Gene Regulation Network (GNR) task at BioNLP Shared Task 2013 Task Proposal

Robert Bossy, Claire Ndellec, Philippe Bessires

October 23, 2012

Useful links

- BioNLP 2013 Shared Task
- Bacteria Biotopes at BioNLP 2013 Shared Task
- Specification of file formats

Task Summary

Task name Gene Regulation Network

Goal in IE Assess the performance of information extraction systems to extract the full genetic regulation network of *Bacillus subtilis* sporulation.

Motivation in biology The gene regulation network task aims at evaluating the quality of the extraction of gene interaction by IE systems with respect to the goals in biology. The automatic design of gene regulation network is one of the main challenges in Biology, because it is a crucial step forward in understanding the cellular regulation system. The sporulation network of the bacteria model *Bacillus subtilis* is very well-studied. There will thus be no need for an adjudication phase, because all predicted interactions that would not be in the reference regulatory network will be true negative.

The goal here is to retrieve all the genic interactions of the reference network –at least one occurrence per interaction– independently of where there are mentioned in the literature. Some of the interactions are not directly mentioned in the text and should be deduced from the extracted events. The inference rules will be given.

Compared to the BI task of BioNLPST'11, the evaluation will measure the capability of the IE systems to reconstruct the target regulatory network.

Description of the entities and events Entities are biomolecules and genetic objects. Events are low-level molecular and genetic events. Finally relations link genic entities where there is a mention of an interaction.

Concrete examples in BioNLP format Available in attached data.

Evaluation and criteria We advance an evaluation based on the genic regulation network formed by the set of interactions available on the corpus.

Task corpus The BI 2011 data set complemented with the gene interactions missing for building the complete sporulation network and with the regulation network.

Existing corpus The BI 2011 data set: the INRA-GI corpus is a set of sentences from abstracts of selected PubMed references concerning the genetic regulation of B. subtilis. Most sentences are the same as the LLL challenge corpus. The annotation was revised and enriched by a joint effort of the Bibliome team of MIG Laboratory at the Institut National de Recherche Agronomique (INRA) and the Laboratoire d'Informatique de Paris Nord at the Universit Paris 13.

Feasibility of task The results on BioNLP Shared Task 2011 Bacteria Genic Interaction were quite good. Other genic interaction and PPI tasks suggest that there are efficient systems.

Annotation process The annotated corpus has been build on top of the BioNLP Shared Task 2011 Bacteria Genic Interactions corpus. A senior annotator is revising the annotation of the whole corpus.

1 Introduction

This document describes the BioNLP Shared Task 2013 Genic Regulation Network (GNR) task. The GNR task consists in predicting a genetic network out of a set of sentences selected from PubMed abstracts.

Genic regulation networks are the primary study object in systems biology. They allow to better understand the relationship between molecular mechanisms and cellular behavior. However one of the bottlenecks in systems biology is the acquisition of an accurate genetic regulation network. In the recent years the BioNLP community has produced performant systems for extracting genic interactions and PPI from the literature. The goal of this task is to present these results as a genic regulation network, and evaluate the IE systems by their capacity to reconstruct the network.

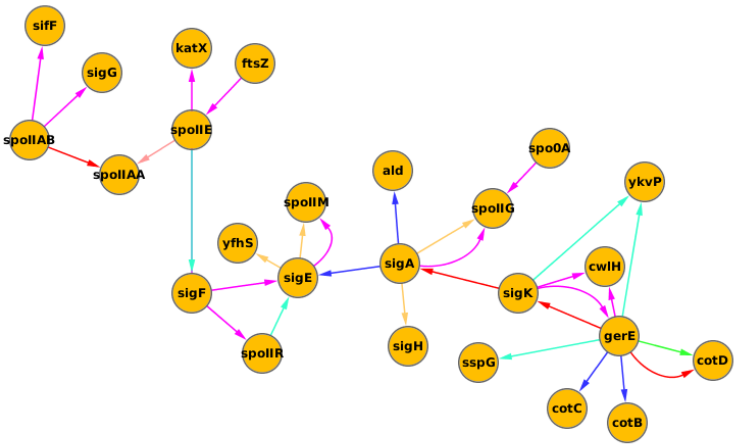
2 Representation and Task Setting

The GNR task is a relation extraction task that follows the BioNLP-ST 2013 frame of representation. The participants are provided a manually curated annotation of the training corpus including entities, events, coreferences and genic interaction. Since the GNR corpus is derived from LLL'05, we also provide gold tokenization and gold syntactic parses.

For training, the participants are provided the genic regulation network that can be reconstructed with interactions mentioned in sentences of the training

corpus. The network is a directed graph where vertexes represent genes, and arcs represent interactions between genes extracted from the text. The arcs are labeled with a type: Regulation, Activation, Inhibition, Transcription, Other.

The following figure represents the network build from the sample data:



In the text, the interactions do not necessarily link two gene entities. These entities may represent a gene product (proteins), a gene superstructure (operons) or a combination of both (regulons). We provide the name of a reference gene for each entity. In the case of superstructure, we provide the name of all the component genes, in this case one arc for each gene is generated in the network. Finally the provided gene name is unique for each gene, we resolve gene synonyms according to SubtiWiki.

The test is provided with gold entities (including reference genes). Participating systems will be evaluated by comparing the shape of the predicted genic regulation network against the gold network. Event predictions will not be evaluated, they are considered in this task as a features that may help participants to extract interactions. The detailed evaluation metrics is still to be defined.

3 Document selection

The corpus is an extension of the BioNLP-ST 2011 BI corpus derived from the LLL challenge corpus. For GNR, it has been expanded in order to cover the complete regulation network of a specific cellular function in *Bacillus subtilis*: the sporulation. This phenomenon is an adaptation of the bacteria to scarce resource conditions (low nutrients), it has been thoroughly studied in the past and the regulation network is stable and suffers no controversy.

4 Annotation Overview

The annotation is inherited from LLL and BioNLP-ST 2011 BI. This annotation is going through an additional revision iteration in order to comply with the proposed annotation schema.

4.1 Entities

Entities are organized in following hierarchy.

```
Action
Gene
GeneProduct
    RNA
        mRNA
    Protein
GeneComplex
GeneFamily
ProteinFamily
ProteinComplex
    PolymeraseComplex
Regulon
Site
    Promoter
```

All except one entity type represent either molecular or genic objects. The *Action* entity type represent potential modifiers applicable to other entities like quantity markers ("presence", "concentration") or molecular processes ("transcription", "translation"). All genic entities are associated to one or several gene names.

4.2 Events and relations

Events are organized in the following hierarchy:

```
Action_Target
    Transcription_by
    Transcription_from
Bind_to
Master_of_Promoter
Master_of_Regulon
Member_of_Regulon
Promoter_of
Site_of
```

Genic interactions are annotated as relations between genic entities. Each interaction corresponds to an arc in the regulation network between the nodes identified by the gene identifier. Interactions are broken into seven types: the label of the corresponding arc in the network is the modality of the corresponding interaction:

`Interaction.Regulation`
: observed interaction between two genes with no additional detail about its modality or mechanism.

`Interaction.Activation`
: positive interaction

`Interaction.Inhibition`
: negative interaction

`Interaction.Requirement`
: the target gene cannot be expressed without the presence and some action of the agent gene product

`Interaction.Transcription`
: the transcription of the target gene is modified by the agent gene product

`Interaction.Binding`
: the agent and target genes (or products) are physically in contact

`Interaction.Other`
: marginal modalities and mechanisms (*e.g.* phosphorylation)

Finally, coreferences are annotated as relations of type *Coreference*. They do not indicate an equivalence class; we include them in order to help participants to resolve anaphoric expressions.

5 Annotation Process

The revision of the annotation is conducted by a senior annotator with a biology and bioinformatics background. The resulting network will be checked by *Bacillus subtilis* specialists to make sure the gold network represents a scientific truth.