BioNLP Shared Task 2011 – Bacteria Gene Interactions and Renaming

**Julien Jourde[1], Alain-Pierre Manine[2], Philippe Veber[1], Karën Fort[3], Robert Bossy[1],
Erick Alphonse[2], Philippe Bessières[1]**

[1]Mathématique, Informatique et
Génome – Institut National de la
Recherche Agronomique
MIG INRA UR1077
F78352 Jouy-en-Josas, France
`forename.lastname@jouy.inra.fr`

[2]PredictiveDB
16, rue Alexandre Parodi
F75010 Paris, France
`{apmanine,alphonse}`
`@predictivedb.com`

[3]LIPN – Université Paris-Nord/
CNRS UMR7030 and
INIST CNRS UPS76 – F54514
Vandœuvre-lès-Nancy, France
`karen.fort@inist.fr`

Abstract

We present two related tasks of the BioNLP
Shared Tasks 2011: Bacteria Gene Renaming (Rename) and Bacteria Gene Interactions
(GI). We detail the objectives, the corpus specification, the evaluation metrics, and we summarize the participants' results. Both issued
from PubMed scientific literature abstracts,
the Rename task aims at extracting gene name
synonyms, and the GI task aims at extracting
genic interaction events, mainly about gene
transcriptional regulations in bacteria.

1 Introduction

The extraction of biological events from scientific
literature is the most popular task in Information Extraction (IE) challenges applied to molecular biology, such as in LLL (Nédellec, 2005), BioCreative
Protein-Protein Interaction Task (Krallinger et al.,
2008), or BioNLP (Demner-Fushman et al., 2008).
Since the BioNLP 2009 shared task (Kim et al.,
2009), this field has evolved from the extraction of a
unique binary interaction relation between proteins
and/or genes towards a broader acceptation of biological events including localization and transformation (Kim et al., 2008). In the same way, the tasks
Bacteria Gene Interactions and Bacteria Gene Renaming deal with the extraction of various molecular events capturing the mechanisms relevant to gene
regulation in prokaryotes. The study of bacteria has
numerous applications for health, food and industry, and overall, they are considered as organisms
of choice for the recent integrative approaches in
systems biology, because of their relative simplicity.

Compared to eukaryotes, they allow easier and more
in-depth analysis of biological functions and of their
related molecular mechanisms.

Processing literature on bacteria raises linguistic and semantic specificities that impact text analysis. First of all, gene renaming is a frequent phenomenon, especially for model bacteria. Hence, the
abundance of gene synonyms that are not morphological variants is high compared to eukaryotes. The
history of bacterial gene naming has led to drastic
amounts of homonyms and synonyms which are often missing (or worse, erroneous) in gene databases.
In particular, they often omit old gene names that
are no longer used in new publications, but that are
critical for exhaustive bibliography search. Polysemy makes the situation even worse, as old names
frequently happen to be reused to denote different
genes. A correct and complete gene synonym table
is crucial to biology studies, for instance when integrating large scale experimental data using distinct
nomenclatures. Indeed this information can save a
lot of bibliographic research time. The Rename Task
is a new task in text-mining for biology that aims at
extracting explicit mentions of renaming relations.
It is a critical step in gene name normalization that
is needed for further extraction of biological events
such as genic interactions.

Regarding stylistics, gene and protein interactions
are not formulated in the same way for eukaryotes and prokaryotes. Descriptions of interactions
and regulations in bacteria include more knowledge
about their molecular actors and mechanisms, compared to the literature on eukaryotes. Typically in
bacteria literature, the genic regulations are more

65

likely expressed by direct binding of the protein, while in eukaryote literature, non-genic agents related to environmental conditions are much more frequent. The bacteria GI Task is based on (Manine et al., 2010) which is a semantic re-annotation of the LLL challenge corpus (Nédellec, 2005), where the description of the GI events in a fine-grained representation includes the distinction between expression, transcription and other action events, as well as different transcription controls (e.g. regulon membership, promoter binding). The entities are not only protein agent and gene target but extend to families, complexes and DNA sites (binding sites, promoters) in order to better capture the complexity of the regulation at a molecular level. The task consists in relating the entities with the relevant relations.

2 Rename Task Description

The goal of the Rename task is illustrated by Figure 1. It consists in predicting renaming relations between text-bound gene names given as input. The only type of event is *Renaming* where both arguments are of type *Gene*. The event is directed, the former and the new names are distinguished. Genes and proteins were not distinguished because of the high frequency of metonymy in renaming events. The relation to predict between genes is a *Renaming* of a former gene name into a new one. In the example of Figure 1, YtaA, YvdP and YnzH are the former names of three proteins renamed CotI, CotQ and CotU, respectively.
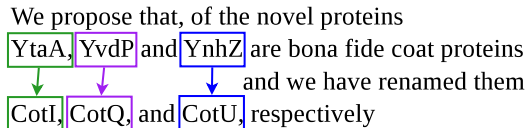

Figure 1: Examples of relations to be extracted.

2.1 Rename Task corpus

The Rename Task corpus is a set of 1,836 PubMed references of bacterial genetic and genomic studies, including title and abstract. A first set of 23,000 documents was retrieved, identifying the presence of the bacterium *Bacillus subtilis* in the text and/or in the MeSH terms. *B. subtilis* documents are particularly rich in renaming mentions. Many genes were re-

named in the middle of the nineties, so that the new names matched those of the *Escherichia coli* homologues. The 1,843 documents the most susceptible to mention renaming were automatically filtered according to two non exclusive criteria:

1. Either the document mentions at least two gene synonyms as recorded in the fusion of seven *B. subtilis* gene nomenclatures. This led to a set of 703 documents.

2. Or the document contains a renaming expression from a list that we manually designed and tested (e.g. rename, also known as). It is an extension of a previous work by (Weissenbacher, 2004). A total of 1,140 new documents not included in the first set match this criteria.

About 70% of the documents (1,146) were kept in the training data set. The rest was split into the development and test sets, containing 246 and 252 documents respectively. Table 1 gives the distribution of genes and renaming relations per corpus. Gene names were automatically annotated in the documents with the nomenclature of *B. subtilis*. Gene names involved in renaming acts were manually curated. Among the 21,878 gene mentions in the three corpus, 680 unique names are involved in renaming relations which represents 891 occurrences of genes.

	Training + Dev.	*Test*
Documents	(1,146 + 246) 1,392	252 (15%)
Gene names	18,503	3,375 (15%)
Renamings	373	88 (24%)

Table 1: Rename Task corpus content.

2.2 Rename Task annotation and guidelines

Annotation procedure The corpus was annotated in a joint effort of MIG/INRA and INIST/CNRS. The reference annotation of the Rename Task corpus was done in two steps, a first annotation step by science information professionals of INIST with MIG initial specifications, a second checking step by people at MIG. Two annotators and a project manager were in charge of the task at INIST. The documents were annotated using the Cadixe editor[1]. We

---

[1] http://caderige.imag.fr/Articles/ CADIXEXML-Annotation.pdf

provided to them detailed annotation guidelines that were largely modified in the process. A subset of 100 documents from the first set of 703 was annotated as a training session. This step was used to refine the guidelines according to the methodology described in (Bonneau-Maynard et al., 2005). Several inter-annotator agreements coefficients were computed to measure the discrepancy between annotators (Fort et al., 2009). With a *kappa* and *pi* scores (for more details on those, see (Artstein and Poesio, 2008)), the results can be considered satisfactory. The manual analysis of the 18 discrepancies led to enrich the annotation guidelines. The first hundreds of documents of the second set did not mention any renaming, leading to concentrate the annotation efforts on the first set. These documents actually contained renamings, but nearly exclusively concerning other kinds of biological entities (protein domains, molecules, cellular ultrastructures, etc.).

Guidelines In order to simplify the task, only short names of gene/protein/groups in *B. subtilis* were considered. Naming conventions set short names of four letters long with an upper case letter at the end for all genes (e.g. gerE) and the same names with the upper case of the initial letter (e.g. GerE) and long names for the proteins (e.g. Spore germination protein gerE). But many irregular gene names exist (e.g. tuf), which are considered as well. It also happens that gene or protein name lists are abbreviated by factorization to form a sequence. For instance queCDEF is the abbreviation of the list of gene names queC, queD, queE and queF. Such aggregations are acceptable gene names as well. In any case, these details were not needed by the task participants since the corpus was provided with tagged gene names.

Most renaming relations involve couples of the same type, genes, proteins or aggregations. Only 18 relations link mixed couples of genes and proteins. In case of ambiguity, annotators would consult international gene databases and an internal INRA database to help them determine whether a given couple of names were actually synonyms.

Multiple occurrences of the same renaming relation were annotated independently, and had to be predicted. The renaming pairs are directed, the former and the new forms have to be distinguished.

When the renaming order was not explicit in the document, the rule was to annotate by default the first member of the couple as the new form, and the second one as the former form. Figure 2 presents the most common forms of renaming.
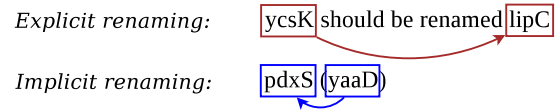

Figure 2: Common types of relations to be extracted.

Revised annotations INIST annotations were systematically checked by two experts in Bioinformatics from INRA. Mainly, encoding relations (e.g. the gene encoding sigma K (sigK)) that are not renaming cases were purged. Given the number of ambiguous annotations, we designed a detailed typology in order to justify acceptance or rejection decisions in seven different sub-cases hereafter presented. Three positive relations figure in Table 2, where the underlined names are the former names and the framed names are the new ones. Explicit renaming relations occur in 261 sentences, synonymy-like relations in 349 sentences, biological proof-based relations in 76 sentences.

Explicit renaming relation is the easiest positive case to identify. In the example, the aggregation of gene names ykvJKLM is clearly renamed by the authors as queCDEF. Although the four genes are con-

---

Explicit renaming
PMID 15767583: Genetic analysis of ykvJKLM mutants in Acinetobacter confirmed that each was essential for queuosine biosynthesis, and the genes were renamed queCDEF .
Implicit renaming
PMID 8002615: Analysis of a suppressor mutation ssb (kinC) of sur0B20 (spo0A) mutation in *Bacillus subtilis* reveals that kinC encodes a histidine protein kinase.
Biological proof
PMID 1744050: DNA sequencing established that spoIIIF and spoVB are a single monocistronic locus encoding a 518-amino-acid polypeptide with features of an integral membrane protein.

Table 2: Positive examples of the Rename Task.

catenated, there is no evidence mentioned of them acting as an operon. Furthermore, despite the context involving mutants of Acinetobacter, the aggregation belongs correctly to *B. subtilis*.

Implicit renaming is an asymmetric relation since one of the synonyms is intended to replace the other one in future uses. The example presents two renaming relations between former names ssb and spo0A, and new names kinC and sur0B20, respectively. The renaming relation between ssb and kinC has a different orientation due to additional information in the reference. Like in the preceding example, the renaming is a consequence of a genetic mutation experiment. Mutation names represent an important transversal issue that is discussed below.

Biological proof is a renaming relation induced by an explicit scientific conclusion while the renaming is not, as in the example where experiments reveal that two loci spoIIIF and spoVB are in fact the same one and then become synonyms. Terms such as "allelic to" or "identical to" usually qualify such conclusions. Predicting biological proof-based relations requires some biological modeling.

The next three cases are negative (Table 3). Underlined gene and protein names are involved in a relation which is not a renaming relation.

Protein encoding relation occurs between a gene and the protein it codes for. Some mentions may look like renaming relations. The example presents the gene yeaC coding for MoxR. No member of the couple is expected to replace the other one.

Homology measures the similarity between gene or protein sequences. Most of the homology mentions involve genes or proteins from different species (orthologues). The others compare known gene or protein sequences of the same species (paralogues). This may be misleading since the similarity mention may look like biological proof-based relations, as between ParA and Soj in Table 3.

Operon, regulon or family renaming involves objects that may look like genes, proteins or simple aggregations of gene or protein names but that are perceptibly different. The objects represent more than one gene or protein and the renaming does not necessarily affect all of them. More problematic, their name may be the same as one of the genes or proteins they contain, as in the example where sigA and rpoD are operons but are also known as gene names. Here, sigA (and so rpoD) represents at least two different genes. For the sake of clarity, operons, regulons and families are rejected, even if all the genes are clearly named, as in an aggregation.

The last point concerns ***mutation*** which are frequent in Microbiology for revealing gene phenotypes. They carry information about the original gene names (e.g., rvtA11 is a mutant name created by adding 11 to rvtA). But partial names cannot be partially annotated, that is to say, the original part (rvtA) should not be annotated in the mutation name (rvtA11). Most of these names are local names, and should not be annotated because of their restricted scope. It may happen so that the mutation name is registered as a synonym in several international databases. To avoid inconsistencies, all renamings involving a mutation referenced in a database were accepted, and only biological proof-based and explicit renamings involving a strict non-null unreferenced mutation (a null mutation corresponds to a total suppression of a gene) were accepted.

2.3 Rename Task evaluation procedure

The evaluation of the Rename task is given in terms of recall, precision and F-score of renaming relations. Two set of scores are given: the first set is computed by enforcing strict direction of renaming relations, the second set is computed with relaxed direction. Since the relaxed score takes into account renaming relations even if the arguments are inverted, it will necessarily be greater or equal than the strict score. The participant score is the relaxed score, the strict score is given for information. Relaxed scores are informative with respect to the ap-

Protein encoding
PMID 8969499: The putative products of ORFs yeaB (Czd protein), yeaC (MoxR), yebA (CNG-channel and cGMP-channel proteins from eukaryotes),
Genetic homology
PMID 10619015: Dynamic movement of the ParA-like Soj protein of *B. subtilis* and its dual role in nucleoid organization and developmental regulation.
Operon | Regulon | Family
PMID 3127379: Three promoters direct transcription of the sigA (rpoD) operon in *Bacillus subtilis*.

Table 3: Negative examples of the Rename Task.

plication goal. The motivation of the Rename task is to keep bacteria gene synonyms tables up to date. The choice of the canonical name among synonyms for denoting a gene is done by the bacteriology community, and it may be independent of the anteriority or novelty of the name. The annotation of the reference corpus showed that the direction was not always decidable, even for a human reader. Thus, it would have been unfair to evaluate systems on the basis of unsure information.

2.4 Results of the Rename Task participants

Final submissions were received from three teams, the University of Turku (Uturku), the University of Concordia (Concordia) and the Bibliome team from MIG/INRA. Their results are summarized in Table 4. The ranking order is given by the overall F-score for relations with relaxed argument order.

Team	Prec.	Recall	F-score
Univ. of Turku	**95.9**	**79.6**	**87.0**
Concordia Univ.	74.4	65.9	69.9
INRA	57.0	73.9	64.4

Table 4: Participant scores at the Rename Task.

Uturku achieved the best F-score with a very high precision and a high recall. Concordia achieved the second F-score with balanced precisions and recalls. Bibliome is five points behind with a better recall but much lower precision. Both UTurku and Concordia predictions rely on dependencies (Charniak-Johnson and Stanford respectively, using McClosky model), whereas Bibliome predictions rely on bag of words. This demonstrates the high value of dependency parsing for this task, in particular for the precision of predictions. We notice that UTurku system uses machine learning (SVM) and Concordia uses rules based on trigger words. The good results of UTurku confirms the hypothesis that gene renaming citations are highly regular in scientific literature. The most frequently missed renamings belong to the Biological Proof category (see Table 2). This is expected because the renaming is formulated as a reasoning where the conclusion is only implicit.

2.5 Discussion

The very high score of Uturku method leads us to conclude that the task can be considered as solved

by a linguistic-based approach. Whereas Bibliome used an extensive nomenclature considered as exhaustive and sentence filtering using a SVM, Uturku used only two nomenclatures in synergy but with more sophisticated linguistic-based methods, in particular syntactic analyses. Bibliome methods showed that a too high dependence to nomenclatures may decrease scores if they contain compromised data. However, the use of an extensive nomenclature as done by Bibliome may complement Uturku approach and improve recall. It is also interesting that both systems do not manage renamings crossing sentence boundaries.

The good results of the renaming task will be exploited to keep synonym gene lists up to date with extensive bibliography mining. In particular this will contribute to enriching SubtiWiki, a collaborative annotation effort on *B. subtilis* (Flórez et al., 2009; Lammers et al., 2010).

3 Gene Interactions Task description

The goal of the Bacteria GI Task is illustrated by Figure 3. The genes cotB and cotC are related to their two promoters, not named here, by the relation *PromoterOf*. The protein GerE is related to these promoters by the relation *BindTo*. As a consequence, GerE is related to cotB and cotC by an *Interaction* relation. According to (Kim et al., 2008), the need to define specialized relations replacing one unique and general interaction relation was raised in (Manine et al., 2009) for extracting genic interactions from text. An ontology describes relations and entities (Manine et al., 2008) catching a model of gene transcription to which biologists implicitly refer in their publications. Therefore, the ontology is mainly oriented towards the description of a structural model of genes, with molecular mechanisms of their transcription and associated regulations.

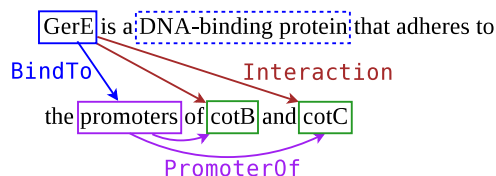The corpus roughly contains three kinds of genic

Figure 3: Examples of relations to be extracted.

interaction mentions, namely regulations, regulon membership and binding. The first case corresponds to interactions the mechanism of which is not explicitly given in the text. The mention only tells that the transcription of a given gene is influenced by a given protein, either positively (activation), negatively (inhibition) or in an unspecified way. The second kind of genic interaction mention (regulon membership) basically conveys the same information, using the regulon term/concept. The regulon of a gene is the set of genes that it controls. In that case, the interaction is expressed by saying that a gene is a member of some regulon. The third and last kind of mention provides with more mechanistic details on a regulation, since it describes the binding of a protein near the promoter of a target gene. This motivates the introduction of *Promoter* and *Site* entities, which correspond to DNA regions. It is thus possible to extract the architecture of a regulatory DNA region, linking a protein agent to its gene target (see Figure 3).

The set of entity types is divided into two main groups, namely 10 genic entities and 3 kinds of action (Table 5). Genic entities represent biological objects like a gene, a group of genes or a gene product. In particular, a *GeneComplex* annotation corresponds to an operon, which is a group of genes that are contiguous in the genome and under the control of the same promoter. The annotation *GeneFamily* is used to denote either genes involved in the same biological function or genes with sequence homologies. More importantly, *PolymeraseComplex* annotations correspond to the protein complex that is responsible for the transcription of genes. This complex includes several subunits (components), combined with a sigma factor, that recognizes specific promoters on the DNA sequence.

The second group of entities are phrases expressing either molecular processes (e.g. sequestration, dephosphorylation, etc.) or the molecular state of the bacteria (e.g. presence, activity or level of a protein). They represent some kind of action that can be performed on a genic entity. Note that transcription and expression events were tagged as specific actions, because they play a specific part in certain relations (see below).

The annotation of entities and actions was provided to the participants, and the task consisted in extracting the relations listed in Table 6.

Name	Example
Gene	cotA
GeneComplex	sigX-ypuN
GeneFamily	class III heat shock genes
GeneProduct	yvyD gene product
Protein	CotA
PolymeraseComplex	SigK RNA polymerase
ProteinFamily	DNA-binding protein
Site	upstream site
Promoter	promoter regions
Regulon	regulon
Action	activity	level	presence
Expression	expression
Transcription	transcription

Table 5: List of molecular entities and actions in GI.

Name	Example
ActionTarget	**expression** of *yvyD*
Interaction	**ComK** negatively regulates *degR* expression
RegulonDependence	*sigmaB* **regulon**
RegulonMember	*yvyD* is member of sigmaB **regulon**
BindTo	**GerE** adheres to the *promoter*
SiteOf	**-35 sequence** of the *promoter*
PromoterOf	the *araE* **promoter**
PromoterDependence	*GerE*-controlled **promoter**
TranscriptionFrom	**transcription** from the *upstream site*
TranscriptionBy	**transcription** of cotD by *sigmaK RNA polymerase*

Table 6: List of relations in GI.

The relations are binary and directed, and rely the entities defined above. The three kinds of interactions are represented with an *Interaction* annotation, linking an agent to its target. The other relations provide additional details on the regulation, like elementary components involved in the reaction (sites, promoters) and contextual information (mainly provided by the *ActionTarget* relations). A formal definition of relations and relation argument types can be found on the Bacteria GI Task Web page.

3.1 Bacteria Gene Interactions corpus

The source of the Bacteria GI Task corpus is a set of PubMed abstracts mainly dealing with the tran-

scription of genes in *Bacillus subtilis*. The semantic annotation, derived from the ontology of (Manine et al., 2008), contains 10 molecular entities, 3 different actions, and 10 specialized relations. This is applied to 162 sentences from the LLL set (Nédellec, 2005), which are provided with manually checked linguistic annotations (segmentation, lemmatization, syntactic dependencies). The corpus was split into 105 sentences for training, 15 for development and 42 for test. Table 7 gives the distribution of the entities and actions per corpus and Table 8 gives the distribution of the relations per corpus.

3.2   Annotation procedures and guidelines

The semantic annotation scheme was developed by two annotators through a series of independent annotations of the corpus, followed by reconciliation steps, which could involve concerted modifications (Manine et al., 2010). As a third and final stage, the corpus was reviewed and the annotation simplified to make it more appropriate to the contest. The final annotation contains 748 relations distributed in nine categories, 146 of them belonging to the test set.

The annotation scheme was generally well suited to accurately represent the meaning of the sentences in the corpus, with one notable exception. In the corpus, there is a common phrasing telling that a protein P regulates the transcription of a gene G by a given sigma factor S. In that case, the only annotated interactions are between the couples (P, G) and (S, G). This representation is not completely satisfactory, and a ternary relation involving P, S and G would have been more adequate.

Additional specific rules were needed to cope with linguistic issues. First, when the argument of a relation had coreferences, the relation was repeated for each maximally precise coreference of the argument. Second, in case of a conjunction like "sigmaA and sigmaX holoenzymes", there should ideally be two entities (namely "sigmaA holoenzyme" and "sigmaX holoenzyme"); however, this is not easy to represent using the BioNLP format. In this situation, we grouped the two entities into a single one. These cases were rare and unlikely affected the feasibility of the task, since entities were provided in the test set.

Entity or action	Train. + Dev.	Test
Documents	(105+15) 120	42
Protein	219	85
Gene	173	56
Transcription	53	21
Promoter	49	10
Action	45	22
PolymeraseComplex	43	14
Expression	29	6
Site	22	8
GeneComplex	19	4
ProteinFamily	12	3
Regulon	11	2
GeneProduct	10	3
GeneFamily	6	5

Table 7: Distribution of entities and actions in GI.

Relation	Train. + Dev.	Test
Interaction	208	64
ActionTarget	173	47
PromoterOf	44	8
BindTo	39	4
PromoterDependence	36	4
TranscriptionBy	36	8
SiteOf	23	6
RegulonMember	17	2
TranscriptionFrom	14	2
RegulonDependence	12	1

Table 8: Distribution of relations in GI.

3.3   Gene Interactions evaluation procedure

The training and development corpora with the reference annotations were made available to participants by December, 1st on the BioNLP shared Task pages together with evaluation software. The test corpus with the entity annotations has been made available by March, 1st. The participants sent the predicted annotations to the BioNLP shared Task organizers by March, 10th. The evaluation results were computed and provided to the participants and on the Web site the same day. The participants are evaluated and ranked according to two scores: F-score for all event types together, and F-score for the *Interaction* event type. In order for a predicted event to count as a hit, both arguments must be the same as in the reference in the right order and the event type must be the same as in the reference.

3.4 Results of GI Task participants

There was only one participant, whose results are shown in Tables 9 and 10. Some relations were not significantly represented in the test set and thus the corresponding results should be considered with caution. This is the case for *RegulonMember* and *TranscriptionFrom*, only represented two times each in the test. The lowest recall, 17%, obtained for the *SiteOf* relation is explained by its low representation in the corpus: most of the test errors come from a difficult sentence with coreferences.

The recall of 56% for the *Interaction* relation certainly illustrates the heterogeneity of this category, gathering mentions of interactions at large, as well as precise descriptions of gene regulations. For instance, Figure 4 shows a complex instance where all of the interactions were missed. Surprisingly, we also found false negatives in rather trivial examples ("*ykuD was transcribed by SigK RNA polymerase from T4 of sporulation.*"). Uturku used an SVM-based approach for extraction, and it is thus delicate to account for the false negatives in a simple and concise way.

Event	U. Turku scores
Global Precision	85
Global Recall	71
Global F-score	77
Interaction Precision	75
Interaction Recall	56
Interaction F-score	64

Table 9: University of Turku global scores.

Event	Prec.	Rec.	F-score
Global	85	71	77
ActionTarget	94	92	93
BindTo	75	75	75
Interaction	75	56	64
PromoterDependence	100	100	100
PromoterOf	100	100	100
RegulonDependence	100	100	100
RegulonMember	100	50	67
SiteOf	100	17	29
TranscriptionBy	67	50	57
TranscriptionFrom	100	100	100

Table 10: University of Turku scores for each relation.

The addition of ClpX to in vitro transcription reactions resulted in the stimulation of RNAP holoenzyme activity, but sigmaH-RNAP was observed to be more sensitive to ClpX-dependent stimulation than sigmaA-RNAP.

Figure 4: Examples of three missed interactions.

3.5 Discussion

The GI corpus was previously used in a relation extraction work (Manine et al, 2009) based on Inductive Logic Programming (Muggleton and Raedt, 1994). However a direct comparison of the results is not appropriate here since the annotations were partially revised, and the evaluation setting was different (leave-one-out in Manine's work, test set in the challenge).

Nevertheless, we note similar tendencies if we compare relative results between relations. In particular, it was also found in Manine's paper that *SiteOf*, *TranscriptionBy* and *Interaction* are the most difficult relations to extract. It is also worth to mention that both approaches rely on syntactic dependencies, and use the curated dependencies provided in the corpus. Interestingly, the approach by the University of Turku reports a slightly lower F-measure with dependencies calculated by the Charniak parser (about 1%, personal communication). This information is especially important in order to consider a production setting.

4 Conclusion

The quality of results for both challenges suggests that current methods are mature enough to be used in semi-automatic strategies for genome annotation, where they could efficiently assist biological experts involved in collaborative annotation efforts (Lammers et al., 2010). However, the false positive rate, notably for the *Interaction* relation, is still too high for the extraction results to be used as a reliable source of information without a curation step.

References

Artstein R., Poesio M. (2008). Inter-coder agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555-96.

Björne J., Heimonen J., Ginter F., Airola A., Pahikkala T., Salakoski T. (2009). Extracting complex biological events with rich graph-based feature sets. BioNLP'09 Proc. *Workshop Current Trends in Biomedical Natural Language Processing: Shared Task*, pp. 10-18.

Bonneau-Maynard H., Rosset S., Ayache C., Kuhn A., Mostefa D. (2005). Semantic annotation of the French Media Dialog Corpus. *Interspeech-2005*, pp. 3457-60.

Demner-Fushman D., Ananiadou S., Cohen K.B., Pestian J., Tsujii J., Webber B. (2008). Themes in biomedical natural language processing: BioNLP08. *BMC Bioinformatics*, 9(Suppl. 11):S1.

Flórez L.A., Roppel S.F., Schmeisky A.G., Lammers C.R., Stülke J. (2009). A community-curated consensual annotation that is continuously updated: The *Bacillus subtilis* centred wiki SubtiWiki. *Database*, 2009:bap012.

Fort K., François C., Ghribi M. (2010). Évaluer des annotations manuelles dispersées : les coefficients sont-ils suffisants pour estimer l'accord inter-annotateurs ? *17ᵉ Conf. Traitement Automatique des Langues Naturelles (TALN 2010)*.

Kim J.D., Ohta T., Tsujii J. (2008) Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10.

Kim J.D., Ohta T., Pyysalo S., Kano Y., Tsujii J. (2009). Overview of BioNLP'09 shared task on event extraction. *BioNLP'09 Proc. Workshop Current Trends in Biomedical Natural Language Processing: Shared Task*, pp. 1-9.

Krallinger M., Leitner F., Rodriguez-Penagos C., Valencia A. (2008). Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, 9(Suppl. 2):S4.

Lammers C.R., Flórez L.A., Schmeisky A.G., Roppel S.F., Mäder U., Hamoen L., Stülke J. (2010). Connecting parts with processes: SubtiWiki and Subti-Pathways integrate gene and pathway annotation for *Bacillus subtilis. Microbiology*, 156(3):849-59.

Manine A.P., Alphonse E., Bessières P. (2008). Information extraction as an ontology population task and its application to genic interactions. *20th IEEE Int. Conf. Tools with Artificial Intelligence (ICTAI'08)*, pp. 74-81.

Manine A.P., Alphonse E., Bessières P. (2009). Learning ontological rules to extract multiple relations of genic interactions from text. *Int. J. Medical Informatics*, 78(12):e31-8.

Manine A.P., Alphonse E., Bessières P. (2010). Extraction of genic interactions with the recursive logical theory of an ontology. *Lecture Notes in Computer Sciences*, 6008:549-63.

Muggleton S., Raedt L.D. (1994) Inductive Logic Programming: Theory and methods. *J. Logic Programming*, 19-20:629-79.

Nédellec C. (2005). Learning Language in Logic – Genic Interaction Extraction Challenge. *Proc. 4th Learning Language in Logic Workshop (LLL'05)*, pp. 31-7.

Weissenbacher, D. (2004). La relation de synonymie en Génomique. *RECITAL 2004 Conference*.