Ontology learning for Information Extraction in functional genomics - the Caderige project

Claire Nédellec

MIG (Mathématiques, Informatique et Génome)
INRA (Institut National de Recherche en Agronomie)
France
nedellec@versailles.inra.fr

Functional genomics, a test-case for IE from the semantic web

Most of the IE tasks as defined in MUC do not require deep analysis

⇨ IE patterns can then be hand-coded.

Finding a price in a text comes to point out the dollar sign and look for a sequence of digits (from [Soderland, 98])

```
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
```

But, in many web documents,

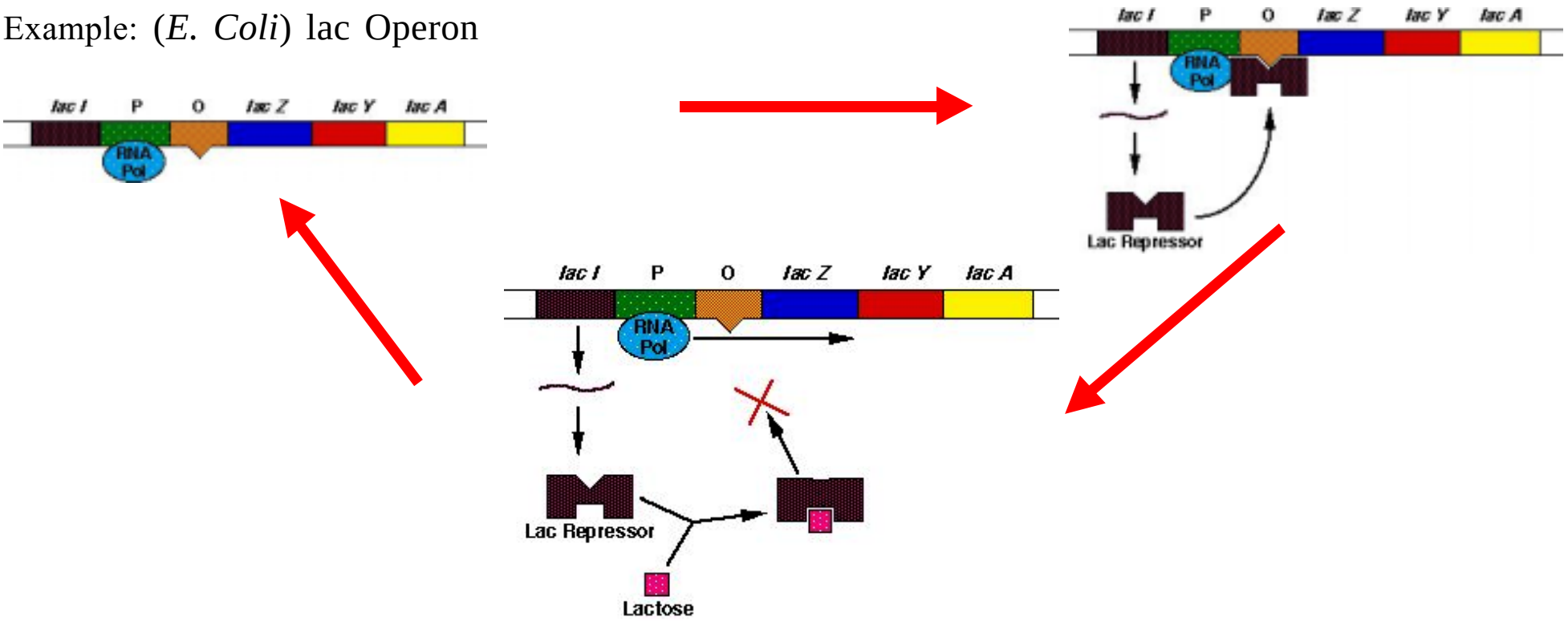
- Learning is needed because of the low coverage of hand-coded patterns
- Deeper understanding linguistic methods are required for identifying the discriminant regularities

It is the case in *functional genomics*

Gene function

- The proteins are molecules that accomplish most of the functions of the living cells (catalysis, defense, acquiring and transforming energy).
- Genes are DNA sequences that express proteins.
- Proteins are also responsible of the gene activation and inhibition

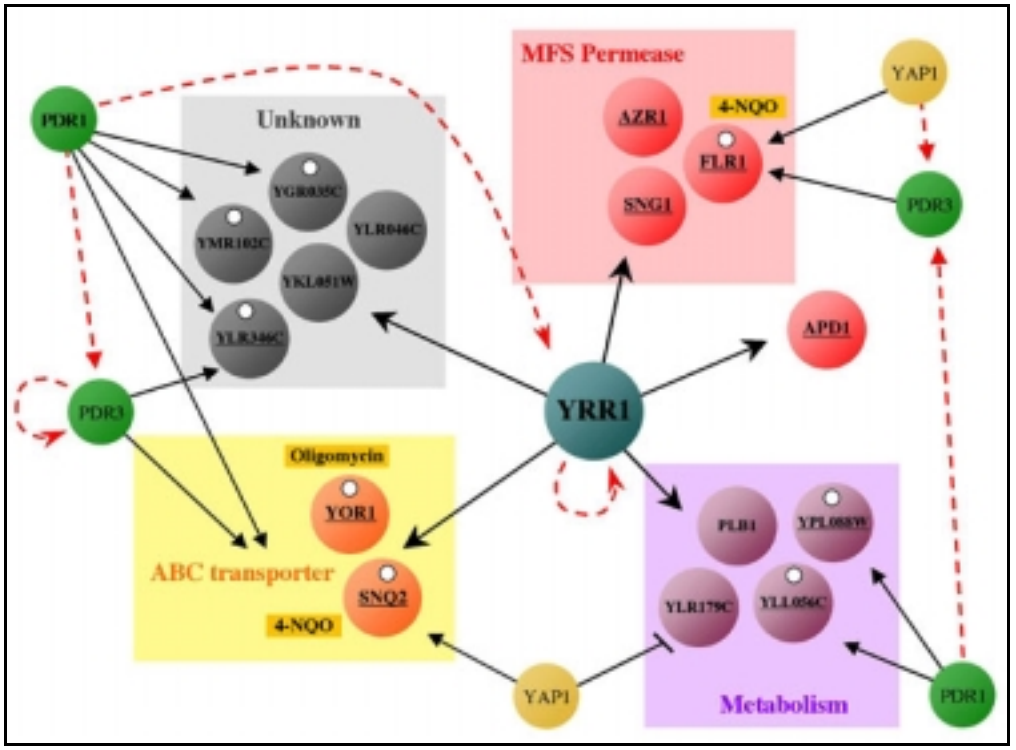
Example: (*E. Coli*) lac Operon



Gene regulation network

• Gene and protein activity may be part of a whole regulation network

Example: Characterization of YRR1 transcription factor regulation system for a better knowledge of the PDR (pleitotropic drug resistance) network in the yeast *Saccharomyces cerevisiae*.



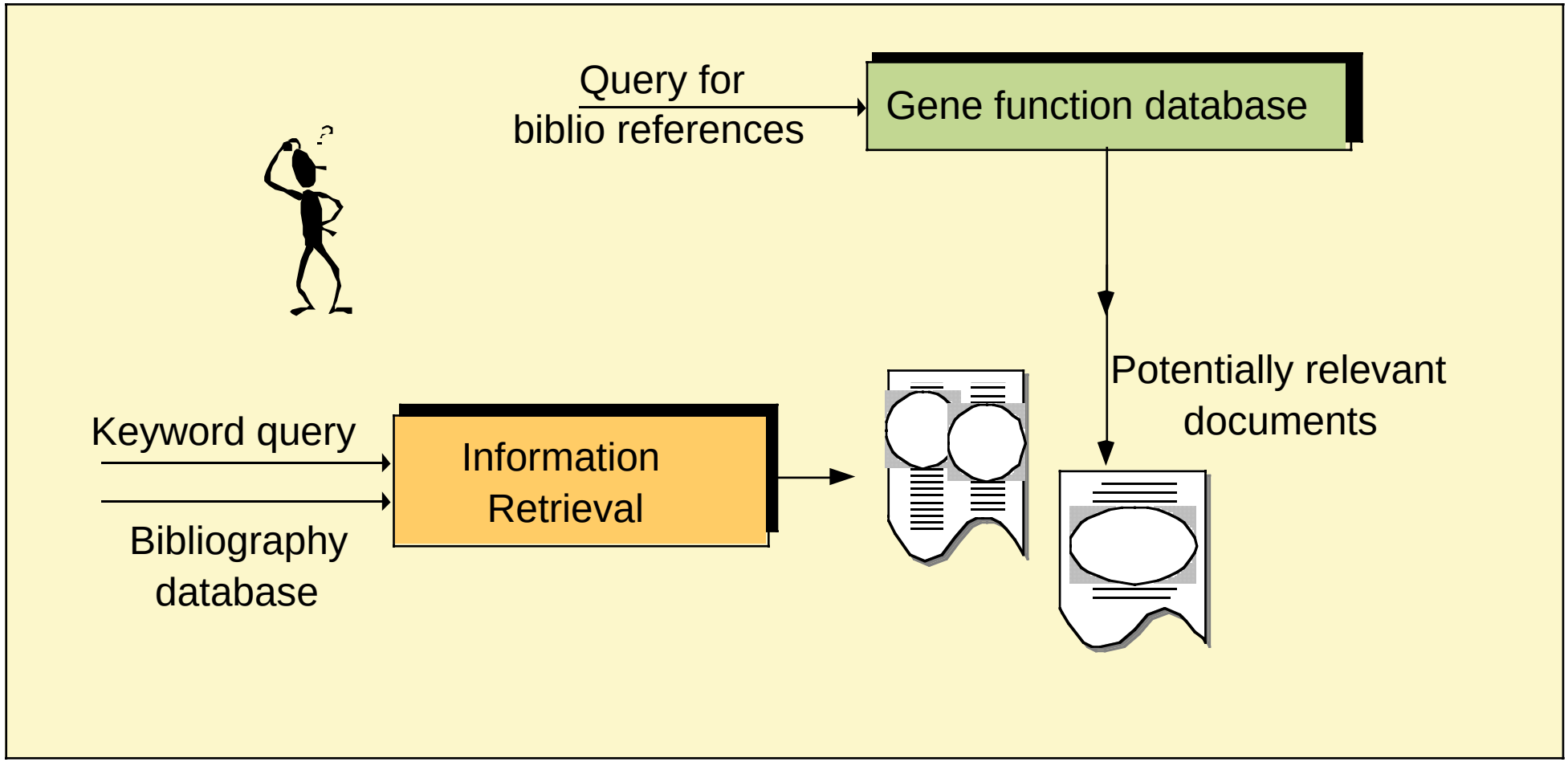
Discovering the gene function

- **Experimental approaches** for few genes at a time (sequencing, functional analysis)
- **Inference by homology** *in silico*: "*my new gene should have the same function as similar genes*" (and what, if many of them?)
- **Large scale study** (regulation network): databases and information extraction in bibliography from the web, of *gene interaction / phenotype / gene reaction to environmental conditions* and experimentation (ADN chips),

Two examples of *web bibliography databases*

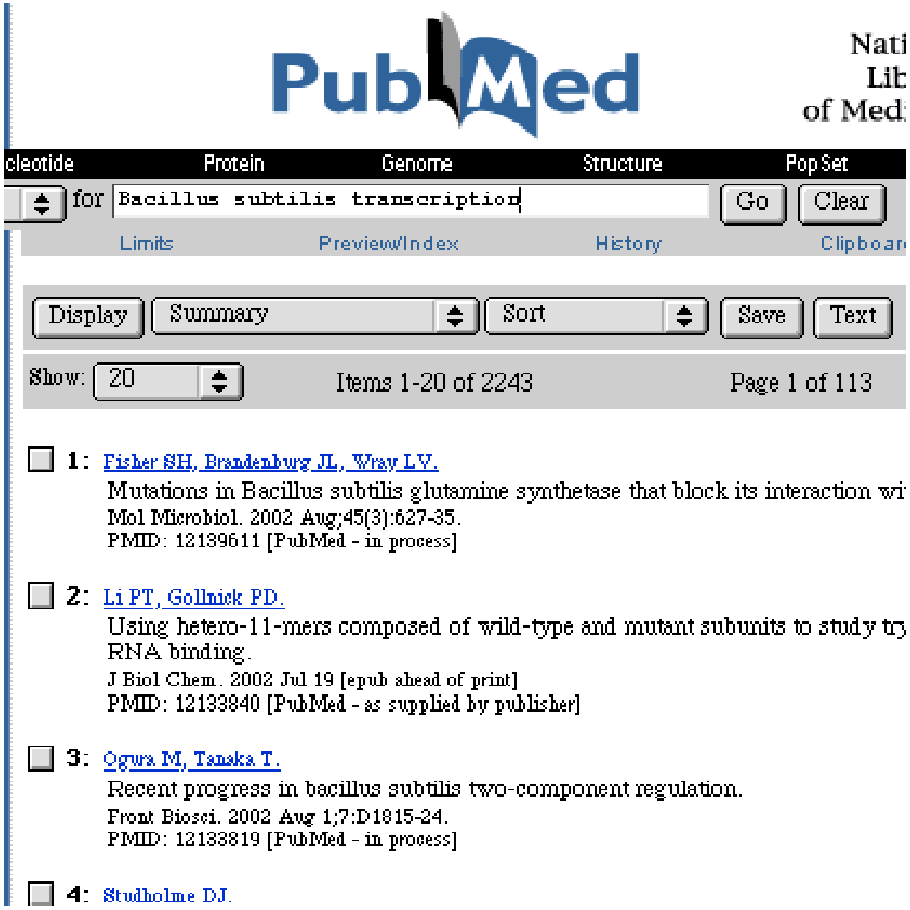
	MedLine	**FlyBase**
DB Size	> 12 millions of refs.	> 9500 genes recorded

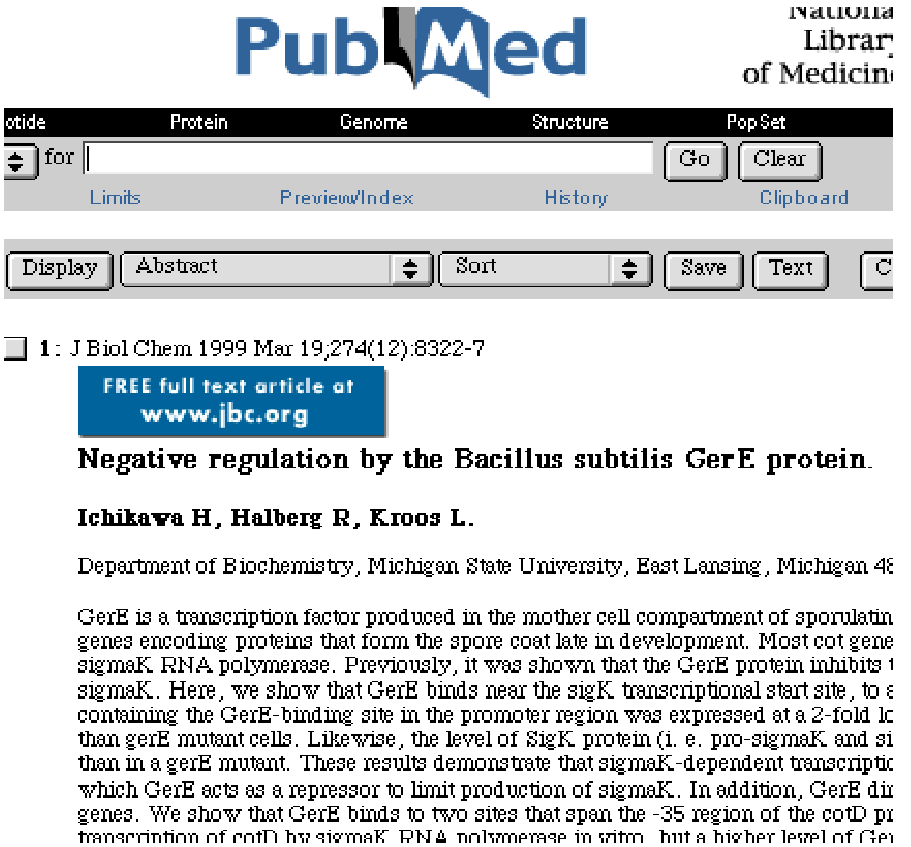
Bibliography access



MedLine content

http://www.ncbi.nlm.nih.gov/entrez/





A MedLine abstract

UI - 99175219
AU - Ichikawa H
AU - Halberg R
AU - Kroos L
TI - Negative regulation by the Bacillus subtilis GerE protein.
..
PT - JOURNAL ARTICLE
..
DA - 19990415
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
MHDA- 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

Discovering the function of the genes: an Information Extraction problem

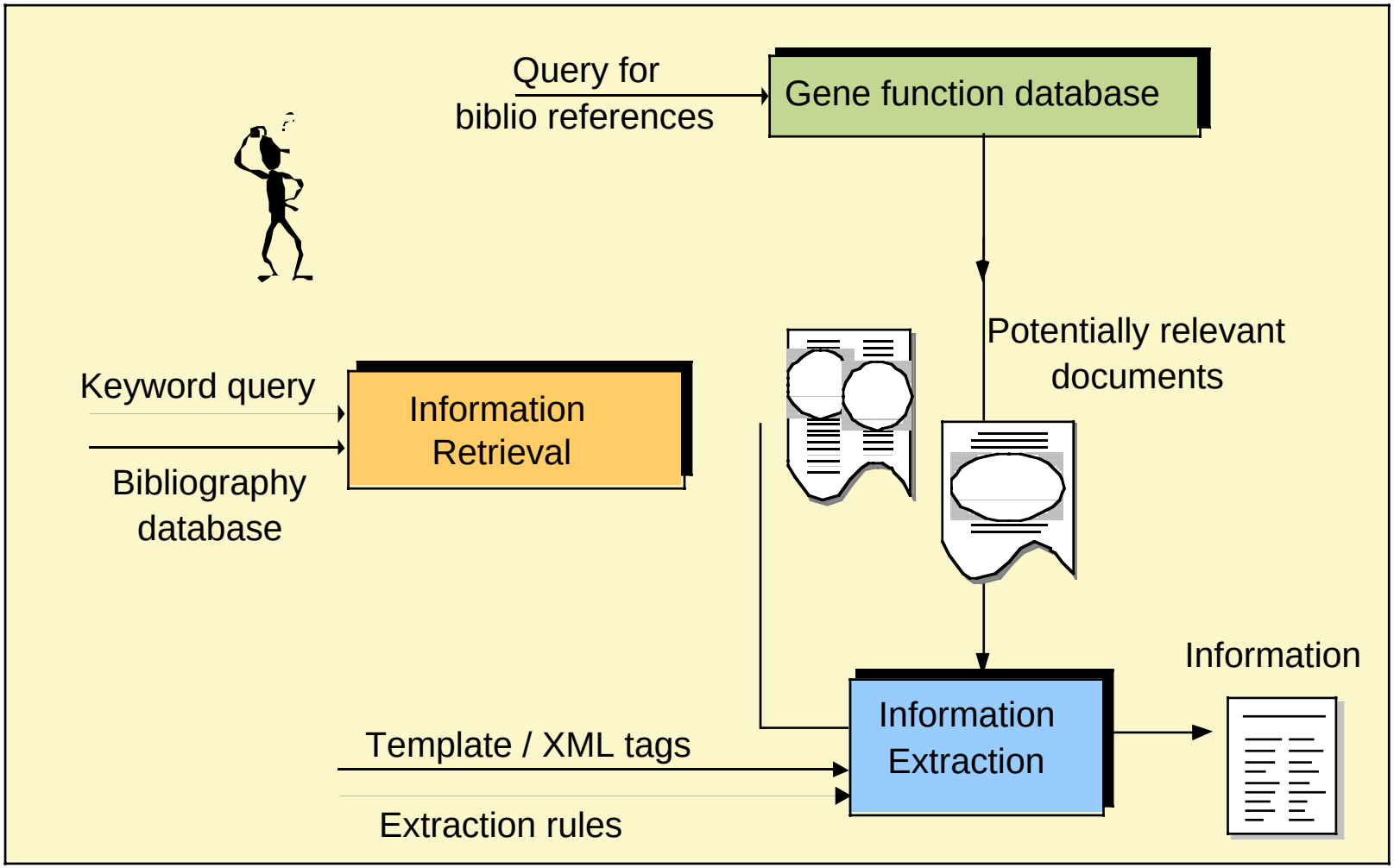
Two examples of the IE problem in functional genomics

- **Inference by homology**: A biologist at MIG annotates 8 genes of *Lactobacillus bulgaricus* per day.
  One third stays unannotated because bibliographic references are too long to process. 2000 genes have to be annotated!

- **Large sacale study:** The query "*Bacillus subtilis transcription*" retrieves 2243 references from MedLine

⇨ Information extraction must be partially automated

Bibliography access



Information to be extracted from a text fragment

Fragment from a Medline abstract

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

Filled form

Interaction	**Type :** negative	
	Agent : GerE protein	
	Target: Expression	**Source :** sigK gene
		Product : sigmaK

Equivalently, XML tagging

The <agent type=protein>GerE protein</agent> <interaction type=negative>inhibits </interaction> <target type=expression>transcription of <source type=gene>the sigK gene</source> encoding <product>sigmaK</product></target>

Information extraction in functional genomics

Three main approaches for automating IE:

1. Hand-coded patterns
2. Learning keyword-based extraction patterns
3. Learning linguistics-based extraction rules

1. Hand-coded IE patterns [Ono *et al.*, 2001]

One simple example from Ono's:

Bnr1p interacts with another Rho family member, Rho4p, but not with Rho1p.

Patterns: Pattern1: Protein1 * interact/stimulate/inhibit <space> with <space> Protein2 *

Pattern2: Pattern1 * but <space> not <space> Protein3

Result:

Interaction	**Type:** positive	**Type :** negative
	Agent: Bnr1p	**Agent:** Bnr1p
	Target: Rho4p	**Target:** Rho1p

1. Hand-coded IE patterns

Counter example from Caderige:

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

Result with pattern1:

Interaction	**Type:** positive	**Type :** negative
	Agent: GerE	**Agent:** GerE, cotD, cotA
	Target: cotD, cotA, sigK	**Target**: sigK

⇨ Additional constraint on the number of words between the elements of the triple: *Distance ≤ 5 words*

⇨ High precision but *low* recall (you can never be exhaustive)

2. Keyword-based rule learning

[Marcotte et al., 2001], [Nedellec *et al.*, 2001]

- **Assumption**

  Sentences with less than two (0, or 1) gene/protein names *do not* mention gene interactions.

- **Learning in 3 phases**

  1. Select the sentences with at least *two gene names*
  2. *Classify* the sentences by hand, as relevant or not
  3. *Learn* a classifier with the learning set (vector representation)

Training examples

Vector representation

Sentences are lemmatized

Stop-words and gene names are suppressed

The words represent the vector attributes, the values are boolean

Training example from *Bacillus subtilis*

Sentence : 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.

Example : addition stimulate transcription inhibit transcription vitro RNA polymerase expected vivo study unexpectedly profoundly inhibit vitro transcription gene encode

Class : Positive

Evaluation of keyword-based rule learning

Example of an evaluation on gene interaction in *Bacillus subtilis*
Four methods with, and without feature selection [Nedellec *et al*, PKDD'01]

Corpus	**Bacillus subtilis corpus**							
Method	BN		IVI		IVI+BN		SVM	
# attributes	all 2340	1800	all 2340	1900	all 2340	1500	all 2340	1800
Recall positive	85,7	90,8	82,6	**91,5**	70,00	90,00	70,21	77,02
Precision positive	66,6	74,1	67,4	78,3	67,14	**85,11**	69,62	78,52
Recall	71,1	77,5	71	82,8	67,60	**84,12**	69,53	75,86
Precision	71,1	79,9	71	83,2	67,60	**87,89**	69,53	78,21

⇨ Good precision and recall but no precise information retrieved (sentence level).

3. Linguistic-based extraction patterns

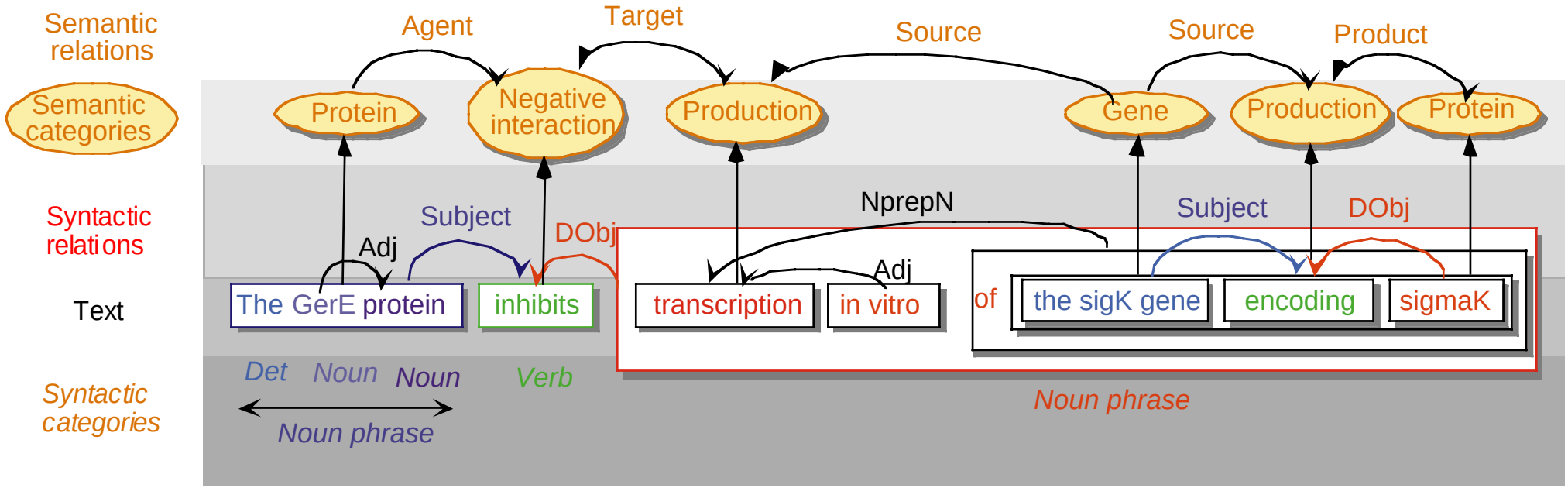
Conclusion on keyword-based learning

For a good precision and a high recall, extraction rules should include conditions on different *text analysis levels*, syntactic and semantic.

⇨ Information extraction *in two phases* [Riloff, Soderland, Freitag, Ciravegna]

1. **Sentence processing**
   Parsing and semantic tagging for an enriched and normalized text representation
2. Application of **extraction patterns** on this representation

Normalized text representation



Semantic relation agent(Ger_protein, inhibit), cible(transcription, inhibit), …

Semantic category concept(Ger_protein,protein),concept(inibit,negative_interaction), …

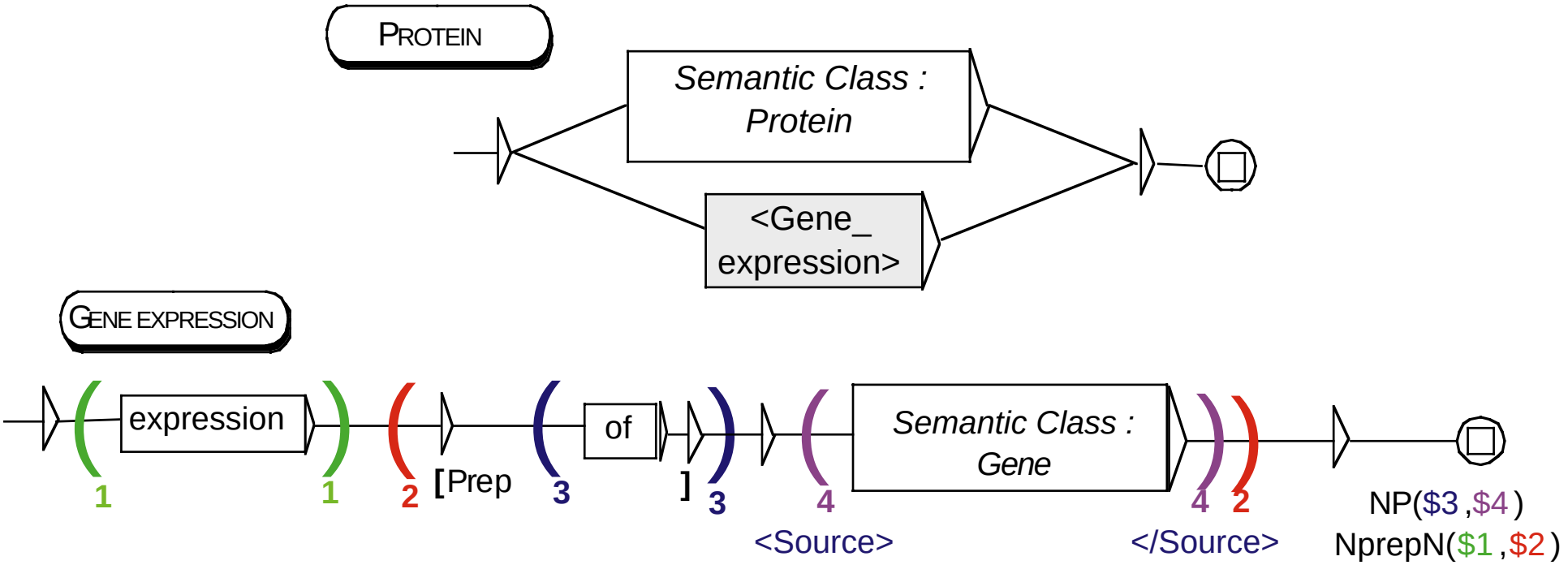
Syntactic relation subject(Ger_protein, inhibit), DObj(transcription, inhibit), …

Text tk(the), tk(Ger_protein), tk(inhibit), tk(transcription), tk(of), …

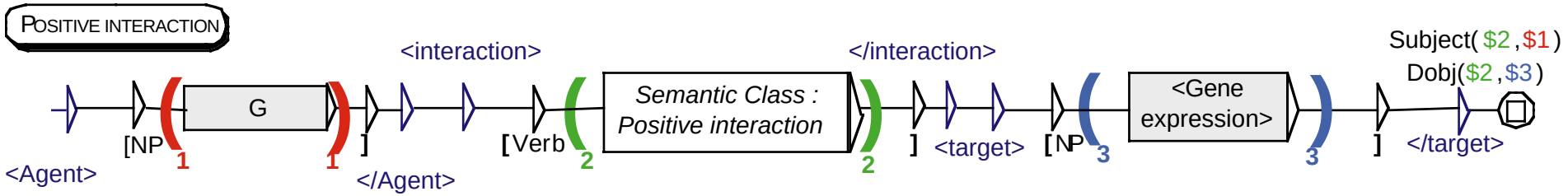
Syntactic category cat(the, det), cat(Ger_protein, term), cat(inhibit, verb), …

Example of EI pattern: a transducer for protein identification and mark up

The automata use the syntactic and semantic information from the parsing phase to identify proteins in the text.



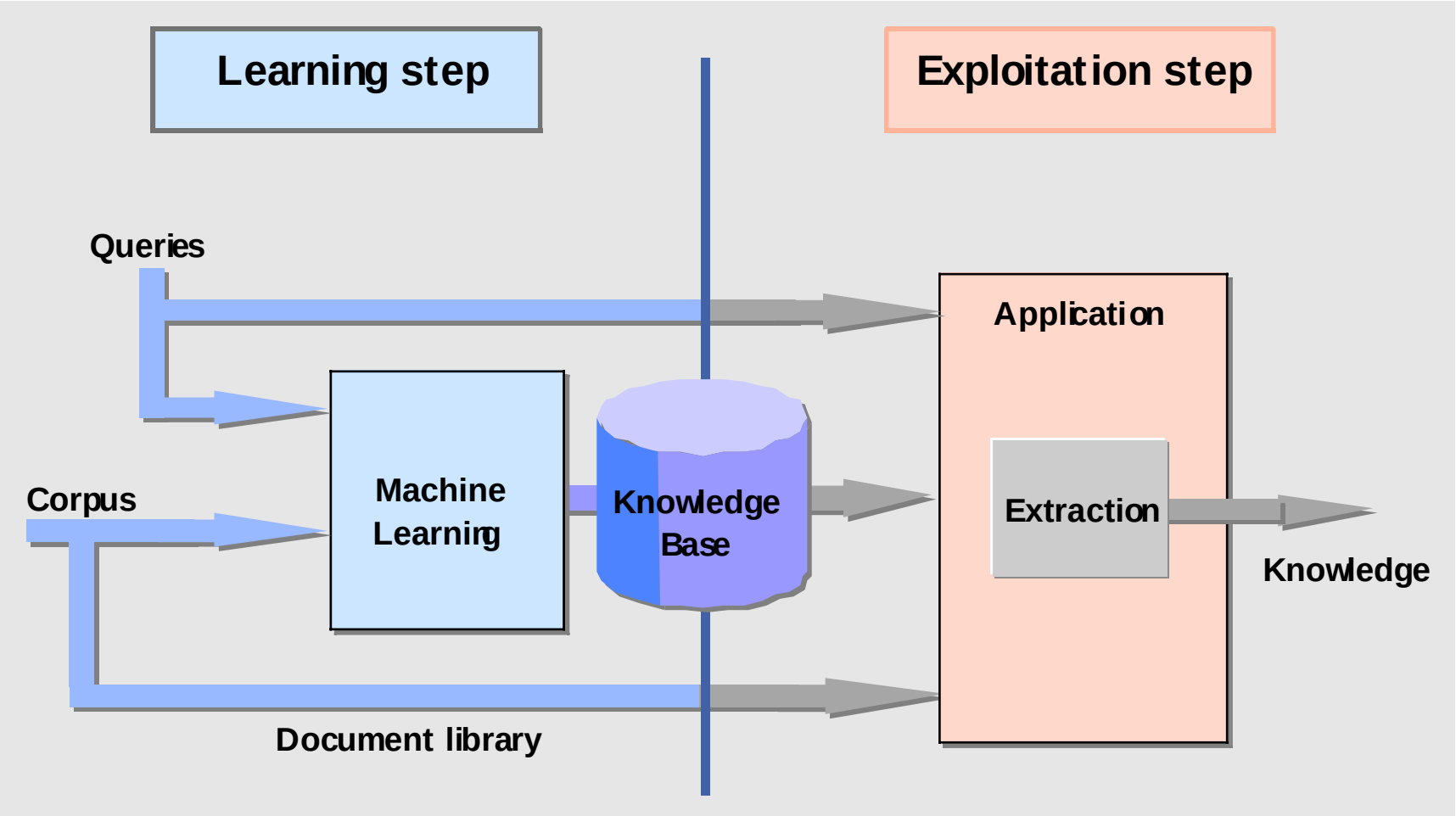
Example of EI pattern: a transducer for gene interaction identification and mark up



What knowledge for extracting information

Types of knowledge needed	How to get it
• Syntactic • Syntactic categories (parts of speech) • Syntactic relations (dependencies)	**Tools** • Morphosyntactic taggers • Syntactic parsers
Semantic • Semantic categories (conceptual hierarchies), restrictions of selection • Subcategorization frames • Predicate schemata	**Background knowledge learned from corpus** • Distributional semantics (clustering) • Grammatical inference, ILP • FOL clustering
Extraction rules	• Relational learning, ILP

Knowledge learning and information extraction



Learning information extraction patterns

Types of knowledge needed	How to get it
• Syntactic • Syntactic categories (parts of speech) • Syntactic relations (dependencies)	**Tools** • Morphosyntactic taggers • Syntactic parsers
Semantic • Semantic categories (conceptual hierarchies), restrictions of selection • Subcategorization frames • Predicate schemata	**Background knowledge learned from corpus** • Distributional semantics (clustering) • Grammatical inference, ILP • FOL clustering
Extraction rules	• Relational learning, ILP

Learning information extraction patterns

Active domain research from the beginning of the nineties (MUC)

- **Learning extraction rules from free and semi-structured texts**
  - AutoSlog [Riloff, 93-99]
  - LIEP [Huffmann, 96]
  - SRV [Freitag, 98]
  - Crystal [Soderland, 95], Whisk [Soderland, 99]
  - WAWE [Aseltine, 99]
  - Pinocchio, LP2 [Ciravegna, 00-02]
  - ILP RHB+ [Sasaki & Matsuo, 00]
- **Learning methods**
  - Attribute-value methods (C4.5, Naïve Bayes), propositional
  - Relational methods bottom-up and top-down (FOIL-like)
  - Grammatical inference (Alergia)

Example

Annotated sentence

The <agent type=protein>GerE protein</agent> <interaction type=negative>inhibits </interaction> <target type=expression>transcription of <source type=gene>the sigK gene</source> encoding <product>sigmaK</product></target>

Positive example of the field agent

```
interaction_agent(GerE protein):-
  tk(the), tk(Ger_protein), tk(inhibit), tk(transcription), tk(of), …
  cat(the, det), cat(Ger_protein, term), cat(inhibit, verb), …
  next-token(the, Ger_protein), next-token(Ger_protein, inhibit), …
  subject(Ger_protein, inhibit), DObj(transcription, inhibit), …
  concept(Ger_protein, protein), concept(inibit, negative_interaction), …
  agent(Ger_protein, inhibit), cible(transcription, inhibit), …
  …
```

Open problems

- **Role of the additional knowledge**
  (syntactic and semantic tags and relations)

⇨ can normalize and increase the textual regularities (e.g. replacing terms by concepts) but potentially suppress discriminant attributes

⇨ potentially add discriminant attributes but drastically augment the search space

⇨ **Feature granularity and selection problem**
Obviously depend on the information to be extracted

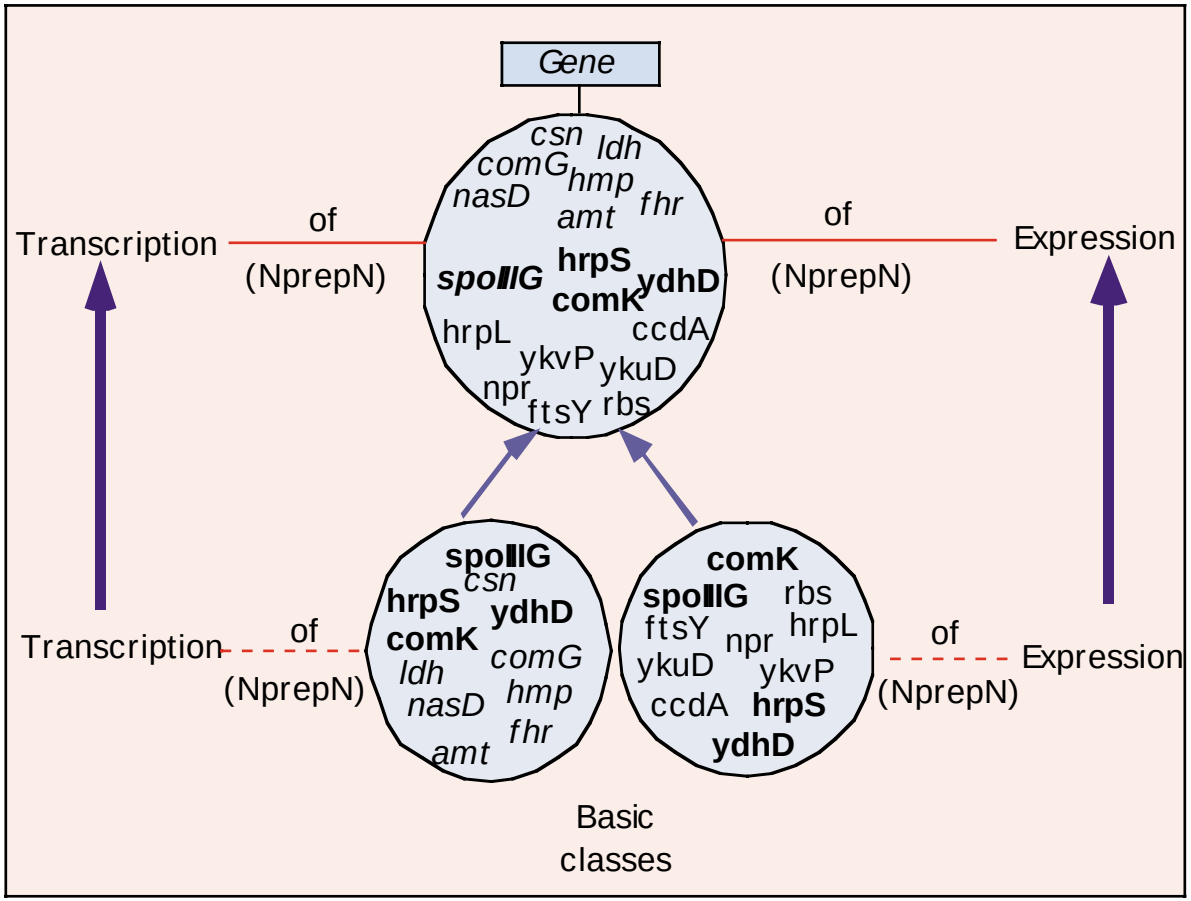
- **Learning the additional knowledge**

Learning background knowledge for IE

Types of knowledge needed	How to get it
• Syntactic • Syntactic categories (parts of speech) • Syntactic relations (dependencies)	**Tools** • Morphosyntactic taggers • Syntactic parsers
Semantic • Semantic categories (conceptual hierarchies), restrictions of selection	**Background knowledge learned from corpus** • Distributional semantics (clustering)
• Subcategorization frames • Predicate schemata	• Grammatical inference, ILP • FOL clustering
Extraction rules	• Relational learning, ILP

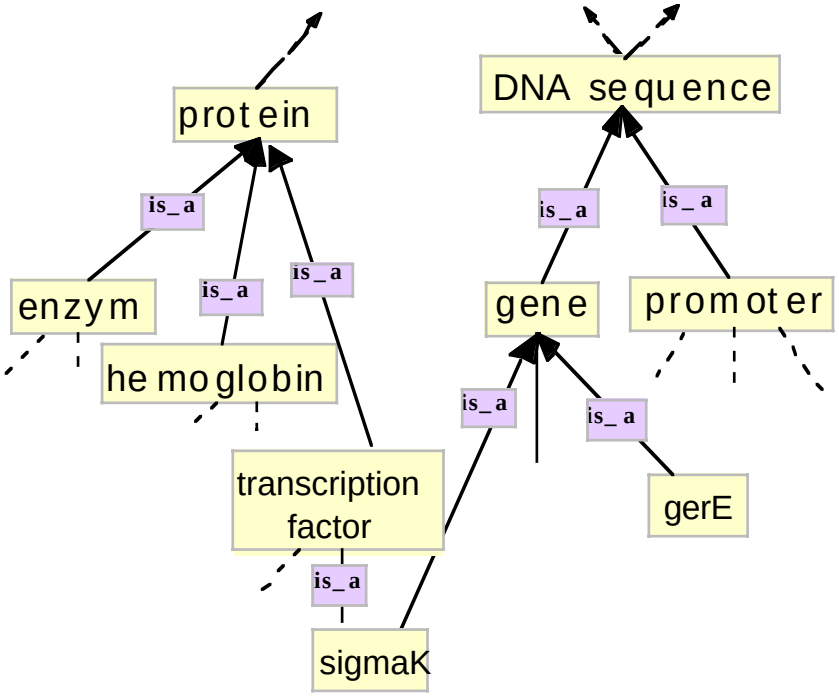
Learning conceptual hierarchies by clustering

Classes of words co-occurring in different syntactic contexts form concepts



Clustering result

Conceptual hierarchy



Restrictions of selection

Verb : activate Subject : protein DObj : expression	Noun : inhibition NprepN (by): protein NprépN (of): gene

Open problems

Evaluation [Nedellec & Bisson, 2000]

Integration of known terminology (Gene Ontology, gene dictionaries)

Learning background knowledge for IE

Types of knowledge needed	How to get it
• Syntactic • Syntactic categories (parts of speech) • Syntactic relations (dependencies)	**Tools** • Morphosyntactic taggers • Syntactic parsers
Semantic • Semantic categories (conceptual hierarchies), restrictions of selection	**Background knowledge learned from corpus** • Distributional semantics (clustering)
• Subcategorization frames	• Grammatical inference, ILP
• Predicate schemata	• FOL clustering
Extraction rules	• Relational learning, ILP

From selectional restrictions to subcategorization frames

Subcategorization frames =
restrictions of selection + syntactic (structure) and semantic constraints

Example

clustering ⇨ binary independent relations	Subcategorization frame
Verb : inhibit Subject : protein Verb : inhibit Subject : stress Verb : inhibit DObj : expression	Verb : inhibit Subject : *repressor XOR* stress DObj : expression

- *Protein* and *stress* are mutually exclusive as subject of *inhibit*
- Among proteins, only *repressors* inhibit gene expressions

Subcategorization frame learning by grammatical inference

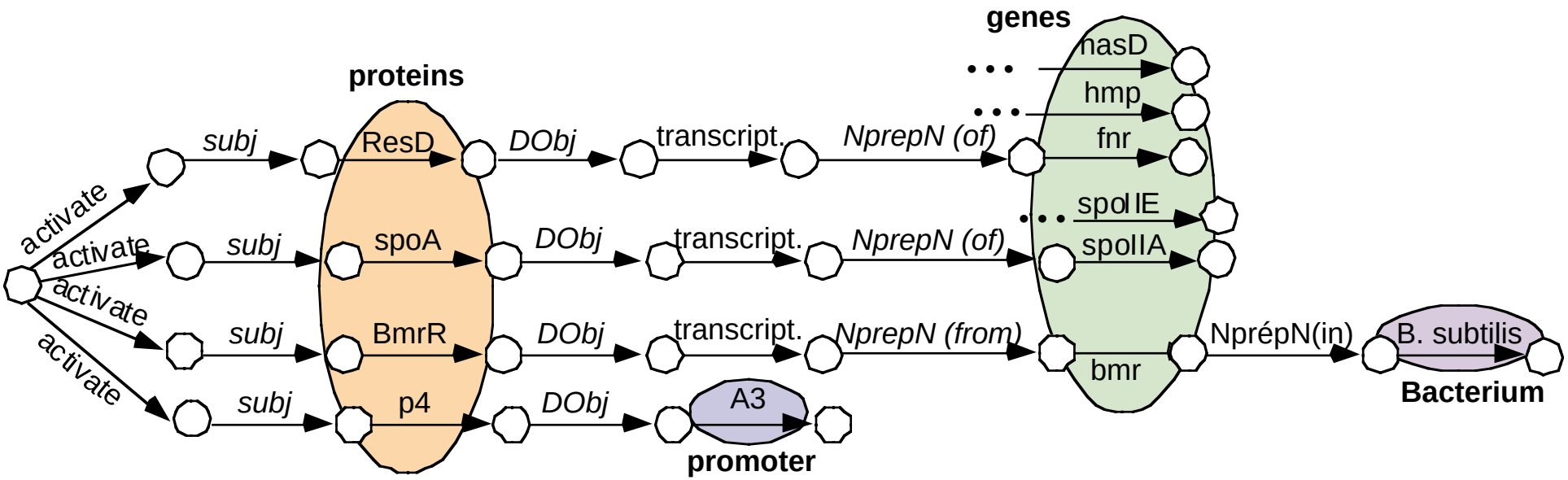
Sentences

ResD activates transcription of fnr, hmp and nasD

The phosphorylated form of Spo0A activates spoIIA and spoIIE transcription.

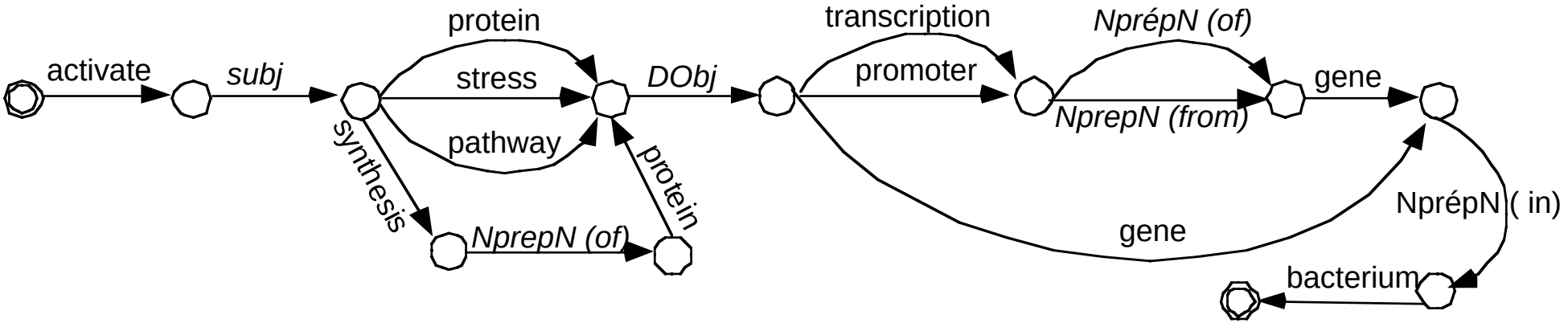
BmrR activates transcription FNR

PTA:



Subcategorization frame learning by grammatical inference

Verb : activate	Subject: protein	DObj: transcription
	Subject: stress	DObj: promoter
	Subject: pathway	DObj: gene
	Subject: syntesis	NprepN (in): bacterium
Noun : synthesis	NprepN (of): protein	
Noun : transcription	NprepN (of): gene	NprepN (from): gene

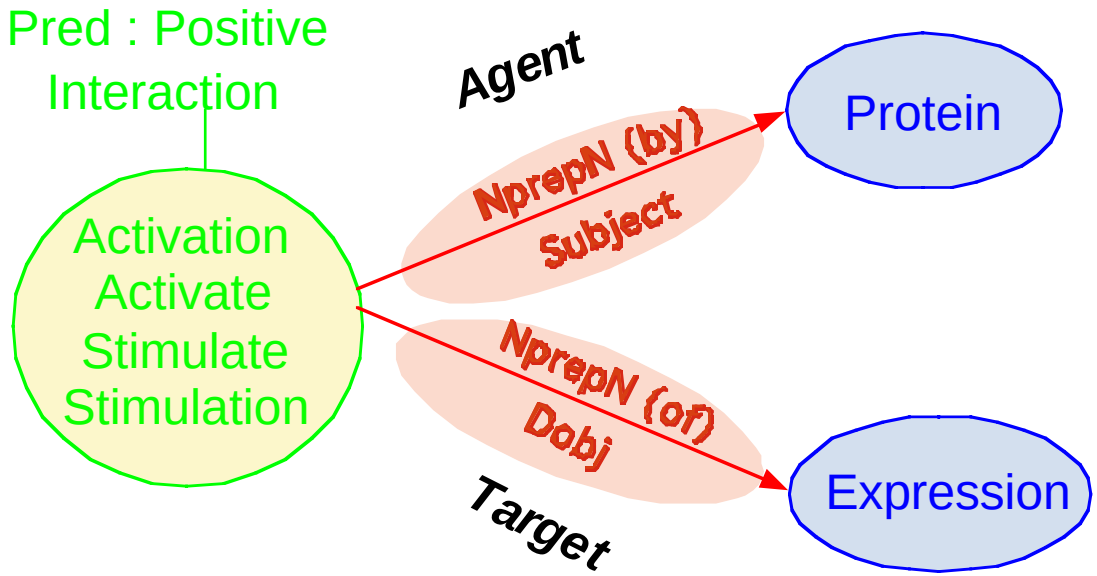


Learning background knowledge for IE

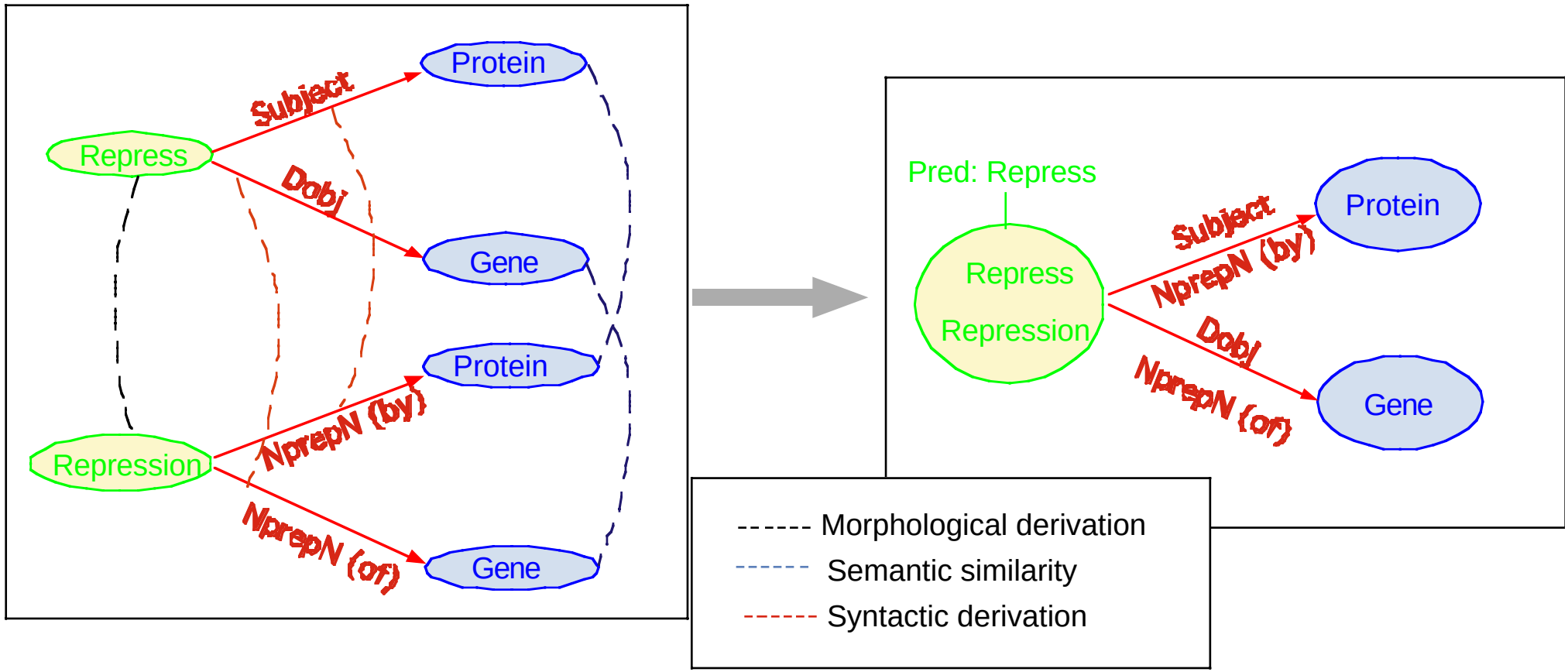
Types of knowledge needed	How to get it
• Syntactic • Syntactic categories (parts of speech) • Syntactic relations (dependencies)	**Tools** • morphosyntactic taggers • syntactic parsers (SP XRCE)
Semantic • Semantic categories (conceptual hierarchies), restrictions of selection • subcategorization frames	**Background knowledge learned from corpus** • distributional semantics (clustering) • grammatical inference, ILP
• Predicate schemata	• FOL clustering
Extraction rules	• relational learning, ILP

Predicate schemata

Predicate schemata = predicate classes and their arguments related by semantic and syntactic dependencies



Learning predicate schemata from subcategorization frames, by clustering



Conclusion

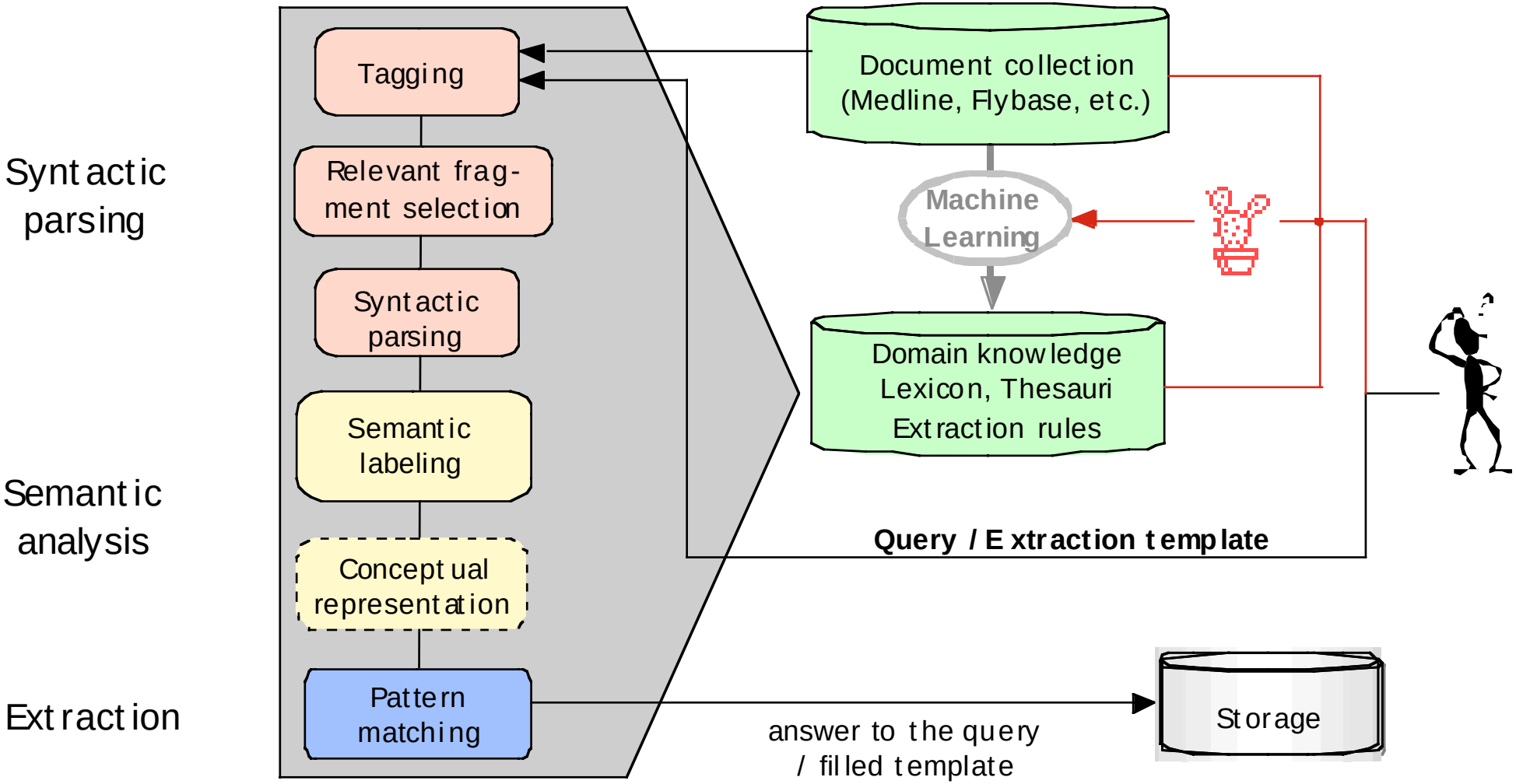
IE in functional genomics, as in many domains requires

- deep understanding methods based on syntactic-semantic analysis
- dictionaries, lexicon, ontologies, thesaurii, IE patterns to be learned by multistrategy learning and integrated with existing knowledge bases.

Learning methods for acquiring the needed knowledge and the IE patterns include relational methods

- *able to handle*
  - background knowledge
  - relations (syntactic dependencies, semantic relations)
- *able to acquire*
  - conceptual hierarchies for semantic labeling
  - subcategorization frames for disambiguating semantic labeling
  - predicate argument structure for semantic relation labeling
  - IE patterns

Architecture of Caderige (http://caderige.imag.fr)



One further step towards semantic normalization

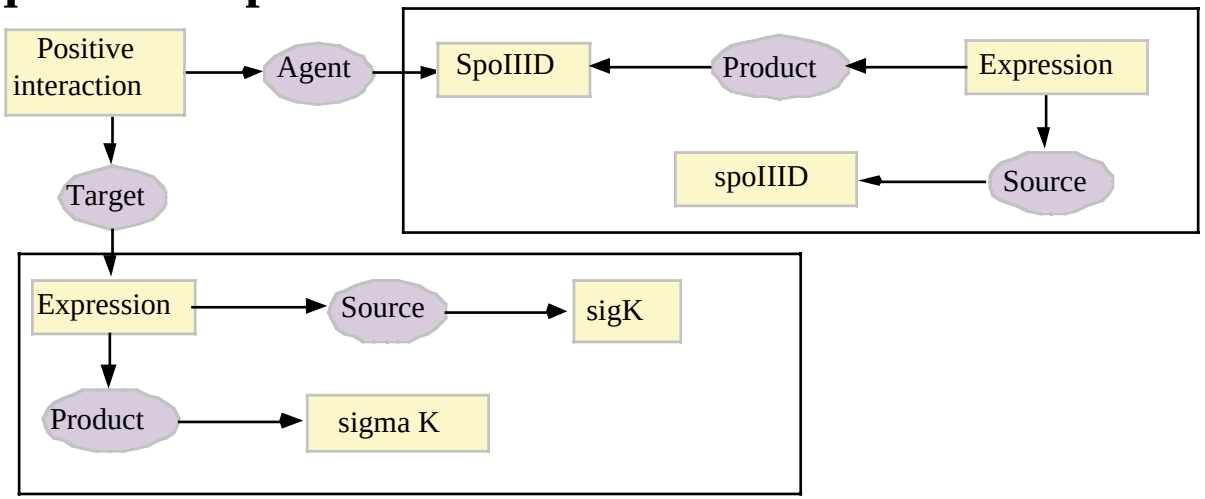
Various expressions ...

The expression of spoIIID
spoIIID expression
The spoIIID gene product
The production of SpoIIID
SpoIID
SpoIIID production

stimulates

the expression of sigK.
sigK expression.
the sigK gene product
the production of sigma K.
sigma K production.

for a unique interpretation



From restrictions of selection to conceptual structures

- **Learning sets of subcategorization frames**

  (nominalization)

- **Learning semantic sets of predicate structures** with their corresponding arguments