Um Sistema Baseado em Software Livre para Anotação Automática de Genomas e Proteínas*

Ana L. C. Bazzan†, Bruno S. Fajardo, Leonardo V. Nascimento, Cássia T. dos Santos, Vitório F. Sassi

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul
Caixa Postal 15064 – 90501-970 Porto Alegre, RS

{bazzan,bsfajardo,lvnascimento,ctsantos,vfcsassi}@inf.ufrgs.br

Abstract. *The present article aims at reporting the experience achieved with the development of the ATUCG tool. The goal of this tool is to provide an annotation tool for the expert in genomics and proteomics. Such tool is open and based on free software.*

Resumo. *Este artigo visa reportar a experiência obtida com o desenvolvimento da ferramenta para anotação de genoma ATUCG. O objetivo deste ambiente é disponibilizar ao especialista em genômica e proteômica uma ferramenta para anotação que seja aberta e baseada em software livre.*

1. Introdução

A Bioinformática tem o propósito de ligar duas ciências que vêm crescendo de forma exponencial nos últimos 20 anos: a Biologia e a Ciência da Computação. O estudo dos genes e proteínas provê informações sobre o crescimento celular, comunicação e sua organização. O projeto genoma humano foi o principal responsável pela identificação da necessidade de ferramentas computacionais para auxiliar os pesquisadores na análise do material decodificado. É importante salientar que o genoma humano é formado de 3×10^9 pares de bases, mas apenas uma pequena parcela desses dados são os responsáveis pela codificação das características humanas.

No Brasil existem várias redes de seqüenciamento e análise de genoma trabalhando com vários organismos (a maioria patogênicos). Quase todos estes projetos têm uma característica fortemente multidisciplinar, incluindo especialistas das áreas de biotecnologia e informática. Este casamento não ocorre por acaso: dado o volume de dados produzidos por cada projeto genoma, torna-se absolutamente necessário gerenciar estes dados e, principalmente, automatizar processos de forma a libertar o especialista de tarefas repetitivas. Em geral, cada rede desenvolve suas próprias ferramentas de integração e anotação, procedimento que poderia ser otimizado se houvesse troca destas entre as redes.

Dentre as principais tarefas associadas a estes projetos, podem ser citadas a descoberta e a anotação das características de cada gene. Anotação é a tarefa de descrever várias particularidades de um genoma ou parte de um genoma ou ainda de seqüências de bases (DNA) ou amino-ácidos (proteínas) e depositar estas informações posteriormente em um banco de dados para que sejam utilizadas no futuro para consultas. Esta tarefa usualmente era feita de forma manual e com velocidade lenta. Entretanto, a corrida para seqüenciamento e compreensão de um genoma tem levado os pesquisadores ao que se denomina seqüenciamento de alta vazão ou seja, um seqüenciamento de um genoma inteiro realizado de forma rápida o que envolve mecanização de algumas tarefas (tendência atual).

* Projeto parcialmente apoiado pelo CNPq e pela FAPERGS
† Autores parcialmente apoiados pelo CNPq

Além do conhecimento necessário (o qual é indispensável), o especialista usa uma série de ferramentas computacionais e programas específicos. Grande parte destes processos pode ser automatizado. Logo, o objetivo principal deste projeto é a implementação de um ambiente para anotação automática, acessível através da Internet, que disponibilize o acesso público as suas funcionalidades e código-fonte. Este ambiente é denominado ATUCG – Ambiente para anoTação aUtomátiCa de Genomas. Como *testbed* estão sendo utilizados dados públicos relativos ao genoma de bactérias relacionadas com o organismo *M. hyopneumoniae* que é o organismo-alvo do projeto PIGS/GENESUL [Zaha, 2001] (apoiado pela Rede Sul de Genoma que conta com financiamento da Fapergs e do CNPq).

A ferramenta proposta possibilita que pesquisadores e especialistas não precisem alocar horas em trabalhos repetitivos e tediosos que exigem constante busca em bases de dados, pois tais atividades podem ser feitas, ao menos parcialmente, de forma automática. Para isso, um sistema multiagente é especialmente indicado pois permite uma modularidade ou distribuição das atividades de forma natural, baseada na função da atividade. Diversos agentes se encarregam de realizar as tarefas repetitivas especificadas pelo especialista. Esta ferramenta tem ainda a vantagem de reunir em um único ambiente as funcionalidades requeridas na tarefa de anotação de um genoma, evitando que o usuário tenha que se adaptar às inúmeras interfaces e *modus operandi* das diversas ferramentas hoje disponíveis de forma isolada para auxílio da anotação.

Este artigo está organizado como segue. Na seção 2, a motivação para o desenvolvimento do ambiente proposto e os trabalhos relacionados são comentados. A seção 3 apresenta a arquitetura do ambiente ATUCG. Por fim, na seção 4, as considerações finais e as propostas para trabalhos futuros são comentadas.

2. Motivação e Trabalhos Relacionados

Existem artigos publicados sobre o uso de métodos para auxílio nas diversas tarefas ligadas a um projeto genoma. Entretanto, muitos destes trabalhos reportam ferramentas para áreas e usos isolados, como por exemplo para predição de estruturas de proteínas. Desta forma, tais ferramentas em geral não atingem seu pleno potencial devido ao fato de que se especializam em determinados nichos de aplicação. Com o crescente volume de dados sendo posto a disposição, todos os dias novas relações de homologia são detectadas, tornando necessário que o técnico da área de biotecnologia conheça todos os métodos para decidir pelo seu uso.

Existe portanto, em projetos genoma, no momento, uma grande carência de ferramentas computacionais integradas que lidem com as várias fases do projeto: desde a descoberta das ORFs (open reading frames ou regiões potencialmente codificantes) até a predição da funcionalidade de determinado gene. A idéia de se basear as várias atividades em sistemas multiagente e distribuir as tarefas entre agentes não é nova (embora recente). No projeto GeneWeaver [Bryson et al., 2000], um sistema multiagente está sendo desenvolvido onde os agentes concentram-se nas tarefas de mais alto nível como análise dos dados. Entretanto, os agentes não lidam com a parte mais repetitiva que é justamente a busca de informações. Em [Decker et al., 2001], é descrito um protótipo que objetiva a anotação automática de um vírus com base em busca de informações em bancos de dados públicos.

Outros trabalhos relacionados com a nossa proposta referem-se ao emprego de técnicas de aprendizado de máquina (machine learning) e descoberta de conhecimento em bancos de dados. Técnicas de aprendizado têm sido usadas largamente em bioinformática. Uma proposta de combinação de aprendizado e anotação automática é feita em [Kretschmann et al., 2001], que usa um algoritmo de aprendizado de máquina (o C4.5) para gerar regras para realizar a anotação automática de um dos campos do banco de dados SWISS-PROT.

No entanto, estes ambientes de modo geral não adotam a filosofia de software livre, embora estejam disponíveis via Internet para uso. Neste trabalho é proposta uma plataforma integrada, baseada em software livre e acesso público, para a anotação e seqüenciamento de genomas e proteínas. A motivação à adoção de software livre para o desenvolvimento deste ambiente está baseada nas seguintes premissas: gratuidade de licenciamento; bom nível de qualidade em relação aos softwares proprietários, devido ao número de revisões do código-fonte; estabilidade, robustez e escalabilidade oferecidas pelos softwares básicos (UNIX/LINUX); sofwares necessários às atividades propostas para o ambiente estão disponíveis em versões estáveis de softwares livres; e vasta documentação das funcionalidades dos softwares e de seus códigos-fonte.

3. Arquitetura Baseada em Software Livre

A arquitetura do ATUCG , ilustrada na Figura 1, é formada por diferentes camadas. Detalhes podem ser obtidos em [Bazzan et al., 2003]. A camada I é responsável pela tarefa de encontrar ORFs, a partir da seqüência de DNA de um determinado organismo, informada pelo usuário. Nesta camada, é feita a análise da sequência informada, gerando uma lista de ORFs não redundantes, repassadas à camada II. Na camada II, núcleo da abordagem, são realizadas as seguintes atividades: coleta de dados na base de dados SWISS-PROT; criação das consultas específicas; criação de formatos de saída adequados; geração de dados para os algoritmos de aprendizado de máquina; avaliação da qualidade das regras geradas; preparo dos dados a serem anotados; aplicação das regras; e anotação automática das palavras-chave. A saída desta camada é um modelo de anotação automático dos campos do banco de dados SWISS-PROT. Finalmente, a camada III objetiva auxiliar o usuário na verificação da corretude do modelo de anotação proposto. Para isto, as regras de anotação obtidas são convertidas para uma linguagem semântica e apresentadas de forma legível ao usuário, que pode indicar quais regras foram aplicadas corretamente e quais não deveriam ter sido consideradas.

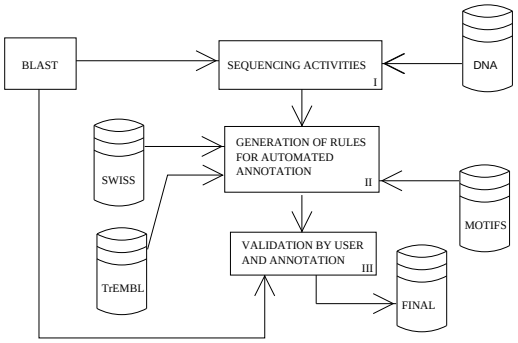

Figura 1: Arquitetura geral do ATUCG

Atualmente, as camadas I e II estão parcialmente implementadas. No futuro, devem ser incorporados os serviços previstos para a camada III, que consiste basicamente da validação pelo especialista. Conforme comentado anteriormente, o sistema atual está completamente implementado adotando-se software livre. Assim que optou-se pelo uso deste tipo de software, foi realizado um estudo dos softwares que possuem compatibilidade, a partir do qual foram selecionados os seguintes softwares/linguagens: Apache, PostgreSQL, PHP e Perl.

O gerenciamento de anotação de seqüências de um organismo é feito através de uma interface web desenvolvida na linguagem PHP, a qual é suportada por um servidor Apache. Os dados obtidos durante o processo de anotação são armazenados em um banco de dados relacional PostgreSQL. Além disso, a linguagem Perl, que oferece ampla facilidade para manipulação de

seqüências de strings (os bancos de dados biológicos armazenam dados neste formato), é utilizada na implementação de scripts que geram os arquivos de entrada para as ferramentas de aprendizado de máquina (C4.5, CN2, T2) e que formatam as regras de anotação. O Perl também é usado na implementação de uma ferramenta de acesso ao banco de dados SWISS-PROT.

Atualmente, o acesso ao sistema restringe-se à rede interna do Instituto de Informática da UFRGS, em função das normas de acesso do Instituto. Futuramente, pretende-se prover o acesso livre ao sistema via Internet, para a comunidade genômica e proteômica do Brasil e Exterior, disponibilizando versões do sistema em Português e Inglês. Além disso, objetiva-se disponibilizar o acesso ao código-fonte do sistema, sob licença GPL.

4. Conclusão e Trabalhos Futuros

Neste artigo foi reportada a experiência obtida com o desenvolvimento da ferramenta para anotação de genoma ATUCG, a qual está baseada em software livre e acesso público. O ambiente foi projetado visando disponibilizar aos pesquisadores de genômica e proteômica ferramentas computacionais integradas para suporte às tarefas de anotação e seqüenciamento de proteínas e genomas. Este ambiente estará disponível publicamente para uso através da Internet, bem como para download, sob os termos GPL. Grande parte dos ambientes que suportam (e integram) as atividades relacionadas ao seqüenciamento e anotação, ainda que raros, não estão amplamente disponíveis. Por outro lado, os ambientes desenvolvidos com software livre ainda não provêm estas atividades de forma integrada. O ambiente ATUCG endereça estes dois aspectos, propondo a integração entre ferramentas para seqüenciamento e anotação, baseada em uma plataforma de software livre e uso público através da Internet.

A próxima atividade refere-se a integração da tecnologia de agentes ao ambiente, onde as tarefas de cada camada serão distribuídas entre diversos agentes, atuando de forma cooperativa (parte destes agentes já encontram-se implementados). Para isto, serão utilizadas ferramentas tais como Apache Tomcat, distribuído sob licença ASF, pacote Java, disponibilizado gratuitamente pela Sun, e plataforma JADE (Java Agent DEvelopment Framework) para suporte a comunicação entre agentes, distribuída sob a licença LGPL. Além disso, devem ser incorporados os serviços propostos para a camada III, que consiste na validação com o especialista. De acordo com esta validação, poderá ser feita a incorporação da anotação ao banco de dados mantido pelo projeto.

Referências

Bazzan, A. L. C., Duarte, R., Pitinga, A. N., F., S. L., Silva, S. C., and Souto, F. A. (2003). ATUCG–an agent-based environment for automatic annotation of genomes. *International Journal of Cooperative Information Systems*, 12(2):241–273.

Bryson, K., Luck, M., Joy, M., and Jones, D. (2000). Applying agents to bioinformatics in GeneWeaver. In *Proc. of the Fourth Int. Workshop on Collaborative Information Agents*, Lect. Notes in Computer Science. Springer-Verlag.

Decker, K., Zheng, X., and Schmidt, C. (2001). A multi-agent system for automated genomic annotation. In *Proc. of the Int. Conf. Autonomous Agents*, Montreal. ACM Press.

Kretschmann, E., Fleischmann, W., and Apweiler, R. (2001). Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, 17:920–926.

Zaha, A. (2001). Projeto rede sul de análise de genomas e biologia estrutural. In Portuguese.