

MASTER THESIS

ONTOLOGY-BASED INFORMATION EXTRACTION OF E. COLI REGULATORY NETWORKS FROM SCIENTIFIC ARTICLES

Alejandra C. López Fuentes

Directors: Dr. Antonio Moreno y Dr. David Isern

June 2011



ITAKA
Intelligent Technologies for
Advanced Knowledge Acquisition

Contents

List of figures	v
List of figures	v
1 Introduction.....	1
1.1 Background.....	1
1.2 RegulonDB and the automatic extraction of biological information	4
1.3 Aims of this work.....	5
1.4 Document organization.....	6
2 State-of-the-Art on Information Extraction	9
2.1 What is Information Extraction?	9
2.2 IE techniques	10
2.2.1 Statistical Information Extraction.....	11
2.2.2 Rule-based Information Extraction	12
2.2.3 Ontology-Based Information Extraction	13
2.2.3.1 <i>Ontology components</i>	14
2.2.3.2 <i>Characteristics of ontology-based IE systems</i>	15
2.2.3.3 <i>Common architecture of OBIE systems</i>	17
2.3 Information Extraction in molecular biology	18
2.3.1 GENIES (2001).....	18
2.3.2 BioRAT (2004).....	20
2.3.3 BioText (2005)	22
2.3.4 String - IE (2005)	24
2.3.5 GRNS (2009)	27
2.4 Comparative analysis of the IE systems	28
3 Ontology model built from RegulonBD	31
3.1 RegulonDB.....	31
3.2 Translating RegulonDB into an ontology.....	34
3.2.1 Classes	34
3.2.2 Properties.....	35
3.2.2.1 <i>Property restrictions</i>	37
3.2.3 Individuals	38
3.3 Lexicon class	40
3.3.1 MultiUnit class.....	40
3.3.2 Biological class.....	41
3.3.3 ParseTerms class	42
3.4 Grammar class.....	43
3.5 Complete ontology.....	44
4 Improving and generalizing the CCG-IE system.....	47
4.1 CCG-IE system architecture.....	47
4.2 System modules: explanation and improvements.....	50
4.2.1 Preprocess	50

4.2.2 POS tagging	54
4.2.3 Semantic tagging	57
4.2.4 Parsing	59
4.2.4.1 <i>Tagset mapping</i>	60
4.2.4.2 <i>Grammar</i>	61
4.2.5 Interaction selection	66
5 Cases of study and results	69
5.1 Abstracts.....	69
5.2 Full-text	78
6 Conclusions and future work.....	81
6.1 Conclusions.....	81
6.2 Future work	82
7 References.....	83
Appendix A:	87
Regulation of Escherichia coli Hemolysin E Expression by H-NS and Salmonella SlyA	87
Appendix B:	97
Influence of DNA geometry on transcriptional activation in Escherichia coli.....	97

List of figures

FIGURE 1: GROWTH OF INTERNET USER.	2
FIGURE 2. GROWTH OF PUBMED JOURNAL ARTICLES	3
FIGURE 3 RULE-BASED IE.....	12
FIGURE 4: GENERAL ARCHITECTURE OF AN OBIE SYSTEM	17
FIGURE 5: GENIES ARCHITECTURE	19
FIGURE 6: BIO RAT ARCHITECTURE	21
FIGURE 7: THE IE SYSTEM ARCHITECTURE.....	22
FIGURE 8: STRING-IE ARCHITECTURE	24
FIGURE 9: CASS GRAMMAR EXAMPLE	26
FIGURE 10: RECOGNIZING TYPES OF RELATIONS.....	27
FIGURE 11. GRNS SYSTEM ARCHITECTURE	27
FIGURE 12: CASS GRAMMAR EXAMPLE	28
FIGURE 13: PART OF THE REGULONDB ENTITY-RELATIONSHIP DIAGRAM	32
FIGURE 14: TABLES CONVERTED INTO ONTOLOGY CLASSES.....	34
FIGURE 15: CLASS HIERARCHY.....	35
FIGURE 16: MAPPING BETWEEN COLUMNS AND PROPERTIES	36
FIGURE 17: CONDITIONS FOR COMPLEX CLASS.....	37
FIGURE 18: PROCESS FOR POPULATING THE ONTOLOGY	39
FIGURE 19: EXAMPLE OF AN INSTANCE CREATED FROM A QUERY RESULT	40
FIGURE 20: EXAMPLE OF A MULTI UNIT INSTANCE	41
FIGURE 21: EXAMPLE OF AN INSTANCE OF BIOLOGICAL CLASS	42
FIGURE 22: ANAPHORA_ GRAMMAR.REG FILE, GRAMMAR RULES.....	43
FIGURE 23: EXAMPLE OF AN INSTANCE OF LEVEL1 CLASS (GRAMMAR SUBCLASS).....	44
FIGURE 24: FINAL STRUCTURE OF THE ONTOLOGY	45
FIGURE 25: PROCESSING PIPELINE OF CCG-IE SYSTEM	48
FIGURE 26: EXAMPLE OF MULTI-UNIT TERMS	53
FIGURE 27: PREPROCESS OUTPUT.....	54
FIGURE 28: POS TAG OUTPUT USING DIFFERENT PARAMETER FILES	57
FIGURE 29: INSTANCES OF THE PHYSICALOBJECT SUBCLASSES	58
FIGURE 30: COMPARISON OF SEMANTIC TAGGING PROCESSES	59
FIGURE 31: GENERATION OF THE TAG MAP FILE FROM THE ONTOLOGY INSTANCES.....	61
FIGURE 32: EXAMPLE OF HOW THE .REG FILE WAS BUILT.....	64
FIGURE 33: COMPARISON OF THE PARSING PROCESS FILES.....	65
FIGURE 34: CASCADED FILE CONVERTED INTO A XML FILE	66
FIGURE 35. FINAL XML FORMAT WITH THE IDENTIFIED INTERACTIONS.....	67
FIGURE 36. SENTENCE FROM THE PARSE FILE REPRESENTED IN XML	74
FIGURE 37: IDENTIFIED REGULATORY INTERACTIONS.....	75
FIGURE 38: FINAL XML FORMAT SHOWING THE FOUNDED INTERACTIONS	76

List of Tables

TABLE 1: CHARACTERISTICS OF BIOLOGICAL ORIENTED IE SYSTEMS 29

TABLE 2: RELATION BETWEEN NAMES IN THE ONTOLOGY CLASSES AND THE REGULONDB TABLES 35

TABLE 3: ONTOLOGY PROPERTIES 38

TABLE 4: PENN TREEBANK TAGSET 55

TABLE 5: CORPUS DESCRIPTION 69

1 Introduction

This chapter starts with the presentation of an overview of the continuous growth of digital information in recent times, which has lead to the development of computational automated technologies to extract specific information. Section 1.2 provides a concrete example of this problem in the area of molecular biology and how it is being addressed. Finally, the last two sections described the aims of this work and the document organization, respectively.

1.1 Background

The Internet has become a vital tool in our Information Society. It has grown exponentially through the 1990's and into the 21st century and it has expanded to serve hundreds of millions of users and a multitude of purposes in all parts of the world (Figure 1).

The Internet carries a vast range of information resources that provide general information which can be easily found on any subject with the simple click of the mouse. Thousands of new Web pages appear every day: news, magazine, and journal articles are constantly being created online. All these trends result in an enormous amount of text available in digital form; however, these repositories of text are mostly untapped resources of information[1], and identifying *specific* desired information in them becomes a difficult task.

“Over 95 percent of the digital universe is unstructured data. In organizations, unstructured data accounts for more than 80 percent of all information.” (Jonathan Martin, HP)

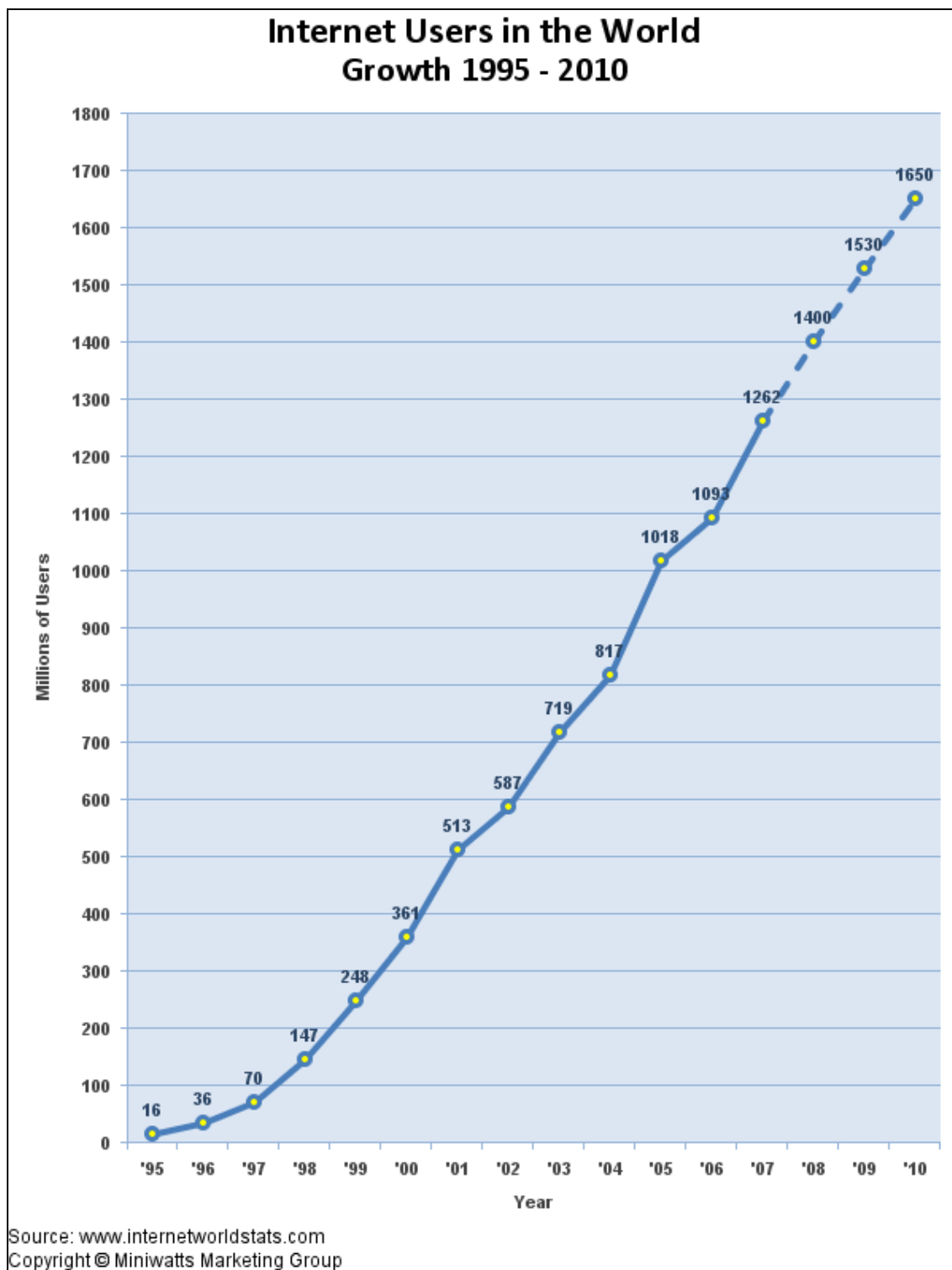


Figure 1: Growth of internet user.

The Web is drowning us with an enormous amount of pages of information related to all domains of discourse. Recently, the fields of Molecular Biology and Medicine have enjoyed an explosive development; as a result, the number of biomedical **journal**

articles has grown exponentially, and an urgent need on the part of biologists to find specific information in the ever expanding biological literature has emerged.

Access to full-text articles is difficult because each journal has its own organization, interface and formatting conventions. Fortunately, in Biology and Medicine abstracts are collected and indexed in MEDLINE, hosted at the National Library of Medicine (NLM). The system at the NLM is called PubMed and it indexes 5,511 different journals in Medicine and Molecular Biology. It currently contains more than 18 million abstracts and it keeps growing¹ (see Figure 2).

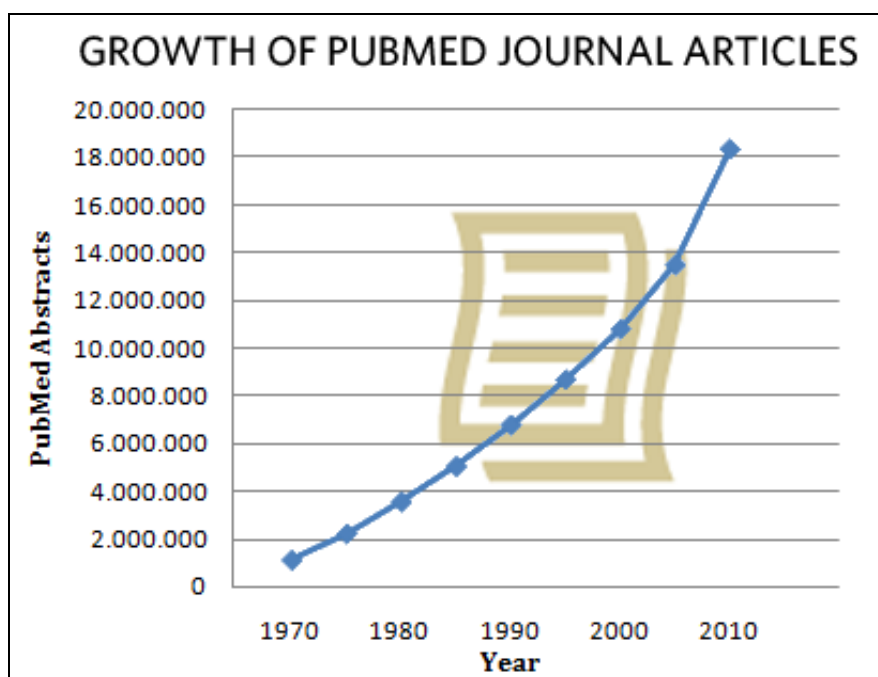


Figure 2. Growth of PubMed journal articles

Today, it is possible to select a number of articles using queries that can be very complex (including combinations of different attributes of the publication and logical operators) and to read a short associated summary (the abstract) usually provided by the authors. With the increasing distribution of journals in electronic format, the full text of any paper will soon be only one click away [2].

¹ http://www.nlm.nih.gov/bsd/index_stats.html (Last access: March 2011)

However, access to literature does not solve the problem of the selection of information. Reading or even browsing all of those publications is something that most researchers will not contemplate. Clearly, there is a necessity of developing methods for automatic extraction of relevant information from any source of scientific data, especially sources such as scientific literature written in human language (also known as natural language).

Molecular Biology is a branch of Biology which primarily deals with functions, characteristics and structures of mainly three major macro-molecules: DNA, RNA and Proteins. Molecular research technologies and developments in Information Technologies have combined to produce a huge amount of information related to the multiple research areas of Molecular Biology such as sequence analysis, genome annotation, analysis of gene expression, analysis of regulation, etc [3].

Several databases have been designed and implemented in order to compile all this information [4-5]. However, the amount of knowledge in the domain is overwhelming and grows at an unprecedented rate. Much of it is published and available in the Internet in the form of scientific texts, therefore, in natural language.

The main hypothesis in this work is that it is necessary to develop improved computer-based tools to aid human experts to extract the facts needed by the research community [6]. In the recent years, a technology called *Information Extraction* (IE) has emerged within the Artificial Intelligence community to face this problem.

1.2 RegulonDB and the automatic extraction of biological information

RegulonDB² is the largest electronically-encoded database of the Escherichia coli K-12 bacteria. It offers curated knowledge about the elements and interactions of the network of transcriptional regulation in Escherichia coli K-12. RegulonDB represents a “gold standard” in the bioinformatics of gene regulation design in bacterial genomics

² <http://regulondb.ccg.unam.mx/> (Last access: June 2011)

[7]. It was developed by the Center of Genomic Sciences (CCG) of the Universidad Nacional Autónoma de México (UNAM).

The information contained in RegulonDB is complete and updated; however, maintaining it in this state has become more complex and compelling. Currently, to obtain information of the components of the regulatory network, *annotations derived from the literature* are generated through a process of high-quality manual curation.

The usual curation process starts by searching for articles that contain information about transcriptional regulation using a set of pertinent keywords in the PubMed database. To select only the most relevant articles, a team of biologists-curators reads the abstracts of these papers. Finally, the curators have to read and review very carefully the selected papers to ensure the reliability of the data to annotate them in the RegulonDB database. This turns the manual curation process into an expensive and labor-intensive endeavour.

The CCG has started to develop processes that allow the automation of the curation process, so that information about the interactions of the regulatory network can be automatically extracted by analyzing scientific articles (abstracts or full texts). Currently, the results obtained with this automated process are limited (between 10 and 45% of the current contents of RegulonBD, depending on the input corpus [8]). The main intention of this research work is to continue this line of study, generalizing and improving the initial ideas developed at the CCG. Considering that it will always be a semi-automatic process since it will be necessary the collaboration of a human expert for validating the results of the Information Extraction process.

1.3 Aims of this work

The purpose of this work is to try to generalize the preliminary ideas proposed at CCG and improve their performance by developing and implementing new techniques of ontology-based Information Extraction (OBIE).

The specific tasks to be completed in order to fulfill the objective are the following:

- Study of the available information in the RegulonDB database and of the automatic curation process proposed by the Center for Genomic Sciences at UNAM. Presentation of a state of the art of works that propose the automatic analysis of scientific articles in the biomedical field.
- Develop a methodology to translate the RegulonDB entity-relation structure to an ontology structure.
- Present a state of the art of existing mechanisms of ontology-based information extraction.
- Development and implementation of generic mechanisms for Information Extraction, which exploit the semantic content of the information in a domain ontology.
- Application of these new methods to the case of extracting information about the regulatory network in *E. coli*.
- Test the new system with several scientific articles to verify its good performance.

1.4 Document organization

This Master Thesis is organized as follows:

Chapter 2 gives an overview of Information Extraction. It is divided in four parts: the first one is intended to explain what Information Extraction is; the second part explains the existing Information Extraction techniques (statistical, rule-based and ontology based), with a special focus on the last one. The third part consists in a State-of-the-Art of Information Extraction systems with a biological approach. Finally, the last part summarizes the main common characteristics of these systems.

Chapter 3 explains some relevant issues of the RegulonDB database such as the field of knowledge to which it belongs, its entity-relationship model and the description of the tables related to regulatory networks. Subsequently, the developed methodology to translate the existing RegulonDB structure into an ontology structure is described. Finally, the resulting ontology is explained as well as its classes, properties and relationships.

Chapter 4 is divided in two parts. The first one gives an overview of the current state of the IE system developed by the CCG. In the second part the tasks realized in all different modules of the system are explained as well as the changes that were implemented in order to improve and generalize the system.

Chapter 5 includes the testing and evaluation of the improved system explained in the previous chapter by considering four cases of study (analyzing two full-text articles and two abstracts). Explanation of the results of each case of study is also included.

Chapter 6 summarizes some conclusions of this work and devises some lines of future work order to improve the results.

Appendixes A and B correspond to the two analyzed full-text articles in Chapter 5.

2 State-of-the-Art on Information Extraction

This chapter provides primarily an introduction to Information Extraction, its definition and existent techniques such as statistical, rule-based and ontology-based. Regarding the ontology-based technique, section 2.2.3 briefly explains the concept of ontology and its components, for a better understanding of the approach.

Section 2.3 gives an overview of five State-of-the-Art Information Extraction systems with a biological approach, and finally section 2.4 contains a comparative table of the mentioned systems.

2.1 What is Information Extraction?

Information Extraction (IE) is concerned with selecting salient facts about a given topic from documents. Typically, these facts are then entered automatically into a database, which may then be used for further processing. IE is a technology based on analyzing natural language.

The following are several definitions of Information Extraction that have been proposed by various researchers:

- "The identification of instances of a particular class of events or relationships in a natural language text, and the extraction of the relevant arguments of the event or relationship." It "involves the creation of a structured representation (such as a data base) of selected information drawn from the text." [9]
- "A subfield of natural language processing that is concerned with identifying predefined types of information from text." [10]

-
- "An emerging Natural Language Processing (NLP) technology whose function is to process unstructured, natural language text, to locate specific pieces of information, or facts in the text, and to use these facts to fill a database." [11]
 - "The task of filling template information from previously unseen text which belongs to a pre-defined domain." [12]
 - "The task of locating specific pieces of data within a natural language document". [13]
 - "A technology based on analyzing natural language in order to extract snippets of information." [14]

2.2 IE techniques

An IE system must be designed to extract the entities and relations appropriate to a specific task [15]. In general, the core of an IE system is an *extractor* which processes text; it overlooks irrelevant words and phrases and attempts to home in on entities and the relationships between them [16]. There are several methods to perform IE which can be classified into *statistical* methods and *rule-based* methods, being the later superior than the former speaking of tasks that require extraction of relationships (expression of a specific product of a certain gene in a certain tissue, specific function of a protein etc) [17] though they are labor-intensive to implement.

Additionally, another approach has recently emerged as a subfield of IE and is called *ontology-based information extraction* (OBIE). Since IE is essentially concerned with the task of retrieving information for a particular domain, specifying the concepts of that domain formally and explicitly through an ontology can be helpful to this process [18]. In section 2.2.3 this approach is explained in detail.

2.2.1 Statistical Information Extraction

Unlike rule-based IE, the statistical approach uses training material which can be annotated or unannotated (through Machine Learning techniques). Several methods are based on the frequency of occurrence of words in a large text corpus that has been previously organized in line with some form of external knowledge [19]. The basic elements of text are words, and their frequencies, co-occurrences and lexical features can be used to cluster and classify text, find documents that treat a similar theme or select significant words that describe a group of documents. More specific methods were developed by Andrade and Valencia [20], who used the characteristics of word distributions in text clusters to extract significant words. These approaches are limited because words are often ambiguous and refer to more than one object. Moreover, different words can have the same meaning (synonyms) and the same word can be part of constructions with very different meanings.

There are many statistical methods that have been applied for specific tasks in the IE process and have also proved to be very effective where there is a large corpora of training data available [21]. Here are some examples:

- For automatic speech recognition, smoothed n-grams are used in order to find the most probable string of words $w_1, \dots, w_n$ out of a set of candidate strings [22].
- Part-of-speech tagging can be done using Hidden Markov Models to find the most probable tag sequence $t_1, \dots, t_n$ given a word sequence $w_1, \dots, w_n$ [23].
- Syntactic parsing using probabilistic grammars can find the most probable parse tree T given a word sequence $w_1, \dots, w_n$ (or tag sequence $t_1, \dots, t_n$) [24].
- For word sense disambiguation, Bayesian classifiers are used to find the most probable sense s for word w in context C [25].

2.2.2 Rule-based Information Extraction

Rule-based Information Extraction is a process by which structured entities are extracted from text based on rules written by human developers that capture syntactical, lexical and semantic knowledge required to identify the entities and the relationships in the domain [17]. The compositional nature of rule-based information extraction also allows rules to be expressed over previously extracted entities [26]. Figure 3 shows the schematic diagram of Rule-based IE, which has five processes and uses dictionaries or lexicons that are created by experts of a specific domain.

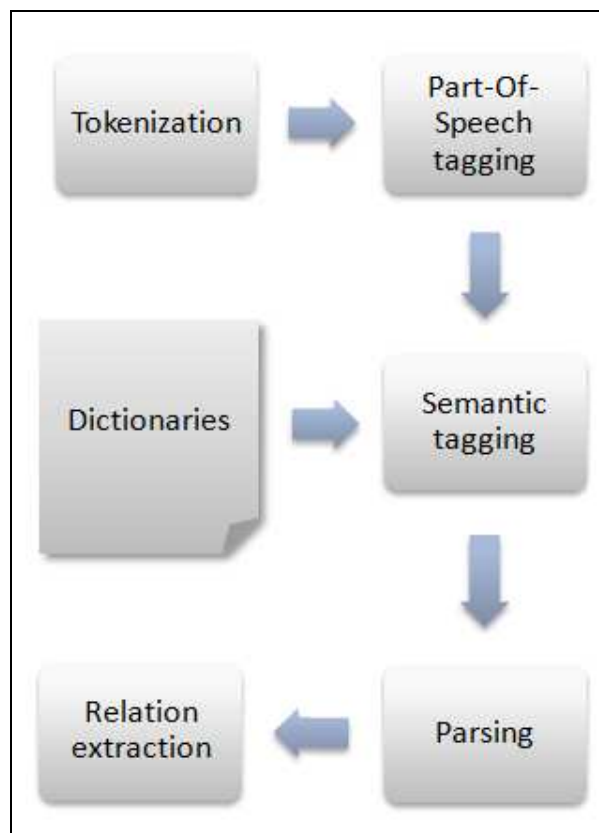


Figure 3 Rule-based IE.

- **Tokenization:** Identifying the boundaries of sentences in a document and decomposing each sentence into tokens. Tokens are obtained by splitting a sentence along a predefined set of delimiters like spaces, commas, and dots. A token is typically a word, a digit, or a punctuation.

-
- Part-Of-Speech (POS) tagging: Assigning to each word a grammatical category coming from a fixed set. The set of tags includes the conventional part of speech such as noun, verb, adjective, adverb, article, conjunct and pronoun.
 - Semantic tagging: Recognizing entities relevant to the domain. For this step it is common to use dictionaries or gazetteers that contain relevant terms and are created by experts of a specific domain.
 - Parsing: Grouping words in a sentence into prominent phrase types such as noun phrases, prepositional phrases, and verb phrases. A context free grammar is typically used to identify the structure of a sentence in terms of its constituent phrase types. The output of parsing is a parse tree that groups words into syntactic phrases.
 - Relationship extraction: The resulting parse trees from the previous step are useful in entity extraction because typically named entities are noun phrases. In relationship extraction they are useful because they provide valuable linkages between verbs and their arguments.

2.2.3 Ontology-Based Information Extraction

The detection and extraction of relevant information from textual documents depends on the proper understanding of text resources. Rule-based IE systems are limited by the rigidity and ad-hoc nature of the manually composed extraction rules. As a result, they present a very limited semantic background.

Ontology-based information extraction (OBIE) has recently emerged as a subfield of IE. Here, ontologies are used by the information extraction process and the output is generally presented through an ontology. In [27] an ontology is defined as “a formal, explicit specification of a shared conceptualization”. Conceptualization refers to an abstract model of some phenomenon in the world by having identified its relevant concepts. Explicit means that the type of identified concepts, and the constraints of their use, are explicitly defined. Formal refers to the fact that the ontology should be

machine-readable. Shared reflects the notion that an ontology captures consensual knowledge, that is, not a personal view of the target phenomenon of some particular individual, but one accepted by a group.

Ontologies capture knowledge about some domain of interest, so they are used in applications that need to process the content of information, as well as to reason about it, instead of just presenting information to humans. An ontology describes the concepts in the domain and also the relationships that hold between those concepts. Different ontology languages provide different facilities. The most recent development in standard ontology languages is OWL from the World Wide Web Consortium (W3C)³. It is based on a logical model which makes it possible for concepts to be defined as well as described. Complex concepts can therefore be built up in definitions out of simpler concepts. Furthermore, reasoners can be used to check whether or not all of the statements and definitions in the ontology are mutually consistent and can also recognize which concepts fit under which definitions.

2.2.3.1 Ontology components

Ontologies share many structural similarities, regardless of the language in which they are expressed. Most ontologies describe individuals (instances), classes (concepts), properties (attributes), and relationships.

- **Classes:** Are interpreted as sets that contain individuals. They are described using formal (mathematical) descriptions that state precisely the requirements for membership of the class. Classes may be organized into a superclass-subclass hierarchy, which is also known as a *taxonomy*.
- **Properties:** Are binary relations on individuals. There are two main types of properties in OWL: Object properties and Datatype properties. Object

³ <http://www.w3.org/TR/owl-guide/> (March 2011)

properties link an individual to an individual. Datatype properties link an individual to an XML Schema Datatype value⁴ or an RDF literal⁵.

- Relationships: Typically a relation is of a particular type that specifies in what sense the object is related to the other object in the ontology. Much of the power of ontologies comes from the ability to describe relations. Together, the set of relations describes the semantics of the domain.
- Individuals (instances): Represent concrete objects in the domain that we are interested in⁶. Strictly speaking, an ontology doesn't need to include any individuals, but one of the general purposes of an ontology is to provide a means of classifying individuals, even if those individuals are not explicitly part of the ontology.

As it can be seen, ontologies represent an ideal knowledge background in which to base text understanding and enable the extraction of relevant information. This may enable the development of more flexible and adaptive IE systems than those relying on manually composed extraction rules.

2.2.3.2 Characteristics of ontology-based IE systems

Before arriving to a definition for an OBIE system, it is useful to identify their key characteristics proposed by [18], according to diverse systems that have been discussed in the literature:

- Process unstructured or semi-structured natural language text: since OBIE is a subfield of IE, which is generally seen as a subfield of natural language processing, it is reasonable to limit the inputs to natural language text. They can be either unstructured (e.g. text files) or semi-structured (e.g. Web pages using a particular template such as pages from Wikipedia).
- Present the output using ontologies: Li and Bontcheva [28] identify the use of a formal ontology as one of the *system inputs* and the target output as an

⁴ <http://www.w3.org/TR/xmlschema-2/> for more information on XML Schema Datatypes (Last access: March 2011)

⁵ <http://www.w3.org/TR/rdf-primer/> for an introduction to Resource Description Framework (RDF) (Last access: March 2011)

⁶ Also known as the *domain of discourse*

important characteristic that distinguishes OBIE systems from IE systems. While this statement holds true for most OBIE systems, there are some OBIE systems that construct the ontology to be used through the IE process itself instead of treating it as an input (e.g. the Kylin system [29]). It is important to mention that constructing an ontology in this way does not disqualify a system from being an OBIE system.

- Use an IE process guided by an ontology: ‘Guide’ is a suitable word to describe the interaction between the ontology and the IE process in an OBIE system. In all OBIE systems, the IE process is guided by the ontology to extract things such as classes, properties and instances. This means that no new IE method is invented but an existing one is oriented to identify the components of an ontology.

An important open question is whether the information extractors, which are the components of an IE system that extract different ontological concepts, should be considered a part of the ontology or not. Several authors have argued that information extractors should be considered a part of an ontology, when linguistic rules are used as the IE technique; authors of OBIE systems that use other IE techniques, such as classification and Web-based search, generally ignore this question. But considering that the IE process of an OBIE system is guided by an ontology, both possibilities can be accepted: the information extractors may be either part of an ontology or may be outside it.

A related issue is the use of the term ‘*ontology-driven information extraction*’, which has been used in several publications [30] [31] [32]. In most cases, this can be seen as a synonym for OBIE, which has emerged due to the lack of a standard terminology. In this work, the term ontology-based information extraction is used since it appears to be the one used by a majority of publications. However, Yildiz and Miksch make a distinction between these two terms [33]. They state that in ontology-driven systems the extraction process is driven by an ontology whereas the ontology is yet another

component in an ontology-based system. This argument too is based on the view that linguistic rules should be considered a part of an ontology.

Combining these factors with the definitions of information extraction presented by [34] and [10], OBIE can be defined as a system that processes unstructured or semi-structured natural language text through a mechanism guided by ontologies to extract certain types of information and presents the output using ontologies [18].

2.2.3.3 Common architecture of OBIE systems

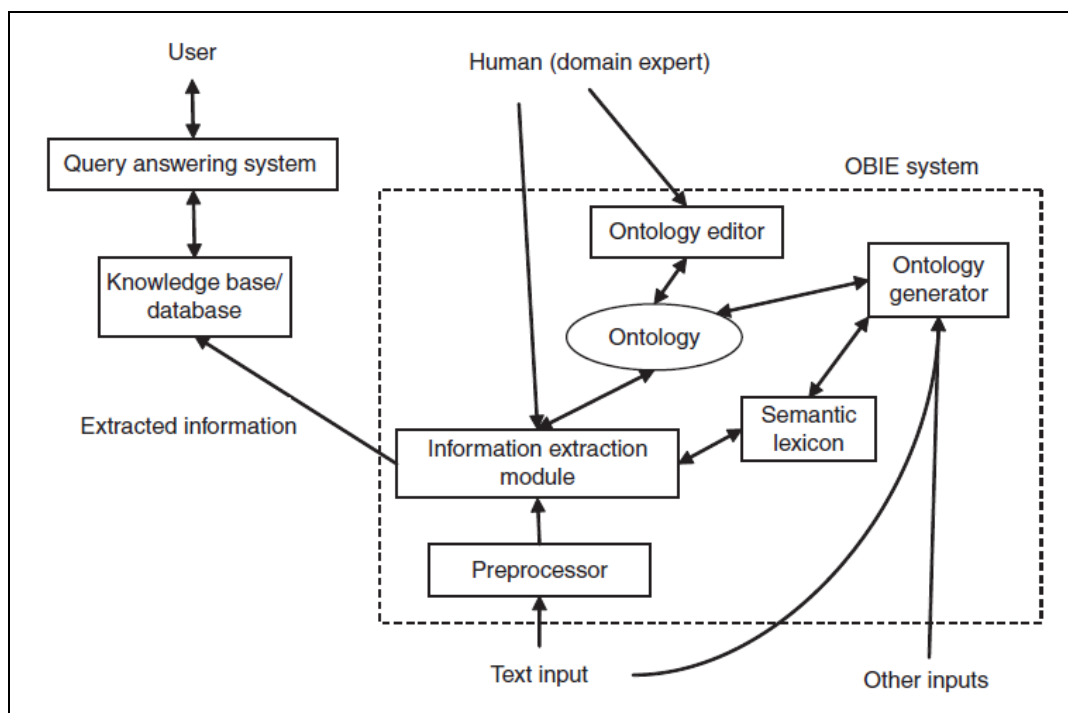


Figure 4: General architecture of an OBIE system

Figure 4 represents the union of different components found in different OBIE systems. As such, many systems do not contain all the components of this architecture, e.g. the systems that use an ontology defined by others, instead of constructing an ontology internally, do not have the 'ontology generator' component.

As represented in the figure, the textual input of an OBIE system first goes through a *preprocessor* component, which converts the text to a format that can be handled by the IE module, e.g. converting a PDF file into a pure text file.

The *information extraction module* is where the actual extraction takes places. This can be implemented using several techniques; no matter what technique is used, it is guided by an ontology. A semantic lexicon for the language is often used to support this purpose.

The ontology that is used by the system may be generated internally by an ontology generator component. This process might also make use of a semantic lexicon. In addition, humans may assist the system in the ontology generation process. This is typically done through an ontology editor such as Protégé⁷. Humans may also be involved in the information extraction process in some systems that operate in a semi-automatic way.

The output of the OBIE system consists of the information extracted from the text. It can be represented using an ontology definition language such as the Web Ontology Language (OWL). In addition, the output might also include links to text documents from which the information was extracted. Also the output of the OBIE process is often stored in a database or a knowledge base.

2.3 Information Extraction in molecular biology

IE recently has become very active in bioinformatics, since a huge amount of interesting papers have been published. Speaking specifically in the field of molecular biology, numerous efforts of Information Extraction from the scientific literature have been made. The most relevant are summarized in this section.

2.3.1 GENIES (2001)

GENomics Information Extraction System [35] is a system that extracts and structures information about pathways from the biological literature in accordance with an existing medical natural language processing system, MedLEE [36]. For GENIES the

⁷ <http://protege.stanford.edu> (Last access: June 2011)

authors used three components of the MedLEE system (preprocessor, parser, and error recovery), modified the knowledge components (lexicon and grammar) and developed a plug-in component, term tagger.

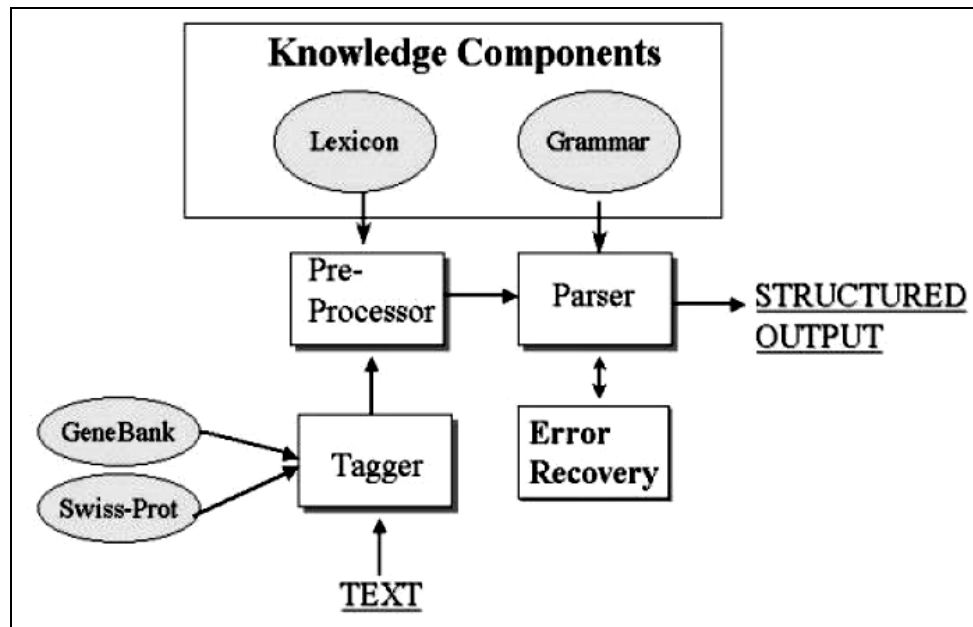


Figure 5: GENIES architecture

The system has two internal knowledge sources (a lexicon and grammar), three processing components (preprocessor, parser, and error recovery component) and a plug-in component term tagger that utilizes two external knowledge sources (Genebank⁸ and Swiss-Prot⁹).

- *Term Tagger*: This plug-in component uses specialized rules, external knowledge (GenBank and Swiss-Prot) and the BLAST [37] pattern-matching algorithm to identify and tag genes and proteins even if they are written with slight variations.
- *Preprocessor*: Is a MedLEE component that determines the sentences, words and phrases, and performs lexical lookup.

⁸ <http://www.ncbi.nlm.nih.gov/genbank> (Last access: June 2011)

⁹ <http://www.expasy.org/> (Last access: February 2011)

-
- *Parser*: Is a MedLEE component that uses a grammar consisting of semantic patterns interleaved with semantic and syntactic constraints to identify relevant relationships and to specify target output forms. For example, the following output is obtained for the sentence “*phosphorylated Cbl coprecipitated with CrkL, which was constitutively associated with the C3G*”:

[action,attach,[protein,Cbl,[state,phosphorylated]],
[protein,CrkL,[action,attach,[protein,CrkL],
[protein,C3G]]]]

- *Error Recovery*: Is a MedLEE component that uses various strategies to parse segments of a sentence.

GENIES parses complete journal articles, whereas other systems only process abstracts. Rather than extracting only binding- or enzyme related interactions, GENIES classifies and captures a complete set of interactions and relationships between biological molecules. It also handles nominalized and agentive forms of verbs of interest.

2.3.2 BioRAT (2004)

Biological Research Assistant for Text mining [38] is an IE tool designed to perform biomedical IE, which is able to locate and analyze both abstracts and full-length papers. The heart of BioRAT is an IE engine, based on the GATE toolbox (*General Architecture for Text Engineering*, [39]). GATE is a general purpose text engineering system. BioRAT uses GATE to label words according to their parts of speech, and then applies a filter that rejects determinants verbs, etc. as not being proteins. It has a molecular biology general orientation and a friendly interface. The components of GATE that BioRAT modifies are gazetteers and templates.

A gazetteer is a list of words identifying members of a particular category. BioRAT incorporates gazetteers from three sources, namely MeSH¹⁰, Swiss-Prot and hand-made lists.

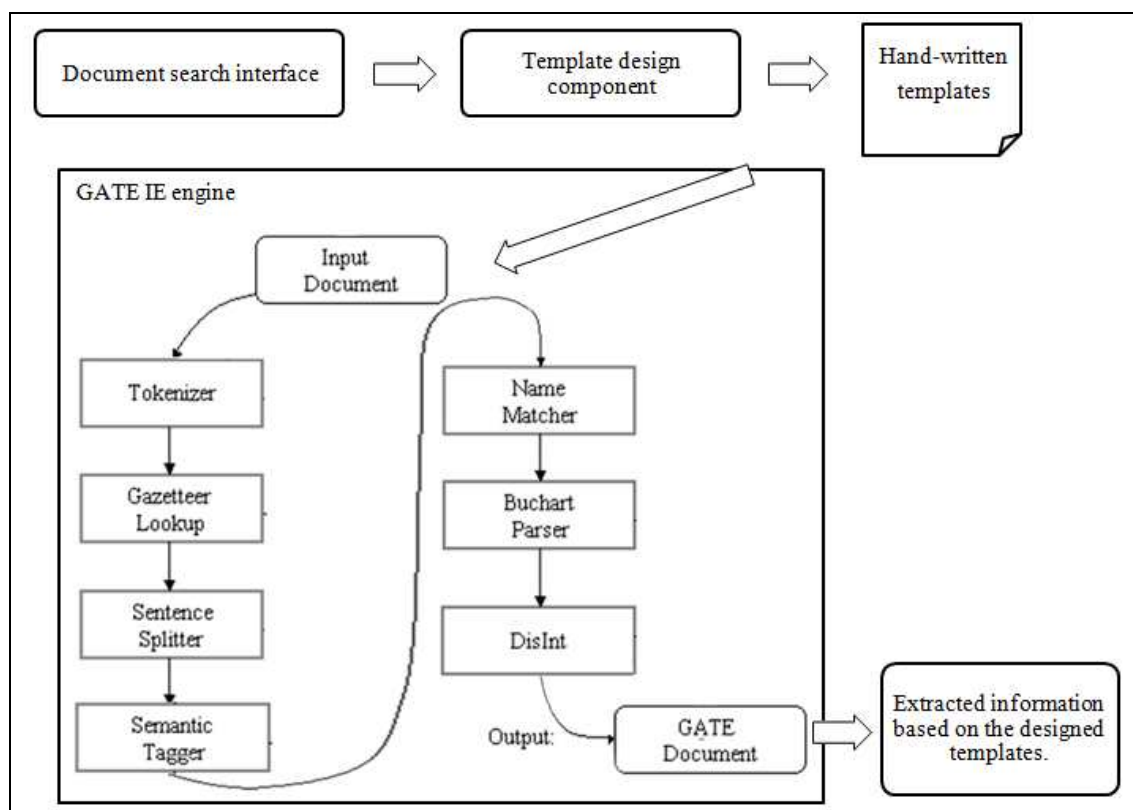


Figure 6: BioRat architecture

A template is a representation of a text pattern that allows extracting information automatically. It consists of a number of predefined slots to be filled by the system from information contained in the text, e.g.

'interaction of' (PROTEIN_1) 'and' (PROTEIN_2)

Here, 'PROTEIN_1' and 'PROTEIN_2' are slots to be filled with names of proteins, as defined by a gazetteer. Templates are written by hand; therefore, BioRAT incorporates a template design tool with a graphical user interface.

¹⁰ Medical Subject Hierarchy. <http://www.nlm.nih.gov/mesh/> (Last access: February 2011)

BioRAT produces data in XML format, which can be imported into existing database query systems. The same data are produced simultaneously as HTML for viewing in applications such as a browser, if that is more convenient for the user. Each record in the resulting database represents a single completed template.

2.3.3 BioText (2005)

It is a system that combines structural natural language processing with Machine Learning Methods to address the general and domain-specific challenges of Information Extraction targeting protein-protein interactions. BioText proposes the use of Information Retrieval (IR), Named Entity (NE) recognition, syntactic analysis, and pattern-based domain analysis [40].

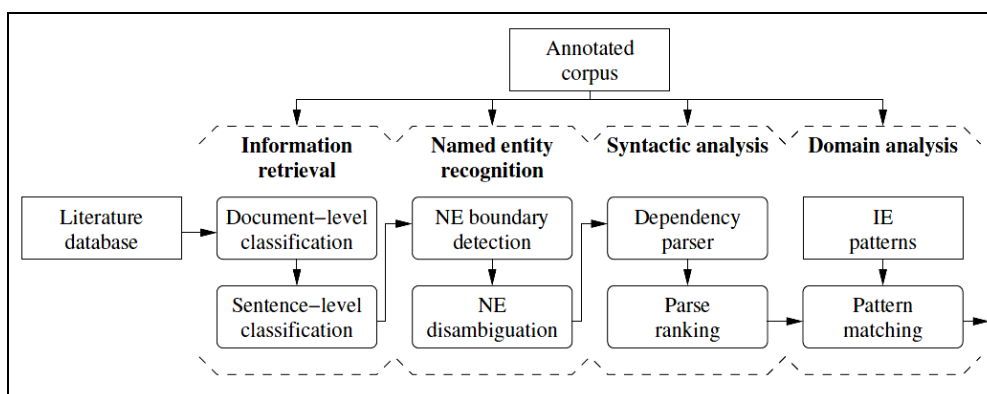


Figure 7: The IE system architecture

- *Annotated corpus*: An annotated domain-language corpus is necessary to facilitate the development and evaluation of the various parts of IE systems. The corpus was created from biomedical English resources, and it is focused on protein-protein interactions. It consists of manually annotated sentences at three levels: named entities, dependency syntax, and entity interactions. The corpus is described in detail in [41].
- *Information retrieval*: The system employs a method described in [42] applicable to the classification of PubMed-indexed articles that retrieve from publications only the sentences which are relevant to the domain. To identify individual sentences likely to discuss protein-protein interactions, a method is

used in which known protein names, verbs specific to protein-protein interactions, and their mutual positions in the sentence are used as features for a rough-set based classifier [43].

- *Named entity recognition and disambiguation*: This phase is divided into two subtasks: determining the boundaries of the NEs and classifying the entities into classes such as genes and proteins. Much of the ambiguity in biomedical text is caused by inconsistent or non-existent naming conventions. Further, capitalization and other surface clues are not reliable indicators of entities in the domain. BioText uses Machine Learning methods with a particular focus on kernel-based learning algorithms [44] to address the problem of Word Sense Disambiguation.
- *Syntactic analysis*: BioText uses the Link Grammar (LG) parser of Sleator and Temperley [45]. The LG parser is a full dependency parser with broad coverage of English. The architecture of the BioText syntactic analysis component built around LG is as follows. First, the input sentences are tokenized in a separate tokenization step. After tokenization, the parsing system is augmented with separate pre-processing and post-processing stages. In pre-processing, input sentences are simplified by replacing detected NEs with single tokens recognized by the parser; post-processing is applied after parsing to restore the original sentence text.

After post-processing, a machine-learning approach for parse ranking is applied in order to recognize the best parses among a set of alternative parses for a single sentence. This is done by applying features such as grammatical bigrams, link types (the grammatical roles assigned to the links), a combination of link length and link type, part-of-speech information and several additional attributes. Each parse is assigned a penalty based on the number of incorrect links.

- *Domain analysis*: BioText uses a set of handwritten patterns to extract factual knowledge from the parsed sentences. Each pattern specifies a substructure of

the linkage (the representation of a LG dependency parse) that is likely to state a protein-protein interaction. A successful match of a pattern in a linkage corresponds to an identified interaction.

2.3.4 String - IE (2005)

String-IE [46-47] is a rule-based system which implements cascaded finite state automata [48] to extract regulatory gene/proteins networks mainly from biomedical abstracts, but it has also been applied successfully in sets of full text journal articles.

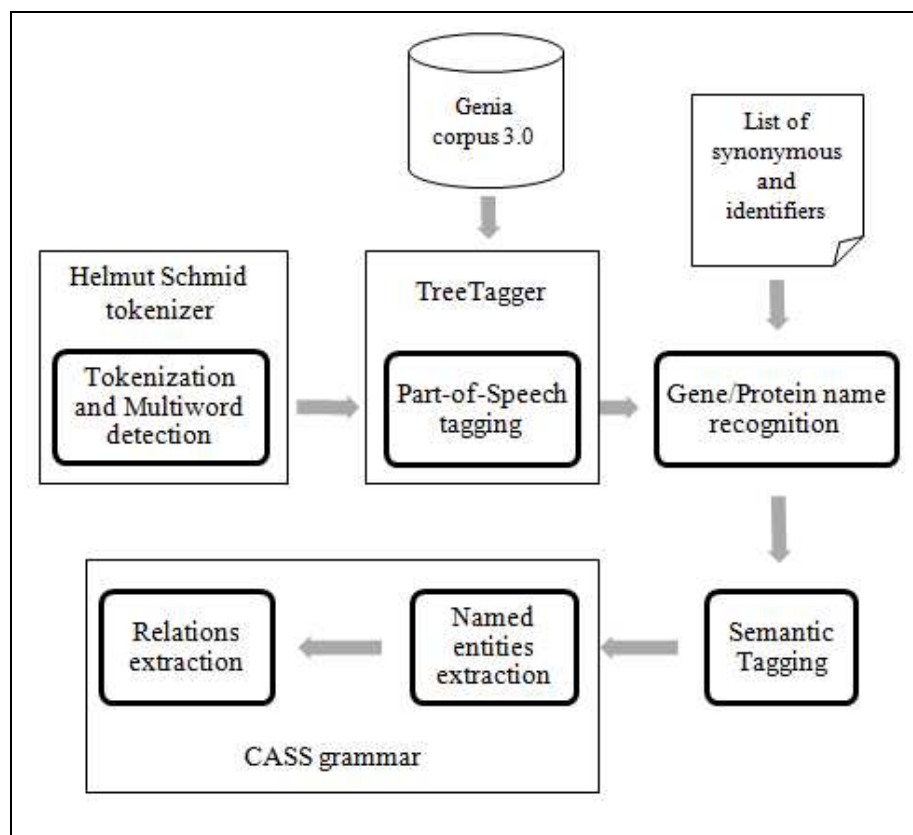


Figure 8: String-IE architecture

String-IE is organized in cascaded modules such that the output of one module is the input of the next module and a speedy partial parser. Each one of the modules is explained below.

- *Tokenization and Multiword detection*: The process of tokenization consists in the segmentation of the input text into a sequence of tokens and the detection

of sentential boundaries. String-IE uses the tokenizer developed by Helmut Schmid at IMS (University of Stuttgart) because it combines a high accuracy with unsupervised learning (i.e. no manually labelled data are needed) [49]. Multiwords were acquired semi-automatically to ensure that terms of interest are captured with high accuracy.

- *Part-of-speech tagging*: To improve the accuracy of POS-tagging on analyzed abstracts, String-IE uses TreeTagger, a tool for automatic annotation of text corpora with part-of-speech and lemma information [50], training it with the GENIA 3.0 corpus [51]. This corpus is a collection of biomedical literature which has been annotated with various levels of linguistic and semantic information.
- *Recognizing gene/protein names*: To be able to recognize gene/protein names as such, String-IE uses a list of synonymous names and identifiers in six eukaryotic model organisms compiled from several sources¹¹. This list of names contains also different orthographic variants of each name (uppercase and lowercase letters: all uppercase, all lowercase, first letter uppercase, and for multiword names, first letter of each word uppercase) and is included in the lexicon used for multiword detection and POS tagging.
- *Semantic tagging*: In addition to the recognition of the gene and protein names, this module recognizes several other terms and annotates them with semantic tags. This set of semantically relevant terms consists of nouns and verbs, and a few prepositions and adjectives. The main set of nominal terms is classified as follows:
 - Nouns representing highly relevant concepts.
 - Nouns triggering experimental or artificial contexts.
 - Enzyme names.

¹¹ <http://www.bork.embl.de/synonyms/> (Last access: February 2011)

- o Species/organism names extracted from the NCBI taxonomy of organisms [52].
- o Relational nouns, like nouns of activation, nouns of repression and nouns of regulation.
- o Verbal terms which are crucial for the extraction of relations between entities: verbs of activation, verbs of repression and verbs of regulation.
- Extraction of named entities: Once the relevant nouns are classified according to semantic criteria, the next step is to chunk noun phrases generalizing over both POS-tags and semantic tags. This syntactic-semantic chunking is performed to recognize named entities using cascades of finite state rules implemented as a CASS grammar [48]. The following simplified example shows how gene noun phrases are recognized using the CASS grammar.

```

[nx_gene
 [dt the]
 [nnp_g CYC1]
 [gene gene]
 [in in]
 [yeast Saccharomyces cerevisiae]]

```

Figure 9: CASS grammar example

- *Extraction of relations between entities*: This step of processing concerns the detection of three types of relations between the recognized named entities: up-regulation, down-regulation, and (unspecified) regulation of expression. To achieve this, syntactic properties and semantic properties of the relevant verbs are combined to map them to one of the three relation types.

For example, the following shows a bracketed structure consisting of three parts, a promoter chunk, a verbal complex chunk, and a UAS chunk (theme):

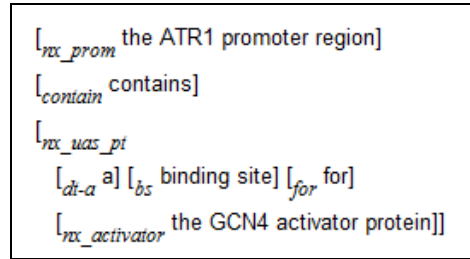


Figure 10: Recognizing types of relations

From this, String-IE extracts that the GCN4 protein *activates* the expression of the ATR1 gene.

2.3.5 GRNS (2009)

The *Gene Regulatory Network System* [53] is the most modern system of the ones presented in this chapter. It consists of four modules: the Knowledge Collection and Creation Module, the Pre-Processing Module, the Information Extraction Module, and the Post-Processing Module. The external database level includes some general database resources, such as GeneBank and Swiss-Prot, as well as some specific organism resources.

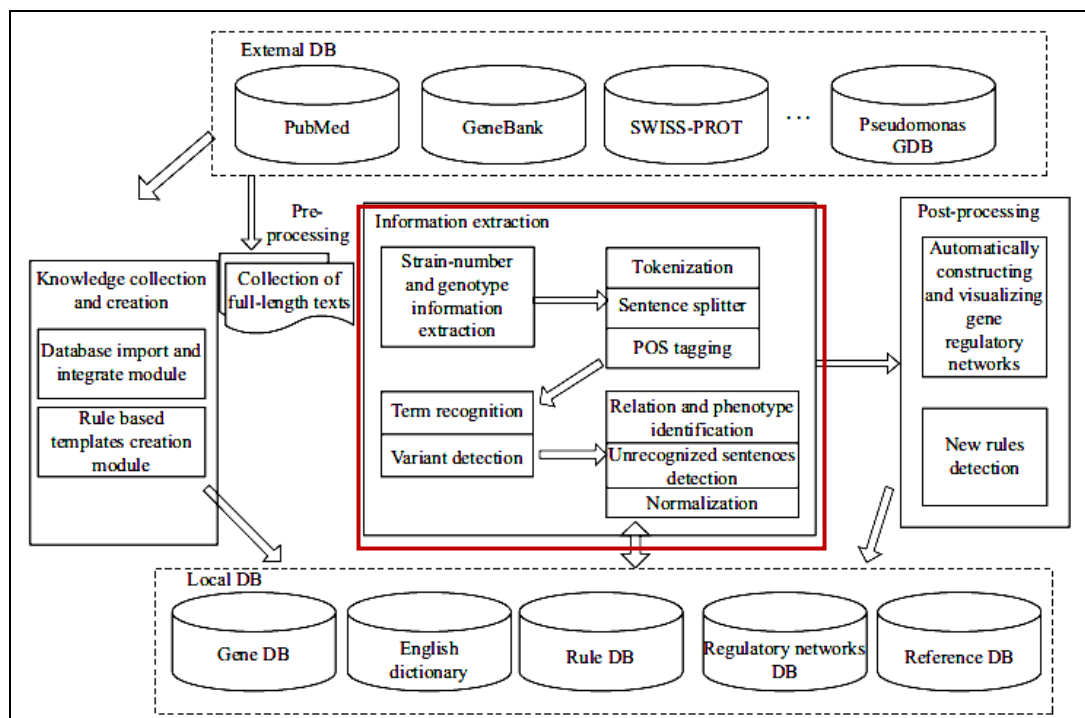


Figure 11. GRNS System Architecture

The IE Module does the tokenization, sentences splitting and Part of Speech tagging. It also recognizes the gene and the protein entities, and extracts the relations and other kinds of entities based on a rule-based approach. For the Part of Speech tagging it uses the Brill part-of-speech tagger [54]; after POS tagging, the IE module performs term recognition and variant detection to recognize gene/protein names. To deal with problems such as expanded form of abbreviation, homology and aliases a gene-dictionary of aliases and abbreviations is used. This dictionary was constructed by combining multi-database resources, such as the Swiss-Prot and the *Pseudomonas aeruginosa* Genome Database.

GNRS used the cascaded finite state automata implemented by a CASS parser [48] to recognize the gene regulatory relation and phenotype information. Figure 12 shows a CASS grammar example. Other systems such as the STRING-IE [47] also use the CASS parser to recognize the regulatory gene/protein relation.

Grammar	ex_reg- >nxp (cma?wdt?) (rb)* (cma?neg?) (vx?) (rb)* (VERB) ownexpr
Explanation	One possible gene relation is gene A regulate its own expression.

Figure 12: CASS grammar example

After the IE process, GRNS automatically constructs and visualizes, in a graphical form, the regulatory networks based on the entity and relation information extracted from biomedical literature.

2.4 Comparative analysis of the IE systems

Although the systems described above have been constructed in different ways, they all fit in a generic IE architecture. The String-IE and GRNS systems tightly follow the ruled-based architecture, thus each module of the systems represents a step in the architecture (tokenization, POS tagging, semantic tagging, parsing and relation extraction). BioRAT also realizes these steps through the GATE IE engine besides having a template design module. GENIES, on the other hand, encompasses tokenization and

POS tagging in its preprocessor module which in turn uses the term tagger module; the rest of the steps are included in the parser module. Finally, BioText carries out the IE task in a different way and order that combines rule-based IE with Machine Learning methods.

The following table summarizes the main characteristics of the five systems.

System	Domain	Domain expert supervision	Scope	IE technique	Domain knowledge sources	Output
GENIES	Molecular pathways	Yes	Full text	Rule-based	GeneBank, Swiss-Prot	Structured text file
BioRAT	Molecular biology in general	Yes	Abstracts and full text	Rule-based	Hand-made gazetteers	Table within a graphical interface
BioText	Protein – protein interactions	Yes	Sets of annotated sentences	Combined rule-based and statistical	Machine learning methods	Non specified
String-IE	Gene – proteins networks	Yes	Mostly abstracts	Rule-based	Lists of named entities	Structured text file
GRNS	Gene regulatory networks	Yes	Abstracts and full text	Rule-based	Gene name dictionary	Tables and interactive visualization of the networks

Table 1: Characteristics of biological oriented IE systems

This chapter has provided an overview of how Information Extraction is being successfully applied to molecular biology. Several systems have been presented explaining their characteristics; it is worth to notice that there are not many systems for this domain that use Ontology Based Information Extraction.

This work will contribute to apply OBIE to an existing system explained in the following chapters in order to improve it and generalize it.

3 Ontology model built from RegulonDB

It has been described what ontologies are and how they are used for Information Extraction since ontologies represent an ideal knowledge background in which to base text understanding and enable the extraction of relevant information. Their use may permit the development of more flexible and adaptive IE systems

In this work, the domain knowledge (molecular biology) is contained in the RegulonDB database, from which an ontology to guide the IE process was created.

Section 3.1 provides some relevant issues of the RegulonDB database such as its model and the description of some tables.

Details about the implemented methodology to translate the RegulonDB database structure into an ontology structure, are explained in section 3.2

Finally sections 3.3 and 3.4 describe how grammar rules and domain lexicons were also included in the ontology, so that the ontology contains all the domain knowledge needed for the developed OBIE system.

3.1 RegulonDB

As mentioned in chapter 1, RegulonDB¹² offers information about the elements and interactions of the network of transcriptional regulation in Escherichia coli K-12.

The information in RegulonDB is structured in an entity-relationship model that is composed by about thirty one tables. Figure 13 shows the part of this model which

¹² <http://regulondb.ccg.unam.mx/> (June 2011)

corresponds to the tables that store most of the information about gene regulation. The tables in orange are the ones that were used to generate the ontology.

The Gene, Terminator, Attenuator, Attenuator Terminator, Riboswitch, Shine Dalgarno, Product, Motif, Transcription Factor, Effector, Promoter and Site tables contain single elements of biological entities, whereas the Transcription Unit and Conformation tables represent a set of entities with biological meaning.

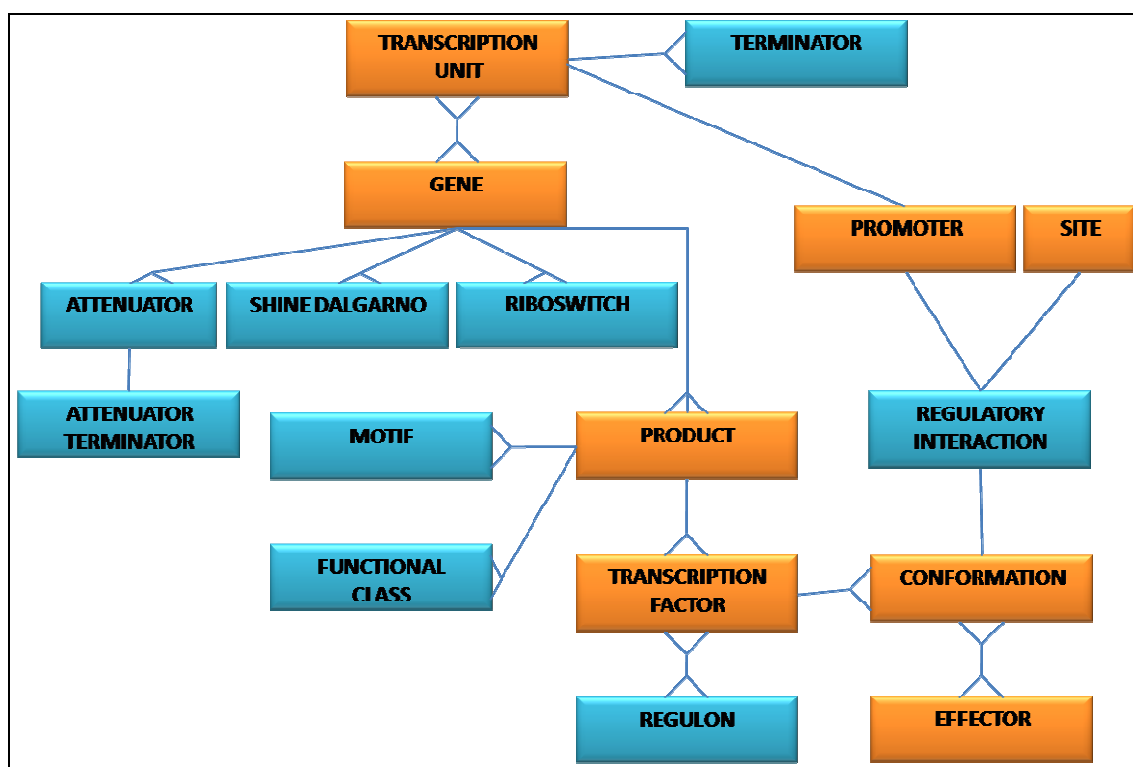


Figure 13: Part of the RegulonDB Entity-Relationship diagram

The following are the biological definitions of what each orange-colored table represents:

- Gene: Is the segment of DNA involved in producing a polypeptide chain or stable RNA; it includes regions preceding and following the coding region (leader and trailer).

-
- Promoter: Is the DNA sequence where RNA polymerase binds and initiates the transcription of a particular gene. Promoters are located near the genes they regulate.
 - Site: The binding sites are physical DNA sites recognized by transcription factors within a genome.
 - Transcription Unit: A Transcription unit (TU) is a set of one or more genes transcribed from a single promoter. A TU may also include regulatory protein binding sites affecting this promoter and a terminator.
 - Product: A product is the RNA or protein produced based on the gene template.
 - Transcription factor: A Transcription Factor (TF) or regulatory protein is a protein (more precisely a complex protein, since it can be a dimer or multimer) that activates or represses the transcription of a TU upon binding to specific DNA sites.
 - Effector: Is a molecule that binds to a protein (forming a conformation) and thereby alters the activity of that protein.
 - Conformation: A physical object whose structure is comprised of other physical objects bound to each other non-covalently, at least one of which is a macromolecule (e.g. protein).
 - Regulatory Interaction: Regulation of the expression reaction by the controlling element such as a transcription factor.

The object synonym table does not represent a biological element, but a list of synonyms of genes, products and transcription units. Synonyms are additional names given to these biological entities in other databases.

e.g. Gene name: *adhE* Synonyms: *adhC*, *ana*, *b1241*

Once the basic concepts of gene regulation have been presented, the process performed to create the ontology is explained.

3.2 Translating RegulonDB into an ontology

The translation of some of the RegulonDB tables to an ontology was an automated process implemented in Java. This section explains how the process was done, starting with the definition of classes, then assigning properties to them and finally generating individuals to populate the ontology.

3.2.1 Classes

First of all, each table (figure 14) was mapped as a class in the ontology so that nine classes were obtained. Then, some adjustments were performed in order to provide a better representation of the knowledge domain:

- Since all the classes (except Regulatory Interaction) represent biological entities or objects, they were grouped into a super class called Physical Entity.
- The Product table in RegulonDB stores two kinds of entities which are proteins and RNAs; thus, two subclasses were added to the product class.
- At the same time, the Transcription Factor class was placed as a subclass of Protein because the TFs are, in fact, a kind of proteins.

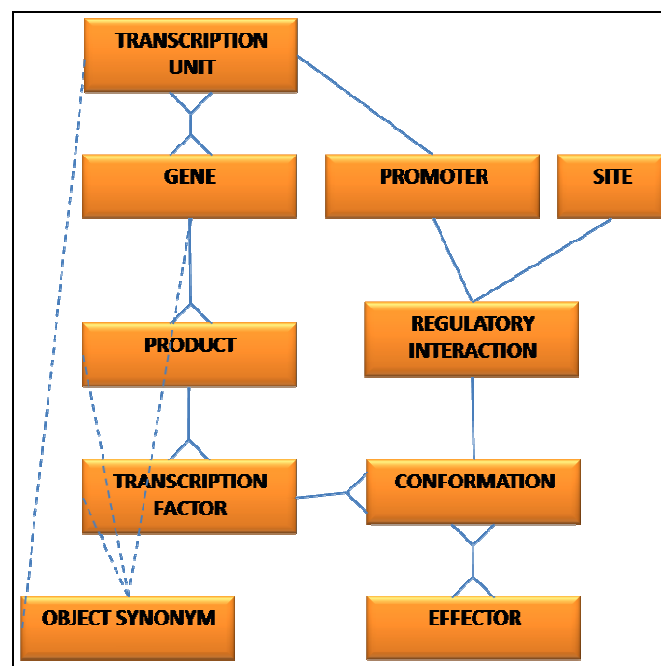


Figure 14: Tables converted into ontology classes

The final hierarchy is shown in the figure 15. The classes that are in the same level (e.g. transcription unit, complex, promoter) are set as *disjoint*, so that an individual cannot be an instance of more than one of these classes.

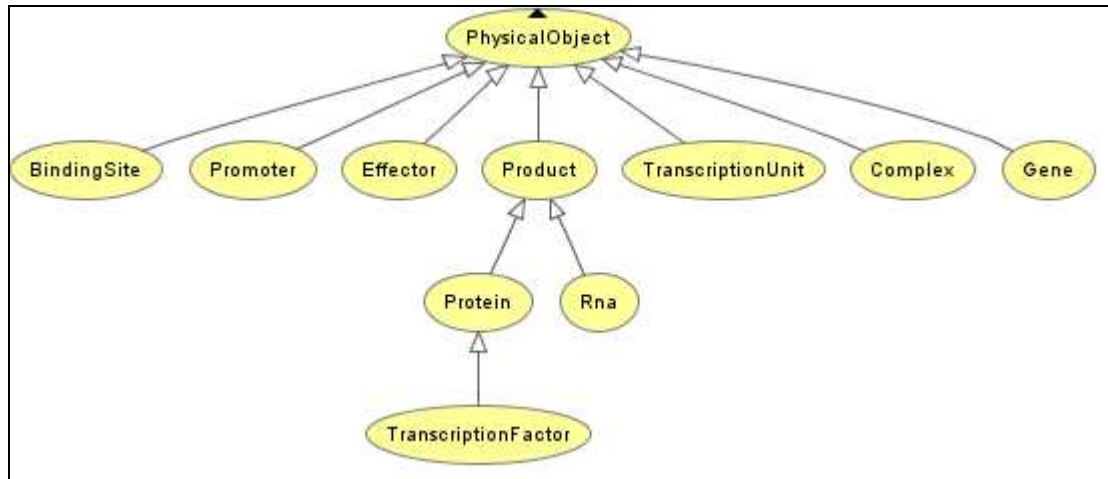


Figure 15: Class hierarchy

As can be seen in the previous figure, some class names are different to the names of the RegulonDB tables; the following table shows the relation between the names in the ontology and in the database.

Class name	RegulonDB table
BindingSite	Site
Complex	Conformation
Effector	Effector
Gene	Gene
Promoter	Promoter
Product	Product
Protein	*
RNA	*
Transcription Factor	Transcription Factor
Transcription Unit	Transcription Unit

* There are no RegulonDB tables, these entities are both stored on the Product table

Table 2: Relation between names in the ontology classes and the RegulonDB tables

3.2.2 Properties

Some of the columns from the RegulonDB tables were taken to create the ontology datatype and object properties.

The columns that were mapped were the ones that are helpful to identify the entities: *id*, *name* and *synonym*. All the others were not relevant for this work since they are related to the specific properties of the entity to which they belong.

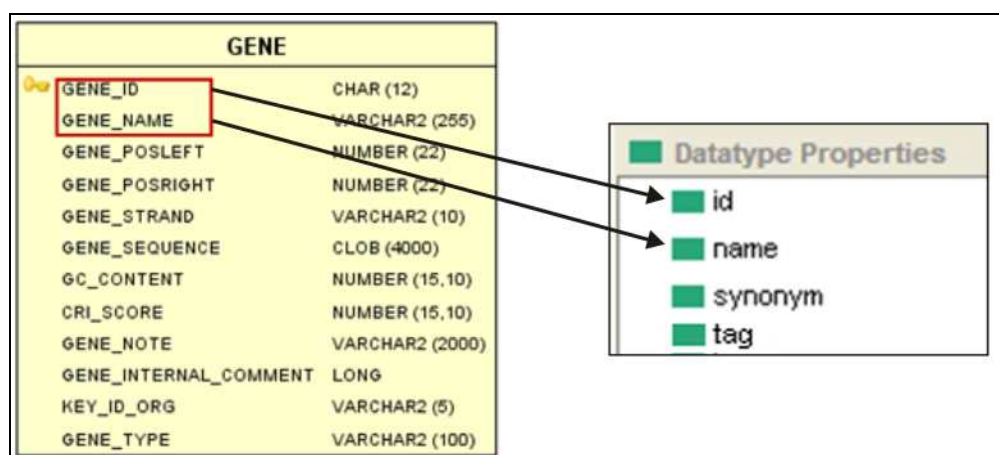


Figure 16: Mapping between columns and properties

Figure 16 shows an example of what was explained before. The left box is the description of the gene table; the fields in the red rectangle are the ones that were taken to create the properties and the others are the ones that are not relevant. The right box is the list of ontology properties.

The *synonym* property is slightly different to *name* and *id* properties, since it was not taken from a field of a table. As figure 14 shows, there is a table (Object_synonym) where all the synonyms are stored.

The values for the *tag* property are the only ones that are not obtained from the database; this property is related to a tagging process which is described in chapter 4

3.2.2.1 Property restrictions

As the name suggests, restrictions are used to constrain the individuals that belong to a class. The following are the restrictions that were applied to the existing *object properties* (see table 3).

Universal Restrictions

- hasComponent property can *only* have components from Effector or Protein classes.
- hasGene property can *only* have components from Gene class.
- hasPromoter property can *only* have components from Promoter class.

Cardinality Restrictions

- hasComponent property has a minimum cardinality restriction of 1.
- hasGene property has a minimum cardinality restriction of 1.
- hasPromoter property has a maximum cardinality restriction of 1.

In accordance with these restrictions, individuals that belong for instance, to the Complex class must fulfill the following conditions:

- Individuals can only have instances of Effector or Protein classes in their hasComponent attribute
- Individuals must have at least one instance in their hasComponent attribute.

Figure 17 shows the Protégé representation of what is stated above.

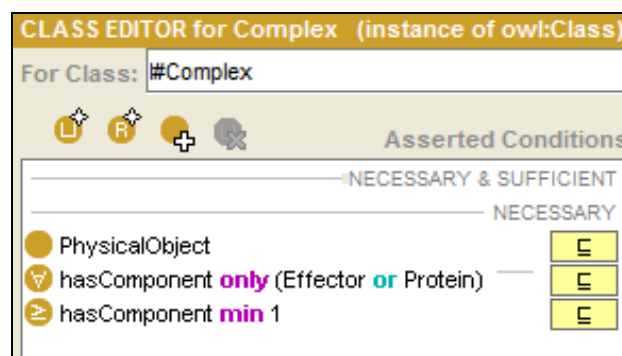


Figure 17: Conditions for Complex class

Table 3 describes the set of defined properties with its domain and ranges as well as a brief description.

Name	Type	Domain	Range	Description
id	Datatype	PhysicalObject	String	The RegulonDB id of the object
name	Datatype	PhysicalObject	String	Name of the entity
synonym	Datatype	Gene, Protein, Transcription Unit	String	Other known names for the entity
tag	Datatype	PhysicalObject	String	The word that will be assigned to the instances in the POS tagging process
hasComponent	Object	Complex	Protein, Effector	Entities that form the complex
hasGene	Object	Transcription Unit	Gene	Genes belonging to a TU
hasPromoter	Object	Transcription Unit	Promoter	Promoter of the TU

Table 3: Ontology properties

3.2.3 Individuals

The ontology was populated with the records of each RegulonDB table of figure 15. Unlike the previous steps that were made manually, this one was an automated process. It consists on connecting to the RegulonDB server and executing a series of queries to get the desired data. Then, instances of the different classes (depending on the executed query) are generated, one per each output record. Figure 18 illustrates the described process.

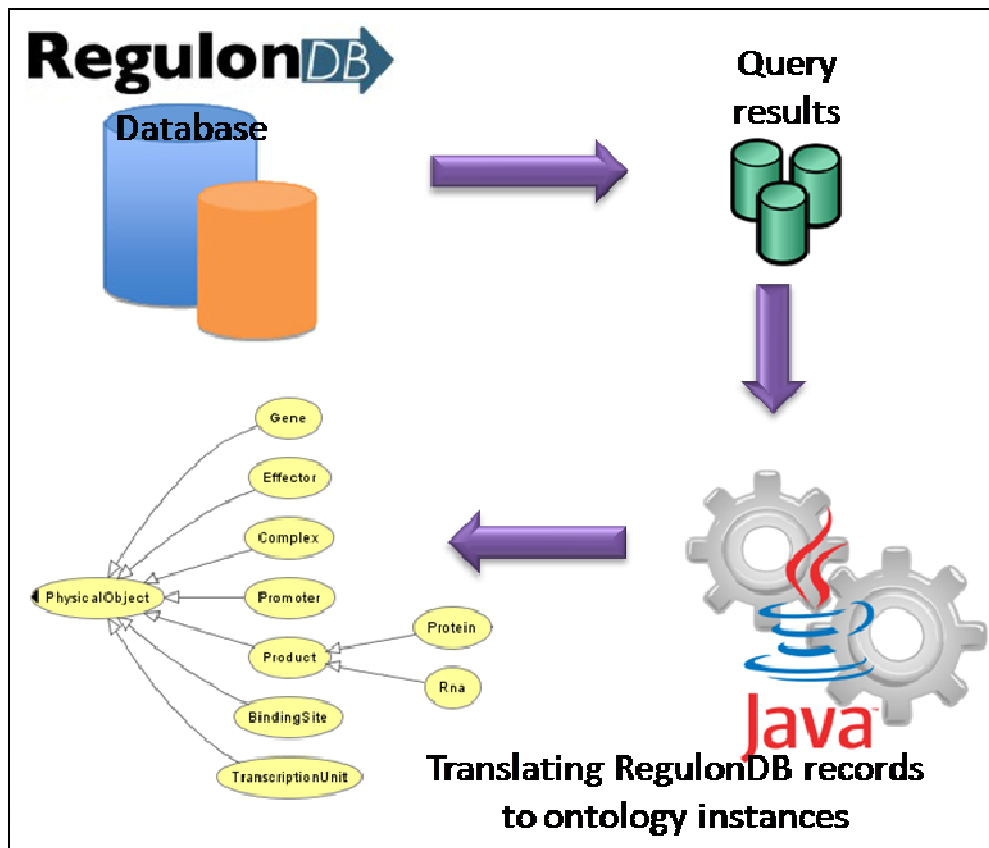


Figure 18: Process for populating the ontology

After the process was finished, the resulting ontology contained:

- 4611 Genes
- 1940 Binding Sites
- 177 Transcription Factors
- 76 Effectors
- 3409 Transcription Units
- 232 Conformations
- 4309 Proteins
- 1878 Promoters
- 167 RNAs

Figure 19 shows an illustrative example about how an individual and its synonyms retrieved from the database are converted into an ontology instance.

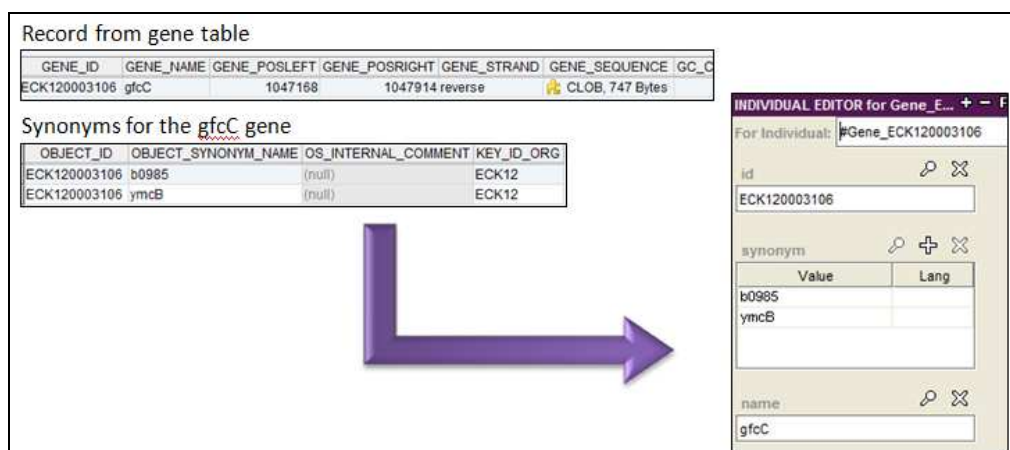


Figure 19: Example of an instance created from a query result

At the end the PhysicalObject class contains all the biological entities that will be recognized in the scientific articles. Additionally, two classes were created to store more domain knowledge words and grammar rules. In the following sections these mentioned classes are explained.

3.3 Lexicon class

The PhysicalObject class described in the previous section contains all the biological entities that will be used to recognize them in the scientific articles. The Lexicon class is a superclass placed at the level of the PhysicalObject class. It was created to store terms and tags from the dictionaries used by the CCG-IE system and thus keep the system domain independent. The Lexicon class has three subclasses: MultiUnit, Biological and ParseTerms which correspond to three dictionaries of the CCG-IE system.

3.3.1 MultiUnit class

The MultiUnit class contains biological terms that are composed of more than one word (e.g. Transcriptional activation). This class was automatically populated with the 'multi-term_lexicon_complete.txt' gazetteer. The gazetteer consists of a list of 4053 terms, one per line. Each term became an ontology individual, giving a total of 4053 instances. Figure 20 shows an example of an individual containing the multi unit term "stable interaction".

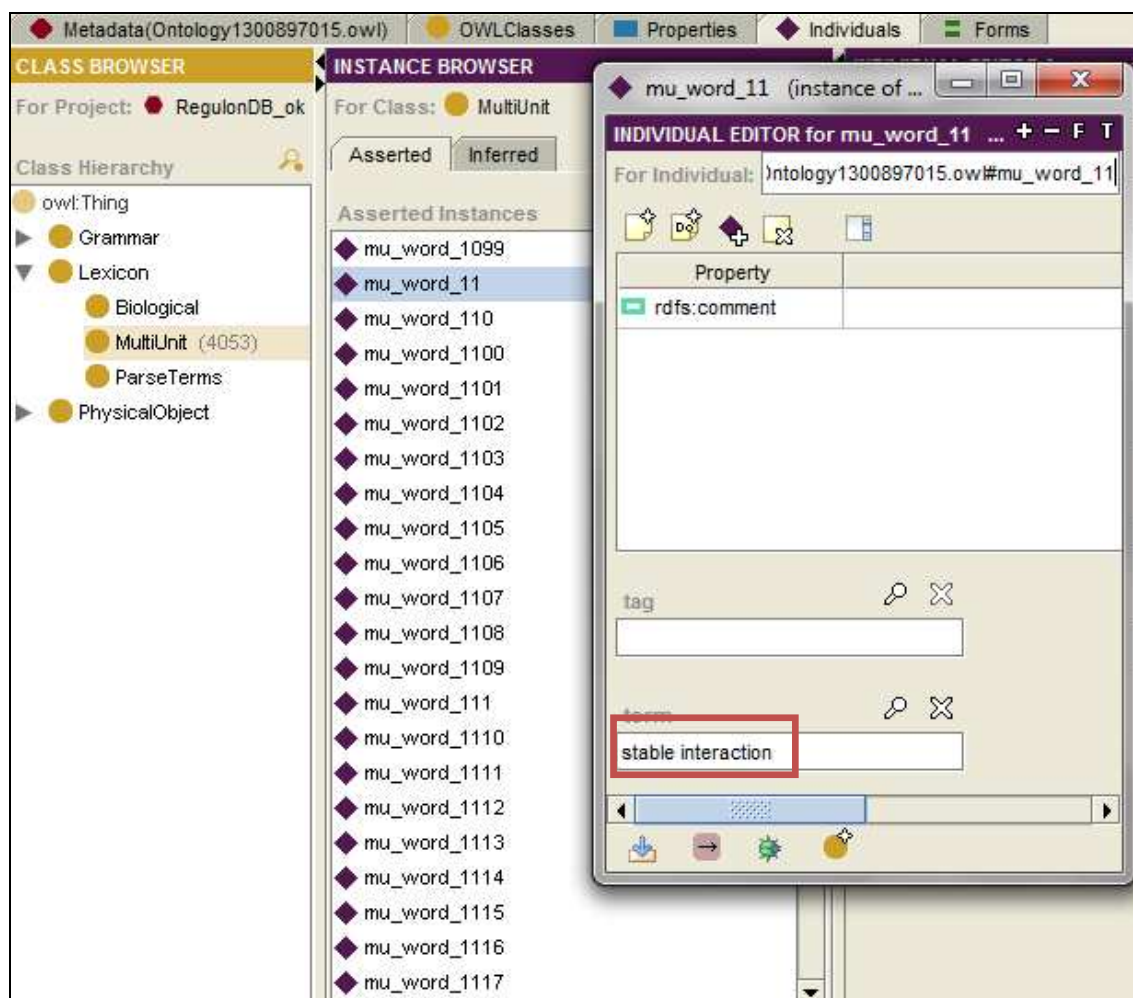


Figure 20: Example of a multi unit instance

3.3.2 Biological class

The Biological class is slightly different to the MultiUnit class. It contains biological terms that, even though they are not physical objects, it is important to identify them as knowledge domain (e.g. regulons, phosphorylated). The class has two properties:

- tag: The label that will be assigned to the instances in the POS tagging process
- term: The biological word.

'dic_for_NewRetag.txt' was the dictionary that was used to automatically populate this class; however, it also contained names for genes, proteins and other biological entities giving a total of 57497 words. All the biological entities names were not needed anymore since the new approach was to take these names directly from the

RegulonDB database, so the dictionary was manually depurated and all the names for entities were eliminated, and in the end 31689 ontology instances were created. Figure 21 illustrates an example of a Biological class individual and how the used dictionary looks like.

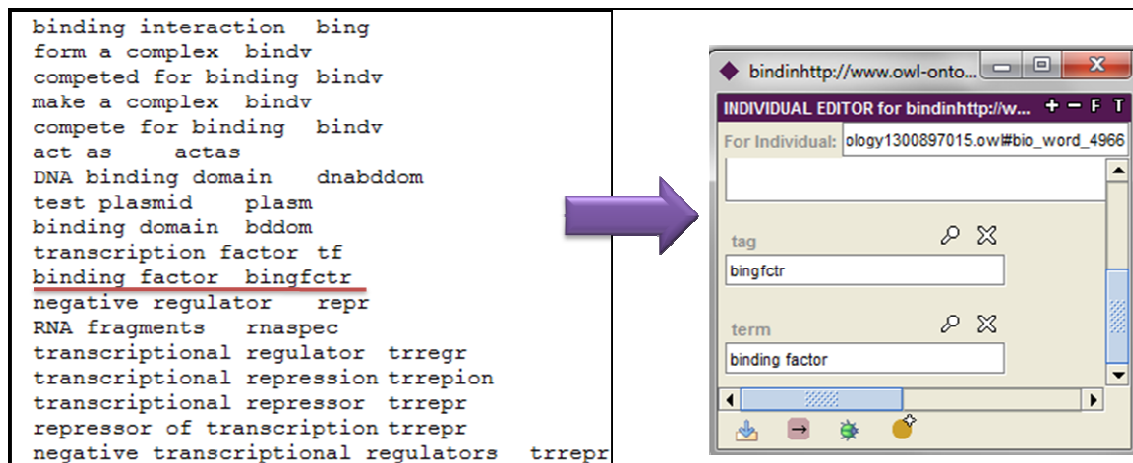


Figure 21: Example of an instance of Biological class

3.3.3 ParseTerms class

The ParseTerms class stores tags needed by the parser to map from a standard tagset to the grammar-specific tagset. The class has three properties:

- previousTag: Tag that was assigned in the semantic tagging process.
- term: A word.
- tag: New tag to be assigned.

Section 4.2.4.1 in chapter 4 explains in detail how the values of this properties are used and what they mean.

The class was automatically populated with the content of the 'other_plus_final.fx' dictionary; however, before the population the dictionary was also depurated because it contained names for biological entities and they were already stored in the ontology in the PhysicalObject class.

3.4 Grammar class

The Grammar class is located in the same level of Lexicon and PhysicalObject. This class contains grammar rules focused on the expression of biological concepts. These rules are classified into a sequence of levels since it is a requirement of the parsing program (Figure 22). In the ontology each level is represented as a subclass of the Grammar class, giving a total of 46 subclasses (46 levels) with their instances which are the grammar rules.

```
:Entities1
cdqlx -> as much as | more than | about | only | well? over;
cdx -> (cdql|cdqlx) cd | (cdql|cdqlx)? cd+ cd;
mx -> (cd|cdx) h=(units|tunits) | (cdql|cdqlx)? dt-a h=(unit|tunit);
ccx -> cma? cc cc+;
cc -> ccx;
nnpqx -> ((cd | sym)* h=nnpq (cd | sym)*)+;
uasx -> uas cd?;
ursx -> urs cd?;
bevvn -> (be|bez|ber|bedz|bedr) vvn;
vvnby -> vvn by;
aasequ -> aa aa+;
:Entities2a
DET = dt | dtp | prps | (cdql | cdqlx)? (dt-a | dt-q | dtp-q);
ADVHD = (rb cc?)+ | cdql | then | well;
ADV = ADVHD | rbr | more | rbs | ql;
NUM = ed | edx;
JX = ADV? (jj | jjr | jjs);
JXC = JX (cma JX)* (cc | cma) JX;
ADJ = JX | JXC | mx;
PTC = ADV* (vbn | vbg);
MOD = (org | ADJ | PTC | vvn | cd | sym);
MODX = (MOD | (nn | nns | np | nps)+ (dependent | independent | responsive)?);
PRE = (NUM ((cc | of | to) NUM)?)? such? DET? ((MODX cma)? MODX)?;
POST = ((from | in | of)? (NUM ((cc | of | to) NUM)?)? such? DET? MODX)+;
nxpg -> PRE (nnpqx cc)? nnpqx POST | PRE nnpqx (cma (nnpqx | nn))+ cma? cc (nnpqx | nn) POST;
nxdep -> PRE nxpg (dependent | independent | responsive) POST;
nxneg -> PRE rb? neg nxpg ((cc? PRE nxpg POST \))? POST;
:Entities2b
DET = dt | dtp | prps | (cdql | cdqlx)? (dt-a | dt-q | dtp-q);
ADVHD = (rb cc?)+ | cdql | then | well;
ADV = ADVHD | rbr | more | rbs | ql;
```

Figure 22: anaphora_grammar.reg file, grammar rules

The Grammar class was automatically populated with the rules from the ‘anaphora_grammar.reg’ file. In this particular case, the order in which the rules are declared is very important; the properties for the class are stated as follows:

- term: The name of the rule or variable (e.g. cdqlx)
- description: Regular expression defining the rule.
- order: Continuous enumeration to know the place of the rule within the level

Figure 23 gives an example of how a grammar rule looks like as an ontology instance. It can be interpreted as:

“Rule number 1 (order property value) of Level1 (subclass name) is cdqix (term property value) -> as much as | more than | about | only | well? over; (description property value)”

Section 4.2.4.2 explains in detail the structure of the grammar.

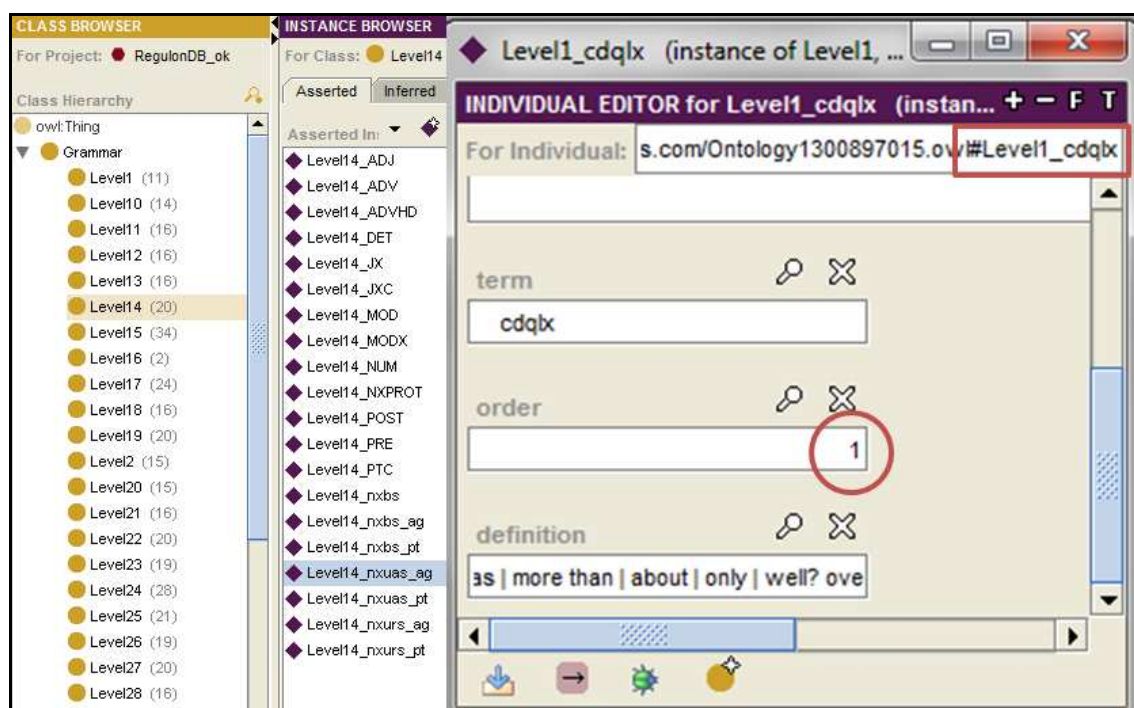


Figure 23: Example of an instance of Level1 class (Grammar subclass)

3.5 Complete ontology

In this way, the dictionaries of the CCG-IE system are included in the ontology together with the instances that participate in regulatory networks taken from RegulonDB, therefore all the domain information needed for the IE process is stored in the ontology as a whole. The resulting ontology diagram is shown on figure 24.

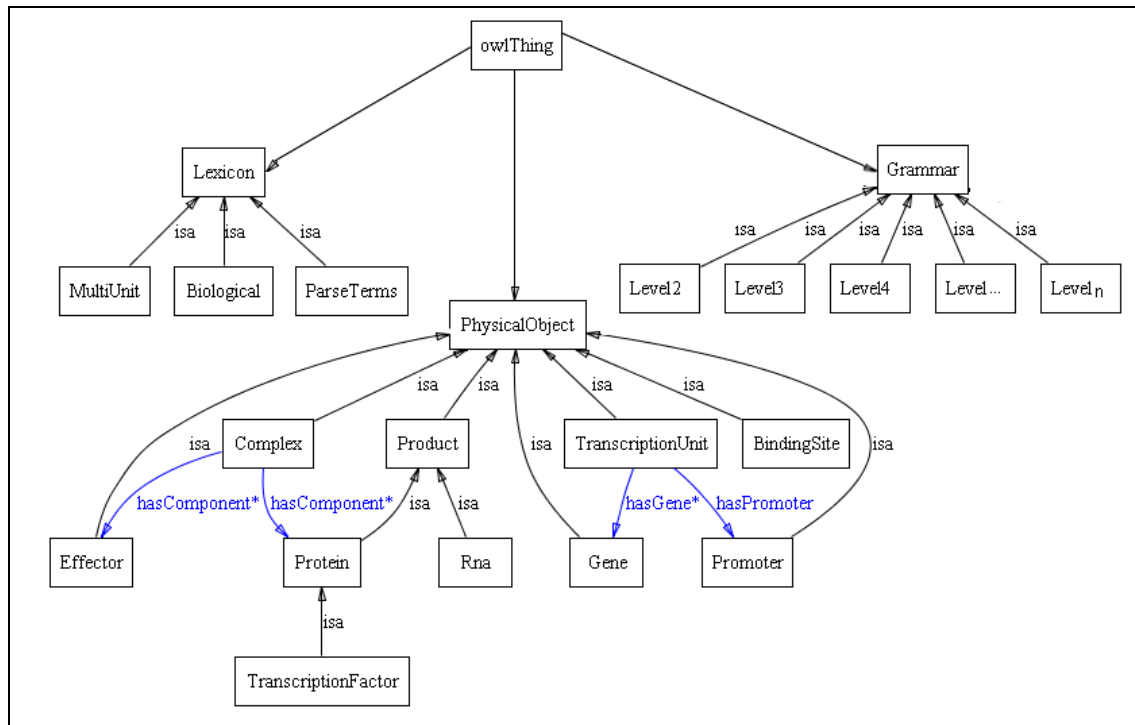


Figure 24: Final structure of the ontology

At this point, all the domain knowledge has been separated from the system and included in an ontology. The next task to generalize the system is to adapt it to this new generated ontology. This topic is addressed in the following chapter.

4 Improving and generalizing the CCG-IE system

As mentioned in chapter 1, the CCG has started to develop processes that allow the automation of the curation process, so that information about the interactions of the regulatory network can be automatically extracted by analyzing scientific articles. The intention of this work is to improve these processes by implementing an OBIE approach, and generalize the system in order to make it applicable to other domains of knowledge.

In section 4.1 an overview of the current state of the IE system developed by the CCG (CCG-IE) is provided. Section 4.2 explains in detail the different modules of the system as well as the changes that were implemented in order to improve and generalize the system.

4.1 CCG-IE system architecture

The CCG-IE system was implemented under the rule-based approach, since in general it is more focused in accuracy (precision) than on coverage (recall), and rule-based systems are more accurate for well-defined tasks (although they are very labor-intensive to implement), while statistically-based approaches, although less exact, are more robust and tolerant to noisy data and errors.

The CCG-IE system creates computer-readable networks of regulatory interactions directly from different collections of abstracts and full-text papers. It was developed by customizing and extending the STRING-IE [47] system for *E. coli* K-12 literature. The basic processing pipeline is shown in Figure 25.

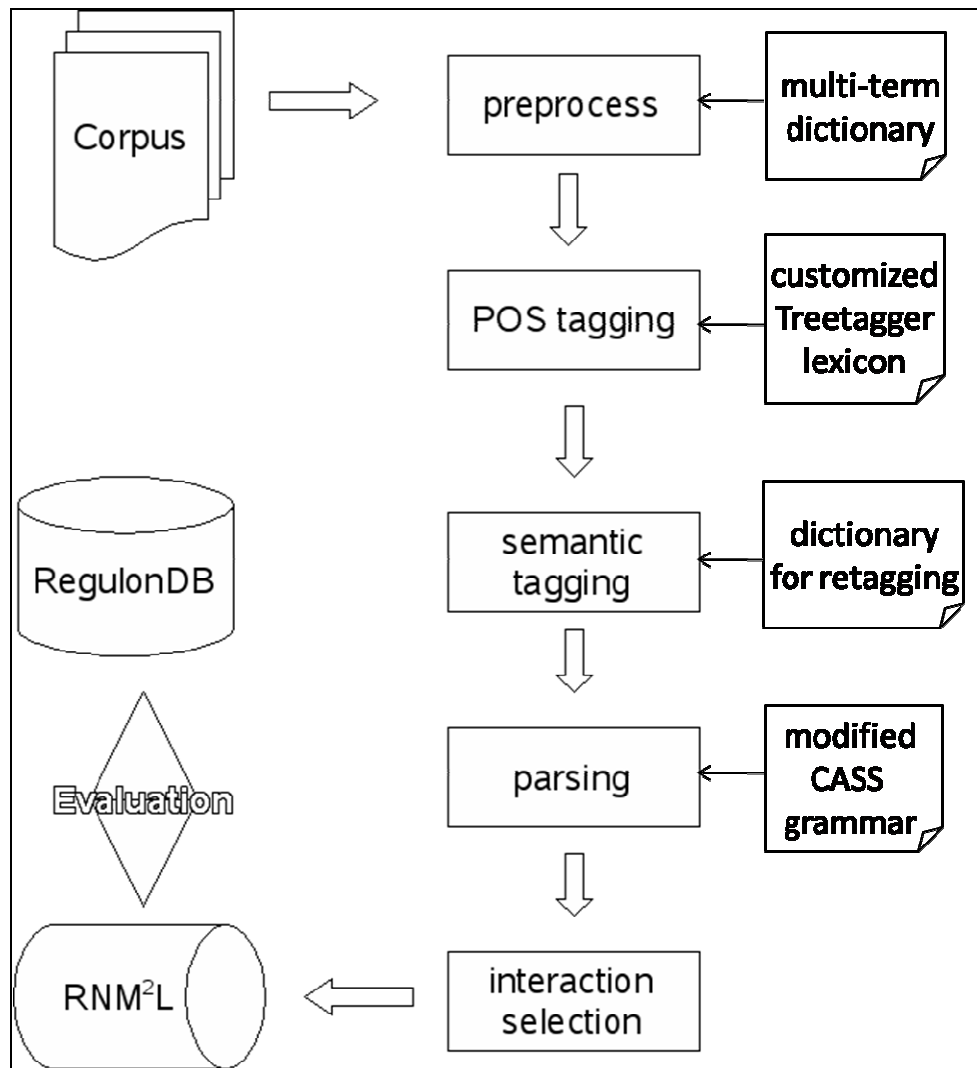


Figure 25: Processing pipeline of CCG-IE system

The system input consists of a set of articles in plain text. The phases of the system are the acquisition of the input corpus, the preprocessing, analysis of lexical components (POS tagging), a primary semantic analysis (semantic tagging, linking semantic tags to the components of a sentence), full syntactic analysis (parsing) and a final step which consists in interactions selection. Although each step will be explained in detail in the following subsections, a brief outline of each one is given in the following list of items:

- **Preprocess:** This module is divided in two steps: first the input corpus is normalized and tokenized, separating all words and terms, dealing with abbreviations and punctuation and identifying sentential boundaries. Second,

multi-unit terms (stored in a dictionary) are recognized and placed on the same line for tagging purposes.

- POS tagging: The part-of-speech of each word is tagged using a customized version of Treetagger [50]. This is made by using a file which contains a fullform lexicon in which each line corresponds to one word form and contains the word form itself followed by a Tab character and a sequence of tag-lemma pairs.
- Semantic tagging: This is a retagging module which substitutes some of the POS tags for more semantically oriented labels, such as org (organism), actv (activation verb), etc. This Named-Entity Recognition task uses a dictionary of biological terms and their corresponding tags.
- Parsing: The resulting output file from the semantic tagging module is fed into the Abney's CASS/SCOL parser [48] that uses the CASS cascading grammar parser to generate a tree-like structure by applying a grammar focused on the expression of biological concepts. This markup allows for the recognition of biological entities and processes in relationships that can be inferred from the grammatical structure of the linguistic phrase. The core CASS grammars that this system uses was adapted from the one developed for the STRING-IE system [55] for transcriptional regulation.
- Interaction selection: For the interaction selection step, the resulting partially-parsed sentences are converted into an intermediate XML format and processed by customized heuristic modules that:
 - a) Identify the regulatory interactions that are to be extracted.
 - b) Identify, when possible, the kind of the interaction extracted (activation or repression).
 - c) Create an XML output file with a regulatory network retrieved from the processed raw text.

The CCG-IE system was developed in python and uses two external programs: TreeTagger¹³ and Steve Abney's SCOL¹⁴ chunk parser. It runs only on UNIX/Linux platforms because the SCOL parser is not a multiplatform software.

4.2 System modules: explanation and improvements

In IE for molecular biology most of the developed systems are rule-based, others combine statistical methods and rules, but very few are ontology-based. In Chapter 3 the process to create and populate the ontology with domain knowledge was explained, making the ontology the only knowledge source for the system.

This section describes how the CCG-IE system was modified into an OBIE system, because in this way it will be improved and generalized. Some of the advantages obtained by doing these modifications are: the system becomes domain independent; it is possible to use it in other contexts by changing the ontology, there is no need to modify the code. The individual modules can also be reused as separate processes.

The programming language chosen for the implementation of this OBIE approach was Java, although some python scripts of the original system are also used. Details of the implementation and how the system works are explained module by module on the following sections.

4.2.1 Preprocess

As mentioned before, the first step of this module is to normalize and tokenize the input plain text file.

Normalization is a process by which text is transformed to make it consistent in a way which it might not have been before. Some examples of text normalization are: removing punctuation, removing accent marks and other diacritics from letters,

¹³ <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/> (June 2011)

¹⁴ <http://www.vinartus.net/spa/> (June 2011)

identifying and expanding contractions, etc. This is done for 'cleaning up' the text and to make easier the processing in the following modules.

Tokenization is the process of demarcating and possibly classifying sections of a string of input characters. As an example, consider the following normalized line taken from an input plain text file:

In vivo binding of the OmpR protein to the ompF and ompC promoter regions was observed using an in vivo dimethyl sulfate DNA Footprinting technique.

The string isn't implicitly segmented on spaces, as an English speaker would do. The raw input, the 148 characters, must be explicitly split into the 25 tokens with a given space delimiter (i.e. matching the string " "). This is represented as:

```
<s pmid='2656704.txt'>  
In  
vivo  
binding  
of  
the  
OmpR  
protein  
to  
the  
ompF  
and  
ompC  
promoter  
regions  
was  
observed  
using  
an  
in  
vivo  
dimethyl  
sulfate  
DNA  
footprinting  
technique  
.  
<\s>
```

This is done to the whole plain text input so that the resulting file is divided by sentences and contains one word or punctuation per line. Since normalization and tokenization are already generic processes, this step remains unmodified.

The second step of this preprocess module is more domain-oriented and consists of recognizing specific biological multi-unit terms, this is, terms that have more than one word (e.g. DNA binding domain, negative regulation, inner membrane protein, etc) and place them in the same line of the file for tagging purposes.

Originally, the CCG-IE system used a dictionary with all the existing multi-unit terms to recognize them in the text file. But since the purpose of this work is to generalize the system and implement an ontology-based approach, this dictionary was mapped into a class (MultiUnit) and automatically populated with its corresponding instances (each multi-unit term) as explained in chapter three, section 3.3.1.

The recognition of the multi-unit terms is performed by a perl script of the TreeTagger software which is called from the developed Java program. TreeTagger is a tool for automatic annotation of text corpora with part-of-speech and lemma information.

To perform the recognition task, the script needs as input a file with the list of terms to be identified. So before executing the TreeTagger script, the Java program generates the mentioned file by navigating through the ontology, looking for the MultiUnit class, then getting all its instances and setting them in the appropriate TreeTagger format. (Figure 26)

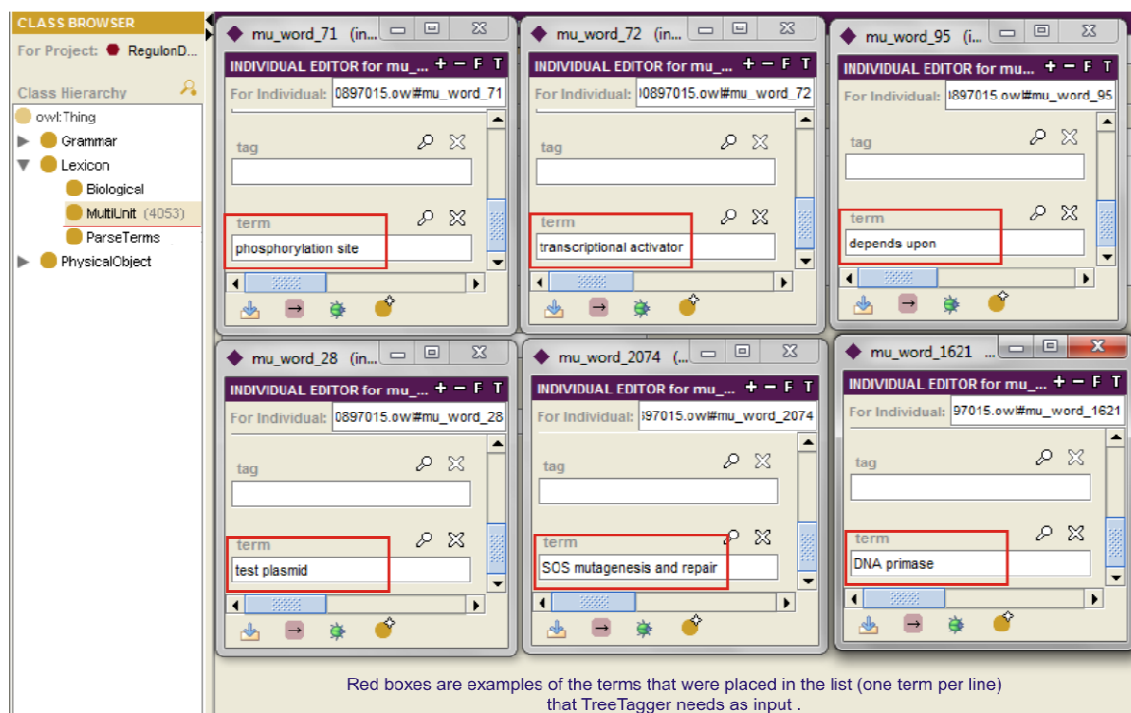


Figure 26: Example of multi-unit terms

After executing the two tasks described before, the final output is a tokenized file (.tok) divided by sentences and with one word per line, except for the multi-unit terms. A section of the output file is shown on Figure 27.

```

<s pmid='2656704.txt'>
In vivo
binding      Multi-unit term
of
the
OmpR
protein
to
the
ompF
and
ompC
promoter
regions
was
observed
using
an
in vivo
dimethyl
sulfate
DNA
footprinting
technique
.
</s>

```

One word per line

Figure 27: Preprocess output

Once that preprocess is done, the control returns to the Java program to continue with the POS tagging module.

4.2.2 POS tagging

As explained in chapter 2, the part-of-speech tagging is the process of assigning a part-of-speech like noun, verb, pronoun, preposition, adverb, adjective or other lexical class marker to each word in a sentence. The input of a tagging algorithm is a *string of words* of a natural language sentence and a specified *tagset* (a finite list of Part-of-speech tags).

In the CCG-IE system, this module is performed by the TreeTagger software which uses the Penn Treebank tagset. The following table shows a list of parts of speech with their corresponding tag.

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NP	Proper noun singular
NPS	Proper noun plural
PDT	Predeterminer
POS	Possessive ending
PP	Personal pronoun
PP\$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	To
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, noun-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP\$	Possessive wh-pronoun
WRB	Wh-adverb

Table 4: Penn Treebank tagset

To do the POS tagging, the TreeTagger software needs as input the file to be tagged and a so called ‘parameter file’ that is a compiled version of the lexicon to be used. The lexicon consists of a file in which each line contains a word form, a tab character and a sequence of tag/lemma-pairs separated by blanks.

e.g.

<i>aback</i>	<i>RB</i>	<i>aback</i>
<i>abacuses</i>	<i>NNS</i>	<i>abacus</i>
<i>abandon</i>	<i>VB</i>	<i>abandon</i>
<i>abandoned</i>	<i>JJ</i>	<i>abandoned</i>
<i>abandoning</i>	<i>VBG</i>	<i>abandon</i>

The parameter file that the CCG-IE uses is a customized version, which includes biological terms. However, what is intended is to generalize the system and therefore to separate the knowledge domain from the program itself. So the customized parameter file was replaced by a general English language parameter file which uses the English morphological database¹⁵. Thus this process is not linked anymore to a specific domain, but uses general English words.

The output of this module file with a single POS tag for each word, Figure 28 shows a comparison between the POS tag outputs using the CCG-IE customized file and the English language parameter file.

¹⁵ <http://www.cis.upenn.edu/~xtag/> (June 2011)

Customized parameter file	English parameter file
<pre> <s pmid='2656704.txt'> Identification NN of IN the DT DNA-binding NN Domain NN of IN the DT OmpR NN Protein NN Required VVN for IN Transcriptional Activation NN of IN the DT ompF NN and CC ompC NN Genes NPS of IN Escherichia coli NN by IN in IN VivoDNA NN Footprinting NN " " ((Received VVN for IN publication NN ' December NP 8 CD </pre>	<pre> <s pmid='2656704.txt'> Identification NN of IN the DT DNA-binding NP Domain NP of IN the DT OmpR NP Protein NP Required NP for IN Transcriptional Activation NP of IN the DT ompF NN and CC ompC NN Genes NPS of IN Escherichia coli NP by IN in IN VivoDNA NP Footprinting NP " " ((Received NP for IN publication NN ' December NP 8 CD </pre>

Figure 28: POS tag output using different parameter files

As can be seen, the non domain specific words were tagged with the same tag in both cases unlike the biological words; however, since the following process makes a semantic tagging, the results using the English file are good enough for this module.

4.2.3 Semantic tagging

The purpose of this module is to recognize the entities that are relevant to the domain. To carry out this process, the CCG-IE system uses a dictionary created by the domain experts which contains relevant terms; this is, a list of biological terms including names for genes, proteins, and other physical entities, also verbs and other nouns.

To keep the system as general as possible, the domain-oriented dictionary that the original system uses for this process was mapped to the ontology as explained in chapter 3 section 3.3.2.

To perform the semantic tagging process, the Java program gets the names of all the biological objects that take part in a regulatory network; in other words, all the ontology instances (including their synonyms) of the different subclasses of the PhysicalObject class (figure 29). This gives as a result a total of 16799 biological entities that together with its synonyms form a part of the domain lexicon and will be used to assign the semantic tag (e.g. *nnpg*).

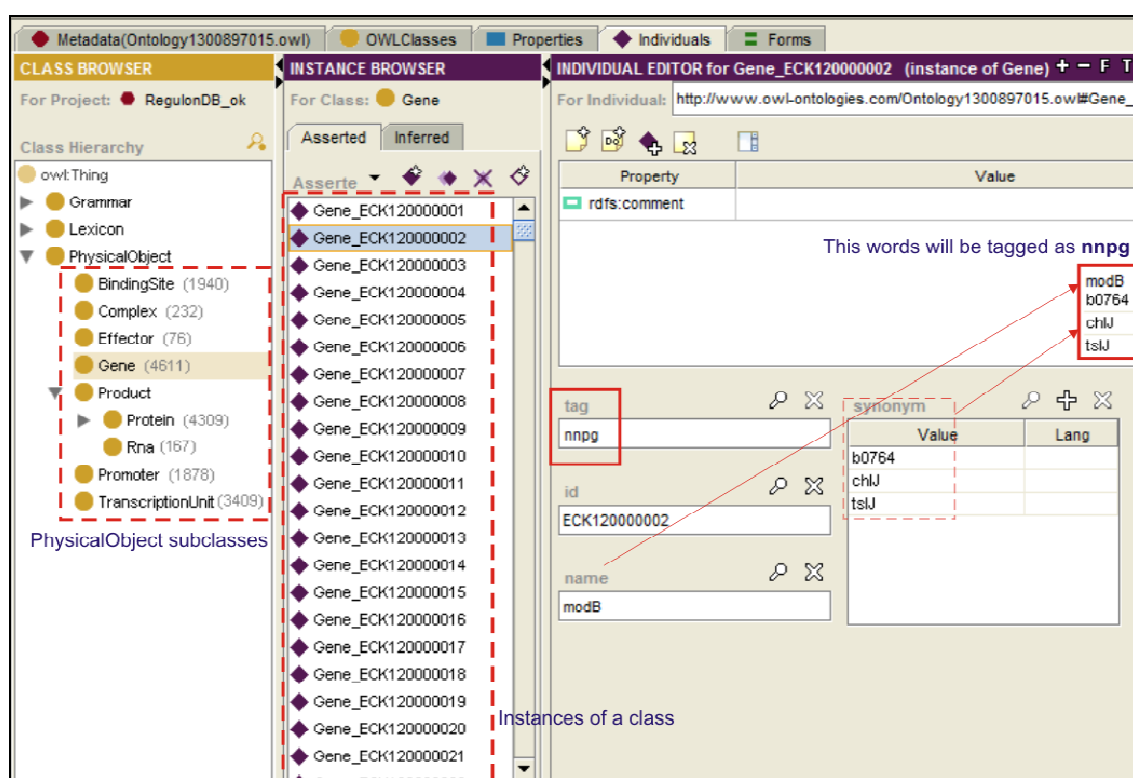


Figure 29: Instances of the PhysicalObject subclasses

A similar process is performed with the 31689 instances of the Biological subclass; as explained in section 3.3.2 this class contains the other part of the domain lexicon; that is, biologically- oriented words that are not names of biological entities.

The instances of all the physical objects and the ones of biological class form a domain lexicon of 48488 terms with an associated tag. This lexicon is fed into a script that in

the first place generates a compiled version of the lexicon, and then uses this compiled lexicon to tag the words that match within the analyzed text.

Figure 30 shows that there is a small difference between the semantic tagging of the CCG-IE system and the one done using the ontology instances (e.g. the word 'Required' was tagged as VVN and as NP respectively); however, for the most important words, which are the biological entities, the tagged output is the same (e.g. OmpR, ompF, Escherichia coli).

Semantic tagging using the dictionary	Semantic tagging using ontology instances
<pre><s pmid='2656704.txt'> Identification NN of of the dt DNA-binding NN Domain dom of of the dt OmpR nnpG Protein prot Required VVN for for Transcriptional Activation NN of of the dt ompF nnpG and cc ompC nnpG Genes gene of of Escherichia coli org by by in in VivoDNA nnpG Footprinting vvg " " ((Received VVN for for publication NN , cma December np 8 cd</pre>	<pre><s pmid='2656704.txt'> Identification NN of of the dt DNA-binding NP Domain dom of of the dt OmpR nnpG Protein prot Required NP for for Transcriptional Activation NN of of the dt ompF nnpG and cc ompC nnpG Genes gene of of Escherichia coli org by by in in VivoDNA nnpG Footprinting vvg " " ((Received NP for for publication NN , cma December np 8 cd</pre>

Figure 30: Comparison of Semantic tagging processes

4.2.4 Parsing

The parsing module of the CCG-IE system is performed by the SCOL parser as mentioned in section 4.1. For this process two files were used: a modified grammar that is focused on the expression of biological concepts and a dictionary of tags that serve as a tag map file to unify the POS tags of the input file with the ones that the CASS grammar uses. These two files (the grammar and the dictionary of tags) were translated to ontology instances in order to keep all the domain knowledge independent from the developed OBIE system.

Further explanation of how the SCOL system parses the tagged file and how the information stored in the ontology is stated below.

4.2.4.1 Tagset mapping

The Cass parser expects its input file to be tagged by other software, so it is necessary to adapt Cass to use the output of any tagger that uses a moderately standard tagset. This process is called *tagset mapping* and uses the UPenn Treebank tagset corpus.

The Cass input should consist of tab-separated fields, with one word per line of input (just like the output of the semantic tagging module). The first field contains a word and the second field contains a part of speech. The part of speech that Cass uses *depends on the grammar*. To avoid retagging with every change in the grammar, the tag mapping is done to map from a standard tagset to the grammar-specific tagset. The mapping is specified by a tag map file which is recognized by Cass as .fx file e.g.

```
aa      ala      *
aa      alanine  *
aa      arg      *
aa      arginine *
according according vbg
actdom activating domain *
by      through  IN
regul   control  NN
regul   regulation *
```

This tag map converts the UPenn tagset into the tagset of the specific grammar. The file consists of mappings. As with Cass input files, the fields are tab-separated. The mapping "y w x" should be read "output new tag y wherever word w appears with old tag x". Either the word or the old tag can be wildcarded with * as in:

*regul regulation **

Hence the interpretation is "word should be tagged as new tag no matter its old tag".

This .fx file was built by the Java program by taking the instances of the ParseTerms class and placing them in the format described above, Figure 31 gives an illustrative example:

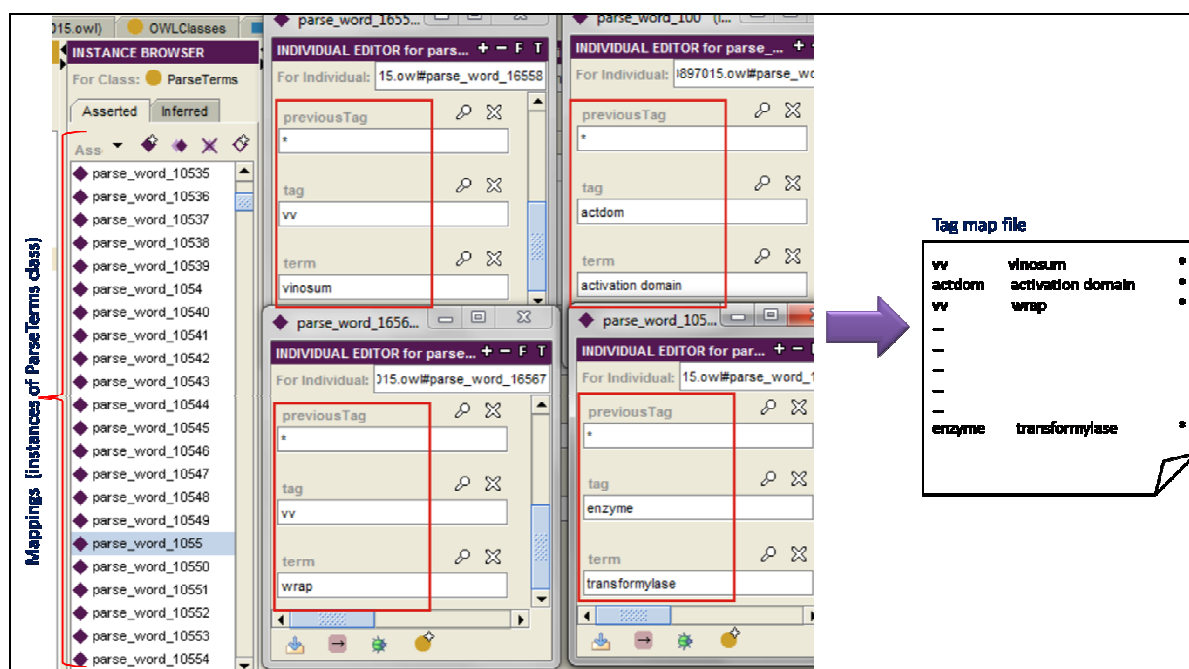


Figure 31: Generation of the tag map file from the ontology instances

After the 23301 instances were placed on the map file, the SCOL parser compiled the .fx file creating the.fxc file which is the one used by the grammar.

4.2.4.2 Grammar

Since the CCG-IE is a domain specific system, the default CASS grammar is not sufficient for parsing, that is why a grammar focused on the expression of biological

concepts is used. In order to generalize the system, this grammar was included as part of the ontology as explained in section 3.4. When the parsing process is performed, the Java program goes to the ontology, extracts all the grammar rules and processes them to build a .reg file. This file is required by SCOL, together with the .fxc file explained in the previous section.

The SCOL parsing process is briefly described below because it is helpful to understand the format of the .reg file in which the ontology instances were placed.

Cass consists of a pipeline of specialized recognizers organized by levels. At the lowest level, level 1, the input consists of words with parts of speech. Level 2 finds all sequences at level 1 that match a given pattern (e.g. date expressions) and it reduces those sequences to single elements with the appropriate category (e.g. Date). The output of level 2 then becomes the input to level 3 and so on.

Levels are defined by a colon followed by the name of the level and they consist of patterns (rules) that are defined by using a Regular Expression Grammar (REG). A pattern consists of a category followed by an arrow followed by a regular expression terminated by a semi colon e.g.:

```

:np1 -----> Level name
NP1 -> ART NN+;
NAME -> NE+;
:vp
VP -> VVFIM | VVIMF;
:pp
PP = NAME | NP1;
PP -> APPR NP;
:np2
VERB = VVFIM | VVPP;
NP2 -> ART (PP* VERB)? (NN | NAME);
```

Diagram illustrating the grammar rules and their corresponding level names and patterns:

- Level name: **:np1**
- Patterns: **NP1 -> ART NN+;** and **NAME -> NE+;**
- Level name: **:vp**
- Patterns: **VP -> VVFIM | VVIMF;**
- Level name: **:pp**
- Patterns: **PP = NAME | NP1;** and **PP -> APPR NP;**
- Level name: **:np2**
- Patterns: **VERB = VVFIM | VVPP;** and **NP2 -> ART (PP* VERB)? (NN | NAME);**

This grammar has 4 levels; thus, compilation process produces 4 automata and the parser runs these automata in a cascading manner, one after the other. Each rule in turn is described below:

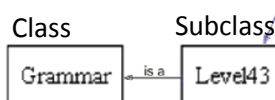
1. NP1 is constructed from an ART followed by at least one NN and possibly more.
2. NAME is constructed from at least one NE and possibly more.

-
3. A VP is constructed by a VFIN or a VINP.
 4. This is not a rule; it is like a local macro that can also be seen as a local variable. It defines a symbol that can be used in the following rules of the level. It makes writing the next rule easier. The symbol is NOT sent on to the next level.
 5. This rule uses the NP symbol defined above in a rule for a PP.
 6. This line defines the symbol VERB.
 7. This rule uses parentheses, (PP* VERB)? to allow complex pieces of a regular expression. It is important to notice that this rule uses a PP in it which has a NP1 in it from an earlier level.

So the first step to generate the grammar file was to access the Grammar subclasses (which represent the levels of the grammar), although the Protégé editor does not order the subclasses in an ascending way, it is very important to access them in strictly ascending order because, as we have seen before, the patterns of one level are linked to the ones of the next level.

The order of the rules and variables (if it exists) defined for each level is very important, so using the 'order' property, the content of each level (first Level1, then Level2 and so on) was used to generate the .reg file which is the non-compiled grammar file. After finishing this process, the Java program sent the reg file to the SCOL parser and it generated a compiled automaton .fxc which is the grammar that is used to generate the output cascades. An example of this process is shown on figure 32.

Class Level43 with its instances; as can be seen, the instances were placed in the .reg file according to the 'order' property



Instances	
Level43_ev_act_expr_rx	
#definition =	-> nxev_uas_in nxev_expr nxev_bind_u...
#order =	1
#term =	ev_act_expr_rx
Level43_ev_act_expr_rx	
#definition =	-> nxev_tractr ev_trregr_act_vp ev_t...
#order =	2
#term =	ev_act_expr_rx
Level43_ev_reg_expr_rx	
#definition =	-> nxev_bs_in nxev_dep_expr nxev_bin...
#order =	5
#term =	ev_reg_expr_rx
Level43_ev_rep_expr_rx	
#definition =	-> nxev_trregr ev_trregr_reg_vp ev_a...
#order =	6
#term =	ev_rep_expr_rx
Level43_ev_rep_expr_rx	
#definition =	-> nxev_urs_in nxev_bind_urs nxev_re...
#order =	3
#term =	ev_rep_expr_rx
Level43_ev_rep_expr_rx	
#definition =	-> nxev_trregr ev_trregr_rep_vp ev_t...
#order =	4
#term =	ev_rep_expr_rx



Grammar .reg file

```

ev_be_tractr_vp -> tractr_pt (cma? wdt)? (bez |
ev_be_tractr_va_n -> NXnnpgrx (cma? wdt)? (bez n
ev_be_tractr_vp_n -> nxtractr_pt (cma? wdt)? (b
ev_be_trrepr_va -> NXnnpgrx (cma? wdt)? (bez | b
ev_be_trrepr_vp -> nxtrrepr_pt (cma? wdt)? (bez
ev_be_trrepr_va_n -> NXnnpgrx (cma? wdt)? (bez n
ev_be_trrepr_vp_n -> nxtrrepr_pt (cma? wdt)? (b
ar_be_trregr_va -> NXnnpgrx (cma? wdt)? (bez b
bez
bez
ev_be_trregr_vp_n -> nxtrregr_pt (cma? wdt)? (b
:Level43
ev_act_expr_rx -> nxev_uas_in | nxev_expr | nxe
ev_act_expr_rx -> nxev_tractr | ev_trregr_act_vp
ev_rep_expr_rx -> nxev_urs_in | nxev_bind_urs |
ev_rep_expr_rx -> nxev_trregr | ev_trregr_rep_vp
ev_reg_expr_rx -> nxev_bs_in | nxev_dep_expr |
ev_reg_expr_rx -> nxev_trregr | ev_trregr_reg_vp
:Level44
events_all -> ev_act_expr_rx | ev_act_expr_rx |
:Level45
PREP = to | with | by | cc | in | of;
CONNECTORS = nxrest | rx | rb | pp | PREP;
ACTIVATION = ev_act_expr_rx | ev_act_expr_rx;
REPRESSION = ev_rep_expr_rx
EVENTS = ev_act_expr_rx | ev_act_expr_rx | ev r
NXnnpgrx = nxpg | nxgene | nxrna | nxprot | nxt
VERBS = actv | codv | contv | dphosv | expv |
VING = acting | assembling | causing | coding

```

Figure 32: Example of how the .reg file was built

The following image compares the resulting output of the parsing modules of CCG-IE system and the output of this work.

Parsing output of the CCG-IE	Parsing output of this ontology approach
<pre> [rb In vivo] [events_all [ev_reg_expr_rx [nxev_bind_prom [bing binding] [of of] [nxprot [nxpg [nxpg [dt the] [nnpgx [nnpg ompR]]]] [prot protein]] [to to] [nxprom [nxpg [nxpg [dt the] [nnpgx [nnpg ompF]] [cc and] [nnpgx [nnpg ompC]]]] [prom promoter] [nns regions]]]]] [bevvn [bedz was] [vvn observed]] [vgx [vvg using]] [dt an] [rb in vivo] [nxrest </pre>	<pre> [rb In vivo] [events_all [ev_reg_expr_rx [nxev_bind_prom [bing binding] [of of] [nxprot [nxpg [nxpg [dt the] [nnpgx [nnpg ompR]]]] [prot protein]] [to to] [nxprom [nxpg [nxpg [dt the] [nnpgx [nnpg ompF]] [cc and] [nnpgx [nnpg ompC]]]] [prom promoter] [nns regions]]]]] [vx [bedz was] [vvd observed]] [vgx [vvg using]] [dt an] [rb in vivo] [nxrest </pre>

Figure 33: Comparison of the parsing process files

As can be seen, the structure of the cascades is practically the same except on the red box; however, the relevant cascades such e.g. *ev_reg_expr_rx* (which denotes an expression of a protein regulating a gene) are identified correctly. Once the parsing has finished, the last step is to select the relations we are interested in (Regulatory expressions in this case).

4.2.5 Interaction selection

This last module consists of selecting the desired interactions, in this particular case, regulatory interactions. The CCG-IE system converts the parsed file into an XML format (figure 34) and processes it using customized heuristics modules that:

- Identify the regulatory interactions that are to be extracted and when possible, the kind of the interaction extracted (activation or repression).
- Create an XML output file with a regulatory network retrieved from the processed raw text (figure 35).

```
<?xml version='1.0' encoding='UTF-8'?>

<RegNet>
<s pmid="XXXX"><nxrest><nn>Identification</nn></nxrest><of><of></of></nxrest><dt>the</dt><np>
DNA-binding</np></nxrest><dom>Domain</dom><of><of></of></nxprot><nxpg><nxpg><dt>the</dt><nnpgx>
<nnpg>OmpR</nnpg></nnpgx></nxpg></nxpg><prot>Protein</prot><np>Required</np></nxprot><for>for</for>
<for><nxrest><np>Transcriptional Activation</np></nxrest><of><of></of></nxgene><nxpg><nxpg><dt>the</dt>
<dt><nnpgx><nnpg>ompF</nnpg></nnpgx><cc>and</cc><nnpgx><nnpg>ompC</nnpg></nnpgx></nxpg></nxpg>
<gene>Genes</gene><of><of></of></org>Escherichia coli</org></nxgene><by>by</by><in>in</in><nxpg>
<nxpg><nnpgx><nnpg>VivoDNA</nnpg></nnpgx></nxpg></nxpg><vgx><vgx><vvg>Footprinting</vvg></vgx>
<punctpunct>"</punctpunct><punct>(</punct><nxrest><np>Received</np></nxrest><for>for</for>
<nxrest><nn>publication</nn></nxrest><cma>, </cma><nxrest><np>December</np></nxrest><nxrest><cd>
</cd></nxrest><cma>, </cma><nxrest><cd>1988</cd></nxrest><punct>)</punct><nxpg><nxpg><np>Kangla</np>
<nnpgx><nnpg>TsungS</nnpg></nnpgx></nxpg></nxpg><cma>, </cma><nxrest><np>Renee</np></nxrest>
<nxrest><np>E.</np></nxrest><nxrest><np>Brissette</np></nxrest><cma>, </cma><cc>and</cc><nxrest>
<np>MasayoriInouye</np></nxrest><from>From</from><nxrest><dt>the</dt><np>Department</np></nxrest>
<of><of></of></nxrest><np>Biochemistry</np></nxrest><cma>, </cma><nxrest><np>Robert</np></nxrest>
<nxrest><np>Wood</np></nxrest><nxrest><np>Johnson</np></nxrest><nxrest><np>Medical</np></nxrest>
<nxrest><np>Schoolat</np></nxrest><nxrest><np>Rutgers</np></nxrest><cma>, </cma><nxrest><np>
University</np></nxrest><of><of></of></nxrest><np>Medicine</np></nxrest><cc>and</cc><nxrest><np>
Dentistry</np></nxrest><of><of></of></events_all><ev_reg_expr_xr><ev_expr_inv_va><nxexpr><jj>New</jj>
<jj><np>Jersey</np><cma>, </cma><np>Piscataway</np><cma>, </cma><jj>New</jj><np>Jersey</np><cd>
08854</cd><expr>Expression</expr><of><of></of></nxgene><nxpg><nxpg><dt>the</dt><nnpgx><nnpg>ompF</nnpg>
</nnpgx><cc>and</cc><nnpgx><nnpg>ompC</nnpg></nnpgx></nxpg></nxpg><gene>genes</gene><of><of>
</of></org>Escherichia coli</org></nxgene><nxexpr><requv>requires</requv><nxprot><nxpg><nxpg><dt>
the</dt><nnpgx><nnpg>OmpR</nnpg></nnpgx></nxpg></nxpg><prot>protein</prot></nxprot></ev_expr_inv_va>
</ev_reg_expr_xr></events_all><for>for</for><traction>transcriptional activation</traction><sent>.</sent></s>
</RegNet>
```

Figure 34: Cascaded file converted into a XML file

```

<?xml version="1.0" ?>
<!DOCTYPE RegulatoryNetwork
  SYSTEM 'regulation.dtd'>
<?xml-stylesheet type='text/xsl' href='ext.xsl'?>
<RegulatoryNetwork>
  <interaction ID="10" from="ev_bind_prom_va" ri_function="unknown" source="XXXX">
    <regulator GenProtID="ECK120000665" org="ecoli" type="nxpg">
      OmpR
    </regulator>
    <regulated GenProtID="xxxx" org="ecoli" type="nxprom">
      theompF
    </regulated>
    <evidence verb="bind"/>
    <sentence pmid="XXXX">
      It is not clear at present if three OmpR in vivo footprinting we could , instead ,
      findtwo types of OmpR molecules are able to bind to theompF promoter at thesame binding motifs
      with different consensus sequences ( Fig. 4B ) .
    </sentence>
  </interaction>
  <interaction ID="3" from="nxev_bind_prom" ri_function="unknown" source="XXXX">
    <regulator GenProtID="ECK120000665" org="ecoli" type="nxprot">
      OmpR
    </regulator>
    <regulated GenProtID="ECK120000664" org="ecoli" type="nxprom">
      ompF
    </regulated>
    <evidence verb="to"/>
    <sentence pmid="XXXX">
      In vivo binding of the OmpR protein to the ompF and ompC promoter regions was
      observed using an in vivo dimethyl sulfate DNA footprinting technique .
    </sentence>
  </interaction>
</RegulatoryNetwork>

```

Figure 35. Final XML format with the identified interactions

The red boxes indicate what the type of interaction was found; then, inside the interaction label, there is the name of the regulator and the regulated and the sentence where the interaction was found. This is useful to the human expert that can corroborate the solution by reading the whole sentence and analyzing if the output is correct or not.

This process is very attached to the domain knowledge since the heuristics in this module were designed very carefully by human experts to detect and extract a set of specific regulatory expression between biological entities. Hence, it is not easy to generalize this module, so that when employing the OBIE system developed in this work using an ontology different to the current one the system will still work, being the parsing module the last one that was generalized. Once that cascaded file is generated, the domain expert should create his module to extract the desired information in the preferred format. Using XML is not mandatory.

The generalization of this interaction selection process has been left as future work. Once the new OBIE system has been finished and explained is it useful to make a system test by applying some cases of study, as chapter 5 explains.

5 Cases of study and results

This chapter shows the results obtained after trying the system with four different cases of study. The corpus used to measure the system performance was obtained from the PubMed website and it consists of a set of two full-papers and two abstracts.

The description of the corpus is given in table 5.

PMID	Document type	Title	Author(s)
7510658	Abstract	Localization of promoters in the fim gene cluster and the effect of H-NS on the transcription of fimB and fimE	Olsen PB, Klemm P.
2464125	Abstract	Mapping of the multiple regulatory sites for putP and putA expression in the putC region of Escherichia coli.	Nakao T, Yamato I, Anraku Y.
14996792	Full-text	Regulation of Escherichia coli hemolysin E expression by H-NS and Salmonella SlyA.	Wyborn NR, Stapleton MR, Norte VA, Roberts RE, Grafton J, Green J.
8895588	Full-text	Influence of DNA geometry on transcriptional activation in Escherichia coli.	Déthiollaz S, Eichenberger P, Geiselmann J.

Table 5: Corpus description

5.1 Abstracts

The “Localization of promoters in the fim gene cluster and the effect of H-NS on the transcription of fimB and fimE.” abstract was chosen to illustrate in a more detailed way, the obtained results.

Abstract:

Localization of promoters in the fim gene cluster and the effect of H-NS on the transcription of fimB and fimE.

Olsen PB, Klemm P.

Source: Department of Microbiology, Technical University of Denmark, Lyngby.

The expression of type 1 fimbriae in *Escherichia coli* undergoes phase variation in which individual bacteria switch between a fimbriated and non-fimbriated state. The transition from one state to the other is caused by inversion of a DNA segment containing the promoter for the fimA gene. The orientation of the invertible segment is controlled by two proteins, FimB and FimE, which mediate an on/off and off only orientation of the segment, respectively. In this study we have mapped the 5' termini of the fimB, fimE and fimA transcripts. Furthermore, we show that expression of fimB and fimE is strongly influenced by the H-NS nucleoid protein.

The outputs of the preprocess, POS tagging and semantic tagging modules is not shown because they consist only in files with words and their corresponding tag.

The following is the output of the parsed file:

```
<s pmid='7510658.txt'>
  [nxrest
    [nn Localization]]
  [of of]
  [nxrest
    [prom promoters]]
  [in in]
  [nxrest
    [dt the]
    [nn fim]
    [gene gene]
    [nn cluster]]
  [cc and]
  [nxrest
    [dt the]
    [nn effect]]
  [of of]
  [nxpg
    [nxpg
      [nnpgx
        [nnpg H-NS]]]]
  [in on]
  [nxexpr
    [dt the]
    [transc transcription]
    [of of]
    [nxpg
      [nxpg
        [nnpgx
          [nnpg fimB]]
        [cc and]
        [nnpgx
          [nnpg fimE]]]]]]
```

```

[sent .]
<\s>
<s pmid='7510658.txt'>
[nxrest
  [np Olsen]]
[nxrest
  [np PB]]
[cma ,]
[nxrest
  [np Klemm]
  [np P.]
  [nn SourceDepartment]]
[of of]
[nxrest
  [nn Microbiology]]
[cma ,]
[nxrest
  [np Technical]]
[nxrest
  [np University]]
[of of]
[nxrest
  [np Denmark]]
[cma ,]
[nxrest
  [nn Lyngby]]
[sent .]
<\s>
<s pmid='7510658.txt'>
[nxrest
  [nn Abstract]]
[nxrest
  [dt The]
  [expr expression]]
[of of]
[nxrest
  [nn type]]
[ls 1]
[nxrest
  [nns fimbriae]]
[in in]
[nxrest
  [org Escherichia coli]]
[vvz undergoes]
[nxrest
  [nn phase]
  [nn variation]]
[in in]
[wdt which]
[nxrest
  [jj individual]
  [org bacteria]]
[vv switch]
[in between]
[nxrest
  [dt-a a]
  [jj fimbriated]
  [cc and]
  [jj non-fimbriated]]

```

```

[nn state]]
[sent .]
<\s>
<s pmid='7510658.txt'>
[nxrest
[dt The]
[nn transition]]
[from from]
[nxrest
[cd one]
[nn state]]
[to to]
[dt the]
[ax
[jj other]]
[vx
[bez is]]
[cauv caused]
[by by]
[nxrest
[nn inversion]]
[of of]
[dt-a a]
[dna DNA]
[nxrest
[nn segment]]
[containing containing]
[nxrest
[dt the]
[prom promoter]]
[for for]
[nxgene
[nxpg
[nxpg
[dt the]
[nnpgx
[nnpg fimA]]]]
[gene gene]]
[sent .]
<\s>
<s pmid='7510658.txt'>
[nxrest
[dt The]
[nn orientation]]
[of of]
[nxrest
[dt the]
[jj invertible]
[nn segment]]
[vx
[bez is]]
[regv controlled]
[by by]
[nxprot
[cd two]
[prot proteins]
[cma ,]
[nxpg
[nxpg

```

```

        [nnpgx
        [nnpg fimB]]
        [cc and]
        [nnpgx
        [nnpg fimE]]]]
    [cma ,]]
[wdt which]
[mediv mediate]
[dt an]
[in on]
[sym /]
[rp off]
[cc and]
[rp off]
[cdqlx
    [only only]]
[nxrest
    [nn orientation]]
[of of]
[nxrest
    [dt the]
    [nn segment]]
[cma ,]
[rb respectively]
[sent .]
<\s>
<s pmid='7510658.txt'>
    [in In]
    [nxrest
        [dtp this]]
    [vv study]
    [pp we]
    [vx
        [hv have]]
    [vvn mapped]
    [dt the]
    [ls 5]
    [pos ']
    [nxrest
        [nns termini]]
    [of of]
    [nxrna
        [nxpg
            [nxpg
                [dt the]
                [nnpgx
                    [nnpg fimB]]
                    [cma ,]
                    [nnpgx
                        [nnpg fimE]]
                        [cc and]
                        [nnpgx
                            [nnpg fimA]]]]]
                [trans transcripts]]
            [sent .]
        ]
    ]
<\s>
<s pmid='7510658.txt'>
    [rb Furthermore]
    [cma ,]

```

```

[pp we]
[showv show]
[wdt that]
[events_all
  [ev_reg_expr_xr
    [ev_reg_expr_vp
      [nxexpr
        [expr expression]
        [of of]
        [nxpg
          [nxpg
            [nnpgx
              [nnpg fimB]]
            [cc and]
            [nnpgx
              [nnpg fimE]]]]]]
        [vx
          [bez is]]
        [rb strongly]
        [regv influenced]
        [by by]
        [nxprot
          [nxpg
            [nxpg
              [dt the]
              [nnpgx
                [nnpg H-NS]]
                [jj nucleoid]]]
            [prot protein]]]]]
      [sent .]
    ]
  ]
</s>

```

This cascade file was analyzed by one of the scripts of the interaction selection process. When the script found some regulatory expression (such as **ev_reg_expr_xr**) it converted the sentence where the expression appeared into an XML format as figure 36 shows.

```

<?xml version='1.0' encoding='UTF-8'?>

<RegNet>
<s pmid="XXXX"><rb>Furthermore</rb><cma>, </cma><pp>we</pp><showv>show</showv><wdt>that</wdt>
<events_all><ev_reg_expr_xr><ev_reg_expr_vp><nxexpr><expr>expression</expr><of>of</of><nxpg>
<nxpg><nnpgx><nnpg>fimB</nnpg><nnpgx><cc>and</cc><nnpgx><nnpg>fimE</nnpg><nnpgx><nxpg><nxpg>
</nxexpr><vx><bez>is</bez></vx><rb>strongly</rb><regv>influenced</regv><by>by</by><nxprot><nxpg>
<nxpg><dt>the</dt><nnpgx><nnpg>H-NS</nnpg><nnpgx><jj>nucleoid</jj><nxpg><nxpg><prot>protein</
prot></nxprot></ev_reg_expr_vp></ev_reg_expr_xr></events_all><sent>.</sent></s>
</RegNet>

```

Figure 36. Sentence from the parse file represented in XML

The second step was to identify the regulatory interactions and if it is possible, the kind of interaction, the output was:

```
<?xml version="1.0" encoding="utf-8"?><!DOCTYPE RegulatoryNetwork SYSTEM 'regulation.dtd'>
<?xml-stylesheet type='text/xsl' href='ext.xsl'?><RegulatoryNetwork><interaction ID="1" from=
"ev_reg_expr_vp" ri_function="activator" source="XXXX"><regulator GenProtID="ECK120000450" org=
"ecoli" type="nxprot">H-NS</regulator><regulated GenProtID="ECK120000303" org="ecoli" type=
"nxpg">fimB</regulated><evidence verb="influenced"/><sentence pmid="XXXX"> Furthermore , we show
that expression of fimB and fimE is strongly influenced by the H-NS nucleoid protein .</
sentence></interaction><interaction ID="2" from="ev_reg_expr_vp" ri_function="activator" source=
"XXXX"><regulator GenProtID="ECK120000450" org="ecoli" type="nxprot">H-NS</regulator><regulated
GenProtID="ECK120000306" org="ecoli" type="nxpg">fimE</regulated><evidence verb="influenced"/>
<sentence pmid="XXXX"> Furthermore , we show that expression of fimB and fimE is strongly
influenced by the H-NS nucleoid protein .</sentence></interaction></RegulatoryNetwork>
```

Figure 37: Identified regulatory interactions

As it can be seen in figure 37, attributes that define the indentified interactions and their participants have been assigned, e.g.:

Attributes for an identified interaction:

- ID: 1
- Form: ev_reg_expr_vp
- Function: activator

Attributes for the participants:

- | <i>Regulator</i> H-NS | <i>Regulated</i> fimB |
|--------------------------|-----------------------|
| • ID: ECK120000450 | • ID: ECK120000303 |
| • Organism: E.Coli | • Organism: E.Coli |
| • Type: nxprot (protein) | • Type: nxpg (gene) |

The last step was just to create single entries for all unique interactions in the previous file and give some format to the final output as the following image shows:

```

<?xml version="1.0" ?>
<!DOCTYPE RegulatoryNetwork
  SYSTEM 'regulation.dtd'>
<?xml-stylesheet type='text/xsl' href='ext.xsl'?>
<RegulatoryNetwork>
  <interaction ID="1" from="ev_reg_expr_vp" ri_function="activator" source="XXXX">
    <regulator GenProtID="ECK120000450" org="ecoli" type="nxprot">
      H-NS
    </regulator>
    <regulated GenProtID="ECK120000303" org="ecoli" type="nxpg">
      fimB
    </regulated>
    <evidence verb="influenced"/>
    <sentence pmid="XXXX">
      Furthermore , we show that expression of fimB and fimE is strongly influenced by
the H-NS nucleoid protein .
    </sentence>
  </interaction>
  <interaction ID="2" from="ev_reg_expr_vp" ri_function="activator" source="XXXX">
    <regulator GenProtID="ECK120000450" org="ecoli" type="nxprot">
      H-NS
    </regulator>
    <regulated GenProtID="ECK120000306" org="ecoli" type="nxpg">
      fimE
    </regulated>
    <evidence verb="influenced"/>
    <sentence pmid="XXXX">
      Furthermore , we show that expression of fimB and fimE is strongly influenced by
the H-NS nucleoid protein .
    </sentence>
  </interaction>
</RegulatoryNetwork>

```

Figure 38: Final XML format showing the founded interactions

Two interactions were founded in the analyzed abstract, both in the same sentence:

“Furthermore, we show that expression of fimB and fimE is strongly influenced by the H-NS nucleoid protein.”

For this case of study the results were completely satisfactory since the two existing interactions were found by the system.

The results of the abstract “Mapping of the multiple regulatory sites for putP and putA expression in the putC region of Escherichia coli” are the following:

Mol Gen Genet. 1988 Nov;214(3):379-88.
 Mapping of the multiple regulatory sites for putP and putA expression
 in the putC region of Escherichia coli.
 Nakao T, Yamato I, Anraku Y.
 SourceDepartment of Biology, Faculty of Science, University of Tokyo,
 Japan.

The effects of regulatory proteins on the expression of putP and putA
 were studied using put-lacZ fusion genes. The expression of the putP-

lacZ gene was activated by the glnG gene product and the catabolite gene activator protein (CAP). The putA gene product inhibited activation of putP-lacZ gene expression by CAP or the glnG gene product and its inhibition was greater in the absence of proline. The expression of the putA-lacZ gene was activated by CAP and repressed by the glnG gene product. The putA gene product acted as a repressor in the absence of proline, but not in its presence. Studies using put-lacZ fusion genes with upstream deletions showed that the region required for the activation of putP by CAP was within 234 bp upstream of the translational initiation site and that that for the activation of putA was within 107 bp upstream of the translational initiation site of the putA gene. This supported the suggested locations of CAP binding sites. The region required for induction of putP and putA expression by proline was located at the HpaI site 182 bp upstream of the translational starting site of putA, suggesting that a sequence of dyad symmetry located 1 bp to the left of the HpaI site is a candidate for the binding site of the putA gene product.

Found interaction(s):

Interaction ID="1" from="ev_act_expr_vp" ri_function="activator"

Regulator: glnG GenProtID="ECK120000378" type="nxgeneprod"

Regulated: putP GenProtID="ECK120000793" type="nxgene"

"The expression of the putP gene was activated by the glnG gene product and the catabolite gene activator protein (CAP)."

Interaction ID="3" from="anaph+ev_act_expr_xr" ri_function="repressor"

Regulator: glnG GenProtID="ECK120000378" type="nxgeneprod"

Regulated: putA GenProtID="ECK120000792" type="nxgene">

"The expression of the putA gene was activated by CAP and repressed by the glnG gene product."

Interaction ID="2" from="ev_act_expr_vp" ri_function="activator"

Regulator CAP GenProtID="ECK120000160" type="nxpg"

Regulated putA GenProtID="ECK120000792" type="nxgene">

"The expression of the putA gene was activated by CAP and repressed by the glnG gene product."

The human expert agreed with these three found interactions.

5.2 Full-text

Due to the length of the full text papers, images of the output results are not included in this work; instead, a table with a comparison between the expressions found by the system and the ones found by an expert is presented below. Besides, appendixes A and B contain the full-text papers with highlighted interactions.

PMID	Interactions found by the system	Interactions found by the expert
8002608	ID: 22 Function: unknown Regulator: H-NS Regulated: hlyE	Function: Interactions Regulator: H-NS Regulated: hlyE
	ID: 1 Function: activator Regulator: FNR Regulated: hlyE	Function: activation Regulator: FNR Regulated: hlyE
	ID: 27 Function: activator Regulator: H-NS Regulated: malT	Function: activation Regulator: H-NS Regulated: malT
	ID: 8 Function: activator Regulator: H-NS Regulated: csiD	Function: activation Regulator: H-NS Regulated: csiD
	ID: 3 Function: Regulator: SlyA Regulated: hlyE	Function Regulator: SlyA Regulated: hlyE
	ID: 7 Function: activator Regulator: crp Regulated: hlyE	Function: activation Regulator: crp Regulated: hlyE

PMID	Interactions found by the system	Interactions found by the expert
6262769	ID: 1 Function: activator Regulator: crp Regulated: malT	Function: activation Regulator: crp Regulated: malT
	ID: 2 Function: activator Regulator: crp Regulated: crp	Function: activation Regulator: crp Regulated: crp (promoter and protein have the same name but they are different entities)

Table 6. Comparison between interaction found by the system and expert

According to table 6, it can be inferred that the system performs considerably well finding interactions, as it detected all the regulatory expressions that the expert did. It just needs to be modified to avoid detecting the same interaction more than one time.

6 Conclusions and future work

As mentioned at the beginning of this work, in recent years Information Extraction has become very active in the field of bioinformatics. This work gave an overview of the current state of IE applied to the molecular biology domain, where most of the developed systems are rule-based and others combine statistical methods and rules.

The implemented system described in this work, was meant to be a domain-independent extension of the CCG-IE system.

6.1 Conclusions

This work was aimed to generalize the IE system proposed by the CCG and to improve their performance by implementing new techniques of ontology-based Information Extraction (OBIE).

The OBIE system was created by designing and creating an ontology to transfer all the domain knowledge of the CCG-IE system into it. The ontology was also automatically populated with the biological entities from RegulonDB by developing a methodology to translate the RegulonDB entity-relation structure to an ontology structure as explained in chapter 3.

Implementation of generic mechanisms for Information Extraction where applied to exploit the semantic content of the information in the created domain ontology. Finally four cases of study (consisting of two abstracts and two full text papers) were used to test the system performance. The interactions that the system found were compared with the ones that a human expert found in the same corpus. The results (chapter 5) show that the system throws good results.

Even the system performs very well; there will always be necessary a human expert to analyze the resulting output to ensure the results are correct.

6.2 Future work

As for future work, an important aspect is to prove the system with a bigger corpus consisting in both abstract and full-text in order to have a more realistic idea of the performance of the system, and also to measure accuracy (precision) and coverage (recall).

Another important improvement is to find a way to generalize the last module of the system, the interaction selection, so that the system could be completely domain independent and the user shall only need to deal with changing the ontology without modifying anything from the system code.

7 References

1. Kauchak, D., J. Smarr, and C. Elkan, *Sources of Success for Information Extraction Methods*. Journal of Machine Learning Research, 2002.
2. Andrade, M.A. and P. Bork, *Automated extraction of information in molecular biology*. FEBS Lett, 2000. **476**(1-2): p. 12-7.
3. Wikipedia. *Bioinformatics*. 2011 [cited 2011 February]; Available from: <http://en.wikipedia.org/wiki/Bioinformatics>.
4. Benson, D.A., et al., *GenBank*. Nucleic Acids Res, 2006. **34**(Database issue): p. D16-20.
5. Keseler, I.M., et al., *EcoCyc: a comprehensive database resource for Escherichia coli*. Nucleic Acids Res, 2005. **33**(Database issue): p. D334-7.
6. Grivell, L., *Mining the bibliome: searching for a needle in a haystack?* EMBO Rep., 2002. **3**(3): p. 200 - 203.
7. Collado-Vides, J., et al., *Bioinformatics resources for the study of gene regulation in bacteria*. J Bacteriol, 2009. **191**(1): p. 23-31.
8. Rodríguez-Penagos, C., et al., *NLP-Based Curation of Bacterial Regulatory Networks*, in *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing*. 2007, Springer-Verlag: Mexico City, Mexico. p. 575-586.
9. Grishman, R., *Information Extraction: Techniques and Challenges*, in *International Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*. 1997, Springer-Verlag. p. 10-27.
10. Riloff, E., *Information extraction as a stepping stone toward story understanding*, in *Understanding language understanding*, R. Ashwin and M. Kenneth, Editors. 1999, MIT Press. p. 435-460.
11. Yangarber, R., *Scenario customization for information extraction*. 2001, New York University: New York. p. 133.
12. Peshkin, L. and A. Pfeffer, *Bayesian information extraction network*, in *Proceedings of the 18th international joint conference on Artificial intelligence*. 2003, Morgan Kaufmann Publishers Inc.: Acapulco, Mexico. p. 421-426.
13. Xiao, L., et al., *Information Extraction from the Web: System and Techniques*. Applied Intelligence, 2004. **21**(2): p. 195-224.
14. Cunningham, H., *Information Extraction, Automatic*, in *Encyclopedia of Language and Linguistics*, K. Brown, Editor. 2005, Elsevier Science Publishers. p. 665-677.
15. Blaschke, C., L. Hirschman, and A. Valencia, *Information extraction in molecular biology*. Brief Bioinform, 2002. **3**(2): p. 154-65.
16. Etzioni, O., et al., *Open information extraction from the web*. Commun. ACM, 2008. **51**(12): p. 68-74.
17. Feldman, R., et al., *Mining Biomedical Literature using Information Extraction*. Current Drug Discovery, 2002: p. 19-23.
18. Wimalasuriya, D.C. and D. Dou, *Ontology-based information extraction: An introduction and a survey of current approaches*. J. Inf. Sci., 2010. **36**(3): p. 306-323.
19. Blaschke, C., et al., *Extracting information automatically from biological literature*. Comp Funct Genomics, 2001. **2**(5): p. 310-3.

-
20. Andrade, M.A. and A. Valencia, *Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families*. Bioinformatics, 1998. **14**(7): p. 600-7.
 21. Bikel, D.M., et al., *Nymble: a high-performance learning name-finder*, in *Proceedings of the fifth conference on Applied natural language processing*. 1997, Association for Computational Linguistics: Washington, DC. p. 194-201.
 22. Bahl, L.R., F. Jelinek, and R.L. Mercer, *A maximum likelihood approach to continuous speech recognition*, in *Readings in speech recognition*, W. Alex and L. Kai-Fu, Editors. 1990, Morgan Kaufmann Publishers Inc. p. 308-319.
 23. Merialdo, B., *Tagging English text with a probabilistic model*. Comput. Linguist., 1994. **20**(2): p. 155-171.
 24. Charniak, E., *Statistical parsing with a context-free grammar and word statistics*, in *Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence*. 1997, AAAI Press: Providence, Rhode Island. p. 598-603.
 25. Gale, W.A., K.W. Church, and D. Yarowsky, *A method for disambiguating word senses in a large corpus*. Computers and the Humanities 1992. **26**: p. 415-439.
 26. Michelakis, E., et al., *Uncertainty management in rule-based information extraction systems*, in *Proceedings of the 35th SIGMOD international conference on Management of data*. 2009, ACM: Providence, Rhode Island, USA. p. 101-114.
 27. Studer, R., V.R. Benjamins, and D. Fensel, *Knowledge engineering: principles and methods*. Data Knowl. Eng., 1998. **25**(1-2): p. 161-197.
 28. Li, Y. and K. Bontcheva, *Hierarchical, perceptron-like learning for ontology-based information extraction*, in *Proceedings of the 16th international conference on World Wide Web*. 2007, ACM: Banff, Alberta, Canada. p. 777-786.
 29. Wu, F. and D.S. Weld, *Autonomously semantifying wikipedia*, in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 2007, ACM: Lisbon, Portugal. p. 41-50.
 30. Yildiz, B. and S. Miksch, *Motivating ontology-driven information extraction*, in *Proceedings of the 2007 International Conference on Semantic Web and Digital Libraries*, A. InPrasad and D. Madalli, Editors. 2007, Springer-Verlag: Indian Statistical Institute Platinum Bangalore, India. p. 45-53.
 31. Wu, F., R. Hoffmann, and D.S. Weld, *Information extraction from Wikipedia: moving down the long tail*, in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008, ACM: Las Vegas, Nevada, USA. p. 731-739.
 32. McDowell, L. and M.J. Cafarella, *Ontology-Driven Information Extraction with OntoSyphon*, in *of the 5th International Semantic Web Conference 2006*, Springer: Berlin, Germany. p. 428-444.
 33. Yildiz, B. and S. Miksch, *ontoX - a method for ontology-driven information extraction*, in *Proceedings of the 2007 international conference on Computational science and its applications - Volume Part III*. 2007, Springer-Verlag: Kuala Lumpur, Malaysia. p. 660-673.
 34. Russell, S. and P. Norvig, *Artificial Intelligence: A Modern Approach*. 2nd ed. 2003, Englewood Cliffs, NJ: Prentice-Hall.
 35. Friedman, C., et al., *GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles*. Bioinformatics, 2001. **17 Suppl 1**: p. S74-82.
 36. Friedman, C., et al., *A general natural-language text processor for clinical radiology*. J Am Med Inform Assoc, 1994. **1**(2): p. 161-74.

-
37. Altschul, S.F., et al., *Basic local alignment search tool*. J Mol Biol, 1990. **215**(3): p. 403-10.
 38. Corney, D.P., et al., *BioRAT: extracting biological information from full-length papers*. Bioinformatics, 2004. **20**(17): p. 3206-13.
 39. Cunningham, H., et al., *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*, in *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics 2002*: Philadelphia PA USA.
 40. Ginter, F., et al., *Information extraction from biomedical text: The BioText project*. 2005.
 41. Ginter, F., et al., *Bio Dependency Bank: a dependency corpus for information extraction in the biomedical domain*. Submitted, 2004.
 42. Ginter, F., et al., *Ontology-based feature transformations: A data-driven approach*, in *Proceedings of EsTAL'04*. 2004, Springer: Berlin, Germany. p. 279-290.
 43. Ginter, F., et al., *Extracting protein-protein interaction sentences by applying rough set data analysis*, in *Proceedings of Proceedings of the Fourth International Conference on Rough Sets and Current Trends in Computing (RSCTC)*. 2004. p. 780-785.
 44. Schölkopf, B. and A.J. Smola, *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. 2002, Cambridge, MA ; London: MIT Press. xviii, 626 p.
 45. Sleator, D.D. and D. Temperley, *Parsing English with a link grammar*. Carnegie Mellon University. School of Computer Science, 1991(CMU-CS-91-196): p. 91 p.
 46. Saric, J., L.J. Jensen, and I. Rojas, *Large-scale extraction of gene regulation for model organisms in an ontological context*. In Silico Biol, 2005. **5**(1): p. 21-32.
 47. Saric, J., et al., *Extraction of regulatory gene/protein networks from Medline*. Bioinformatics, 2006. **22**(6): p. 645-50.
 48. Abney, S., *Partial parsing via finite-state cascades*. Nat. Lang. Eng., 1996. **2**(4): p. 337-344.
 49. Schmid, H., *Unsupervised learning of period disambiguation for tokenisation Technical report*. Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, 2000.
 50. Schmid, H., *Probabilistic part-of-speech tagging using decision trees*, in *International Conference on New Methods in Language Processing*. 1994: Manchester, UK. p. 44-49.
 51. Kim, J.D., et al., *GENIA corpus--semantically annotated corpus for bio-textmining*. Bioinformatics, 2003. **19 Suppl 1**: p. i180-2.
 52. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information*. Nucleic Acids Res, 2006. **34**(Database issue): p. D173-80.
 53. Song, Y.L. and S.S. Chen, *Text mining biomedical literature for constructing gene regulatory networks*. Interdiscip Sci, 2009. **1**(3): p. 179-86.
 54. Brill, E., *Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging*. Comput. Linguist., 1995. **21**(4): p. 543-565.
 55. Saric, J., et al., *Extracting regulatory gene expression networks from PubMed*, in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. 2004, Association for Computational Linguistics: Barcelona, Spain. p. 645-650.

Appendix A:

Regulation of *Escherichia coli* Hemolysin E Expression by H-NS and *Salmonella* SlyA

Neil R. Wyborn, Melanie R. Stapleton, Valia A. Norte, Ruth E. Roberts,
Jamie Grafton, and Jeffrey Green*

*The Krebs Institute for Biomolecular Research, Department of Molecular Biology and Biotechnology,
University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom*

Received 30 June 2003/Accepted 5 December 2003

The *Escherichia coli* *hlyE* gene (also known as *clyA* or *sheA*) codes for a novel pore-forming toxin. Previous work has shown that the global transcription factors FNR and CRP positively regulate *hlyE* expression by binding at the same site. Here in vivo transcription studies reveal that FNR occupies the *hlyE* promoter more frequently than CRP, providing a mechanism for the moderate upregulation of *hlyE* expression in response to two distinct environmental signals (oxygen and glucose starvation). It has been reported that H-NS interacts with two large regions of the *hlyE* promoter (*PhlyE*), one upstream of the σ_{35} element and one downstream of the σ_{10} element. Here we identify two high-affinity H-NS sites, H-NS I, located at the 3' end of the extended upstream footprint, and H-NS II, located at the 5' end of the extended downstream footprint. It is suggested that these high-affinity sites initiate the progressive formation of higher order complexes, allowing a range of H-NS-mediated regulatory effects at *PhlyE*. Finally, the identification of a SlyA binding site that overlaps the H-NS I site in *PhlyE* suggests a mechanism to explain how SlyA overproduction enhances *hlyE* expression by antagonizing the negative effects of H-NS.

Recently, a novel pore-forming toxin, designated HlyE, ClyA, or SheA, was identified for *Escherichia coli* and *Salmonella enterica* serovars Typhi and Paratyphi A (2, 6, 7, 10, 16, 17, 21, 22, 35). The three-dimensional (3-D) structure of HlyE shows that it is a mostly α -helical, long (~ 100 Å), rod-shaped molecule with a hydrophobic two-stranded antiparallel β -sheet at one end that is thought to allow an interaction between HlyE and target membranes (35). The HlyE protein forms pores in target membranes that appear in electron microscopy (EM) images as ring-shaped structures with internal diameters of 50 to 55 Å, when viewed from above, and as 100- to 105-Å spikes in a side view, suggesting that HlyE does not undergo large conformational changes during pore formation (35). Two members of the CRP family of transcription factors control the expression of *hlyE* in *E. coli* K-12 by binding at the same site centered at ~ 61.5 bp upstream of the *hlyE* transcriptional start (10, 36). Thus, CRP enhances *hlyE* expression in response to glucose starvation (36) and FNR enhances *hlyE* expression in response to oxygen starvation (10, 11, 23). Both CRP and FNR are ~ 50 -kDa homodimers that bind related inverted repeats with a TGANNNNNTCA core motif (12). At the *hlyE* promoter (*PhlyE*), this site (TTTGATATTTATC ATA) most closely resembles an FNR site (9 of 10 nucleotides match the FNR consensus, TTGATNNNNATCAA, compared to 8 of 10

matches to the CRP consensus, TGTGANNNN NNTCACA; discriminatory bases are underlined). However, it has been shown that in certain circumstances CRP can recognize FNR sites, although the affinity of CRP for an FNR site is 50-fold lower than that for an equivalent CRP site (27). A further layer of regulation is provided by the nucleoid structuring protein H-NS (36). The H-NS protein influences the expression of many genes in *E. coli* K-12. It is a small ~ 15 -kDa protein that forms higher order complexes in a concentration-dependent manner (30, 34). The H-NS protein was shown to interact with a large region of *PhlyE* (from ~ 137 to ~ 172 , relative to the transcription start site) to repress *hlyE* expression (36) after the observation that an *hns* mutant strain has a hemolytic phenotype (9). A hemolytic phenotype was also conferred upon *E. coli* K-12 by the overproduction of either the *E. coli* or *Salmonella enteric* transcription factor SlyA (16, 21). The SlyA protein is a member of the MarR family of transcription factors that includes MarR and EmrR (*E. coli*), PecS (*Erwinia chrysanthemi*), HprR (*Bacillus subtilis*), and RovA (*Yersinia enterocolitica*) (20,24). The 3-D structures of *E. coli* MarR and a SlyA-like protein from *Enterococcus faecalis* provide the structural archetypes for this family of proteins. The 3-D structures show that they are homodimers in which each subunit possesses a winged-helix DNA-binding domain (1, 37). A recent characterization of the *Salmonella* SlyA protein

revealed that it is also a homodimer (~32 kDa) that recognizes an inverted repeat sequence in target promoters (32). Site-directed mutagenesis of *PhlyE* led to the suggestion that a GC-rich sequence located between an unusual heptameric -10 element (TATGAAT) and a conventional -35 element might

be the site of SlyA action (17). Here we show that the regulation of *hlyE* expression by H-NS is more complex than was previously thought and that the overproduction of SlyA enhances *hlyE* expression by antagonizing the negative effects of H-NS.

TABLE 1. Bacterial strains and plasmids used in this study

Strain or plasmid	Relevant characteristics	Source or reference
Strains		
M182	<i>Δlac</i>	5
JRG4830	M182 <i>Δlac Δcrp</i>	5
JRG4864	M182 <i>Δlac Δhns</i> ; Cm ^r	This work
JRG2631	M182 <i>Δlac Δfnr Δcrp</i> ; Tet ^r	Jeff Cole
JRG4865	M182 <i>Δlac Δfnr Δhns</i> ; Cm ^r Tet ^r	This work
JRG4732	M182 <i>Δlac Δcrp Δhns</i> ; Cm ^r	This work
JRG4866	M182 <i>Δlac Δfnr Δcrp Δhns</i> ; Cm ^r Tet ^r	This work
JRG4747	M182 <i>Δlac Δfnr</i> ; Cm ^r	This work
JRG3212	M182 <i>Δlac Δfnr Δcrp</i> ; Cm ^r	This work
JRG4385	JM109 (pGS1482); Ap ^r	32
Plasmids		
pRS415	Medium-copy-number vector (15 to 20 copies per cell; pBR322 origin of replication) for construction of <i>hlyE::lacZ</i> reporter plasmid pGS1629; Ap ^r	29
pGS1629	pRS415 containing <i>PhlyE</i> (-299 to +82); Ap ^r	This work
pRW50	Low-copy-number (2 to 5 copies per cell; pSC101 origin of replication), <i>lac</i> reporter vector; Tet ^r	15
pGS1065	pRW50 containing <i>PhlyE</i> (-97 to +61); Tet ^r	10
pGS1064	pUC118 containing <i>PhlyE</i> (-97 to +61, start codon is +1); Ap ^r	10
pGS1482	pGEX-KG (13) containing the <i>slyA</i> coding region; Ap ^r	32
pGS1657	pBluescript containing <i>S. enterica slyA</i> coding region and promoter; Ap ^r	This work

MATERIALS AND METHODS

Bacterial strains, plasmids, and microbiological methods. Relevant characteristics of the bacterial strains and plasmids used are given in Table 1. Isogenic derivatives of M182 were constructed by P1vir-mediated transduction. Bacteria were grown in Lennox broth (L-broth; contains yeast extract, 5 g liter⁻¹; tryptone, 10 g liter⁻¹; and NaCl, 5 g liter⁻¹) at 37°C supplemented with glucose (0.2% [wt/vol]), ampicillin (150 μg ml⁻¹), tetracycline (35 μg ml⁻¹), and chloramphenicol (20 μg ml⁻¹), as appropriate. For -galactosidase activity measurements (19), anaerobic cultures were grown in sealed bottles filled to the neck with medium, and aerobic cultures were grown in conical flasks (250 ml) containing medium (10 ml), with vigorous shaking (250 rpm) at 37°C. Hemolytic activities were estimated by measuring the areas of hemolysis surrounding individual colonies after 16 h of growth at 37°C on blood agar under aerobic conditions.

Gel shift assays. Initial H-NS gel shifts used a PCR-amplified 696-bp fragment of *PhlyE* containing all of the previously reported sequences protected by H-NS from DNase I digestion (36) plus additional upstream and downstream sequences. In subsequent experiments, this fragment was further resolved into four

subfragments by restriction digestion with *SspI*, *BsaMI*, and *DraI*. Target DNAs (200 to 500 ng) were incubated with H-NS (0 to 6 μM) in 25 mM HEPES, pH 7.6, containing 0.1 mM EDTA, 10% (vol/vol) glycerol, 5 mM dithiothreitol, 50 mM KCl, and 0.01 mg of poly(dI-dC) ml⁻¹ for 15 min at 25°C in a total volume of 10 or 20 μl. Complexes were separated in 6% (wt/vol) nondenaturing Trisborate-EDTA-buffered polyacrylamide gels for the long *PhlyE* fragment or in 12.5% (wt/vol) gels for the dissected

PhlyE fragments. DNA was visualized by staining with ethidium bromide. The interaction of H-NS and SlyA with *PhlyE* was also investigated by using *PhlyE* DNA amplified by PCR with pGS1064 (10) (Table 1) as the template and with pUC/M13 forward and reverse primers. The product was digested with *BamHI* and radiolabeled with Klenow enzyme and [32P]-dGTP (26). For H-NS, 10 ng of DNA and 0.04 to 2.0 μM H-NS were coincubated for 5 min at 25°C in a solution containing 10 mM Tris-HCl (pH 7.6),

50 mM MgCl₂, 2 mM spermidine, and 15 mM potassium glutamate (total incubation volume, 10 μl). For SlyA, 10 ng of DNA and 0.06 to 0.63 μM SlyA protein were coincubated for 2 min at 25°C in a solution containing 10 mM

Tris-HCl (pH 9.0), 50 mM KCl, and 0.1% (vol/vol) Triton X-100 (total incubation volume, 10 μl). Reactions were then loaded onto gels for autoradiographic analysis.

DNase I footprinting. For H-NS, the reactions (total volume, 10 μl) contained radiolabeled *PhlyE* (10 ng), H-NS (1 μM), 10 mM Tris-HCl (pH 7.6), 50 mM MgCl₂, 2 mM spermidine, and 15 mM potassium glutamate. For SlyA, the reactions (total volume, 10 μl) contained radiolabeled *PhlyE* (10 ng), SlyA (2.0 and 4.0 μM), 20 mM Tris-HCl (pH 8.0), 10 mM MgCl₂, 10 mM dithiothreitol, and 5% (vol/vol) glycerol. The mixtures were incubated for 2 to 5 min at 25°C, followed by digestion with DNase I (1 μl of a 1-U μl⁻¹ solution for 5 to 60 s at 25°C). Reactions were stopped by the addition of 200 μl of 0.3 M sodium acetate (pH 5.2) containing 20 mM EDTA, followed by phenol-chloroform extraction. The DNA was ethanol precipitated and resuspended in 10 μl of loading buffer (80% [vol/vol] formamide, 0.1% [wt/vol] sodium dodecyl sulfate, 10% [vol/vol] glycerol, 8 mM EDTA, 0.1% [wt/vol] bromophenol

blue, and 0.1% [wt/vol]) xylene cyanol) for electrophoretic fractionation on 6% (wt/vol) polyacrylamide– urea gels and autoradiographic analysis. Maxam and Gilbert G tracks were used to provide a calibration (18).

Other methods. The H-NS protein was provided by C. F. Higgins (MRC Clinical Sciences Centre, Hammersmith Hospital, London, United Kingdom), and the SlyA protein was isolated from *E. coli* strain JRG4385 (Table 1) as previously described (32). The manipulation of DNA, PCR, and plasmid constructions was achieved by conventional methods (26). Plasmid copy numbers were estimated by the method of Taylor and Brose (33).

CRP made significant positive contributions to *hlyE* expression, whereas the *hns* mutation resulted in slightly reduced *hlyE* transcription under these conditions (Fig. 1a). It was predicted that the addition of glucose (0.2% [wt/vol]) to the medium would abolish the effects of CRP. Accordingly, *hlyE* expression was reduced in cultures of the parental strain containing glucose compared to those without glucose (Fig. 1b). Moreover, while *hlyE* expression was reduced in the *fnr* strain compared to that in the parent strain in the presence of glucose, it was, as expected, unaffected in the *crp* mutant (Fig. 1b). The transcription of *hlyE* in the *hns* strain under anaerobic conditions in the presence of glucose is presumably

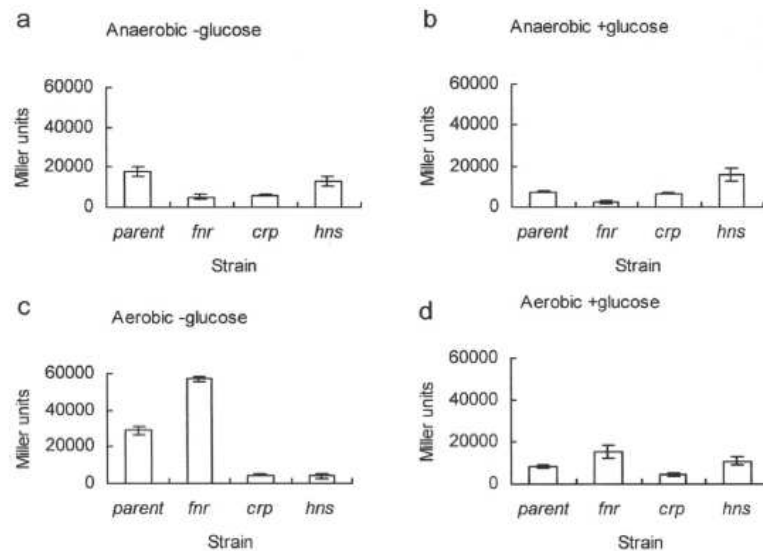


FIG. 1. Effects of FNR, CRP, and H-NS on in vivo transcription of *hlyE*. All cultures were grown in L-broth at 37°C for ~16 h under the indicated conditions. The activity of the *hlyE* promoter was estimated by measuring the β -galactosidase activity associated with each culture carrying the *hlyE::lacZ* plasmid pGS1629. Strains: parental, M182; *fnr*, JRG4747; *crp*, JRG4830; *hns*, JRG4864 (Table 1). Error bars indicate 2 standard deviations from the means ($n = 3$).

RESULTS

Effects of the global transcription factors FNR, CRP, and H-NS on *hlyE* expression in liquid culture. Previous studies have shown that FNR, CRP, and H-NS contribute to *hlyE* expression when *E. coli* is grown on a solid medium (10, 36). Both FNR and CRP make positive contributions, whereas H-NS acts negatively under these conditions. Because bacteria grown on solid media are exposed to different microenvironments due to the generation of concentration gradients of nutrients and oxygen within colonies, the starting point for this work was to investigate the effects of all three global regulators on *hlyE* expression in liquid cultures by using a plasmid-based *hlyE::lacZ* gene fusion (pGS1629; Table 1). The reporter plasmid contained a fragment of DNA stretching from _299 bp to _82 bp relative to the *hlyE* transcript start site ligated with pRS415 (Table 1). Initially, transcriptional activity from *PhlyE* was estimated for anaerobic cultures of parental (M182), *fnr*, *crp*, and *hns* strains transformed with pGS1629 (Table 1), after growth in L-broth at 37°C for 16 h. The data revealed that both FNR and

driven by FNR alone. Under these conditions, *hlyE* expression was greater for the *hns* strain than for the corresponding parental cultures, suggesting that H-NS has a negative effect on FNR-driven *hlyE* expression (Fig. 1b). Thus, we concluded that CRP, FNR, and H-NS all contribute towards the regulation of *hlyE* expression. Similar experiments were done with aerobic cultures. For these conditions, we predicted that the glucose-responsive effects of CRP would be retained, whereas the positive regulation by FNR would be abolished. The prediction for the CRP response was confirmed by the data (Fig. 1c and d). However, the data for the parental and *fnr* cultures suggested that even under aerobic growth conditions, some active FNR was present in the bacteria, and that even this small amount of active protein (relative to anaerobic conditions) was sufficient to interfere with CRP-mediated *hlyE* expression (Fig. 1c). Thus, these data suggest that under such conditions, FNR and CRP are in competition for binding at *PhlyE*, and that the simplest explanation of these observations is that FNR is more efficiently bound than CRP at *PhlyE* but is not as effective as CRP in activating *hlyE* transcription. Such an interpretation is consistent with

the sequence of the FNR/CRP box in *PhlyE*, which lacks the discriminatory G-C base pairs that promote CRP specificity, and with the observation that, unlike CRP, FNR is a poor activator of class I promoters such as *PhlyE* (reviewed by Green et al. in reference 12). The effect of the *hns* mutation on *hlyE* expression in aerobic cultures was also interesting (Fig. 1c and d). In the absence of glucose, H-NS appeared to have a positive effect on *hlyE* expression (Fig. 1c). However, in the presence of glucose, H-NS appeared to have little effect on *hlyE* expression (Fig. 1d). Thus, under aerobic conditions, H-NS has a positive influence on CRP-mediated *hlyE* expression. This positive effect of H-NS was not observed under anaerobic conditions, presumably because FNR and not CRP occupies *PhlyE* under these conditions. To further investigate the effects of the regulators on *hlyE* expression, we monitored the activity of the *hlyE::lacZ* fusion in pGS1629 in strains carrying multiple relevant mutations. Under anaerobic

conditions in the absence of glucose, *hlyE* expression in an *fnr crp hns* triple mutant was low compared to that of the parental strain (Fig. 2a). Restoring *hns* did not affect *hlyE* expression. Similarly, restoring CRP did not enhance *hlyE* expression, despite the absence of glucose (Fig. 2a). However, the restoration of *fnr* produced a significant increase in *hlyE* expression under these conditions (Fig. 2a). This suggests that during anaerobic growth in liquid cultures in the absence of glucose, FNR is the major regulator of *hlyE* expression, and that CRP-mediated *hlyE* expression requires the presence of H-NS. The pattern of expression obtained in anaerobic cultures in the presence of glucose was similar to that obtained in the absence of glucose, except that *hlyE* expression in the *fnr crp hns* triple mutant and the *crp hns* double mutant was significantly higher than that for the corresponding cultures that lacked glucose (Fig. 2b). The *hlyE* expression patterns for equivalent aerobic cultures in the presence

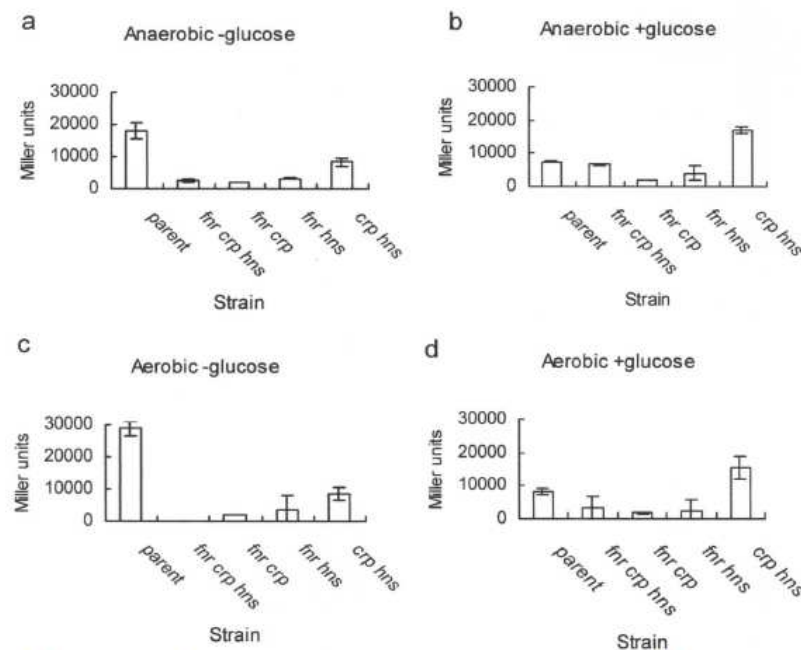


FIG. 2. Effects of FNR, CRP, and H-NS on in vivo transcription of *hlyE*. All cultures were grown in L-broth at 37°C for ~16 h under the indicated conditions. The activity of the *hlyE* promoter was estimated by measuring the β -galactosidase activity associated with each culture carrying the *hlyE::lacZ* plasmid pGS1629. Strains: parental, M182; *fnr crp hns*, JRG4866; *fnr crp*, JRG2631; *fnr hns*, JRG4865; *crp hns*, JRG4732. Error bars indicate 2 standard deviations from the means ($n = 3$).

and absence of glucose were similar to those obtained under anaerobic conditions (Fig. 2c and d). In summary, the data presented here suggest that FNR and CRP are positive regulators of *hlyE* expression in liquid culture in response to oxygen and glucose starvation, respectively. This is in agreement with observations made with cultures grown on a solid medium (36). However, rather than the approximately eightfold enhancement in *hlyE* expression observed in an *hns* mutant on solid medium (36), in liquid culture H-NS has a positive effect on CRP-driven *hlyE* expression, a negative effect on FNR-driven *hlyE* expression, and little intrinsic regulatory activity in the absence of FNR and CRP.

H-NS binds at two regions of the *hlyE* promoter with high affinity. Footprinting studies have shown that FNR and CRP activate *hlyE* expression from the same site centered 61.5 bp upstream of the SlyA-associated transcription start site (10, 36). A further footprinting analysis indicated that H-NS protects a large region of the *hlyE* promoter, extending from 137 to 172 (relative to the transcription start site) (36). To investigate how much of this extensive H-NS protection is required for the observed regulation of *hlyE* expression, we used two approaches as follows.

Firstly, a 696-bp region of the *hlyE* promoter region was amplified by PCR (Fig. 3a). Gel shift assays revealed that

H-NS retarded the mobility of this fragment upon electrophoretic separation in Tris-borate-EDTA-

buffered polyacrylamide gels (Fig. 3b). In the presence of 1 μ M H-NS, retardation of some of the DNA target was observed as a smear in the gel. Upon the addition of further H-NS (2 μ M), the free DNA was replaced entirely by a retarded smear. Only at 4 μ M H-NS was a distinct retarded complex observed (Fig. 3b). The amplified fragment contains restriction sites for *Ssp*I, *Bsa*MI, and *Dra*I, such that a triple digest yields four fragments (Fig. 3a). Gel shift assays using a mixture of *hlyE* DNA fragments as the target DNA revealed that H-NS interacts with all four *PhlyE* fragments to some extent. However, the 147-bp *Dra*I-*Bsa*MI fragment was bound with the highest affinity, with 50% of this fragment retarded at H-NS concentrations of $\sim$ 3 to 4 μ M (Fig. 3c and d). Thus, this region of *PhlyE* ($\sim$ 91 to $\sim$ 56) contains both the FNR/CRP box ($\sim$ 61.5) and the highest affinity H-NS site(s).

Secondly, for testing of whether the region of *PhlyE* containing the high-affinity H-NS site(s) was sufficient to account for the pattern of *hlyE* regulation observed in vivo, another *hlyE::lacZ* reporter plasmid was used (Table 1) (10). This second gene fusion (pGS1065) contains a minimal *hlyE* promoter beginning 18 bp downstream of the *Bsa*MI site and ending 25 bp downstream of the *Dra*I site and thus extends from $\sim$ 97 to $\sim$ 61 relative to the transcript start site in the low-copy-number (two to five copies per cell) vector pRW50 (Table 1). Cultures of strains M182 and M182 *hns* transformed with pGS1065 were grown under anaerobic conditions in the presence and absence

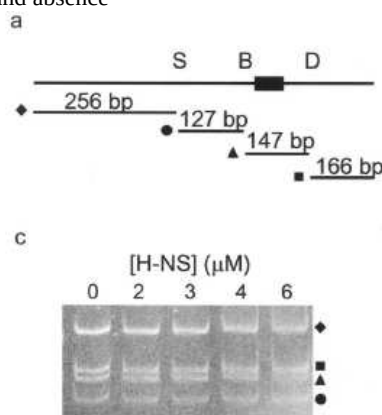


FIG. 3. Identification of high-affinity H-NS sites within *PhlyE*. (a) Schematic of the 696-bp *PhlyE* fragment. The locations of relevant restriction sites are indicated. (b) Gel shift assay showing H-NS binding to the *PhlyE* fragments. The concentration of H-NS is indicated at the top of each lane. (c) Gel shift assay showing H-NS binding to the *PhlyE* fragments. The amount of each *PhlyE* fragment remaining unretarded in panel c was estimated by quantitative densitometry and plotted against the concentration of H-NS in the assay. The symbols relate to the restriction fragments, as indicated in panel a. (d) Quantification of H-NS binding to different regions of *PhlyE*. The amount of each *PhlyE* fragment remaining unretarded in panel c was estimated by quantitative densitometry and plotted against the concentration of H-NS in the assay. The symbols relate to the restriction fragments, as indicated in panel a.

of glucose at 37°C for $\sim$ 16 h. The transcriptional activity in vivo, as estimated by the measurement of β -galactosidase activity, indicated that this (pGS1065) low-copy-number reporter containing a minimal *hlyE* promoter responded in the same way as the longer medium-copy-number *hlyE::lacZ* fusion

(pGS1629) to the presence or absence of H-NS (Table 2).

Thus, in the presence of glucose, *hlyE* expression was enhanced in an *hns* mutant (365 Miller units) relative to the parent (271 Miller units) (compare with Fig. 1b). In the absence of glucose, *hlyE* expression was slightly lower in the *hns* mutant (399 Miller units) than in the parent (420 Miller units) (compare with Fig. 1a). Furthermore, there were no significant H-NS-related changes in reporter plasmid copy number (not shown), as judged by the method of Taylor and Brose (33). Thus, it was concluded that the *Bsa*MI-*Dra*I fragment of *PhlyE* contains all of the significant regulatory elements that control *hlyE* expression under the growth conditions used here and that the effects on *hlyE* expression observed with pGS1629 were not significantly affected by the copy number of the reporter plasmids.

Identification of the high-affinity H-NS sites in *PhlyE*. For determination of the number of H-NS sites within the minimal *PhlyE* sequence from pGS1065, further gel shift and footprinting assays were undertaken. The gel shifts showed that H-NS bound to *PhlyE* at concentrations close to those observed for other H-NS-regulated genes ($\sim$ 1 μ M) (3) and that two distinct retarded complexes were formed (Fig. 4a). The locations of the H-NS binding sites were determined by DNase I footprinting (Fig. 4b). Two protected regions were detected: H-NS I, consisting of an A-T-rich region (20 A-T/24 bp) stretching from $\sim$ 75 to $\sim$ 37, which overlaps the FNR/CRP box centered at $\sim$ 61.5, and H-NS II, between $\sim$ 22 and $\sim$ 11, overlapping the $\sim$ 10 element of *PhlyE*. The H-NS II site contains a hypersensitive base at position $\sim$ 17 that is separated by several unprotected bases ($\sim$ 16 to $\sim$ 18) from another protected A-T-rich region (8 A-T/10 bp) (Fig. 4b). Based upon the size and properties of H-NS and the extent of the DNase I footprint, it is likely that several H-NS molecules bind *PhlyE*.

TABLE 2. Effect of SlyA and H-NS on transcription from a minimal *hlyE* promoter in vivo

Strain ^a	Data without glucose		Data with glucose	
	Amt of β -galactosidase (Miller units) ^b	Ratio ^c	Amt of β -galactosidase (Miller units) ^b	Ratio ^c
Parent + vector	420 $\pm$ 11	1.46	271 $\pm$ 6	1.14
Parent + SlyA	614 $\pm$ 10		310 $\pm$ 21	
<i>hns</i> + vector	399 $\pm$ 10	0.91	365 $\pm$ 15	0.82
<i>hns</i> + SlyA	365 $\pm$ 17		299 $\pm$ 38	
<i>crp fmr</i> + vector	257 $\pm$ 6	0.87	152 $\pm$ 2	1.11
<i>crp fmr</i> + SlyA	223 $\pm$ 7		169 $\pm$ 2	

^a Strains containing the low-copy-number minimal *hlyE::lacZ* promoter fusion plasmid (pGS1065) were grown under anaerobic conditions for 16 h at 37°C in L-broth with or without glucose (0.2% [w/vol]) as indicated. The strains used were as follows: parent, M182(pGS1065); *hns*, JRG4864(pGS1065); *crp fmr*, JRG3212(pGS1065) (Table 1). SlyA indicates the presence of multicopy *slyA* (pGS1657), and vector indicates the presence of the vector (pBluescript).

^b Means ($n = 3$) and standard deviations are given.

^c Ratio of β -galactosidase activities obtained with multicopy *slyA* and vector of transformants.

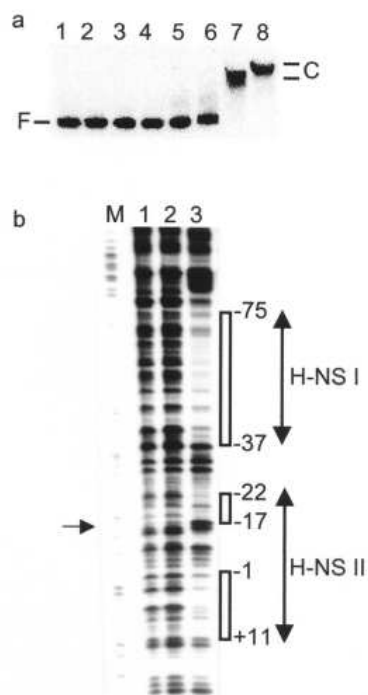


FIG. 4. Location of H-NS binding sites in *PhlyE*. (a) H-NS binds at *PhlyE*. Lane 1, no protein; lanes 2 to 8, H-NS (0.04, 0.08, 0.12, 0.25, 0.5, 1.0, and 2.0 μ M, respectively). The positions of the retarded complexes (C) and free *PhlyE* DNA (F) are indicated. (b) DNase I footprints of H-NS at *PhlyE*. Lane M, Maxam and Gilbert G track; lanes 1 and 2, no protein; lane 3, H-NS (1 μ M). The regions of H-NS protection (open boxes) and the hypersensitive base (arrow at the left) are indicated. All numbering is relative to the SlyA-associated transcription start site (36).

Effects of H-NS on SlyA-promoted *hlyE* expression.

Overproduction of the *E. coli* or *S. enterica* SlyA protein confers a hemolytic phenotype on *E. coli* K-12 by enhancing *hlyE* expression (16, 17, 21). To test the effects of H-NS on SlyA-driven *hlyE* expression, we measured β -galactosidase activities from cultures of isogenic parental and *hns* strains carrying the *hlyE::lacZ* plasmid pGS1065 and either a multicopy *S. enterica* SlyA expression plasmid (pGS1657; Table 1) or, as a control, the vector (pBluescript). As expected, irrespective of the addition of glucose, the presence of multicopy *slyA* increased *hlyE* expression in cultures of the parental strain, although the enhancement was greater for cultures lacking the glucose supplement (Table 2). In contrast, in the absence of *hns*, SlyA overproduction did not enhance *hlyE* expression (Table 2). This pattern of expression was supported by qualitative studies in which the same strains were grown on blood agar plates. These revealed that multicopy *slyA* enhanced hemolytic activity associated with the parental strain but not with the *hns* strain (not shown). These results confirm and extend the observations of Westermarck et al. (36) and suggest that, either directly or indirectly, SlyA activates *hlyE* expression by modulating H-NS binding and ultimately by counteracting the negative effects of H-NS. This suggestion was supported by the observations that multicopy *slyA* did not enhance *hlyE* expression in a *crp fnr* double mutant (Table 2) and that SlyA did not promote open complex

formation at *PhlyE* in vitro (not shown). Thus, it was concluded that SlyA, like H-NS, has no intrinsic activator role at *PhlyE* in the absence of FNR and CRP.

Interaction of SlyA with the *E. coli hlyE* promoter.

The pattern of regulation reported above suggests that SlyA, like H-NS, interacts directly with the *E. coli hlyE* promoter. Therefore, gel shift assays were used to investigate the interaction of purified SlyA with *PhlyE*. These experiments revealed that the addition of 0.16 to 0.31 μ M SlyA was sufficient to retard the mobility of *PhlyE* (Fig. 5a). A consensus SlyA DNA target was recently suggested to exist (32), and an inspection of the *PhlyE* sequence revealed at least two possible matches. Therefore, the location of the SlyA binding site(s) within *PhlyE* was investigated by DNase I footprinting (Fig. 5b). A protected region, -70 to -38 relative to the previously determined SlyA-associated transcript start site (17, 36), was observed in the presence of 2 μ M SlyA. This region of protection was extended further downstream to -31 when the concentration of SlyA was increased to 4 μ M (Fig. 5b). Thus, SlyA protects a region of *PhlyE* similar to the H-NS I region (-75 to -37 ; Fig. 4b). The cross sections of the *E. coli* MarR and *E. faecalis* SlyA-like protein dimers are -70 Å² (1, 37) and therefore would be expected to protect ~ 20 bp of DNA. Thus, the protection observed here (up to 40 bp) suggests that two SlyA dimers are bound at *PhlyE*. An inspection of the DNA sequence of the protected region revealed two related sequences with partial dyad symmetry that re-

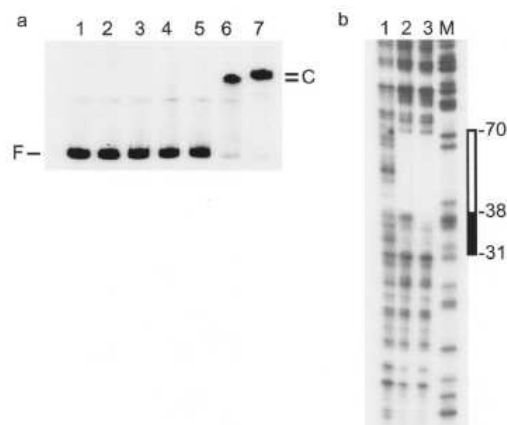


FIG. 5. Interaction of SlyA with the *hlyE* promoter. (a) SlyA retards the mobility of *PhlyE* in band shift assays. SlyA concentrations: lane 1, 0 μ M; lane 2, 0.06 μ M; lane 3, 0.08 μ M; lane 4, 0.1 μ M; lane 5, 0.16 μ M; lane 6, 0.31 μ M; lane 7, 0.63 μ M. The positions of the free DNA (F) and the SlyA-DNA complexes (C) are indicated. (b) DNase I footprint of SlyA at *PhlyE*. Lane 1, no protein; lane 2, 2.0 μ M SlyA; lane 3, 4.0 μ M SlyA; lane M, Maxam and Gilbert G track. The region of SlyA protection is indicated to the right (box). At the higher SlyA concentration, protection was extended, as indicated by the filled box.

semble the proposed SlyA binding consensus; they are SlyAI ($_{61}$ TTATCATATTAA $_{50}$, with 8 of 12 bases (bold) matching the SlyA binding site consensus TTAGCAAGCTAA) and SlyAII ($_{50}$ ATAGAAATAAAG $_{39}$, with 6 of 12 bases matching the consensus). The putative presence of two SlyA dimers at *PhlyE* was consistent with the

two distinct complexes with retarded mobilities in the gel shift assays (Fig. 5a). The site centered at -55.5 is the better match to the SlyA binding site consensus and may be occupied in preference to the site centered at -44.5 , which accordingly is less similar to the consensus.

DISCUSSION

The starting point for the work described here was the observation that lesions in *hns* or the overproduction of *S. enteric* or *E. coli* SlyA confers a hemolytic phenotype on *E. coli* K-12 by enhancing the expression of the pore-forming toxin HlyE (9, 16, 17, 21). Here we have shown that H-NS and SlyA interact directly with the *hlyE* promoter. Both proteins occupy a common region of *PhlyE* that overlaps the binding site for the global transcription factors FNR and CRP, which are known to activate *hlyE* expression (10, 36). In vivo and in vitro evidence

suggests that SlyA activates *hlyE* expression by antagonizing H-NS-mediated repression. The in vivo transcriptional evidence presented here suggests that FNR occupies the *hlyE* promoter more frequently than CRP even under aerobic conditions. This is evident from enhanced *hlyE::lacZ* expression in aerobic cultures (compare parent, *fnr*, and *crp* strains in Fig. 1c and d). These observations may be explained by a mechanism in which FNR recognizes *PhlyE* efficiently but acts only as a relatively poor activator of *hlyE* transcription (10, 11), whereas, conversely, the recognition of *PhlyE* by CRP is poor, but once CRP is bound, CRP-mediated activation of *hlyE* expression is efficient. This reciprocity of binding site recognition and transcriptional efficiency provides a mechanism for the moderate upregulation of *hlyE* expression in response to two distinct environmental signals (oxygen and glucose starvation) rather than, for example, the much larger degree of upregulation when FNR and CRP act synergistically by binding at different sites within the *ansB* promoter (28). In principle, any pair of transcription factors that recognize similar DNA sequences but have different transcriptional efficiencies could adopt this strategy.

Under some of the conditions studied here, the effects of an *hns* lesion on *hlyE::lacZ* expression in liquid cultures were different from those reported by Westermarck et al. for bacteria grown on a solid medium (36). This is perhaps not surprising considering the significant physiological differences between cultures grown in liquid and on solid media. In the previous report (36), H-NS acted as a strong repressor of *hlyE* expression, whereas here H-NS appears to have a positive effect on *hlyE* expression in the absence of glucose but an attenuated or negative effect on *hlyE* expression in the presence of glucose (Fig. 1). This suggests that, in general, H-NS inhibits FNR-driven

hlyE expression (Fig. 1) but enhances CRP-driven expression in liquid cultures (compare Fig. 1b and c). Moreover, any enhancement of *hlyE::lacZ* expression in the *hns* strain observed here did not approach the

eightfold increase observed on solid medium (36). Thus, it appears that H-NS modulates *PhlyE* activity both positively and negatively in response to the prevailing growth and/or environmental conditions, such as growth in liquid and on solid media. This may be a reflection of the formation of different H-NS-*PhlyE* complexes in response to environmental signals. This idea is supported by the identification of two regions within the larger previously described H-NS-*PhlyE* complex (36) that bind H-NS with high affinities. The first region (H-NS I) overlaps with, and extends downstream from, the FNR/CRP site (Fig. 6). The extent of protection associated with the H-NS I site (38 bp) suggests that more than one H-NS dimer is bound, and the formation of different subcomplexes could potentially act to modulate FNR/CRP-driven transcription activation either positively (occupation of the downstream portion of H-NS I; thick arrow in Fig. 6) or negatively (occupation of sequences overlapping the FNR/CRP-binding site; thin arrow in Fig. 6). While H-NS usually acts as a negative regulator of gene expression, *hlyE* is not the only example of a CRP-regulated gene whose expression is positively regulated by H-NS; for example, expression of both the *malT* and *csiD* genes in *E. coli* is stimulated by H-NS (8, 14). Thus, it would appear that a CRP family transcription factor plus H-NS is a versatile combination, providing the means to generate a range of regulatory effects. The second region of H-NS protection (H-NS II) overlaps the basic promoter elements, and H-NS bound at this region would be expected to repress *hlyE* expression by promoter occlusion. The findings that SlyA does not enhance *hlyE* expression in *hns* cultures grown in liquid medium under anaerobic conditions and that SlyA appears unable to activate *hlyE* expression in the absence of FNR and CRP and the determination of the location of the region of *PhlyE* occupied by SlyA (Fig. 6) suggest that SlyA activates *hlyE* expression by antagonizing the negative action of H-NS. Previous studies have indicated that the intracellular levels of H-NS change with the growth phase. However, whereas one proteomics-based study suggested a fivefold increase in H-NS in stationary-phase cultures (31), a more recent immunological analysis suggested a twofold decrease (4). In addition, we have shown that intracellular H-NS levels increase as growth temperature decreases, as evidenced by an approximately twofold increase in H-NS immunoblot signal intensity from cultures grown at 20°C compared to those grown at 37°C (N. R. Wyborn and J. Green, unpublished data). Thus, it appears that intracellular H-NS levels are influenced by the environment. Therefore, we suggest that when environmental conditions dictate a low intra-



FIG. 6. Schematic representation of the *hlyE* promoter region. The locations of the FNR/CRP box and SlyA sites are shown. The high-affinity H-NS binding regions observed in DNase I footprints of *PhlyE* are also shown. The transcript start site (+1) is marked. The region of the H-NS I site that extends downstream of the FNR/CRP box is indicated by a thick arrow, and the part overlapping the FNR/CRP box is indicated by a thin arrow.

cellular concentration of H-NS, the H-NS I region of *PhlyE* is occupied to modulate the activity of the upstream activator (FNR or CRP). As the intracellular levels of H-NS increase, H-NS II is occupied and *hlyE* expression is downregulated by promoter occlusion. The footprinting studies described by Westermarck et al. (36) indicated that further increases in H-NS levels result in the formation of higher order complexes that occupy *PhlyE* from the $_{10}$ element to position $_{172}$ and from just upstream of the $_{35}$ element to position $_{137}$. Thus, it appears that these higher order complexes extend from the primary sites of interaction (H-NS I and H-NS II) identified here. Such higher order complexes, which effectively silence *hlyE* expression, appear to have directionality in that they extend upstream from H-NS I (to as far as position $_{137}$) to occlude the FNR/CRP site and downstream from H-NS II (to as far as $_{172}$) to occlude the basic promoter elements. This leaves the region between the two primary sites of interaction unoccupied. Thus, we suggest that the formation of different H-NS-*PhlyE* complexes offers the opportunity for H-NS, in combination with FNR and CRP, to control *hlyE* expression both positively and negatively. This type of behavior, in which specific patterns of protection are replaced by general protection as the concentration of H-NS increases, has been observed for footprints of *gal* promoter variants (25). It was also shown previously that this general protection was dependent on the polymerization of H-NS on the DNA and that this was more likely if there was an initial nucleation event on the DNA at specific sites at low H-NS concentrations (25). It appears that the H-NS I and H-NS II regions of *PhlyE* are such nucleation sites. By competing with H-NS for the region of *PhlyE* downstream of the FNR/CRP site, SlyA may prevent the formation of the negatively acting H-NS-*PhlyE* complexes by blocking a primary interaction between H-NS and *PhlyE*, allowing FNR and CRP to continue to activate *hlyE* expression. Presumably, the action of the *E. coli* SlyA protein will prove to be similar to that of the *Salmonella* protein studied here, given that their primary structures are 89% identical and that the overproduction of *E. coli* SlyA also enhances *hlyE* expression (16, 21). In summary, we have shown that the H-NS-mediated regulation of *hlyE* expression in *E. coli* K-12 is more complex than was previously suggested because H-NS can contribute positively as

well as negatively to *hlyE* expression. This range of regulatory activity is probably associated with the progressive formation of higher order H-NS-*PhlyE* complexes that extend from two primary sites of interaction to ultimately bring about silencing of the *hlyE* gene. The H-NS and SlyA footprints suggest that H-NS regulation is overcome by the overproduction of SlyA, which by binding at a site that overlaps one of the primary H-NS sites (H-NS I), prevents the formation of repressive H-NS-*PhlyE* complexes. Further detailed in vitro analyses will be required to fully analyze the complex relationships between H-NS, SlyA, FNR, and CRP and their consequences for *hlyE* expression.

ACKNOWLEDGMENTS

We acknowledge A. J. G. Moir for DNA sequencing. We thank Jeff Cole and Steve Busby (University of Birmingham, Birmingham, United Kingdom) and Bernt Eric Uhlin (Umea University, Umea, Sweden) for providing bacterial strains. We thank Chris Higgins (Hammersmith Hospital, London, United Kingdom) for providing the H-NS protein. This work was supported by the Biotechnology and Biological Sciences Research Council, Swindon, United Kingdom.

REFERENCES

- Alekshun, M. N., S. B. Levy, T. R. Mealy, B. A. Seaton, and J. F. Head. 2001. The crystal structure of MarR, a regulator of multiple antibiotic resistance, at 2.3 Å resolution. *Nat. Struct. Biol.* **8**:710–714.
- Atkins, A., N. R. Wyborn, A. J. Wallace, T. J. Stillman, L. K. Black, A. B. Fielding, M. Hisakado, P. J. Artymiuik, and J. Green. 2000. Structure-function relationships of a novel bacterial toxin, hemolysin E. *J. Biol. Chem.* **275**:41150–41155.
- Atlung, T., and H. Ingmer. 1997. H-NS: a modulator of environmentally regulated gene expression. *Mol. Microbiol.* **24**:7–17.
- Azam, T. A., A. Iwata, A. Nishimura, S. Ueda, and A. Ishihama. 1999. Growth phase-dependent variation in protein composition of the *Escherichia coli* nucleoid. *J. Bacteriol.* **181**:6361–6370.
- Busby, S., D. Kotlarz, and H. Buc. 1983. Deletion mutagenesis of the *Escherichia coli* galactose operon promoter region. *J. Mol. Biol.* **167**:259–274.
- del Castillo, F. J., S. C. Leal, F. Moreno, and I. del Castillo. 1997. The *Escherichia coli* K-12 *sheA* gene encodes a 34-kDa secreted haemolysin. *Mol. Microbiol.* **25**:107–115.
- Fernandez, S. V., J. Xing, V. Kapur, S. J. Libby, R. G. Barletta, and R. A. Moxley. 1998. Regulation of the *Escherichia coli* *sheA* gene and characterization of its encoded hemolytic activity. *FEMS Microbiol. Lett.* **168**:85–90.
- Germer, J., G. Becker, M. Metzner, and R. Hengge-Aronis. 2001. Role of activator site position and a distal UP-element half-site for sigma factor selectivity at a CRP/H-NS-activated s-dependent promoter in *Escherichia coli*. *Mol. Microbiol.* **41**:705–716.
- Gomez-Gomez, J. M., J. Blazquez, F. Baquero, and J. L. Martinez. 1996. *Hns* mutant unveils the presence of a latent haemolytic activity in *Escherichia coli* K-12. *Mol. Microbiol.* **19**:909–910.
- Green, J., and M. Baldwin. 1997. The molecular basis for the differential regulation of the *hlyE*-encoded haemolysin of *Escherichia coli* by FNR and HlyX lies in the improved activating region 1 contact of HlyX. *Microbiology* **143**:3785–3793.
- Green, J., and M. Baldwin. 1997. HlyX, the FNR homologue of *Actinobacillus pleuropneumoniae*, is a [4Fe-4S]-containing oxygen-responsive transcription regulator that anaerobically activates FNR-dependent class I promoters via an enhanced AR1 contact. *Mol. Microbiol.* **24**:593–605.
- Green, J., C. Scott, and J. R. Guest. 2001. Functional versatility in the CRP-FNR superfamily of transcription factors: FNR and FLP. *Adv. Microb. Physiol.* **44**:1–34.

-
13. Guan, K., and J. E. Dixon. 1991. Eukaryotic proteins expressed in *Escherichia coli*—an improved thrombin cleavage and purification procedure of fusion proteins with glutathione-S-transferase. *Anal. Biochem.* **192**:262–267.
 14. Johansson, J., B. Dagberg, E. Richet, and B. E. Uhlin. 1998. H-NS and StpA proteins stimulate expression of the maltose regulon in *Escherichia coli*. *J. Bacteriol.* **180**:6117–6125.
 15. Lodge, J., R. Williams, A. Bell, B. Chan, and S. Busby. 1990. Comparison of promoter activities in *Escherichia coli* and *Pseudomonas aeruginosa*—use of a new broad host range promoter probe plasmid. *FEMS Microbiol. Lett.* **67**:221–225.
 16. Ludwig, A., C. Tengler, S. Bauer, A. Bubert, R. Benz, H. Mollenkopf, and W. Goebel. 1995. SlyA, a regulatory protein from *Salmonella typhimurium*, induces a haemolytic and pore-forming protein in *Escherichia coli*. *Mol. Gen. Genet.* **249**:474–486.
 17. Ludwig, A., S. Bauer, R. Benz, B. Bergmann, and W. Goebel. 1999. Analysis of the SlyA-controlled expression, subcellular localization and pore-forming activity of a 34 kDa haemolysin (ClyA) from *Escherichia coli* K-12. *Mol. Microbiol.* **31**:557–567.
 18. Maxam, A. M., and W. Gilbert. 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.* **65**:499–560.
 19. Miller, J. H. 1972. Experiments in molecular genetics, p. 352–355. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
 20. Miller, P. F., and M. C. Sulavik. 1996. Overlaps and parallels in the regulation of intrinsic multiple-antibiotic resistance in *Escherichia coli*. *Mol. Microbiol.* **21**:441–448.
 21. Oscarsson, J., Y. Mizunoe, B. E. Uhlin, and D. J. Haydon. 1996. Induction of haemolytic activity in *Escherichia coli* by the slyA gene product. *Mol. Microbiol.* **20**:191–199.
 22. Oscarsson, J., Y. Mizunoe, L. Li, X.-H. Lai, A. Weislander, and B. E. Uhlin. 1999. Molecular analysis of the cytolytic protein ClyA (SheA) from *Escherichia coli*. *Mol. Microbiol.* **32**:1226–1238.
 23. Ralph, E. T., J. R. Guest, and J. Green. 1998. Altering the anaerobic transcription factor FNR confers a hemolytic phenotype on *Escherichia coli* K-12. *Proc. Natl. Acad. Sci. USA* **95**:10449–10452.
 24. Revell, P. A., and V. L. Miller. 2000. A chromosomally encoded regulator is required for expression of the *Yersinia enterocolitica* *inv* gene and for virulence. *Mol. Microbiol.* **35**:677–685.
 25. Rimsky, S., F. Zuber, M. Buckle, and H. Buc. 2001. A molecular mechanism for the repression of transcription by the H-NS protein. *Mol. Microbiol.* **42**:1311–1323.

Appendix B:

Influence of DNA geometry on transcriptional activation in *Escherichia coli*

Sylvie Dethiollaz, Patrick Eichenberger and
Johannes Geiselmann¹

Departement de Biologie Moléculaire, Université de Genève,
30 Quai Ernest Ansermet, CH-1211 Genève 4, Switzerland

¹Corresponding author

Transcription from many *Escherichia coli* promoters can be activated by the cAMP-CRP complex bound at different locations upstream of the promoter. At some locations the mechanism of activation involves direct protein-protein contacts between CRP and the RNA polymerase. We positioned the CRP binding site at various distances from the transcription start site of the *malT* promoter and measured the *in vivo* activities of these promoter variants. From the activation profiles we deduce that the protein-protein interactions involved in transcriptional activation are rather rigid. A heterologous protein (IHF) that bends the DNA to a similar degree as does CRP activates transcription when bound at sites equivalent to activating positions for CRP. DNA geometry makes a major contribution to the process of transcriptional activation and DNA upstream of the activator binding site participates in this process. Removal of this DNA decreases the capacity of the *malT* promoter to be activated by CRP *in vitro*. We conclude that both DNA topology and direct protein-protein contacts contribute to transcriptional activation and that the relative importance of these two modes of activation depends on the nature of the activator and on the location of the activator binding site.

Keywords: cyclic AMP receptor protein/DNA bending/ DNA-protein interactions/integration host factor/RNA polymerase

Introduction

Transcription can be regulated negatively by repressors and positively by activators. Whereas the molecular mechanisms of repression are often based on steric exclusion of protein binding to the DNA (Schlax et al., 1995) and are relatively well understood, many different ways exist to activate transcription. Even a single activator, such as the cyclic-AMP receptor protein (CRP), can utilize different mechanisms depending on the promoter context. A sound understanding of the functioning of

this particular protein will therefore shed light on several activation mechanisms at once.

CRP controls transcription of a large number of *Escherichia coli* genes in response to carbon source limitation (for a review, see Kolb et al., 1993a). The protein is a dimer of two identical subunits composed of 209 amino acids (Aiba et al., 1982). In the presence of cAMP, the cAMP-CRP complex (hereinafter called CRP) will bind to a 22 bp 2-fold-symmetric recognition site located within or near target promoters (Berg and von Hippel, 1988), and thereby activate transcription from these promoters.

The structure of CRP bound to DNA is known to atomic resolution (Schultz et al., 1991) and confirms solution studies that have shown that CRP bends the DNA by about 90° upon binding to its specific binding site (Liu- Johnson et al., 1986).

The mechanism by which CRP increases the rate of transcription initiation has been intensely studied at particular promoters, but many molecular details remain unknown. It is intriguing that the centre of the DNA binding site for CRP is located at different distances from the transcription start site, e.g. -41.5, -61.5 and -70.5 bp, for the *gal*, *lac* and *malT* promoters, respectively (Taniguchi et al., 1979; Schmitz, 1981; Chapon and Kolb, 1983). Promoters can be divided into two classes with respect to activation by CRP: class II comprises promoters where the CRP binding site overlaps the -35 hexamer of the promoter; in class I promoters the CRP binding site is located upstream of the binding site for RNA polymerase, i.e. around position -61 or further upstream (Ebright, 1993). Despite the loose distance requirement, activation of transcription can take place only when CRP and RNA polymerase bind to the same face of the DNA helix (Gaston et al., 1990; Ushida and Aiba, 1990).

The precise step of transcription initiation that is activated in each case also depends on the particular promoter. Whereas CRP accelerates isomerization to

the open complex at the gal promoter (Herbert et al., 1986), it increases the affinity of the polymerase for the DNA at lac (Malan et al., 1984), and facilitates promoter escape at malt (Menendez et al., 1987). The possible physical interactions between CRP and the RNA polymerase that result in an increased rate of transcription include direct protein-protein contacts (Igarashi and Ishihama, 1991; Chen et al., 1994) and distortion of the DNA between CRP and RNA polymerase (Ryu et al., 1994). Due to the large bend induced in the DNA upon binding of CRP, interactions of the DNA upstream of the CRP binding site with the 'backside' of the RNA polymerase are conceivable (Bracco et al., 1989; Gartenberg and Crothers, 1991).

In the past few years, genetic experiments have yielded important insights into the molecular mechanism underlying transcriptional activation. Mutants of CRP that bind normally to the target site on the DNA and do not perturb bending of the DNA, but fail to activate transcription (positive control or pc-mutants), have been isolated by several laboratories. These studies show that a surface exposed loop of CRP constitutes an activating region (activating region I or ARI) for class I and class II CRP-dependent promoters (Bell et al., 1990; Eschenlauer and Reznikoff, 1991; Zhou et al., 1993, 1994). Other regions (activating regions II and III) at the surface of CRP are involved exclusively in activation at class II promoters (Williams et al., 1996). Likewise, mutants of RNA polymerase subunits have been used to establish that the C-terminal domain of the alpha (α)-subunit is the target for activation from class I promoters and both the σ and sigma (σ)-subunits interact with CRP at class II promoters (Zou et al., 1992; Ishihama, 1993). Further experiments using intergenic suppressors will certainly delineate the precise amino acids involved in these protein-protein contacts. A recent model proposed that a flexible linker connecting the C-terminal domain and the N-terminal domain of the σ -subunit could extend to maintain the interaction with the activator when the CRP binding site is moved further upstream (Zhou et al., 1994).

The cases of the lac and gal-type promoters have been intensively studied. However, much less is known about promoters where the CRP binding site is located further upstream of the transcription start site, such as the malt promoter. A direct contact between CRP and RNA polymerase could be more difficult to establish due to the increased distance between the proteins. Other types of interactions, such as bringing the DNA upstream of the CRP binding site into the proximity of RNA polymerase, may make major contributions to transcriptional activation. Very little is known about such a mechanism relying mainly on the overall geometry of the DNA within the transcription initiation complex.

In order to evaluate the possibility of such interactions, we have constructed variants of the malt promoter that move the CRP binding site by small increments along the DNA upstream of the

transcription start site. Using a very sensitive luciferase assay (Meighen, 1991, 1993), we have measured the *in vivo* activities of these variant promoters.

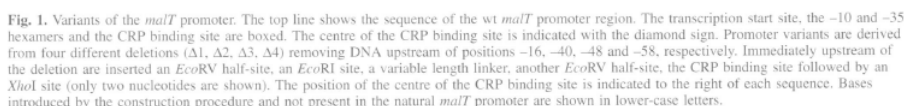
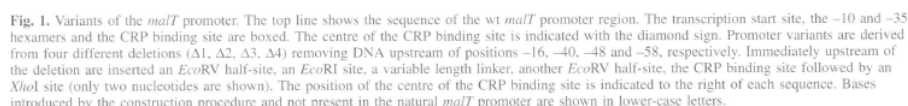
To assess directly the contribution of DNA geometry on transcriptional activation, we have replaced the CRP site by a binding site for the integration host factor (IHF). IHF is a heterodimeric protein that belongs to a class of prokaryotic histone-like proteins and is implicated in many cellular processes involving DNA (Friedman, 1988). IHF has the ability to bend DNA to a similar degree as does CRP (Kosturko et al., 1988) and thus allowed us to analyse the respective contributions of protein-protein contacts and DNA curvature to the mechanism of transcriptional activation by CRP. Finally, using malt promoter variants where the DNA sequence upstream of the CRP binding site was deleted, we assessed the importance of contacts between this DNA region and the back-side of RNA polymerase *in vitro*. Our results suggest that DNA topology and direct protein-protein contacts contribute to a comparable degree to the process of transcriptional activation by CRP at the malt promoter.

Results

Construction of variants of the malt promoter. The malt promoter directs transcription of the central activator of the maltose operons, the malt protein. The activity of the malt promoter is regulated by the activator protein CRP which binds to a specific recognition site centred at -70.5 bp from the start site of transcription. All elements necessary and sufficient for the regulation of this promoter are located within the region encompassing the promoter and the CRP binding site (Raibaud et al., 1991). Variants of the malt promoter were constructed by deletion of DNA upstream of the promoter (deletions A1 to A4), insertion of short linkers and reconstitution of the natural CRP binding site (Figure 1). The sequence of the CRP binding site of malt starting 11 bp upstream of the first pentamer and ending with the downstream pentamer is cloned just upstream of the linkers. Promoters are identified by the position of the centre of the activator binding site. The same deletions (A1 to A3 of Figure 1) were used to place a binding site for the heterodimeric IHF of *E. coli* upstream of the malt promoter. We chose the core of the well-characterized HI-site within the att region of bacteriophage lambda as a representative IHF binding sequence (Yang and Nash, 1989). The IHF binding site thus obtained possesses an affinity for IHF very close to the affinity of HI in its natural sequence context (data not shown). The site is occupied *in vivo* in an IHF⁺ strain (Engelhorn et al., 1995). The centre of the pseudo 2-fold symmetric recognition site of the heterodimeric IHF has been determined to high precision (Yang and Nash, 1989; Nunes-Diuby et al., 1995) and simplifies the comparison with DNA bends induced by CRP. IHF and CRP provoke qualitatively equivalent geometrical distortions of the DNA. The diamond signs in Figures 1 and 2 indicate the centres of the DNA bends directed towards CRP or IHF.

Effect of sequences upstream of -35 on activator-independent transcription In order to assess the effect of the distance between the activator binding site and the transcription start site on transcriptional activation, we had to establish the reference level of transcription in the absence of activator. This basal transcriptional activity (Figure 3) was measured in variants of E.coli strain pop2491 carrying deletions of either the *crp* or *hip* genes (coding for CRP and the α -subunit of IHF, respectively). Deletion of the α -subunit of IHF (*hip*) completely abolishes IHF activity in vivo (Granston and Nash, 1993). The

Do sequences outside the core promoter region influence the basal transcription activity of the *malT* promoter? If the DNA sequence upstream of the -35 hexamer did not influence the basal promoter activity, all deletions affecting this region (A2, A3 and A4) should be equally active (deletions with smaller identification numbers correspond to less wt DNA being retained). We observe, however, a 5-fold higher basal transcription activity for promoters derived from the A2 deletion, which replaces base pairs upstream of -40. Deletions A3 or A4, which



preserve the same sequence between -40 and -49 but they differ widely wt sequences at least up to base pair -48, show the same in the sequence further upstream (due to different length basal promoter activity as the wt *malT* promoter (Figure spacers, insertion of a CRP or IHF binding site). Despite 3). Promoters derived from the A2 deletion possess the these differences, they all show the same basal promoter -40 and -49 decreases the basal activity of the wt *malT* promoter. In order to obtain the -59.5 promoter variant, we deleted the central AATT sequence of the *EcoRI* restriction site from the -63.5 promoter variant derived from the A2 deletion, thereby modifying the sequence between -45 and -49 (Figure 1). This replacement results in a further 4-fold increase in basal promoter activity (Figure 3A). All sequence replacements within the -40 to -49 region that decrease the AT-content lead to an increase of basal promoter activity. The α -subunit of RNA polymerase contacts the DNA in this region (P.Eichenberger, S.Dethiollaz, N.Fujita, A.Ishihama and J.Geiselmann, in preparation). Since the α -subunit is thought to bind preferentially to AT-rich sequences (Galas et al., 1985; Ross et al., 1993), it is likely that the effect on basal promoter activity is mediated by this polypeptide (see Discussion).

Activation of transcription by CRP located at variable distances upstream of the promoter.

Having established basal transcription levels, we determined the magnitude of transcriptional activation by CRP *in vivo* by comparing the promoter strength of the constructions when transformed into the wt *E.coli* strain pop2491 or the isogenic strain pop2492 carrying a deletion of the chromosomal CRP gene. The results are shown in Figure 4. Activation is very strong at certain positions and the figure represents activation factors on a logarithmic scale (see Materials and methods) for easier comparison. Activation by CRP is strongest at the positions corresponding to the natural promoters *lac* and *malT*. Comparison of the -81.5 activation peak with natural promoters is difficult because such promoters are generally activated by CRP only in conjunction with a secondary activator (Kolb et al., 1993a).

The first peak of activation of this class I promoter series is observed when the CRP binding site is located at -60.5. The CRP binding site at the natural *lac* promoter is centred around -61.5. However, the *lac* promoter has spacing of 18 bp between the -10 and -35 hexamers whereas the *malT* promoter has a spacing of 17 bp. The activation maximum for our semi-artificial promoters occurs therefore at exactly the same position of the CRP binding site relative to the -35 hexamer as the one found in the *lac* promoter.

We observe another strong peak of activation when CRP is bound at approximately the position of the wt *malT* promoter (-70.5). Activation from this position is even stronger than from the *lac* position. This may in part be due to the lower basal activity of this promoter (see above). The precise location of the activation peak is at -71.5, i.e. 1 bp upstream of the

position found in the wt *malT* promoter, but exactly one helical turn upstream of the preceding activation peak. The activation factor obtained with the wt *malT* promoter is -10-fold higher than the

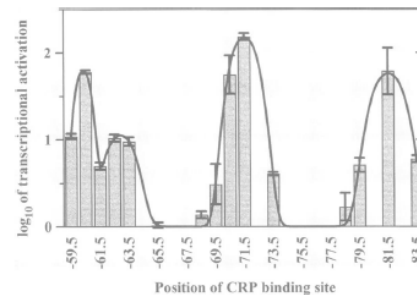


Fig. 4. Transcriptional activation of *malT* promoter variants. Activation factors are plotted on a logarithmic scale for easier visualization of largely different activation factors. (The logarithm of the activation factor may also be interpreted as a measure of the interaction free energy between CRP and RNA polymerase, according to the relationship $\Delta G = RT \ln K$) The line represents the best fit of the data using a twisting potential varying quadratically with the deviation from the optimal location of the activator binding site. The peaks are at -60.5, -63, -71.5 and -81.5 with respect to the start site of transcription. Error bars indicate the standard deviation of independent measurements.

activation factor of the -70.5 *malT* construct (data not shown). These two promoters differ in the sequence from -49 to -62 and in the sequence upstream of the CRP binding site, suggesting the involvement of either one or both regions in the mechanism of transcriptional activation. This result represents the first indication concerning the importance of the upstream sequence in the process of activation. The next activation peak is observed when the CRP binding site is moved one helical turn further upstream to position -81.5, with an activation factor almost as strong as for the more promoter-proximal positions.

Surprisingly, we observe yet another, smaller, peak of activation at -63, a position where CRP would be about 900 out of phase with respect to the canonical sites. The steep decrease of transcriptional activation when the CRP binding site is moved away from the -60.5 position (Figure 4), as well as the calculation of the free energy cost of twisting a short piece of DNA (see Discussion), preclude the possibility that the proteins within the two structures make identical interactions by deforming (twisting) the intervening DNA to different extents. Thus, the nucleoprotein complex responsible for activation from -63 appears to be distinct from the activation complex of the -60.5 promoter (possibly the opposite face of CRP making contact with the RNA polymerase, see Discussion). *In vivo* transcription start site mapping (data not shown) shows that the same start sites are used by all promoter variants and that no new promoter was created accidentally during the construction.

The difference between the activation factor at the optimal position and adjacent position differing by insertion or deletion of 1 bp are 7-fold, 3-fold and 2-fold (as judged from the curves of Figure 4) when the

CRP binding site is centred around positions -60.5, -71.5 and -81.5, respectively. The shape of these curves indicates that the interaction between CRP and RNA polymerase is rather rigid, at least when the CRP binding site is located at position -60.5 (see Discussion).

When CRP is bound around -40.5, the activation factor is -10 (Table I) and the basal promoter activity is very low. This is not surprising since promoters derived from the A1 deletion have lost the -35 region and the -10 region of the *malT* promoter is far from the consensus (2/6), contrary to usual class II promoters. CRP is nevertheless capable of stabilizing the RNA polymerase to some extent on these truncated promoters.

Table I. Activation factors for class II promoter variants

Position	CRP (activation)	Position	IHF (inhibition)
-36.5	8.7 ± 3	-37.5	6.0 ± 1
-37.5	9.3 ± 3	-38.5	2.5 ± 1
-39.5	9.2 ± 4	-40.5	3.8 ± 2
-41.5	6.2 ± 1	-42.5	4.8 ± 1

Activation factors are given as the ratio of luminosities in wt strains compared with strains deleted for the activator (CRP or IHF). IHF inhibits transcription and the values represent inhibition factors, i.e. the ratio luminosity in IHF⁻ strain/luminosity in IHF⁺ strain. Values are mean ± standard deviation.

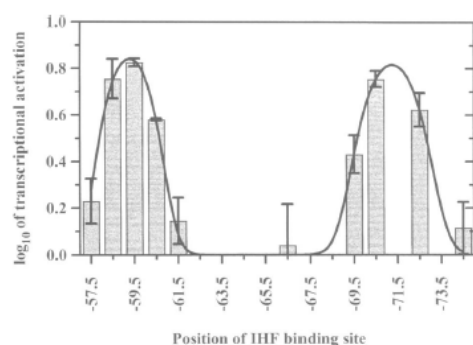


Fig. 5. Transcriptional activation of IHF variants of the *malT* promoter. The peaks are at -59.5, and at -71 with respect to the start site of transcription. Promoter variants with IHF binding sites at positions -62.5 and -64.5 show a slight (~30%) inhibition in the presence of IHF and are therefore not visible on this graph. The centre of the IHF binding site is as indicated in Figure 1. Error bars represent one standard deviation.

Activation of transcription by IHF

In order to assess the contribution of DNA bending to transcriptional activation by CRP, we replaced CRP by the heterologous DNA-bending protein IHF. The in vivo promoter activities (Figure 5) show peaks of transcriptional activation when IHF is centred at positions -60 and -71. These locations correspond to peaks of activation by CRP. IHF derivatives of the A4 deletion (sites centred around -80) have not been constructed. These results suggest that curvature alone is sufficient to activate the *malT* promoter (see Discussion for alternative interpretations). The magnitude of IHF activation is smaller than the activation obtained with CRP bound at the same positions. DNA curvature makes an important contribution to the mechanism of transcriptional activation but, by itself, is not sufficient for full activation.

When IHF is bound around position -40.5 we observe an inhibition of transcription (Table I). In this configuration, IHF may prevent the binding of RNA polymerase to the -35 region. Whereas CRP stabilizes the RNA polymerase-promoter complex in this situation (Table I), IHF apparently cannot establish favourable protein-protein contacts with the polymerase and acts rather as a repressor by steric exclusion of RNA polymerase (see Discussion).

Participation of ARI in transcriptional activation

The in vivo activities of the *malT* promoter variants (Figure 4) show four peaks of activation, at positions -60.5, -63, -71.5 and -81.5. Does the same activating region on the surface of CRP participate in the activation process? In the case of the *lac* promoter, a surface-exposed loop around amino acids 158-161 of CRP constitutes an activating region (ARI) essential for transcriptional activation (Bell et al., 1990; Eschenlauer and Reznikoff, 1991; Zhou et al., 1993). Mutants of CRP mapping in this region, called positive control or pc-mutants, are (nearly) completely unable to activate transcription, but bind DNA with the same affinity and bend DNA to the same extent as does wt CRP. We assayed promoter activity of our constructs in *E. coli* strain pop2492 carrying a chromosomal deletion of the gene coding for CRP and containing a plasmid expressing either

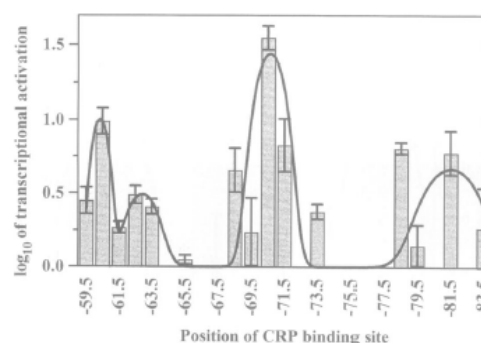


Fig. 6. Transcriptional activation of the *malT* promoter variants by CRP mutant H159L. The nomenclature and the peaks positions are as in Figure 4.

wt CRP or a pc mutant of CRP, H159L (replacing His 159 with leucine). Smaller activation values and the use of a two-plasmid system cause larger experimental errors but the data clearly show that the activation profile (Figure 6) resembles the one obtained with wt CRP (Figure 4). All activation factors are between 7- and 12-fold lower for H159L than for wt CRP. ARI thus appears to participate in transcriptional activation at all class I promoters. Even in the absence of a favourable contact between His159 and RNA polymerase, transcriptional activation persists. The activation factors, however, approach the levels obtained with IHF. CRP deficient in ARI and IHF (which does not possess an ARI) may activate transcription in a similar manner (see Discussion).

Role of the DNA sequence upstream of the CRP binding site

The most direct way to demonstrate that upstream sequences participate in contacts with RNA polymerase is to remove the sequences completely. Such an experiment can only be performed *in vitro*. Figure 7 shows *in vitro* transcription reactions using templates that retain different lengths of upstream DNA. Both templates are transcribed to the same extent in the absence of CRP (lanes 2 and 3). In the presence of CRP (lanes 4 and 5) the longer fragment yields several-fold more transcripts than the fragment truncated at -93, just upstream of the CRP binding site. Footprinting experiments confirm that both proteins are bound to the DNA fragments and that RNA polymerase does not interfere by binding to the end of the fragments under our experimental conditions (data not shown). Two run-off transcripts, that differ in length by about six nucleotides, are observed in these gels. *In vivo* transcription start site mapping shows the same doublet. Either the *malT* promoter possesses a genuine second transcription start site at +6 or RNA polymerase processes the long transcript by removing six nucleotides from its 5'-end. Even though we do not know the reason for the double band, the interpretation of our results is not affected by this phenomenon because the two transcripts are strictly co-regulated.

Discussion

Transcriptional activation uses a panoply of mechanisms. Transcriptional activation by CRP constitutes an ideal model system because the geometry of the activation complexes is variable and possibly different modes of activation can be studied in the same system. As a starting point for analysing the mechanisms of transcriptional activation, we have to define the factors that set the activator-independent promoter strength. An activator should improve the limiting step of the process.

The strength of a promoter in the absence of an activator, the basal promoter activity, is determined mainly by the sequence of the recognition hexamers around -10 and

-35 (Hawley and McClure, 1983). In addition to these well-documented sequences, our variants of the *malT* promoter clearly implicate the region between -40 and -49 for setting the basal promoter activity. The A2 variants of the *malT* promoter (Figure 3) are stronger than variants that leave the -40 to -49 region untouched. Changes upstream of -49 do not affect the basal transcriptional activity of the promoter. A further perturbation of this sequence, by removal of the central AATT tetramer within the EcoRI site (the -59.5 promoter), reduces the AT content of this region and basal transcription further increases 4-fold.

Evidence has accumulated that the σ -subunit of RNA polymerase binds to this segment of DNA at certain promoters (Galas et al., 1985; Ross et al., 1993; Blatter et al., 1994; Rao et al., 1994). We have confirmed biochemically that analogous interactions

are established at the *malT* promoter (P.Eichenberger, S.Dethiollaz, N.Fujita, A.Ishihama and J.Geiselmann, in preparation). The sequence changes in the -40 to -50 region could either favour or disfavour the interaction with the σ -subunit. It appears unlikely that we fortuitously introduced a strong recognition site for the σ -subunit. Furthermore, it has been shown that the σ -subunit preferentially binds to ATrich sequences (Galas et al., 1985; Ross et al., 1993; Blatter et al., 1994; Rao et al., 1994). Taken together, these observations strongly suggest that the sequence of the wt *malT* promoter in the region between -40 and -50 favours binding of the σ -subunit(s). Paradoxically, replacement of this sequence leads to a stronger basal transcription despite the (probably) weaker binding of RNA polymerase. The paradox can be resolved, knowing that the *malT* promoter is limited at promoter escape *in vitro* (Menendez et al., 1987). Abolishing (or weakening) interactions that stabilize RNA polymerase on the promoter, and that have to be relinquished during escape, is therefore expected to increase promoter activity (Ellinger et al., 1994). In the course of constructing the A2 promoter variants, we probably destroyed favourable binding sequences for the σ -subunit and thus accelerated promoter escape. While basal promoter activity is dictated by the sequences of the promoter and of the region just upstream of -35, transcriptional activation has to result from interactions (direct or indirect) between the activator and RNA polymerase. Direct protein-protein interactions between CRP and the σ -subunit of RNA polymerase have been convincingly shown to exist at the *lac* promoter (Igarashi and Ishihama, 1991; Kolb et al., 1993b; Chen et al., 1994). Such protein-protein contacts depend on a precise spatial alignment of the components; generally, we find strong peaks of activation whenever the CRP binding site is in the correct helical phase with respect to RNA polymerase (Figure 4). There is one exception to this pattern. A smaller, secondary activation peak is observed when the centre of the CRP binding site is at -63, 2.5 bp upstream of the activation peak corresponding to a lactate promoter. CRP is now 90° out of phase compared with all other activation peaks. This result is interpreted in the most straightforward manner by assuming that the left side of CRP (left and right being defined by looking in the direction of transcription) touches the right side of RNA polymerase when CRP is centred at -60.5 and that the right side of CRP contacts the left side of RNA polymerase when moved 2-3 bp further upstream. At intermediate positions, transcriptional activation diminishes due to a steric clash between CRP and RNA polymerase. Experiments are underway to verify this interpretation. No doublets are observed for peaks at more promoter-distal positions of the CRP binding site because the binding sites for CRP and RNA polymerase no longer overlap. If transcriptional activation is based on protein-protein contacts, how can these contacts be maintained when one of the protein partners is moved 34 or even 70 Å further upstream? It has been suggested that the C-terminal domain of the σ -

subunit of RNA polymerase is attached to the rest of the enzyme via a flexible protein linker of at least 13 amino acids (Blatter et al., 1994). A very long stretch of amino acids in a completely extended conformation would be necessary to explain almost equal activation factors when CRP is placed at -60 or -80 solely on the basis of identical protein-protein interactions between CRP and RNA polymerase. A pronounced bending of the DNA could approach the two components, but would be too expensive energetically. In addition, we do not observe any footprinting signals that would indicate such a strong bending (P.Eichenberger, S.Dethiollaz, N.Fujita, A.Ishihama and J.Geiselmann, in preparation). The assumption of such great flexibility leads to the following prediction: small misalignments of the proteins should be tolerated without a major decrease in transcriptional activation. The activation profiles obtained from promoters, where the CRP binding site is shifted away from the optimal position by only 1 bp, contradict this prediction. If we assume that transcriptional activation depends on the establishment of a precise protein-protein contact between rigid proteins, we can calculate the cost in free energy of realigning CRP with RNA polymerase by twisting the DNA segment between the two anchor points: the -35 hexamer recognized by RNA polymerase, and the CRP binding site. The length of this twistable DNA is therefore only -10 bp for the lac-type promoters. The energetic cost of twisting an average 10 bp segment by 340 is -1 kcal/mol, assuming a torsional rigidity of 1.1XO-19 erg-cm (Law et al., 1993). The decrease in transcriptional activation by CRP upon shifting the CRP binding site 1 bp away from the optimal position is indeed of this magnitude. Any flexibility in either one of the proteins would lead to a less dramatic decrease in transcriptional activation. We conclude that the protein-protein contacts between CRP and RNA polymerase are rather rigid. As expected, the decrease in transcriptional activation upon misaligning the CRP binding site is less dramatic for the activation peaks around -70 and -80. The longer stretch of twistable DNA decreases the cost in free energy of realigning the two proteins. Additional interactions may participate in transcriptional activation at these promoter variants.

Candidates for such interactions are contacts between the DNA upstream of the CRP binding site with the backside of RNA polymerase, since CRP is known to bend the DNA to a large degree (Liu-Johnson et al., 1986). In order directly to investigate this possibility and to minimize possible contributions of direct protein-protein contacts, we reconstructed promoters that replace the CRP binding sites with recognition sites for IHF. This protein should bend the DNA to a similar degree as does CRP (Kosturko et al., 1988) without making the exact same protein-protein contacts with RNA polymerase.

We observe peaks of activation with a helical periodicity, as was the case for the CRP-activated promoters. IHF has been shown to activate

transcription at the *ilvGMEDA* promoter through a long-range distortion of the DNA helix. Transcriptional activation by this mechanism is completely independent of the phase of the helix (Parekh and Hatfield, 1996). We therefore exclude the DNA distortion mechanism of activation for our promoter constructs. We conclude, on the contrary, that transcriptional activation by IHF is due to bending of the DNA. It is well documented that sequence-directed bends alone can lead to transcriptional activation at the *gal* and *lac* promoters (Bracco et al., 1989; Gartenberg and Crothers, 1991). Our experiments make use of a protein-induced bend and thus eliminate the possibility that activation is a result of unknown cellular factors binding to the bent

DNA sequences. We show furthermore that direct contacts between RNA polymerase and regions of IHF corresponding to ARI, ARII or ARIII of CRP are not responsible for activation: the class II promoter variants of *malT* are repressed by IHF binding, whereas the homologous CRP promoters are activated by CRP.

Two other promoters, activated by IHF, are comparable with our constructs: the PL promoter of phage lambda (Giladi et al., 1990, 1992a,b) and the Pe promoter of bacteriophage Mu (van Rijn et al., 1988). IHF binds upstream of these promoters on the same side of the DNA helix, but in the opposite orientation, as on our *malT*-derived constructs (Goosen and van de Putte, 1995). Since different subunits of the heterodimeric IHF point towards the RNA polymerase it appears unlikely that identical, direct interactions could be established with RNA polymerase in all cases (the sequence identity between the two subunits of IHF is only 25%). The common characteristic of all these promoters is the helical phase of the bending centre of IHF. The most parsimonious mechanistic explanation, applicable to all examples cited, would attribute the same effect (transcriptional activation) to a common cause (geometry of the DNA). Binding of the activator induces a bend in the DNA that facilitates contacts of the upstream DNA with the back-side of RNA polymerase. Any of the subunits of RNA polymerase, including the σ -subunits, could be involved in this additional interaction. If IHF can activate transcription by bending the DNA it is quite likely that DNA-bending by CRP has the same effect, namely transcriptional activation.

Even though direct protein-protein contacts certainly participate in transcriptional activation by CRP, DNA bending may provide a major contribution at certain promoters. In vitro transcription experiments require the upstream DNA region for transcriptional activation by CRP at the *malT* promoter (Figure 7). We have detected an upstream signal at -94 by UV-laser footprinting that is visible only when CRP is present (P.Eichenberger, S.Dethiollaz, N.Fujita, A.Ishihama and J.Geiselmann, in preparation). Such

strong indications for far-upstream contact are not unique to the *malT* promoter. A similar signal has previously been observed at the *lac* promoter (Buckle et al., 1992). Further evidence that transcriptional activation does not rely solely on protein-protein interactions comes from experiments using pc-mutants of CRP. The H159L mutation in ARI drastically affects transcriptional activation by CRP (Bell et al., 1990; Zhou et al., 1993). When we measure the activation factors of this CRP mutant with our *malT* promoter variants, we obtain the same activation profile as for the wt CRP, but all activation factors are decreased -10-fold. This result confirms that His159 contributes to activation but also shows that it is not the sole source of interactions. Very specific contacts between His 159 and RNA polymerase would probably be destroyed by most amino acid substitutions at this position. However, at least six very different amino acids can replace His 159 without dramatically decreasing transcriptional activation (Williams et al., 1991; Niu et al., 1994). The phenotype of ARI-mutants could be explained in an alternative way by less-specific, 'long-range' effects of these mutations on the bendability (the parameter important for facilitating contacts of the upstream DNA with RNA polymerase, and not to be confused with DNA bending) of the DNA bound to CRP. We have initiated experiments that will test this prediction.

All experimental observations can be reconciled by the following model. Two major kinds of interactions contribute to transcriptional activation: direct protein protein contacts and binding of upstream DNA to the back-side of RNA polymerase. The latter process is facilitated by activator-induced bending of DNA. Such additional contacts could stabilize specific intermediates of transcriptional initiation. At class II promoters, transcriptional activation is due almost exclusively to favourable protein-protein contacts between different RNA polymerase subunits and CRP. IHF cannot make these contacts and cannot replace CRP at class II promoters. At class I promoters (CRP centred around -61 or further upstream), ARI makes favourable protein-protein contacts that increase the bimolecular binding constant (KB) of the promoter. At the same time, upstream DNA is brought into the vicinity of the back-side of RNA polymerase. At the *lac* promoter, where escape is not limiting, the major effect of CRP is on closed complex formation. At the *malT* promoter, closed complex formation is efficient and CRP primarily stimulates promoter escape through backside contacts with RNA polymerase and/or by destabilizing RNA polymerase-promoter interactions that impede escape (e.g. interactions between the α -subunit and DNA just upstream of -35). The contribution of DNA bending to transcriptional activation is more important when CRP is bound further upstream and protein-protein contacts are more difficult to establish. A protein like IHF, that cannot make favourable protein-protein interactions, would activate transcription by modifying DNA geometry in a way that favours interactions between

upstream DNA and RNA polymerase (possibly the α -subunits). The activation factors are smaller than for CRP either because additional interactions exist between the genuine activator CRP and RNA polymerase, or because the geometrical distortions introduced by IHF do not exactly mimic the DNA conformation induced by CRP.

We cannot exclude the possibility that protein binding (CRP or IHF) in the vicinity of the promoter modifies the chromatin structure of the promoter (e.g. by excluding the association of histone-like proteins such as H-NS or HU), and that the modified chromatin structure in itself leads to activation of transcription. Experiments are underway in our laboratory to test this possibility along with in vitro and in vivo footprinting experiments that probe the physical interactions predicted by the model (P.Eichenberger, S.Dethiollaz, N.Fujita, A.Ishihama and J.Geiselman, in preparation).

Materials and methods

General methods

Standard methods of molecular biology (Sambrook et al., 1989) were used unless otherwise specified. Enzymes were purchased from New England Biolabs, Biofinex, Promega, United States Biochemicals, or Pharmacia. DNA oligonucleotides used as primers and for the construction of synthetic promoter sequences were synthesized on a Perkin- Elmer ABI 392 synthesizer. Plasmids were sequenced using the dideoxymediated sequencing reaction adapted as described in the sequenase manufacturer's instruction manual (United States Biochemicals).

Bacterial strains and plasmids The strain of *E. coli* used for the cloning steps of this work is the XL1-blue [*recA1*, *lac-⁻*, *endA1*, *gyr A96*, *thi*, *hsdR17*, *supE44*, *relA1*, (*F'* *proAB*, *lacIq*, *lacZAM15*, *Tn10*)] developed by Stratagene. All strains used for the bioluminescence assay are derived from the pop2491 strain (*F*- *aroB* *relA* *araD139* *A/acU169* *nal/Tp⁺* *Ama/IA102* *Amiia/A549*). The isogenic pop2492 strain (*F*- *aroB* *relA* *araD139* *A/acU169* *nal/* *1Tp*- *AmnalA102* *AinalA549* *AcrpT8*) carries a deletion of the gene coding for CRP (Raibaud et al., 1991). P1 transduction was used to interrupt the *hip* gene (coding for the 3-subunit of IHF) in the strain pop2491 yielding pop3491. The parental strain for P1 transduction was MC252 (Gamas et al., 1986), carrying an insertion of the chloramphenicol acetyltransferase gene into the deleted *hip* gene. For the measurements of transcriptional activation by the positive control mutant of CRP, we transformed pop2492 with plasmid pDW300 (expressing the wt *cep* gene) (Bell et al., 1990). Or with plasmid pDW301 (expressing the HL159 pc-mutant of CRP) (West et al., 1993).

Plasmid construction

The sequence of the *ma/Tfl* promoter from positions -122 to +90 with respect to the transcription start site (Cole and Raibaud, 1986) was cloned into the

polylinker of the pBluescript SKII+ vector (Stratagene) between the EcoRI-HindIII restriction sites and introducing SmaI and PstI restriction sites. The *inaIT* promoter is an up-mutant of the wt *inaIT* promoter containing a G to T transversion at position -12 within the -10 hexamer (Chapon, 1982). The resulting sequence at the junctions is (vector sequence in lower-case letters):

gaattcccgGGATCAGCG-...CTAATCGTCCGacctgcagccaagctt. In order to construct the *malIT* promoter variants that differ in the location of the CRP binding site, we first produced a series of *ExoIII*Mung Bean nuclease deletions starting at the upstream EcoRI site. After digestion with *HindIII*, the mixture of different-length fragments was cloned into the pBluescript SKII+ vector between the EcoRV and *HindIII* restriction sites. Five deletions that conserve the original sequence up to -23, -40, -48, -58 and -89 (A1, A2, A3, A4 and A5, respectively) were selected to design promoter variants. Four linkers of different lengths, L0, L1 (linker 0 + 1 nucleotide), L3 (linker 0 + 3 nucleotides) and L5 (linker 0 + 5 nucleotides), containing an EcoRI site were cloned into another pBluescript SKII+ vector between the EcoRV and PstI sites. A synthetic DNA fragment that reconstitutes the original CRP binding site was introduced upstream of the four linkers between the *XbaI* and EcoRV sites. Finally, the four *malIT* deletion fragments (A1-A4) were cloned downstream of these constructs between the EcoRI site of the linkers and the PstI site of the vector (Figure 1).

An additional promoter placing the centre of the CRP binding site at -59.5 was created by cutting the -63.5 promoter (promoters are identified by the position of the centre of the CRP or IHF binding site as indicated in Figures 1 and 2) with EcoRI, digesting the four bases overhanging and re-ligating the vector. The -62.5 and -70.5 promoters were constructed by removing the segment including the CRP site and the linker (between *XbaI* and EcoRI) from the appropriate plasmids and replacing the entire fragment with a synthetic DNA fragment containing the CRP site and a linker of length two (L2, linker 0 + 2 nucleotides) (Figure 1).

In order to obtain the corresponding *mtalIT* constructs, a T-to-G mutation was introduced at position -12 of each *malIT* construct using the Chameleon double-stranded site-directed mutagenesis kit of Stratagene.

A full-length wt promoter fragment was obtained as a SmaI-EcoRI fragment from plasmid pOM35 (Raibaud et al., 1991) and cloned into the polylinker of the pBluescript KS+ vector (Stratagene). This plasmid now contains the wt *inaIT* sequence from position -122 to position +90. The sequence at the junctions is (vector sequence in lower-case letters):

gtcgacGGATCAGCG...CTAATCGTCCGaattc. The wt *malIT*, as well as the *mtalIT* promoter variants, were then transferred to the pSB377 vector.

a derivative of pHG177 (described in Stewart et al., 1986), containing a 5.82 kbp promoterless luxCDABE cassette cloned as an EcoRI-EcoRV fragment between the EcoRI and *HpaI* sites of pHG

177. The lux cassette contains the XLLABCDEA sequence (accession number M90093 of the EMBL sequence data bank) between 459 and 6257 with a short linker (shown in bold) introduced upstream of the *lucC* gene: 5'-GAATTCAGGCTTGGAGGATACGT ATG ACT AAA AAA-3'. The EcoRI site was expanded into a multiple cloning site comprising recognition sequences for *XbaI*-*BamHI*-*SmaI*-*SmaI*. The resulting vector (also referred to as 'luciferase vector' in the text) contains the complete promoterless luciferase operon of the terrestrial bacterium *X. luminiscens* consisting of the two subunits of the luciferase as well as the genes for aldehyde production (for a review, see Meiahen, 1991, 1993). All promoter variants (see above) were cut with *XbaI* and *SspI* and transferred to this vector between *XbaI* and *SmaI*. The wt *malIT* promoter was cut with *BamHI* and *SspI* and transferred to the luciferase vector between the *BamHI* and *SmaI* restriction sites. The IHF variants of the *mtalIT* promoter were constructed using a synthetic DNA fragment containing the IHF binding site HI from the bacteriophage lambda att region (Yang and Nash, 1989). The IHF oligonucleotide was introduced upstream of the four linkers, already cloned into the pBluescript SK vector. Between the *XbaI* and EcoRV sites. The *XbaI*-EcoRI fragments of these constructs were introduced into the different luciferase vector constructs in place of the corresponding *XbaI*-EcoRI fragment containing the CRP binding site. Four additional linkers were used to construct promoter variants with the IHF binding site located between -57.5 and -60.5 as shown in Figure 2. The integrity of all promoters was verified by dideoxy-mediated sequencing.

Luciferase assay to measure in vivo transcriptional activity and data treatment

After transformation at least three colonies were picked and grown overnight with aeration at 37°C in L-broth. They were diluted 500-fold in M9 liquid medium (Sambrook et al., 1989) supplemented with 0.4% glucose (w/v) or 0.4% glycerol (V/v) and grown again overnight at 37°C. The cultures were diluted again 500-fold in 5 ml of the same medium and grown for 6 h at 37°C (OD600 = 0.6). Bioluminescence was measured using a Biorbit 1250 Luminometer. The OD600 of each sample culture was determined immediately after the measurement of bioluminescence using a Beckman DU-5 spectrophotometer. The relative light intensity (*lumrel*) is calculated as the ratio of the luminometer reading (in arbitrary light units) and the OD600 of the culture. The lowest promoter activities measured are more than 20-times above the background level of luminosity determined for the promoterless luciferase vector.

Activation factors (*Act*) are calculated as the ratio of *lumrel* in a wt strain and in the isogenic strain deleted for the activator (and corrected for activator-independent transcription): $Act = (lum_{wt} / lum_{\Delta act})$ in activator minus strain). Activation

factors for CRP span a large range and the data are therefore plotted on a logarithmic scale by taking the log10 of the activation factor. The variation of activation factors with the position of the activator binding site is modelled by assuming that activation decreases in an harmonic fashion when the activator is moved away from the optimal position. The curves in Figures 4-6 thus assume that an optimal activating structure can be restored by twisting the DNA when the activator is misaligned relative to RNA polymerase. The energy of DNA twisting is considered to be the only factor impeding optimal activation when the activator and RNA polymerase are out of helical phase.

In vitro run-off transcriptions

DNA fragments were prepared from the pBluescript KS' vector carrying the truncated A5 version of the mltI promoter (see above) lacking the malT sequence upstream of position -89. The fragments prepared from this plasmid, their total length and the amount of upstream DNA were: (i) a XbaI/HindIII fragment (223 bp long, 123 bp upstream); and (ii) an EcoRI/HindIII fragment (193 bp long, 93 bp upstream). The upstream end is defined as the double-strand extremity of the DNA. In order to obtain standard transcripts from a CRP-independent promoter, a 189 bp DNA fragment was prepared by PCR from the pBluescript KS+ vector carrying the truncated A2 version of the ma/Tpl promoter (see above) lacking the na/Tpl sequence upstream of position -40. The following oligonucleotides were used as primers: 5'-AGTTAATCACTTCACTGTGG-3' complementary to the ina/Tpl sequence between nucleotides +40 and +60 downstream of the ica/Tpl transcription start site and 5'-AATACGACTCACTATAGGGC-3' complementary to the pBluescript KS- sequence between nucleotides -129 to -109 upstream from the ma/Tpl transcription start site. After separation on a polyacrylamide gel the restriction or PCR fragments were recovered by cutting the band and extracting the DNA by electroelution or using the QIAquick gel extraction kit from Qiagen. Respectively, after precipitation, the DNA was resuspended in H₂O and the concentration determined spectrophotometrically.

The *in vitro* run-off transcriptions are performed in a total volume of 20 μ l in TxMg buffer (20 mM HEPES-KOH, pH 7.5, 150 mM K⁺-glutamate, 10 mM Mg₂+glutamate, 1 mM EDTA). The reactions contain 10 nM malT promoter DNA and the reference promoter DNA, 200 PM each of ATP, CTP and GTP, 10 pM UTP and 70 nM [α -³²P]UTP (3000 Ci/mmol, Amersham). When present, CRP is at a concentration of 100 nM and 200 pM cAMP is included in the reaction. The samples are preincubated for 10 min at 37°C. The transcription reaction is started by the addition of RNA polymerase (Sigma) to a final concentration of 40 nM. The reaction is terminated after 5 min of incubation by the addition of 60 μ l of 96% ethanol and precipitation in a dry-ice/ethanol bath. After centrifugation, the pellet is resuspended in 12 μ l of 96% formamide, 20 mM

EDTA, heat-denatured for 2 min at 80°C and loaded on a 10% denaturing polyacrylamide gel (Sambrook et al., 1989).

Acknowledgements

We gratefully acknowledge Steve Busby for generously providing the CRP expression plasmids, Michael Winson for providing the luciferase reporter plasmid, Nicolas Roggli for expert photographic work. We keep a vivid memory of Olivier Raibaud, who provided the CRP deletion strains, and whose genetic insights, helpful suggestions and constructive, pertinent criticism remain invaluable. We thank Michel Clerget for help in constructing IHF deletion strains. We thank Pascal Damay for help with laboratory procedures. Special thanks go to Dany Rifat for help with the plasmid constructions and for contributing her expertise of bacteriological techniques. This work was supported by grants 31-32582.91 and 31-42501.94 of the Swiss National Fund.

References

- Aiba,H., Fujimoto,S. and Ozaki,N. (1982) Nuicleic Acids Res., 10, 1345-1361.
- Bell,A., Gaston,K., Williams,R., Chapman,K., Kolb,A., Buc,H., Minchin,S., Williams,J. and Busby,S. (1990) Nucleic Acids Res., 18, 7243-7250.
- Berg,O.G. and von Hippel,P.H. (1988) J. Mol. Biol., 200, 709-723.
- Blatter,E.E., Ross,W., Tang,H., Gourse,R.L. and Ebright,R.H. (1994) Cell, 78, 889-896.
- Bracco,L., Kotlarz,D., Kolb,A., Diekmann,S. and Buc,H. (1989) EMBO J., 8, 4289-4296.
- Buckle,M., Buc,H. and Travers,A.A. (1992) EMBO J., 11, 2619-2625.
- Chapon,C. (1982) EMBO J., 1, 369-374.
- Chapon,C. and Kolb,A. (1983) J. Bacteriol., 156, 1135-1143.
- Chen,Y., Ebright,Y.W. and Ebright,R.H. (1994) Science, 265, 90-92.
- Cole,S.T. and Raibaud,O. (1986) Genie, 42, 201-208.
- Ebright,R.H. (1993) Mol. Microbiol., 8, 797-802.
- Ellinger,T., Behnke,D., Bujard,H. and Gralla,J.D. (1994) J. Mol. Biol., 239, 455-465.
- Engelhorn,M., Boccard,F., Murtin,C., Prentki,P. and Geiselmann,J. (1995) Nucleic Acids Res., 23, 2959-2965.
- Eschenlauer,A.C. and Reznikoff,W.S. (1991) J. Bacteriol., 173, 5024- 5029.
- Friedman,D.I. (1988) Cell, 55, 545-554.
- Galas,D.J., Eggert,M. and Waterman,M.S. (1985) J. Mol. Biol., 186, 117-128.
- Gamas,P., Burger,A.C., Churchward,G., Caro,L., Galas,D. and Chandler,M. (1986) Mol. Gen. Genet., 204, 85-89.
- Gartenberg,M.R. and Crothers,D.M. (1991) J. Mol. Biol., 219, 217-230.
- Gaston,K., Bell,A., Kolb,A., Buc,H. and Busby,S. (1990) Cell, 62, 733-743.
- Giladi,H., Gottesman,M. and Oppenheim,A.B. (1990) J. Mol. Biol., 213, 109-121.
- Giladi,H., Igarashi,K., Ishihama,A. and Oppenheim,A.B. (1992a) J. Mol. Biol., 227, 985-990.

- Giladi,H., Koby,S., Gottesman,M.E. and Oppenheim.A.B. (1992b) *J. Mol. Biol.*, 224, 937-948.
- Goosen,N. and van de Putte,P. (1995) *Mol. Microbiol.*, 16, 1-7.
- Granston,A.E. and Nash,H.A. (1993) *J. Mol. Biol.*, 234, 45-59.
- Hawley,D.K. and McClure,W.R. (1983) *Nucleic Acids Res.*, 11, 2237- 2255.
- Herbert,M., Kolb,A. and Buc,H. (1986) *Proc. Natl Acad. Sci. USA*, 83,2807-2811.
- Igarashi,K. and Ishihama,A. (1991) *Cell*, 65, 1015-1022.
- Ishihama,A. (1993) *J. Bacteriol.*, 175, 2483-2489.
- Kolb,A., Busby,S., Buc,H., Garges,S. and Adhya,S. (1993a) *Annu. Rev. Biochem.*, 62, 749-795.
- Kolb,A., Igarashi,K., Ishihama,A., Lavigne,M., Buckle,M. and Buc.H. (1993b) *Nucleic Acids Res.*, 21, 319-326.
- Kosturko,L.D., Daub,E. and Murialdo,H. (1988) *Nucleic Acids Res.*, 17, 317-334.
- Law,S.M., Bellomy,G.R., Schlax,P.J. and Record,T.M.,Jr (1993) *J. Mol. Biol.*, 230, 161-173.
- Liu-Johnson,Huei-N., Gartenberg,M.R. and Crothers,D.M. (1986) *Cell*, 47, 995-1005.
- Malan,T.P., Kolb,A., Buc,H. and McClure,W.R. (1984) *J. Mol. Biol.*, 180, 881-909.
- Meighen,E.A. (1991) *Microbiol. Rev.*, 55, 123-142.
- Meighen,E.A. (1993) *FASEB J.*, 7, 1016-1022.
- Meighen,E.A. and Szittner,R.B. (1992) *J. Bacteriol.*, 174, 5371-5381.
- Menendez,M., Kolb,A. and Buc,H. (1987) *EMBO J.*, 6, 4227-4234.
- Niu,W., Zhou,Y., Dong,Q., Ebright,Y.W. and Ebright,H.E. (1994) *J. Mol. Biol.*, 243, 595-602.
- Nunes-Duby,S.E., Smith-Mungo,L.I. and Landy,A. (1995) *J. Mol. Biol.*, 253, 228-242.
- Parekh,B.S. and Hatfield,W.G. (1996) *Proc. Natl Acad. Sci. USA*, 93, 1173-1177.
- Raibaud,O., Vidal-Ingigliardi,D. and Kolb,A. (1991) *Res. Microbiol.*, 142, 937-942.
- Rao,L., Ross,W., Appleman,J.A., Gaal,T., Leirno,S., Schlax,P.J., Record,M.T.,Jr and Gourse,R.L. (1994) *J. Mol. Biol.*, 235, 1421-1435.
- Ross,W., Gosink,K.K., Salomon,J., Igarashi,K., Zou,C., Ishihama,A., Severinov,K. and Gourse,R.L. (1993) *Science*, 262, 1407-1413.
- Ryu,A., Garges,S. and Adhya,S. (1994) *Proc. Natl Acad. Sci. USA*, 91, 8582-8586.
- Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Schlax,P.J., Capp,M.W. and Record,M.T.,Jr (1995) *J. Mol. Biol.*, 245,331-350.
- Schmitz,A. (1981) *Nucleic Acids Res.*, 9, 277-292.
- Schultz,S.C., Shields,G.C. and Steitz,T.A. (1991) *Science*, 253, 1001- 1007.
- Stewart,G.S.A.B., Lubinsky-Mink,S., Jackson,C.G., Cassel,A. and Kuhn,J. (1986) *Plasmid*, 15, 172-181.
- Taniguchi,T., O'Neill,M. and de Crombrughe,B. (1979) *Proc. Natl Acad. Sci. USA*, 76, 5090-5094.
- Ushida,C. and Aiba,H. (1990) *Nucleic Acids Res.*, 18, 6325-6330.
- van Rijn,P.A., Goosen,N. and van de Putte,P. (1988) *Nucleic Acids Res.*, 16, 4595-4605.
- West,D., Williams,R., Rhodius,V., Bell,A., Sharma,N., Zou,C., Fujita,N., Ishihama,A. and Busby,S. (1993) *Mol. Microbiol.*, 10, 789-797.
- Williams,R.M., Bell,A.I., Sims,G. and Busby,S. (1991) *Nucleic Acids Res.*, 19, 6705-6712.
- Williams,R.M., Rhodius,V.A., Bell,A.I., Kolb,A. and Busby,S.J.W. (1996) *Nucleic Acids Res.*, 24, 1112-1118.
- Yang,C.-C. and Nash,H.A. (1989) *Cell*, 57, 869-880.
- Zhou,Y., Zhang,X. and Ebright,R.H. (1993) *Proc. Natl Acad. Sci. USA*, 90, 6081-6085.
- Zhou,Y., Merkel,T.J. and Ebright,R.H. (1994) *J. Mol. Biol.*, 243,603-610.
- Zou,C., Fujita,N., Igarashi,K. and Ishihama,A. (1992) *Mol. Microbiol.*, 6, 2599-2605. Received on April 24, 1996; revised on June 20, 1996