Ontology learning for Information Extraction in genomics – the Caderige Project

Philippe Bessières

MIG -INRA
Jouy-en-Josas

philb@biotec.jouy.inria.fr

Gilles Bisson

Leibniz – IMAG
CNRS Grenoble

gilles.bisson@imag.fr

Adeline Nazarenko

LIPN – Université Paris-Nord
& CNRS

nazarenko@lipn.univ-paris13.fr

Claire Nédellec

LRI
Université Paris-Sud &
CNRS

cn@lri.fr

Mohammed Ould Abdel Vetah

LRI & Valigen

Mohammed.Ould-Abdel-Vetah@lri.fr

Thierry Poibeau

Thalès Group

thierry.poibeau@thalesgroup.com

Outline

1. Overall approach: *from scientific abstracts to gene interaction database*
2. A knowledge-based extraction method
3. Building classes for semantic tagging
4. Learning extraction rules
5. Towards a conceptual representation of texts

An Information Extraction problem

Functional Genomics: gene interaction discovery

- Experimental approaches (sequencing, functional analysis)
- *Information Extraction in Genomics literature*

Examples of bibliography databases

	MedLine	**FlyBase**
DB Size	> 16 millions of refs.	> 9500 genes recorded
Abstract length	10 sentences	2 - 3 sentences

Example: a MedLine abstract

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 region of the cotD promoter. A low level of GerE activated transcription of cotD by sigmaK RNA polymerase in vitro, but a higher level of GerE repressed cotD 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 [..]

Example of information extracted from a text fragment

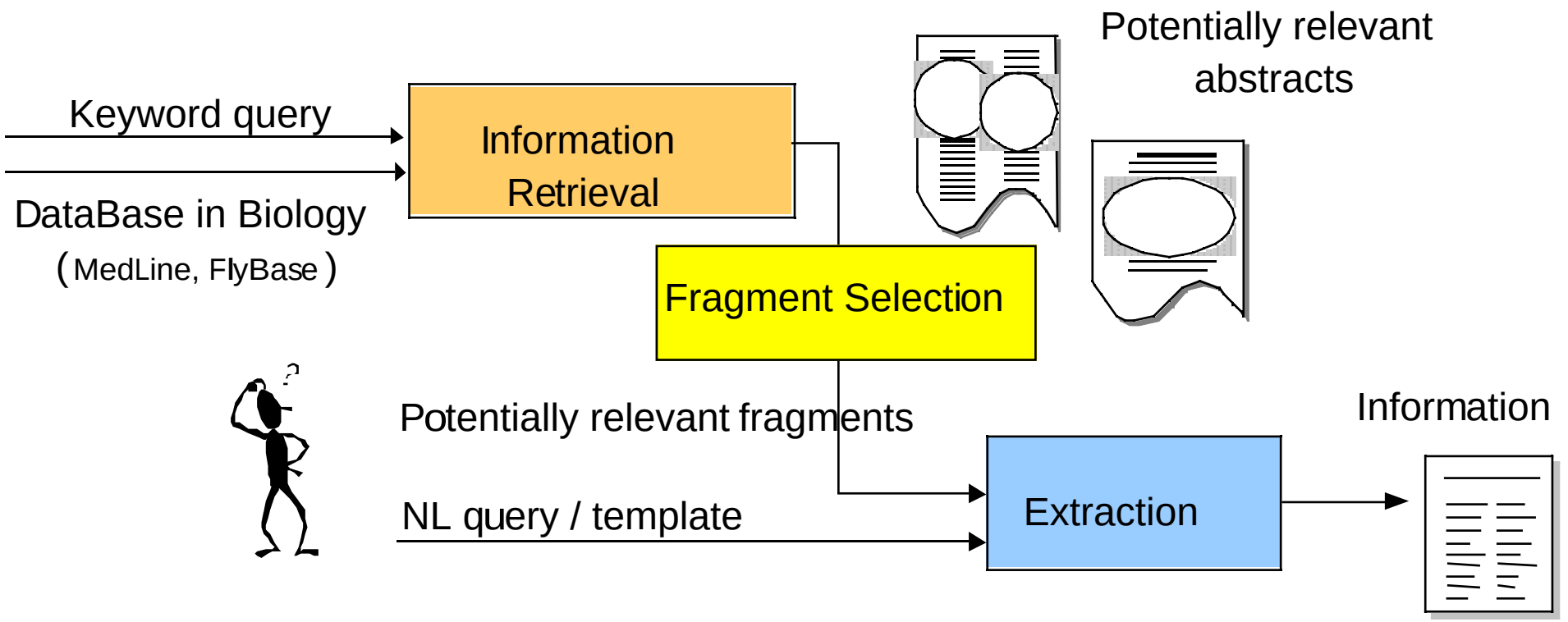
Fragment from a Medline abstract

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

Filled form

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

Information Extraction in Genomics



Overall approach

As information is scattered (around 3 % of the abstract sentences are relevant for the discovery of gene interactions), a full text analysis is too costly

A two step approach: “selection first, then extraction”

- Relevant fragment selection

  A fast and robust processing based on surface clues and key words

- Knowledge extraction

  Apply extraction rules on “normalized” texts

Limitations of keywords based approaches (1)

Identifying the presence of interaction between 2 genes using word weights

- 80 % Recall and precision for sentences including 2 gene names
- Few information is extracted (classification based approach)

$$\text{Recall}(\text{Class}_i) = \frac{|\text{Ex} \in \text{Class}_i \text{ and classified in Class}_i|}{|\text{Ex} \in \text{Classe}_i|}$$

$$\text{Precision}(\text{Class}_i) = \frac{|\text{Ex} \in \text{Class}_i \text{ and classified in Class}_i|}{|\text{Ex classifed in Classe}_i|}$$

Limitations of keywords based approaches (2)

Identifying interaction triples (gene name/protein, interaction verb, gene name/protein)

more information, but low precision

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

Constraint on the number of words between the elements of the triple

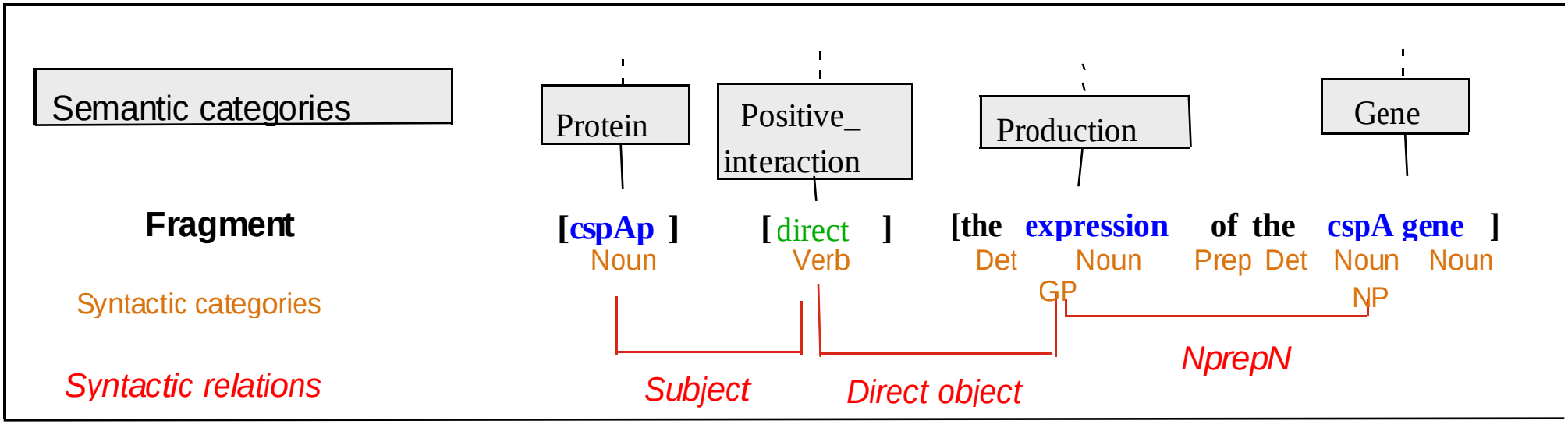
- ❶*Distance ≤ 5 words: good precision but low recall*
- ❶*Distance > 5 words: lower precision*

Combining different level of textual analysis

For a good precision and a large recall, extraction rules should include conditions on different textual analysis levels

1. Sentence processing

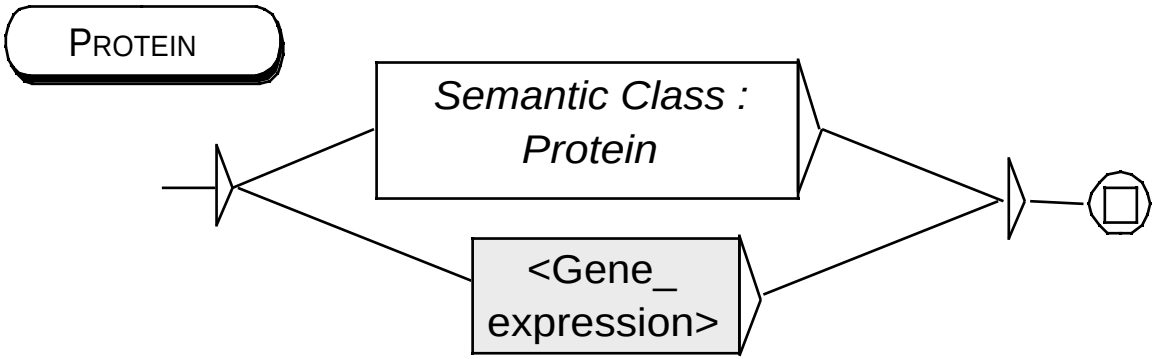
Parsing and semantic tagging lead to an enriched and normalized text representation

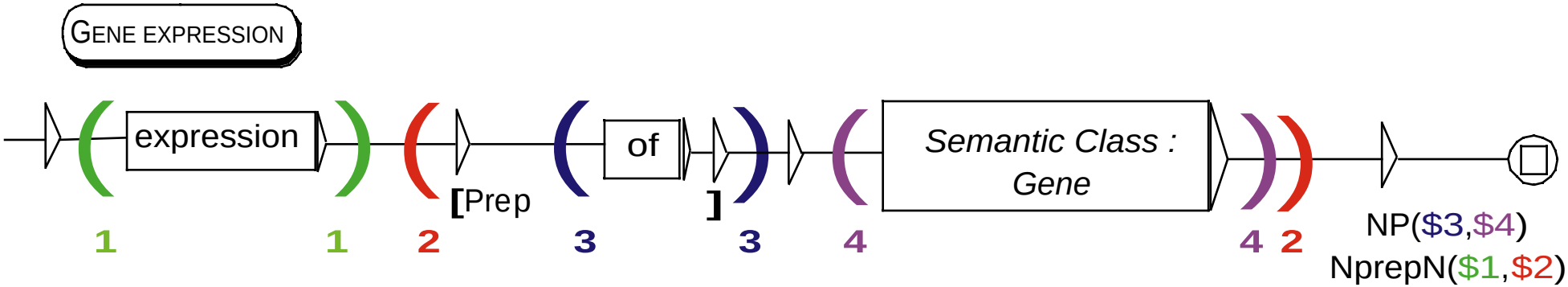


2. Application of extraction rules (automata) on the resulting interpretation

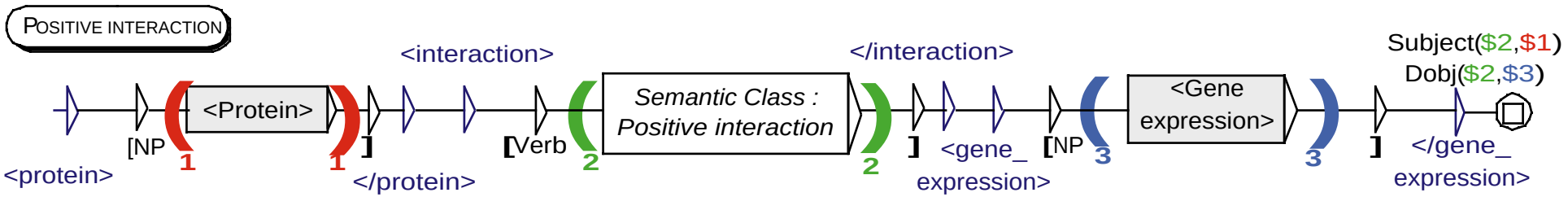
Automata examples: protein identification

The automata use the syntactic and semantic information from the parsing phase to recognize interactions

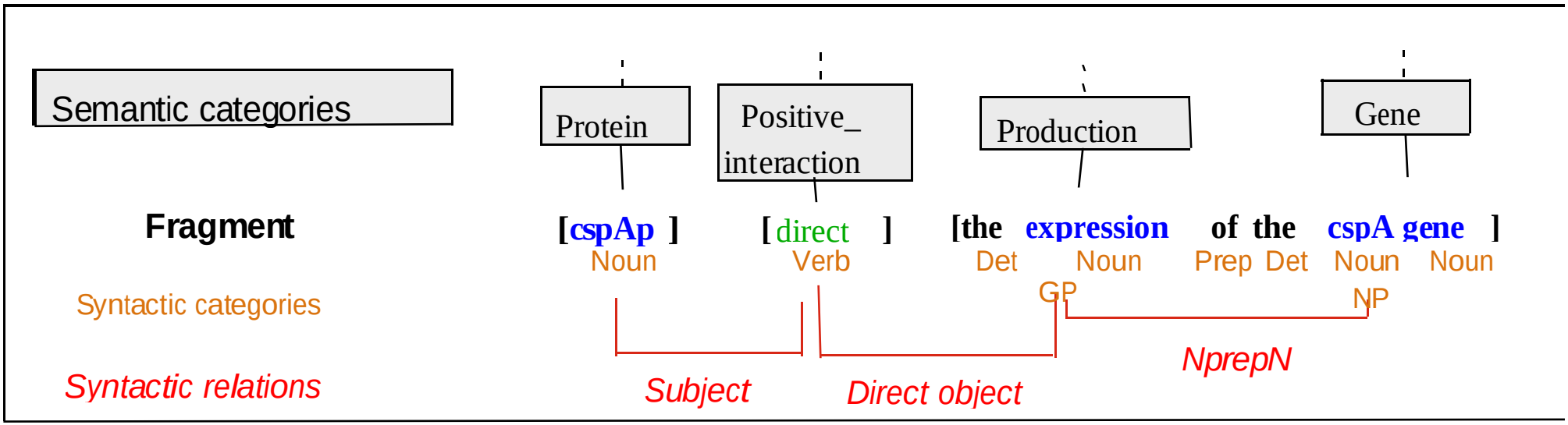




Automaton example: interaction identification and mark up

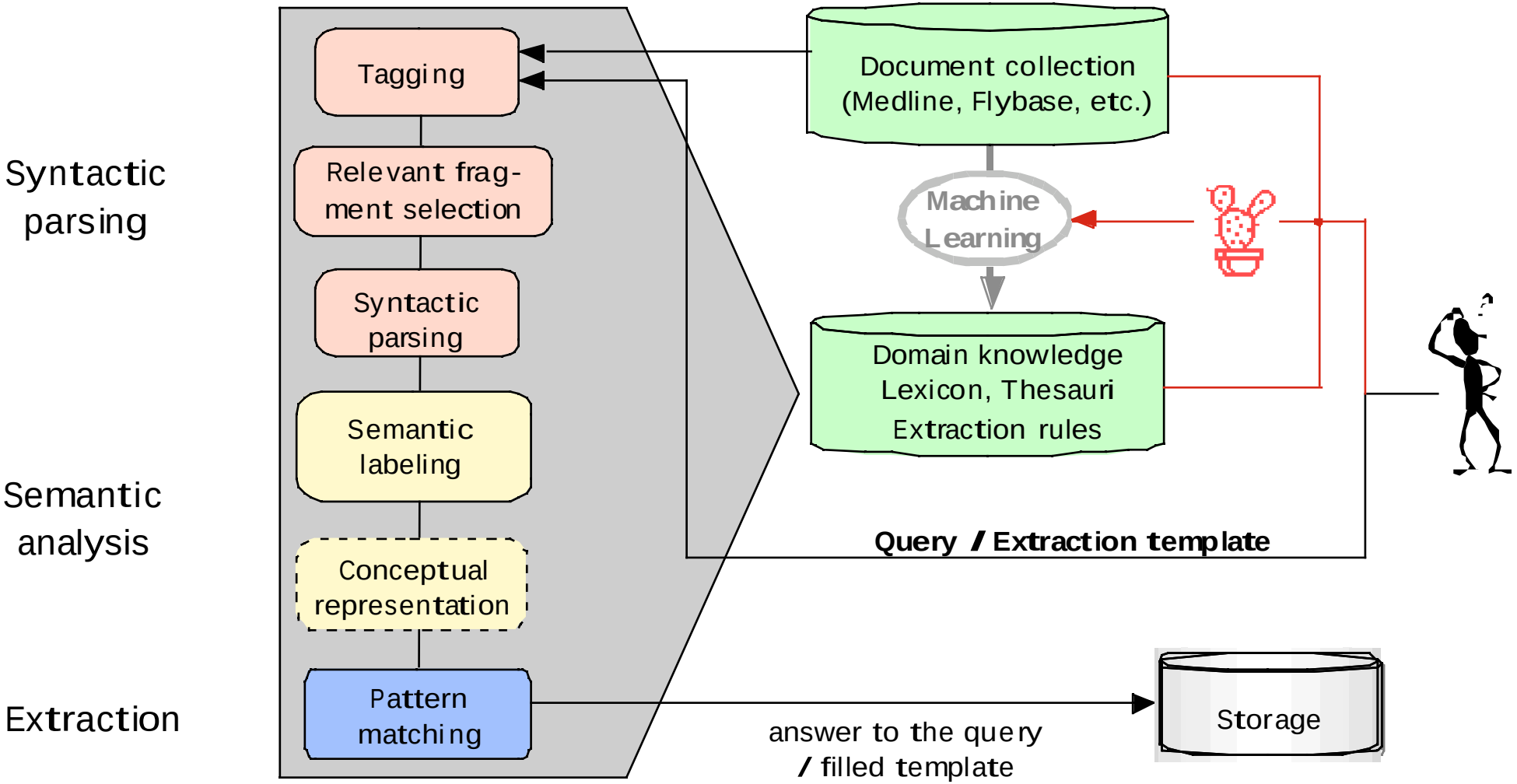


Syntactic and semantic knowledge needed



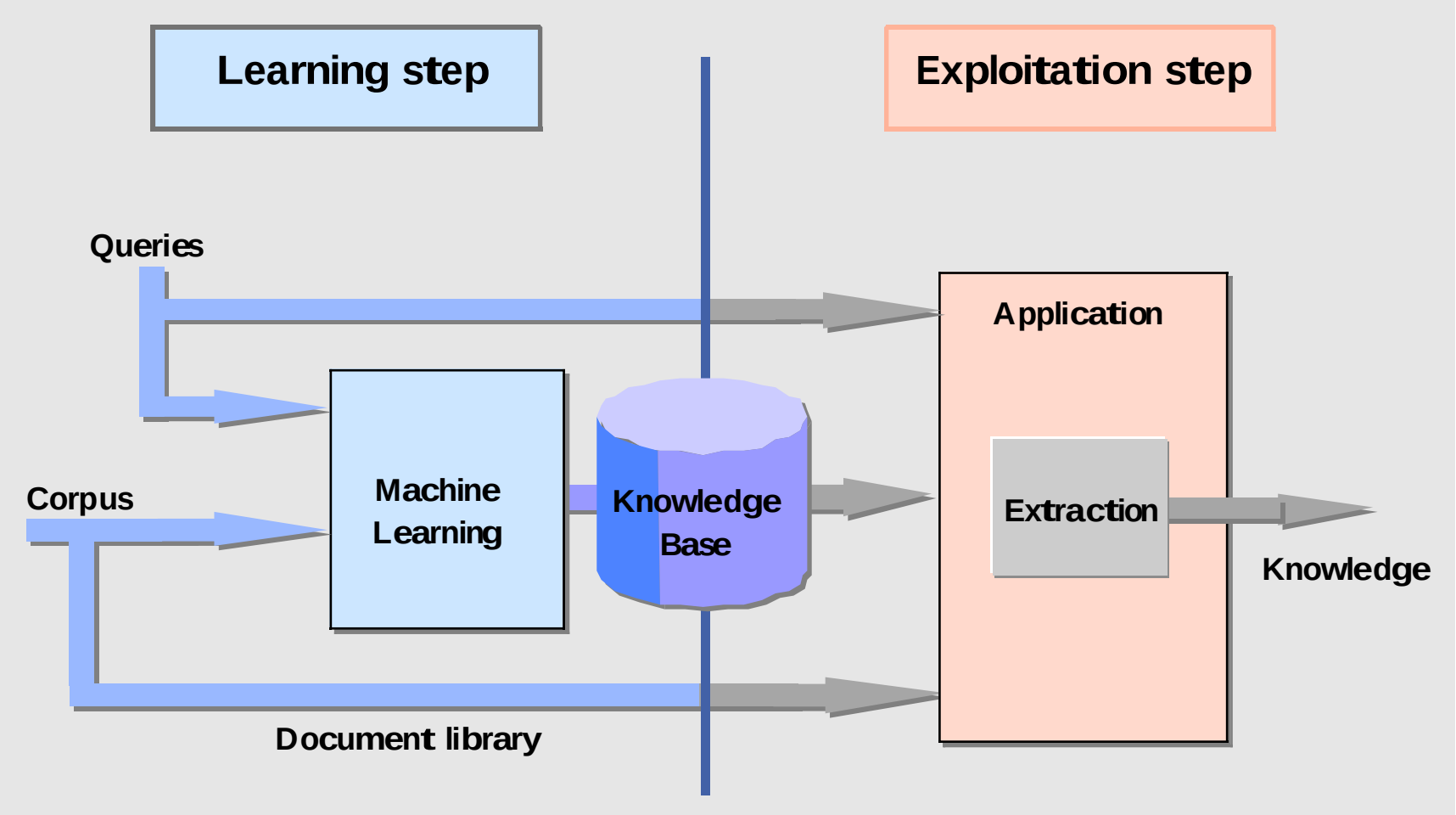
Types of knowledge needed	How to get it
Syntactic categories (parts of speech) Syntactic relations (dependencies)	Tools exist: • morphosyntactic taggers • syntactic parsers (SP XRCE)
Semantic categories (conceptual hierarchies) Extraction rules *Predicate schemata*	Knowledge can be learned from the corpus

Architecture of Caderige

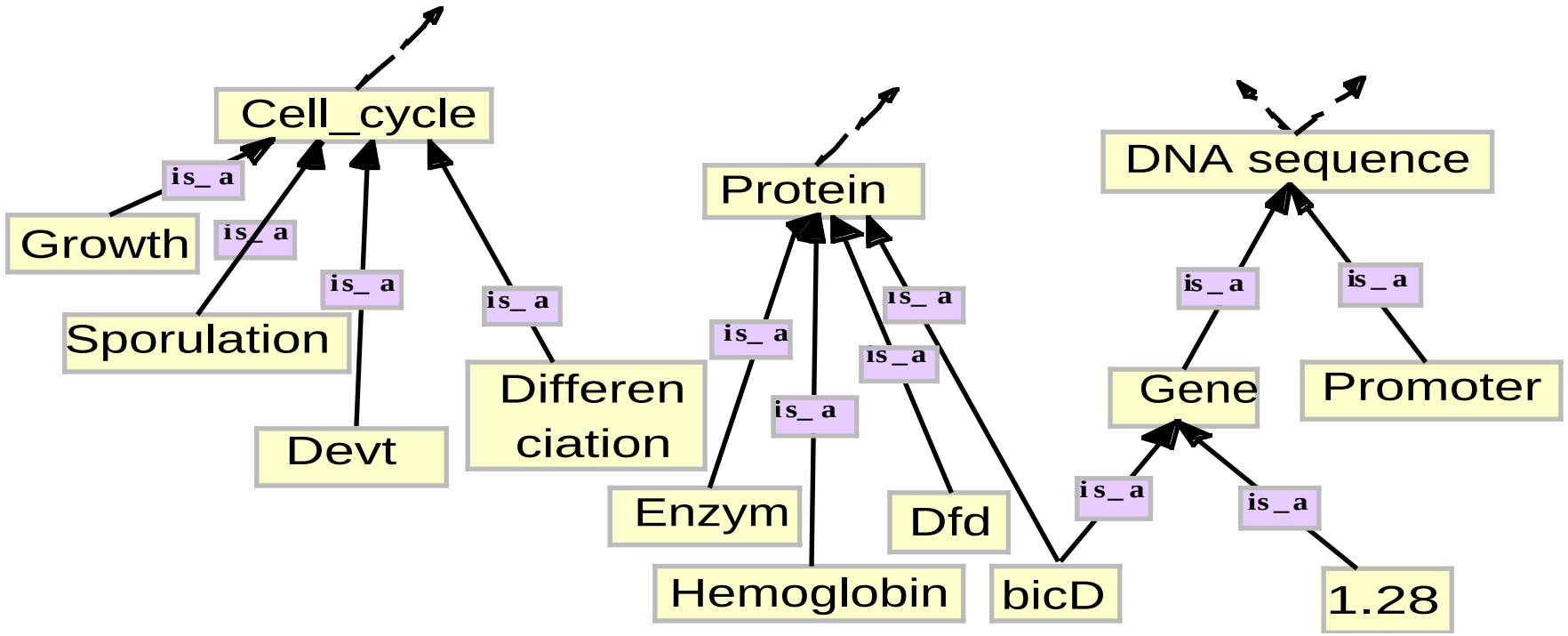


Knowledge learning and exploitation

(Information Extraction task)



Learning conceptual hierarchies for semantic tagging



Hierarchies of semantic classes can be learned if the following conditions are sastified:

- from an homogeneous corpus, written in a specialized language
- using a robust parser
- with the help of an expert (or user)

Classical approaches to word classes building

Harris' assumption of distributional semantics

The semantics is reflected by the syntax in specific domain corpora

Some semantics can be learned by observing syntactic regularities

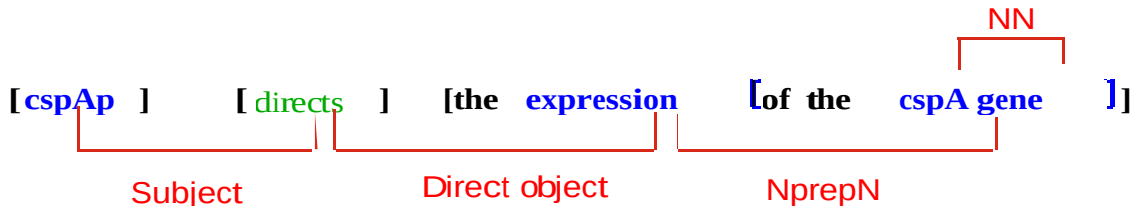
- The classes are based on the semantic proximity between words
- The similarity measure of two words is based on the number of their *common contexts* of in the training corpus
- Traditional context definitions

  ➢Word co-occurrences within a window, or in a document.

  ➢Co-occurrences of words relation of syntactic dependancy

Similarity based on the syntactic context

- Parsing gives syntactic relations between the predicates (verb/noun) and their arguments
- Syntactic dependencies are represented as triplets (predicate, relation, argument)
- These triplets are the **learning examples**



Expression ***NprepN (of)*** *N*

[Expression**] [of** spoIIIG**].**

[Expression**] [of** ykuD**].**

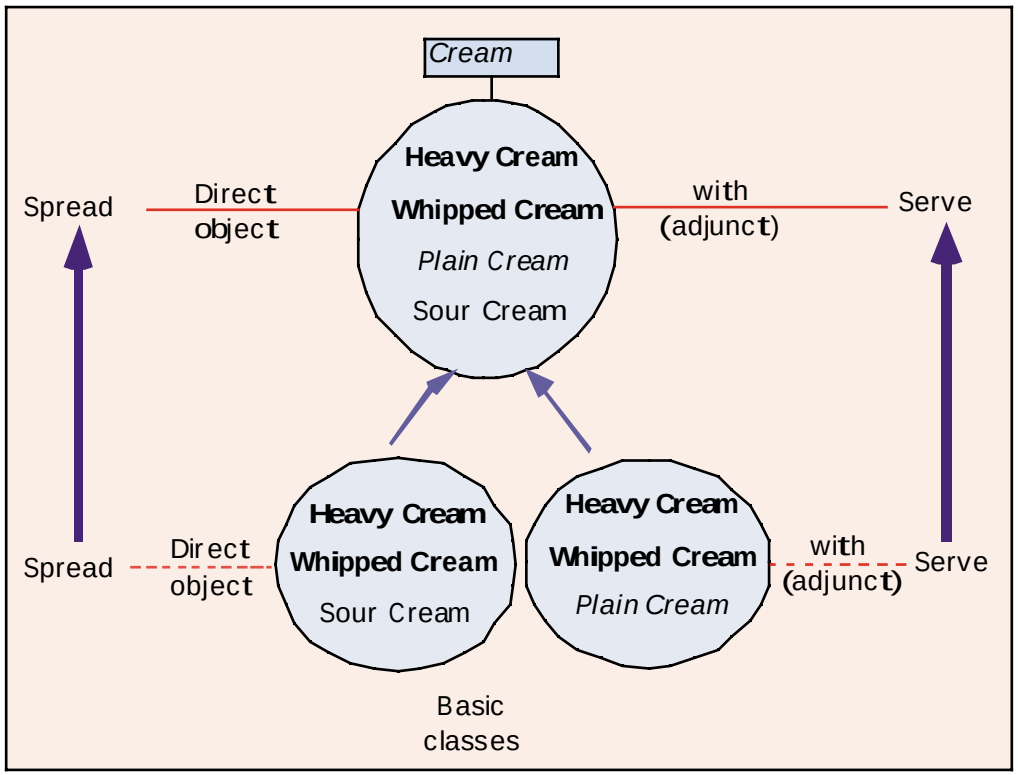
Transcription ***NprépN (of)*** *N*

[Transcription**] [**SpoIIIG**].**

[Transcription**] [**comG**].**

[Transcription**] [**ydhD**].**

Classes of words co-occurring in different syntactic contexts form a concept



- Builds words classes along with their selectional restrictions (predicates or arguments which the words can occur with)
- Generalizes the syntactic dependencies observed in the corpus

From word classes to term classes

Limitations of word classes

- The terms (domain relevant semantic units) are often multi-word expressions
- Single word expressions are often polysemous and difficult to interpret
- Working with complex terms reduces syntactic ambiguity and therefore increases
  distributional evidence

Problem for building term classes

- How to identify terms which result from domain expert agreement?
- How to process terms of heterogeneous size (up to 5 or 6 words) in a distributional analysis?

Building term classes

Term extraction using ACABIT [Daille 95]

- List of potential terms and variants

`acid synthase deficient`
`stationary phase phenomena`
`new tangible evidence`
`fatty acid ↔ fatty acids`
`chromosomal map`
`several genes`
`further distinctive conformational change`
`unsaturated acid ↔ unsaturated fatty acid`
`stable RNA`
`alpha-oxo acid`
`map of Piggot and Hoch`
`set of single-gene replacement`

- Relevance sorting criteria (logLike)

Term filtering using

- Stop lists to filter out noise (~~`futher`~~, ~~`several`~~, ~~`set of`~~ …)
- Existing keyword lists and glossaries (SwissProt, JouyINRA…) to choose a relevance threshold

Redefinition of ASIUM distributional analysis to take complex terms into account

Class building experimentations and parameter tuning using Mo'K

Methods for the design of extraction rules

Manual design

Time consuming and difficult to tune the precision/recall balance

Semantic class learning and rule manual design

30% time gained with the help of semantic class learning [Faure & Poibeau, 2000].

Next step

Learning extraction rules from annotated and semantically tagged texts [Riloff, 93], [Freitag, 98], [Soderland, 99].

Extraction rule learning from a training corpus

Building a training corpus with interaction markup

Enriching and normalizing the training corpus

- Syntactic tagging and parsing
- Term identification
- Semantic tagging

Learning extraction rules from the training corpus, parsed and tagged

Normalization increases phrasing homogeneity and makes it easier to learn extraction rules

Building a training corpus

1. Fragment selection

2. Definition of annotation guidelines

3. Biologists must mark up relevant information in the training corpus

The GerE protein inhibits transcription of the sigK gene encoding sigmaK

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

- Training corpus of annotated examples

Extraction rule learning

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

- 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 [Ciravegna, 00]
  - ILP RHB+ [Sasaki & Matsuo, 00]

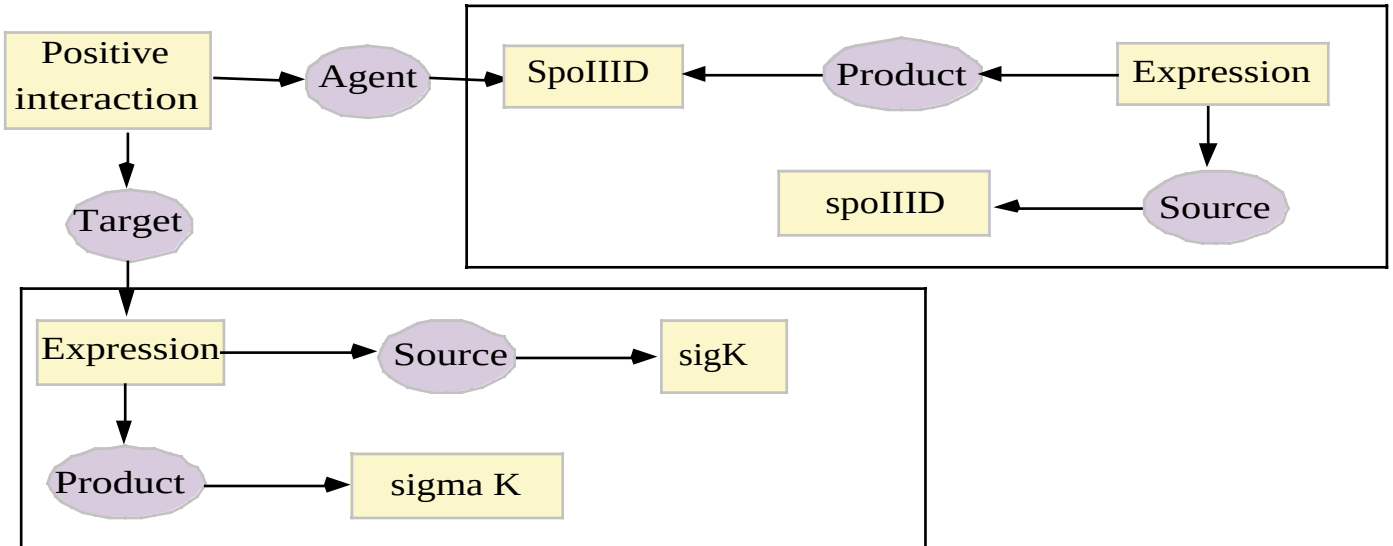
- Learning methods
  - Relational methods (ILP), bottom-up and top-down (FOIL-like)
  - Grammatical inference (Alergia)
  - Attribute-value methods (C4.5, Naïve Bayes) and propositional

One further step towards semantic normalization

Various expressions ...

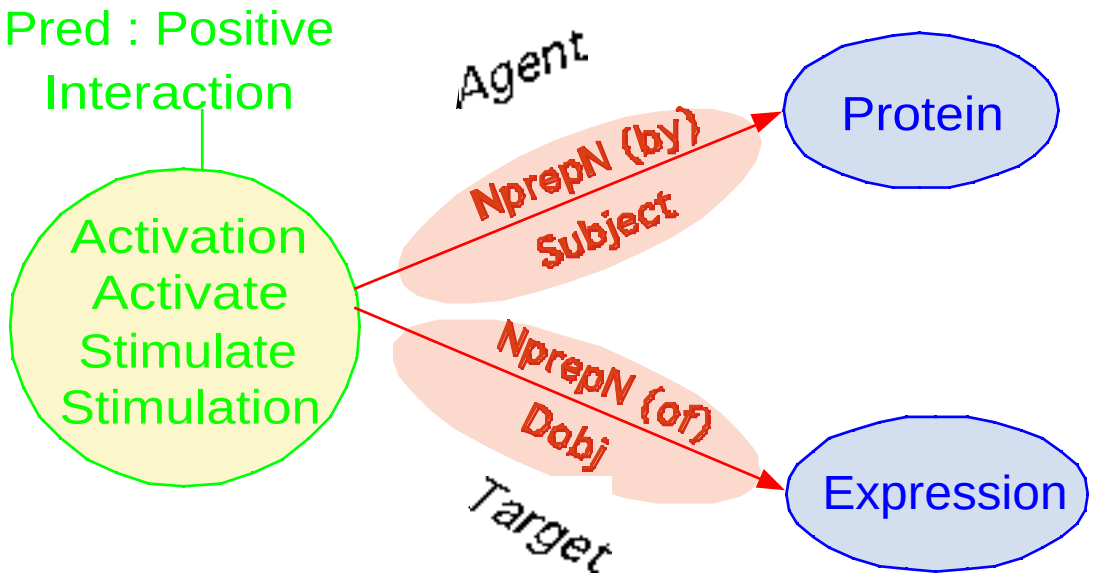
The expression of spoIIID		the expression of sigK.
spoIIID expression		sigK expression.
The spoIIID gene product	stimulates	the sigK gene product
The production of SpoIIID		the production of sigma K.
SpoIID		sigma K production.
SpoIIID production		

for one interpretation



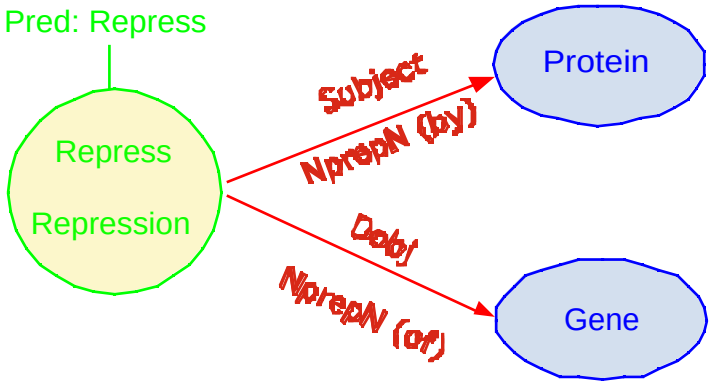
Additional knowledge: Predicate schemata

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

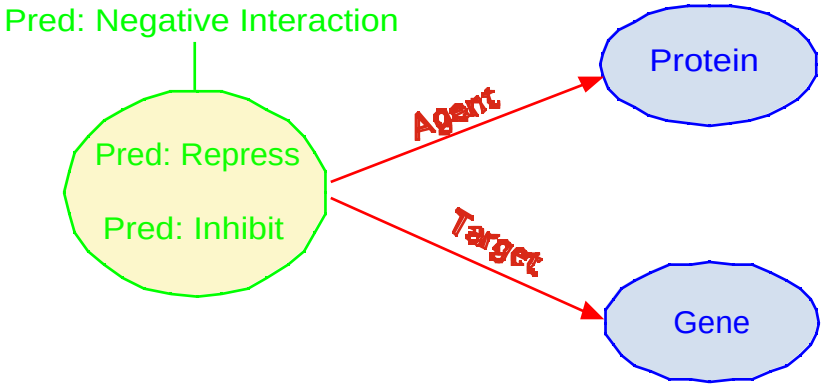


From restrictions of selection to conceptual structures

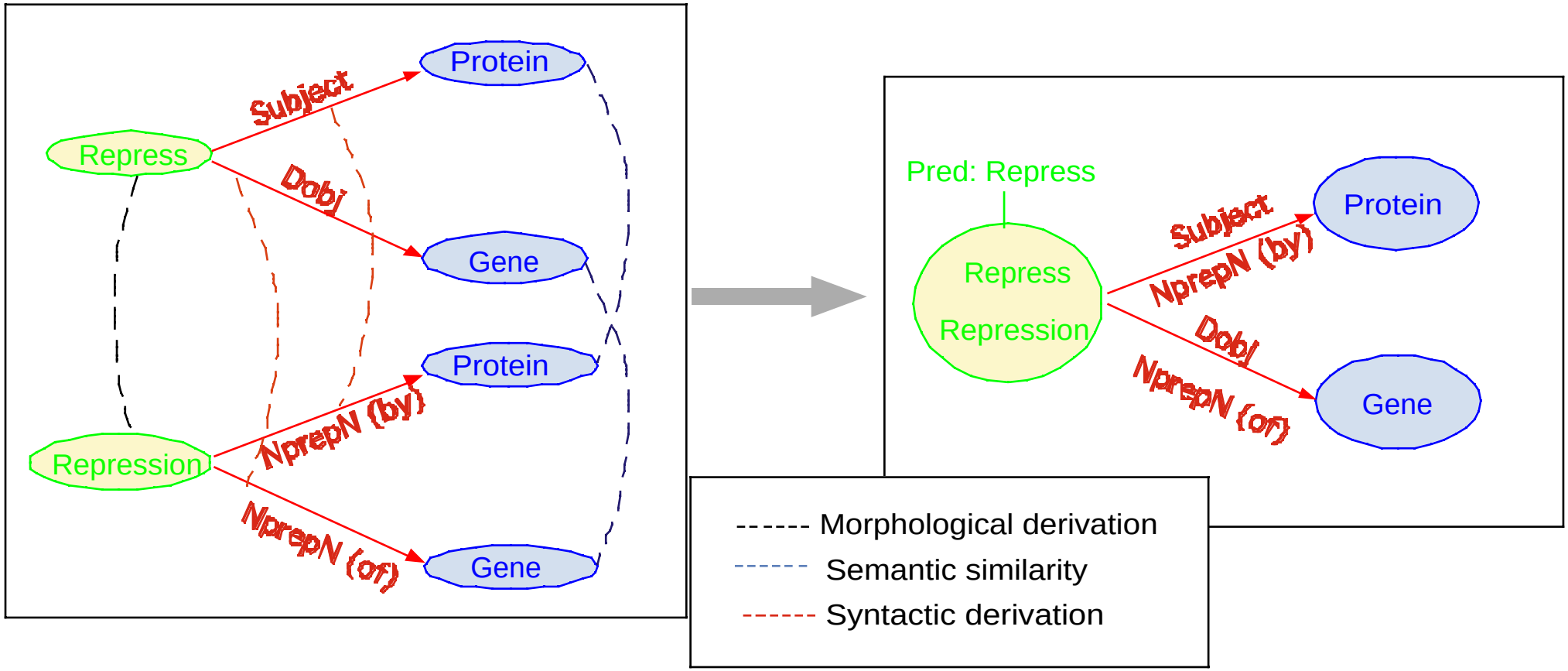
- **Selectional restrictions** are learned along with the semantic classes.
- **Learning subcategorization frames**
  Organizing and specializing the lists of selection restrictions with respect to the meaning and usage (`to perform an operation` / `to perform in a play`)
- **Learning sets of predicates which are morphologic derivations** with their corresponding arguments



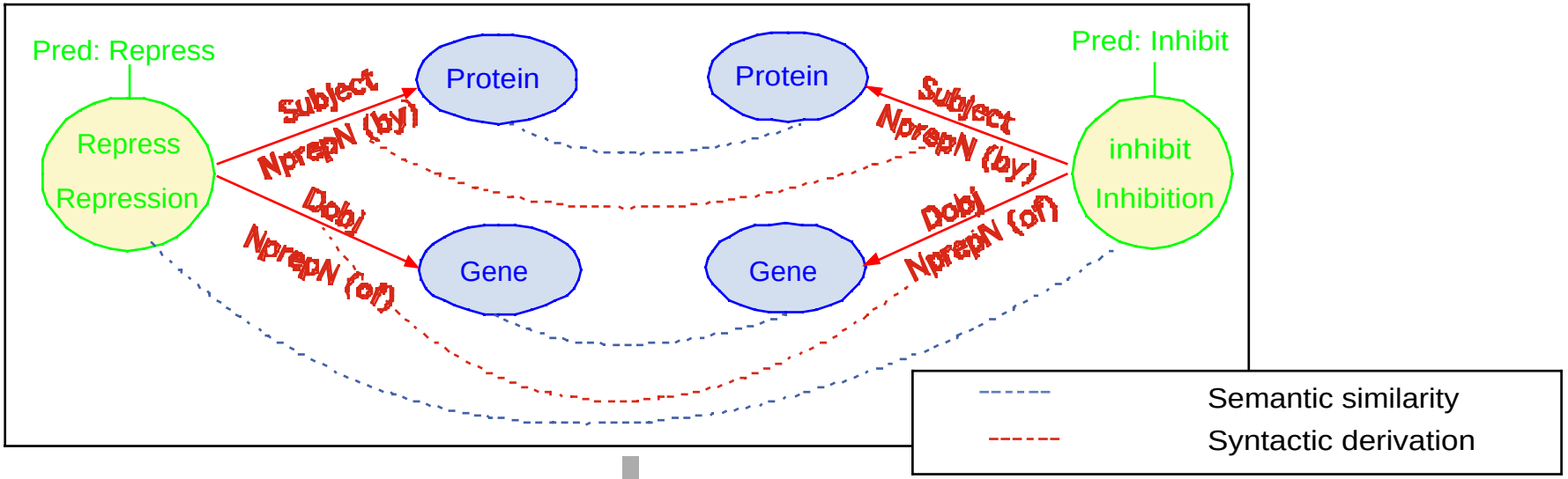
- **Learning semantic sets of predicates** with their corresponding arguments

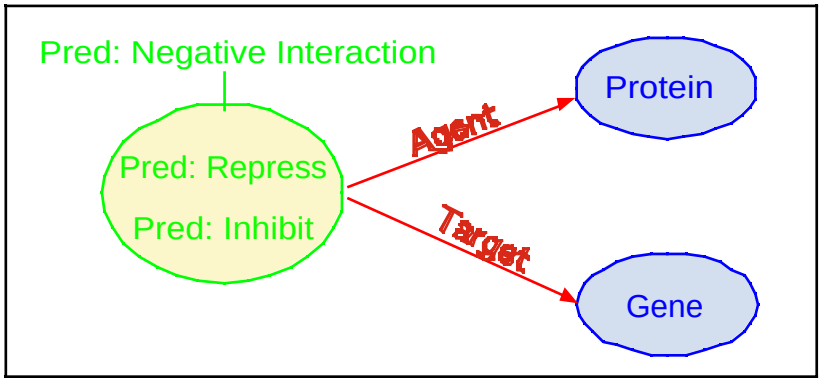


Learning predicate-argument structures



Learning conceptual structures





More conceptual interpretation

"The sigma factor controls the expression of gene dacB"

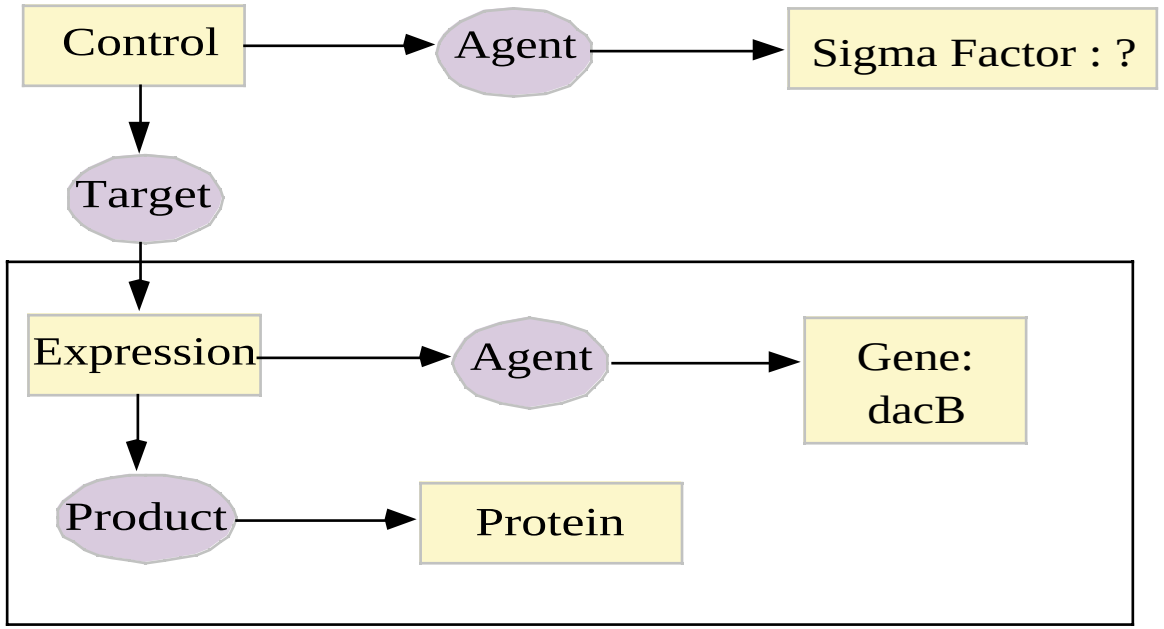
- **At the syntactic level**

Verb : control Subject : Sigma factor DObj : expression of gene dacB	Noun : Expression Noun Modifier (of) : dacB gene

- **At the predicate level**

Action = Control	(= to control, *verb*)
Agent = Protein	(= sigma factor, *subject*)
Object = Protein production	(= expression of gene dacB, *DObj*)
Action = Express	(= expression, *Noun*)
Agent = Gene	(= gene dacB, *Noun Mod*)

And the resulting interpretation



Open problems

- Co-reference resolution, negation
- Exploit the biological models (cascades, sequences, cycle, etc.)

Conclusion

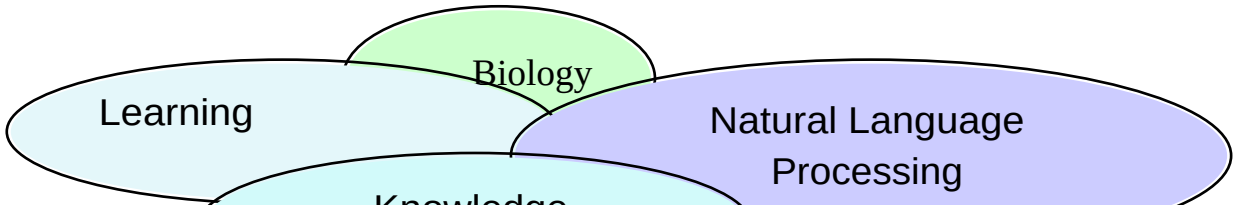
Information Extraction requires tools and linguistic/conceptual knowledge for building more abstract and conceptual representations of the text

- **Robust tools are available**: morphosyntactic taggers, syntactic parsers, term extractors…
- **Linguistic and conceptual knowledge can be automatically learned:**

  Today: semantic classes, selectional restrictions

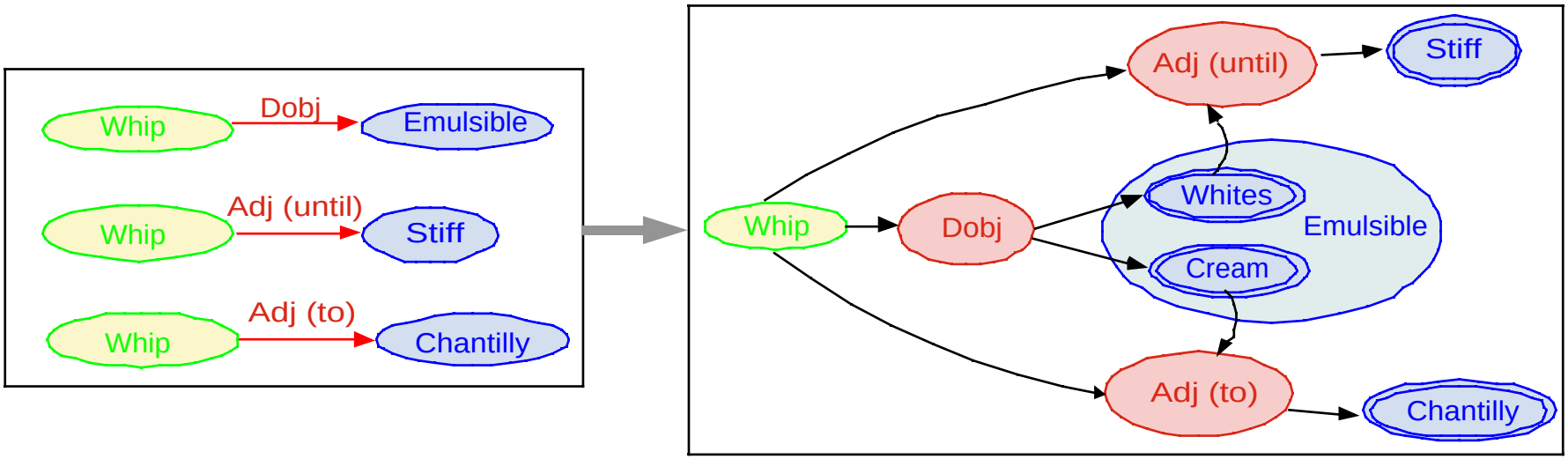
  Tomorrow: term classes, predicate schemata …

Building such resources call for multidisciplinary research and concern many other tasks than IE: Information Retrieval, Translation, Lexicography, Writing Assistance…



Subcategorization frames (SCF) learning

- From conceptual hierarchies, restrictions of selection and parsed corpus



- Learning **structural constraints**: optionality, mutual exclusion, etc.
  - ➔ Syntactic desambiguation of the attachments
- Learning **conceptual dependencies** between complements (restrictions of selection are overgeneral).
  - ➔ Semantic desambiguation: & efficiency in IR (❍ expansion of the queries)

➢ *Required for learning predicate argument structures*

The approach to learning SCF: ILP plus DL

- Hybrid method: combining Description Logic and Inductive Logic Programming for a good expressivity and a low complexity.

"Whip" has at least one direct object. They are all either Cream, or Whites.

Schemata1(X) :- Whip(X), ≥ 1 DObj(X), $\forall$ DObj.Cream(X).
Schemata2(X) :- Whip(X), ≥ 1 DObj(X), $\forall$ DObj.Whites(X).

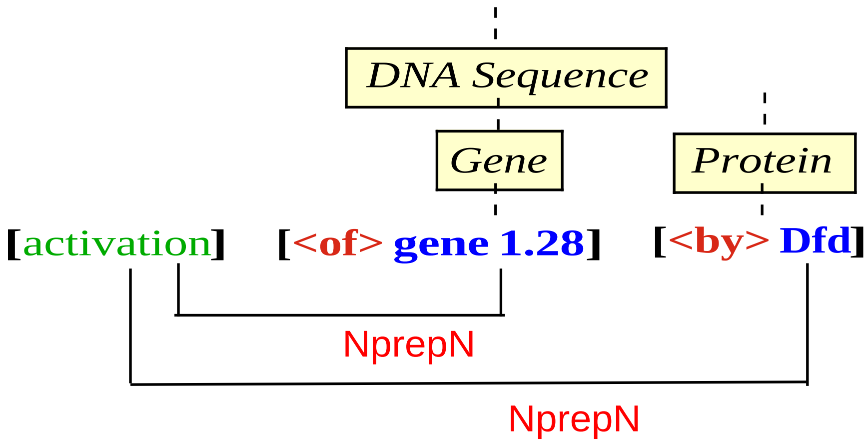
If "Whip" has a complement starting with "until", its head is of "stiff" type and there is no complement starting with "to".

Schemata3(X) :- Whip(X), $[\geq 1$ until(X) $]$, $\forall$ until.stiff(X), ≤ 0 to(X).
Schemata4(X) :- Whip(X), $[\geq 1$ to(X)$]$, $\forall$ to.Chantilly(X), ≤ 0 until(X).

- A complementary approach: Grammatical Inference.

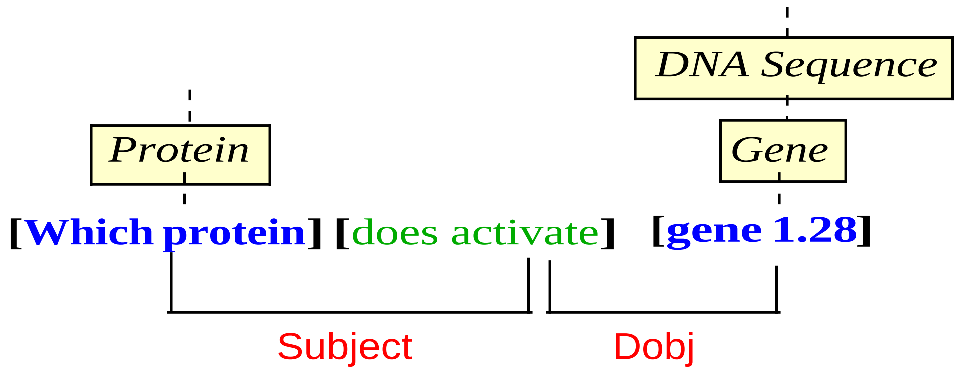
Fragment

Activation of gene 1.28 by Dfd […]



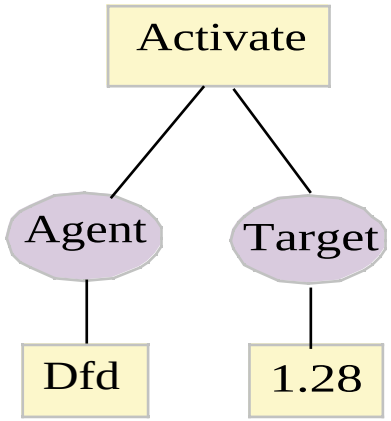
Question

Which protein does activate gene 1.28?



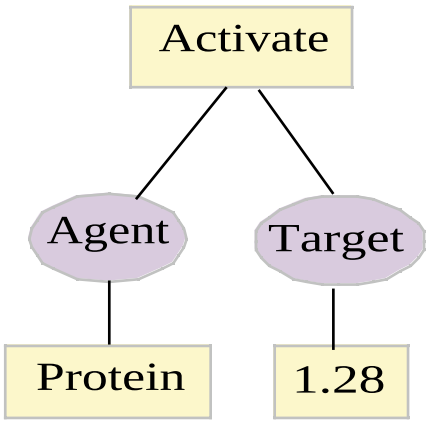
Fragment

Activation of gene 1.28 by Dfd […]

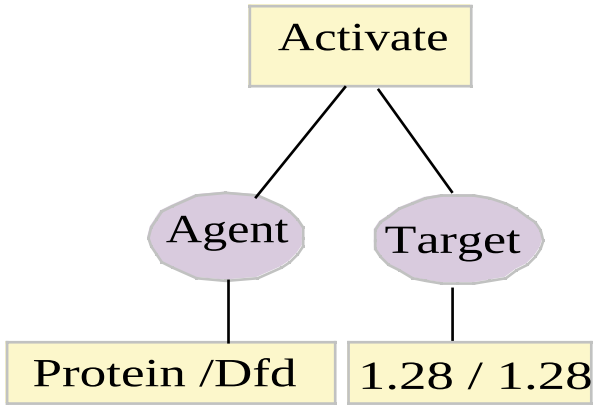


Question

Which protein does activate gene 1.28?



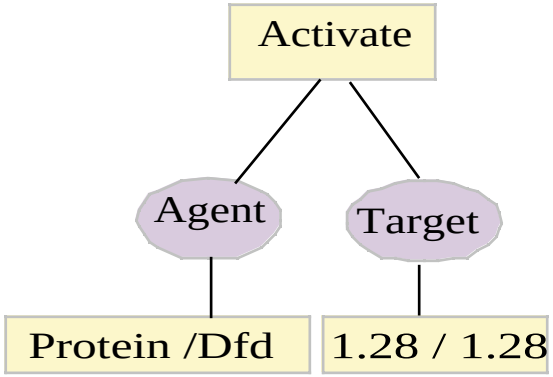
Projection



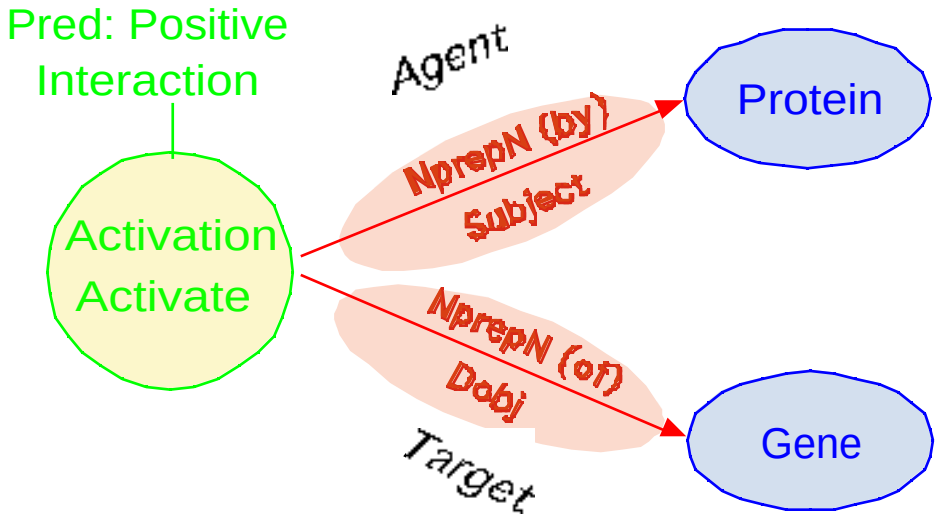
The conceptual structures required

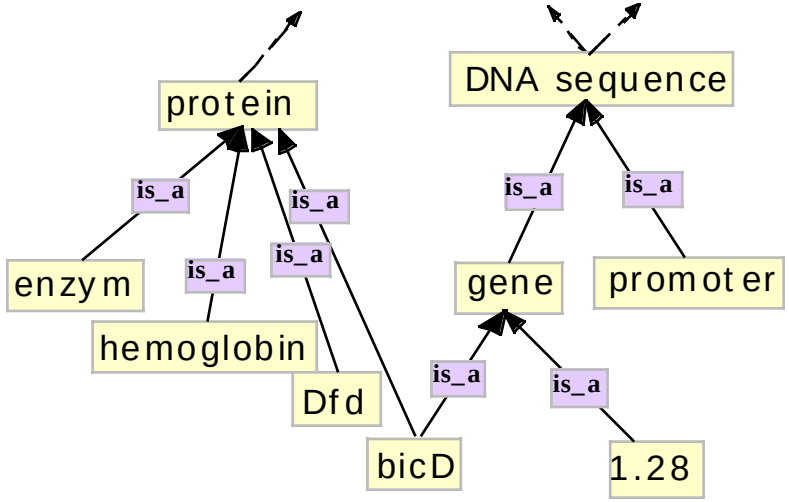
Activation of gene 1.28 by Dfd [...]

Which protein does activate gene 1.28?



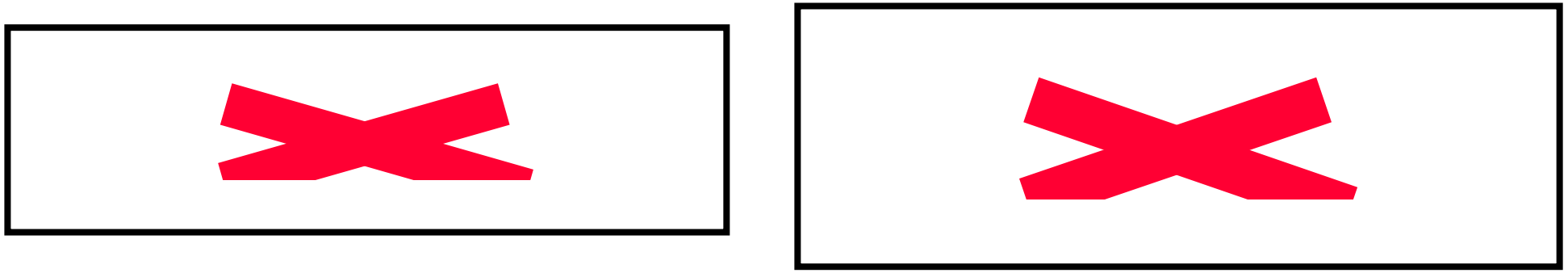
- Additional conceptual knowledge is needed **to interpret the sentences**.



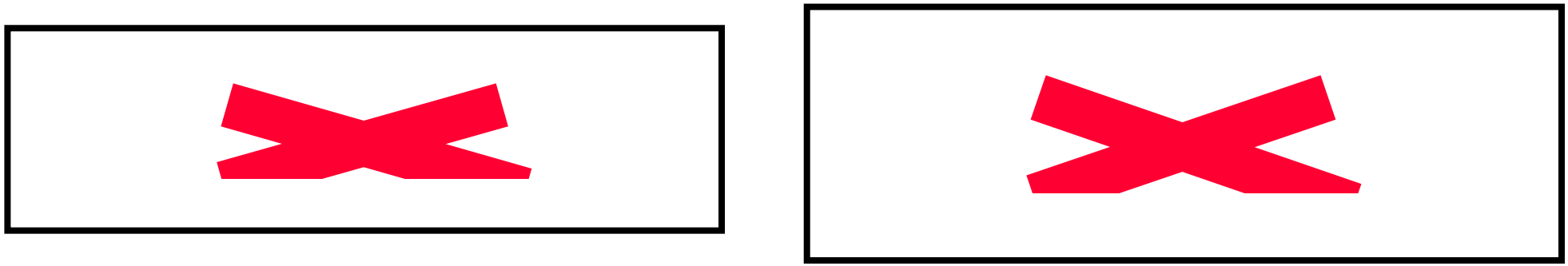


Item to classify: Predicates or Modifiers

- A dual point of view of the examples
  - Objects: predicate; Attributes: modifier



  - Objects: modifier; Attributes: predicate



```
[Dry]
   Dobj          <food>       required
   [(Adj mean)  by <air>         XOR      with <tambour>] optional
   [(Adj duration) during <duration>   XOR for  <duration>]
   optionnel
```

Medline experiment in a keyword based representation

- Total number of "biterms" sentences: 313 classed by biologists.

- 104 / 313 = 33,3 % *with* interaction
- 209 / 313 = 66,7 % *without* interaction

Recall rate: 74 %
Precision rate: 51,7 %

➔ Half of the sentences classed positively are negative.
➔ 1/ 3 of the interactions are recognized.

Recall is OK but precision is very poor.

Example of MedLine abstract

Other Formats: [Citation Format]
Links: [98 medline neighbors] [Journal of Bacteriology]

UI - 98348468
AU - Qi Y
AU - Hulett FM
TI - Role of PhoP approximately P in transcriptional regulation of genes involved in cell wall anionic polymer biosynthesis in bacillus subtilis [In Process Citation]
LA - Eng
DA - 19980801
DP - 1998 Aug
IS - 0021-9193
TA - J Bacteriol
PG - 4007-10
SB - M
CY - UNITED STATES
IP - 15
VI - 180
JC - HH3
AA - AUTHOR

Example of MedLine abstract

AB - tagA, tagD, and tuaA operons are responsible for the synthesis of cell wall anionic polymer, teichoic acid, and teichuronic acid, respectively, in Bacillus subtilis. Under phosphate starvation conditions, teichuronic acid is synthesized while teichoic acid synthesis is inhibited. Expression of these genes is controlled by PhoP-PhoR, a two-component system. It has been proposed that PhoP approximately P plays a key role in the activation of tuaA and the repression of tagA and tagD. In this study, we demonstrated the role of PhoP approximately P in the switch process from teichoic acid synthesis to teichuronic acid synthesis, by using an in vitro transcription system. The results indicate that PhoP approximately P is sufficient to repress the transcription of the tagA and tagD promoters and also to activate the transcription of the tuaA promoter.
AD - Laboratory for Molecular Biology, University of Illinois at Chicago, Chicago, Illinois 60607, USA.
RO - O:099
PMID- 0009683503
SO - J Bacteriol 1998 Aug;180(15):4007-10

```
SUBJ(8@P 9@play)
SUBJPASS(1@it 4@propose)
DOBJ(9@play 12@role)
VMODOBJ(9@play 21@of 24@tagD)
VMODOBJ(9@play 16@of 20@repression)
VMODOBJ(9@play 13@in 15@activation)
ADJ(22@tagA 24@tagD)
ADJ(17@tuaA 20@repression)
   _It has been proposed  that PhoP approximately 8@P plays a key role in the
activation of tuaA and the repression of tagA and 24@tagD .
  [SC  [NP _It NP]/SUBJ :v has been proposed SC]  [SC that [AP PhoP AP]
approximately [NP 8@P NP]/SUBJ :v plays SC] [NP a key role NP]/OBJ [PP in the
activation PP] [PP of tuaA and the repression PP] [PP of tagA and 24@tagD PP] .
NN(11@key 12@role)
NNPREP(20@repression 21@of 24@tagD)
NNPREP(15@activation 16@of 20@repression)
NNPREP(12@role 13@in 15@activation)
NUNSURE([N [NP a key role NP] [PP in the activation PP] [PP of tuaA and the
repression PP] [PP of tagA and tagD PP] N])
NUNSURE([N [NP P NP] N])
```

Learning examples

activation $ of (Nom-Prep-Nom) $ P. $ 1
activation $ of (Nom-Prep-Nom) $ repression $ 1
activation $ of (Nom-Prep-Nom) $ promoter $ 19
activation $ of (Nom-Prep-Nom) $ some $ 1
activation $ of (Nom-Prep-Nom) $ expression $ 8
activation $ of (Nom-Prep-Nom) $ Spo0A $ 1
activation $ of (Nom-Prep-Nom) $ tuaA $ 1
repression $ of (Nom-Prep-Nom) $ tagA $ 1
activation $ of (Nom-Prep-Nom) $ PA3 $ 1
activation $ of (Nom-Prep-Nom) $ phoA $ 1
activation $ of (Nom-Prep-Nom) $ lichenysin $ 1
activation $ of (Nom-Prep-Nom) $ transcription $ 9
activation $ of (Nom-Prep-Nom) $ phoB $ 1
activation $ of (Nom-Prep-Nom) $ pro-sigmaE $ 1
activation $ of (Nom-Prep-Nom) $ RocR $ 1
activation $ of (Nom-Prep-Nom) $ sigma $ 14
activation $ of (Nom-Prep-Nom) $ PrfA $ 1
activation $ of (Nom-Prep-Nom) $ set $ 1
activation $ of (Nom-Prep-Nom) $ regulator $ 1
activation $ of (Nom-Prep-Nom) $ narGHJI $ 1
activation $ of (Nom-Prep-Nom) $ enzyme $ 1
activation $ of (Nom-Prep-Nom) $ FNR $ 1
activation $ of (Nom-Prep-Nom) $ gltC $ 1
activation $ of (Nom-Prep-Nom) $ autoregulation $ 1
activation $ of (Nom-Prep-Nom) $ gene $ 4

activate $ COD $ transcription $ 5
activate $ COD $ e. $ 1
activate $ COD $ promoter $ 5
activate $ COD $ b $ 1
activate $ COD $ expression $ 6
activate $ COD $ catabolism $ 1
activate $ COD $ sequence $ 1
activate $ COD $ phosphorelay $ 2
activate $ COD $ operons $ 1
activate $ COD $ function $ 1
activate $ COD $ 29 $ 1
activate $ COD $ PA3 $ 1
activate $ COD $ gene $ 3
activate $ COD $ map $ 1
activate $ COD $ 86 $ 1