Applied Soft Computing 35 (2015) 373–385

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

# A rule-based expert system for inferring functional annotation

Daniela Xavier[a,*,1], Berta Crespo[b], Rubén Fuentes-Fernández[c]

[a] GARP (Genomic and RNA Profiling Core), Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, 77030 Houston, USA
[b] Department of Fish Physiology and Biotechnology, Instituto de Acuicultura de Torre la Sal, Consejo Superior de Investigaciones Científicas (CSIC), Torre la Sal, C/ Ribera de Cabanes s/n, 12595 Castellón, Spain
[c] Department of Software Engineering and Artificial Intelligence, Universidad Complutense de Madrid, C/ Profesor José García Santesmases 9, 28040 Madrid, Spain

ARTICLE INFO

*Article history:*
Received 23 January 2014
Received in revised form
18 December 2014
Accepted 9 May 2015
Available online 3 July 2015

*Keywords:*
Functional annotation
Rule-based expert system
Bioinformatics

ABSTRACT

Functional annotation is the process that assigns a biological functionality to a deoxyribonucleic acid (DNA) sequence. It requires searching in huge data sets for candidates, and inferring the most appropriate features based on the information found and expert knowledge. When humans perform most of these tasks, results are of a high quality, but there is a bottleneck in processing; when experts are largely replaced by automated tools, annotation is faster but of poorer quality. Combining the automatic annotation with expert systems (ESs) can enhance the quality of the annotation, while effectively reducing experts' workload. This paper presents INFAES, a rule-based ES developed for mimicking the human reasoning in the inference stage of the functional annotation. It integrates knowledge on Biology and heuristics about the use of Bioinformatics tools. Its development adopts state-of-the-art methodologies to facilitate the acquisition and integration of new knowledge. INFAES showed a high performance when compared to the systems developed for the first large-scale community-based critical assessment of protein function annotation (CAFA) [1].



## 1. Introduction

One of the most challenging tasks of Genomics is predicting the biological function of deoxyribonucleic acid (DNA) sequences, a procedure called *functional annotation*. Its outcome has to be as reliable and accurate as possible, as it will be used in further researches, including to predict new annotations.

The functional annotation is currently a complex, labor-intensive, and time-consuming task for experts. It requires a high degree of expertise to use the proper tools, algorithms, and databases in order to collect relevant information, and to make the pertinent decisions. The amount of genomic data that has been produced, especially in the last years, makes this *manual approach* feasible just for small data sets or reference genomes. Besides, it may produce conflicting interpretations of the analysis [2].

The alternative *automatic approach* uses tools able to process large volumes of data consistently without (almost any) user intervention. Its main drawback is that tools only use limited expert knowledge, so their results are less precise than those of human experts. For instance, only a few tools, like Figenix [3], take orthology knowledge into consideration, what increases the reliability of the annotation. Moreover, tools usually lack the flexibility to adapt to different needs and an ever evolving environment. For example, many of them restrict the kind of query sequences they support (e.g., sequences inside the genomic context [2] or bacterial sequences [4]), and they integrate only a limited and fixed set of data sources to search [5].

A possible way to preserve the quality of the manual annotation without running into its drawbacks is applying expert systems (ESs) to emulate the expert reasoning in certain parts of the process. Among the variety of ESs for annotation [6], rule-based ones [7] are particularly well-suited because of several reasons. First, rules are a natural way of representing knowledge about procedures and heuristics [7], as that applied to a large extent in functional annotation. Second, there are multiple knowledge elicitation techniques [8] to guide rule specification with experts. Third, since rules are more easily understandable by experts than code, their usage promotes system evolution through user involvement [9].

Despite these advantages, existing rule-based ESs (RBESs) for annotation present several issues. Their development is not usually related to standard good practices. Literature does not report

* Corresponding author. Tel.: +34 91 394 7548.
*E-mail addresses:* xavier@bcm.edu (D. Xavier), ruben@fdi.ucm.es (R. Fuentes-Fernández).
[1] This work was partially done while working at the Department of Biochemistry and Molecular Biology I, Universidad Complutense de Madrid, Spain.

http://dx.doi.org/10.1016/j.asoc.2015.05.055
1568-4946/

the application of any engineering methodology, so key decisions are neither explained nor documented. Moreover, they frequently rely on ad-hoc technologies poorly supported. For instance, there is no documentation on the development approach of the Ensembl Analysis Pipeline (EAP) [2], and it uses its own inference engine. This way of working makes systems difficult to maintain and evolve. Regarding the traditional limitations of annotation tools previously mentioned, RBESs facilitate their solving, but designers need to address them explicitly. For instance, EAP [2] and Figenix [3] are designed to integrate other tools, but only considering dataflow management. There are no guidelines, either general for RBESs or particular for these systems, on how to integrate the new tools in their ESs regarding knowledge, so this integration relies on designers' expertise.

To address these issues, this work proposes a RBES for inferring the functional annotation of DNA sequences called INFAES. INFAES is part of a wider research project called MASSA, a multiagent system (MAS) to Support functional Annotation. This MAS is a community of Intelligent Agents (IAs) [10] that work together. They implement a flexible pipeline of Bioinformatics tools that collects candidates and clues for the prediction task. Then, INFAES uses this information to evaluate the candidates and infer the most likely function.

Although there are already some ESs and RBESs for this task, INFAES was specifically developed to overcome several of their limitations. In particular, it provides an integration of knowledge and analyses previously scattered among different tools, and mechanisms (i.e., an architecture and development guidelines) to facilitate further evolution of the system in order to keep it up to date with emerging research.

As for the annotation process, INFAES is capable of assigning accurate functional annotations to DNA sequences regardless of the species, and whether they are or not complete genomes. Moreover, INFAES rules comprise knowledge that other systems do not consider, what increases the annotation effectiveness. Its rules are able to mine additional data related to the information from the pipeline, and compare the candidate annotations to come to a conclusion. These comparisons apply heuristics that integrate analysis variables from the pipeline tools (e.g., e-value, bit score, identity, and homology likelihood), and Biological knowledge (e.g., the orthology relationship between sequences, the domains and families, and the level of conservation of important sites). This knowledge has been extracted from several sources [11], and pursues modeling the Biologist expertise at the inference stage. These biological concepts are explained later in Section 2.1.

Since INFAES has a special focus on evolution, its architecture and development consider requirements for maintenance. These have not been explicitly taken into account in related works, but they must be in order to keep tools up to date in a domain with a fast changing pace.

INFAES knowledge is structured around the computation of scores and their interpretation. This facilitates considering new knowledge. It appears as new sets of rule to compute additional scores, that specific rules combine with existing ones.

Addressing evolution is not only an issue of system design. The development process also has to consider it. INFAES improves this aspect compared to existing tools for annotation. It follows CommonKADS [12], a well-known methodology for ESs, to build its Knowledge Base (KB) and document the process [11]. Moreover, it adopts widespread technologies, such as Java and Drools [13], which reduces development costs. These decisions allow designers and experts to focus on eliciting and managing the specific domain knowledge required for the annotation process, while facilitating the examination and validation of results.

The rest of the paper discusses these aspects in detail. Section 2 presents the background of the functional annotation problem. MASSA is briefly described in Section 3, while INFAES and the methodology used to tackle the problem are introduced in Section 4. Section 5 exemplifies the annotation and evaluates the system performance. The state of the art in systems for annotation is reviewed in Section 6. Finally, Section 7 discusses conclusions about the work and its results.

## 2. Background

In most living beings, the hereditary information is stored in macromolecules of DNA. Such molecules comprise two long complementary strands composed of small molecules called nucleotides. *Genes* are sections of these strands that detain the information to produce the proteins that participate in different biological processes. Determining the correspondences between genes and biological processes is the goal of *functional annotation*. This is a complex task, as there is not a one-to-one correspondence between genes and their functions, and the involved techniques are generally expensive in effort and resources.

Next sections provide further insights into this problem. Section 2.1 presents its biological basis, and Section 2.2 the current techniques applied in functional annotation and their main features.

### 2.1. Biological concepts

DNA molecules, and in particular their genes, contain the genetic information required to synthesize functional cellular components. This process is called *gene expression*, and has two parts. The first one is the *transcription*, where a complementary strand of nucleotides of a gene is transcribed into a messenger ribonucleic acid (mRNA); the second one is the *translation*, where the mRNA is translated into proteins. Since the DNA can be transcribed from both strands, a total of six *reading frames* (three from each strand, as they are always translated grouped by triplets) are possible for further translation into proteins.

During the transcription, different parts of the gene can be used to form distinct mRNAs. This phenomenon is known as *alternative splicing*, and it is the reason why a single gene can code different proteins.

In a simplistic way, a protein can be seen as a large molecule composed of amino acid (AA) chains. The *primary structure* of a protein is the linear sequence of these AAs. This structure can fold into itself forming two-dimensional organizations (e.g., helices, sheets, and turns) known as *secondary structure*. The components of this structure in turn, are folded into compact globules that form the *tertiary structure*.

The protein function is directly linked to its tertiary structure. However, some portions of the primary structure can vary substantially without changing the protein role. In fact, the AA sequence contains sections called *conserved residues* or *regions*, which are responsible for the functionality of the protein. Among them are *domains*.

Domains are compact, local, semi-independent units in proteins. Their existence and function is not tied to specific proteins, and their sequences tend to be more conserved than those of other regions. For these reasons, they are widely used to establish the functions of proteins. A protein may have one or more domains, and proteins that have the same domains are generally classified into the same *family*.

A protein family is a set of evolutionary-related proteins that share a significant degree of similarity. The members of a family are called *homologs*, and they descend from the same ancestor. Usually, homologs are 25% or more identical throughout their sequences. Homologs can be classified as *orthologs* or *paralogs*. The first ones