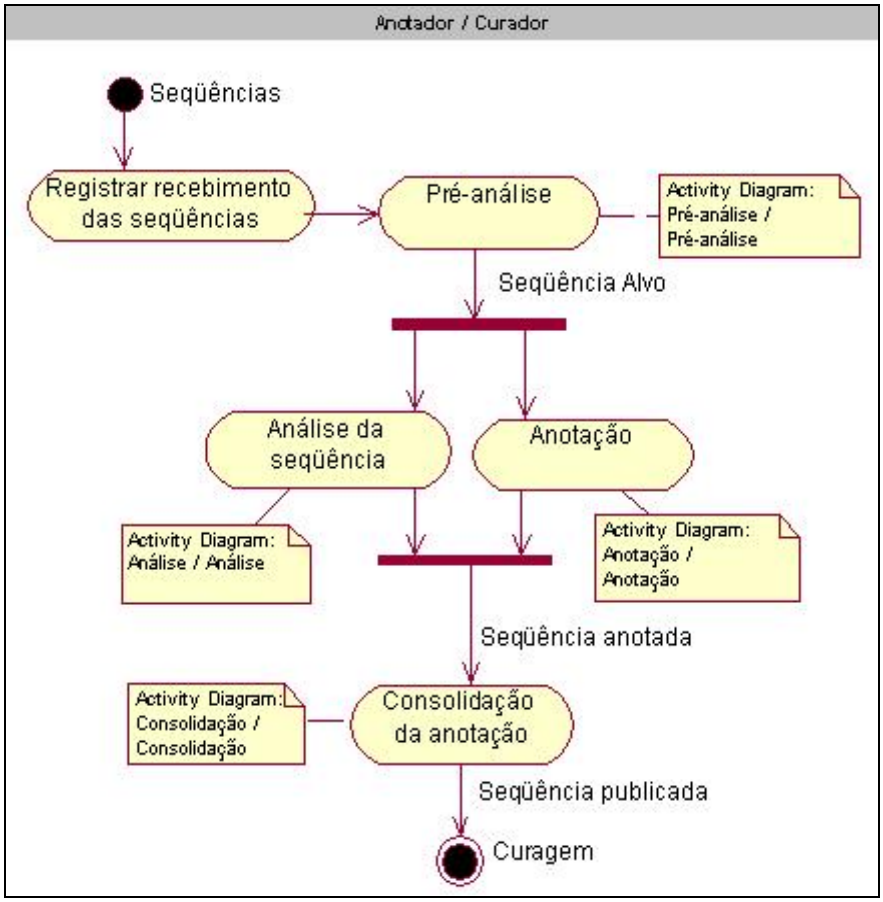


FIG. 3.5 Visão geral dos processos para estudo de seqüências genômicas

3.2.4.1 REGISTRAR RECEBIMENTO DAS SEQÜÊNCIAS

A atividade possui como objetivo fazer o registro de todas as seqüências que vieram da bancada de experimentos biológicos para serem estudadas. O registro é feito nos sistemas de apoio ao estudo das seqüências, como por exemplo, um sistema de anotação. Em geral, este registro consta da atribuição de um nome para cada seqüência, de acordo com projeto, o grupo de pesquisa e o organismo em questão. A atividade possui como entradas as seqüências fragmentadas (representadas através de arquivos digitais que contêm a seqüência de nucleotídeos na forma binária) e como saída, as seqüências registradas com a denominação específica.

3.2.4.2 PRÉ-ANÁLISE

Após o registro do recebimento das seqüências vindas da bancada de experimentos biológicos, as seqüências de nucleotídeos passam a ser entradas de um processo denominado de “pré-análise”, como mostrado na FIG. 3.6. Pode-se dizer que as atividades que compõem este processo servem para preparar as seqüências de forma que se consiga um conjunto de seqüências não redundantes, as quais possam ser estudadas no sentido de descobrir o que elas representam. A saída deste processo são as seqüências alvo, e as atividades acontecem como descritas a seguir:

Limpar seqüências: é a primeira atividade do processo de pré-análise e tem por objetivo fazer a limpeza da seqüência fragmentada retirando as partes que correspondem à seqüência de vetores. As seqüências fragmentadas são as entradas desta atividade e a saída são as mesmas, limpas dos nucleotídeos referentes aos vetores.

Avaliar qualidade das seqüências: o objetivo desta atividade é verificar se a seqüência possui uma boa qualidade para ser posteriormente analisada. Seqüências pequenas em geral, não possuem boa qualidade e então são descartadas, assim como as seqüências redundantes. A entrada desta atividade são as seqüências fragmentadas, limpas de vetores e a saída são as seqüências de boa qualidade.

Agrupar seqüências semelhantes: esta atividade possui como objetivo juntar todas as seqüências cuja maioria dos nucleotídeos seja igual. Isto acontece, pois as técnicas de replicação do DNA (tarefa que ocorre na bancada de experimentos biológicos) geram muitas seqüências que podem ser semelhantes, codificando assim o mesmo gene. Desta forma, estas sequências não precisam ser estudadas separadamente. A entrada desta atividade são as seqüências de boa qualidade e a saída são as seqüências agrupadas ou *clusters*. Para cada grupo de seqüências clusterizadas, uma seqüência denominada de seqüência alvo é representativa do cluster. Portanto, como são diversos clusters, gera-se então várias seqüências alvo.

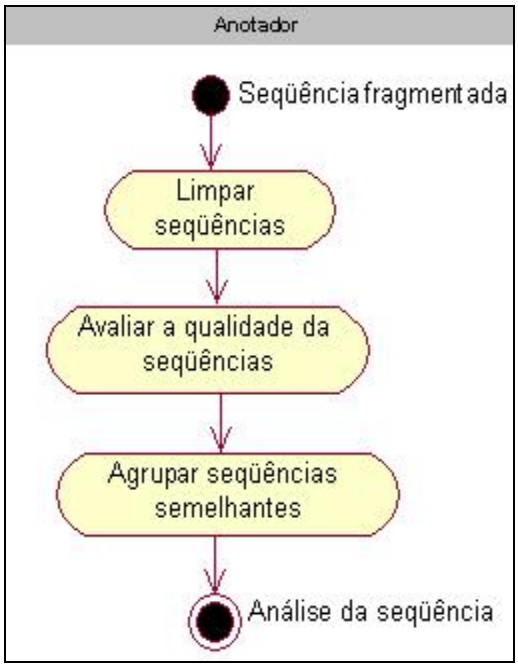


FIG. 3.6 Atividades da pré-análise

3.2.4.3 ANÁLISE DA SEQÜÊNCIA

Na continuidade do processo do estudo de seqüências genômicas, encontra-se o processo de “análise da seqüência”, cujas atividades estão ilustradas na FIG. 3.7. Cada atividade tem como objetivo identificar as características das seqüências. A partir destas características é possível saber se a seqüência codifica algum gene e qual a sua função, por exemplo. A entrada deste processo são as seqüências alvo e a saída são as mesmas analisadas, de acordo com as atividades a seguir:

As atividades iniciais da análise da seqüência são: “fazer predição de genes” e “fazer busca por similaridade”. São atividades que podem ser realizadas em paralelo, pois elas independem uma da outra.

Fazer predição de genes: possui como objetivo encontrar os possíveis genes codificantes de proteínas nas seqüências. A entrada da atividade é a seqüência alvo e as saídas são os genes possivelmente identificados e o seu posicionamento na seqüência.

Fazer busca de similaridade: esta atividade tem como objetivo fazer buscas por similaridades com outras seqüências já estudadas e armazenadas em banco de dados públicos ou privados. Também possui como entrada a seqüência alvo, e como saída todas as seqüências similares a ela.

Fazer alinhamentos múltiplos: os resultados da predição de genes e uma seleção das seqüências similares, juntamente com a seqüência alvo são as entradas para esta

atividade, que possui como objetivo alinhar todas as seqüências (similares e alvo) entre si, de forma a estudar posteriormente a origem comum de todas. Como saída, tem-se então este alinhamento.

Fazer análise filogenética: os alinhamentos múltiplos das seqüências passam a ser entradas da atividade seguinte, “fazer análise filogenética”, a qual tentará identificar a origem comum de todas as seqüências, caso exista. A saída desta atividade é uma árvore filogenética.

Desta forma, cada seqüência está analisada de acordo com estes distintos aspectos e então essas características podem ser observadas, criticadas e utilizadas, a fim de anotar a seqüência.

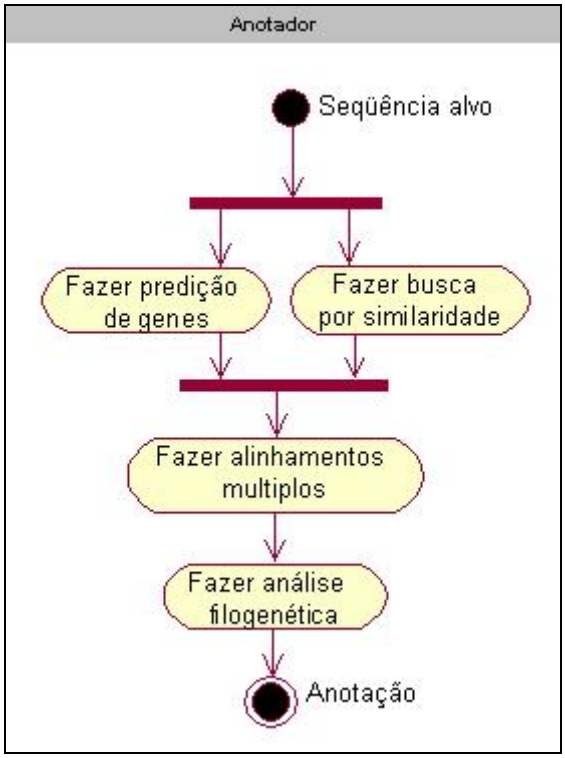


FIG. 3.7 Atividades da análise das seqüências

3.2.4.4 ANOTAÇÃO

Com o resultado do processo de análise das seqüências, informações podem ser inferidas sobre as seqüências, constituindo o processo denominado de “anotação”, cujo objetivo é descrever cada seqüência de acordo com as características que nela foram encontradas. Assim, a entrada deste processo é a seqüência alvo que passou por programas de análise e a saída é a seqüência anotada. O processo de anotação, foco

deste trabalho, foi modelado conforme mostra a FIG. 3.8 e é descrito de acordo com as atividades a seguir:

Observar e comparar resultados: a primeira atividade do processo é “observar e comparar os resultados”, advindos da análise da seqüência. Isto acontece para verificar se estes resultados estão satisfatórios ou não. Nesta atividade há uma interação entre anotador e curador a fim de chegarem a um consenso sobre o resultado. A entrada da atividade é a seqüência alvo e a saída, a conclusão dos resultados observados, se foram satisfatórios ou não.

Registrar anotação semi-automatizada: se os resultados dos programas de análise das seqüências são satisfatórios, então a atividade a ser realizada é “registrar anotação semi-automatizada”. Esta denominação deve-se ao fato de que durante esta atividade, além dos resultados automáticos gerados pelos programas do processo de análise da seqüência, há a verificação dos resultados um a um, e freqüentemente há a necessidade de intervenção manual, como por exemplo, buscas na literatura para confirmar tais resultados. O objetivo desta atividade é fazer o registro desta anotação, a qual é a principal para os grupos de pesquisa. A entrada da atividade é a seqüência alvo cujos resultados foram suficientes e a saída é a seqüência anotada.

Se os resultados não forem satisfatórios, as seqüências voltam para o processo de “análise da seqüência”, e são submetidas a novas execuções dos programas de análise (com outros parâmetros de entrada, por exemplo), sendo então, geradas novas seqüências alvo. Estes programas são citados no Anexo 1.

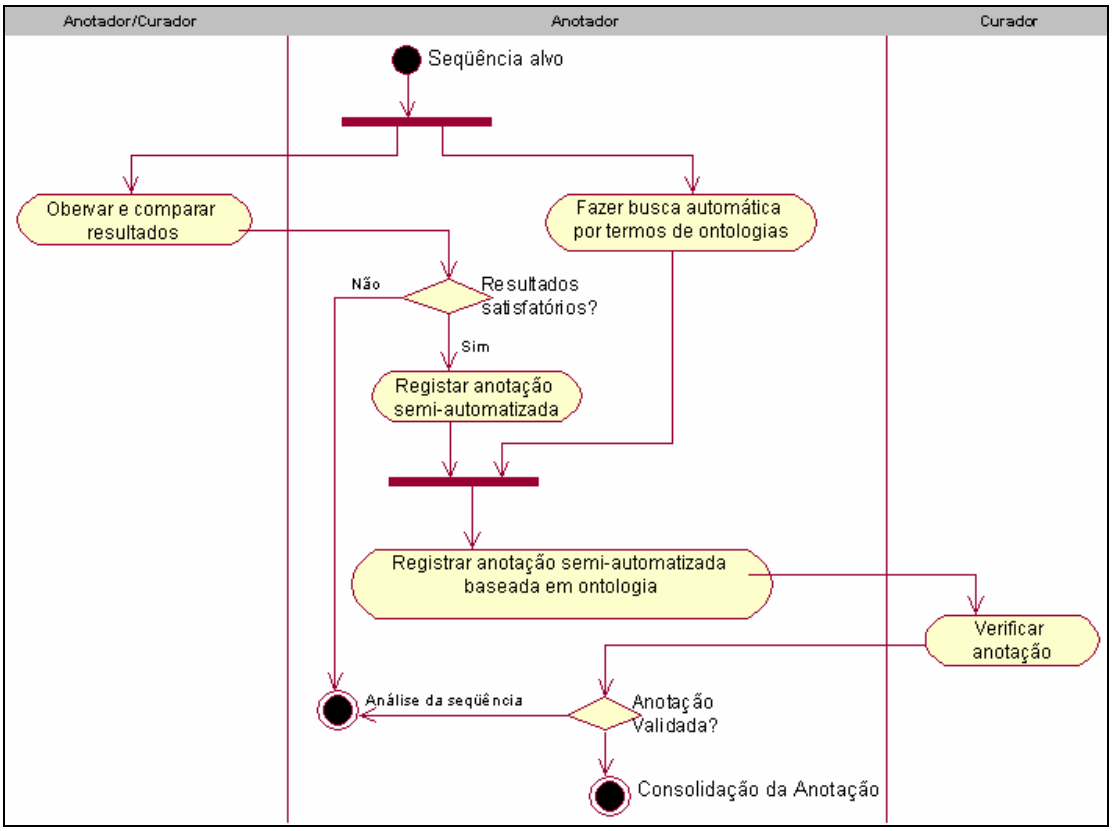


FIG. 3.8 Atividades da anotação

Fazer busca automática por termos de ontologia: paralelamente à observação dos resultados dos programas de análise da seqüência, outra atividade a ser realizada é "fazer busca automática por termos de ontologias", cujo objetivo é buscar automaticamente termos para anotar as seqüências através de ontologias. A entrada desta atividade é também a seqüência alvo, e a saída, a seqüência anotada por termos da ontologia. Uma das formas de se buscar o termo adequado é fazer uma busca de similaridade contra bancos de dados que a ontologia, por ventura, mantenha com organismos anotados com tais termos, ou contra bancos de dados curados, que disponibilizem suas seqüências anotadas com termos da ontologia. Os bancos de dados curados são utilizados por possuírem anotações que são de qualidade e por isso, confiáveis. Portanto, quando uma seqüência pesquisada é similar a outra seqüência destes bancos, os anotadores podem basear-se nos termos da ontologia usados nas anotações destas seqüências similares para anotar as seqüências que estão pesquisando. Nesta atividade, pode-se ainda fazer um refinamento dos termos encontrados, de modo a utilizar o termo mais adequado possível.

Registrar anotação semi-automatizada baseada em ontologia: a atividade seguinte é "registrar anotação semi-automatizada baseada em ontologia", na qual o

objetivo é unir estas duas formas de identificação das regiões codificantes da seqüência estudada: anotação semi-automatizada e anotação por termos da ontologia. Logo, estas duas anotações são as entradas desta atividade, e a saída é junção das anotações.

Verificar anotação: com as anotações realizadas parte-se então para a atividade "verificar anotação", onde é feita uma verificação da anotação para garantir que a mesma esteja correta. A verificação é feita pelo curador que orienta o anotador ou o grupo de pesquisa. A entrada desta atividade é a anotação semi-automatizada baseada em ontologia e a saída, esta anotação validada. Caso a anotação seja validada, inicia-se um novo processo, a "consolidação da anotação", caso contrário, as seqüências devem passar novamente pelo processo de "análise da seqüência".

3.2.4.5 CONSOLIDAÇÃO DA ANOTAÇÃO

Com a anotação semi-automatizada baseada em ontologia, verificada e validada, parte-se para o processo de "consolidação da anotação", mostrado na FIG. 3.9. O objetivo deste processo é produzir um artigo a ser submetido para revistas especializadas da área, assim como submeter a seqüência pesquisada e anotada para bancos de dados públicos primários. A entrada deste processo é a seqüência anotada e a saída, a publicação da seqüência. As atividades são apresentadas a seguir:

Preparar artigo: uma das atividades iniciais é "preparar artigo", cujo objetivo é escrever um artigo no qual conste todas as descobertas sobre as seqüências do organismo pesquisado. Este artigo é usualmente escrito pelo grupo de anotadores e curadores envolvidos. A entrada desta atividade são as seqüências anotadas e a saída, o artigo escrito.

Formatar a anotação: a atividade "formatar a anotação", ocorre em paralelo com a atividade da preparação do artigo, pois o objetivo desta é formatar a anotação das seqüências de acordo com modelos pré-estabelecidos, como *Feature Table* (FT) ou *General Feature Format* (GFF), descritos no Capítulo 2 e que são aceitos pelos bancos de dados públicos. A entrada desta atividade também são as seqüências anotadas e a saída, a anotação formatada.

Submeter seqüências e anotações: uma vez que a anotação está formatada, a atividade seguinte é "submeter seqüências e anotações". Isto ocorre porque a submissão/publicação do artigo depende da submissão das seqüências com suas respectivas anotações a algum banco de dados público primário, pois, feito isso, o grupo de pesquisa receberá números identificadores das seqüências no banco em questão, os

quais devem ser ditos no artigo. Contudo, a visualização destas seqüências nos bancos de dados públicos é condicionada à efetivação da publicação, ou a um tempo determinado, que costuma ser de no máximo um ano. A atividade possui como entrada a seqüência anotada nos formatos pré-definidos e como saída os números identificadores.

Submeter artigo: o artigo escrito e os números identificadores do banco de dados passam a ser entradas da próxima atividade, “submeter artigo”, na qual os anotadores e curadores irão submeter o artigo com as descobertas das seqüências para eventos e revistas especializadas da área. A saída da atividade é o artigo submetido.

Disponibilizar seqüências e anotações em bancos de dados públicos: se o artigo for aceito, então a atividade seguinte é “disponibilizar seqüências e anotações em bancos de dados públicos”, pois a disponibilização que estava condicionada à publicação do artigo, pode neste momento ocorrer. A entrada desta atividade é a informação do artigo submetido e a saída são as seqüências e anotações disponibilizadas nos bancos de dados públicos.

Rever artigo: caso o artigo não seja aceito, a atividade a ser realizada é “rever artigo”. Nela, o anotador e o curador irão arrumar o artigo nas questões em que foi criticado para posterior submissão. Esta atividade acontece até que o artigo seja aceito e possui como entrada o artigo produzido inicialmente, e como saída o artigo revisto e re-arrumado.

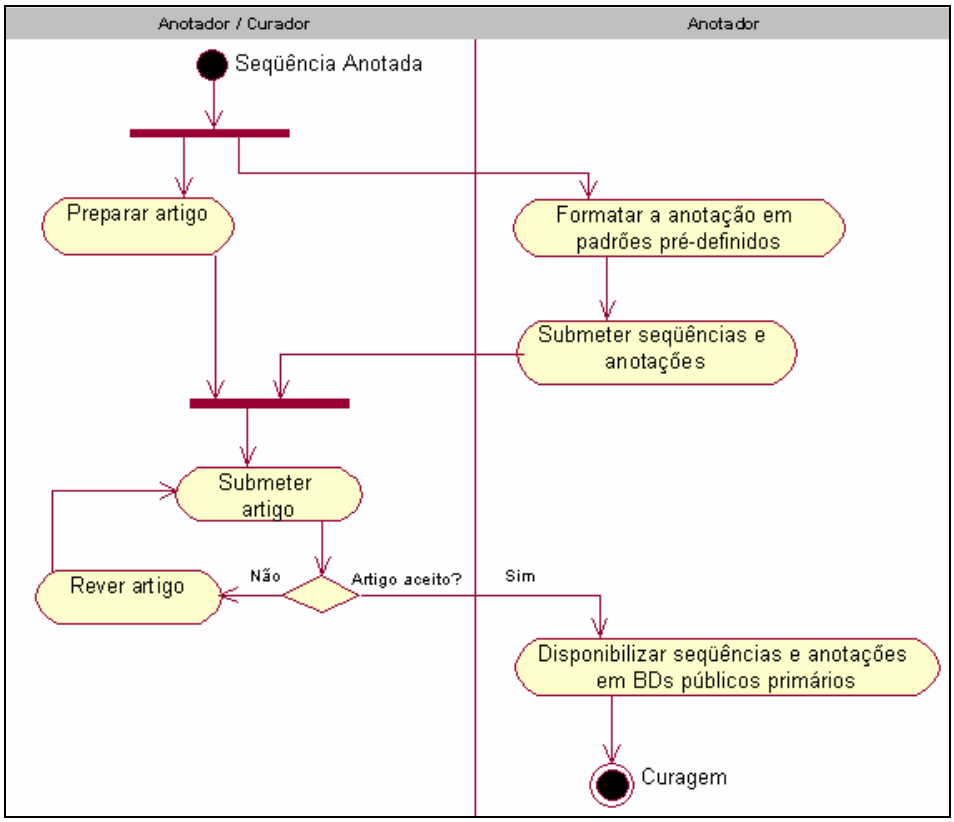


FIG. 3.9 Atividades da consolidação da anotação

Uma vez que as seqüências já estejam disponibilizadas nos bancos de dados primários, elas podem vir a ser disponibilizadas em bancos de dados secundários, passando para o processo que denominamos de curagem. Porém, para que isto seja possível, é necessário que haja o interesse dos administradores destes bancos no estudo realizado com as seqüências, e caso isto aconteça, as anotações passam inicialmente pela validação do curador do banco de dados. É um processo demorado, devido a uma análise maior sobre a anotação e ao pouco número de curadores que fazem essa análise em relação à quantidade de seqüências submetidas. Este processo não é aqui detalhado, devido a este ser de responsabilidade de centros de pesquisa mantenedores dos bancos de dados, e não mais do grupo de pesquisa.

3.2.5 A ONTOLOGIA COMO RECURSO NO PROCESSO DE ANOTAÇÃO GENÔMICA

As atividades descritas anteriormente fazem parte dos experimentos *in-silico* sobre as seqüências. Logo, eles fazem uso de recursos computacionais para apoiar o seu desenvolvimento. O principal recurso a ser descrito são as ontologias, contudo

exemplos de *softwares* que são utilizados para cada atividade do processo são descritos no Anexo 1.

O processo de anotação pode fazer uso de recursos computacionais como as ontologias. As ontologias são úteis no sentido que criam um vocabulário comum do que se deseja interpretar, no caso as características das regiões codificantes das seqüências.

A ontologia mais disseminada e utilizada na Bioinformática é a *Gene Ontology* (GO) cujos termos servem para realizar anotação genômica. Uma das formas de se encontrar o termo da GO é fazer uma busca de similaridade contra bancos de dados curados e que possuem seqüências anotadas com os termos da GO e contra o banco de dados que o Consórcio da GO mantém sobre organismos modelos anotados desta forma.

Todavia, além da busca por similaridade, diversas outras abordagens já foram iniciadas e constam na literatura sobre a escolha dos termos, para que estes sejam os melhores, os mais representativos. Essas abordagens fazem uso de grafos orientados, matriz de distância entre termos, e outros. Algumas delas são encontradas em (JOSLYN et al, 2004) (JONES et al, 2005) (VERSPOOR et al, 2006). O objetivo é fazer a busca da forma mais refinada possível. Contudo, apesar do refinamento através destas abordagens, ainda observações e intervenções manuais devem ser feitas para verificar se existe algum termo que não condiz com a seqüência e assim ser retirado, ou escolher o melhor se mais de um termo foi encontrado. O ideal é que somente um termo represente a região codificante da seqüência.

A próxima seção descreve, com base num sistema real, o GARSA, os resultados gerados a partir da execução de algumas das ferramentas citadas no Anexo 1, incluindo o uso de ontologia.

3.3 ANOTAÇÃO NO GARSA

A modelagem de processos, assim como a descrição das estruturas de apoio foram melhor identificadas a partir de observações e reuniões realizadas com o grupo de pesquisa do IOC/FIOCRUZ. Para entender melhor como cada processo acontece e quais são os resultados, esta exemplificação baseada no uso do sistema GARSA, desenvolvido pelo grupo em questão, foi realizada.

Como comentado anteriormente, o GARSA é um sistema desenvolvido para facilitar o processo de análise e anotação de seqüências genômicas. Assim, descrevemos com exemplos estes processos com base neste sistema, buscando através da captura de telas elucidar o processo explicado ou mostrar os resultados. Para isto foi utilizado exemplos

dos processos de análise e anotação do estudo já concluído e armazenado no GARSA sobre o organismo *Trypanosoma rangeli*. (WAGNER, 2006).

3.3.1 RECEBIMENTO DAS SEQÜÊNCIAS E PRÉ-ANÁLISE

Inicialmente, o administrador do projeto cria as bibliotecas do sistema, informando qual o vetor de clonagem utilizado, seu código, nome e descrição. Assim o usuário pode carregar os seus cromatogramas (*reads*), que recebem uma nomenclatura específica para cada um. Essa nomenclatura segue uma máscara do tipo AABBXXXYYYZZZ.P, cujo significado é:

AA: código do projeto

BB: código do laboratório

XXX: identificação da biblioteca

YYY: identificação do número da placa utilizada no seqüenciamento

ZZZ: posição da placa de seqüenciamento

P: identificação da fita seqüenciada (g-*reverse* e b-*forward*)

É importante conhecer essa nomenclatura, pois as análises e anotações no GARSA são realizadas para cada cromatograma, portanto, serão referenciados desta forma.

As seqüências no processo de pré-análise passam pelos programas Phred e Crossmatch para avaliação da qualidade e remoção de seqüências dos vetores. Logo após é feito o agrupamento de seqüências similares pelo programa CAP3, que formam um conjunto de seqüências não redundantes.

Os programas seguintes, para análise e anotação recebem como entrada esse conjunto de seqüências não redundantes (*clusters*) e não as seqüências isoladamente.

3.3.2 ANÁLISE DAS SEQÜÊNCIAS

Diversos programas são executados para a análise das seqüências a fim de identificar características nas seqüências. Eles possuem finalidades específicas, como visto no Anexo 1.

A análise inicia com a execução dos programas *geecee* (Rice et al, 2000) do pacote EMBOSS (*European Molecular Biology Open Software Suíte*) que faz uma estimativa do conteúdo G+C das seqüências não redundantes, e tRNA-Scan (LOWE et al, 1997) que busca seqüências de tRNA (*transfer RiboNucleic Acid*).

Em seguida, para o estudo do *Trypanosoma rangeli*, somente o programa Glimmer3 foi utilizado para a predição de genes (encontrar genes codificantes de proteínas). O GARSA ainda possui no seu *workflow* o programa YACOP para a mesma finalidade. A FIG. 3.10 mostra uma tabela do sistema GARSA que traz a predição de genes de acordo com os clusters. Os clusters são mostrados através de seu nome (*Cluster Name*), os quais seguem a nomenclatura descrita anteriormente. Para cada cluster a tabela informa qual o programa que foi utilizado para encontrar a predição de genes (*Program*), no caso somente o Glimmer3, um identificador para o Gene (*Gene id*), o valor (Value), a porcentagem do conteúdo G+C encontrado (*GC Content*), calculado anteriormente pelo programa *geecee*, o gene inicial (*Gene Start*), o gene final (*Gene End*), qual o *frame* de leitura (*Cluster Frame*), o tamanho do gene (*Length*) e ainda uma possibilidade de visualizar graficamente todos esses resultados para cada cluster (*View* – mostrado na FIG. 3.11).

Gene Prediction Table

Cluster Name	Program	Gene Id	Value	GC Content	Gene Start	Gene End	Cluster Frame	Length	View
TGEG101003F12.g	Glimmer3	1	6.24	54%	281	48	-3	233	
TGEG101004B09.g	Glimmer3	1	4.29	55%	413	6	-3	407	
TGEG101004E10.g	Glimmer3	1	5.55	47%	418	161	-2	257	
TGEG101007C12.g	Glimmer3	1	2.71	54%	72	374	+3	302	
TGEG101008B05.g	Glimmer3	1	2.08	56%	371	111	-3	260	
TGEG101008G07.g	Glimmer3	1	7.15	31%	240	82	-1	158	
TGEG101011B04.g	Glimmer3	1	3.51	48%	408	226	-1	182	
TGEG101012C12.g	Glimmer3	1	11.56	36%	204	22	-1	182	
TGEG101013E12.g	Glimmer3	1	4.44	48%	410	228	-3	182	
TGEG101014H06.g	Glimmer3	1	4.86	46%	449	189	-3	260	

FIG. 3.10 Predição de genes

Os resultados mostrados graficamente na FIG. 3.11 correspondem ao primeiro cluster da FIG. 3.10, TCEG101003F12.g. A linha traçada no FRAME -3 que mostra as posições iniciais e finais e conteúdo GC equivale ao gene predito encontrado. Possui também a visualização deste gene através de seqüências de nucleotídeos e aminoácidos.

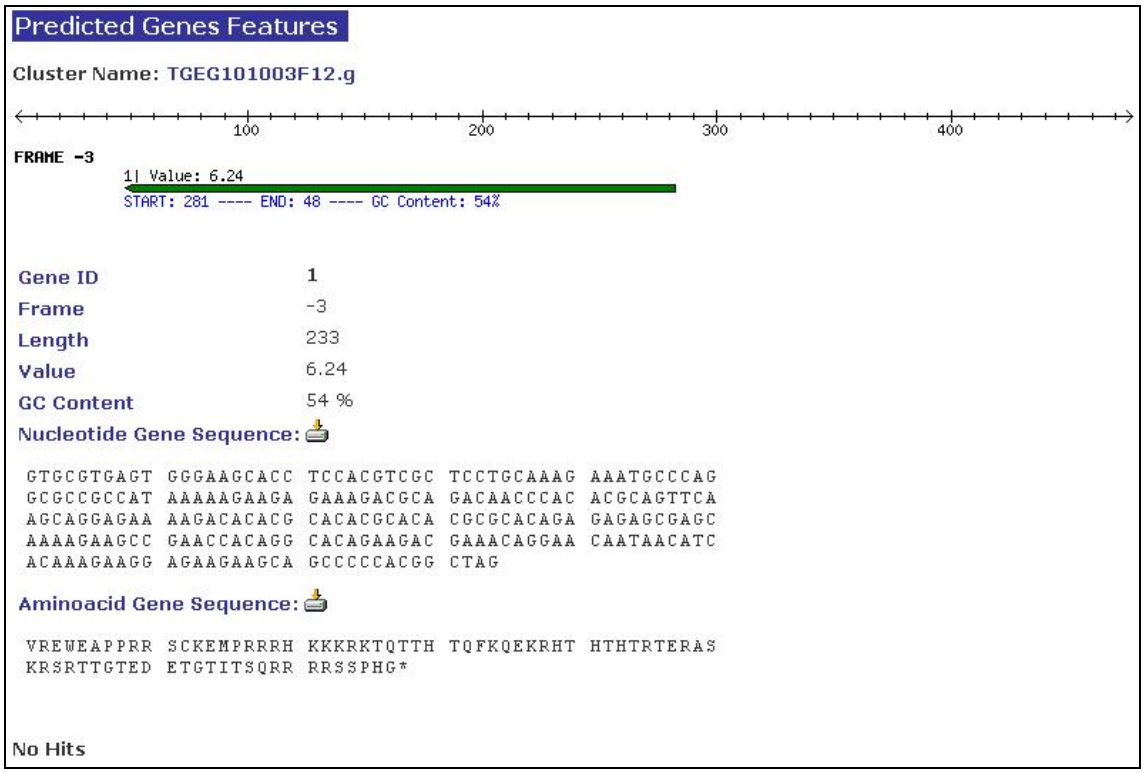


FIG. 3.11 Característica do gene predito

Para busca de similaridade, foram utilizados os programas do pacote BLAST: *blastn*, *blastx* e *tblastx* fazendo comparação dos clusters contra diversos bancos de dados de nucleotídeos e de proteínas.

A FIG. 3.12 apresenta uma síntese de alguns resultados da análise de similaridade gerados após a execução do BLAST. É possível escolher opções de configuração para mostrar somente o dado desejado (*Show*). Os resultados constam de um número identificador (*No.*), qual o programa do pacote BLAST foi utilizado (*Program*), qual o banco de dados ou arquivos de seqüências que foram utilizados (*Databank*), o número de entradas (*No. Entries*), a quantidade de *hits* (boa similaridade) e a porcentagem sobre esse *hits* (*Hits - %*), a quantidade que não obteve nenhum *hit* e sua porcentagem (*No Hits - %*) e o total (*Total*) das entradas que geraram resultados.

Para cada resultado, tem-se como visualizar os clusters que tiveram *hits* e quais não tiveram, e para cada cluster pode-se observar quais foram os melhores resultados de similaridade e assim fazer uma anotação.

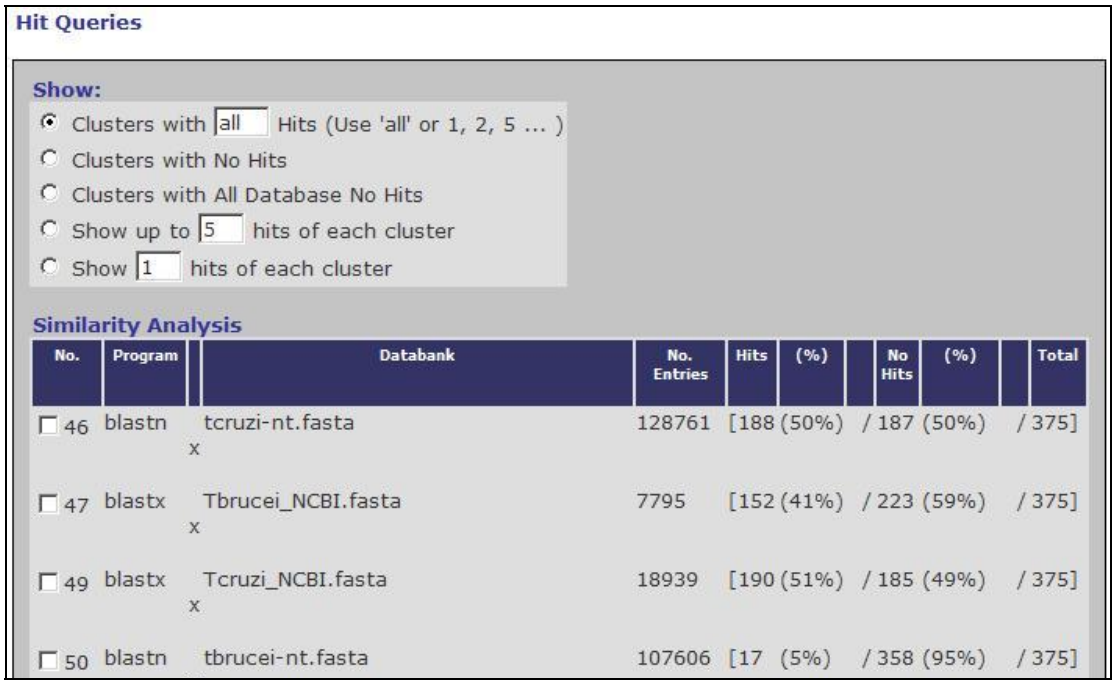
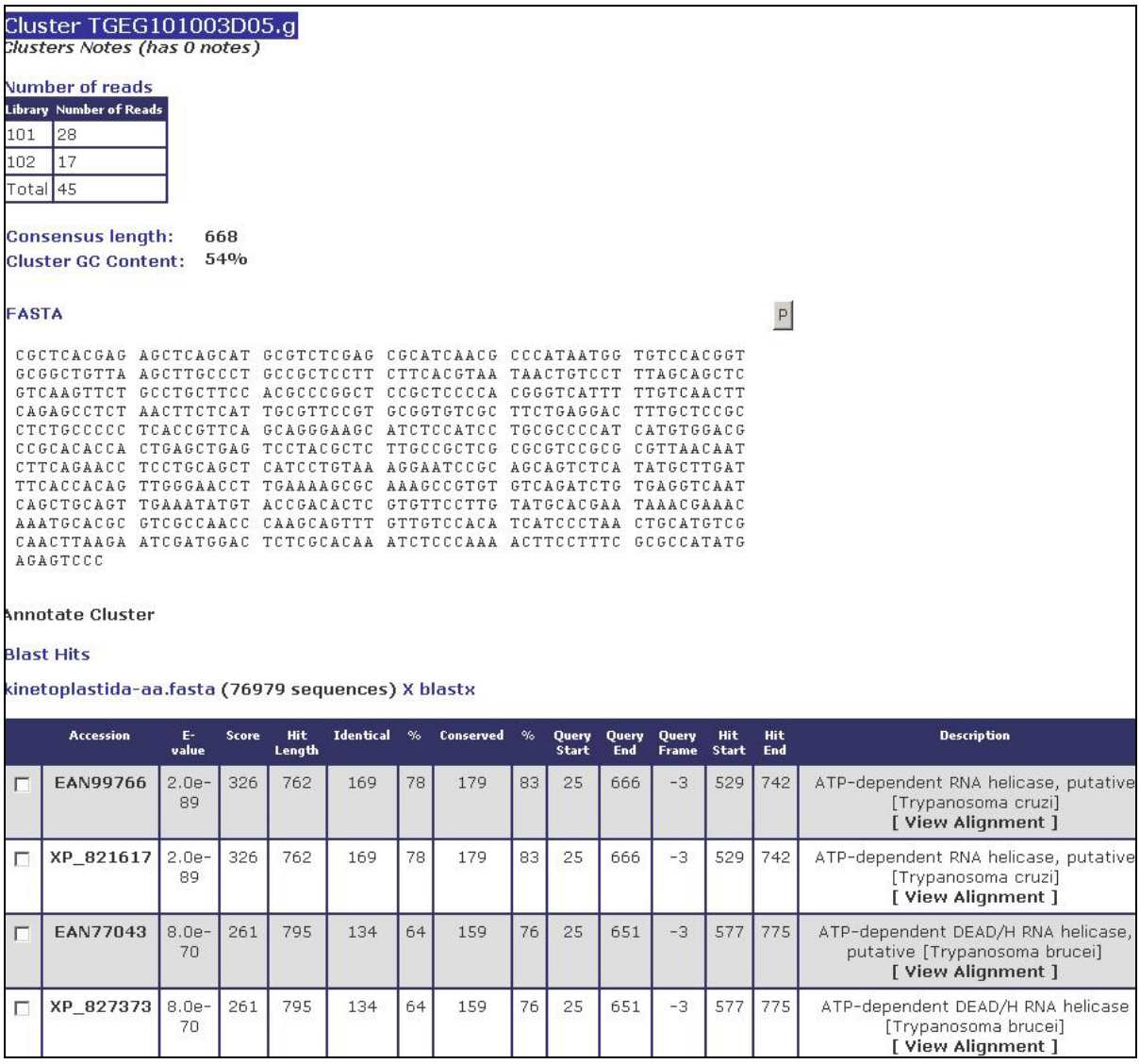


FIG. 3.12 Resultados do programa BLAST

Dando continuidade ao processo, para a busca de domínios e famílias de proteínas utilizou-se os programas *InterProScan* e o rpsBLAST para a busca nos bancos de dados CDD, Pfam, Smart, COG e KOG.

Para ilustrar, a FIG. 3.13 mostra os *hits* obtidos com o BLAST para o cluster TGEG101003D05.g. Este apresentou como melhores resultados de similaridade a execução do programa *blastx* contra o conjunto de seqüências do *kinetoplastida-aa.fasta*. Os outros resultados não são mostrados aqui.

Analisando os resultados, foi inferido que este cluster é identificado como um possível *ATP-dependent RNA helicase*, e desta forma foi feita uma anotação.



Cluster TGEG101003D05.g

Clusters Notes (has 0 notes)

Number of reads

Library	Number of Reads
101	28
102	17
Total	45

Consensus length: 668

Cluster GC Content: 54%

FASTA

```
CGCTCACGAG AGCTCAGCAT GCGTCTCGAG CGCATCAACG CCCATAATGG TGTCCACGGT
GCGGCTGTTA AGCTTGCCCT GCCGCTCCTT CTTCACGTAA TAACTGTCCT TTAGCAGCTC
GTCAAGTTCT GCCTGCTTCC ACGCCCGGCT CCGCTCCCCA CGGGTCATTT TTGTCAACTT
CAGAGCCTCT AACTTCTCAT TGCGTTCCGT GCGGTGTCGC TTCTGAGGAC TTTGCTCCGC
CTCTGCCCCC TCACCGTTCA GCAGGGAAGC ATCTCCATCC TGCGCCCCAT CATGTGGACG
CCGCACACCA CTGAGCTGAG TCCTACGCTC TTGCCGCTCG CGCGTCCGCG CGTTAACAAT
CTTCAGAACC TCCTGCAGCT CATCCTGTAA AGGAATCCGC AGCAGTCTCA TATGCTTGAT
TTCACCACAG TTGGGAACCT TGAAAAGCGC AAAGCCGTGT GTCAGATCTG TGAGGTCAAT
CAGCTGCAGT TGAAATATGT ACCGACACTC GTGTTCCTTG TATGCACGAA TAAACGAAAC
AAATGCACGC GTCGCCAACC CAAGCAGTTT GTTGTCCACA TCATCCCTAA CTGCATGTCG
CAACTTAAGA ATCGATGGAC TCTCGCACAA ATCTCCCAAA ACTTCCTTTC GCGCCATATG
AGAGTCCC
```

Annotate Cluster

Blast Hits

kinetoplastida-aa.fasta (76979 sequences) X blastx

	Accession	E-value	Score	Hit Length	Identical	%	Conserved	%	Query Start	Query End	Query Frame	Hit Start	Hit End	Description
☐	EAN99766	2.0e-89	326	762	169	78	179	83	25	666	-3	529	742	ATP-dependent RNA helicase, putative [Trypanosoma cruzi] [View Alignment]
☐	XP_821617	2.0e-89	326	762	169	78	179	83	25	666	-3	529	742	ATP-dependent RNA helicase, putative [Trypanosoma cruzi] [View Alignment]
☐	EAN77043	8.0e-70	261	795	134	64	159	76	25	651	-3	577	775	ATP-dependent DEAD/H RNA helicase, putative [Trypanosoma brucei] [View Alignment]
☐	XP_827373	8.0e-70	261	795	134	64	159	76	25	651	-3	577	775	ATP-dependent DEAD/H RNA helicase [Trypanosoma brucei] [View Alignment]

FIG. 3.13 Resultado da análise de similaridade cluster TGEG101003D05.g

Foi também executado o programa psiBLAST para as seqüências não redundantes que não apresentaram similaridade com nenhum dos bancos de dados na execução dos programas do pacote BLAST. Este programa promove a busca de similaridade entre seqüências protéicas a partir de diversas buscas ou iterações, utilizando o resultado de cada iteração para montar uma matriz de substituição que é usada na iteração posterior (WAGNER, 2006).

Após as análises de similaridades terem sido feitas, pode-se escolher fazer a análise filogenética de cada cluster. Contudo, alinhamentos múltiplos das seqüências necessitam ser feitos anteriormente. Na execução do programa psiBLAST, quando encontradas mais de cinco seqüências similares ao final de cada análise, estas foram utilizadas para a construção de alinhamentos múltiplos, através do programa *ProbCons*. O GARSA ainda disponibiliza a utilização dos programas *Clustalw* e *Muscle* para esta

tarefa. Os alinhamentos foram utilizados para a busca de homologia em outros clusters através do programa HMMER.

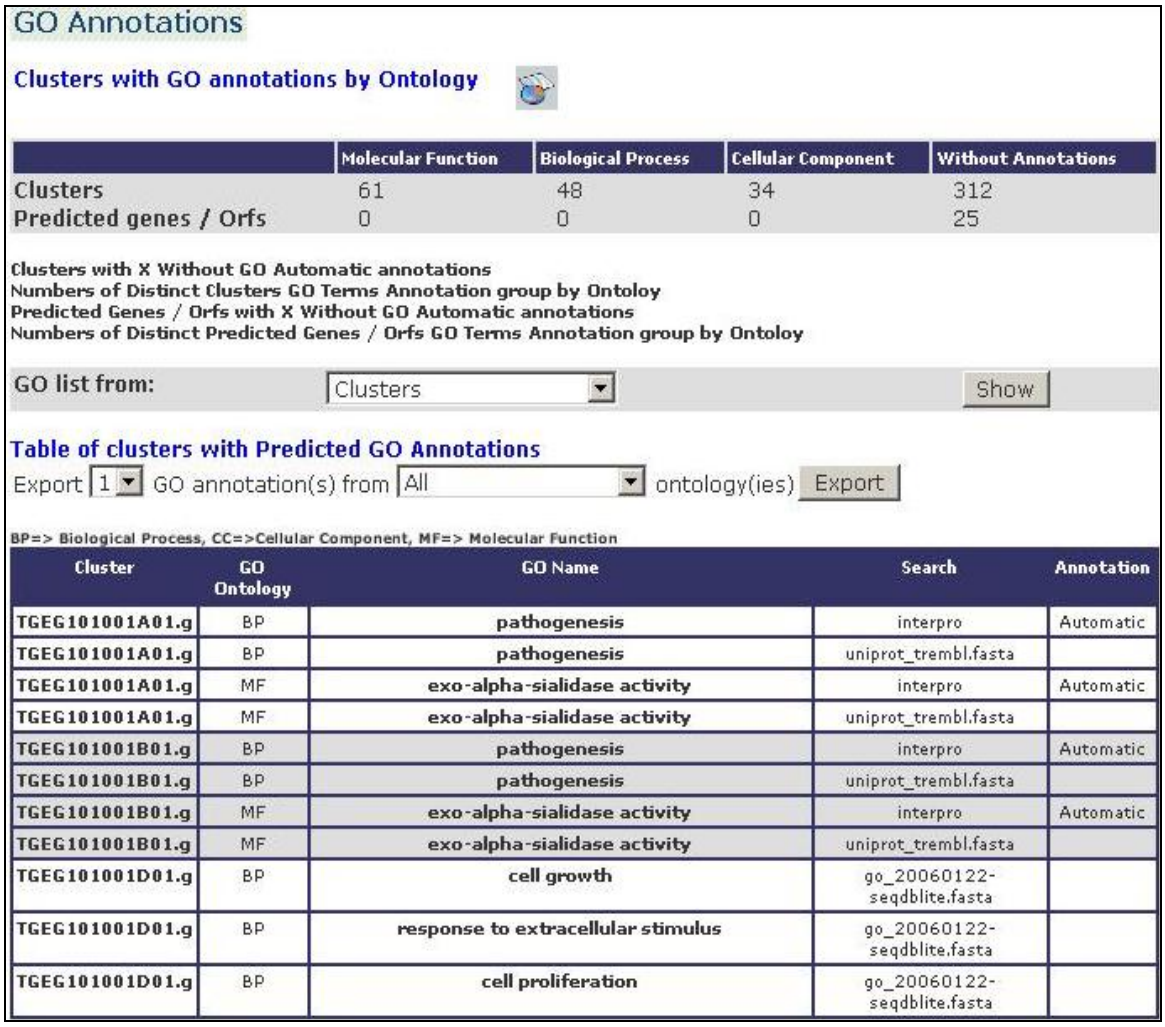
A análise filogenética foi realizada utilizando os programas *Phylip e MEGA.* Para cada cluster o GARSA oferece a opção de executar os programas com estas finalidades tanto para seqüências de aminoácidos quanto para seqüências de nucleotídeos.

De acordo com WAGNER (2006), as seqüências não redundantes, após a identificação das regiões codificantes, foram confrontadas todas contra todas utilizando o programa *blastp* para identificar seqüências repetitivas ou possíveis genes parálogos. Foi também realizada a identificação de genes ortólogos compartilhado entre os organismos *T. rangeli*, *T. cruzi*, *T. brucei* e *L. major* através do programa *OrthoMCL*, o qual não pertence ao *workflow* do GARSA.

3.3.3 ANOTAÇÃO

Enquanto os dados gerados através das análises de similaridade são observados para interpretá-los, outra atividade pode ser feita, que é a busca de termos da *Gene Ontology* (GO) para a escolha dos melhores termos que representam as seqüências estudadas.

São escolhidos bancos de dados anotados com a GO para confrontar com as seqüências de estudo. Para o organismo *T. rangeli* os bancos confrontados foram: banco de termos da GO, InterPro, UniProt/TrEMBL e UniProt/Swiss-Prot.



GO Annotations

Clusters with GO annotations by Ontology

	Molecular Function	Biological Process	Cellular Component	Without Annotations
Clusters	61	48	34	312
Predicted genes / Orfs	0	0	0	25

Clusters with X Without GO Automatic annotations
Numbers of Distinct Clusters GO Terms Annotation group by Ontoloy
Predicted Genes / Orfs with X Without GO Automatic annotations
Numbers of Distinct Predicted Genes / Orfs GO Terms Annotation group by Ontoloy

GO list from: Clusters Show

Table of clusters with Predicted GO Annotations

Export 1 GO annotation(s) from All ontology(ies) Export

BP=> Biological Process, CC=>Cellular Component, MF=> Molecular Function

Cluster	GO Ontology	GO Name	Search	Annotation
TGEG101001A01.g	BP	pathogenesis	interpro	Automatic
TGEG101001A01.g	BP	pathogenesis	uniprot_trembl.fasta	
TGEG101001A01.g	MF	exo-alpha-sialidase activity	interpro	Automatic
TGEG101001A01.g	MF	exo-alpha-sialidase activity	uniprot_trembl.fasta	
TGEG101001B01.g	BP	pathogenesis	interpro	Automatic
TGEG101001B01.g	BP	pathogenesis	uniprot_trembl.fasta	
TGEG101001B01.g	MF	exo-alpha-sialidase activity	interpro	Automatic
TGEG101001B01.g	MF	exo-alpha-sialidase activity	uniprot_trembl.fasta	
TGEG101001D01.g	BP	cell growth	go_20060122-seqdblite.fasta	
TGEG101001D01.g	BP	response to extracellular stimulus	go_20060122-seqdblite.fasta	
TGEG101001D01.g	BP	cell proliferation	go_20060122-seqdblite.fasta	

FIG. 3.14 Resultado da anotação pelos termos da GO

A FIG. 3.14, extraída parcialmente, apresenta uma tabela com a escolha dos termos da GO. Para cada cluster (*Cluster*), esta tabela indica qual a ontologia a que o termo pertence (*GO Ontology*), o nome do melhor termo encontrado para tal ontologia (*GO Name*), o banco sobre o qual foi realizada a busca (*Search*) entre um dos quatro citados anteriormente, e se o melhor termo foi encontrado automaticamente ou precisou de intervenção manual (*Annotation*).

Os resultados são observados individualmente e intervenções manuais podem ser feitas caso algum termo não represente corretamente a anotação sobre o organismo estudado ou nos casos em que mais de um termo é encontrado. Casos em que o termo encontrado não represente adequadamente a anotação sobre a seqüência em questão ou seqüências para as quais não foram encontrados nenhum termo, não são tratados.

A FIG. 3.14 também mostra um resumo da quantidade de *clusters*, a quantidade de termos encontrados para cada ontologia (*Molecular Function*, *Biological Process* e *Cellular Component*) e a quantidade de clusters sem anotação.

As anotações obtidas através dos resultados dos programas de análise das seqüências são unidas às anotações feitas com os termos da *Gene Ontology*, caracterizando a anotação semi-automatizada e baseada em ontologia. As anotações de todos os clusters são sintetizadas numa tabela como mostra a FIG. 3.15, apresentada parcialmente.

Cluster	Description	GC Content	CDS Length	Aa. Length	Molecular Functional	Biological Process	Cellular Component
TGEG101003D05.g.1	ATP-dependent RNA helicase	53%	642	214	ATP-dependent helicase activity	protein biosynthesis	nucleus
TGEG101004C10.g.1	kinetoplasto DNA (minicircle replication)	31%	1323	0			
TGEG101003D07.g.1	Hypothetical Protein	54%	651	217			
TGEG101004E05.g.1	kinetoplasto minicircle DNA	30%	633	0			
TGEG101004H07.g.1	kinetoplasto minicircle DNA	30%	804	0			
TGEG101003H12.g.1	Hypothetical Protein	57%	612	204	binding		
TGEG101038G02.b.1	Hypothetical Protein	51%	576	192			
TGEG101047C02.b.1	kinetoplasto minicircle DNA	32%	585	0			
TGEG101038A09.b.1	ATP-dependent RNA helicase	58%	234	78	ATP-dependent RNA helicase activity	ribosome biogenesis and assembly	nucleus
TGEG101003C04.g.1	Hypothetical Protein, Conserved	55%	414	138			cytoskeleton
TGEG101004B09.g.1	Hypothetical Protein	56%	390	130			
TGEG101003B10.g.1	Hypothetical Protein	49%	585	195			

FIG. 3.15 Interface de anotação das seqüências

A anotação sobre os *clusters* mostrada na FIG. 3.15 consiste dos seguintes itens: nome do cluster (*Cluster*), a anotação sobre o que representam os genes contidos no cluster (*Description*), seu identificador (Id), o conteúdo de GC's (*G+C Content*), o tamanho da região codificiante (*CDS Length*), o tamanho da cadeia de aminácidos (*Aminoacid Length*), e os termos encontrados para as ontologias da GO (*Molecular Functional, Biological Process* e *Cellular Component*).

Para cada cluster pode-se conferir a exibição com detalhes dos resultados dos programas de onde foram inferidas todas as informações anotadas, como mostrado anteriormente na FIG. 3.13.

Com a anotação realizada, dados estatísticos são gerados com as informações utilizadas durante todo o processo de pesquisa das seqüências. A FIG. 3.16 apresenta a estatística para o estudo do organismo *T. rangeli*. Esta é subdividida em três partes: informações sobre as bibliotecas utilizadas (*All Library*), similaridade (*Similarity*) *e* anotação (*Annotation*). Estes dados podem ainda ser visualizados de forma gráfica.

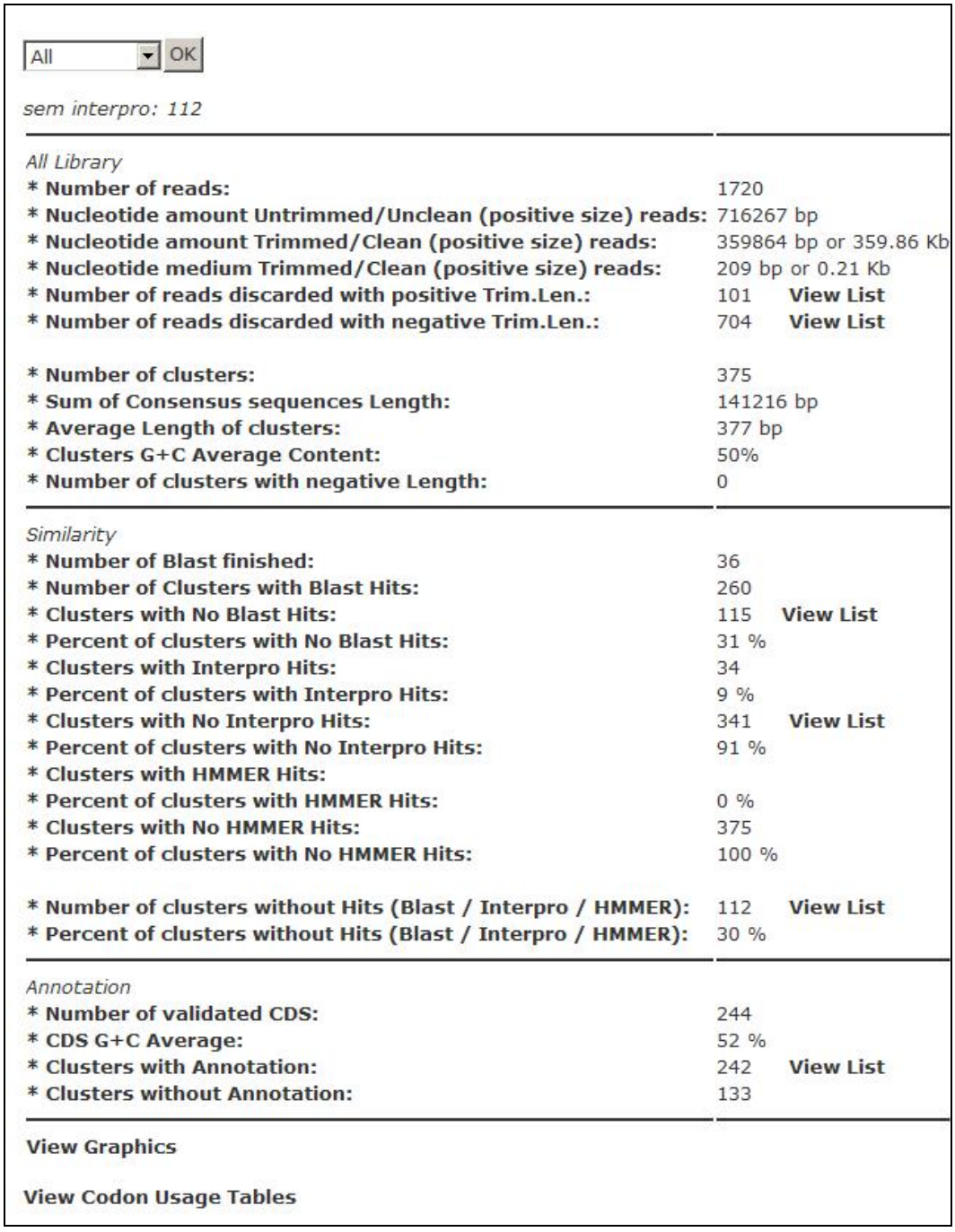


FIG. 3.16 Dados estatísticos

3.4 CONSIDERAÇÕES

Embora o uso de termos de ontologias no processo de anotação genômica seja importante para a geração de anotações de qualidade e para facilitação da busca e recuperação por similaridade, o uso destas ontologias ainda apresenta problemas. Foi possível observar no GARSA que mesmo com as formas refinadas e automáticas de busca de termos, os que são encontrados e usados corretamente ainda representam um percentual muito baixo.

Por exemplo, na pesquisa genômica realizada com o *Trypanosoma rangeli*, das 375 seqüências não redundantes estudadas, somente 63 encontraram termos da GO em alguma das três ontologias, ou seja, apenas 17% do total das seqüências tiveram anotação através de um vocabulário controlado. Isso acontece devido a três fatores:

Desatualização da Ontologia: este fator leva a não anotação baseada em ontologia, pois apesar de ser um gene ou produto de gene já descoberto por outros, a Ontologia ainda não possui tal referência, devido ao fato da descoberta do novo gene não lhe ter sido reportada.

Inexistência de termo: a seqüência a qual deseja encontrar um termo, possuir um gene inédito. Sendo assim, este ainda deve ser reportado aos curadores da Ontologia para serem inseridos os termos representantes do novo gene.

Falhas nos algoritmos de busca e refinamento de termo: diversas abordagens para a busca de termos existem (citadas na secção 3.2.5). Contudo, na prática são abordagens que não se mostram completamente confiáveis, pois trazem muitos termos considerados como falso-positivos, o que requer uma intervenção manual por parte do pesquisador; ou não encontram o termo, apesar dele existir na Ontologia;

Cabe ressaltar que estes fatores não estão relacionados somente com a GO. Qualquer que seja a ontologia utilizada pelos sistemas de anotação, estes três problemas estarão passíveis de acontecer. Entretanto, estes são somente alguns dos problemas que ocorrem no processo de anotação genômica e inclusive no uso das ontologias. Estes problemas serão caracterizados no capítulo seguinte.

O processo de anotação genômica ajudou a identificar o momento em que é realizada a anotação baseada em ontologia e como ela é obtida. Este fato é importante, pois através desta identificação foi possível sugerir uma solução de forma a expandir o processo para o estudo da anotação baseada em ontologia. Isto devido a acreditarmos que esta anotação é de bastante relevância para a ampliação da colaboração existente e

consequentemente do compartilhamento de conhecimento, bem como para a evolução das ontologias e assim a diminuição dos problemas citados anteriormente.

4 EXPANDINDO O USO DA ONTOLOGIA NO PROCESSO DE ANOTAÇÃO GENÔMICA

De acordo com a hipótese deste trabalho, se ficar bem caracterizado o processo de anotação genômica e a colaboração inerente a ele, é possível sugerir soluções de apoio à colaboração para ampliar o uso e a evolução de ontologias deste domínio, compartilhando conhecimento, estimulando o processo de anotação baseado em ontologia e melhorando a qualidade das anotações.

Para caracterizar o processo de anotação genômica, houve a necessidade de entender como o mesmo é desempenhado pelos centros de pesquisa, e a partir de observações e entrevistas foi possível modelar o processo de forma generalizada como mostrado no capítulo anterior. Diante das descrições das atividades que fazem parte dos processos, identifica-se que ocorre colaboração ao longo dos mesmos.

Verifica-se também que esta colaboração pode ser ampliada através do uso das ontologias neste processo, devido ao fato deste tipo de anotação possuir um vocabulário comum, utilizado por diversos grupes de pesquisa da comunidade de Bioinformática, o que viabiliza um entendimento maior entre os pesquisadores e assim aumentam as chances de colaboração e parcerias entre centros de pesquisa genômica de todo o mundo. A ontologia torna-se, portanto, um importante veículo de comunicação entre os diversos papéis envolvidos no processo.

Contudo, a anotação baseada em ontologia tem sido subutilizada, pois ela ainda acontece como uma segunda forma de anotar, uma vez que a anotação semi-automatizada sempre existe. Há também os problemas encontrados sobre as ontologias, como citados no final do capítulo anterior: desatualização das ontologias, inexistência de termo, falha nos algoritmos de busca e refinamento de termo. Outra questão se refere a quando é encontrado algum tipo de problema com o termo, ou casos de inexistência de termos, pois fica a critério do anotador reportar tais questões ou não para os curadores da ontologia, constituindo assim uma tarefa fora do processo de anotação.

Um problema também encontrado com a caracterização do ambiente se refere aos cenários e papéis envolvidos no processo, os quais não são bem reconhecidos e desta maneira a colaboração torna-se pouco explorada.

Diante da falta de compartilhamento do conhecimento pertinente a cada membro do grupo de pesquisa, a anotação baseada em ontologia acaba sendo somente mais uma

forma de anotar e não a principal, impactando a troca de informações entre os participantes do grupo, assim como pesquisadores do mundo inteiro, no sentido em que suas anotações próprias (através de um vocabulário usual do grupo) podem não ser corretamente entendidas. Impacta também no conteúdo armazenado em bancos de dados públicos, principalmente os primários, no sentido em que as anotações armazenadas perdem em termos de qualidade.

Logo, pode-se concluir que ao se criar meios de apoio e incentivo à anotação baseada em ontologia, não só a colaboração pode ser ampliada, como também pode-se contribuir para melhoria da própria ontologia e das anotações futuras.

Este capítulo apresenta inicialmente, na seção 4.1, uma descrição de alguns exemplos onde ocorre colaboração entre grupos e pesquisadores da área. Esta colaboração se manifesta nas interações entre anotadores e curadores durante o processo de estudo de seqüências genômicas, que muitas vezes se dá no contexto de consórcios formados na área de Bioinformática. A partir destes exemplos, foi possível identificar cenários de colaboração e os papéis envolvidos (seção 4.2).

A partir da modelagem de processo apresentada no capítulo anterior, este trabalho procurou identificar pontos onde a colaboração pudesse ser incentivada e apoiada de modo a enriquecer as anotações realizadas e ampliar o compartilhamento de experiências entre os pesquisadores. Nesse sentido, considerando-se o uso da ontologia como parte essencial para o crescimento da colaboração, revisitou-se o processo de anotação genômica e este foi estendido com vistas a compreender todo um estudo sobre a anotação baseada em ontologia, o qual tem por objetivo a captura, a estruturação e a recuperação das ações realizadas por todos que anotam. Esta extensão é apresentada na seção 4.3.

4.1 COLABORAÇÃO NA BIOINFORMÁTICA

A colaboração na área de Bioinformática acontece não somente entre membros de um grupo de pesquisa envolvidos no estudo de seqüências genômicas, mas também entre grupos, no contexto de consórcios formados na área.

A formação de consórcios caracteriza-se pela junção de universidades e centros de pesquisa de biologia e informática com o objetivo de realizarem, além de estudos genômicos, a criação de bancos de dados para a área e o desenvolvimento de ferramentas e recursos para apoiar os processos da área. Estes consórcios enfatizam a

tendência da colaboração entre diversos grupos e que por ventura podem então serem beneficiados com propostas como a apresentada neste trabalho. Então, são exemplos de alguns destes consórcios: myGrid, TCruziDB, Consórcio Gene Ontology e Consórcio BioWebDB.

4.1.1 MYGRID

O projeto myGrid (STEVENS et.al., 2003) foi fundado pelo EPSRC (EPSRC, 2007) e envolve cinco universidades britânicas, o EBI (EBI, 2007) e muitos colaboradores da indústria. Permite explorar interesses na tecnologia de *Grid*, com ênfase em *Grid* de Informação e provê camadas de *middleware*[5] para apoiar experimentos *in-silico* personalizados de acordo com as necessidades imediatas da Bioinformática, oferecendo um alto nível de integração para dados e aplicações da biologia. Um *Grid* é útil para a área de Bioinformática devido ao oferecimento de maior poder de processamento para tarefas custosas dos processos da área.

O myGrid está ligado com uma série de outros projetos colaborativos que o provê com recursos ou que o integra em seus próprios projetos, a saber: Gowlab (SENGER, 2005), Seqhound (MICHALICKOVA et al., 2002), BioMoby (WILKINSON e LINKS, 2002), BioMart (DURINCK et al., 2005), Utopia (PETTIFER et al., 2007) entre outros, os quais habilitam integração de ferramentas de Bioinformática, disponibilizam recursos como *data warehouse*, soluções para recursos computacionais distribuídos e outras funcionalidades.

4.1.2 TCRUZIDB

TCruziDB (LUCHTAN, 2004) é um banco de dados integrado para o armazenamento do genoma do parasita *Trypanosoma cruzi*, o agente causador da doença de chagas.

A caracterização inicial do genoma do *T. cruzi* surgiu em 1997 depois dos estudos feitos com o clone CL Brener e a partir de dados EST seqüenciados. A informação relativa a estes estudos foi disponibilizada em um sistema baseado na *Web*, localizados na Fiocruz, Brasil. O sequenciamento do genoma inteiro do *T. cruzi* iniciou-se em 2000

[5] *Software* que faz a mediação entre outros *softwares*, ocultando do programador diferenças de protocolos de comunicação, plataformas e dependências do sistema operacional.

pelo consórcio internacional TSK-TSC (EL-SAYED et al., 2005) representando um esforço colaborativo entre pesquisadores do Brasil e outros países para trazer à comunidade de pesquisa detalhes sobre o estudo genômico do organismo.

O consórcio TSK-TSC compreende os seguintes institutos: The Institute for Genomic Research (TIGR, USA) (TIGR, 2007), Seattle Biomedical Research Institute (SBRI) (SBRI, 2007) e Karolinska Institute (KI) (KI, 2007), e possui como objetivo a geração das seqüências, montagem e anotação sobre o parasita, para a atualização do banco de dados.

4.1.3 CONSÓRCIO GENE ONTOLOGY

O Consórcio da GO é formado por um conjunto de bancos de dados de proteínas e organismos modelos e comunidades de pesquisa biológica todos envolvidos no desenvolvimento e aplicação da Gene Ontology. A colaboração com a GO acontece de duas formas: através de membros do consórcio e dos associados à GO.

Os membros do consórcio são oficialmente comprometidos com a evolução da Ontologia, participando de submissões regulares de anotações, contribuições para o conteúdo da GO, participações nas reuniões do Consórcio, assim como a promoção de novas reuniões. Alguns destes membros são: FlyBase (CROSBY, et al., 2007), GeneDB (HERTZ-FOWLER et al., 2004), Gene Ontology Annotation (GOA) (CAMON et al., 2004), The Institute for Genomic Research (TIGR).

Os associados à GO são grupos que fazem importantes contribuições para a GO, em pelo menos uma das seguintes formas: contribuição de anotações para o banco de dados da GO, contribuições com aplicações de código aberto para o uso da GO e colaboração no desenvolvimento do conteúdo da GO. Entre os associados estão AgBase (BRIDGES et al., 2005) e um outro consórcio, o Plant-Associated Microbe Gene Ontology (PAMGO) (PAMGO, 2007).

Os pesquisadores, que utilizam a GO em seus projetos, em geral também podem fazer contribuições como sugestões de novos termos ou atualizações, bem como críticas e sugestões. Estas contribuições acontecem principalmente através de listas de discussões ou do Curator Requests Tracker[6] disponibilizado pelo Consórcio.

[6] http://geneontology.sourceforge.net

4.1.4 CONSÓRCIO BIOWEBDB

O Consórcio BioWebDB, financiado pelo CNPq, inicialmente estabeleceu colaborações com o TcruziDB (BIOWEBDB, 2007). As pesquisas do grupo encontram-se concentradas em dois principais focos: no desenvolvimento de ferramentas para a Bioinformática para análises de genomas e transcriptomas, e em análises dos genomas de tripanosomatídeos. A iniciativa do Consórcio é construir plataformas flexíveis, integradas e amigáveis, capazes de serem compartilhadas com diferentes conjuntos de dados e projetos de genoma.

Os pesquisadores do Consórcio BioWebDB, envolvidos em estudos de genômica comparativa e bancos de dados genômicos, fazem parte das seguintes instituições: Fundação Oswaldo Cruz (FIOCRUZ), Universidade Federal de Santa Catarina (UFSC), Universidade Federal do Rio de Janeiro (UFRJ), Universidade Federal do Estado do Rio de Janeiro (UNIRIO) e Instituto Militar de Engenharia do Rio de Janeiro (IME/RJ).

4.2 CENÁRIOS DE COLABORAÇÃO E PAPEIS

A partir da análise da forma de trabalho de alguns projetos (seção 3.1 do capítulo 3) que participam de consórcios como os mencionados na seção anterior, foi possível identificar os cenários onde ocorrem tais colaborações. Assim, uma descrição dos cenários de colaboração existentes no processo é apresentada a seguir:

Intra-projetos: cenário onde se dá a colaboração envolvendo pesquisadores do mesmo centro de pesquisa genômico. Em geral, as pesquisas são distintas e a colaboração ocorre no sentido de discutirem as melhores soluções para seus projetos.

Inter-projetos: cenário onde se dá a colaboração entre pesquisadores de mais de um centro de pesquisa genômico. São centros de pesquisas parceiros e cujas pesquisas tendem a ser de caráter complementar, logo, é necessária a interação entre os pesquisadores para verificação de resultados.

Extra-projetos: cenário onde se dá a colaboração com centros de pesquisas genômicas de referência da comunidade de Bioinformática, como os centros mantenedores de bancos de dados como *Swiss-Prot* (BAIROCH, et. al, 2000) e ontologias como a *Gene Ontology*. A colaboração acontece no sentido de ajudar na evolução destes centros, fornecendo novas seqüências e anotações para o caso de

bancos de dados ou na inserção e atualização de novos termos para o caso das ontologias.

Nestes cenários percebemos a participação de anotadores e curadores no desenvolvimento dos projetos. Entretanto, estes papéis apresentam uma atuação diferenciada em cada um deles. Assim novos papéis também foram caracterizados:

Anotador (intra-projeto): anotador que realiza os experimentos nas bancadas biológica e computacional.

Anotador colaborador (intra-projeto): anotador que realiza experimentos nas bancadas biológica e computacional, e que colabora com os anotadores intra-projeto, sendo seus experimentos distintos.

Anotador externo (inter-projeto): anotador de outros projetos que realiza experimentos nas bancadas biológica e computacional, e que também colabora com os anotadores intra-projeto, sendo seus experimentos complementares.

Orientador (intra-projeto): anotador que orienta os experimentos realizados pelo anotador intra-projeto, fazendo acompanhamentos e avaliações sobre os experimentos pesquisados. Considerado também como o curador do grupo de pesquisa.

Orientador externo (inter-projeto): anotador de outros projetos que colabora na orientação dos anotadores intra e inter projetos, mas não participa diretamente dos experimentos pesquisados.

Curador de banco de dados (extra-projeto): responsável pela “curagem” da anotação antes de ser disponibilizada pelo banco de dados, pelo qual é responsável.

Curador de ontologia (extra-projeto): responsável pelas decisões tomadas referentes aos termos das ontologias.

Apesar do foco deste trabalho estar sobre o processo de anotação genômica, os cenários e os papéis descritos podem ser aplicados para o processo de estudo de seqüências genômicas e outros processos da área.

4.3 EXTENSÃO DO PROCESSO DE ANOTAÇÃO GENÔMICA COM FOCO NO USO DA ONTOLOGIA

Durante o processo do estudo de seqüências genômicas, a colaboração típica se dá nos momentos em que os anotadores interagem entre si e com os orientadores, buscando confirmar ou discutir suas anotações. Baseando-se no modelo do processo descrito no capítulo anterior, estes momentos se dão nas seguintes atividades:

Colaboração entre anotadores: acontece principalmente nas atividades da análise da seqüência. Por exemplo, um bioinformata pesquisa sobre o organismo *Trypanosoma rangeli* e outro pesquisa sobre *Phytomonas serpens* - (anotador e anotador colaborador intra-projeto). Procedimentos iguais terão de ser realizados para as pesquisas de tais organismos, logo, a colaboração acontece no sentido de realizarem esse trabalho juntos, discutindo itens como os parâmetros utilizados nos programas de análise e a necessidade de refazer algum procedimento. Outro exemplo diz respeito à colaboração entre anotadores de instituições diferentes porém parceiras (anotador externo inter-projeto) como: em uma instituição, um bioinformata estuda sobre o organismo *Trypanosoma rangeli* utilizando a metodologia de sequenciamento GSS (*Genome Survey Sequence*), e, na outra instituição, um outro bioinformata também realiza estudos sobre o mesmo organismo, contudo utilizando a metodologia de sequenciamento EST (*Expressed Sequence Tags*), onde a colaboração deve existir para que um saiba do andamento das pesquisas do outro.

Colaboração entre anotadores e orientadores: acontece nas atividades do processo de anotação como "observar e comparar resultados", a fim de verificar se os resultados da análise da seqüência são satisfatórios; e atividades da consolidação da anotação, nas atividades de "preparar artigo" e "rever artigo" onde participam o anotador e o orientador intra-projetos. Caso o projeto em estudo esteja sendo desenvolvido por mais de uma instituição, participam da colaboração os anotadores e orientadores de ambas (anotador e orientador externos, inter-projeto).

Colaboração com grupos externos: acontece quando uma seqüência e suas anotações são submetidas aos bancos de dados públicos e estas serão avaliadas pelo curador de banco de dados (extra-projeto) e/ou quando atualizações ou inserções de novos termos são solicitadas aos curadores das ontologias (extra-projeto).

A colaboração com grupos externos como os curadores das ontologias, como já comentado, é dependente da ação do anotador em reportar as sugestões de atualização e inserção de termos bem como informações de problemas ou apresentação de sucessos. Logo, se a anotação baseada em ontologia é pouco trabalhada, então esta colaboração fica diminuída.

De modo que a anotação baseada em ontologia seja uma forma de anotar mais efetiva no contexto dos projetos de pesquisa, entende-se que um estudo maior deve ser feito sobre uso dos termos das ontologias encontrados para cada seqüência, ou, o caso

da inexistência de termos, de maneira que este estudo colabore para o uso e evolução das ontologias, o compartilhamento de conhecimento e anotações futuras de melhor qualidade.

Para que este estudo aconteça, é ideal, portanto, que o processo de anotação genômica seja monitorado, pois é nesse contexto que se identificam os problemas no uso dos termos da ontologia. Assim, a partir da inserção de novas atividades no processo de anotação, os anotadores, orientadores e colaboradores passam a fornecer informações sobre o uso de termos da ontologia, ricas para o estudo em questão. Deste modo, é proposta uma reformulação no modelo do processo de anotação genômica descrito no capítulo anterior, para que este possa conter as atividades que englobam o uso dos termos das ontologias, constituindo assim, o modelo do processo de anotação genômica com foco no uso da ontologia, como mostra a FIG. 4.1 e explicado a seguir.

As atividades iniciais foram mantidas: “observar e comparar resultados”, “fazer busca automática por termos da ontologia” e “registrar anotação semi-automatizada”. O desenvolvimento dessas atividades segue do mesmo modo como explicado no processo de anotação genômica.

Após o resultado das atividades de busca automática de termos e anotação semi-automatizada, dá-se início a uma série de atividades que têm por objetivo capturar as ações realizadas pelos pesquisadores em relação à anotação baseada em ontologia. As atividades são descritas a seguir:

Analisar e classificar termo: inicialmente é feita uma verificação se a atividade de busca automática de termos da ontologia encontrou algum termo. Se o termo foi encontrado, este será a entrada da atividade “analisar e classificar termo”, juntamente com a anotação semi-automatizada. Esta atividade tem como objetivo fazer um estudo sobre o termo encontrado, analisando-o em relação à sua seqüência e assim fazer uma classificação sobre o mesmo. Logo, a saída da atividade consta das ações realizadas para a análise do termo e a sua classificação, além da anotação semi-automatizada.

Na análise dos termos, o anotador, com o apoio do orientador, executa algumas ações para embasar a classificação dos mesmos. Além desse embasamento, as informações registradas em cada ação podem ser utilizadas na tomada de decisões.

a) Consulta a um pesquisador experiente: muitos bioinformatas, devido ao tempo dedicado à pesquisa, conseguem identificar se o termo está condizente ou não com a seqüência estudada. Assim os anotadores entram em contato com pesquisadores mais experientes, que em geral são os orientadores intra e inter-projetos, e estes

auxiliam os anotadores a esclarecerem qualquer dúvida relativa aos termos. Os orientadores podem inclusive solicitar que outras verificações (como as descritas a seguir) sejam feitas também. Todas as conversas entre os participantes são registradas.

b) Busca na literatura: esta ação visa encontrar na literatura especializada da área, informações sobre o gene contido na seqüência, tais como, verificar se é um gene já descoberto e se existem termos de ontologias cadastrados para ele. Caso esta ação seja realizada, registra-se a literatura consultada.

c) Revisão da análise de similaridade: o objetivo desta revisão é identificar qual(is) anotação(ões) das seqüências similares, justifica(m) a anotação para os termos encontrados para as seqüências pesquisadas. Neste caso, registra-se a anotação da seqüência mais similar.

d) Pesquisa a bancos de dados específicos: bancos de dados genômicos como os de domínio ou família de proteínas trazem informações mais significativas, podendo assim confirmar o termo escolhido das ontologias. Neste caso, registra-se o banco consultado e o(s) identificador(es) relacionado(s).

e) Discussões entre pesquisadores: uma ação a ser efetuada são as discussões entre os diversos anotadores do grupo de pesquisa (anotador e anotador colaborador intra-projetos) ou anotadores externos (inter-projetos). Pontos de vista diferentes são analisados, de modo a se chegar a um consenso sobre o termo da ontologia.

Uma ou mais ações podem ser realizadas e assim, elas irão ajudar a classificar o termo como: adequado ou inadequado, sendo que a classificação do termo como inadequado significa que o termo pode ser:

a) Incorreto: termo que não corresponde ao organismo estudado. Exemplo: foi encontrado o termo *spermatogenesis* para anotar uma seqüência de um organismo que é parasita, como os *trypanosomas*. Esse termo está incorreto devido ao fato de que parasitas não produzem espermas.

b) Pouco específico: o termo não corresponde totalmente à funcionalidade do gene. Exemplo: o termo *ATP-dependent helicase activity* é uma representação significativa, pois denomina a função de um determinando gene contido numa seqüência; contudo, em algumas seqüências, observa-se que é uma atividade presente no RNA, logo o melhor termo seria *ATP-dependent RNA helicase activity*, o que torna o primeiro pouco específico.

c) Inconsistente: ocorre devido a problemas estruturais das ontologias, como os problemas referentes aos relacionamentos “part of”, e “is a” da GO (SMITH ET AL., 2004) [10]. Exemplo: na GO o termo *carboxypeptidase A activity* possui três pais de acordo com o grafo direcionado acíclico que compõe a estrutura das ontologias da GO. Isso causa um problema referente ao relacionamento “is a”, pois não se sabe ao certo a qual dos três ramos, *carboxypeptidase activity*, *metalloexopeptidade activity e metallopeptidade activity*, o termo *carboxypeptidase A activity* pertence.

De acordo com a classificação que o termo recebe, outras atividades serão realizadas durante o processo de anotação genômica visando capturar as ações realizadas pelos pesquisadores sobre o termo escolhido.

Registrar ocorrência para termo adequado: se o termo encontrado for classificado como adequado, significa que a busca automática de termos foi bem sucedida. Então, a atividade a ser realizada é “registrar ocorrência para termo adequado”, pois futuras anotações poderão se apoiar neste registro e ficará claro para o anotador que uma anotação similar anterior foi realizada com um termo considerado adequado, além das ações realizadas que confirmam a escolha do termo. A entrada desta atividade são as ações que foram feitas para confirmar a classificação do termo, a própria classificação e o termo. A saída consta do registro dos elementos da entrada e o termo adequado da ontologia.

Registrar anotação semi-automatizada baseada em ontologia: o termo adequado deve ser usado para também anotar a seqüência alvo, e então será a entrada da próxima atividade, “registrar anotação semi-automatizada baseada em ontologia”, a qual possui como objetivo unir as duas formas de anotação: a semi-automatizada e a que faz uso de termos da ontologia. Portanto, a saída da atividade é constituída desta anotação.

Registrar ocorrência para termo inadequado ou inconsistente: se o termo encontrado não for classificado como adequado, ele se encontra em uma das classificações: incorreto, pouco específico ou inconsistente. Então, é realizada a atividade “registrar ocorrência para termo inadequado e problemas”. Este registro também apoiará as futuras anotações. Neste momento são também registrados quais os problemas que ocorrem com o termo para serem classificados de tal forma:

a) Incorreto: registrar o porquê de o termo estar incorreto e por isso inadequado para uso na anotação.

b) Pouco específico registrar o porquê da pouca representatividade do termo.

c) Inconsistente: registrar qual a inconsistência.

A entrada desta atividade são também as ações que foram feitas para a classificação do termo, a classificação e o termo. A saída consta do registro dos elementos de entrada e o termo inadequado da Ontologia.

Fazer busca manual no banco de termos da ontologia: após o registro da ocorrência para termo inadequado e o problema de cada um, a próxima atividade a ser realizada é “fazer busca manual no banco de termos da ontologia”. Esta atividade será também realizada para o caso em que a atividade de busca e refinamento de termos da ontologia não encontrou termo para a anotação da seqüência. As atividades seguintes referem-se então aos dois casos.

Buscar manualmente no banco de termos da ontologia: essa atividade tem por objetivo verificar se o termo correto existe através de uma busca manual feita sobre um banco de dados remoto, em geral disponível no *site* da ontologia. A entrada desta atividade inclui o termo inadequado e a anotação semi-automatizada, para o caso da existência de termo, e somente a anotação semi-automatizada, caso contrário. Além disso, a entrada inclui também as ações realizadas para a classificação do termo, que servirão de apoio para a busca no banco de termos da ontologia. Estes bancos geralmente encontram-se disponíveis através de sistemas na *Web*, o que possibilita uma busca manual por meio da navegação no sistema. Na saída da atividade, portanto, constará o termo correto caso este seja encontrado.

Registrar falha na busca de termos: se com a busca manual de termos da ontologia, o termo adequado foi encontrado, isto evidencia que houve uma falha nos algoritmos de busca automática de termos da ontologia. Assim, para este caso, a atividade a ser feita é “registrar falha na busca de termos”, que tem por objetivo capturar esta informação. Esta falha pode ocorrer quando a busca automática por termos for realizada sobre uma versão desatualizada da ontologia local em uso. Logo, a entrada desta atividade é o termo correto juntamente com a anotação semi-automatizada, e na saída, além dos itens da entrada, consta ainda a informação sobre a falha na busca do termo. Este registro indica ao grupo de pesquisa a necessidade de arrumar ou melhorar o sistema de anotação que o apóia. O termo adequado e a anotação semi-automatizada são as entradas para a atividade seguinte, “registrar anotação semi-automatizada baseada em ontologia", a qual foi descrita anteriormete.

Se não tiver sido encontrado o termo adequado com a busca manual por termos da ontologia, é feita uma verificação referente a se algum termo existe na anotação para decidir quais atividades devem ser seguidas.

Registrar desatualização: se não foi encontrado o termo adequado na busca manual, mas a anotação consta de um termo inadequado, então a atividade "registrar desatualização" é realizada. A entrada desta atividade é o termo inadequado e a anotação semi-automatizada e possui como objetivo registrar a informação de desatualização. Este registro significa que, apesar de ser, por exemplo, um gene ou produto de gene já descoberto, e com termos para anotação já concebidos e presentes na literatura, a ontologia em uso ainda não possui tal referência, devido ao fato do termo correspondente ainda não constar em seu banco de termos. Esta situação costuma ocorrer, pois é difícil, para os curadores de ontologias, acompanhar a contento as descobertas nesta área (por exemplo, a descoberta de novos genes). Além do registro da informação, na saída desta atividade consta o termo inadequado e a anotação semi-automatizada.

Registrar ineditismo: se as buscas automática e manual não tiveram sucesso, e não existir nenhum termo na anotação, a atividade a ser feita é "registrar ineditismo", cuja entrada é a anotação semi-automatizada. O objetivo desta atividade é registrar a informação de ineditismo, o que significa que a seqüência pesquisada, a qual deseja encontrar um termo, possui um gene inédito. Sendo assim, este ainda deve ser reportado aos curadores da ontologia para serem inseridos os termos representantes do novo gene na mesma. A saída da atividade consta do registro da informação e da anotação semi-automatizada.

Uma vez que o termo adequado não foi encontrado com a busca manual no banco de termos da ontologia, e após ser feito o registro da desatualização ou do ineditismo, o anotador decide se há uma sugestão de termo a ser feita.

Registrar sugestão de termo: se existe a sugestão de termo, a próxima atividade "registrar sugestão de termo" é realizada. As entradas dessa atividade são: o registro sobre a desatualização ou o ineditismo, a anotação semi-automatizada e o termo inadequado; este último é desconsiderado para o caso da inexistência de termo. Com base nas informações de entrada, na análise do termo e através de suas experiências profissionais, o anotador e o orientador estão aptos a fazerem a sugestão de um novo termo para anotação, por exemplo, de um gene contido na seqüência estudada. O objetivo desta atividade é fazer o registro deste novo termo. Logo, as saídas dessa

atividade constam dos elementos de entrada adicionados do termo sugerido. A sugestão do termo é importante, pois ao ser reportada aos curadores da ontologia, ajudará na atualização de termos existentes ou até mesmo na inserção de um novo termo. Entretanto, o termo sugerido não faz parte da anotação sobre a seqüência estudada.

Decidir quanto ao uso do termo: se não há sugestão de termo, mas há alguma anotação como a de termos inadequados, então parte-se diretamente para a atividade "decidir quanto ao uso do termo", a qual também será realizada caso haja termo sugerido. As entradas desta atividade são: o registro sobre a desatualização ou o ineditismo, a anotação semi-automatizada, o termo inadequado, desconsiderado para o caso da inexistência de termo, e a sugestão do termo se houve. O objetivo desta atividade é decidir se o termo existente na anotação, mesmo sendo inadequado, pode ser utilizado para fazer a anotação. Isto porque as classificações do termo como pouco específico ou inconsistente (para termo inadequado) podem não ser a melhor escolha para a anotação da seqüência, porém estes termos ainda podem ser utilizados para anotar. Caso se opte por usar o termo, mesmo que inadequado, então a atividade seguinte é "registrar anotação semi-automatizada baseada em ontologia", já descrita anteriormente.

Verificar anotação: a atividade "verificar anotação" tem por objetivo fazer uma verificação na anotação a fim de validá-la juntamente com orientadores intra e inter projetos. A entrada dessa atividade, portanto, é a anotação semi-automatizada baseada em ontologia ou somente a anotação semi-automatizada para o caso da inexistência de termo. Nesta atividade, há uma discussão entre o orientador do estudo e o orientador externo sobre a anotação realizada, verificando se a mesma está correta, se necessita de complementação ou se possui problemas. Se a anotação não for validada, o processo de "análise da seqüência" deve ser refeito como comentado no processo de anotação genômica descrito no capítulo anterior.

Notificar anotação: se a anotação for validada, deve-se então "notificar a anotação", cujo objetivo é notificar aos projetos parceiros/colaboradores que uma anotação foi realizada. Uma proposta de sistema para notificação é descrita por (ALMEIDA, 2006). Através das observações do processo feitas neste trabalho foi possível perceber a necessidade de suporte a colaboração entre o grupo de pesquisa e seus consorciados, no que diz respeito a atualização de dados referentes a anotações relacionadas. Logo, há uma necessidade freqüente de notificar os colaboradores inter e

intra-projetos, como por exemplo, membros dos grupos do consórcio, residentes em outros centros de pesquisa, sobre novas anotações que impactam em sua pesquisa.

Após a notificação das anotações passa-se para o processo seguinte que é a “consolidação da anotação”, descrito no capítulo anterior.

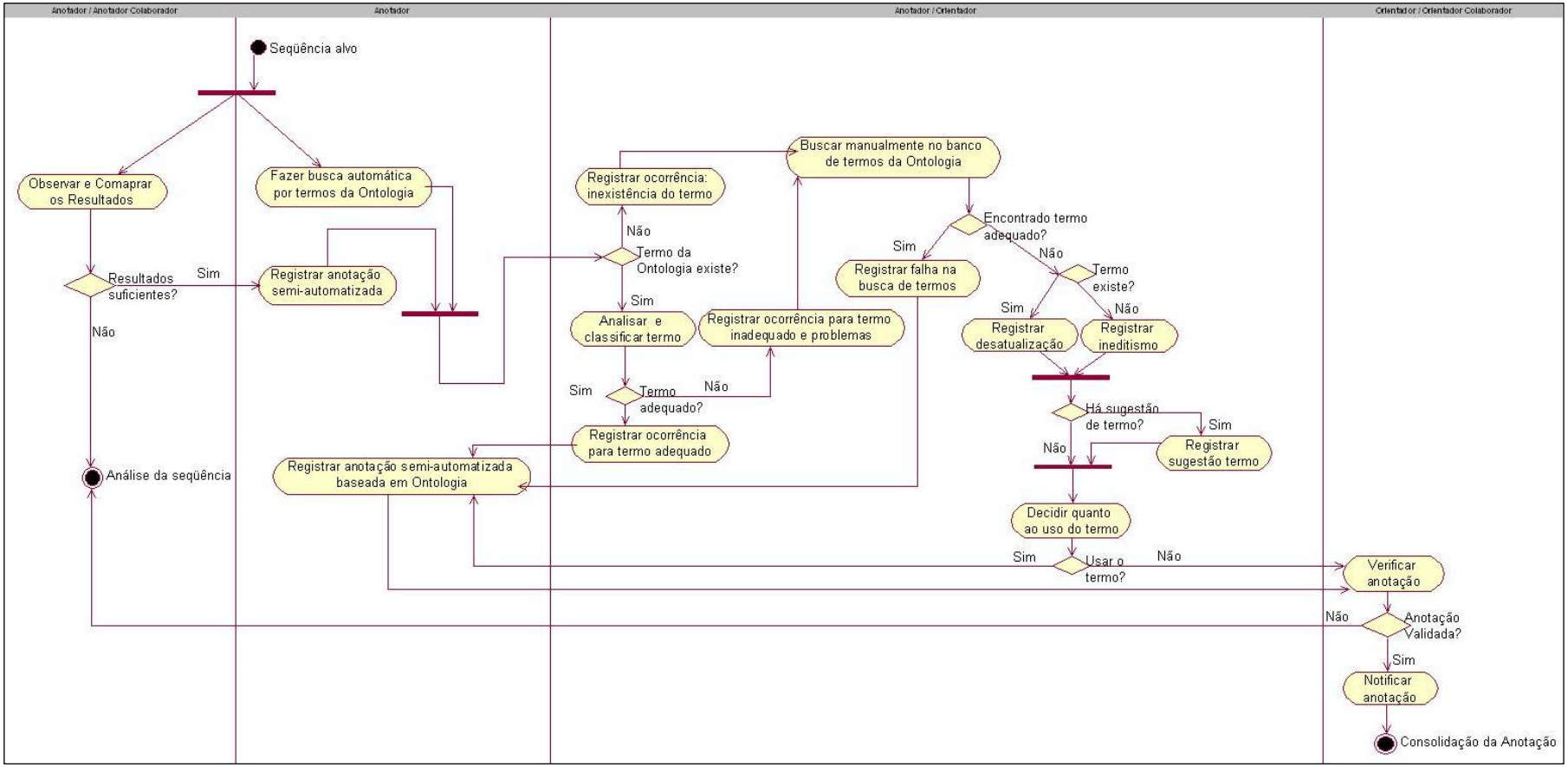


FIG. 41 Processo de anotação genômica com foco no uso da Ontologia

4.4 CONSIDERAÇÕES

O objetivo na descrição dos processos do estudo de seqüências e anotações genômicas parte não somente de sua caracterização realizada através da modelagem, mas também da identificação de momentos onde ocorre colaboração, seja entre as pessoas do mesmo grupo, entre grupos distintos ou através de contatos com instituições de referência da comunidade de Bioinformática. A modelagem e a descrição de cada atividade detalhadamente permitiram que estes momentos se tornassem mais fáceis de serem identificados, e que uma forma de apoio à colaboração fosse sugerida, a qual envolve um incentivo maior ao uso das ontologias.

Para tanto, novas atividades que envolvem a monitoração sobre a anotação baseada em ontologia foram introduzidas no processo de anotação genômica, de modo que esta informação não se perca. A partir do registro destas informações é possível realizar o estudo sobre o uso das ontologias, tão importante para a ampliação da colaboração.

As novas atividades têm a finalidade de registrar o raciocínio feito na execução das mesmas. Em futuras anotações, os anotadores então, já teriam um maior conhecimento sobre os termos escolhidos, o que facilitaria suas anotações. Para novos anotadores, este registro serviria como um apoio em suas próprias ações e anotações. E para todos os casos, as informações de sucesso, de inadequação do termo ou de termo inexistente servirão para comporem um documento formal para ser remetido aos curadores da Ontologia de forma que as informações estarão bem organizadas e detalhadas.

No caso de termos inadequados, por exemplo, o problema estará bem caracterizado, e assim torna-se mais fácil sugerir a inclusão de novos termos, de forma documentada, bem como solicitar atualizações de termos existentes. Desta maneira, fica evidenciado que se o centro de pesquisa tornar estes problemas visíveis aos curadores de ontologias é possível concluir anotações de maior qualidade mais rapidamente, além de facilitar a colaboração com a evolução da Ontologia, através da sugestão de novos termos ou atualização.

Com a anotação baseada em ontologia monitorada e registrada, a tendência é que o uso desta seja cada vez maior, e por conseqüência facilitará a troca de informações entre todos que colaboram e o compartilhamento do conhecimento. Aumentando-se o uso deste tipo de anotação, melhora-se, consequentemente, a qualidade das mesmas.

Com o registro das ações realizadas pelos anotadores em relação à anotação baseada em ontologia, cria-se um ciclo produtivo, que resulta em melhores anotações. A FIG.

4.2 mostra uma visão geral sobre o que este registro proporciona. A partir da anotação baseada em ontologia e do registro do raciocínio da escolha dos termos, o conhecimento sobre estes se amplia, uma vez que cada anotação e os termos são avaliados isoladamente e deste modo, pode-se colaborar com os curadores da ontologia, reportando tais informações, o que ajudará na atualização da Ontologia e, por conseguinte permitirá que mais termos sejam encontrados quando for realizada uma nova busca de termos.

FIG. 4.2 Ciclo produtivo de anotação

Para avaliar melhor os benefícios da extensão do processo de anotação genômica, foi construída uma ferramenta onde o raciocínio sobre a anotação baseada em ontologia pode ser registrado. Esta ferramenta é descrita em detalhes no capítulo seguinte.

5 GARSA NOTES

O GARSA Notes foi desenvolvido para viabilizar a validação da modelagem proposta no capítulo anterior, a qual tem por objetivo monitorar a anotação baseada em ontologia onde novas funcionalidades capturam o raciocínio feito pelos pesquisadores sobre esta anotação para cada seqüência estudada.

O desenvolvimento permitiu mostrar uma forma prática de registro das ações dos anotadores sobre esta monitoração contribuindo assim para que as informações sobre problemas encontrados assim como as informações de sucesso sejam registradas no momento em que ocorrem. O protótipo GARSA Notes foi então desenvolvido baseando-se na expansão do processo de anotação genômica descrito no capítulo anterior.

O GARSA Notes foi desenvolvido como uma extensão do sistema GARSA. Esta escolha foi devida à facilidade de acesso ao código e a desenvolvedores do mesmo. Outra razão refere-se ao GARSA cobrir funcionalidades previstas na modelagem como as anotações: automática e manual, e principalmente a anotação baseada em ontologia.

Este capítulo traz inicialmente uma visão geral do protótipo GARSA Notes mostrando seus módulos principais na seção 5.1. A seção 5.2 detalha cada módulo do protótipo, apresentando as funcionalidades que permitem a captura das ações dos anotadores por trás da anotação baseada em ontologia A seção 5.3 apresenta o modelo de dados que dá suporte ao registro dos dados. A seção 5.4 apresenta o ambiente de desenvolvimento do protótipo, e por fim a seção 5.5 descreve algumas considerações relativas ao sistema GARSA.

5.1 VISÃO GERAL DO PROTÓTIPO

Para compreensão de como o GARSA Notes está inserido no sistema GARSA, esta seção inicialmente descreve o módulo deste sistema onde foi inserida a chamada ao GARSA Notes.

Como descrito no capítulo 3, o *workflow* implementado pelo sistema GARSA engloba diversos programas de análise que realizam anotações semi-automatizada e também baseada em ontologia. Para ter acesso às anotações semi-automatizadas, é preciso utilizar o módulo de busca de sequências do GARSA, cuja interface é mostrada na FIG. 5.1.

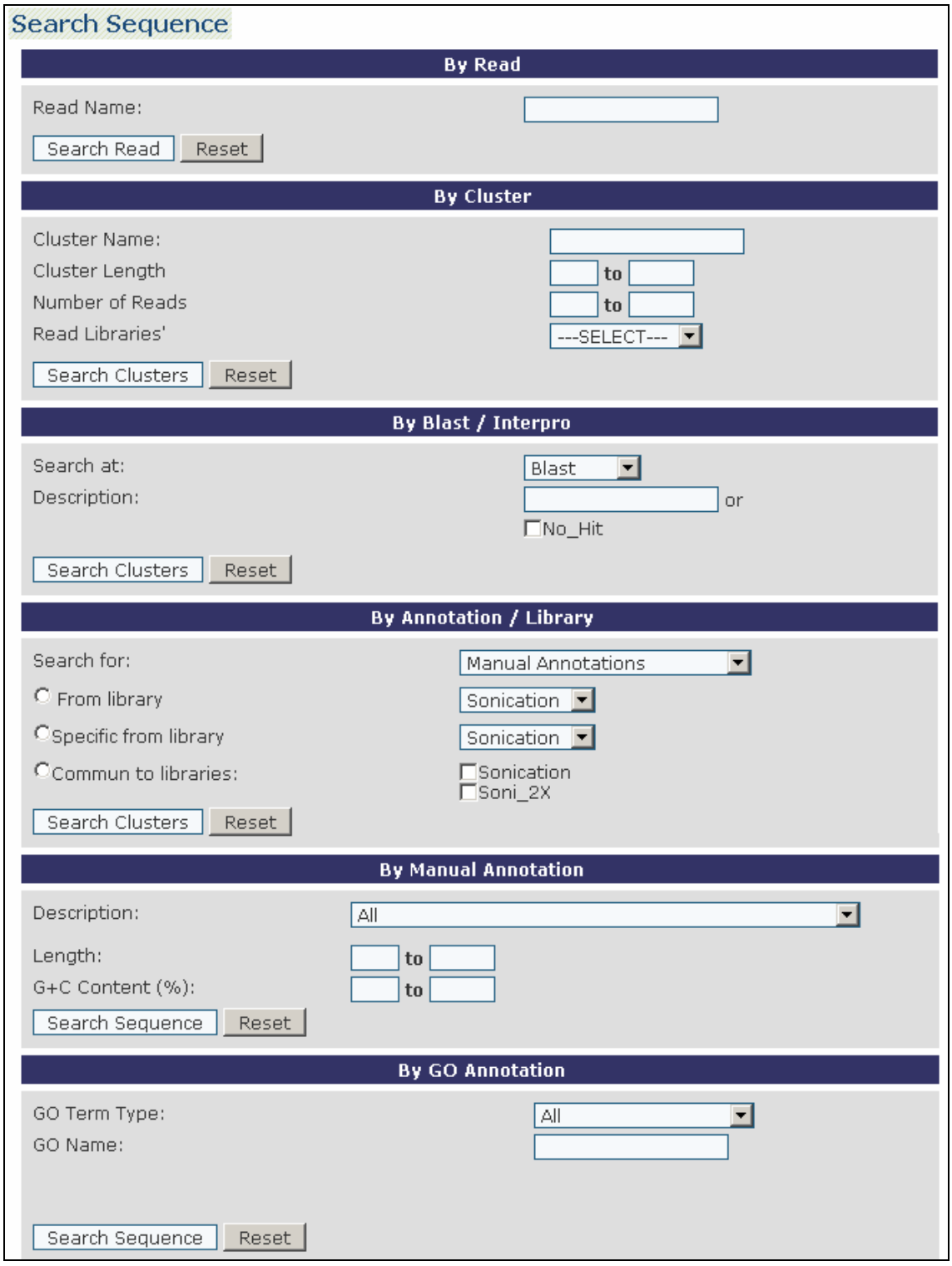


FIG 5.1 Buscas de seqüências (GARSA, 2007)

As anotações podem ser buscadas por identificação de *read* e de cluster, sendo que a busca por cluster traz uma lista destes que pode ser referenciada por análises de Blast ou Interpro, por anotações referentes às bibliotecas, por anotação manual ou por anotação da GO. Na busca de seqüências por cluster, pode-se escolher listar todos os clusters e

verificá-los isoladamente, pois os clusters contêm os resultados dos programas de análise e as anotações correspondentes. Estas anotações, por sua vez são verificadas se estão adequadas ou se necessitam de alguma complementação. Assim, caso o anotador ache necessário fazer um complemento ou uma correção, estas podem ser alteradas. Por exemplo, a FIG. 5.2 mostra a tela do GARSA que permite a alteração de uma anotação semi-automática para o cluster TGEG101003D05.g. Os dados se referem ao estudo do organismo *Trypanosoma rangeli* (WAGNER, 2006), armazenado no GARSA e que será utilizado como exemplo no decorrer do capítulo. Note que para esta anotação, a busca automática por termos da ontologia gerou três sugestões de termos, um para cada ontologia da GO: Função Molecular, Processo Biológico ou Componente Celular.

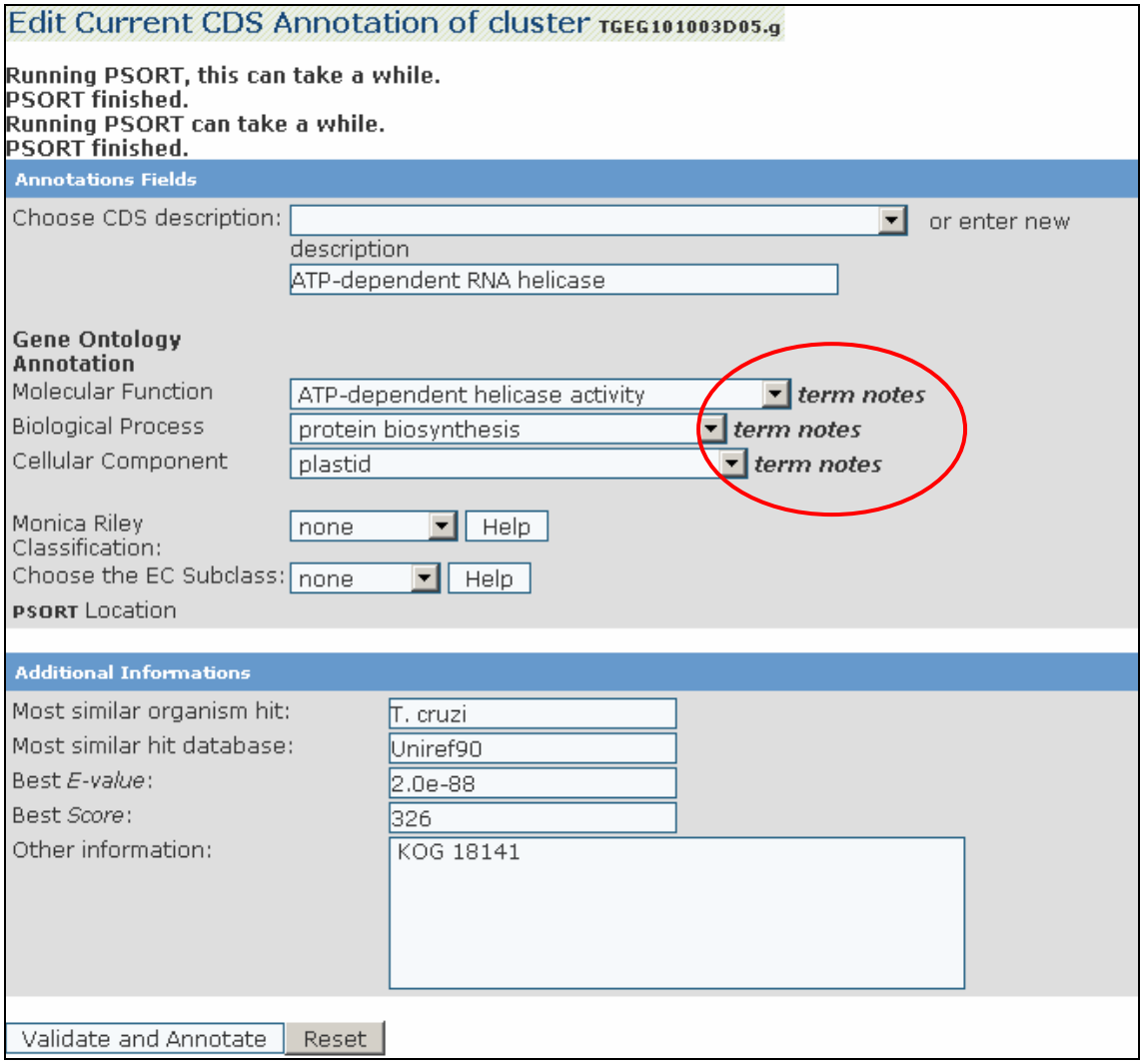


FIG. 5.2 Edição das anotações

Para cada termo representante de cada ontologia da GO foi inserida a opção *term notes* que faz a ligação com o GARSA Notes. A partir da ativação deste *link*, o módulo

do GARSA Notes a ser ativado depende do termo ter sido ou não encontrado. Se houver termo na anotação, ativará o módulo que trata termo encontrado, se não houver termo (representado nas ontologias pela palavra *none*) então o módulo a ser ativado é o que trata de termo não encontrado. Estes dois módulos interagem com o módulo que trata da geração de relatórios.

Portanto, o GARSA Notes é subdividido em três módulos. De forma simplificada, a FIG. 5.3 mostra estes módulos e o módulo do GARSA dá acesso ao GARSA Notes.

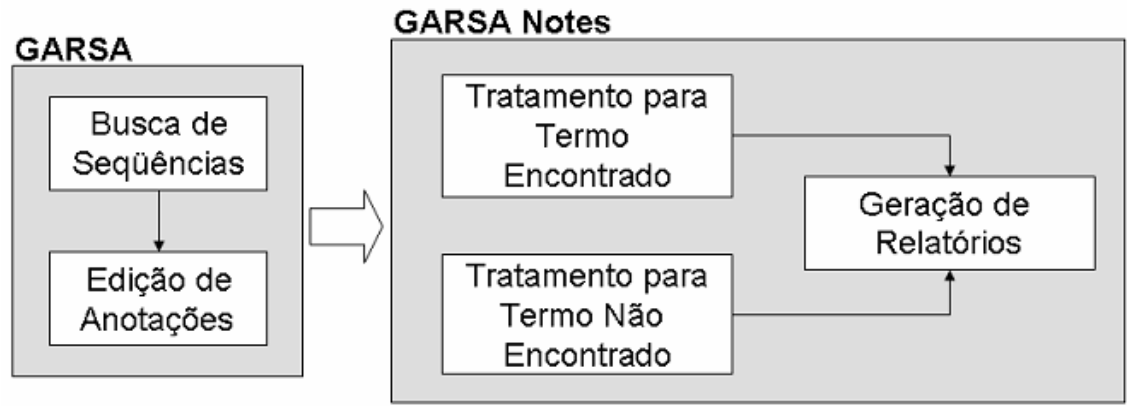


FIG. 5.3 Módulos do GARSA Notes

Tratamento de termo encontrado: este módulo contém os itens referentes às atividades de análise e classificação do termo: consulta a um pesquisador experiente, busca na literatura, pesquisa a bancos de dados específicos, revisão da análise de similaridade e discussão entre pesquisadores. Após ser feito o registro de pelo menos um destes itens é possível realizar a classificação do termo como: adequado, incorreto, pouco especifico ou inconsistente.

Tratamento de Termo não encontrado: neste módulo é sugerida a busca manual através de navegação no banco de termos da ontologia disponível no site oficial da ontologia a fim de verificar se o termo realmente não existe, ou se este não foi encontrado por falha dos algoritmos de busca. Supõe-se que o banco de termos da ontologia replicado localmente, para a realização da busca automática, está atualizado em relação a sua fonte. Caso o termo representante daquela seqüência tenha sido encontrado, a falha é registrada; caso contrário, há a possibilidade da sugestão de um novo termo.

Geração de relatórios: este módulo consiste na geração de relatórios onde possam ser mostradas todas as informações relevantes de acordo com o que se deseja: apoio à anotação ou envio para os curadores da ontologia. Isto porque o objetivo final de todo o

registro do raciocínio feito pelos pesquisadores sobre a anotação baseada em ontologia é poder utilizar as informações registradas de modo a facilitar as anotações futuras e a colaborar para a evolução da ontologia. Desta maneira, torna-se necessário a geração de relatórios. Assim, independente se o termo foi ou não encontrado, um relatório referente ao *cluster* em estudo pode ser gerado. Outras opções de relatórios englobam um relatório geral contendo todos os clusters, outro que traz os termos e todas as informações referentes a cada um deles e outro contendo todas as sugestões de termos.

5.2 FUNCIONALIDADES DO GARSA NOTES

As funcionalidades do GARSA Notes foram desenvolvidas seguindo a extensão do modelo de processo de anotação genômica proposta no capítulo 4. De acordo com os módulos citados anteriormente, os quais englobam esta extensão, algumas das funcionalidades são descritas a seguir.

5.2.1 TERMO ENCONTRADO

Na tela de edição de anotação do sistema GARSA (FIG. 5.2), ao acessar a opção *term notes* para cada termo encontrado referente a uma das três ontologias sobre um determinado cluster, leva-se à funcionalidade de análise e classificação do termo do GARSA Notes mostrada na FIG. 5.4.

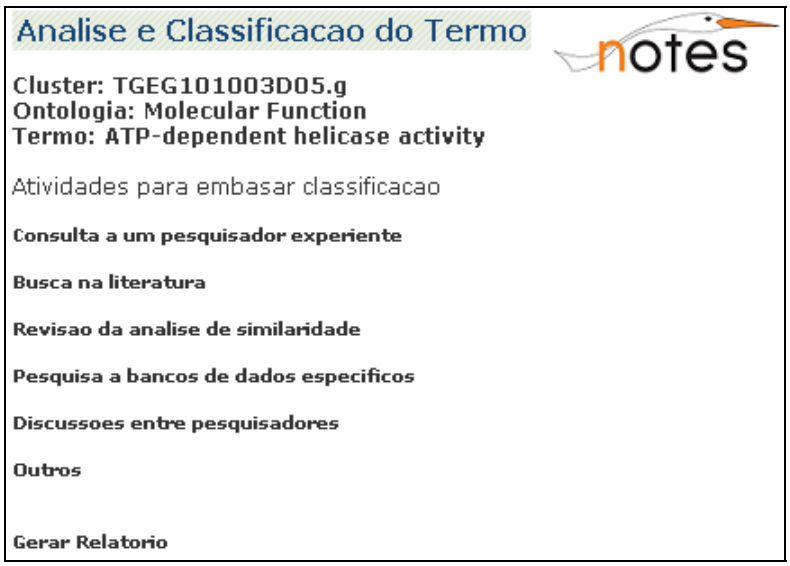


FIG. 5.4 Análise e classificação do termo encontrado

Observe na figura que os itens *cluster, ontologia e termo* vêm preenchidos (no exemplo, de acordo com os estudos do organismo *Trypanosoma rangeli*). Esta funcionalidade está presente em todas as atividades de análise, de modo que o anotador esteja sempre atento a qual anotação está se referenciando. Os itens a seguir apresentam as funcionalidades de cada atividade.

A) CONSULTA A UM PESQUISADOR EXPERIENTE

Uma das atividades mais significativas ou importantes é a consulta a um pesquisador experiente. Através dela, os anotadores podem interagir com diversos especialistas, como o orientador do grupo de pesquisa, o orientador colaborador ou outro especialista cadastrado. A FIG. 5.5 mostra o detalhamento da funcionalidade.

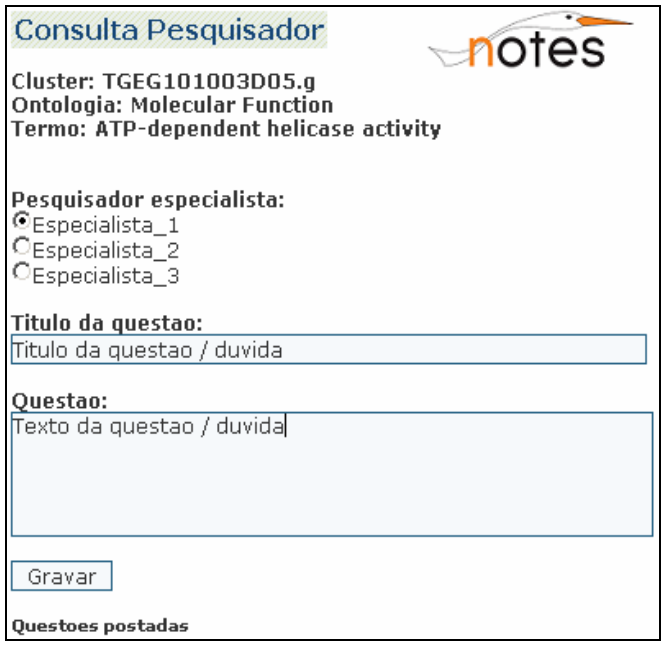


FIG. 5.5 Consulta pesquisador – perguntas

Ao acessar esta atividade são listados todos os especialistas cadastrados que podem opinar sobre o projeto em estudo (usuários cadastrados no banco de dados Seqonsql do sistema GARSA do sistema GARSA que têm acesso ao projeto). O anotador então seleciona qual especialista deseja consultar e informa o título (assunto) de sua dúvida ou qualquer tipo de esclarecimento e o texto em questão.

Ao ser gravada a informação, um e-mail é automaticamente enviado para o especialista selecionado informando sobre a nova postagem, como mostra a FIG. 5.6.

O e-mail traz as seguintes informações: nome do anotador, cluster, ontologia, termo, título e questão. Possui também um *link* que abrirá o GARSA Notes diretamente na tela onde a resposta deve ser postada, de acordo com a FIG. 5.7.

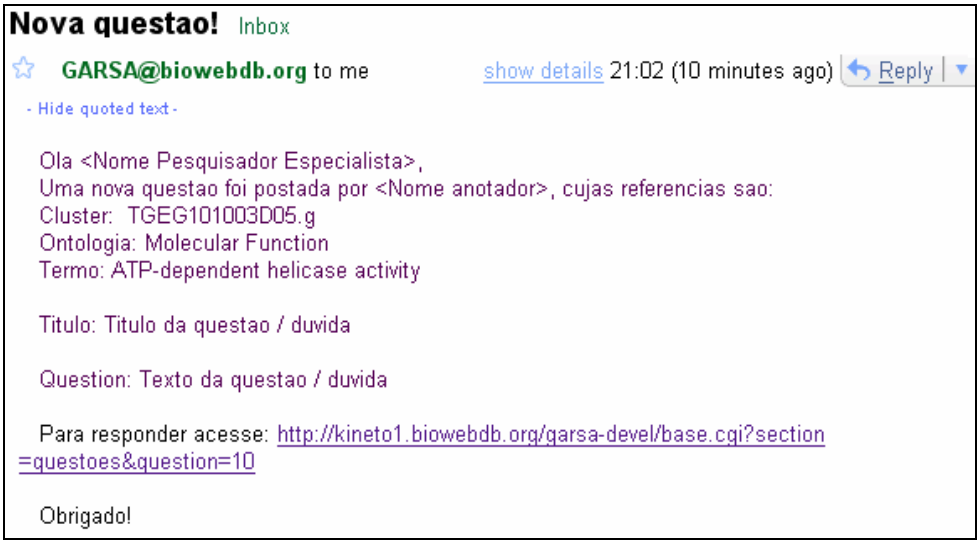


FIG. 5.6 E-mail enviado ao pesquisador especialista

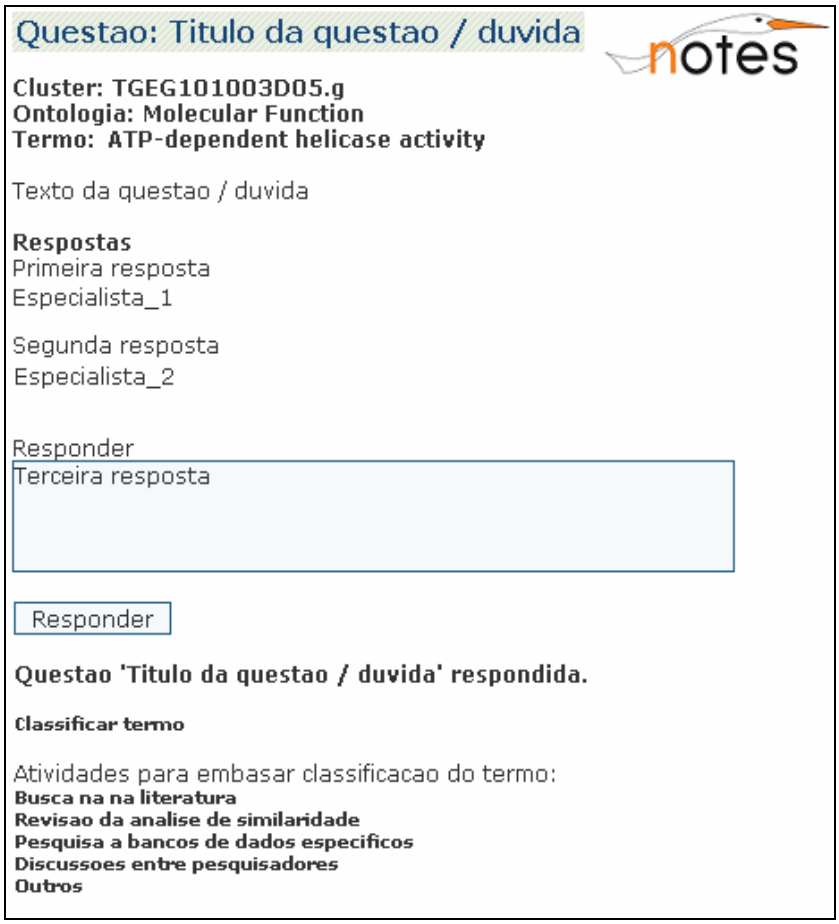


FIG. 5.7 Consulta pesquisador – respostas

A tela de respostas é acessível também através da listagem das questões postadas, que por sua vez podem ser acessadas através do item *Questões postadas* da tela de cadastro de questões (FIG. 5.5).

Logo que uma resposta é postada, é habilitada a opção de *Classificar termo* e também as outras opções para análise do termo. Se o anotador já tiver esclarecido sua dúvida, pode diretamente classificar o termo e passar para o estudo do termo de outra ontologia. Caso contrário, ele pode acessar qualquer uma das outras atividades, a fim de obter maior embasamento. Estas opções aparecem em todas as atividades de análise, mediante a inserção de pelo menos um registro. Isto é feito para garantir que o anotador tenha embasado, de acordo com pelo menos uma atividade, a classificação do termo.

B) BUSCA NA LITERATURA

Esta atividade permite registrar todas as referências buscadas na literatura para confirmação do termo. A FIG. 5.8 ilustra esta atividade.

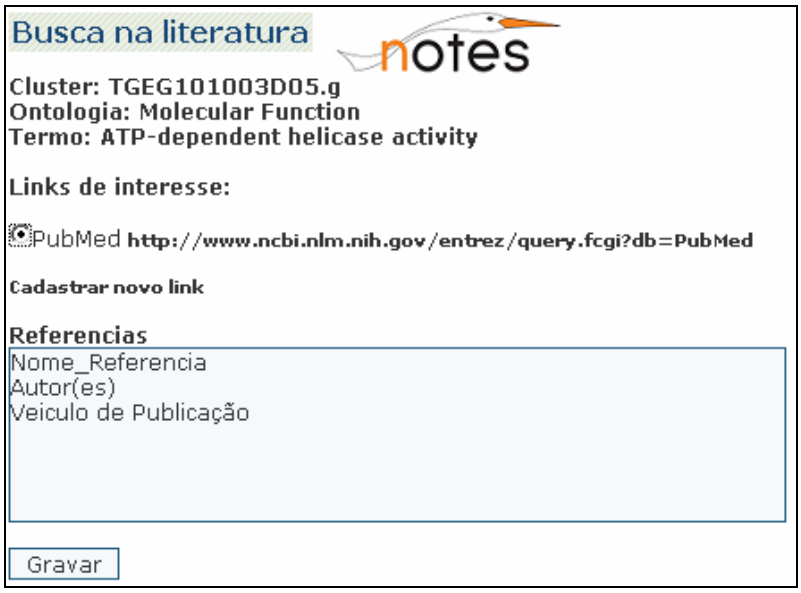


FIG. 5.8 Busca na literatura

De forma a facilitar a utilização por parte dos anotadores, esta funcionalidade traz uma lista de *links* de interesse para referências na *Web*. A marcação deste item não é obrigatória, uma vez que a referência consultada pode ser algum livro ou outro material somente impresso. Podem ser registradas quantas referências forem necessárias.

Os *links* de interesse podem ser cadastrados pelo anotador a medida que surjam novas referências. Basta acessar a opção *Cadastrar novo link.* O cadastro consta do nome da referência e a sua URL, como mostra a FIG. 5.9.

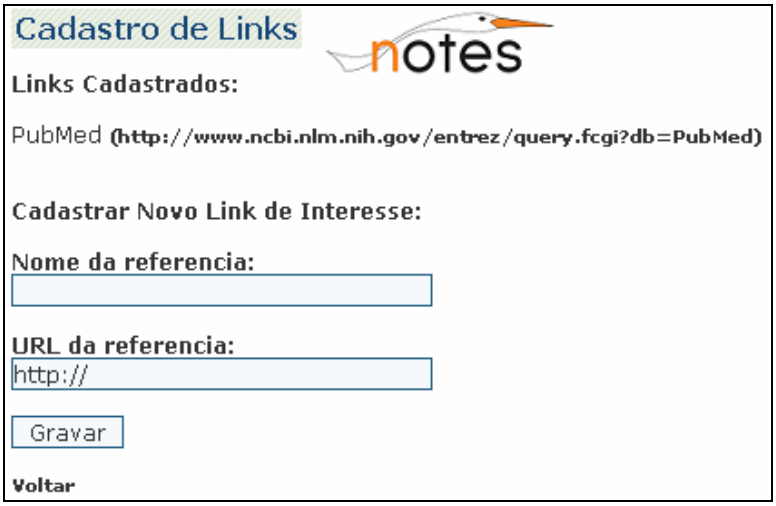


FIG. .5.9 Cadastro de links de interesse

C) PESQUISA A BANCOS DE DADOS ESPECÍFICOS

Esta atividade permite armazenar os bancos de dados que foram pesquisados para a confirmação do termo, conforme pode ser visto na FIG. 5.10.

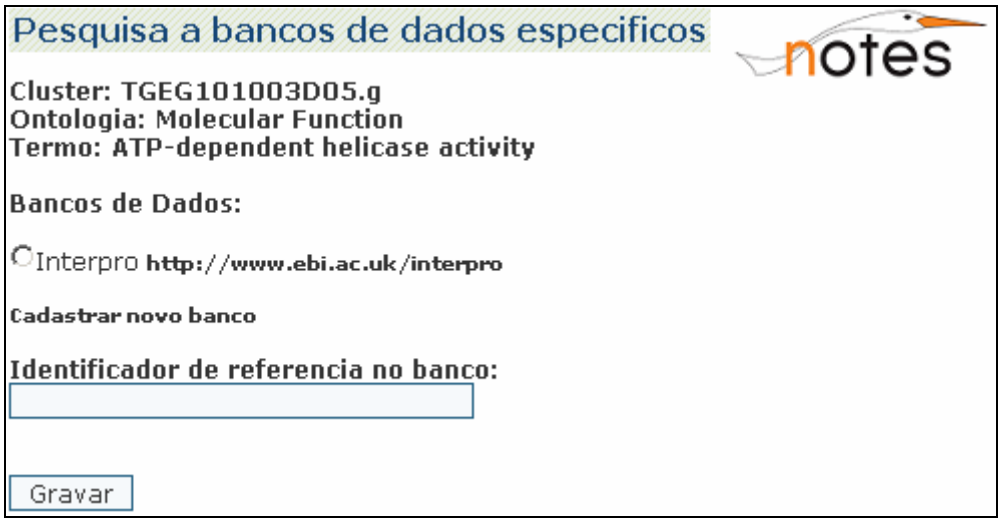


FIG. 5.10 Pesquisa a bancos de dados específicos

O anotador quando realiza esta atividade deve registrar qual o banco de dados onde se encontrou a informação para a confirmação do termo da ontologia e registrar também qual o identificador desta informação no banco de dados em questão. Semelhante à atividade de busca na literatura, também podem ser registradas quantas informações relativas a pesquisas em bancos de dados forem necessárias.

Caso o banco de dados não esteja listado com seu respectivo *link*, o anotador deve necessariamente cadastrá-lo. A FIG. 5.11 ilustra esta funcionalidade. O cadastro consta do nome do banco de dados e URL para acesso a ele.

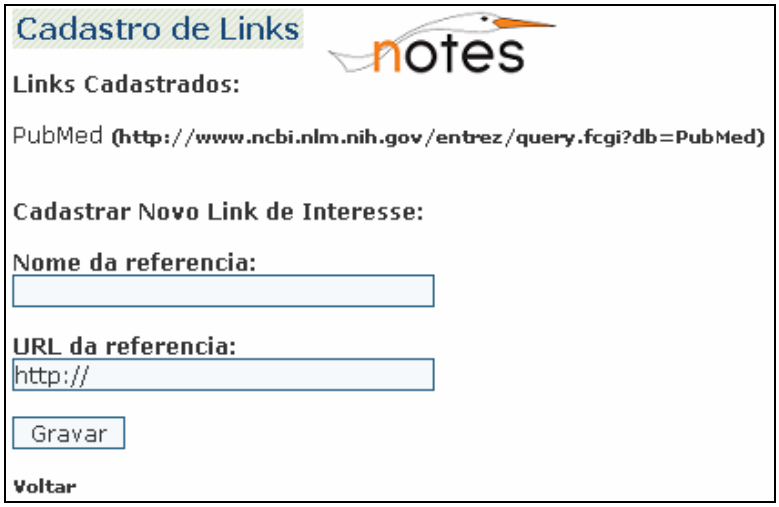
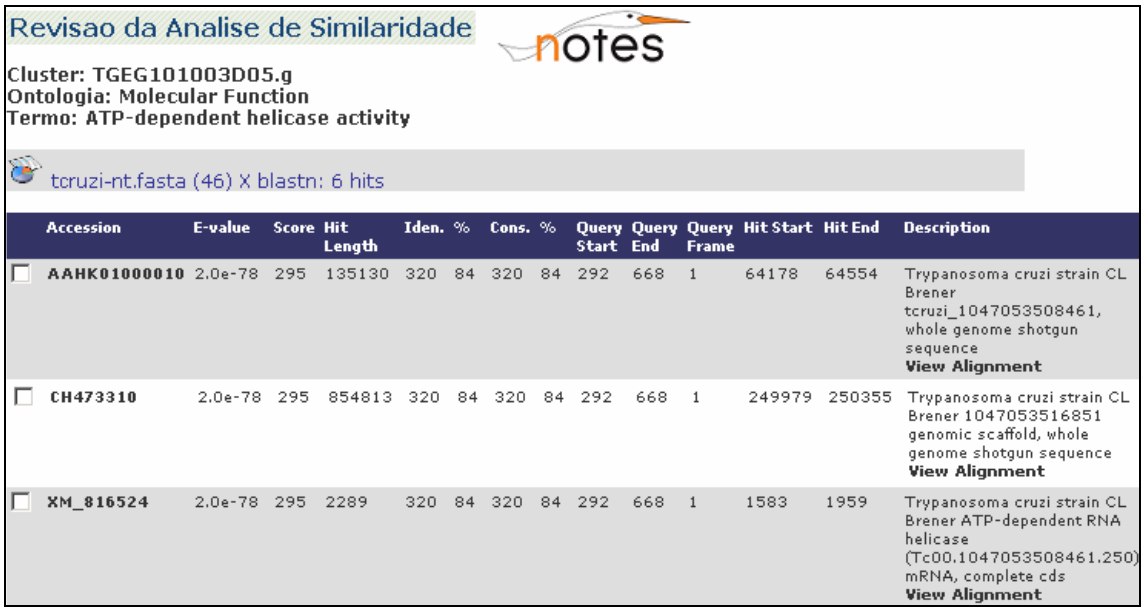


FIG. 5.11 Cadastro de bancos de dados

D) REVISÃO DE SIMILARIDADE

A atividade de revisão de similaridade oferece ao anotador visualizar os resultados dos programas de busca por similaridade entre seqüências de outros bancos de dados, como pode ser visto na FIG. 5.12.



	Accession	E-value	Score	Hit Length	Iden. %	Cons. %	Query Start	Query End	Query Frame	Hit Start	Hit End	Description
☐	AAHK01000010	2.0e-78	295	135130	320 84	320 84	292	668	1	64178	64554	Trypanosoma cruzi strain CL Brener tcruzi_1047053508461, whole genome shotgun sequence View Alignment
☐	CH473310	2.0e-78	295	854813	320 84	320 84	292	668	1	249979	250355	Trypanosoma cruzi strain CL Brener 1047053516851 genomic scaffold, whole genome shotgun sequence View Alignment
☐	XM_816524	2.0e-78	295	2289	320 84	320 84	292	668	1	1583	1959	Trypanosoma cruzi strain CL Brener ATP-dependent RNA helicase (Tc00.1047053508461.250) mRNA, complete cds View Alignment

FIG. 5.12 Revisão de Similaridade

Na figura os resultados para o exemplo são mostrados parcialmente. Porém todas as informações das seqüências similares e a qual banco de dados se referem são

apresentadas. Logo, o anotador pode selecionar mais de um resultado sobre esta análise de similaridade para a confirmação do termo.

E) DISCUSSÃO ENTRE PESQUISADORES

A atividade discussão entre pesquisadores permite registrar todas as interações/colaborações existentes entre diversos pesquisadores no estudo da anotação baseada em ontologia. A FIG. 5.13 apresenta um exemplo.

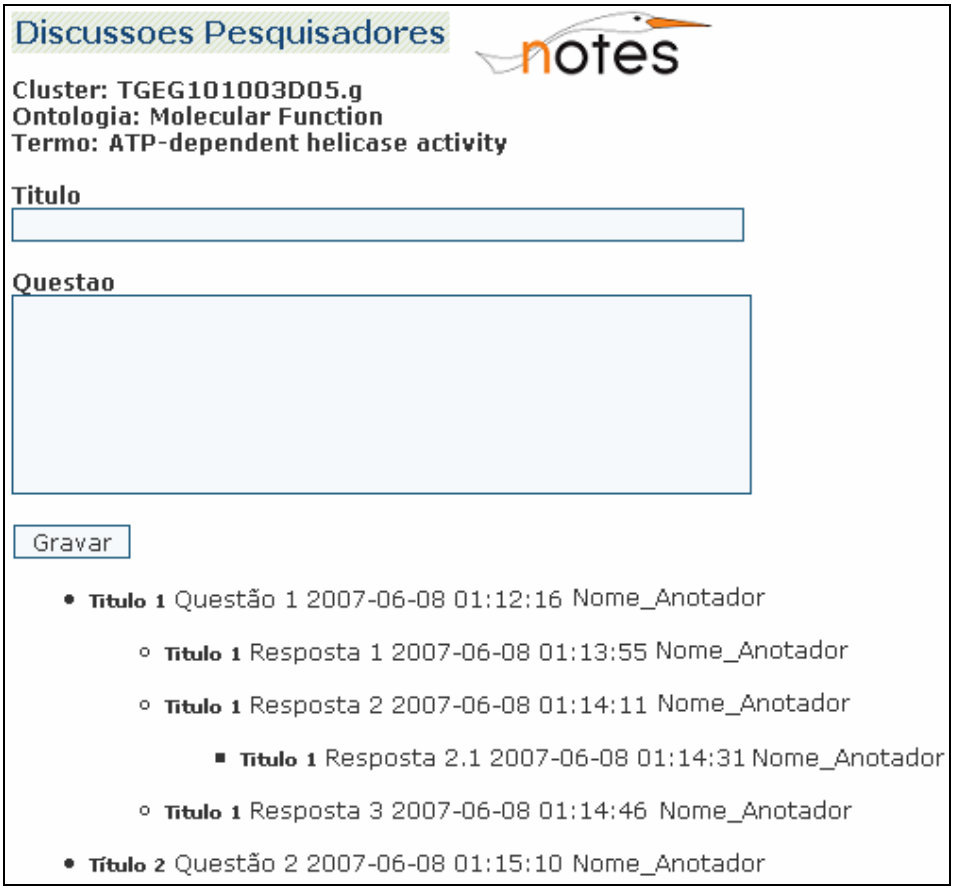


FIG. 5.13 Discussão entre pesquisadores

Esta atividade é baseada na concepção de um fórum. Uma questão principal é cadastrada sendo informada um Título (assunto) e o texto referente à questão. Logo, os diversos anotadores que colaboram entre si podem responder a esta questão, ou a respostas já postadas. A funcionalidade armazena todas as questões e respostas juntamente com a data em que foram postadas e o nome do anotador que está postando.

F) OUTROS

A opção outros consta somente de um campo livre para serem registrados outros itens a serem referenciados, por não se enquadrarem em alguma das atividades sugeridas pelo GARSA Notes, como pode ser visto na FIG. 5.14.

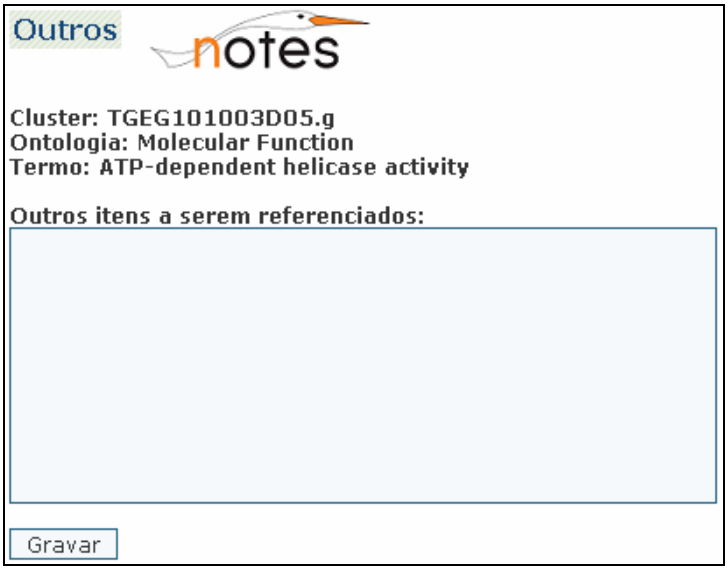


FIG. 5.14 Referência a outros itens

5.2.2 TERMO NÃO ENCONTRADO

Quando não há termo (*none*) para alguma das três ontologias, ao acessar a opção *term notes*, leva-se ao módulo para os registros considerando então a opção de termo não encontrado, como pode ser visto na FIG. 5.15.

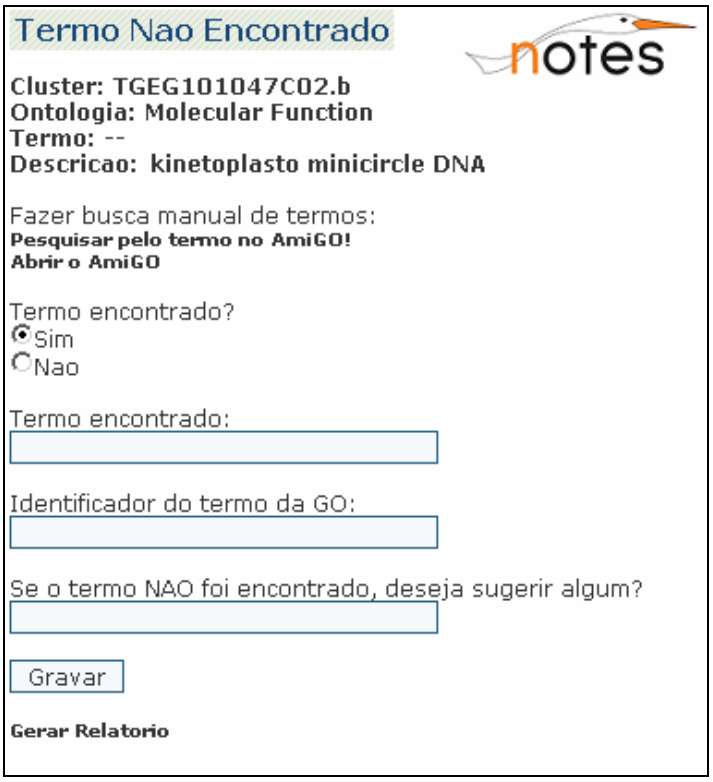


FIG. 5.15 Funcionalidades para termo não encontrado

Uma funcionalidade deste módulo, acessada através da opção *pesquisar pelo termo no AmiGO* é a possibilidade de fazer uma busca manual através de navegação no banco de termos da ontologia de modo a verificar se o termo correspondente existe. Essa navegação é feita baseando-se na anotação semi-automatizada. Dessa forma, além das informações sobre o cluster e a ontologia, a descrição da anotação semi-automatizada é também inserida. O AmiGO, já comentado no capítulo 2, é o *Web-site* oficial para vizualização dos termos do banco de dados GO. Outra possibilidade, *abrir o AmiGO*, é também oferecida caso o anotador queira fazer uma busca através de outras palavras-chaves.

Caso o termo tenha sido encontrado, é feita uma busca no banco de termos local ao sistema a fim de verificar se este existe. Caso o termo exista no banco local, é provável que tenha ocorrido uma falha no algoritmo usado para busca automática por termos (atividade “fazer busca automática por termos da ontologia” do processo de anotação, descrita no capítulo 3). Então, é feita uma atualização nas tabelas que guardam a referência do termo sobre a ontologia em questão e registrada a informação sobre o termo e seu identificador na ontologia. Se com a navegação no AmiGO o termo foi encontrado, mas este não consta no banco de termos local do sistema, então uma mensagem solicitando a atualização deste banco é apresentada ao usuário.

Se com a navegação também não foi encontrado algum termo, essa informação é também registrada, pois pode evidenciar desatualização no banco de dados da ontologia ou ineditismo da descoberta (por exemplo, um gene inédito contido na seqüência).

5.2.3 GERAÇÃO DE RELATÓRIOS

A geração de relatórios pode ser realizada da seguinte maneira: referente a termos, encontrados, referente a termos não encontrados e relatórios gerais.

O relatório referente aos termos encontrados é acessado diretamente do módulo de termos encontrados, logo é criado de forma exclusiva para o termo em questão, numa determinada ontologia e para somente um cluster. A FIG. 5.16 mostra as opções para a geração deste relatório.

O relatório informa o cluster, o termo, a ontologia e oferece opções relativas às atividades para a análise do termo, e também a opção sobre a classificação. O anotador pode optar por gerar o relatório contendo todas as informações, selecionando todos os itens, ou fazer a seleção de somente o item que deseja. Um exemplo de relatório gerado é mostrado na FIG. 5.17.

O relatório apresenta as informações de acordo com as opções selecionadas pelo anotador: Consulta Pesquisador, Busca na Literatura, Pesquisa a Bancos de Dados Específicos e a Classificação do Termo. Para este exemplo, o item Busca na Literatura não possuía registros, e então uma mensagem passando essa informação foi apresentada. Para os demais itens, o relatório mostra quais as informações que foram registradas e também o nome do anotador que efetuou tais análises do termo.

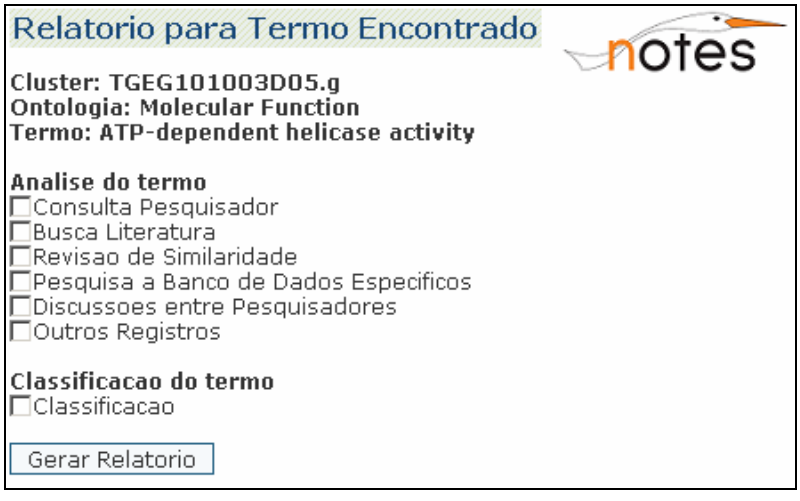


FIG. 5.16 Opções para geração de relatório para termo encontrado

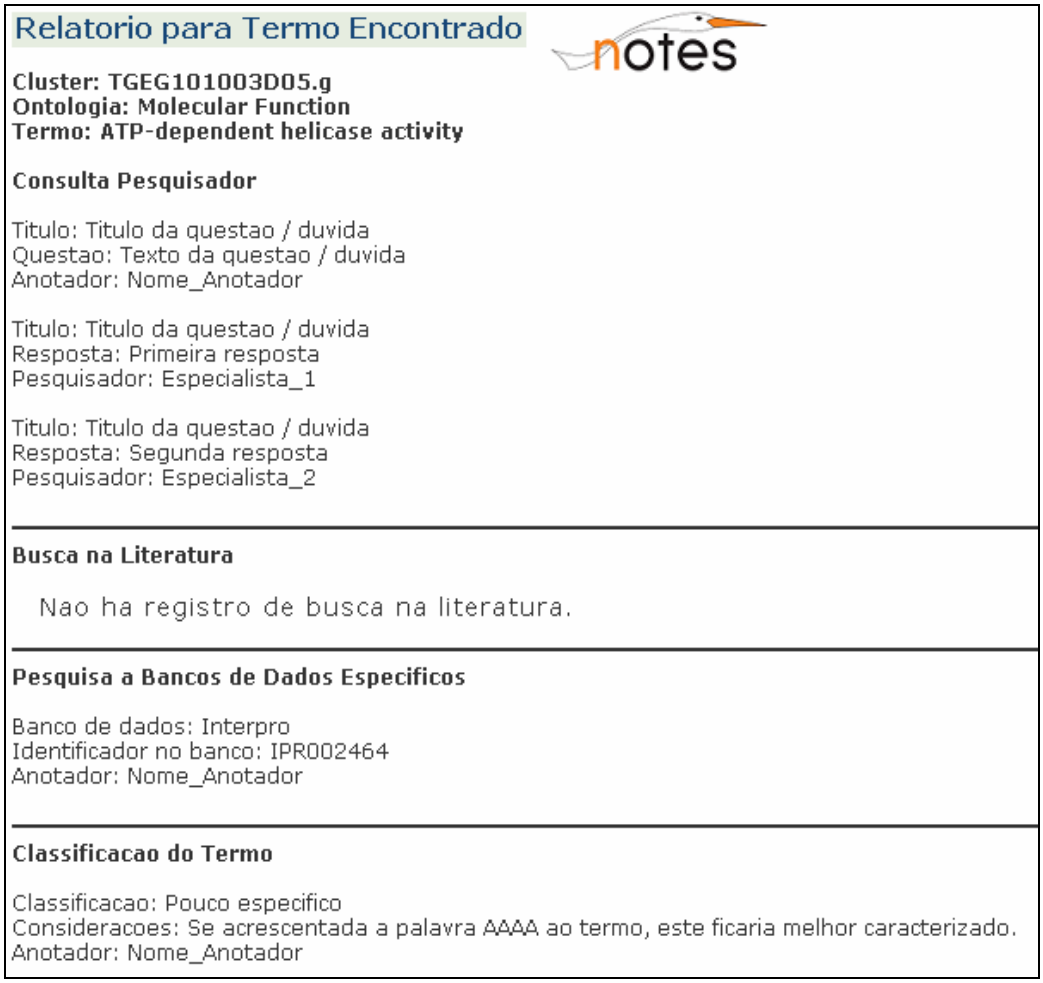


FIG. 5.17 Relatório gerado para termo encontrado

No módulo de termo não encontrado também é possível gerar um relatório com as informações que foram cadastradas. Semelhante ao relatório de termos encontrados, este também pode ser gerado diretamente do módulo de termo não encontrado, e apresenta as opções como mostradas na FIG. 5.18:

FIG. 5.18 Opções para geração de relatório para termo não encontrado

O relatório apresenta as informações sobre o cluster, ontologia e a descrição obtida através da anotação semi-automatizada. O anotador pode optar por quais informações deseja que sejam apresentadas no relatório. A FIG. 5.19 apresenta um exemplo de relatório gerado.

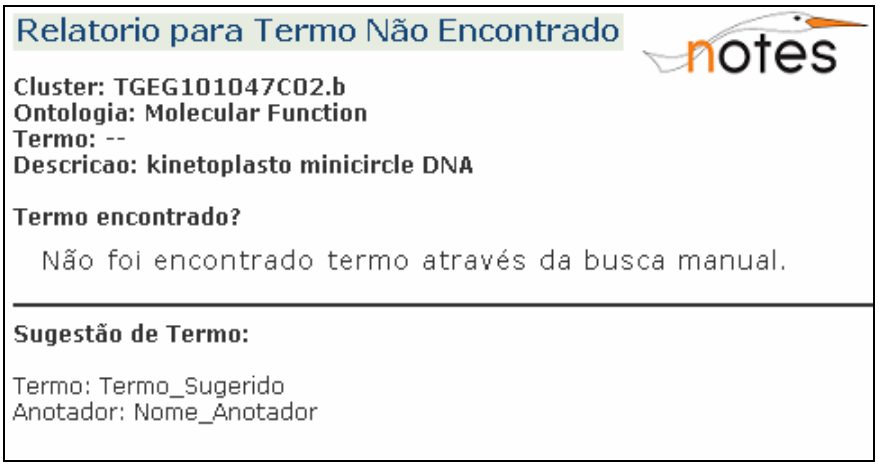


FIG. 5.19 Relatório gerado para termo não encontrado

Para o caso do relatório da FIG. 5.19, não foi encontrado termo a partir da busca manual no banco de termos da ontologia. Logo, o relatório mostra esta informação uma vez que não possui resultados da busca manual para serem apresentados. Mostra também a informação da sugestão de um termo para representação na ontologia em questão.

O GARSA Notes ainda oferece a opção de relatórios gerais como visto na FIG. 5.20. Esta opção é acessada diretamente no menu principal do sistema GARSA, item *GARSA Notes Report*.

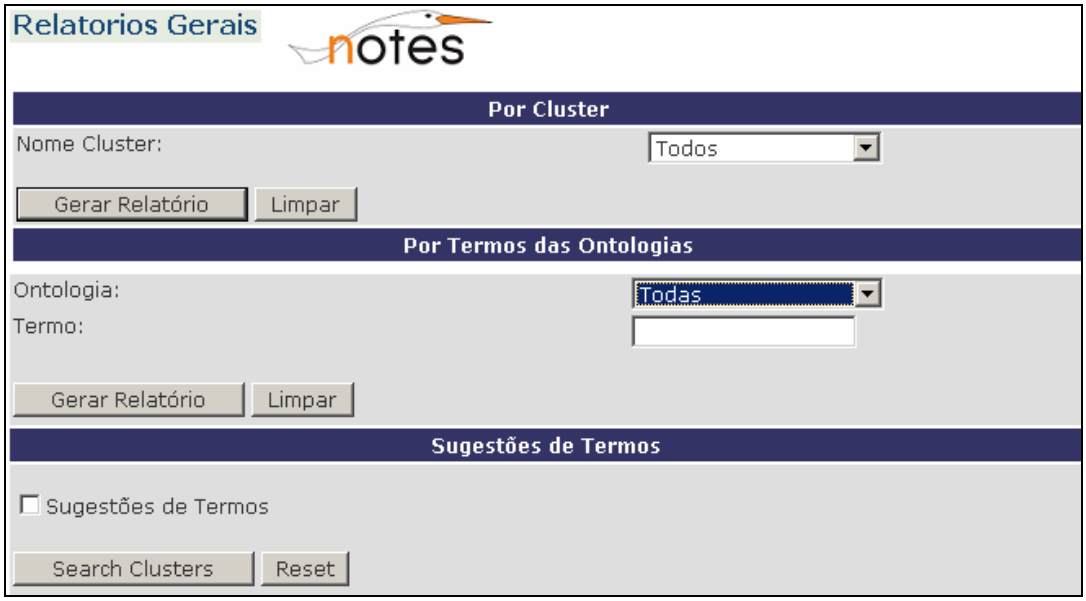


FIG. 5.20 Opção de relatórios gerais

Existem três opções para a geração deste relatório: por cluster, por termos das ontologias e as sugestões de termos. A geração por cluster oferece a possibilidade de gerar o relatório para todos os clusters ou através da escolha de somente um. São mostradas todas as informações registradas sobre os clusters, sejam informações para termo encontrado ou para termo não encontrado.

A geração por termos das ontologias, permite escolher os termos de todas as ontologias ou relativo a uma das três ontologias: função molecular, processo biológico e componente celular. De acordo com a escolha feita, os termos são exibidos juntamente com as informações registradas sobre ele, como a classificação e as considerações feitas.

A ultima opção, sugestões de termos, exibe uma lista de todas as sugestões de termos que foram registradas. Cada sugestão possui a identificação do cluster, ontologia e descrição (anotação semi-automática) a qual está relacionada.

5.3 MODELO DE DADOS

O SGBD utilizado foi o MySQL, escolhido por causa do GARSA utilizá-lo, e também pelo fato de ser eficiente e gratuito. Como comentado anteriormente, também para não alterar o esquema do banco de dados do GARSA, o banco GARSAN foi criado exclusivamente para o protótipo.

Servem ao sistema GARSA sete bancos de dados distintos (WAGNER, 2006):

- Seqonsql: informações sobre usuários do sistema;
- UniVec: seqüências de vetores;
- Contaminant: seqüências consideradas contaminantes e seqüências mitocondriais de alguns organismos;
- EC_Database: números e descrições de cada número EC (*EC number*) para classificação de enzimas;
- Taxonomy: nomes científicos de algumas espécies;
- Gene_Ontology: dados do *Gene Ontology Consortium*;
- "Projetos": cada projeto a ser pesquisado possui um nome distinto e armazena toda as informações de seqüências, resultados de análise e anotação.

Dos bancos citados anteriormente, para a implementação do GARSA Notes foi preciso fazer acesso a informações dos seguintes bancos de dados: Seqonsql, Gene_Ontology e "Projetos".

- Seqonsql: utilizado de modo a identificar o anotador do projeto sobre o qual o registro da anotação baseada em ontologia será feito, bem como os especialistas que participarão da atividade de análise do termo Consulta Pesquisador e os anotadores que irão colaborar na atividade Discussão entre Pesquisadores;
- Gene_Ontology: utilizado para armazenar termos que foram encontrados através da busca manual no banco de dados da ontologia.
- "Projetos": utilizado para obter as informações sobre as seqüências, como os resultados das anotações semi-automatizada e baseada em ontologia.

A FIG. 5.21 mostra o modelo de dados do GARSA Notes e uma descrição sobre o mesmo é feita a seguir. Cada registro feito sobre a anotação baseada em ontologia, se refere ao termo encontrado ou a informação de que não foi encontrado, relativos a uma das três ontologias da e sobre um determinando cluster de um determinado projeto. Desta forma, a entidade *Identificação* guarda essas informações obtidas a partir do banco de dados "Projetos", e também qual o pesquisador que está fazendo o registro em questão, obtido a partir do banco de dados Seqonsql do GARSA. Esta entidade é referenciada (chave estrangeira *id_identificação*) por quase todas as demais entidades do modelo.

A entidade *Questão* representa o registro de análise do termo referente à atividade Consulta Pesquisador. Todas as perguntas e respostas são registradas nesta mesma entidade. Estas são diferenciadas pelo atributo *tipo*, que armazena 'Q' (*question*) para as perguntas e 'A' (*answer*) para as respostas e pelo atributo *referencia_questao* que faz a

referência correta das respostas a sua respectiva pergunta (o auto-relacionamento na entidade demonstra esta referência).

A entidade *DiscussaoPesquisador* registra a atividade de Discussão entre Pesquisadores, baseadas na concepção de um fórum. Desta forma o atributo *referencia_discussao_pesquisador* guarda a referência de uma resposta. Então, pode-se ter resposta de uma questão ou resposta de uma outra resposta já cadastrada. O atributo *data_criacao* registra a data e a hora em que as perguntas ou as respostas foram postadas.

A entidade *BuscaLiteratura* registra a atividade de Busca na Literatura para análise do termo. Possui como chave estrangeira o atributo *id_cadastro_referencia*, vindo da entidade *CadastroReferencia*, que aponta para o site onde realizou a busca. Os atributos *referencia_link* e *referencia_literatura* armazenam o sitio de busca onde se encontrou referencias na literatura e a publicação relevante encontrada na busca. A entidade *CadastroReferencia* armazena os sitios de busca por referências mais utilizados pelos pesquisadores, como por exemplo o PubMed.

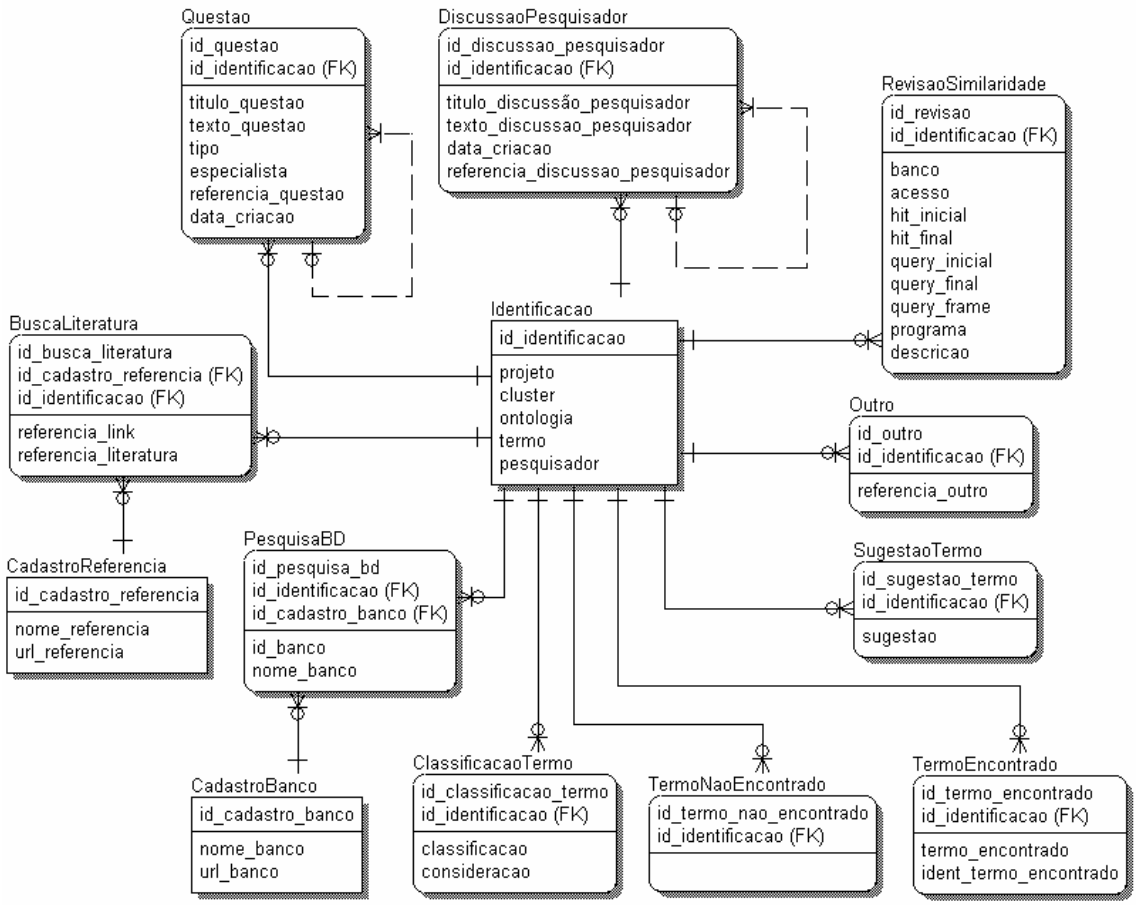


FIG. 5.1 Modelo de dados do GARSA Notes

Semelhante a *BuscaLiteratura*¸ a entidade *PesquisaBD* que representa o registro da análise do termo referente à atividade de Pesquisas em Bancos de Dados Específicos possui como chave estrangeira o atributo *id_cadastro_banco*, relativo a entidade *CadastroBanco* que tem por objetivo armazenar os bancos de dados mais utilizados. Os atributos *id_banco* e *nome_banco* registram o item encontrado na busca.

As entidades *RevisaoSimilaridade* e *Outro* respectivamente registram a análise do termo referente a atividade Revisão de Similaridade e a informação sobre um outro tipo de análise distinta das previstas pelo GARSA Notes.

As entidades *ClassificacaoTermo* e *SugestãoTermo* representam, respectivamente, o registro da classificação do termo feito ao final das atividades de análise e o registro da sugestão de termo, caso este não seja encontrado por buscas automática e manual.

A entidade *TermoEncontrado* registra o termo e seu identificador na ontologia quando é encontrado um termo através de busca manual. Já a entidade *TermoNaoEncontrado* irá armazenar a referência (*Identificação*) para a qual não houve sucesso na busca manual.

5.4 AMBIENTE DE DESENVOLVIMENTO

Com o objetivo de desenvolver um protótipo funcional sobre o estudo da anotação baseada em ontologia a partir do sistema GARSA, o ambiente de desenvolvimento ideal seria o mesmo adotado pelo sistema em questão.

Desta forma, a linguagem de programação utilizada foi a Perl e a interface gráfica do protótipo foi desenvolvida em CGI através de *scripts*, para visualização em navegadores para *Web*. Foi utilizado um servidor da unidade da Fiocruz que é desenvolvedora do GARSA para armazenamento do protótipo. Recursos já instalados como o servidor Apache e o SGBD MySql foram também usados para o desenvolvimento do protótipo. A máquina servidora, sobre a qual roda o sistema GARSA e o GARSA Notes, possui sistema operacional Linux (Fedora Core 2).

No desenvolvimento do protótipo, houve a preocupação de alterar minimamente os *scripts* do sistema GARSA. Isto foi feito para evitar modificações desnecessárias nos seus *scripts*, e na sua interface, bem como para reduzir os riscos de inclusão de erros no sistema. Para tanto, a intenção foi produzir um protótipo que fosse o menos dependente possível do GARSA, contendo seus próprios *scripts* e banco de dados. A interface, por

sua vez, segue o estilo utilizado pelo GARSA, de modo a causar pouco impacto na navegação do usuário.

5.5 CONSIDERAÇÕES REFERENTES AO GARSA E GARSA NOTES

Como mencionado anteriormente, o desenvolvimento do GARSA Notes foi realizado de forma que seus módulos não dependessem muito dos módulos do sistema GARSA. Contudo, o GARSA Notes faz uso das seguintes informações vindas do sistema GARSA: nome projeto, nome cluster, ontologia, termo da ontologia, descrição da anotação semi-automatizada e usuários.

As alterações que envolvem o sistema GARSA referem-se a:

- Inserção da opção *GARSA Notes Report* no menu principal;
- Atualização de registro nas tabelas *GO_HIT* e *CDS* do banco de dados de "Projetos". Esta inserção ocorre na situação em que foi feita uma busca manual pelo termo representante da seqüência no banco de dados da ontologia e este foi encontrado. Se o termo em questão consta na tabela *TERM* do banco de dados Gene_Ontology, houve uma falha na busca automática deste termo. Logo, se ele existe, pode representar a seqüência. Assim, é feita uma atualização das tabelas *GO_HIT* e *CDS* para guardarem o termo representante.

Para atender ao GARSA Notes de forma como este foi desenvolvido é sugerida a seguinte mudança:

- No banco de dados Seqonsql, de informações dos usuários do sistema, atualizar a tabela *USER*, de forma a inserir o campo *especialista*, o qual irá informar se o usuário cadastrado é um pesquisador especialista, pois assim sendo, o nome deste usuário aparecerá na lista de pesquisadores especialistas a serem consultados na atividade Consulta Pesquisador de análise do termo.

Além das alterações sugeridas no GARSA, algumas melhorias no GARSA Notes são também apontadas para seu melhor funcionamento como:

- Armazenar a data em que foram feitos os registros para todas as atividades;
- Associar atividades que demandam colaboração, como: Consulta Pesquisador ou Discussões entre pesquisador, a tecnologias como *blogs* e *wikis*;
- Envio de mensagem ao administrador do sistema quando a monitoração da anotação identificar que o banco de termos local da ontologia está desatualizado;

- Opção de relatórios que cubram grupos de termos da ontologia, mostrando a sua posição no grafo acíclico direcionado da GO;
- Padronização de relatórios e nomeação de *scripts*, em conformidade com o sistema GARSA;
- Adotar a língua utilizada pelo sistema GARSA, no caso a língua inglesa;
- Anexar ao GARSA Notes um sistema de notificação de (re) anotação como proposto por ALMEIDA (2006). Pois quando um termo é encontrado através da busca manual e este está localizado no banco de termos contido no sistema, então automaticamente este termo entra para a anotação da seqüência em questão. Assim, uma vez que a proposta do sistema é também ampliar a colaboração, os grupos de pesquisa parceiros não ficariam desinformados da mudança.

Com o uso do GARSA Notes espera-se que o registro do raciocínio feito pelos pesquisadores na monitoração da anotação baseada em ontologia contribua para o entendimento destes termos e assim para anotações futuras melhores, e facilite a comunicação com os mantenedores da GO. Algumas melhorias sobre o sistema são apontadas na conclusão deste trabalho.

6 OBSERVAÇÃO SOBRE O USO DO GARSA NOTES

O capítulo anterior apresentou diversas funcionalidades do GARSA Notes tendo como base o projeto do estudo do organismo *Trypanosoma rangeli* armazenado no sistema. As anotações sobre este estudo, ajudaram a propor a expansão do processo de anotação genômica, pois num processo reverso de descoberta da anotação, verificou-se que o anotador responsável realizou diversas atividades sugeridas neste trabalho para o estudo da anotação baseada em ontologia.

Contudo, de forma a mostrar a utilização do protótipo em funcionamento com outro projeto armazenado no sistema GARSA, o exemplo do uso do GARSA Notes apresentado neste capítulo se refere ao estudo do organismo *Phytomonas serpens (P. serpens)* (COSTA, 2006) desenvolvido no IOC/FIOCRUZ. A anotadora responsável realizou a monitoração da anotação baseada em ontologia para algumas de suas seqüências (visualizadas através dos clusters), a fim de viabilizar a observação do uso do GARSA Notes. Cabe ressaltar que a anotadora reviu as anotações para esta monitoração, uma vez que estas já foram realizadas.

O objetivo desta observação foi verificar o uso do protótipo na prática com exemplos reais, melhorias a serem feitas, bem como colher as impressões de um anotador quanto ao seu uso, uma vez que são atividades inovadoras dentro do processo de anotação. São mostrados no capítulo exemplos englobando os casos de termos encontrados e não encontrados

6.1 EXEMPLO – PROJETO *P.serpens*

As funcionalidades dos módulos do GARSA Notes estão descritas no capítulo 5 deste trabalho. Portanto, as informações cadastradas pela anotadora são apresentadas a partir dos relatórios gerados para cada cluster.

Primeiramente a anotadora conectou-se no sistema GARSA, fazendo a escolha do projeto *P. serpens*. Após visualizar a lista de seqüências representadas através de clusters, ela fez a escolha de alguns clusters para este estudo, considerando a questão de ontologias com termo encontrado e sem termo. Contudo, para este trabalho foi restringida a apresentação a dois exemplos.

6.1.1 EXEMPLO 1

O primeiro exemplo de cluster a ser feito para o estudo da anotação baseada em ontologia pode ser visto na FIG. 6.1, a qual representa a tela de edição das anotações do sistema GARSA e através da qual o GARSA Notes é acessado. As anotações referentes são para o cluster PSADSAU001A12.g. As ontologias Função Molecular e Processo Biológico encontraram termo. Já a ontologia Componente Celular não obteve sucesso na busca automática. Desta forma, para os termos das duas primeiras ontologias, a opção *term notes* abriu o módulo de termo encontrado e para a terceira ontologia, a opção abriu o módulo de termo não encontrado.

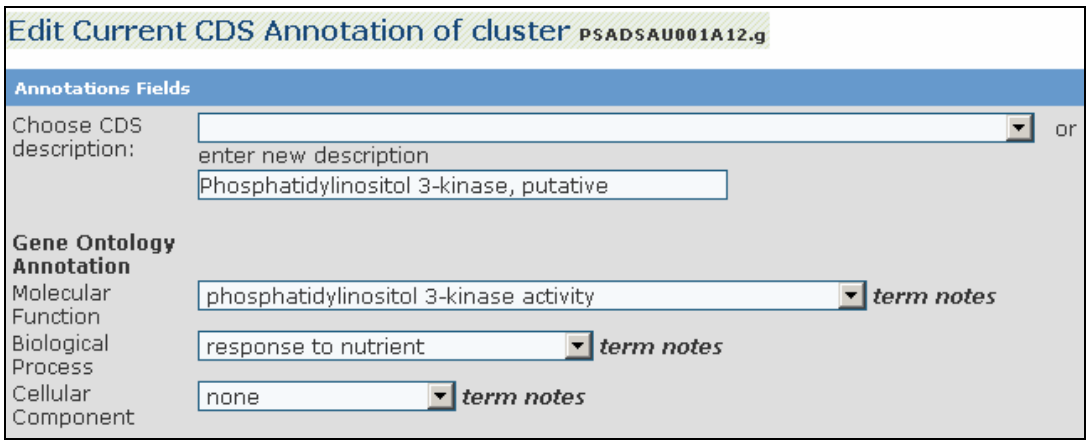


FIG. 6.1 Edição da anotação do cluster PSADSAU001A12.g (GARSA,2007)

A FIG. 6.2 apresenta o relatório referente a análise do termo para a ontologia Função Molecular (*Molecular Function*).

Para o termo **phosphatidylinositol 3-kinase activity** apenas a atividade Revisão de Similaridade foi suficiente para fazer a classificação do termo. O relatório mostra as informações sobre as seqüências mais similares, como as suas descrições (*Description*), e também mostra qual o banco de dados que possui estas seqüências (*Database*). O relatório apresenta também qual foi a classificação dada ao termo e possíveis considerações feitas a respeito.

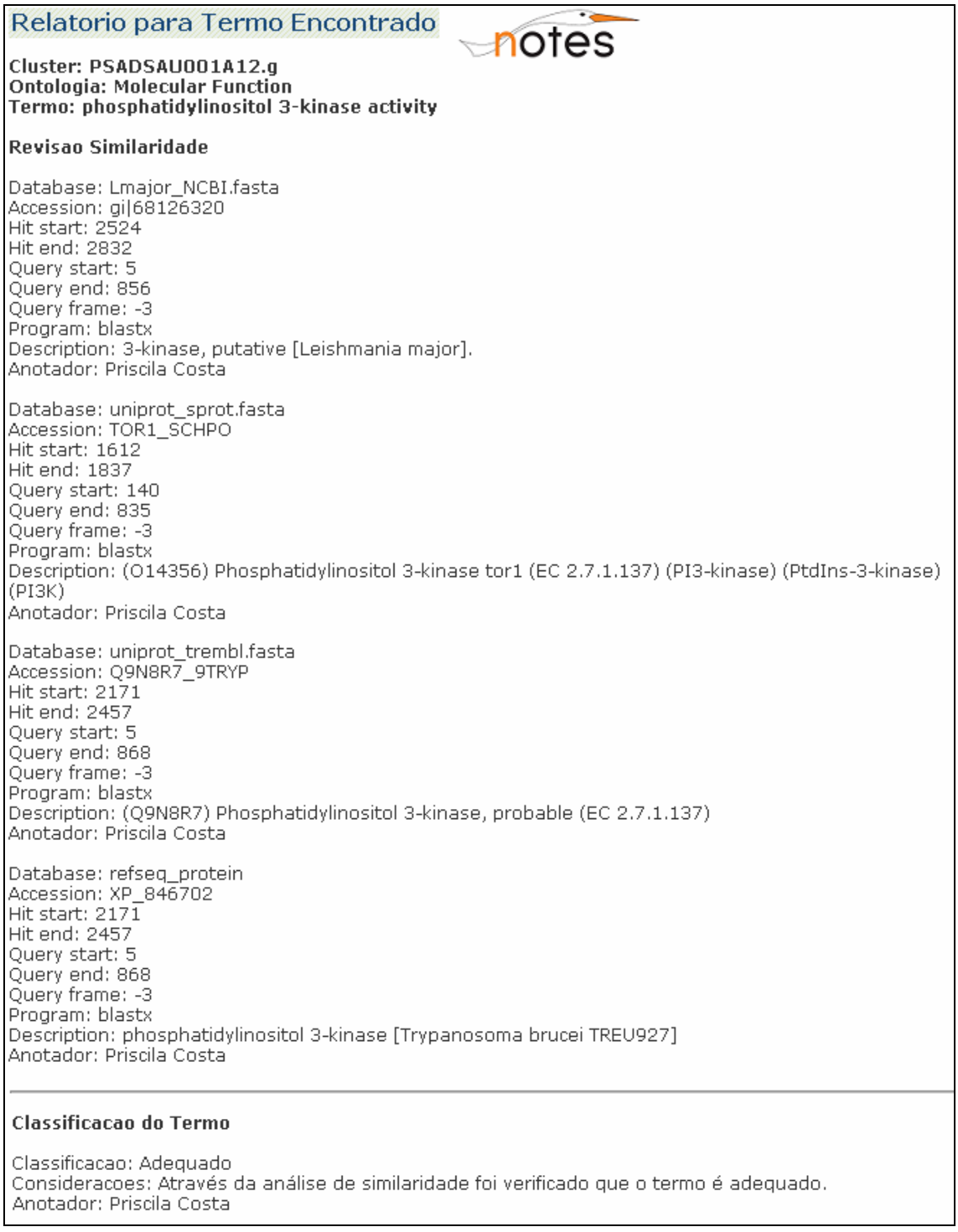


FIG. 6.2 Revisão de similaridade para o termo 'phosphatidylinositol 3-kinase activity'

Para o termo **response to nutrient** da ontologia Processo Biológico (*Biological Process*) foi realizada uma pesquisa em bancos de dados específicos e assim feita sua classificação. A FIG. 6.3 apresenta os resultados.

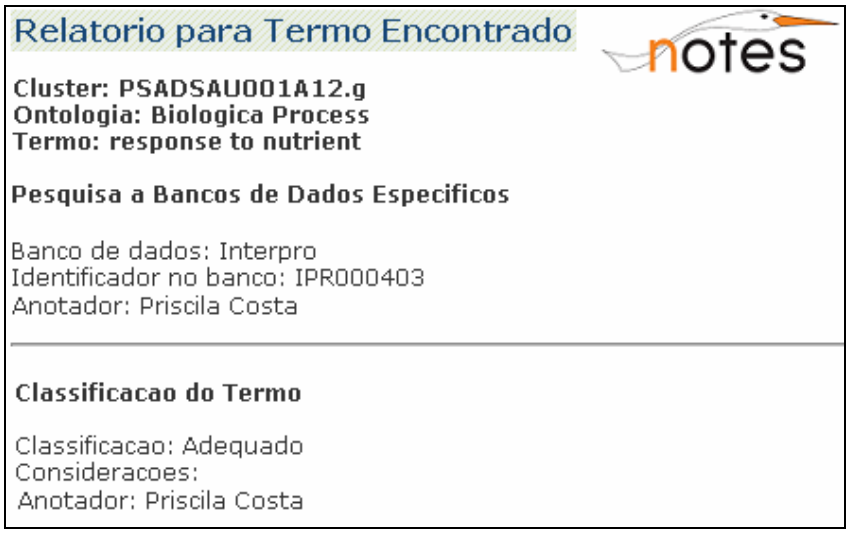


FIG. 6.3 Pesquisa a bancos de dados para o termo 'response to nutrient'

O relatório informa que o banco de dados Interpro possui um registro, o qual permite classificar o termo **response to nutrient.** O campo Identificador guarda qual é este registro no banco. A classificação foi feita como sendo um termo adequado e nenhuma consideração foi adicionada.

Para a ontologia Componente Celular (*Cellular Component*) não foi encontrado termo. Logo uma busca no banco de termos da ontologia através do AmiGO foi realizada, utilizando a descrição da anotação semi-automatizada, **Phosphatidylinositol 3-kinase, putative**. Contudo, nenhum termo foi encontrado e também não houve sugestão para o termo. A FIG. 6.4 apresenta o relatório com as informações.

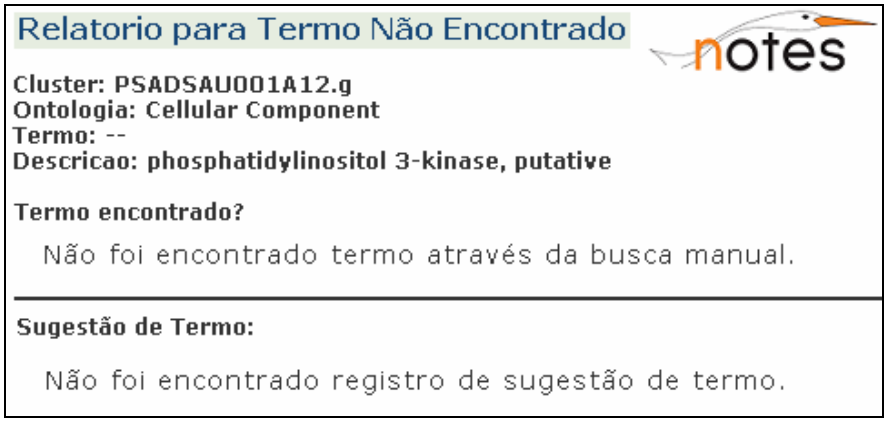


FIG. 6.4 Informações sobre o termo não encontrado

6.1.2 EXEMPLO 2

A FIG. 6.5, de edição de anotações do sistema GARSA mostra o cluster e seus respectivos termos:

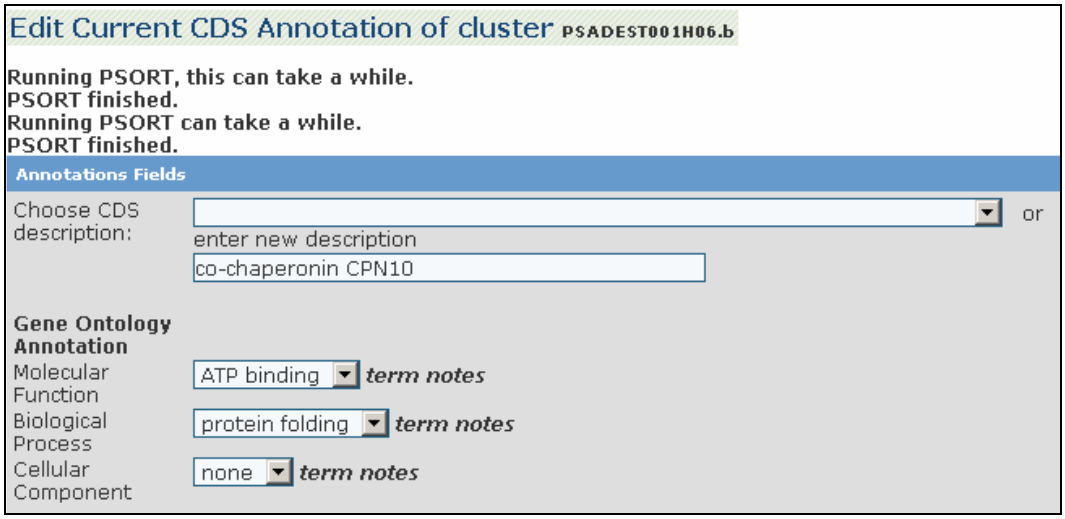


FIG. 6.5 Edição da anotação do cluster PSADEST001H06.b (GARSA,2007)

As anotações se referem ao cluster PSADEST001H06.b. De forma semelhante ao exemplo 1, para este cluster foram encontrados termos para as ontologias Função Molecular e Processo Biológico e não foi encontrado termo para ontologia Componente Celular.

O termo **ATP binding** da ontologia Função Molecular foi verificado através de uma Busca na Literatura como mostra a FIG. 6.6. O relatório mostra a referência de uma literatura a qual foi verificada se o termo corresponde à seqüência estudada. Para este caso, o temo foi classificado como pouco específico. Não houve registro de considerações.

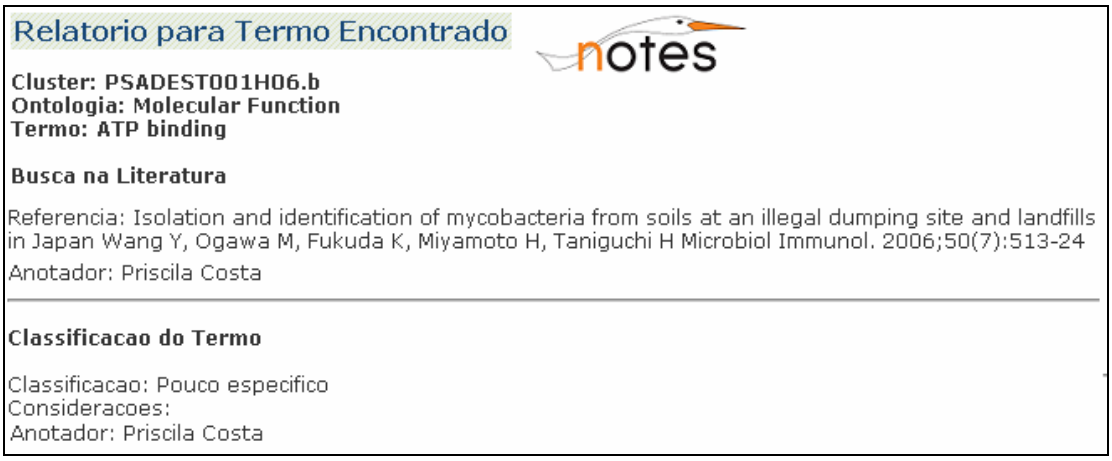


FIG. 6.6 Busca na literatura para o termo 'ATP binding'

O termo **protein folding** da ontologia Processo Biológico também foi confirmado através de uma Busca na Literatura, como ilustrado na FIG. 6.7. Este termo foi classificado como adequado e também não houve registro de considerações.

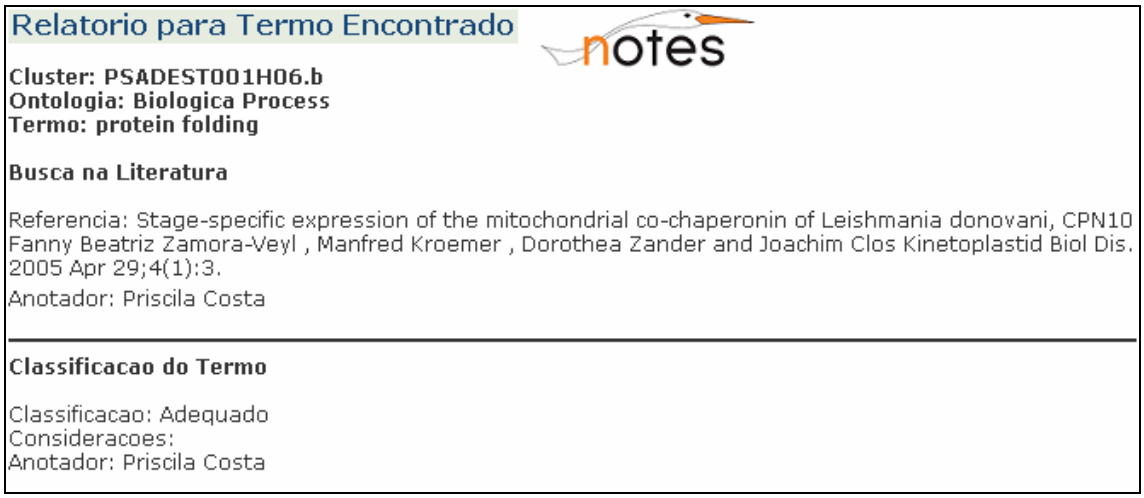


FIG. 6.7 Busca na literatura para o termo 'protein folding'

Para a ontologia Componente Celular não foi encontrado termo. Contudo, através da busca manual no AmiGO com base na descrição da anotação semi-automatizada, **co-chaperonin CPN10**, foi encontrado um termo correspondente e que pode ser utilizado. A FIG. 6.8 ilustra estes resultados.

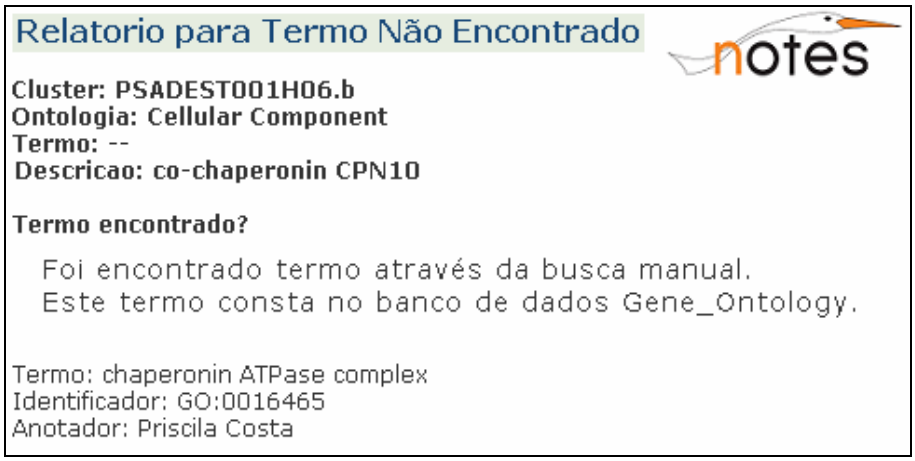


FIG. 6.8 Informações sobre o termo encontrado

O relatório mostra a informação de que o termo foi encontrado pela busca manual e que este consta no banco de dados de termos da ontologia que o sistema mantém. Assim, apresenta qual foi o termo e seu identificador na ontologia. Esta informação representa que houve uma falha no algoritmo de busca de termos do sistema.

6.2 CONSIDERAÇÕES

Os exemplos citados neste capítulo tiveram como objetivo ilustrar o funcionamento do GARSA Notes. A partir desses exemplos, foi possível validar, embora que de forma limitada, o protótipo quanto à sua utilização, colher as impressões de um anotador sobre o uso do mesmo e registrar a monitoração da anotação baseada em ontologia para um estudo genômico. Para que este seja completamente validado o ideal seria que experimentos controlados fossem planejados com a participação de diversos anotadores envolvidos em diferentes projetos genômicos fizessem uso do protótipo e observassem os resultados gerados. Outra validação significante seria um anotador que ainda se encontra em processo de análises e anotações das seqüências fazer o uso do GARSA Notes.

A anotadora, em seu estudo sobre o organismo *P.serpens* revisitou suas anotações para efetivar o registro sobre o uso dos termos da ontologia. Não havia registro das pesquisas que ela havia feito para a confirmação de sua anotação e desta forma as pesquisas tiveram de ser refeitas. Isto demonstra o quanto é importante o registro destas informações no momento em que ocorrem, ou seja, no momento em que se passa pelo processo de anotação genômica.

As atividades Consulta a um Pesquisador Experiente e Discussão entre Pesquisadores, por exemplo, apesar de serem importantes, não foram citadas neste

estudo. A anotadora comentou que fez diversas interações com seu orientador e com outros anotadores nos momentos de análises das seqüências e confirmação das anotações, principalmente as baseadas em ontologia. Porém, como não foram registradas as dúvidas e a solução a qual se chegou, no momento em que ocorreram, não houve como levantar este histórico, fato que também evidencia a importância do registro das informações e da existência de um canal disponível para as discussões.

Para os termos que não foram encontrados através da busca automática, nenhuma atividade foi efetuada. A anotadora achou importante a recomendação de se fazer uma verificação de navegação no site da ontologia, como a oferecida pelo GARSA Notes. O mesmo ela disse em relação à possibilidade de sugestão de um novo termo.

De forma geral, após uma explicação do protótipo, a anotadora conseguiu utilizar e navegar facilmente pelas funcionalidades do mesmo. Ela considerou relevantes as atividades de cada módulo e julgou ser uma importante ferramenta para os anotadores no sentido de que toda a pesquisa envolvida para afirmar a anotação baseada em ontologia estará registrada e com detalhes. Logo, sempre que for necessário rever tal anotação, todas as informações de como se chegou ao termo anotado já estarão disponibilizadas.

A partir de seus comentários, verificou-se que o registro sobre a anotação baseada em ontologia proposto no GARSA Notes, na primeira vez que for realizado por um usuário demanda um tempo maior, devido à pouca familiaridade do mesmo em relação à novidade da funcionalidade do protótipo. Porém, a medida que se faz uso do mesmo, e a partir da navegação sobre o que já foi registrado, a tendência é que se ganhe agilidade.

7 CONCLUSÃO

O processo de anotação genômica é parte fundamental do processo de estudos de seqüências genômicas, pois através dele as regiões codificantes das seqüências são identificadas, e assim conseguimos conhecer todas as características dos possíveis genes contidos nas seqüências.

As anotações em geral são feitas através de um vocabulário próprio utilizado pelo anotador ou pelo grupo de pesquisa, o que pode causar ambigüidades ou uma compreensão equivocada por parte de outros pesquisadores que visualizem estas anotações, assim como dificultar a colaboração para os trabalhos deste processo e interação com grupos de pesquisas parceiros e grupos de referências da comunidade de Bioinformática, como os grupos mantenedores de bancos de dados e ontologias.

Na tentativa de reverter esta situação e para se ter mais uma opção de anotação, diversos sistemas, como GARSA, SABIA e Artemis têm utilizado a anotação baseada em ontologia. Esta anotação garante um vocabulário controlado através dos termos específicos voltados para o domínio da biologia molecular. Contudo, o que foi percebido através do estudo exploratório do sistema GARSA é que esta anotação pode não estar completamente adequada, devido a termos que são encontrados para cada região codificante não representarem com exatidão o gene. Logo, com termos não adequados, fica também dificultada a colaboração e completa compreensão dos resultados por parte de outros pesquisadores.

Ao encontrar problemas com os termos da ontologia, fica a critério do anotador reportar estes erros ou não aos curadores das ontologias, constituindo assim uma tarefa dissociada do processo de anotação genômica. Então, quando esta informação é repassada, em geral não é feita no momento em que o problema ocorreu, o que pode causar a perda de informações importantes.

De modo a sugerir soluções de apoio à colaboração para ampliar o uso e a evolução de ontologias no domínio da biologia molecular, o processo de anotação genômica foi caracterizado sendo identificada cada atividade, os papéis participantes e a colaboração inerente a ele. Entendemos então que para facilitar e promover a colaboração, o uso da ontologia para anotar possui extrema relevância.

Dado que os grupos de pesquisa estudados não utilizam mecanismos que permitam a monitoração do uso dos termos da ontologia encontrados e o registro de todos os

problemas envolvendo seu uso, este trabalho propôs uma extensão do processo de anotação genômica para englobar os estudos sobre os termos das ontologias encontrados e sugestões de atividades que podem ser feitas para quando um termo não é encontrado. Os anotadores, realizando estes estudos podem aumentar seus conhecimentos sobre os termos das ontologias e desta forma tornar mais amplo o compartilhamento de conhecimento, bem como o estímulo à anotação baseada em ontologia.

Outro ponto a comentar se refere à colaboração para a evolução da própria ontologia no sentido em que todo o estudo realizado pode ser reproduzido através de um documento formal a ser remetido aos curadores das ontologias informando problemas, solicitando atualizações e passando sugestões. Desta forma, com as atualizações das ontologias e seu banco de termos, a tendência a cada anotação é que mais termos serão encontrados e estes por usa vez estarão adequados.

O protótipo GARSA Notes foi desenvolvido para apoiar o desenvolvimento deste estudo, em específico para o grupo de pesquisa da FIOCRUZ. Desta forma seu desenvolvimento baseou-se nas tecnologias adotadas pelo grupo para o desenvolvimento do sistema GARSA. Devido a limitações de tempo para conclusão deste trabalho e a falta de disponibilidade de anotadores que estivessem em estudos ainda para entrar na fase de anotação, não foi feito um estudo de caso sobre o protótipo a fim de melhor validá-lo. Apresentou-se desta forma, um exemplo (com dados reais) de uso do mesmo e as validações referentes à observação do uso do protótipo por um anotador.

Em linhas gerais, o protótipo apresentou-se como sendo relevante no apoio ao anotador para monitorar a anotação baseada em ontologia, registrando todas as ações realizadas pelo mesmo e desta forma não perdendo conteúdos importantes ao processo de anotação. Este conteúdo, a partir da geração de relatórios, apóia os anotadores intra e inter projetos na realização de suas anotações e no compartilhamento deste conhecimento, assim como também a colaboração extra-projeto, como ocorre com os curadores das ontologias.

Uma das limitações dessa proposta está relacionada com a adesão ou não por parte dos anotadores. É importante que eles compreendam a necessidade dessa monitoração para o benefício das próprias anotações.

Outra limitação refere-se à dificuldade de implantar o protótipo em outros sistemas, devido a este ter sido desenvolvido apoiando-se em um sistema específico. Sendo assim, há a necessidade de adaptações de acordo com a linguagem de programação

utilizada, tecnologias, padrões e outros. Contudo, essas adaptações tornam-se mais facilitadas, devido a existência do modelo de expansão do processo e da especificação através do protótipo de telas e da navegação entre elas.

Cabe ressaltar que um novo sistema, denominado STINGRAY (WAGNER et al., 2007), o qual é uma evolução do sistema GARSA, está na fase de término de seu desenvolvimento. O novo sistema incluirá as funcionalidades do protótipo desenvolvido neste trabalho, e esse será denominado de STINGRAY Notes (BELLOZE et. al, 2007).

7.1 CONTRIBUIÇÕES

A expansão do processo de anotação genômica para englobar os estudos a serem feitos sobre a anotação baseada em ontologia é a principal contribuição deste trabalho. Tendo estes estudos associados ao processo, todas as ações realizadas pelos anotadores em relação à anotação baseada em ontologia são registradas no momento em que ocorrem, e assim, evita-se a perda de importantes informações. O ciclo produtivo de anotação que se cria, ao fazer esses estudos, ao enviar informações relevantes aos curadores das ontologias, e ao obter termos atualizados das ontologias implicam na melhoria da qualidade da anotação.

Uma tendência que tem surgido nas organizações é a necessidade de evidenciar o desenvolvimento de seus processos, sendo detalhadas as atividades dos mesmos, bem como descrevendo cada papel participante e suas responsabilidades sobre cada atividade. Assim, outra contribuição refere-se à modelagem do processo do estudo de seqüências genômicas e de seus sub-processos de forma genérica, e também a identificação dos cenários de colaboração e a caracterização de novos papéis nestes cenários. Essa contribuição para a comunidade de Bioinformática acontece no sentido de mostrar que diversos grupos de pesquisas genômicas, desempenham atividades semelhantes, mesmo diferenciado-se no organismo estudado, e assim formas de colaboração entre os grupos podem ser sugeridas.

O protótipo GARSA Notes é uma contribuição no sentido de visualizar como o estudo da anotação baseada em ontologia pode ser feito. Devido ao estudo, espera-se que os anotadores possam inferir anotações de maior qualidade por conhecerem melhor os termos das ontologias e facilitar as anotações futuras com base no registro já existente. Para novos anotadores, o registro pode servir como base e espécie de um guia para seus estudos. Os registros também permitirão identificar quando o banco de dados de termos do sistema está desatualizado, bem como se houve alguma falha nos

algoritmos de busca automática de termos e ineditismos de genes. Seus relatórios facilitarão a passagem das informações sobre os termos e sugestões para os curadores das ontologias, contribuindo assim para a evolução das mesmas.

Por ser um protótipo funcional, outra contribuição se refere ao GARSA Notes estar disponível através do sistema GARSA para utilização do grupo de pesquisa do IOC/FIOCRUZ assim como para todos os pesquisadores que fazem parte do Consórcio BioWebDB.

Por fim, esta proposta evidencia que o compartilhamento de conhecimento e o trabalho colaborativo ficam facilitados através do uso de ontologias, devido ao vocabulário comum imposto por estas. Contudo, as ontologias podem apresentar problemas nos seus termos. Este fato ocorre para os diversos domínios, não somente para o de biologia molecular. Como exemplo, pode ser citado o uso das ontologias na Web Semântica, que surge como um apoio computacional importante para prover o conteúdo semântico exigido (FERNÁNDES et al., 2006). Na inteligência artificial, o uso de ontologias vem trazendo benefícios com sua utilização, como por exemplo, o compartilhamento e estruturação do conhecimento através de vocabulários controlados, descrição exata, entre outros (GUIMARÃES, 2002). Logo, os termos também podem apresentar problemas e assim não representarem corretamente o conteúdo semântico ou a descrição exata, como citados nos exemplos anteriormente.

Desta forma, a proposta do estudo dos termos, de acordo com a aplicação de cada ontologia, pode ser estendido para os outros domínios, de modo que a partir destes estudos, trabalhos colaborativos e compartilhamento de conhecimento tendem a serem melhorados, assim como também a evolução das ontologias.

7.2 MELHORIAS E TRABALHOS FUTUROS

7.2.1 MELHORIAS NA PROPOSTA

Como melhoria na proposta, sugere-se analisar o processo de estudo de seqüências genômicas obtido, e em especial o de anotação genômica, seguindo propostas que sistematizam a identificação de requisitos de apoio computacional à colaboração (MAGDALENO, 2006), (MIRANDA, ARAUJO E BORGES 2007), buscando ampliar ainda mais o apoio à colaboração oferecido pelo GARSA Notes.

Outra melhoria seria avaliar a colaboração existente sendo verificados os quatro princípios da área de trabalho colaborativo apoiado por computador: coordenação, comunicação, percepção e memória de grupo. Dessa maneira, novas atividades para a expansão do processo de forma a ampliar a colaboração existente podem ser sugeridas.

7.2.2 MELHORIAS NO GARSA NOTES

Para o GARSA Notes, foram citadas diversas melhorias no final do capítulo 5, como a padronização de acordo com o sistema GARSA, em itens como: a exibição de alguns relatórios e a língua utilizada. Outra melhoria é anexar ao GARSA Notes um sistema de notificação de (re) anotação.

Outra questão diz respeito a melhorias que podem ser feitas, mas somente após o uso do GARSA Notes pelos anotadores, pois desta forma serão identificadas, como: a melhor navegação entre as funcionalidades do protótipo e geração de outros relatórios que atendam outras necessidades dos anotadores e para envio aos curadores das ontologias.

7.2.3 TRABALHOS FUTUROS

Como trabalho futuro sugere-se a descrição de como é feito o processo de curagem dos bancos de dados secundários, pois devido à dificuldade de encontrar essa informação disponibilizada, a modelagem deste processo não foi contemplada neste trabalho.

Outro trabalho refere-se ao retorno dos curadores das ontologias sobre os problemas encontrados nos termos ou sugestões de novos termos e atualizações. Para isso é necessário o trabalho por parte de alguns anotadores no uso do GARSA Notes, para que se possa gerar relatórios e assim efetivar a colaboração com os curadores das ontologias.

Aponta-se também como trabalhos futuros, estudos de casos sobre o GARSA Notes, os quais devem englobar a utilização do protótipo por diferentes anotadores e envolvendo os cenários intra e inter-projetos. O estudo pode ser feito sobre projetos que têm as suas anotações concluídas e também por projetos que ainda irão entrar no processo de anotação genômica.

Outro trabalho seria generalizar a solução, ou seja, desenvolver um ou vários módulos para atender à proposta, tal que possa ser integrado(s) em outros sistemas e também em plataformas como o GUS.

Por fim, outro trabalho futuro é fazer a monitoração do uso de ontologias em outros contextos. Para que isso seja possível deve ser levado em consideração o domínio do trabalho, e a partir de um levantamento de requisitos para este novo contexto, as atividades devem ser repensadas e/ou adaptadas.

8 REFERÊNCIAS BIBLIOGRÁFICAS

AGÜERO, F., ZHENG, W., BRENT WEATHERLY, D. et al. **TcruziDB: an integrated, post-genomics community resource for Trypanosoma cruzi.** Nucleic Acids Research, Vol. 34, D428-D43, 2006

ALMEIDA, L. G. P., PAIXÃO, R., SOUZA, R. C. et al. **A System for Automated Bacterial (genome) Integrated Annotation – SABIA**. Bioinformatics. Nov 1; 20 (16):2832-3, 2004.

ALMEIDA, A. C. **BIOANOT: Um sistema multi-agentes para notificação de (re) anotações de seqüências em bancos de dados genômicos**. Dissertação de Mestrado, Instituto Militar de Engenharia, Rio de Janeiro, 2006.

ALTMAN R., BADA, M., CHAI, X. J. et al. **RIBOWEB: An ontology-based system for collaborative molecular biology**. IEEE Intelligent Systems, 14(5):68-76, 1999.

ALTSCHUL, S. F, GISH, W., MILLER, W. et al. **Basic local alignment search tool.** J. Mol. Biol. 1990; 215:403-410.

AMIGO. **AmiGO.** Disponível: http://amigo.geneontology.org/cgi-bin/amigo/go.cgi.

APOLLO USER GUIDE. **Apollo User Guide.** Disponível: http://www.fruitfly.org/annot/apollo/userguide.html#ReadingGenbank. [Consultado em: 06/10/2005].

ARTEMIS EXAMPLES. **Artemis Examples and Screenshots.** Disponível em: http://www.sanger.ac.uk/Software/Artemis/Examples. [Consultado em: 15/12/2005].

BAIROCH, A.., APWEILER, R. **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000**. Nucleic Acids Res. 28, 45–48, 2000.

BAKER, P. G., BRASS, A., BECHHOFER, S. et al. **TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources**. In Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB'98), pages 25-34, Menlow Park, California, June 28-July 1, 1998.

BARTON, G. J. **SCANPS version 2.3.9 user guide**, University of Dundee, UK, 2002.

BBOP. **Berkeley Bioinformatics and Ontologies Project**. Disponível: http://www.berkeleybop.org. [Consultado em: 11/06/2007].

BELLOZE, K. T., CAVALCANTI, M. C., ARAUJO, R. M., ALBRECHT, F. F., DÁVILA, A. M. R. **Capturing ontology-based annotation history.** International Workshop on Genomic Database. Angra dos Reis, Rio de Janeiro, 2007.

BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J. et al. **GenBank.** Nucleic Acids Res., D34–D38, 2005.

BERGERON, B. **Bioinformatics computing.** Prentice Hall PTR, 2002. Disponível em: http://www.ineedahosting.com/~john/Bioinf/Prentice%20Hall%20PTR%20%20-%20Bioinformatics%20Computing.pdf. Acesso em: 18/10/2005.

BDGP. **Berkeley Drosophila Genome Project** Disponível: http://www.fruitfly.org/about/pubs/index.html. [Consultado em: 10/06/2007]

BIONOTES. **BioNotes.** Disponível: http://139.82.24.24/BioNotes. [Consultado em: 06/12/2005].

BIOWEBDB. **BiowebDB.** Disponível: www.biowebdb.org. Consultado em: [11/06/2007].

BONFIELD, J. K., SMITH, K. F, STADEN, R. **A new DNA sequence assembly program.** Nucleic Acids Res. 24, 4992-4999, 1995.

BORODOVSKY, M., MCININCH, J. **Recognition of genes in DNA sequence with ambiguities.** Biosystems. 30:161-171, 1993.

BRIDGES, S. M, MCCARTHY, F. M., LUTHE, D.S. et al. **AgBase: Targeted gene ontology annotation databases for agriculture**, Fifth Annual Gene Ontology Users Meeting. Bergen, Norway. Sept 14 – 15, 2005.

BURGE, C. KARLIN, S. **Prediction of complete gene structures in human genomic DNA.** J. Mol. Biol. 268, 78-94, 1997.

BURGE, C. B. **Modeling dependencies in pre-mRNA splicing signals.** In Salzberg, S., Searls, D. and Kasif, S., eds. Computational Methods in Molecular Biology, Elsevier Science, Amsterdam, pp. 127-163, 1998

CAMON, E., MAGRANE, M., BARRELL, D. et al. **The Gene Ontology Annotation (GOA)**. **Database: sharing knowledge in Uniprot with Gene Ontology**. Nucleic Acids Research 32(1): D262-D266, 2004.

CARVER, T. J., RUTHERFORD, K. M., BERRIMAN, M. et al. **ACT: the Artemis Comparison Tool,** Bioinformatics, 2005.

CASTRO, E., CHRISTIAN, J. A, SIGRIST, A. G., et al. **ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins.** Nucleic Acids Res. 34. (Web Server issue):W362-W365, 2006.

CHADO. **Chado.** Disponível: http://www.gmod.org/schema/index.shtml. [Consultado em: 18/10/2005].

CHEN, N., HARRIS, T. W., ANTOSHECHKIN, I. et al. **WormBase: a comprehensive data resource for Caenorhabditis biology and genomics**. Nucleic Acids Res 33 DATABASE ISSUE:D383-D389, 2004.

COG. **Clusters of Orthologous Groups**. Disponível: http://www.ncbi.nlm.nih.gov/COG. [Consultado em: 19/04/2007].

COSTA, P. M. O. **Exploração do genoma de *Phytomonas serpens***. Dissertação (Mestrado em Biologia Parasitária) - Instituto Oswaldo Cruz – FIOCRUZ, Rio de Janeiro, 2006.

CROSBY, M. A., GOODMAN, J. L., STRELETS, V.B. et al. **FlyBase: genomes by the dozen.** Nucleic Acids Research 35: D486-D491, 2007.

DAVID, P. H., BLAKE J. A., RICHARDSON, J. E., RINGWALD, M. **Extension and integration of the Gene Ontology (GO): combining GO vocabularies with external vocabularies.** Genome Res.Vol. 12, Issue 12, 1982-1991, December 2002.

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

DDBJ/EMBL/GENBANK. **The DDBJ/EMBL/GenBank feature table: definition.** DNA Data Bank of Japan, Mishima, Japan; EMBL Nucleotide Sequence Database, Cambridge, UK; GenBank, NCBI, Bethesda, MD, USA, 2005. Disponível: http://www.ddbj.nig.ac.jp/fromddbj-e.html. [Consultado em: 28/11/2005].

DELCHER, A.L., HARMON, D. et al. **Improved microbial gene identification with Glimmer**. Nucleic Acids Res. 27(23): 4636 – 4641, 1999.

DO, C. B., MAHABHASHYAM M. S. P., BRUDNO, M., BATZOGLOU, S. **ProbCons: Probabilistic consistencybased multiple sequence alignment.** Genome Res.; 15: 330-340, 2005.

DOWELL, R. D, JOKERST, R. M., DAY, A. et al. **The distributed annotation system.** BMC Bioinformatics, 2:7, 2001

DURBIN, R., HAUSSLER, D. **GFF (General Feature Format) Specifications Document.** Disponível: http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml. [Consultado em: 15/10/2005].

DURINCK, S., MOREAU, Y., KASPRZYK, A. et al. **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.** Bioinformatics. Aug 15;21(16):3439-40, 2005

EBI. **European Bioinformatics Institute**. Disponível: http://www.ebi.ac.uk. [Consultado em: 19/04/2007].

EDDY, S. **HMMER - profile hidden Markov models for biological sequence analysis Version 2.3.2.** Howard Hughes Medical Institute and Dept. of Genetics Washington University School of Medicine, St. Louis, USA, 2003.

EDGAR, R.C. **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** BMC Bioinformatics, 5:113, 2004

EILBECK, K., LEWIS, S., MUNGALL, C.J., YANDELL, M. et al. **The Sequence Ontology: A tool for the unification of genome annotations.** Genome Biology, 6:R44, 2005.

EL-SAYED, N. M., MYLER, P. J., BARTHOLOMEU, D. C. et al. **The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease**. Science, 309, 409–415, 2005

EPSRC. **Engineering and Physical Sciences Research Council**. Disponível: http://www.epsrc.ac.uk/default.htm. [Consultado em: 19/04/2007].

EWING, B., HILLIER, L., WENDL, M. C., GREEN, P. **Base-Calling of Automated Sequencer Traces Using Phred. I**. Accuracy Assessment. Genome Res.; 8:175–185, 1998a.

EWING, B., GREEN, P. **Base-calling of automated sequencer traces using Phred. II. ErrorProbabilities.** Genome Res, 8:186–194, 1998b.

FELSENSTEIN, J. **PHYLIP (Phylogenetic Inference Package), version 3.65.** http://evolution.genetics.washington.edu/phylip.html. Departament of Genetics, University of Washington, Seatle, USA, 2005.

FERNÁNDEZ, M., CANTADOR, I., CASTELLS, P. **CORE: A tool for collaborative ontology reuse and evaluation.** In: 4th International EON Workshop, Evaluation of Ontologies for the Web - EON 2006, Edinburgh, United Kingdom, 2006.

FINN, R. D., MISTRY, J, SCHUSTER-BÖCKLER, B. et al. **Pfam: clans, web tools and services**. Nucleic Acids Research Database Issue 34:D247-D251, 2006.

GALPERIN M. Y. **The Molecular Biology Database Collection: 2005 update National Center for Biotechnology Information.** National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA , 2005.

GALTIER, N., GOUY, M., GAUTIER, C. **SeaView and Phylo_win, two graphic tools for sequence alignment and molecular phylogeny.** Comput. Applic. Biosci., 12, 543-548, 1996.

GARSA. **Genomic Analysis Resources for Sequence Annotation**. Disponível: http://www.biowebdb.org/garsa. [Consultado em: 11/06/2007].

GENBANK**. GenBank**: Disponível: http://www.ncbi.nih.gov/Genbank. [Consultado em: 19/12/2005].

GENOPAR. **Genoma do Paraná.** Disponível: http://www.genopar.org. [Consultado em: 18/10/2005.

GMOD. **Generic Model Organism Database Toolkit.** Disponível: http://www.gmod.org. [Consultado em: 19/12/2005].

GO. **The Gene Ontology.** Disponível: http://www.geneontology.org. [Consultado em: 11/06/2007].

GO SOURCE FORGE. **Gene Ontology**. Disponível: https://sourceforge.net/tracker/?func=add&group_id=36855&atid=440764. [Consultado em: 04/08/2006].

GRUBER, T. R. **Toward principles for the design of ontologies used for knowledge sharing.** International Journal of Human-Computer Studies, v. 43, p. 907-928, 1995.

GUIMARÃES, F. J. Z. **Utilização de ontologias no domínio B2C**. Dissertação de Mestrado em Informática, Pontifícia Universidade Católica, Rio de Janeiro, 2002.

GUS. **The Genomics Unified Schema.** Disponível: http://www.gusdb.org. [Consultado em 18/10/2005].

HERTZ-FOWLER, C., PEACOCK, C. S., WOOD, V. et al. **GeneDB: a resource for prokaryotic and eukaryotic organisms**. Nucleic Acids Research, Vol. 32, Database issue D339-D343, 2004

HUANG, X, MADAN, A. **CAP3: A DNA sequence assembly program**. Genome Res.;9:868–877, 1999.

HUBBARD, T. J. P., AKEN, B. L., BEAL K. et al. **Ensembl 2007**. Nucleic Acids Res.; Database issue, Jan 2007.

INCYTE. **Incyte.** Disponível : http://www.incyte.com/about_press_releases_20010801-2.html. [Consultado em: 07/05/2007].

INF PUC-RIO. **Departamento de Informática da Pontifica Universidade Católica do Rio de Janeiro.** Disponível: http://www.inf.puc-rio.br. [Consultado em: 19/12/2005].

IOC/FIOCRUZ. **Instituto Oswaldo Cruz / Fundação Oswaldo Cruz.** Disponível: http://www.ioc.fiocruz.br. [Consultado em: 13/06/2007].

JONES, C. E et al. **Automated methods of predicting the function of biological sequences using GO and BLAST.** Bioinformatics, 6:272, 2005.

JOSLYN, C.A et al. **The gene ontology categorizer.** Bioinformatics, vol. 20, suppl. 1, pp. i169-i177, 2004.

KAROLINSKA. **Karolinska Institutet**. Disponível: http://cruzi.cgb.ki.se/Databases.php. [Consultado em: 19/04/2007].

KEGG. **KEGG**: Kyoto Encyclopedia of Genes and Genomes. Disponível em: http://www.genome.ad.jp/kegg [Consultado em: 13/06/07].

KORF, I., FLICEK, P., DUAN, D., BRENT, M. R. **Integrating genomic homology into gene structure prediction.** Bioinformatics Vol. 17 no. 90001, Pages S140-S148 2001.

KULIKOVA T., AKHTAR R., ALDEBERT P. et al. **EMBL Nucleotide Sequence Database in 2006**. Nucleic Acids Research 35: D16-D20, 2007.

KUMAR, S., TAMURA, K., NEI, M. **MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment.** Brief Bioinform. 5(2):150 – 163, 2004.

LABINFO. **Laboratório de Bioinformática.** Disponível: http://www.lncc.br/~labinfo/labinfo/labinfo_port/800x600/index.htm. [Consultado em: 12/12/2005].

LASKOWSKI, R A, MACARTHURM, M. W., SMITH, D. K. et al. **PROCHECK v.3.0 - Program to check the stereochemistry quality of protein structures.** Operating instructions, 1994.

LEMOS, M., SEIBEL, L. F. B., CASANOVA, M. A. **Sistemas de anotações em biossequências.** Disponível: www.inf.puc-rio.br/~melissa/publicação/download/mcc_melissa/MCC04-03.pdf. [Consultado em: 31/10/2005]

LEMOS, M., SEIBEL, L. F. B., CASANOVA, M. A. **BioNotes: a system for biosequence annotation,** dexa, p. 16, 14th International Workshop on Database and Expert Systems Applications (DEXA'03), 2003.

LEMOS, M., **Workflow para bioinformática**, Tese de Doutorado, Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, RJ, Brasil, 2005

LEWIS, S. E., SEARLE, S. M. J., HARRIS, N. et al. **Apollo: a sequence annotation editor** Genome Biology, 3(12):research0082, 2002.

LOPEZ, R., SILVENTOINEN, V., ROBINSON, S., KIBRIA, A., GISH, W. **WU-Blast2 server at the European Bioinformatics Institute.** Nucleic Acids Res.; 31(13): 3795 – 3798., 2003.

LOWE, T. M., EDDY, S. R. **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** Nucleic Acids Res.; 35(5): 955 – 964, 1997.

LUCHTAN, M., WARADE, C., WEATHERLY, D. B. et al. **TcruziDB: an integrated Trypanosoma cruzi genome resource.** Nucleic Acids Res. Jan 1;32(Database issue):D344-D346, 2004

MAGDALENO, A. M. **Um modelo de maturidade volaborativo para BPM.** Dissertação de Mestrado em Informática, Universidade Federal do Rio de Janeiro, RJ, Brasil, 2006.

MARCHLER-BAUER, A., ANDERSON, J. B., CHERUKURI, P. F. **CDD: a Conserved Domain Database for protein classification**. Nucleic Acids Research 33 D192-D196, 2005.

MICHALICKOVA, K., BADER, G. D., DUMONTIER, M., LIEU, H., BETEL, D., ISSERLIN, R., HOGUE, C. W. **Seqhound: biological sequence and structure database as a platform for bioinformatics research.** BMC Bioinformatics.3(1):32, 2002

MIRANDA, I. S., ARAUJO, R. M., BORGES, M. R. S. **Discovering Group Communication Requirements.** 10th Workshop Iberoamericano de Ingenieria de Requisitos y Ambientes de Software (aceito para publicação), Venezuela, 2007.

MISO. **miSO sequence ontology term browser**. Disponível em: http://www.sequenceontology.org/miSO/index.html. [Consultado em: 19/04/2007].

MULDER, N. J., APWEILER, R., ATTWOOD, T. K. et al. **InterPro, progress and status in 2005.** Nucleic Acids Res. 33: D201–D205, 2005.

NCBI. **National Center for Biotechnology Information.** Disponível: htp://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html. [Consultado em: 15/05/2005].

NCBO. **The National Center for Biomedical Ontology.** Disponível em: http://bioontology.org. [Consultado em: 11/06/2007].

NIG. **National Institute of Genetics (NIG).** Disponível em: http://www.nig.ac.jp/index-e.html. [Consultado em: 18/05/2007].

O'Reilly XML: GAME DTD (Genome Annotation Markup Elements), 2000. Disponível em: http://www.xml.com/pub/r/946. Acesso em: 27/04/2006.

ORG. **Ontology Research Group**. Disponível em: http://org.buffalo.edu. [Consultado em: 11/06/2007].

KARP, P. D., RILEY, M., SAIER, M. et al. **The EcoCyc and MetaCyc databases.** Nucleic Acids Research, 28:56-59, 2000.

PAMGO. **Plant-Associated Microbe Gene Ontology (PAMGO)**. Interest Group. Disponível: http://pamgo.vbi.vt.edu. [Consultado em: 10/01/2007].

PETTIFER S., WOLSTENCROFT, K., ALPER, P., et al. **myGrid and UTOPIA: an Integrated Approach To Enacting And Visualising In Silico Experiments** in the Life Sciences Lecture Notes in Bioinformatics, 06/2007, 2007.

PFAM. Pfam. Disponível: http://www.sanger.ac.uk/Software/Pfam. [Consultado em: 18/10/2005] a.

PFAM. **Pfam.** Disponível em: ftp://selab.janelia.org/pub/Pfam/userman.txt. [Consultado em: 05/01/2007] b.

PHRAP. **Laboratory of PHIL GREEN.** Disponível: http://www.phrap.org. [Consultado em: 13/06/07].

PRUITT, K. D., TATUSOVA, T., MAGLOTT, D. R. **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. Nucleic Acids Research 33 D501-D504, 2005.

RÖSSLE, S. **Desenvolvimento de um Sistema Computacional para Modelagem Comparativa em Genômica Estrutural: Análise de Seqüências do Genoma da Gluconacetobacter Diazotrophicus**. Tese de Doutorado, IBCCF/UFRJ, Brasil, 2004.

RUTHERFORD, K., PARKHILL, J., CROOK, J. et al. **Artemis: sequence visualisation and annotation**. Bioinformatics 16 (10) 944-945, 2000.

SALI, A. **MODELLER: A Program for Protein Structure Modeling Release 6.** Rockefeller University, 2001.

SANTOS, R. T, **O ambiente 10+c para a definição e execução de workflows in silico através de serviços web.** Dissertação de Mestrado em Informática, Universidade Federal do Rio de Janeiro, Brasil, 2004.

SBRI. **Seattle Biomedical Research Institute.** Disponível em: http://www.sbri.org/Home. [Consultado em: 19/07/2007].

SCHIEX, et al. FrameD**: a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences**, Nucl. Acids Res.. 31 (13): 3738, 2003.

SCHULZE-KREMER, S. **Ontologies for molecular biology.** In Proceedings of the Third Pacic Symposium on Biocomputing, pages 693-704. AAAI Press, 1998.

SENGER, M. **Gowlab: Web Pages as Web Services**. March 2005. Disponível em: http://www.ebi.ac.uk/soaplab/Gowlab.html.

SHARP, A.., MCDERMOTT, P. **Workflow Modeling: Tools for Process Improvement and Application Development**. Boston: Artech House, 2001.

SMITH, B., KÖHLER, J., KUMAR, A. **On the application of formal principles to life science data: a case study in the Gene Ontology**. DILS 2004: Data Integration in the Life Sciences, *124-139, 2004.*

SMITH, B., WILLIAMS J., SCHULZE-KREMER, S., **The Ontology of the Gene Ontology**. Published in Proceedings of AMIA Symposium, 2003.

SNOWDON, R. **Overview of Process Modelling. Informatics Process Group.** Manchester University, UK, 2006.

STEVENS, R., GOBLE, C. A., BECHOFER, S. **Ontology-based knowledge representation for bioinformatics.** Brief Bioinform ;1(4):398-414, 2000.

SUTTON, G., WHITE, O., ADAMS, M., KERLAVAGE, A. **TIGR Assembler: A new tool for assembling large shotgun sequencing projects.** Genome Science & Technology 1:9-19, 1995.

TATENO, Y., FUKAMI-KOBAYASHI, K., MIYAZAKI, S. **DNA Data Bank of Japan at work on genome sequence data.** Nucleic Acids Res. 26, 1, 16-20, 1998.

TATUSOV R. L., GALPERIN M. Y., NATALE, D. A., KOONIN, E. V. **The COG database: a tool for genome-scale analysis of protein functions an evolution.** Nucleic Acids Res.; 28(1): 33–36, 2000.

TECH, M., MERKL, R. **YACOP: Enhanced gene prediction obtained by a combination of existing methods.** In Sil. Biol.; 3:441-51, 2003.

THE GENE ONTOLOGY CONSORTIUM. **Gene Ontology: tool for the unification of Biology.** Nature Genetics, 25:25-29, 2000.

THE GENE ONTOLOGY CONSORTIUM. **The Gene Ontology (GO) project in 2006.** Nucleic Acids Res.; 34: D322-D326, 2006.

THE SANGER INSTITUTE. **The Sanger Institute**. Disponível: http://www.sanger.ac.uk. [Consultado em: 19/12/2005].

THOMPSON, J. D., HIGGINS, D. G., GIBSON, T.J. **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** Nucleic Acids Res.; 22(22): 4673 – 4680, 1994.

TIGR **TIGR Database.** Disponível em: http://www.tigr.org/tdb/e2k1/tca1/intro.shtml. [Consultado em: 19/04/2007]

VASCONCELOS, S. S. **Uma investigação: ESTs (Expressed Sequence Tags) podem ser usados no desenvolvimento de marcadores moleculares baseados em introns?** Dissertação de Mestrado Universidade Católica de Brasília, 2003.

VERSPOOR, K., et al. **A categorization approach to automated ontological function annotation**. Proteoin Sci., 15: 1544-1549, 2006.

WAGNER, G. **Geração e análise comparativa de seqüências genômicas de *Trypanosoma rangeli*.** Dissertação de Mestrado, Instituto Oswaldo Cruz, Rio de Janeiro, 2006.

WAGNER, G., SORIANO, K., JUCÁ, H.C.L., BELLOZE, K.T., el al. **STINGRAY: System for Integrated Genomic Resources and Analysis.** In preparation.

WERNERSSON, R., PEDERSEN A. M. RevTrans**: multiple alignment of coding DNA from aligned amino acid sequences.** Nucleic Acids Research, Vol. 31, No. 13 3537-3539, 2003.

WESTHEAD, D. R., PARISH, J. H., TWYMAN, R. M. **Bioinformatics.** InstantNotes. Bios Scientific Publishers Limited, 2002.

WILKINSON, M. D., LINKS, M. **BioMOBY: an open-source biological web services proposal.** Briefings In Bioinformatics 3:4. 331-341, 2002.

XML COVER PAGES. **Genome Annotation Markup Elements (GAME).** 2002. Disponível: http://xml.coverpages.org/game.html. [Consultado em: 27/04/2006].

YUSIM, J. J., SZINGER, I. H., CALEF, C. et al. **Enhanced motif scan: A tool to scan for HLA anchor residues in proteins.** HIV Immunology and HIV/SIV Vaccine Databases, pp. 25–36. Los Alamos National Laboratory, Theoretical Biology & Biophysics, Los Alamos, New Mexico. LA-UR 04-8162, 2003

9 ANEXOS

9.1 ANEXO 1: RECURSOS COMPUTACIONAIS DE APOIO AO PROCESSO DO ESTUDO DE SEQÜÊNCIAS GENÔMICAS

Os programas referenciados para cada processo não são os únicos que realizam as funções em questão. Diversos outros existem na comunidade de bioinformática, porém, para exemplificar, somente alguns serão citados, os quais são em geral, os mais utilizados.

Pré-análise: neste processo, diversos programas com finalidades diferentes apóiam a realização das atividades. Para a limpeza de contaminação de vetores e avaliação da qualidade da seqüência, são utilizados os programas Crossmatch (EWING et al., 1998a) e PHRED (EWING et al., 1998b) respectivamente, e para o agrupamento/clusterização das seqüências, é utilizado o programa CAP3 (HUANG et al., 1999). Outros programas para montagem de seqüências ainda incluem o Gap4 (Bonfield et al, 1995), e TIGR Assembler (SUTTON et al., 1995). Com a execução destes programas de acordo com cada atividade do processo de pré-análise, o conjunto de seqüências não redundantes ou seqüências alvo é montado e então levado para o processo seguinte.

Análise da Seqüência: neste processo outros programas são executados, com o objetivo de identificar características presentes nas seqüências, de acordo com:

- Predição de genes: utilizados para encontrar genes codificantes de proteínas, como FrameD (SCHIEX et al, 2003), Genscan (BURGE et al, 1997), GeneMark (BORODOVSKY et al., 1993), Glimmer (DELCHER et al, 1999), Twinscan (KORF et al, 2001), YACOP (TECH et a., 2003);
- Buscas por similaridade: fazem buscas por similaridade com seqüências já estudadas e armazenadas em bancos de dados públicos. Os principais programas são da família BLAST. Outros como WU-BLAST (LOPEZ et al, 2003) e Scanps (BARTON, 2002) são também utilizados. Programas como o Interpro (MULDER et al, 2005), ScanProsite (CASTRO et al, 2006), MotifScan (YUSIM et al, 2003) que fazem busca de domínios e famílias, e HMMER (EDDY, 1998) que busca homologia entre *clusters* possuem funções semelhantes e são também utilizados. Uma outra especialização da busca por similaridade é a busca de domínio conservado. Para esta modalidade é utilizado um dos programas da família BLAST, o RPSBlast (ALTSCHUL et al, 1997) que faz a busca em bancos como: CDD (*Conserved Domain Database*) (MARCHLER-BAUER et al, 2001), Pfam (FINN et al, 2006), COG e KOG (COG, 2007). Os domínios conservados

designam motivos geralmente funcionais que se repetem numa mesma proteína ou em proteínas diferentes;

- Alinhamentos múltiplos: após todas as buscas por similaridade serem feitas, os *clusters* são alinhados utilizando os programas ClustalW (THOMPSON et al, 1994), Muscle (EDGAR et al, 2004), Probcons (DO et al, 2005), RevTrans (WERNERSSON & PEDERSEN, 2003) e outros;
- Análise filogenética: o programa mais comum utilizado para esta finalidade é o Phylip (FELSENSTEIN, 2005), contudo existem outros como MEGA (KUMAR et al, 2003) e Phylo_win (GALTIER et al, 1996). A análise filogenética depende dos alinhamentos múltiplos para ser realizada.

Em especial, os programas que buscam similaridades entre seqüências são os que geram resultados mais significativos. Portanto, diversas buscas por similaridades contra seqüências de diferentes bancos de dados são realizadas.

Os programas para a predição de genes podem ser executados em paralelo com os programas de busca por similaridade, pois um não depende do resultado do outro. Já para a execução dos programas para alinhamentos múltiplos e posteriormente os de análise filogenética, são necessárias que as execuções das análises de similaridade já tenham ocorrido, pois os alinhamentos múltiplos dependem de tais resultados.

Após a execução dos programas de análise, as seqüências alvo estão analisadas de acordo com as características presentes em cada uma. Essas características devem ser, portanto identificadas e anotadas.

O processo de anotação é o que mais exige trabalho por parte do anotador e curador, pois cabe a eles interpretar os resultados dos programas de análise e desta forma fazer as anotações de cada seqüência. Neste processo, as ontologias são recursos computacionais bastante importantes, pois seus termos representativos para as seqüências irão complementar os resultados anotados.

Para o processo de consolidação da seqüência, os recursos computacionais de apoio utilizados são as disponibilizados pelos centros de pesquisas mantenedores dos bancos de dados públicos para a submissão de seqüências. Em geral, a submissão é feita através de uma interface na *Web*, onde podem ser enviados os arquivos que contém as informações.

GLOSSÁRIO
DE TERMOS TÉCNICOS E EXPRESSÕES USADAS

- ANOTAÇÃO BASEADA EM ONTOLOGIA. Obtida através da busca automática por termos de ontologias.

- ANOTAÇÃO SEMI-AUTOMATIZADA. Obtida através da execução dos programas de análise nas seqüências. Os anotadores observam os resultados e fazem a anotação correspondente.

- BLAST. Programa que busca similaridade entre as seqüências.

- BLASTN. Modalidade do programa BLAST que compara seqüências de nucleotídeos contra um banco de dados de nucleotídeo.

BLASTX**.** Modalidade do programa BLAST que compara seqüências de nucleotídeos contra um banco de dados de proteínas.

- CEPA. Conjunto de indivíduos de uma espécie existente em uma colônia ou cultivo.

- CLUSTER. Segmentos de seqüências genômicas sobrepostas.

- CONTIGS E READS. Para ser feito o sequenciamento, o DNA é subdividido em várias partes denominadas *contigs,* que por usa vez são também subdivididos em parte menores chamadas *reads.* Estas subdivisões são feitas para que o estudo da seqüência possa ser realizado em partes.

- FRAMEPLOT. Programa que analisa presença de prováveis regiões codificadoras de proteínas.

- GENE. Cada molécula de proteína constitui um gene Os genes guardam as informações para a síntese de proteínas do nosso corpo que vão desde simples substâncias reagentes a hormônios reguladores de complexos sistemas vitais. O conjunto de genes de um indivíduo recebe o nome de genoma.

- GENES ORTÓLOGOS: genes provindos de um gene ancestral presente em uma espécie ancestral das espécies e não necessariamente possuem a mesma função.

- GENES PARÁLOGOS: genes encontrados numa mesma espécie originados a partir de um evento de duplicação e não possuem a mesma função.

- GENSCAN. Programa utilizado para a identificação e o mascaramento de regiões repetitivas freqüentemente encontradas em genomas.

- PROCESSOS DE TRANSCRIÇÃO E TRADUÇÃO. O processo de transcrição envolve a cópia da seqüência de uma das fitas do DNA na forma de um ácido

ribonucléico mensageiro (mRNA). Esta molécula, que é bastante similar ao DNA, possui também quatro tipos de bases nitrogenadas, porém a base (U) uracila substitui a (T) timina. No processo de tradução a informação no mRNA é traduzida com a ajuda de moléculas de RNA de transferência (tRNA), utilizando uma tabela de código genético para determinar a seqüência de aminoácidos. Na tradução, cada grupo de três nucleotídeos, ou códon, especifica um aminoácido em particular.

- PROTEÍNA. São moléculas grandes (macromoléculas) formadas por aminoácidos. Existem diversos tipos de proteínas entre elas as enzimas (que aceleram e regulam os processos vitais), os hormônios (que levam informações químicas que regulam os ciclos vitais), os anticorpos (que defendem o organismo contra substâncias invasoras), além de proteínas de reserva (albuminas), transporte (hemoglobina) e contração de músculos (actina e miosina).

- REGIÃO CODIFICANTE. Região da seqüência em que é encontrado um gene.

- SEQÜÊNCIA ALVO. Seqüência está sendo estudada no momento.

- SEQÜÊNCIA CONSENSO. Seqüência teórica de nucleotídeos ou de aminoácidos representativos em que cada unidade nucleotídeo ou aminoácido é a que ocorre com mais freqüência naquele sítio nas diferentes seqüências onde ocorrem na natureza.

- SEQÜÊNCIA GENÔMICA. Em geral é uma seqüência de DNA, que é formada por nucleotídeos ou bases, que podem ser de quatro tipos: (A) adenina, (C) citosina, (T) timina e (G) guanina. A seqüência linear destes nucleotídeos na molécula de DNA é a fonte básica da informação genética. Após passar por processos de transcrição e tradução, a seqüência passa a ser de aminoácidos.

- SEQÜÊNCIA HIT. A seqüência que obteve maior similaridade com a seqüência alvo.

- SIMILARIDADE. Presença de trechos idênticos entre duas ou mais seqüências, o que pode significar presença de características comuns entre elas.

- VETORES. São organismos onde é inserido o DNA cortado em pedaços, e estes são introduzidos em células hospedeiras onde o DNA pode ser produzido em quantidade.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Administração
Baixar livros de Agronomia
Baixar livros de Arquitetura
Baixar livros de Artes
Baixar livros de Astronomia
Baixar livros de Biologia Geral
Baixar livros de Ciência da Computação
Baixar livros de Ciência da Informação
Baixar livros de Ciência Política
Baixar livros de Ciências da Saúde
Baixar livros de Comunicação
Baixar livros do Conselho Nacional de Educação - CNE
Baixar livros de Defesa civil
Baixar livros de Direito
Baixar livros de Direitos humanos
Baixar livros de Economia
Baixar livros de Economia Doméstica
Baixar livros de Educação
Baixar livros de Educação - Trânsito
Baixar livros de Educação Física
Baixar livros de Engenharia Aeroespacial
Baixar livros de Farmácia
Baixar livros de Filosofia
Baixar livros de Física
Baixar livros de Geociências
Baixar livros de Geografia
Baixar livros de História
Baixar livros de Línguas

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo