mig

Unité Mathématique, Informatique et Génome

Extraction de régulation à partir d'articles

Claire Nédellec et Equipe *Bibliome*

Réseau méthodologique Inférence de réseaux

Paris – 9 février 2012

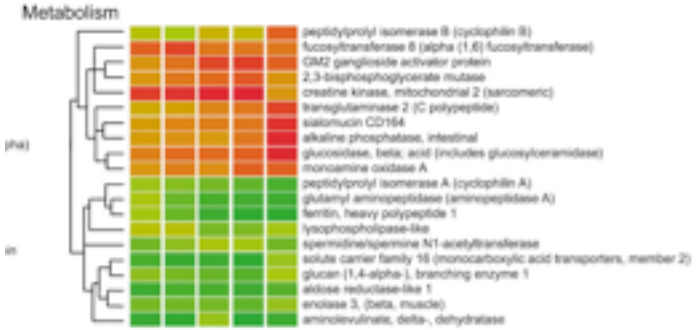
ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

INRA

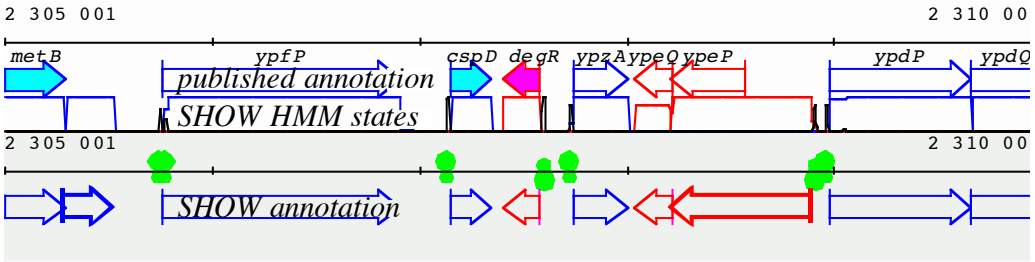
Documents textuels, une source majeure d'information en biologie

Expériences à grande échelle

sequençage, expression transcriptome et protéome, recherche de phénotype



Annotation
of sequences
(genes, function)

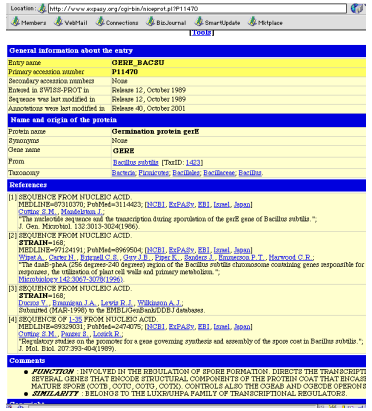


Modèles prédictifs

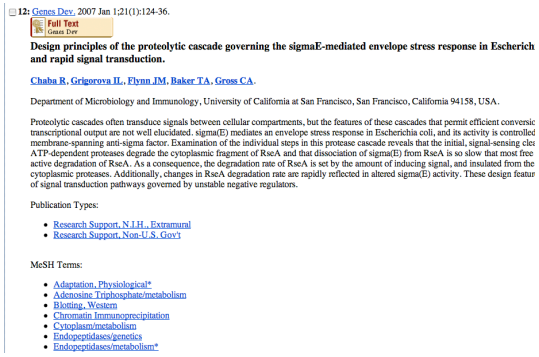
Math-info : analyse des séquences, structures, données d'expression

inférence de réseau de régulation

découverte de motifs



Textes en langage naturel



Design principles of the proteolytic cascade governing the sigmaE-mediated envelope stress response in Escherich and rapid signal transduction.

Chaba R, Grigorova IL, Flynn JM, Baker TA, Gross CA.

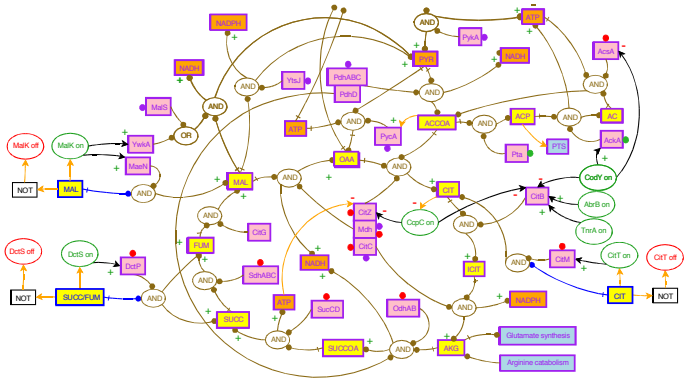
Department of Microbiology and Immunology, University of California at San Francisco, San Francisco, California 94158, USA.

Publication Types:

- Research Support, N.I.H., Extramural
- Research Support, Non-U.S. Gov't

MeSH Terms:

- Adaptation, Physiological*
- Adenosine Triphosphate/metabolism
- Blotting, Western
- Chromatin Immunoprecipitation
- Cytoplasm/metabolism
- Endopeptidases/genetics
- Endopeptidases/metabolism*



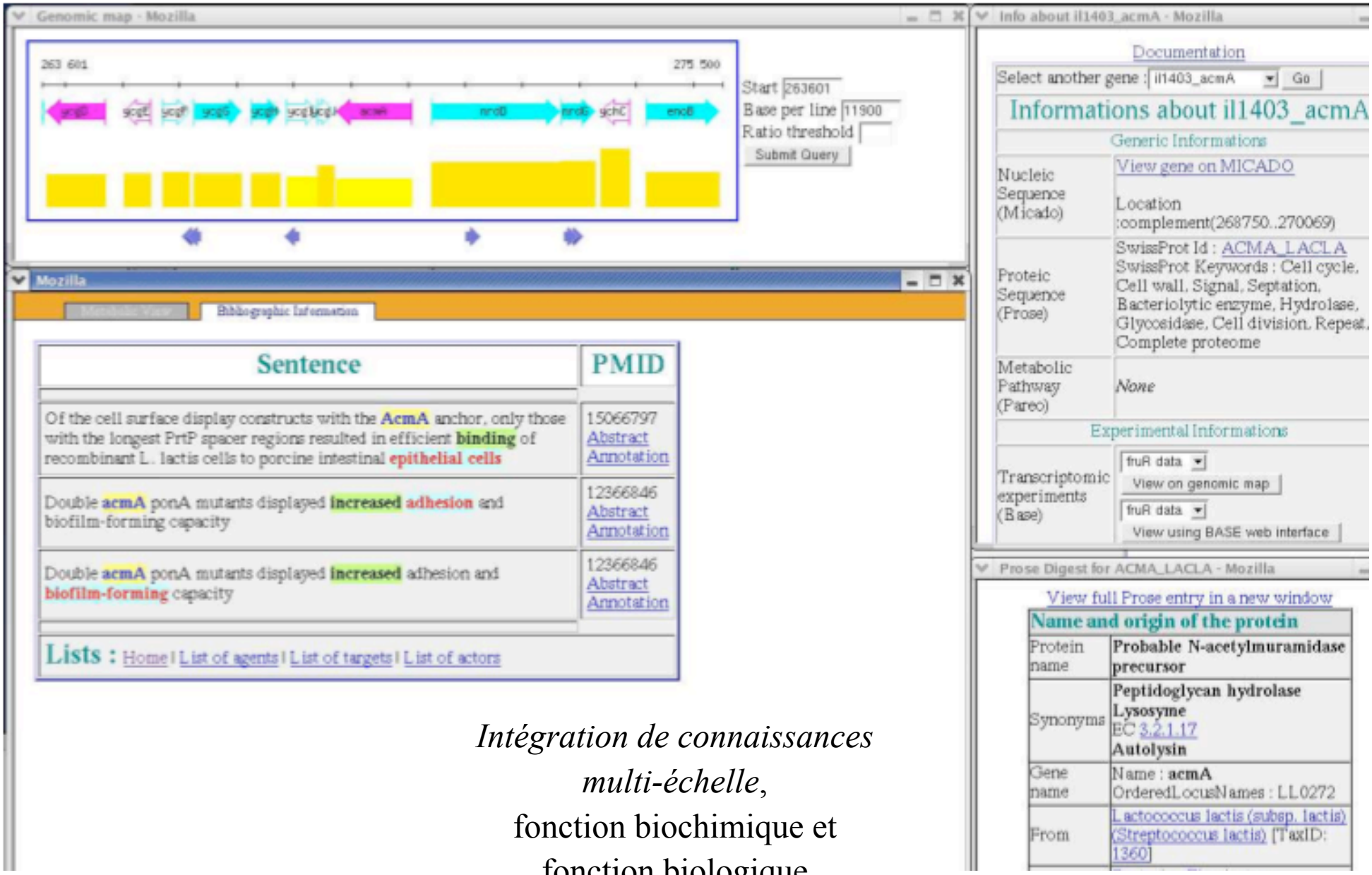
La bibliographie, une source riche d'information en biologie

Des connaissances essentielles, comme les régulations, sont décrites uniquement dans le texte des publications scientifiques.

- **Bibliographie** en biologie **centralisée**: *(> 21 millions de références dans MedLine)*
  Références et résumés libres d'accès. Texte de l'article accessible sous réserve d'abonnement.
- **Base de données avec commentaires** : ex. : champ comment *function* de SwissProt

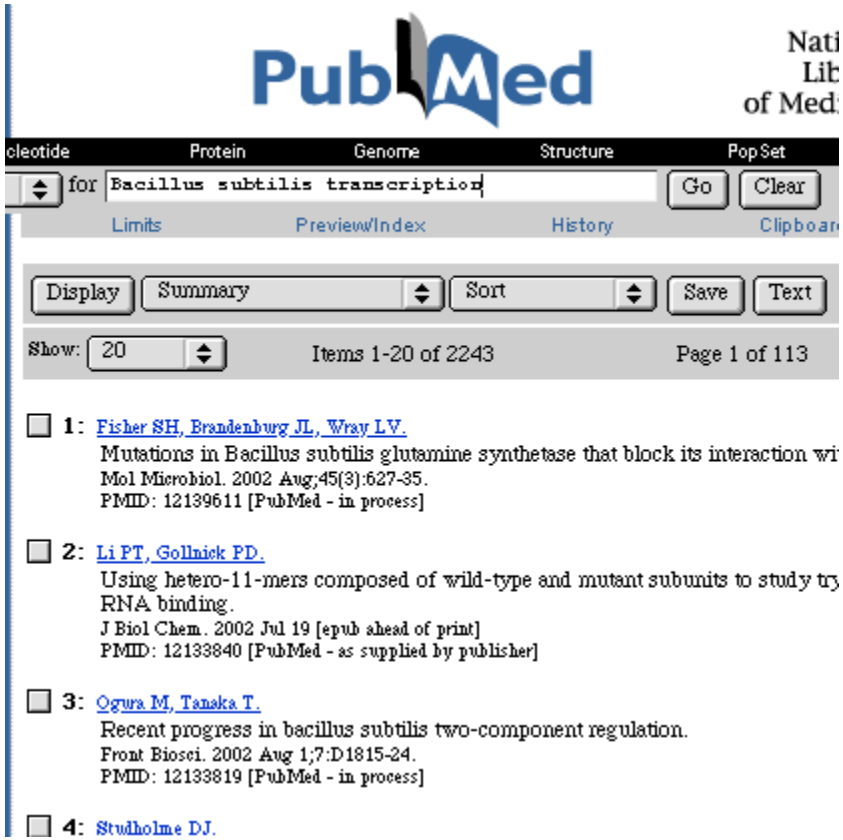
➔ Intérêt de l'automatisation de l'extraction d'information à partir de texte libre.

- Le retour au texte constitue de loin l'étape la plus limitante de l'annotation de gènes nouvellement séquencés [Bessières, 2011] ;
- Cette connaissance est nécessaire pour construire, valider et interpréter les expériences à grande échelle ;
- En biologie des systèmes, pas de modèle réaliste de la cellule sans régulation.



Intégration de connaissances multi-échelle,
fonction biochimique et
fonction biologique

PubMed


PubMed
Nati
Lib
of Med
cleotide
Protein
Genome
Structure
Pop Set
for
Bacillus subtilis transcription
Go
Clear
Limits
Preview/Index
History
Clipboar
Display
Summary
Sort
Save
Text
Show:
20
Items 1-20 of 2243
Page 1 of 113
1: Fisher SH, Brandenburg JL, Wray LV.
Mutations in Bacillus subtilis glutamine synthetase that block its interaction wi
Mol Microbiol. 2002 Aug;45(3):627-35.
PMID: 12139611 [PubMed - in process]
2: Li PT, Gollnick PD.
Using hetero-11-mers composed of wild-type and mutant subunits to study try
RNA binding.
J Biol Chem. 2002 Jul 19 [epub ahead of print]
PMID: 12133840 [PubMed - as supplied by publisher]
3: Ogura M, Tanaka T.
Recent progress in bacillus subtilis two-component regulation.
Front Biosci. 2002 Aug 1;7:D1815-24.
PMID: 12133819 [PubMed - in process]
4: Studholme DJ.


PubMed
National
Library
of Medicine
otide
Protein
Genome
Structure
Pop Set
for
Go
Clear
Limits
Preview/Index
History
Clipboard
Display
Abstract
Sort
Save
Text
1: J Biol Chem 1999 Mar 19;274(12):8322-7
FREE full text article at
www.jbc.org
Negative regulation by the Bacillus subtilis GerE protein.
Ichikawa H, Halberg R, Kroos L.
Department of Biochemistry, Michigan State University, East Lansing, Michigan 48
GerE is a transcription factor produced in the mother cell compartment of sporulatin
genes encoding proteins that form the spore coat late in development. Most cot gene
sigmaK RNA polymerase. Previously, it was shown that the GerE protein inhibits t
sigmaK. Here, we show that GerE binds near the sigK transcriptional start site, to a
containing the GerE-binding site in the promoter region was expressed at a 2-fold lo
than gerE mutant cells. Likewise, the level of SigK protein (i. e. pro-sigmaK and si
than in a gerE mutant. These results demonstrate that sigmaK-dependent transcriptio
which GerE acts as a repressor to limit production of sigmaK. In addition, GerE dir
genes. We show that GerE binds to two sites that span the -35 region of the cotD pr
transcription of cotD by sigmaK RNA polymerase in vitro, but a higher level of Ger

Un exemple de référence PubMed

UI - 99175219
AU - Ichikawa H
AU - Halberg R
AU - Kroos L
TI - Negative regulation by the Bacillus subtilis GerE protein.
..
PT - JOURNAL ARTICLE
..
DP - 1999 Mar 19
IS - 0021-9258
TA - J Biol Chem
AB - GerE is a transcription factor produced in the mother cell compartment of sporulating Bacillus subtilis. It is a critical regulator of cot genes encoding proteins that form the spore coat late in development. Most cot genes, and the gerE gene, are transcribed by sigmaK RNA polymerase. Previously, it was shown that the GerE protein inhibits transcription in vitro of the sigK gene encoding sigmaK. Here, we show that GerE binds near the sigK transcriptional start site, to act as a repressor. A sigK-lacZ fusion containing the GerE-binding site in the promoter region was expressed at a 2-fold lower level during sporulation of wild-type cells than gerE mutant cells. Likewise, the level of SigK protein (i. e. pro-sigmaK and sigmaK) was lower in sporulating wild-type cells than in a gerE mutant. These results demonstrate that sigmaK-dependent transcription of gerE initiates a negative feedback loop in which GerE acts as a repressor to limit production of sigmaK. In addition, GerE directly represses transcription of particular cot genes. We show that GerE binds to two sites that span the -35 transcription. The upstream GerE-binding site was required for activation but not for repression. These results suggest that a rising level of GerE in sporulating cells may first activate cotD transcription from the upstream site then repress transcription as the downstream site becomes occupied. Negative regulation by GerE, in addition to its positive effects on transcription, presumably ensures that sigmaK and spore coat proteins are synthesized at optimal levels to produce a germination-competent spore.

AD - Department of Biochemistry, Michigan State University, East Lansing,
      Michigan 48824, USA.PMID- 0010075739
EDAT- 1999/03/13 03:11
URL - http://www.jbc.org/cgi/content/full/274/12/8322
SO - J Biol Chem 1999 Mar 19;274(12):8322-7

Les mentions d'interaction géniques sont abondantes

Construire des réseaux de régulation à partir de texte

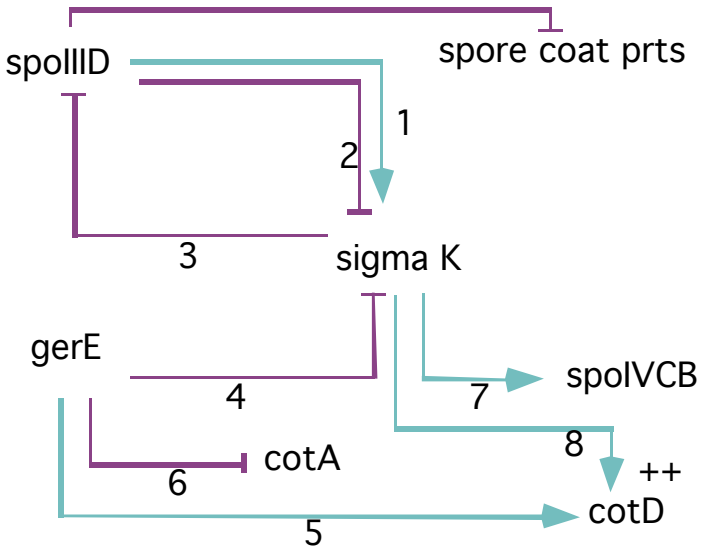
Depuis les années 2000 [Bessières, Nedellec]

positive interact.

negative interact.

1. SpoIIID is needed to produce sigma K

2. SpoIIID is capable of altering the specificity of RNAP-sigma K



3. Production of sigma K leads to a decrease in the level of spoIIID

4. GerE profoundly inhibits in vitro transcription of sigK encoding sigma K

5. GerE stimulates cotD transcription

6. ... and inhibits cotA transcription.

7. sigma K has been found that causes weak transcription of spoIVCB

8. ... and strong transcription of cotD.

Extraction d'information à partir de texte

- Domaine émergeant dans les années 90. Communauté structurée par la série des conférences MUC (*Message Understanding Conference*), compétitions internationales de la DARPA.

Objectif : remplir automatiquement un formulaire *d'événement*

Texte [Soderland, 95]

Capitol Hill - 1 br twnhme. fplc D/W W/D. Undrgrnd pkg incl $675. 3BR, upper flr or turn of ctry HOME. incl gar, grt N. Hill loc $995. (206) 999-9999

Formulaires

Neighborhood: Capitol Hill	**Neighborhood**: Capitol Hill
Bedrooms: 1	**Bedrooms**: 3
Price: 675	**Price**: 995

.

- De très nombreux systèmes. AutoSlog [Riloff, 93-99], SRV [Freitag, 98], Whisk [Soderland, 99], Pinocchio [Ciravegna, 00]

- Distinguant en particulier la reconnaissance d'entités nommées et les relations

Exemple d'extraction d'interaction génique

À partir d'une phrase

[..] the GerE protein inhibits transcription in vitro of the sigK gene encoding sigmaK [..]

Connaissance structurée (formulaire)

Interaction	**Type :** négative		
	Agent : GerE		
	Cible :	**Expression**	**Source :** sigK **Produit :** sigmaK

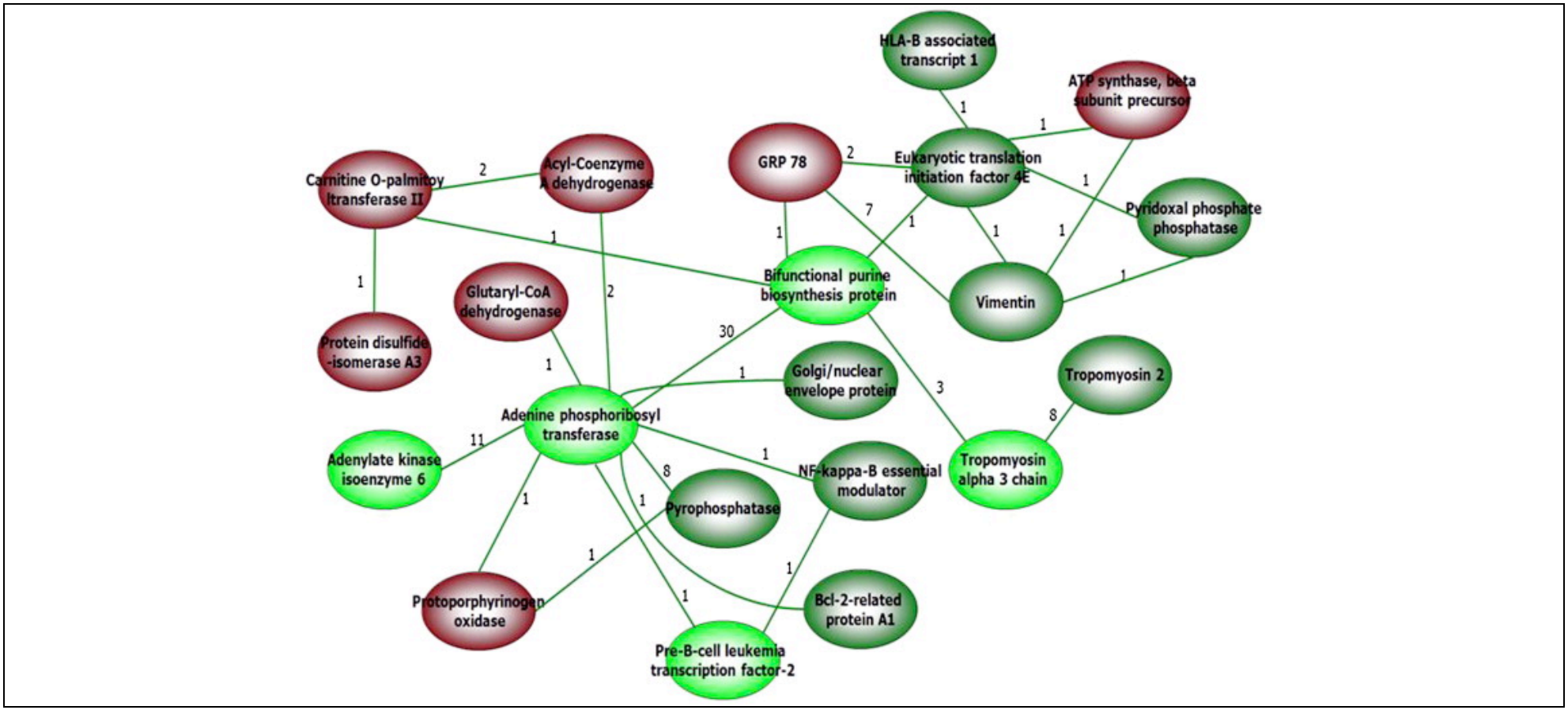
Réseau d'interaction génique, à partir de réseau de cocitation de noms

Les interactions géniques entre la protéine A et le gène B se manifestent par des cocitations fréquentes de A et B dans les phrases, résumés et articles.

(voir les outils MedMiner, PubGene, Ihop, etc.)

Etapes

- Constituer une liste des noms de gènes et protéines d'intérêt (*SwissProt*, *GenBank*, *RefSeq*)
- Constituer un corpus de documents (*PubMed*, *WoS*)
- Segmenter les documents en phrases et en mots
- Filtrer les phrases qui contiennent au moins deux noms de gènes ou de protéines
- Compter les cocitations
- Développer une interface d'interrogation des données de cocitation
  - Visualisation du réseau de cocitation
  - Recherche documentaire de cocitation



Le réseau obtenu avec **PubGene** en cherchant les associations de protéines cocitées avec *adenine phosphoribosyltransferase*. [Kang et al., 09]

Le nombre de références Medline où les protéines sont cocitées est porté sur les arcs.

http://genome.jouy.inra.fr/migale/alemacon/public_html/cgi-bin/cocitations_inter/index.cgi

COCITATIONS DATABASE

Reset

==> ENTITY TYPE : gene or protein back
==> FOR SPECIES : Bacillus subtilis back
==> SELECTED ENTITY : BdbC back

RESULTS

Regular name: BdbC ? Help

Number of cocited entitie(s): 15

Results sorting	⦿ alphabetic ○ occurence by sentence ○ occurence by abstract	SORT
Results grouping	⦿ ungroup ○ by sentence ○ by abstract	GROUP

gene or protein name	in sentences	in abstracts
bdbA (BSU21460)		2
BdbA	2	2
bdbB (BSU21440)	1	2
BdbB	6	2
bdbC (BSU33470)	1	3
BdbD	7	3
bdbD (BSU33480)		2
CcdA	1	1
ComEA		1
ComGA		1
ComGC	1	1
ComK		1
htrA (BSU12900)		1
PhoA	2	2
SunT		1

MIG Mathématique Informatique & Génome

INRA

Cocitations for gene or protein BdbC with BdbD

Bacillus subtilis

Sentence	PMID
Four enzymes of this type, termed **BdbA**, **BdbB**, **BdbC**, and **BdbD**, have been identified in the Gram-positive eubacterium Bacillus subtilis.	11872755
BdbC and **BdbD** have been shown to be critical for the folding of a protein required for DNA uptake during natural competence.	11872755
BdbC and **BdbD** are orthologs of enzymes known to be involved in extracytoplasmic disulfide bond formation.	11744713
Taken together, these observations imply that in the absence of either **BdbC** or **BdbD**, **ComGC** is unstable and that **BdbC** and **BdbD** catalyze the formation of disulfide bonds that are essential for the DNA binding and uptake machinery.	11744713
Consistent with this, **BdbC** and **BdbD** are needed for the secretion of the Escherichia coli disulfide bond-containing alkaline phosphatase, **PhoA**, by B. subtilis.	11744713
BdbC and **BdbD** are thiol-disulfide oxidoreductases.	11844773
Mutations in the thiol-disulfide oxidoreductases **BdbC** and **BdbD** can suppress cytochrome c deficiency of **CcdA**-defective Bacillus subtilis cells.	11844773

Première étape, constituer le corpus de documents

Les documents à traiter sont,

A identifier

Par exemple, définir une requête sur le site Web de la collection.
Ex. PubMed *transcription and Bacillus subtilis.*

Déjà identifiés

Ex. résultat d'une étude bibliographique préalable. Ensemble des documents référencés par une base de données. Les champs commentaire d'une base de donnée.

A télécharger

De nombreux sites sont interrogeables par programme, web services (*Esp@ceNet*, *Web of Science*), URL (*Google*, *PubMed*), bases de données (*Prose*).
Ex. Constituer un corpus de références à partir de PubMed : 3 857 références.

A transformer en texte

Abby, *Acrobat*, *pdf2txt*. Eventuellement nettoyage des balises html, des tableaux et figures.

Segmenter en phrase ou en mots

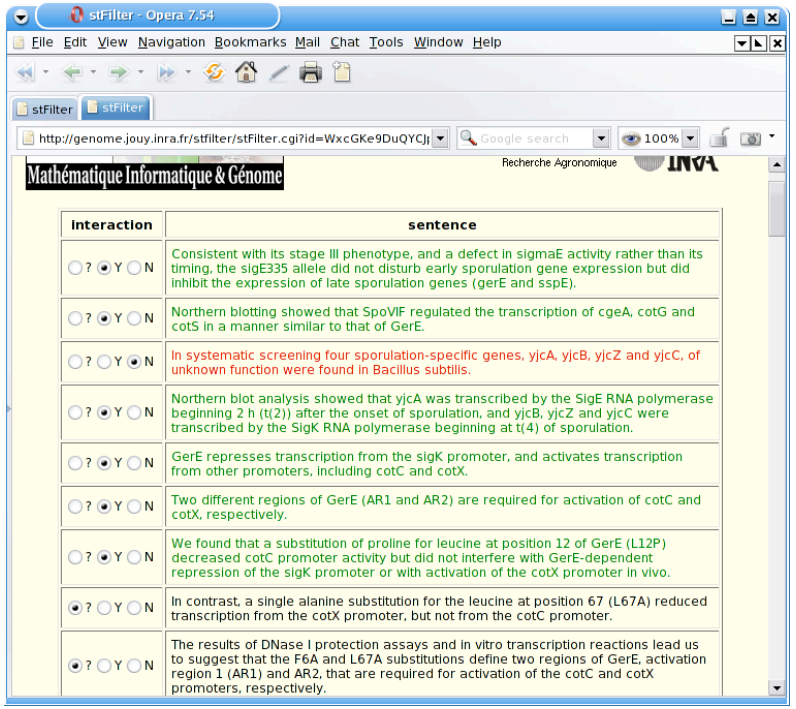
Utiliser un outil spécialisé (tokenizer) en raison des ambiguités des séparateurs (point, tiret).

Filtrage de phrases d'interaction géniques

La cocitation n'indique pas toujours une interaction génique. Le **filtrage** sélectionne les phrases. Si les critères de filtrage sont inconnus, ils peuvent être appris par classification supervisée.

Méthode

1. Classer manuellement les phrases comme pertinentes ou non
2. Produire une représentation appropriée des textes (par ex., *sac de mots lemmatisés*)
3. Entraîner un classifieur
4. Utiliser le classifieur pour filtrer de nouveaux textes



Outil MIG, STFilter [Manine, 2005]

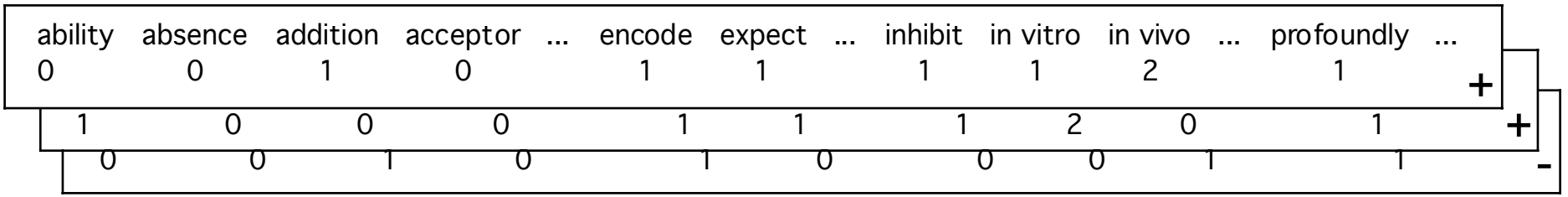
Représentation des phrases

Représentation vectorielle discrète ou booléenne

- Les mots lemmatisés forment le dictionnaire décrivant les exemples.
- Les mots peu discriminants ou trop spécifiques sont supprimés, mots *grammaticaux*, noms de gène
- Si le dictionnaire est trop grand par rapport au nombre d'exemple, les mots sont filtrés par sélection d'attribut (*feature selection*)

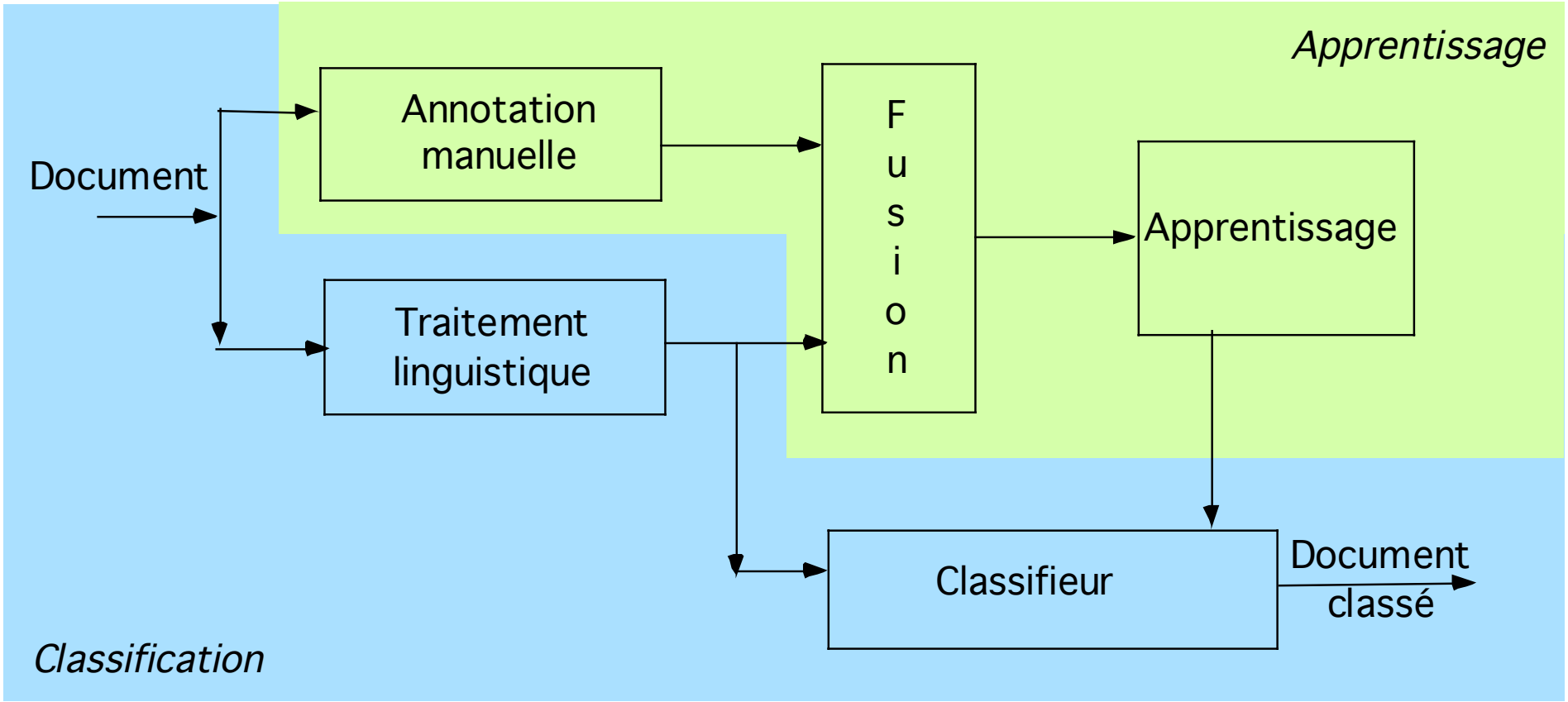
Document : *In addition, GerE stimulates cotD transcription and inhibits cotA transcription in vitro by sigma K RNA polymerase, as expected from in vivo studies, and, unexpectedly, profoundly inhibits in vitro transcription of the gene (sigK) that encodes sigma K.*

Exemple



Règle apprise : présence ou l'absence de mots: *expect*, *inhibit*, *in vitro,* ...dans la phrase à classer

Classification supervisée de phrases d'interaction



Apprentissage de classifieurs pour filtrer les phrases d'interaction

[Craven & Kumlien, 99], [Marcotte *et al.*, 2001], [Nedellec *et al.*, 2001]

Évaluation du filtrage classification, un exemple

Évaluation sur la tâche d'identification des interactions géniques pour 5 corpus, sporulation chez *Bacillus subtilis*, *Lactococcus lactis*, développement chez la drosophile, synthèse des lipides chez le poulet.

Méthode d'apprentissage

Bayésien naïf avec présélection des attributs (par gain d'information)

- Rappel des exemples positifs : 85 %
- Précision des exemples positifs : 74 %

⇒ Bonne précision et bon rappel, mais l'information extraite n'est ni précise ni structurée.

⇒ Le filtrage sert à focaliser les traitements coûteux en temps sur les zones de texte pertinentes.

Limitations de l'approche à base de cocitations pour la détection d'interactions

- **Cas idéal, mais peu fréquent** : la phrase contient un couple de noms unique

  The GerE protein inhibits transcription in vitro of the sigK gene.

- **Cas le plus fréquent :** de nombreux couples dans la même phrase, dont seul un petit nombre indique une interaction génique.

  GerE stimulates cotD transcription and cotA transcription […], and, unexpectedly, inhibits […] transcription of the gene (sigK) […]

  GerE, *cotD*, *cotA* et *sigK* sont cocités,
  mais seuls les couples (*GerE*, *cotD*), *GerE*, *cotA*) et (*GerE*, *sigK*) interagissent.
  Plus de la moitié des phrases avec 2 noms de gène ne mentionnent pas d'interaction.

- La cocitation dénote bien d'autres relations que les *interactions géniques* : homologie, implication dans une même voie métabolique. La sémantique des cocitations est imprécise.

Limitations de l'approche de cocitations, la reconnaissances des noms

Les comptes sont très biaisés par

- **Les homonymes** : des noms de gènes et protéines de plusieurs espèces sont mentionnés dans la même phrase (orthologues). (1/3 des noms de gène de *Bs* appartiennent aussi à *Ec*)

- **Les synonymes.** Les gènes et protéines sont dénotés par différents noms (synonymes)
  - Renommage (KinE / ykrQ)
  - Variantes de nommage qui ne sont pas connues des nomenclatures (sigma K / sigma (K))

- **Les anaphores.** La cocitation au niveau de la phrase ne permet pas de repérer les interactions dont l'antécédent est dans une autre phrase. Pronoms : *it*, fonctions : *the spore coat proteins*, sorte : *the gene*

  this gene *can be* ***transcribed*** *by RNA polymerase associated with either* ***sigmaF*** *or* ***sigmaG***

  La grande majorité des cocitations relevées *au niveau du document* n'indiquent pas des interactions. Le seuillage filtre les cocitations pertinentes et moins fréquentes.

- Pour des comptes corrects, des dictionnaires et des traitements linguistiques sont nécessaires.

Extraction de relation d'interactions géniques

Méthodes

- Reconnaissance des noms de gène et de protéines
- Normalisation des noms : rattachement à l'espèce ou à la souche et assignement de l'identifiant GenBank ou SwissProt
- Traitement des anaphores
- Prédiction de la relation d'interaction génique entre protéines et gènes

Evaluation

- Production de données annotées
- Benchmarks en Extraction d'Information
  - Langue générale (MUC, SemVal)
  - En biologie (LLL, NLPBA, BioCreative, BioNLP)
- Intégration dans des applications, à venir.

Reconnaissance de noms de gènes et de protéines

Entités nommées (EN)

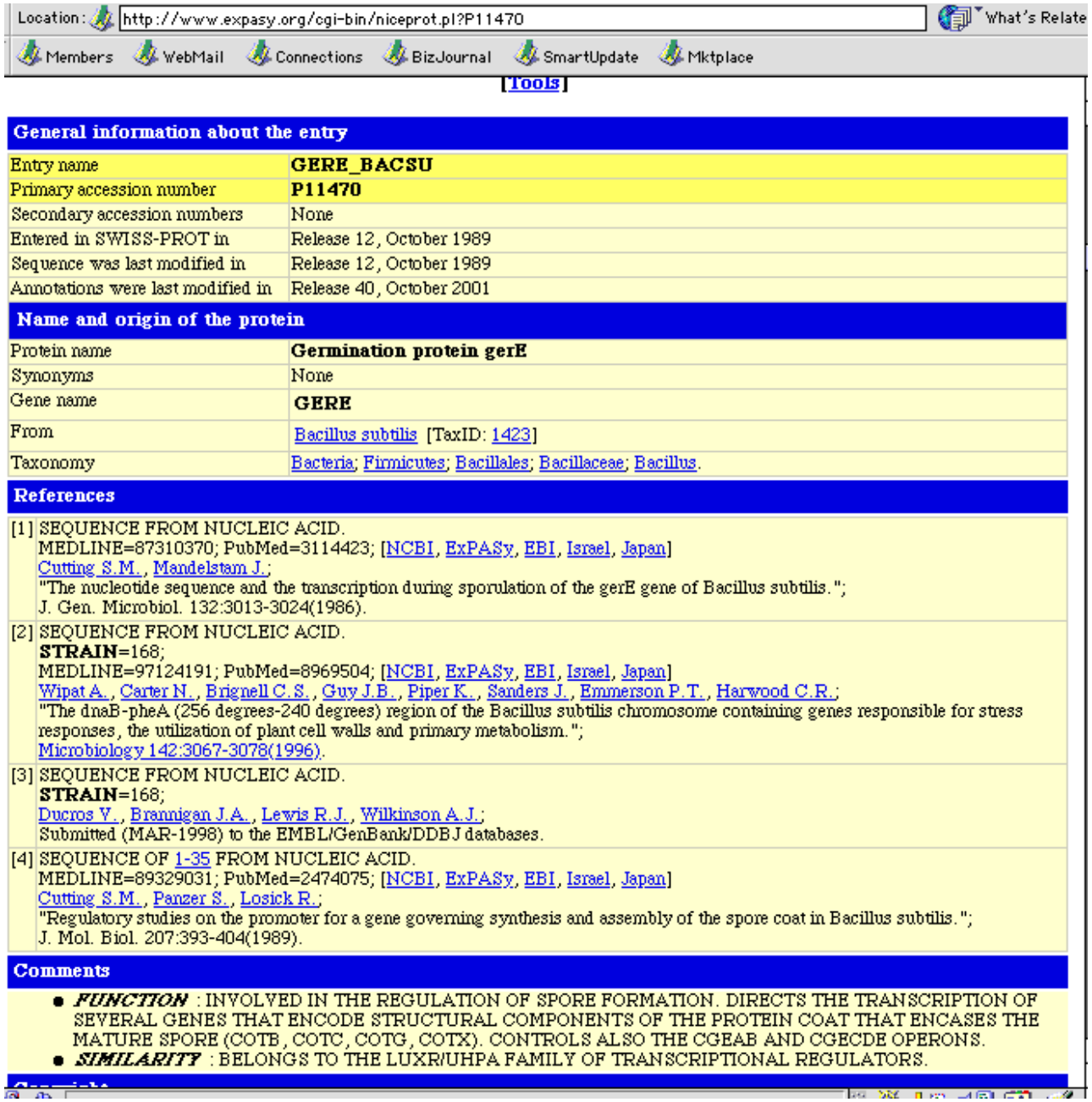
- Les entités nommées désignent des objets particuliers, en général sous la forme de noms propres (ex : noms de gènes et de protéines) ou plus généralement de *formes figées* : *Sigma 32*

- Par opposition aux termes, noms communs de concepts.
  sigma transcription factor, transcription factor

Reconnaissance des EN (REN) : *reconnaître, normaliser et typer* pour indexer

Étape préliminaire et cruciale pour tout traitement documentaire ultérieur.

- Recherche documentaire, extraction d'information, question/réponse, résumé.
- Acquisition d'ontologie
- Interopérabilité entre bases de données
  Ex. : Intégrer les informations des articles de PubMed, GenBank, SwissProt, KEGG et GO


Location: http://www.expasy.org/cgi-bin/niceprot.pl?P11470
What's Relate
Members
WebMail
Connections
BizJournal
SmartUpdate
Mktplace
[Tools]
General information about the entry
Entry name GERE_BACSU
Primary accession number P11470
Secondary accession numbers None
Entered in SWISS-PROT in Release 12, October 1989
Sequence was last modified in Release 12, October 1989
Annotations were last modified in Release 40, October 2001
Name and origin of the protein
Protein name Germination protein gerE
Synonyms None
Gene name GERE
From Bacillus subtilis [TaxID: 1423]
Taxonomy Bacteria; Firmicutes; Bacillales; Bacillaceae; Bacillus.
References
[1] SEQUENCE FROM NUCLEIC ACID.
MEDLINE=87310370; PubMed=3114423; [NCBI, ExPASy, EBI, Israel, Japan]
Cutting S.M., Mandelstam J.;
"The nucleotide sequence and the transcription during sporulation of the gerE gene of Bacillus subtilis.";
J. Gen. Microbiol. 132:3013-3024(1986).
[2] SEQUENCE FROM NUCLEIC ACID.
STRAIN=168;
MEDLINE=97124191; PubMed=8969504; [NCBI, ExPASy, EBI, Israel, Japan]
Wipat A., Carter N., Brignell C.S., Guy J.B., Piper K., Sanders J., Emmerson P.T., Harwood C.R.;
"The dnaB-pheA (256 degrees-240 degrees) region of the Bacillus subtilis chromosome containing genes responsible for stress responses, the utilization of plant cell walls and primary metabolism.";
Microbiology 142:3067-3078(1996).
[3] SEQUENCE FROM NUCLEIC ACID.
STRAIN=168;
Ducros V., Brannigan J.A., Lewis R.J., Wilkinson A.J.;
Submitted (MAR-1998) to the EMBL/GenBank/DDBJ databases.
[4] SEQUENCE OF 1-35 FROM NUCLEIC ACID.
MEDLINE=89329031; PubMed=2474075; [NCBI, ExPASy, EBI, Israel, Japan]
Cutting S.M., Panzer S., Losick R.;
"Regulatory studies on the promoter for a gene governing synthesis and assembly of the spore coat in Bacillus subtilis.";
J. Mol. Biol. 207:393-404(1989).
Comments
FUNCTION : INVOLVED IN THE REGULATION OF SPORE FORMATION. DIRECTS THE TRANSCRIPTION OF SEVERAL GENES THAT ENCODE STRUCTURAL COMPONENTS OF THE PROTEIN COAT THAT ENCASES THE MATURE SPORE (COTB, COTC, COTG, COTX). CONTROLS ALSO THE CGEAB AND CGECDE OPERONS.
SIMILARITY : BELONGS TO THE LUXR/UHPA FAMILY OF TRANSCRIPTIONAL REGULATORS.

Objectifs de la reconnaissance d'entités nommées

- **Typage**

  Associer un *type*, une *catégorie sémantique* aux noms dans les textes

  La liste des catégories est prédéfinie.

  Il existe en général des dictionnaires partiels (*ex. taxonomie d'espèce de GenBank*)

- **Désambiguisation**

  En cas d'homonymie, exploiter le contexte du mot pour lui attribuer la catégorie appropriée.

  Exemples

  l'homonymie entre noms de gène et des mots de la langue courante *map*, *the*, *has*

  CAT est-il *mammalian, enzyme* ou *gene* ?

Objectifs de la reconnaissance d'entités nommées

- **Reconnaissance de nouvelles entités nommées**

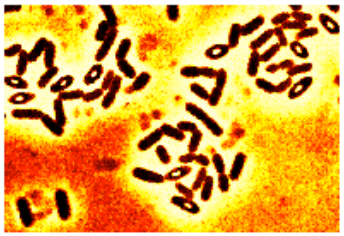
  Aussi complètes que soient les nomenclatures, elles ne peuvent pas être à jour. Il faut les enrichir.
  Exemple : les *nouveaux noms de gène*

- **Reconnaissance de variantes et de synonymes**

  Les acronymes *chloramphenicol acetyltransferase* / *CAT*
  Les abréviations *Bacillus subtilis* / *B. subtilis*
  Les ellipses *EPO mimetic peptide* / *EPO*
  Les variations typographiques *sigma K* / *sigma(K)* / *sigma-K*
  La synonymie due au renommage *SpoIIIG* / *sigma G*.

Automatisation

Automatiser ces traitements à partir de corpus d'apprentissage et de nomenclatures incomplètes et souvent mal structurées.
De nombreuses ressources disponibles en biologie : *GenBank, UniProt, bases de données spécialisées, UMLS*.



spoIIIG sporulation bacillus subtilis Search

Concept all [+]

- gene function
  - sporulation gene
    - forespore gene
      - forespore-spec
- gene concept

Species all [+]

Bacillus subtilis

Genes all [+]

sigG

Authors all [+]

Setlow P

Dates all [+]

2005

Query details: bacillus(lemma) subtilis(lemma) sporulation(Subtilist functional classification/cell envelope and

1-10 among 65 results in 611 categories

Expression of the Bacillus subtilis spoIVB gene is under dual sigma F/sigma G cont

However, during **sporulation**, only sigma G directs significant levels of spoIVB expression.

sigG sigF spoIVB Subtilist functional classification/information pathway/RNA synthesis/initiation of RNA synth organisms/Eukaryota/Fungi-Metazoa group/Metazoa/Eumetazoa/Bilateria/Coelomata/Protostomia/Panarthropoda/Arthropoda/Mandibulata/Pancru subtilis Subtilist functional classification/information pathway/RNA synthesis/initiation of RNA synthesis/sigF S function/enzyme/polymerase/RNA-polymerase Subtilist functional classification/cell envelope and cellular pro Factor Gene Expression Regulation, Bacterial Transcription Factors Base Sequence Molecular Sequence Data

Analysis of the interaction between the transcription factor sigmaG and the anti-s

The activation of sigma(G), a transcription factor, in **Bacillus subtilis** is coupled to the completion o

SpoIIAB SpoIIAA sigG sigF Subtilist functional classification/cell envelope and cellular process/sporulation/Sp Subtilis Molecular Biology Concept/regulator/transcription factor/anti-sigma-factor Subtilis Molecular Biology process/sporulation/SpoIIAA Subtilis Molecular Biology Concept/cell concept/cell-cycle/sporulation J Bacteriol Bacterial 2003 Evans Louise Errington Jeff Feucht Andrea Clarkson Joanna Yudkin Michael D Bacillus subtilis

Transcription of spoIVB is the only role of sigma G that is essential for pro-sigma K

Activation of pro-sigma K processing in the mother cell at late stages of **sporulation** in **Bacillus sub**

Règles de reconnaissance d'entités nommées

État de l'art

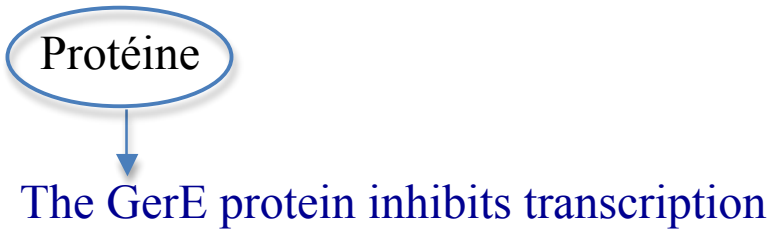
- Exploitation de **patrons morpho-syntaxiques et de dictionnaires**
- De grandes avancées en apprentissage automatique de *patrons* combinant,
  - des descripteurs linguistiques des EN candidates et de leur contexte,
  - des connaissances du domaine,
  - diverses méthodes d'apprentissage.

Les patrons contiennent des conditions

- typographiques, sur la casse, le nombre de lettres, *etc*. (*GerE*),
- sur le voisinage (*GerE* suivi de *protein*),
- sur la catégorie syntaxique (*GerE doit être nom*)

Un exemple de règle

Un *nom*, suivi du mot *protein*, long de *4 lettres*, commençant et finissant par une *majuscule*, est un nom de protéine.



Apprentissage de patrons pour la REN

Conception du corpus d'apprentissage

- Eventuellement, préannoter le corpus d'apprentissage en projetant un dictionnaire (liste de noms de la catégorie).

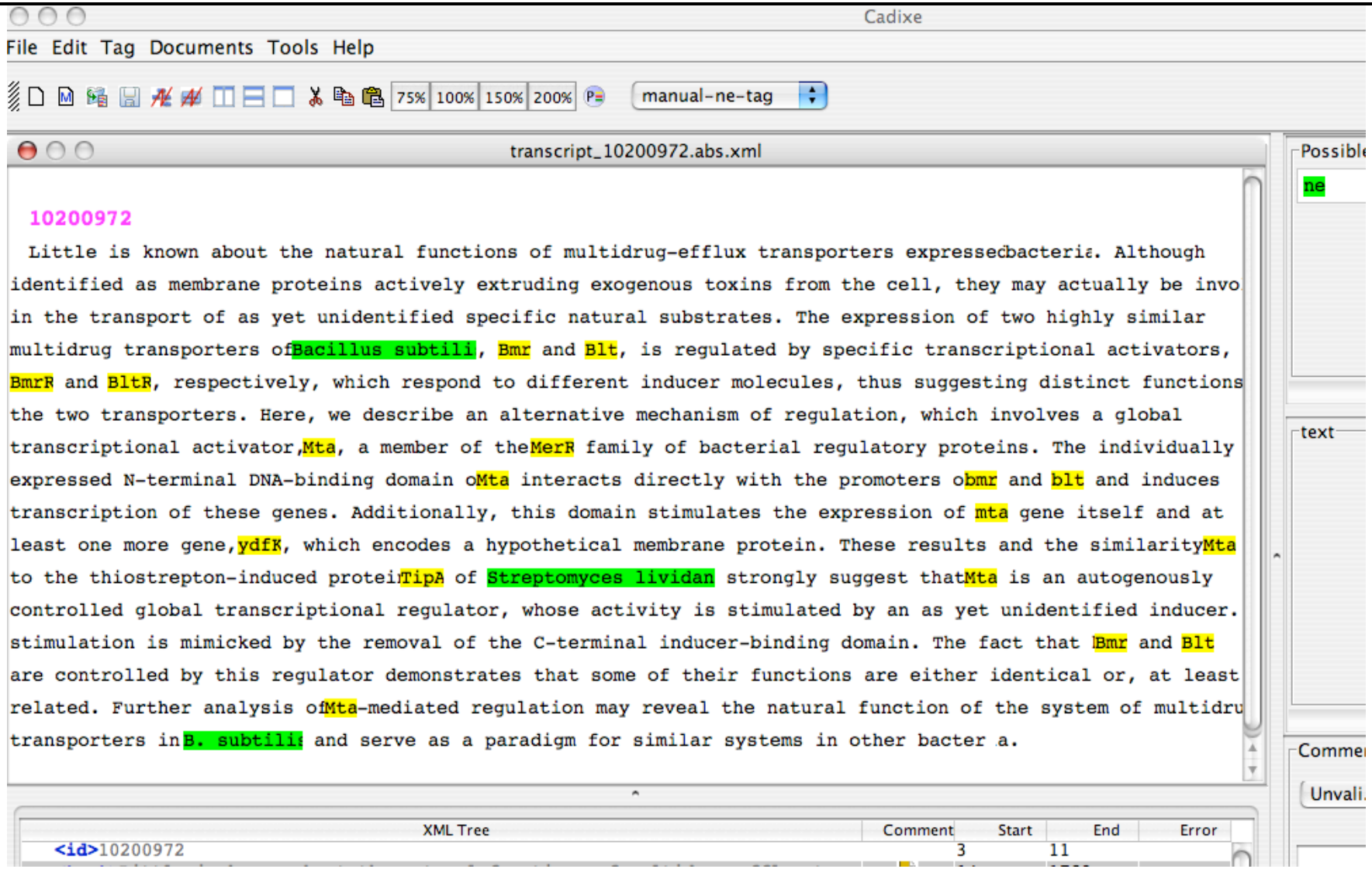
- Les noms appartenant à la catégorie sont étiquetés manuellement. Ils représentent les **exemples positifs** (ex. ***GerE*** *protein inhibits transcription in vitro of the* ***sigK*** *gene*).

- Les **exemples négatifs** sont déduits des exemples positifs : par exemple, des termes de 3 noms au plus qui ne sont pas étiquetés EN (*ex. coat protein, lytic enzymes*)

- Construire une **représentation pertinente** des exemples (*ex. nb lettres, présence de chiffres, mots du voisinage*).

Apprendre automatiquement les attributs discriminants *cad* apparaissant dans les exemples positifs et dans aucun exemple négatif, à l'aide d'un algorithme d'apprentissage (SVM, ME, C4.5).

Annotation d'exemples positifs



Editeur d'annotation Cadixe

Exemple d'expérimentation (projet Quaero)

[Galibert et al., LREC 2010]

Données

A partir de PubMed, sur *Bacillus subtilis* et la transcription.
422 références sélectionnées aléatoirement parmi 22 397 références
6 674 noms de gènes et protéines annotés manuellement, correspondant à 1,647 noms distincts.

4 étapes d'annotation du corpus

1. Projection des noms de gène et de protéines de GenBank *7 049 (1286 distincts)*
   plus les variations typographiques, moins les noms très ambigus
2. Correction par des biologistes (avec l'éditeur XML *Cadixe*)
3. Consensus par un comité *7 185* (+ 361 nouveaux noms)

Evaluation de la stratégie de projection du dictionnaire (baseline)

Précision	Rappel
76,1	78,1

Résultats de l'évaluation

- Expérimentation avec des patrons (Synapse, LIMSI) et des algorithmes d'apprentissage : Induction d'arbre de décision (C4.5) (INRA), CRF (Jouve).

- La méthode basée sur les CRF obtient les meilleurs résultats.

	Précision	Rappel	F-mesure
Inra	93,1	75,3	83,2
Jouve	93,8	77,3	84,7
Limsi	88,6	80,4	84,3
Synapse	69,7	82,6	75,6

- Les résultats obtenus avec le dictionnaire seul montrent que la conception d'un dictionnaire approprié à partir des ressources existantes est productif.

Extraction de *relation* d'interaction génique

Régulation chez *Bacillus subtilis* [Zheng et al., 92]

In addition,
GerE stimulates cotD transcription
[..]

Que doit savoir la machine ?

Que les protéines activent la transcription des gènes
Que les gènes *expriment* des proteins
Les noms des gènes et protéins
Comment se formulent *l'activation de la transcription* et *l'expression des gènes*
…

→ Un problème critique d'analyse de texte, d'acquisition et de modélisation.

Extraction d'information par patron

Un exemple simple [Ono *et al.*, 2001]

Le patron d'extraction d'information

Protein1 .* *interact* <space> *with* <space> Protein2 .*

⇒ Interaction (Type = positif ; Agent = Protein1 ; Cible = Protein2)

appliqué à

Bnr1p interacts with another Rho family member, Rho4p, ...

Interaction	**Type :** positif
	Agent : Bnr1p
	Cible : Rho4p

Les protéines qui *précèdent* un verbe d'interaction sont les agents et les gènes qui *suivent* le verbe sont les cibles.

Extraction à base de patrons

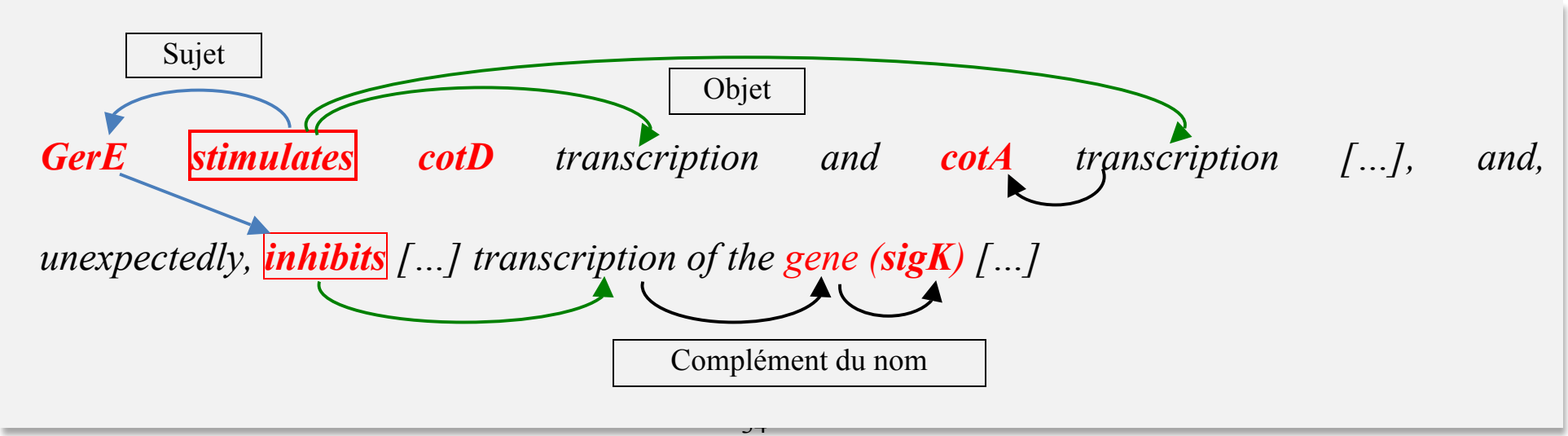
Contre-exemple [Nédellec, 2002] **:**

GerE ***stimulates*** *cotD transcription and* ***cotA*** *transcription […], and, unexpectedly,* ***inhibits*** *[…] transcription of the gene* ***(sigK)*** *[…]*

3 couples sur 6 n'interagissent pas !

⇒ Pour une meilleure qualité de l'extraction, des *traitements linguistiques* sont nécessaires, en particulier les dépendances syntaxiques.

Elles donnent des informations sur le rôle sémantique des arguments de la relation.



Patron d'extraction *avec* dépendances syntaxiques

GerE stimulates cotD *transcription […]*

La règle s'applique à (*GerE*, *cotD*)

Interaction_positive (X, Z):-

is-a(X,protein), sujet(X,Y), verbe(Y), is-a(Y,pos_interaction),
Obj(Z,Y), is-a(Z,gene-transcription).

Interprétation

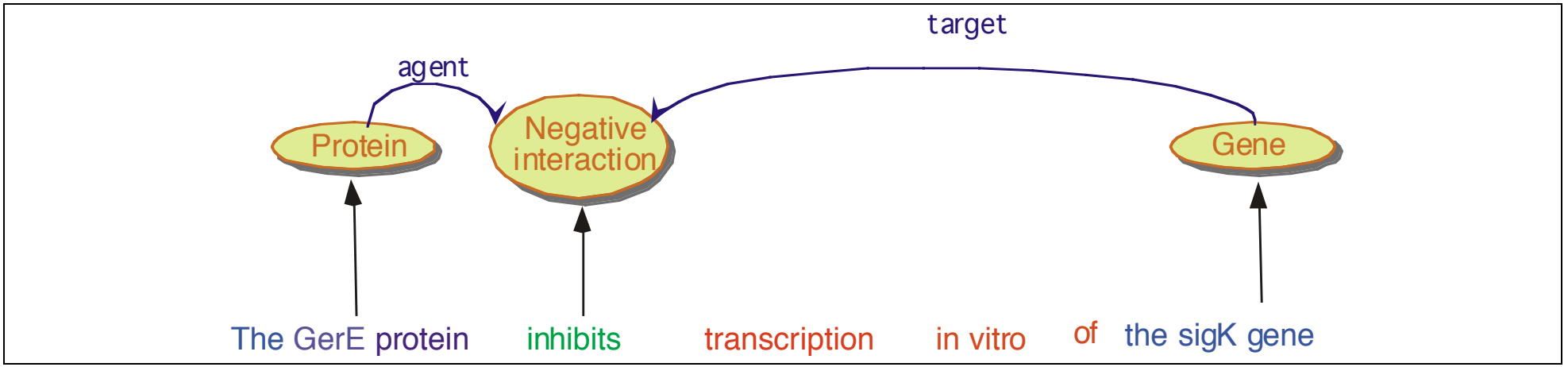
Si le sujet X d'un verbe d'interaction positive, Y est un nom de protéine et que le complément d'objet Z est une transcription de gène,

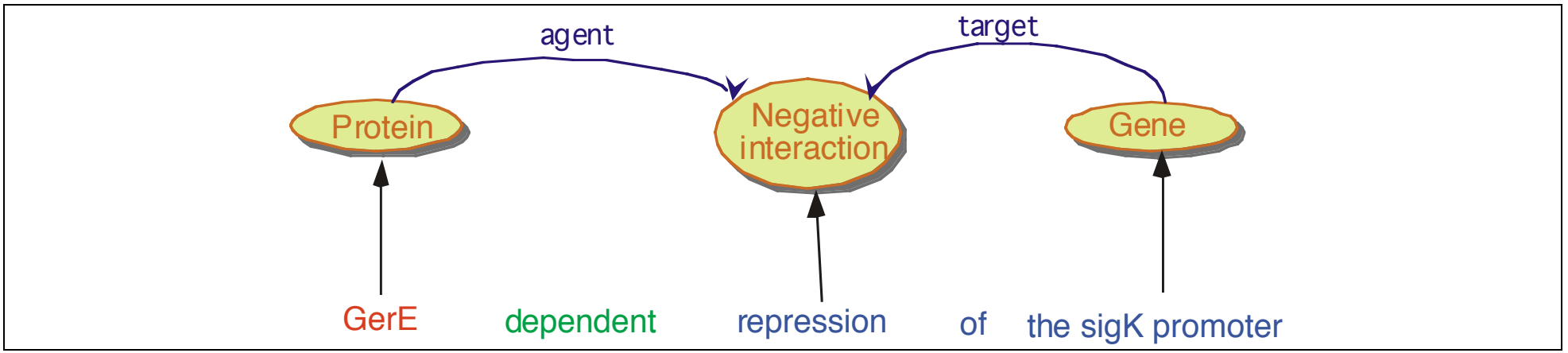
Alors, X est l'agent et Z est la cible de l'interaction génique.

Bien meilleure précision. Mais la couverture de cette règle faible : seulement 0 à 30 % des interactions sont formulées avec un verbe d'interaction dans nos corpus.

⇒ Même dans les cas simples, une analyse syntaxique et sémantique est nécessaire.

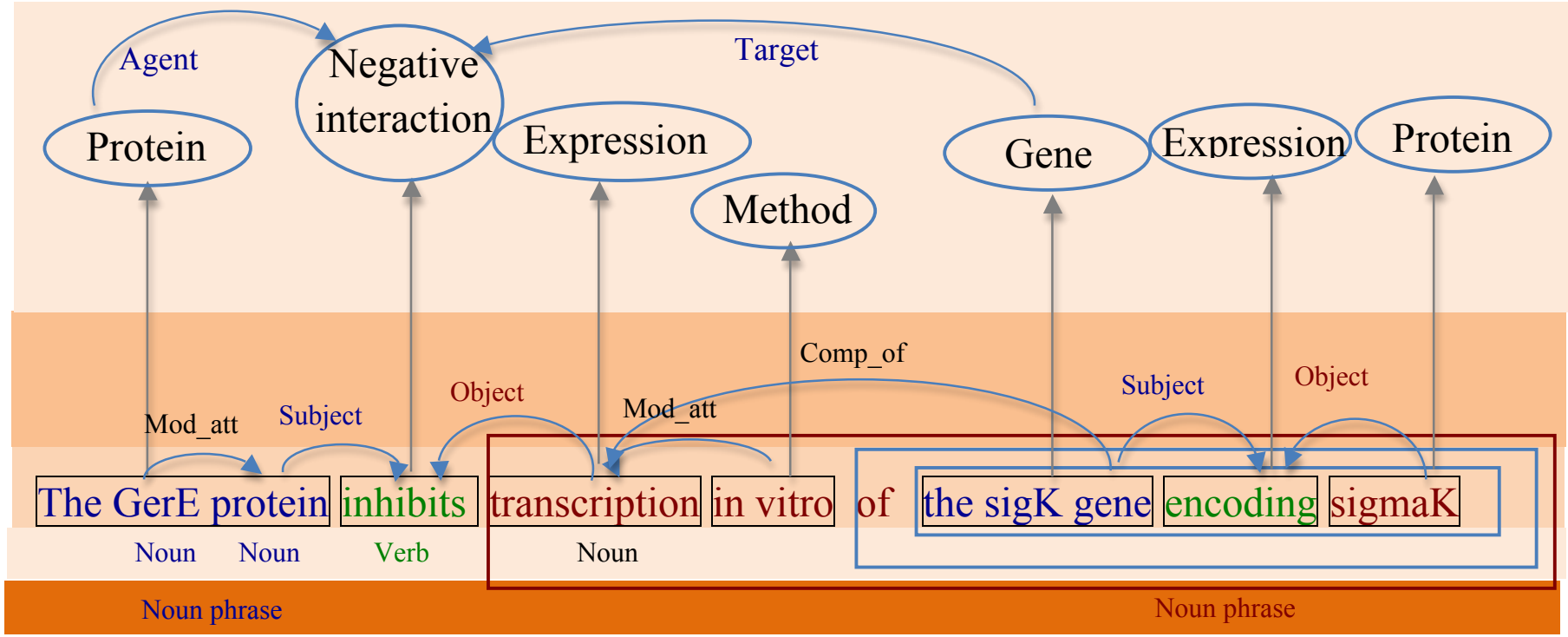
Une même représentation sémantique pour des formulations différentes.





La compréhension automatique, vieux rêve du traitement automatique de la langue.

Normaliser le texte pour rendre l'apprentissage plus simple



Catégorie sémantique	is_a(Ger_protein, protein), is_a(inibit, negative_interaction), …
Dépendance syntaxique	sujet(Ger_protein, inhibit), obj(transcription, inhibit), …
Terme	terme(GerE protein), terme(in vitro) terme(sigK gene)
Catégorie syntaxique	cat(the, det), cat(Ger_protein, term), cat(inhibit, verb), …
Entité nommée	entité(GerE), entité(sigK), entité(sigma K)
Texte segmenté	mot(the), mot(Ger_protein), mot(inhibit), mot(transcription), …

Apprentissage pour l' extraction d'interaction génique

L'apprentissage pour l'extraction d'information relationnelle est vu comme un problème de **classification des arguments**

- étant donnés tous les couples de candidats,
- et les informations linguistiques de leur contexte
- dire s'ils sont ou non en relation (relation orientée).

- **Analyse linguistique**
  - Segmentation, lemmatisation, étiquetage morpho-syntaxique
  - Reconnaissance des arguments de la relation (gènes et protéines)
  - Calcul des dépendances syntaxiques
  - Résolution des anaphores

- **Apprentissage automatique**
  - Calcul de la représentation des exemples
  - Ajustement des paramètres en validation croisée
  - Apprentissage
  - Evaluation sur les données de test.

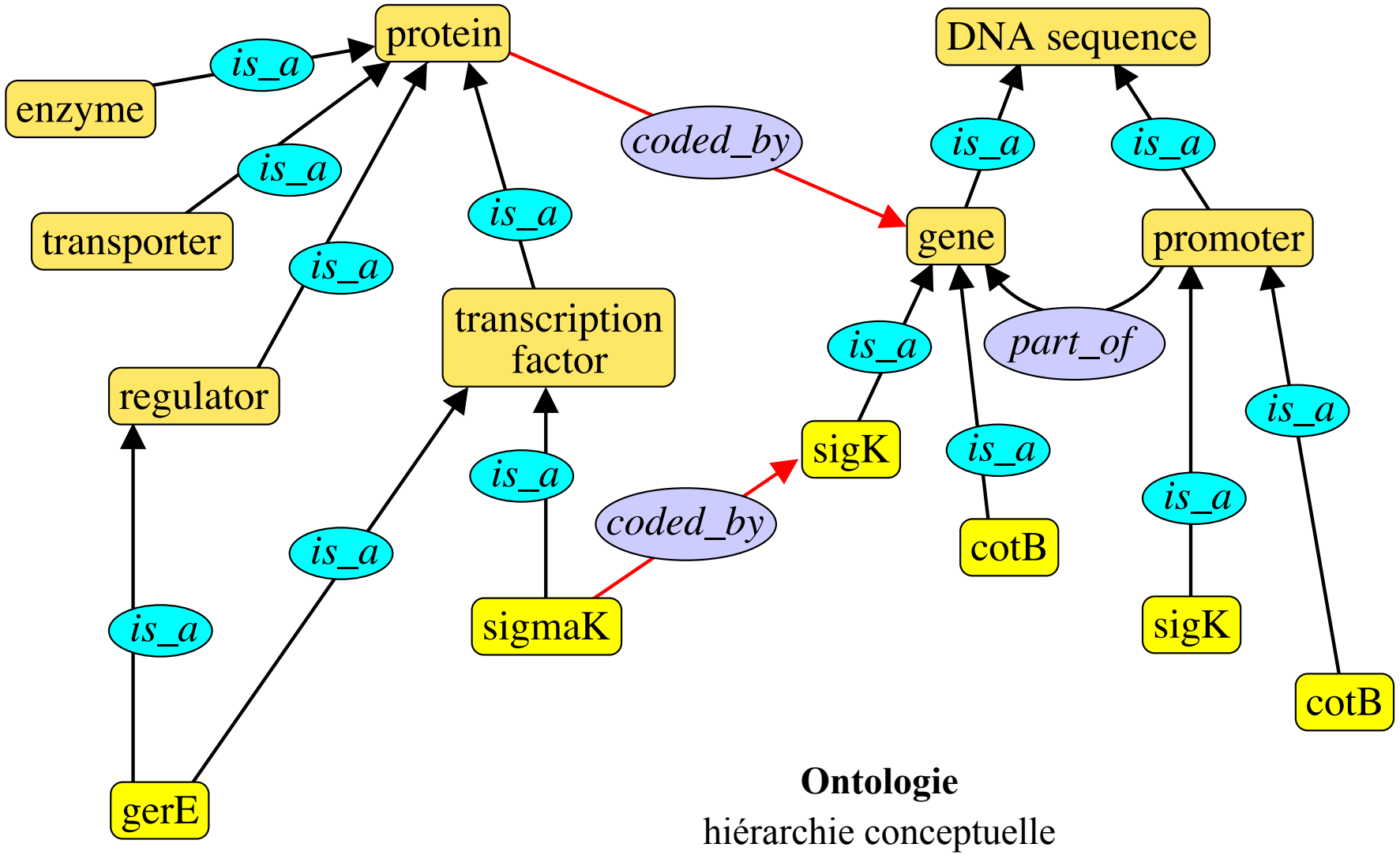
Approches : programmation logique inductive, méthodes à noyau, SVM, plus proches voisins,

Programmation logique inductive pour l'IE

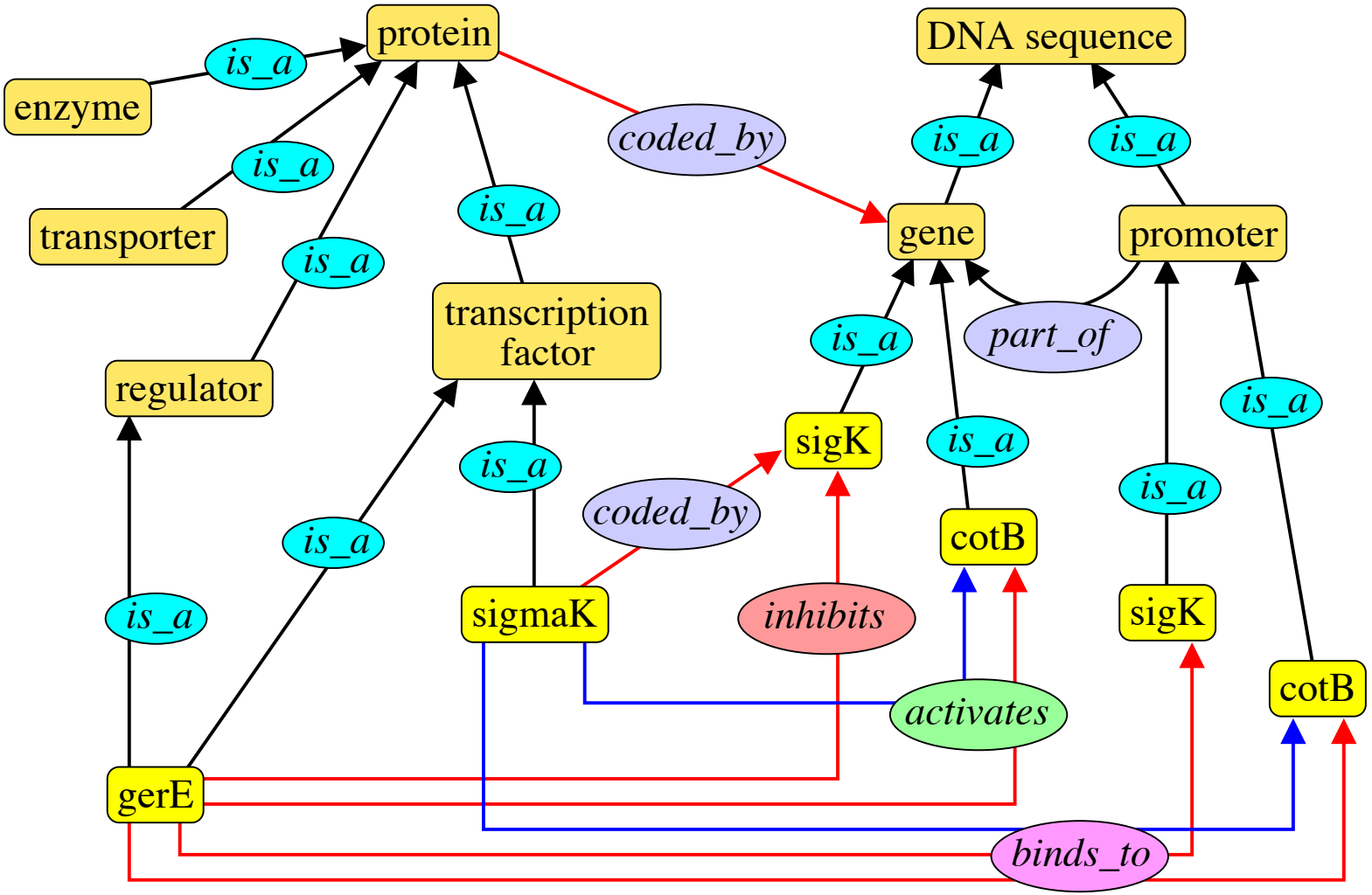
[Manine et al, ICTAI 2009]

Avantages de la programmation logique inductive

- Exploitation de la connaissance du domaine sous la forme de règles d'inférences (clauses de Horn)
- Traitement des données dans une représentation relationnelle
- Apprentissage de programme (plusieurs règles interdépendantes à la fois)



Ontologie
hiérarchie conceptuelle
pour la normalisation du texte



Ontologie

hiérarchie conceptuelle

avec les relations extraites du texte

Ontologie des interactions

Structure de l'ontologie

- 30 objets biologiques
  - gène, ARN, protéine, familles, complexes,
  - opéron, régulon, promoteur de transcription, site, etc.
- 10 relations spécifiques
- Règles d'inférence pour le raisonnement automatique

Transcription des gènes et ses régulations

L'ontologie définit :

- la structure d'un gène
- sa transcription
- les régulations la contrôlant
- modèle implicitement partagé par la description des publications

Transcription des gènes et ses régulations

Relation	***Exemples de texte***
event	expression of *yvyD*
interaction bind_to regulon_dependency regulon_member	**KinC** was responsible for Spo0A ˜P *production* **GerE** binds near the sigK *transcriptional start site* *sigmaB* **regulon** *yvyD* is a member of sigmaB **regulon**
transcription_from transcript_by promoter_dependent promoter_of site_of	**transcription** from the Spo0A-dependent *promoter* **transcription** by final *sigma(A)-RNA polymerase* *sigmaA* recognizes **promoter elements** the *araE* **promoter** *-35 sequence* of the **promoter**

Inférence de relations entre de nouveaux arguments

SpoIIID binds strongly to two sites in the cotC promoter region.

1. Analyse syntaxique et sémantique

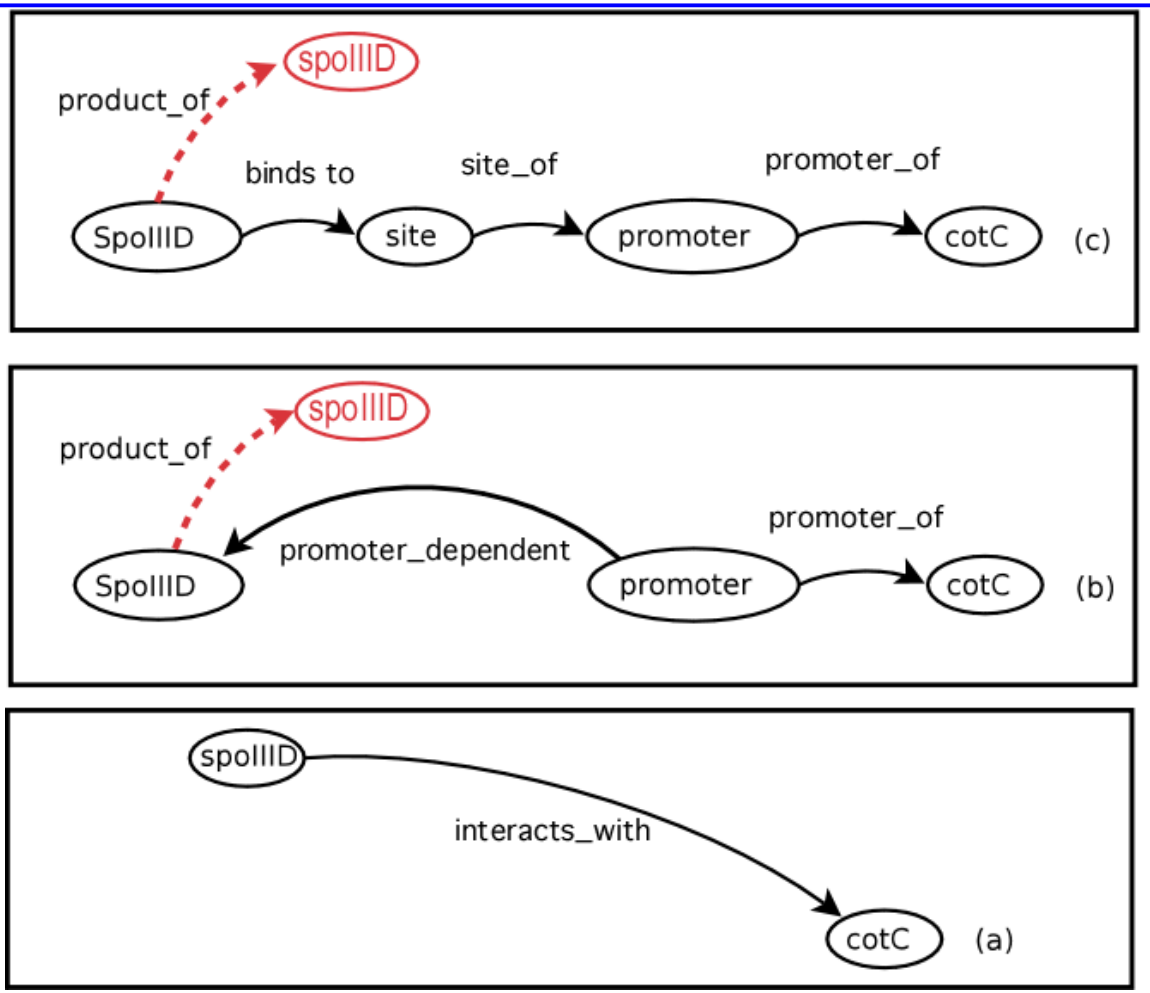
2. Application de règles d'inférence de l'ontologie

Dérivation de la relation promoter-dependent

promoter_dependent(W, X) ←
protein(X), binds_to(X,Y),
site(Y), site_of(Y,W), promoter(W),
promoter_of(W, Z), gene(Z)

Dérivation de la relation d'interaction génique

interact_with(X, Z) ←
protein(X), promoter(W),
promoter_dependent(W, X),
promoter_of(W, Z), gene(Z)



.

Résultats de LP-Propal

Les relations sont différemment reconnues. La précision est meilleure que le rappel.

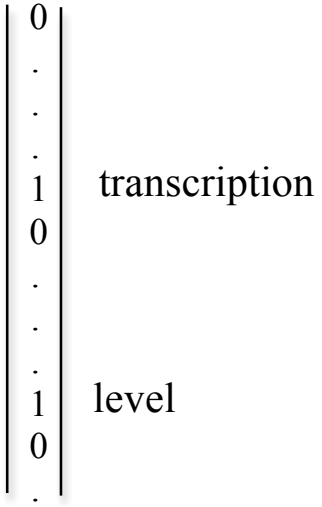
Relation	Rappel (%)	Précision (%)	Qté
interaction	76,4	73,5	161
bind_to	75	**90**	14
regulon_depend	**95**	**100**	12
regulon_member	**90**	**90**	17
transcript_from	85	**96,7**	18
transcript_by	65,5	82,6	44
promoter_depend	**91,5**	**94,3**	47
promoter_of	87,5	85,2	39
site_of	61,7	80,7	21

(Non comparables à BioNLP 2011 GI shared Task)

SVM, string kernel

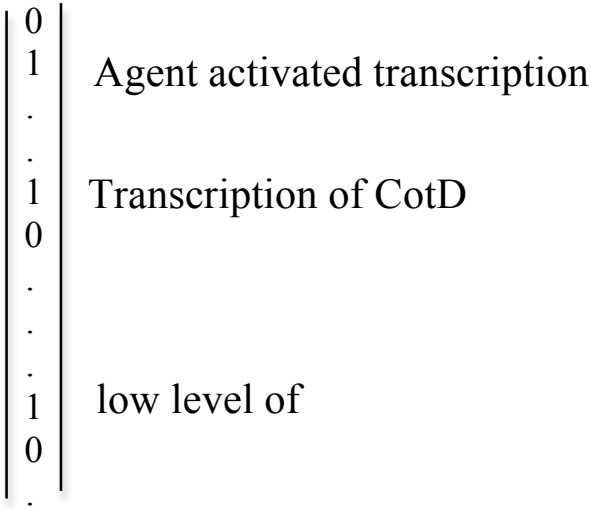
Représentation en sac de mots

A low level of AGENT activated transcription of CotD by TARGET in vitro



Généralisation à des sous-séquences de longueur N (n-grams)

A low level of AGENT activated transcription of CotD by TARGET in vitro



Evaluation de SPSK sur les données LLL

[Veber, Quaero report 2010]

Données LLL : phrases annotées de résumés PubMed sur la transcription chez *Bacillus subtilis*
[Nedellec, ICML whsp 2005]

Données d'apprentissage :

55 phrases
578 exemples, 103 positifs

Cocitation baseline

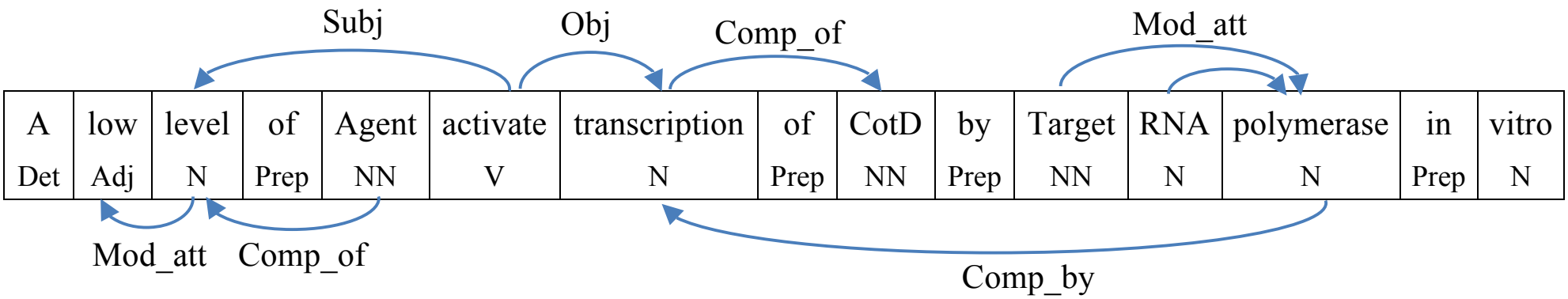
R	P	F
100	18	30,3

String kernel

R	P	F
57,7 (±3,4)	47,8 (±4)	52,2 (±3,1)

Scores sur l'ensemble de test. Comparables aux scores obtenus par les meilleurs systèmes participant à la compétition LLL05

Des séquences aux dépendances syntaxiques

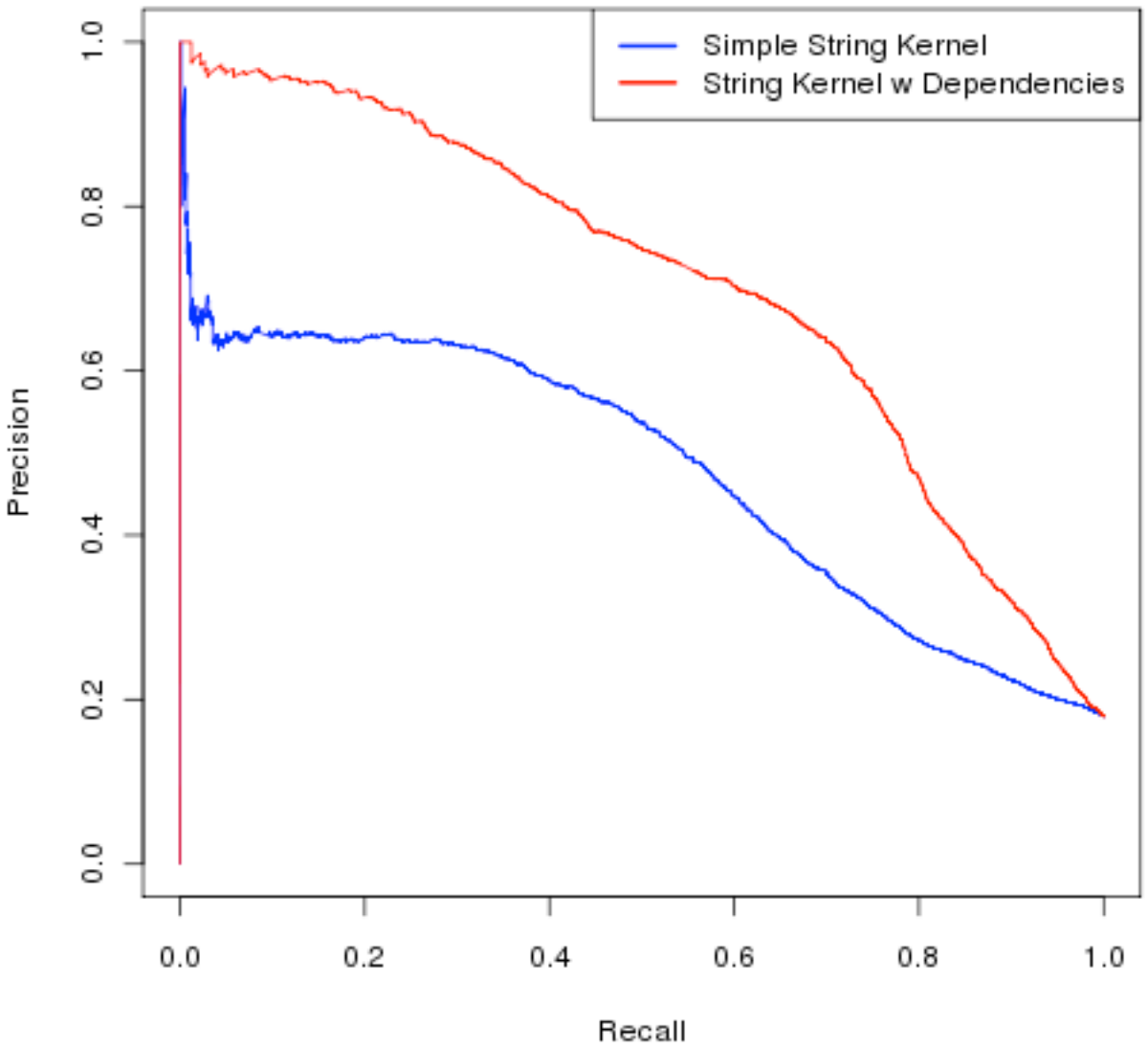


Transformation en séquence : chemin syntaxique le plus court entre l'agent et la cible

Agent NN	Mod_att ←	level N	Subj ←	activate V	Obj →	transcription N	Comp_of →	Target NN

Intuition : découper la phrase en syntagmes nominaux et verbaux comparables

Ex. Nominalisation : *transcription of X by Y / X is transcribed by Y*



R	P	F
76,9 (±2,8)	55 (±2,1)	64,4 (±1,8)

Normalisation à l'aide de classes sémantiques

[Warnier, Quaero report 2011]

Principe : Généraliser en remplacçant les mots par des *classes sémantiques*
Classes sémantiques = classes de mots, construites manuellement ou par,

Sémantique distributionnelle

Hypothèse harrissienne : les mots qui apparaissent dans les mêmes contextes sont sémantiquement proches

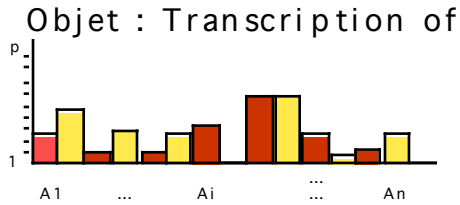
- Compter les occurrences de mots dans les contextes (ici syntaxiques)
- Calculer une mesure de similarité basée sur la comparaison des distributions des contextes
- Former les classes.

synthesis ***Comp:N-N(during) Term***		*transcribe* ***Comp :V_Pass-N(during) Term***	
[synthesis] **[during** growth**].**	**5**	[transcribe] **[during** vegetative growth**].**	**1**
[synthesis] **[during** germination**].**	**3**	[transcribe] **[during** sporulation**].**	**8**

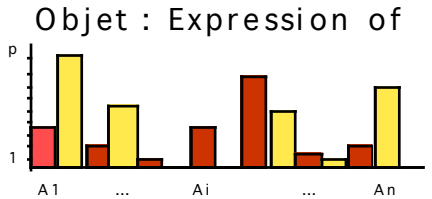
Sémantique distributionnelle et *clustering*

- Matrice de contingence

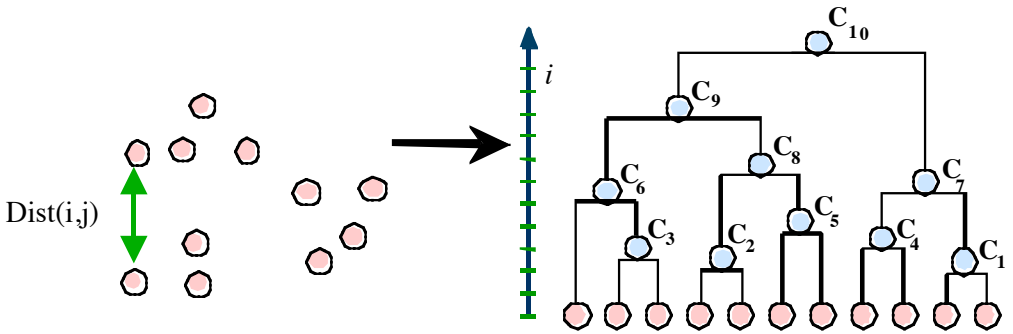
	<attribut 1>	...	SpoIIIG	ComK	...	<attribut n>
Objet 1	$\#occ_{11}$					
Transcription of			3	5		
Objet n						$\#occ_{nn}$

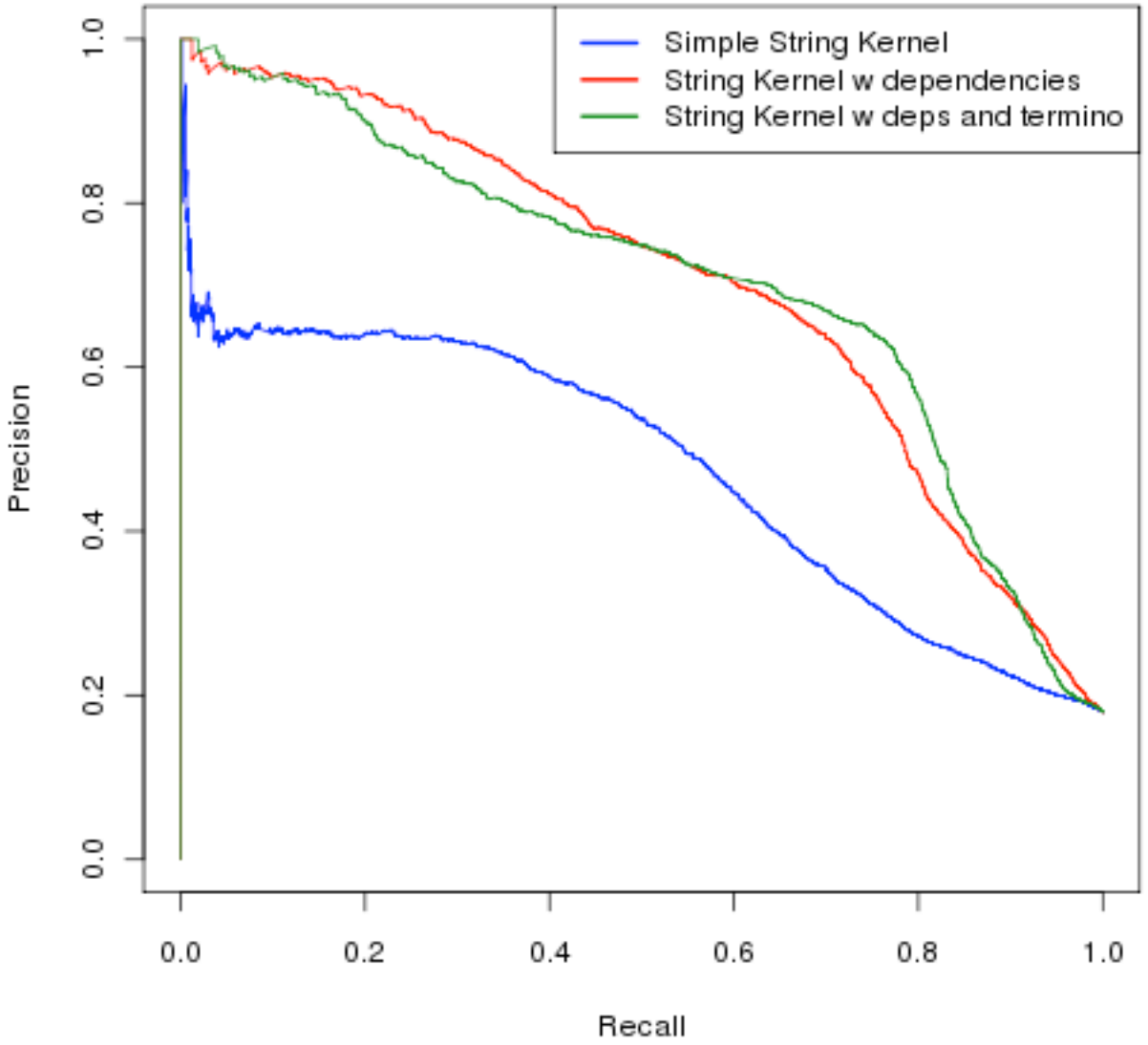


- Le calcul de distance est basé sur la comparaison entre les distributions des contextes semantico-syntaxiques des prédicats.



- Puis les distances sont utilisées par un algorithme de "clustering" hiérarchique ascendant.





R	P	F
79,6 (±2,3)	59,9 (±2,9)	68,4 (±2,3)

Représentation vectorielle alternative

Principe : Utiliser un **alignement global** pour comparer les séquences

A low level of AGENT activated transcription of CotD by TARGET in vitro

AGENT	MOD_ATT	level	SUBJ	activate	OBJ	transcription	COMP_OF	*TARGET*
|			|		|		|	
AGENT	-	-	SUBJ	control	OBJ	expression	MOD_ATT	*TARGET*

AGENT appears to control TARGET expression

Paramètres d'alignement des séquences

Coût de substitution

- Nul si appartenance à la même classe sémantique
- Modéré si même catégorie morpho-syntaxique
- Maximum dans les autres cas

Coût de délétion

- Constant
- Expérimentalement de faibles coûts donnent de meilleurs résultats

Intuition : L'ordre des syntagmes dans la phrase est important et est capturé par l'alignement global. Les ellipses ne changent pas le sens.

Généralisation à d'autres tâche ? Ex. inversion en anglais :

Resistance in PLANT to PLANT DISEASE PATHOGEN
The gene, Rpg1, conferring stable resistance in barley to the wheat stem rust pathogen

DISEASE resistance in PLANT
R gene conferring anthracnose resistance in narrow-leafed lupin

Resistance to PLANT DISEASE PATHOGEN
Ryd4 (Hb) [..] conferring complete and dominant resistance to the barley yellow dwarf virus

Performances de SPSK sur les données LLL

String Kernel		**Analyse syntaxique**		
		Sans	Syntaxe auto	Syntaxe manu.
Cl. sémantiques	sans	52.2 ± 3.1	64.4 ± 1.8	69.0 ± 2.3
	avec	52.4 ± 3.7	68.4 ± 2.3	75.4 ± 2.6

Global Alignment Kernel		Analyse syntaxique		
		Sans	Syntaxe auto	Syntaxe manu.
Cl. sémantiques	sans	-	61.0 ± 4.1	77.0 ± 2.4
	avec	-	59.4 ± 5.4	**79.1 ± 2.8**

Conclusion

Différentes approches de l'extraction d'interactions géniques

Les plus prometteuses : analyse linguistique et apprentissage automatique supervisé
Sont aujourd'hui opérationnelles

Feuille de route pour l'EI pour la biologie

- Meilleure prise en compte des anaphores
- Mieux utiliser la redondance des informations.
- Etendre l'extraction aux conditions environnementales, aux phases du cycle de vie, aux phénotypes, à la localisation.

Utilisation de ces informations pour la construction de réseaux de régulation

- Validation des modéles prédictifs construits à partir des données expérimentales
- Plus généralement, confrontation des données expérimentales aux informations extraites du texte
- Evaluation de l'information extraite en terme de « service rendu ».