An Agent-Based System for Re-annotation of Genomes*

Leonardo Vianna do Nascimento[1] and Ana L. C. Bazzan[1]**

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

Abstract. Genome annotation projects may produce wrong results since they may be based on obsolete data or wrong models. This work aims to develop an automatic re-annotation system that use agents to perform repetitive tasks and reports the results to the user. These tasks involve BLAST searchs on biological databases (GenBank) and the use of detection tools (Genemark and Glimmer) to identify new ORFs. Several agents execute these tools and combine their results to produce a list of ORFs that are sent back to the user. The goal of this work is to reduce the manual work develop on re-annotation, executing most tasks automatically by computational tools.

1 Introduction

Computers have a important role in DNA analysis. The use of computational tools reduced the analysis time either through the analysis of great amounts of data or through the analysis process integration. The annotation aids by computational methods aims to execute repetitive and time consuming tasks, speeding up the analysis of biological data.

Current computational annotation methods are based mainly on comparative approaches. Computational tools, such BLAST [1] and variants, search for homologous genes information stored in public biological databases. Positive hits are used in functional information inference that are executed by human experts to generate the annotation results reported. Other computational tools, like Genemark [5] and GRAIL [12], use human knowledge models about DNA organization and signals for gene identification.

The human knowledge about gene structure and DNA code organization is incomplete. Moreover, the information stored in biological databases used in annotation may be updated. Thus, genome annotation data can become obsolete and must be re-analysed. These facts are stimulating many biologists to initiate *re-annotation* projects where the information acquired from original annotation is revised and compared with new models and data available.

Despite the development of integrated annotation projects, there is a few efforts on the automation of re-annotation processes. The existing projects have

* Project supported by FAPERGS
** Authors partially supported by CNPq

been based in the manual use of computational tools such as BLAST. These tools executed comparisions with new information available in biological databases for new homologous genes and functional information. The results reported by these tools were analysed and integrated by human experts manually.

The use of a automatic re-annotation module can make the work easier and much faster. A high number of the search and analysis tasks can automatically be done by specialized agents, leaving the user responsible only for the verification of the results.

This work aims to develop an integrated and automatic re-annotation system based on software agents. These agents execute bioinformatics tools, searching on biological databases and identifying new information. The user must register to a service which informs, via e-mail, when a significant change in annotation was found.

This paper is organized as follows. The next section briefly discusses the re-annotation projects. Section 3 presents the proposed approach, showing the organization of the agents and the integration of results. The results of the initial experiments are discussed in Section 4. We conclude with Section 5.

2 Related Work

Re-annotation projects for individual species have been reported in the literature by a handful of groups. Most of them have used computational tools to identify new genes and to extend the information about annotated ones.

The *Haemophilus influenzae* re-annotation project [11] use the BLASTX program to compare intergenic regions with entries in the Genbank database. The results revealed a number of highly significant sequence similarities, indicating that these regions may contain aditional genes.

Given these new information, a revision of the set of proteins encoded by the *H. influenzae* genome was done. This re-annotation process combined sequence similarity searches with statistical analisys of DNA sequence using the Genemark program. This aproach produced a new set of 1703 putative protein-enconding genes containing 23 new ORFs and 107 modified ORFs. Moreover, 47 genes were eliminated because their existence could not be corroborated by any of the applied methods.

In another recent project, the *Mycoplasma pneumoniae* whole genome was re-annotated [6]. The tasks involved in this project included comparisions with other genomes (in particular to *M. genitalium*) and searches on biological databases using tools such as PSI-BLAST [2]. The verification of results was made using similar programs, such as HMMER [7] and FASTA [10], and also complementary tools and methods, such as domain analysis, phylogenetic analysis, analysis of context and clusters of orthologous genes. Experimental techiniques, such as mass spectrometry and mRNA expression have been used too.

The final result of the *M. pneumoniae* had 12 new proteins identified by the analysis of intergenic regions (2 proteins have been identified by mass spectrome-

try, 6 hypothetical proteins and 4 with predicted functional features). Five other ORFs were eliminated because they contained pseudo-genes.

The re-annotation process made from the complete genome of *Thermotoga maritima* [9] compared the 1877 original ORFs with the corresponding new predictions available. After discarding all cases where the two independent analysis agreed, cases of apparent disagreement and hypothetical proteins were analysed in detail. The analysis used several computational tools, including:

- More of five iterations of PSI-BLAST algorithm
- Presence of PROSITE patterns
- Searches on Pfam and COG databases, verifying protein families related to the sequences
- Functional domains organization using searches on PRODOM database.

The final analysis demonstrated that 90% of functional assignments agreed with the original ones. In total, 193 new cases of conflicting annotations were identified (10.3% of the whole genome), of which 164 were new functions identifications and the remaining 29 cases were amendments to previously proposed functions. The total number of functional assignments increase from 1014 (54%) to 1178 (63%), a 16% increase over the original results.

2.1 The ATUCG System

The re-annotation module proposed in this work will be included in a annotation system called ATUCG (Agent-based environmenT for aUtomatiC annotation of Genomes) [3]. The basic architecture of the ATUCG system is shown in Figure 1.

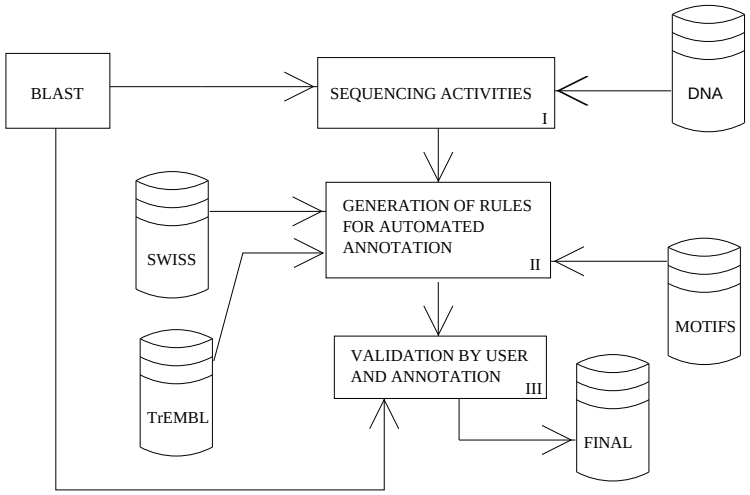

Fig. 1. The ATUCG architecture.

The system is composed by three layers. Layer I is responsible for building a nonredundant ORF list from DNA sequence. This task is accomplished by the execution of several detection tools made by individual agents. The results reported by all agents are merged and sent to user for verification.

The ORFs proceeding from Layer I are analysed in Layer II. This layer executes a partial functional annotation of the ORFs, assigning a keyword list to each ORF. This goal is reached by classification rules that associate keywords found in the Swiss-Prot database with motifs in the ORF sequence. The result is passed to Layer III, where the user can validate the annotation.

The re-annotation module is inserted in Layer I. Its function is re-analyse the annotated ORFs stored in the database and execute the re-annotation process periodically. New annotations found that are confirmed by the user are then stored in the system database.

3 The Re-annotation System

The recent re-annotation projects have used isolated computational tools such as BLAST. The integration of the several tools used in the process was made manually by the human experts. There is a lack of computational re-annotation tools that automatically execute the analysis programs and integrate their results.

This work aims to develop an automatic re-annotation system, based on the multiagent systems technology. The system analyse the annotated ORFs stored in a local database. The agents search for new annotations stored in public biological databases as well as identify possible new genes. The analysis of annotated data is distributed and each agent executes a single task. The organization of the several agents is shown on Figure 2.

The re-annotation approach proposed here use three types of analysis:

- *BLAST agents* run searches comparing annotated ORFs stored in local databases with entries stored in the GenBank database [4]. If different entries are found in the returned list, they are reported to user who is registered in the system as someone interested in getting this kind of information. Only the most significant hits are analysed. The BLAST agents search for similar ORFs that present modifications in sequence or new annotation information.
- Similarly, the intergenic regions detected in the DNA sequence of the organism are "BLASTed" against entries on GenBank in order to find possible new genes. ORFs stored in system database contains position-specific information that are used by the *Intergenic Agents* to extract intergenic regions of DNA sequence. Significant hits will be reported to the user for verification and validation.
- The user can choice to run Genemark and Glimmer detection tools. Each tool is executed by an agent. These agents generate additional ORFs reported in the final list.

The BLAST execution is configurable. The user can choose the BLAST variant (blastx, blastn, blastp, tblastx,tblastn), significance threshold and the sequence database used in the search.

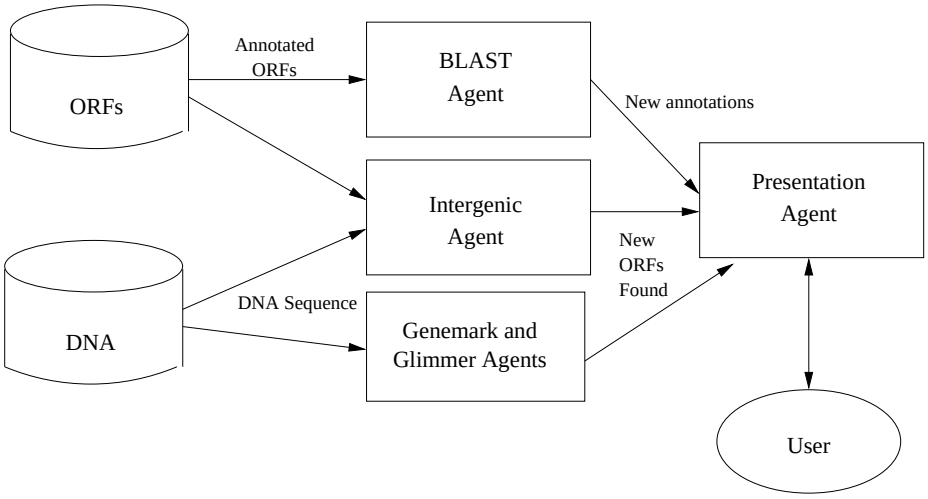

Fig. 2. Agents organization in the system.

The agents described above make periodic analysis of annotated data. The results of the several agents will be merged in a list reported to the user. This task is performed by the *Presentation* agent that groups the BLAST hits by position relative to DNA sequence. Overlapped BLAST hits are placed in the same groups. Each group represents a possible re-annotation to a gene or a new ORF. A hit list in a group is ordered by significance using the *E value* returned by BLAST. An analysis is made on group hits to determine the limits of pseudo-ORFs. The user can choose to discard ORFs, manually edit the data obtained or accept the re-annotation proposed.

Each agent follows the structure shown in Figure 3. The *Knowledge* module contains the information used by the agent as well as the actions results. These actions are executed by the *Action* module that processes the execution requisitions. These requisitions are sent periodically by the system. These actions can be BLAST actions (that execute the NCBI BLAST tool), Intergenic processing actions or gene detection actions (Genemark, Glimmer). The *Communication* module receives the messages sent by the other agents and by the system. This module processes the incoming messages and send them to the apropriate destinations. Action execution requisitions are sent to the Action module while the data messages are sent to the Knowledge module. Agents exchange messages using FIPA ACL communication language.

4 Results

We use the *Mycoplasma pneumoniae* original ORFs reported in [8] (Some of them shown in Table 1) for validation and evaluation of system results. The *M.*

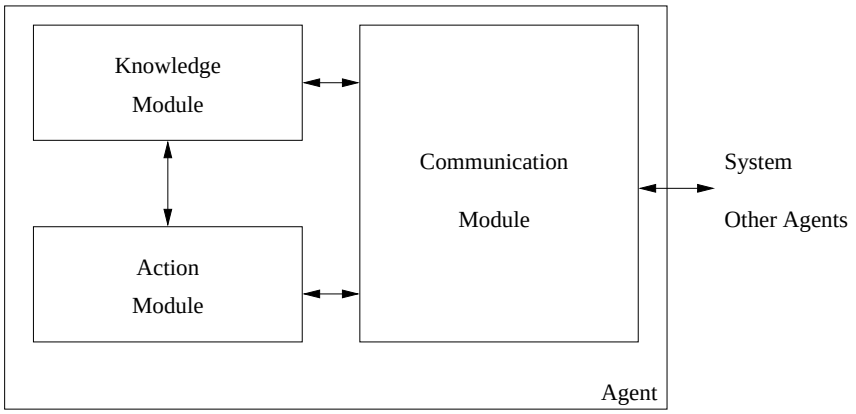

Fig. 3. Agents structure.

pneumoniae genome was recently re-annotated. The re-annotated ORFs reported in [6] are available in Genbank database (GI: 13507739).

Table 1. Original annotated ORFs of *M. pneumoniae*.

ORFs	**Description**
ORF: 77..3418 GI: 1673646	conserved hypothetical protein, MG140 homolog
ORF: 3594..5978 GI: 1673647	conserved hypothetical protein
ORF: 6261..6662 GI: 1673648	hypothetical protein
ORF: 7145..7819 GI: 1673649	hypothetical protein
ORF: 7647..8951 GI: 1673650	hypothetical protein
...	

The original *M. pneumoniae* ORFs were submited to the BLAST and Intergenic agents. The BLAST tool was executed with an E value of 10^{-6} by both agents. The results reported by these agents were analysed by the Presentation agent. Hits from re-annotated *M. pneumoniae* genome were ignored and the final results were similar to the ones reported in [6]. Table 2 shows some ORFs detected using the results reported by the re-annotation system and the correspondent *M. pneumoniae* original re-annotated ORFs. The first column shows the re-annotated ORFs detected by the system. This column contains the ORF position and the function description of most similar hits. The second column shows the correspondent ORF in the original re-annotated genome.

Table 2. Results reported by the system.

Predicted ORFs	**Original Re-annotated ORFs**
ORF: 759..1832 Hits: DNA Polimerase	ORF: 682..1834 GI: 13507740
ORF: 1839..2765 Hits: dnaJ-like proteins	ORF: 1838..2767 GI: 13507741
ORF: 2870..3418 ORF: 3422..3595 ORF: 3596..4780 Hits: DNA gyrase subunit B	ORF: 2869..4821 GI: 13507742
ORF: 4822..5979 ORF: 5980..6261 ORF: 6262..6663 ORF: 6664..7146 Hits: DNA gyrase subunit A	ORF: 4821..7340 GI: 13507743
ORF: 7313..7819 ORF: 7649..8560 Hits: seryl-tRNA synthetase	ORF: 7312..8574 GI: 13507744
...	

In the final results 1339 pseudo-ORFs are found. The ORFs detected using the groups proposed by the system and the original re-annotated ORFs have similar functions. Some ORFs detected by the system are associate with the same original re-annotated ORF. These ORFs are regions of the same re-annotated ORF found in neighbor genes on the original annotated genome. These ORFs presented the a high number of similar BLAST hits what suggests that these sequences can be merged in a single ORF. Thus, an approach to treat these cases is under development.

5 Conclusion and Future Work

The automatic re-annotation is an useful process to detect changes in genome annotation. The human experts can analyse the results reported by several tools executed by software agents. These agents can run the tools in a distributed and integrated way, searching new entries on sequence databases and executing other tools over the original annotated sequences.

The re-annotation module proposed in this work uses the multi-agent technology to re-analyse ORFs annotated using the ATUCG system. The module uses several agents running BLAST, Genemark and Glimmer tools over the annotated genome to identify new annotations. Intergenic regions are analysed to identify new genes.

The *Mycoplasma pneumoniae* genome has been used for evaluation purposes. The results show that the new ORFs reported by the system are similar to the ones reported in [6]. More tests will be executed in the future using Genemark

and Glimmer agents over the DNA sequence and other annotated genomes, such as *H. influenzae*.

We intend to use multiple alignment tools, such as HMMER, to extract motif and pattern information from BLAST hits in groups. The results will be compared to motif databases and will be used to add some functional information about new sequences found.

Conflicts can occur when the results reported by the agents are merged. New ORFs identified can overlap other annotated ORFs. Which are the correct ORFs? Which is the criteria to be used to select the correct ones? Unfortunately the most part of the ORF selection work will be done by user. However this work can be improved by some pre-processing tasks done by the system. We are studying an approach to treat the possible conflicts that can occur between original annotated ORFs and new re-annotated ORFs found.

References

1. Altschul, S. *et al*: Basic local alignment search tool. Journal of Molecular Biology. **215** (1990) 403-410
2. Altschul, S. *et al*: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. **25** (1997) 3389-3402
3. Bazzan, A. *et al*: ATUCG - An Agent-Based Environment for Automatic Annotation of Genomes. International Journal of Cooperative Information Systems. **12** (2003) 241-273
4. Benson, D.; Karsch-Mizrachi I.; Lipman D.; Ostell J.; Wheeler D.: GenBank: update. Nucleic Acids Research. **32** (2004) Database issue D23-D26
5. Borodovsky M. and McIninch J.: GeneMark: parallel gene recognition for both DNA strands. Computers & Chemistry. **17** (1993) 123-133
6. Dandekar, T. *et al*: Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames. Nucleic Acids Research. **28** (2000) 3278-3288
7. Durbin, R. *et al*: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. 1.ed. Cambriedge University Press. (1998)
8. Himmelreich, R.; Hilbert, H.; Plagens, H.; Pirkl, E.; Li, B. and Herrmann,R.: Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Research. **24** (1996) 4420-4449
9. Kyrpides, N. *et al*: Analysis of the *Thermotoga maritima* genome combining a variety of sequence similarity and genome context tools. Nucleic Acids Research. **28** (2000) 4573-4576
10. Person, W. and Lipman, D.: Improved tools for biological sequence comparision. Proc. Natl. Acad. Sci. **85** (1998) 2444-2448
11. Tatusov, R. *et al*: Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparision with *Escherichia coli*. Current Biology. **6** (1996) 279-291
12. Uberacher, E. and Mural R.: Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. Proc. Natl. Acad. Sci. **28** (1991) 11261-11265.