Machine learning for ontology population of a domain specific relation

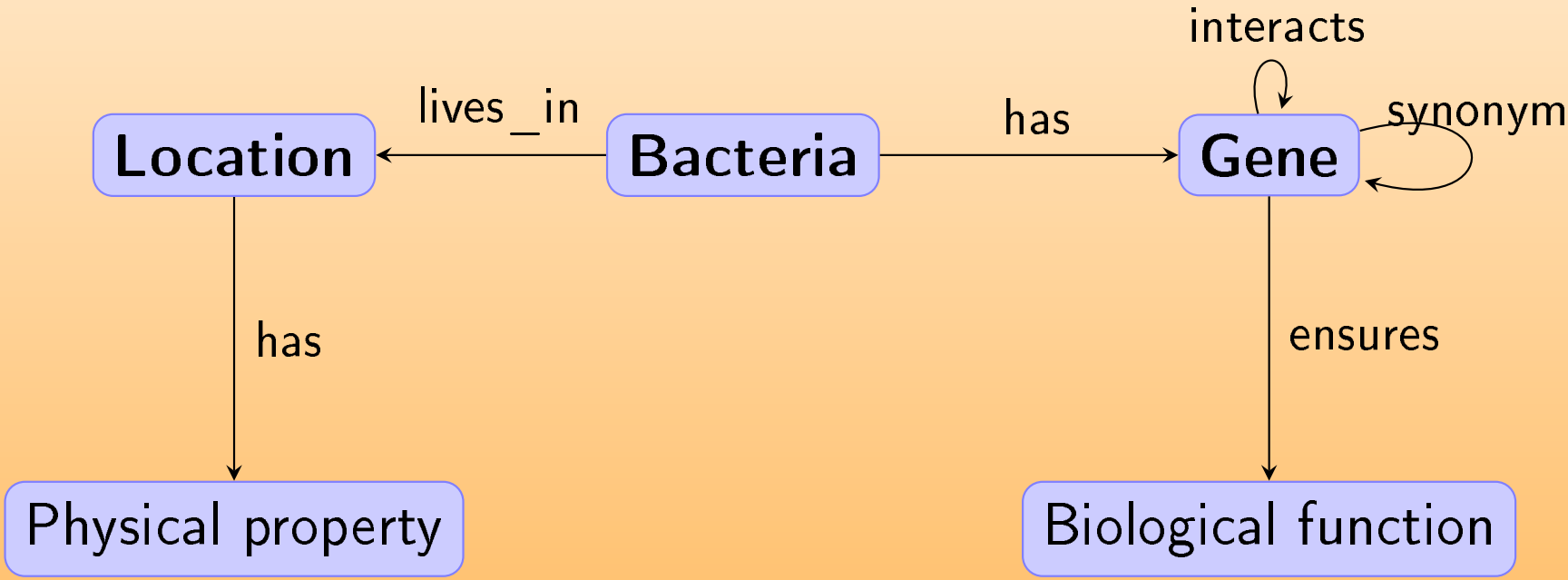
Philippe Veber

Unité Mathématique, Informatique et Génome, INRA

October 18th, 2010

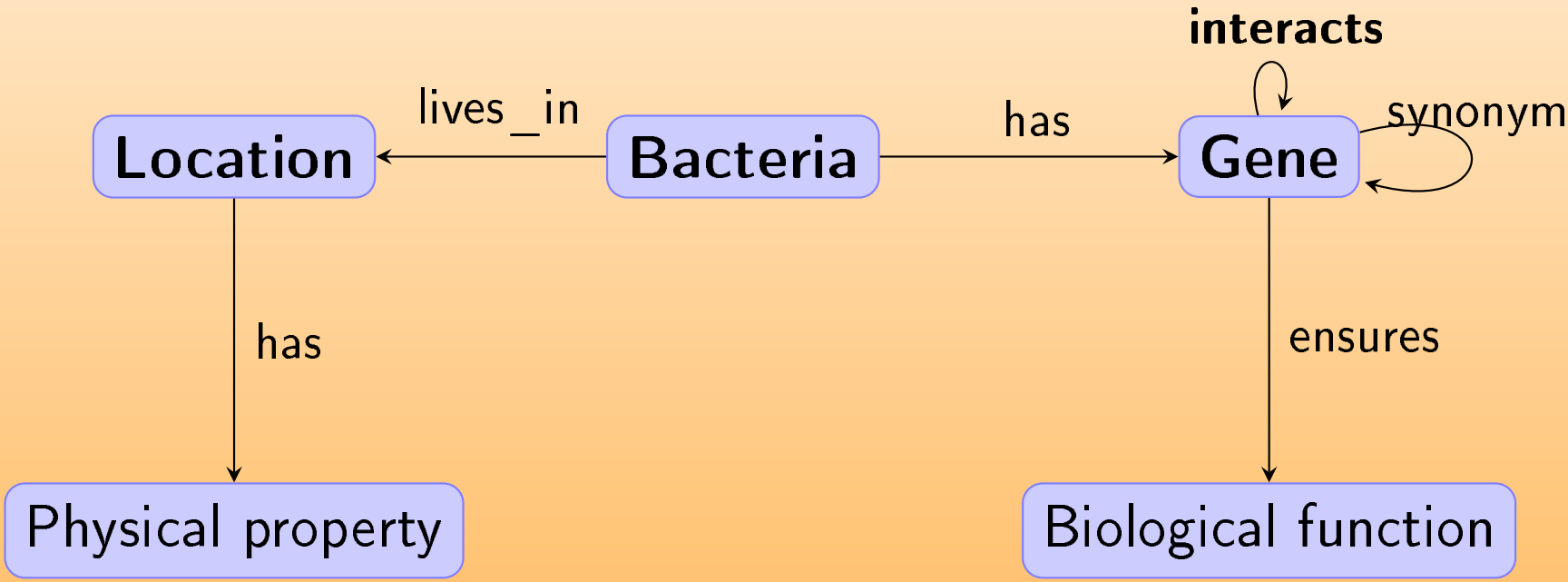
Ontology acquisition with text-mined information

- Application domain: microbiology ontology



Ontology acquisition with text-mined information

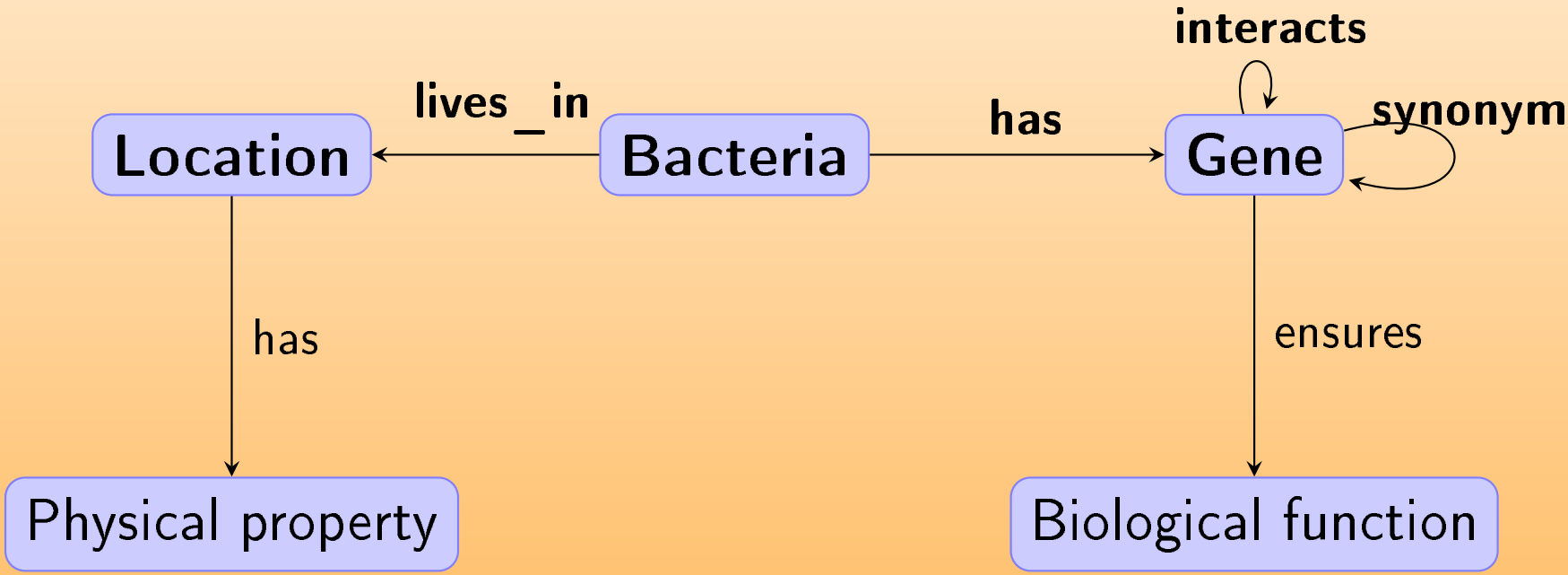
- Application domain: microbiology ontology



- This year's challenge

Ontology acquisition with text-mined information

- Application domain: microbiology ontology



- This year's challenge
- **BIONLP Shared task**
  - international challenge jointly organized with Tsujii Lab, NaCTeM, DBCLS and others
  - open participation!

`http://sites.google.com/site/bionlpst/`

Overview of the challenge (task 3.3)

- Organized by LIPN, based on LLL-Genic Interactions (INRA 2005)
- corpus: 578 examples, 55 sentences from Pubmed abstracts
- rich, manually curated linguistic annotations (named entities, lemmatization, syntactic dependency parsing)
- task: find pairs of interacting genes

> *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 encode sigma K.*

  - Gene names are given as a dictionary

Overview of the challenge (task 3.3)

- Organized by LIPN, based on LLL-Genic Interactions (INRA 2005)
- corpus: 578 examples, 55 sentences from Pubmed abstracts
- rich, manually curated linguistic annotations (named entities, lemmatization, syntactic dependency parsing)
- task: find pairs of interacting genes

> *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 encode sigma K.*

- Gene names are given as a dictionary
- 5 correct answers among 30 pairs,
- pairs are oriented!

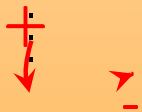
Extraction as a classification task

- supervised classification setting
- instances:

⋮

The expression of rsfA is under the control of both sigma(F) and sigma(G).

The expression of rsfA is under the control of both sigma(F) and sigma(G).



- 102 positive examples, 474 negative examples
- evaluation by 10-fold cross-validation: precision, recall, F-measure
- classification algorithm: SVM

Main issue: find a suitable kernel function

Kernel function using global alignment

Approach:

- represent interacting pairs as paths of words (after lemmatization)

The rocG gene of B. subtilis requires for its expression RocR, a member of . . .

The synthesis of cotD gene product depends on the activity of gerE.

Kernel function using global alignment

Approach:

- represent interacting pairs as paths of words (after lemmatization)

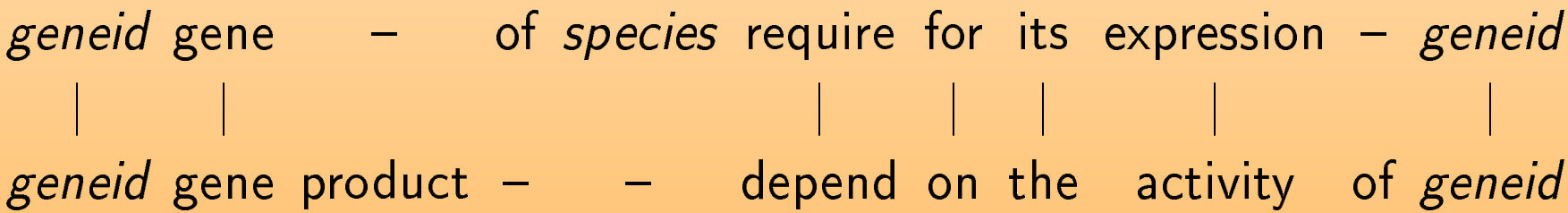
geneid gene of *species* require for its expression *geneid*

geneid gene product depend on the activity of *geneid*

Kernel function using global alignment

Approach:

- represent interacting pairs as paths of words (after lemmatization)
- use global alignment (edit/Levenstein distance)

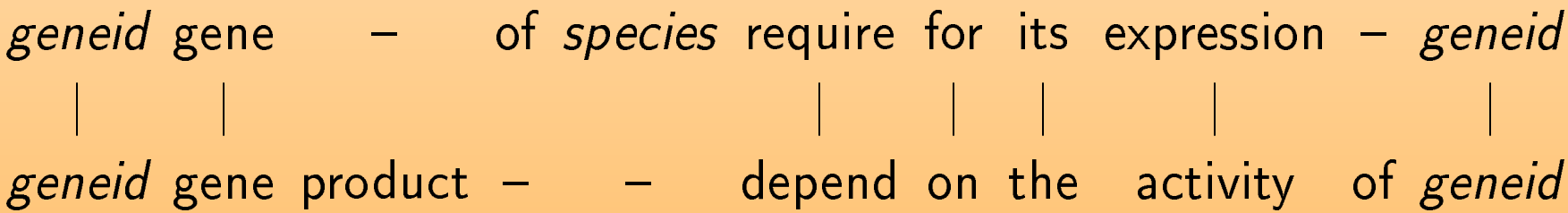


- optimal alignment, given a substitution function and a gap penalty, is easy to compute (dynamic programming)

Kernel function using global alignment

Approach:

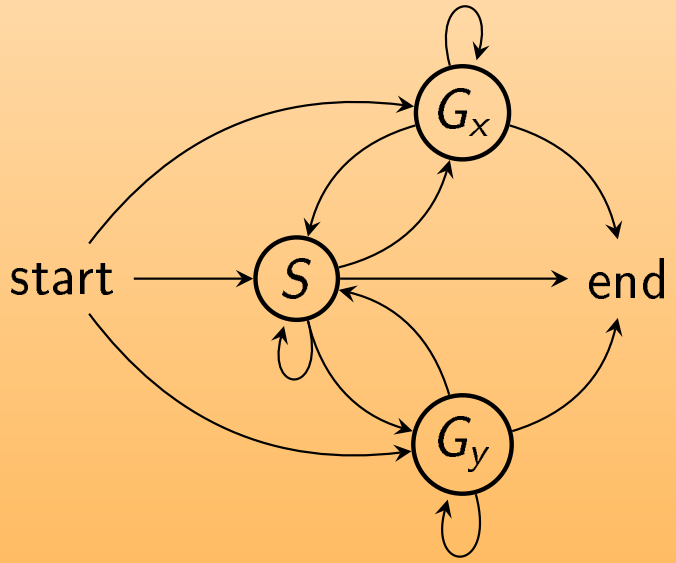
- represent interacting pairs as paths of words (after lemmatization)
- use global alignment (edit/Levenstein distance)



- optimal alignment, given a substitution function and a gap penalty, is easy to compute (dynamic programming)
- however, it is **not a valid kernel function**

Workaround [Watkins 2001]

- a joint distribution $P(X, Y)$ is a kernel if X and Y are conditionally independent.
- in particular, the joint distribution of two sequences x and y under a pair HMM is a kernel.

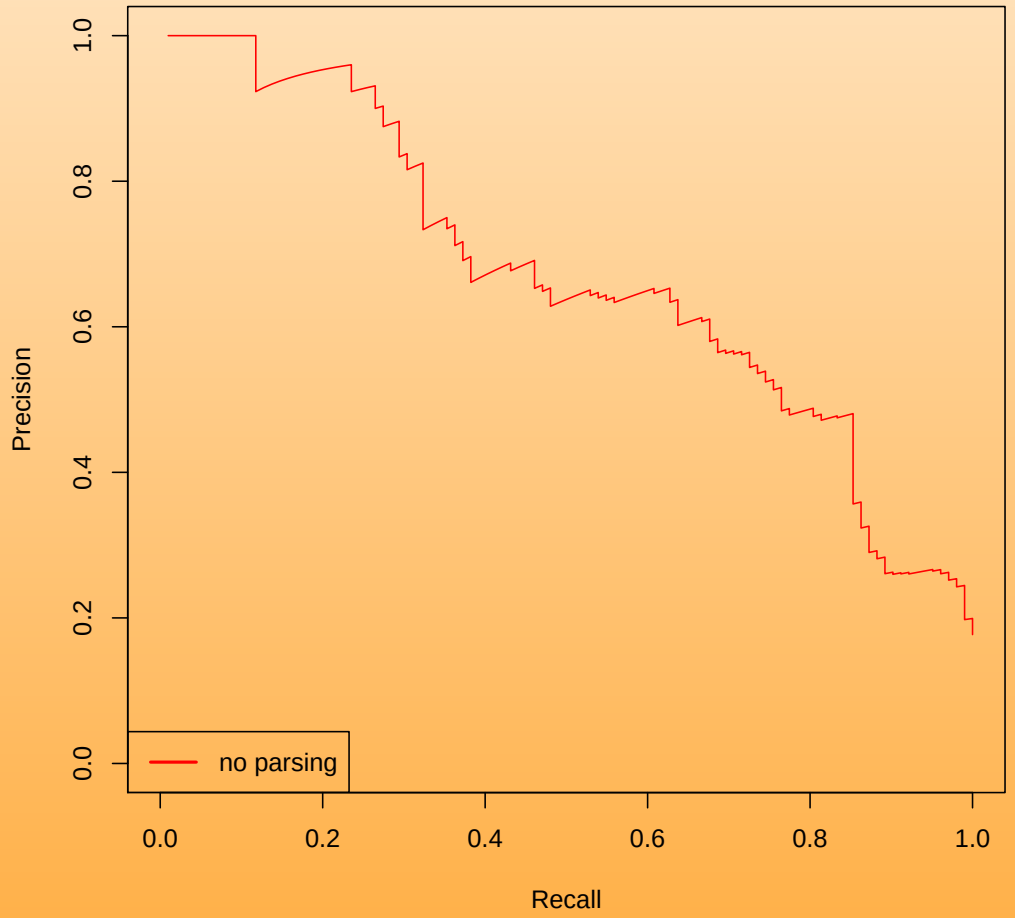


- joint generation of two sequences x and y
- hidden path = global alignment between x and y

- edit distance between x and y $\equiv$ probability of the most probable hidden path that generate x and y
- our kernel is

$$P(X, Y) \equiv \text{sum of alignment scores over all global alignments}$$

First results



Kernel	Rec.	Prec.	F-meas.
no parsing	**0.44**	**0.69**	**0.54**

Using syntactic parsing

- Challenge data include accurate syntactic parsing

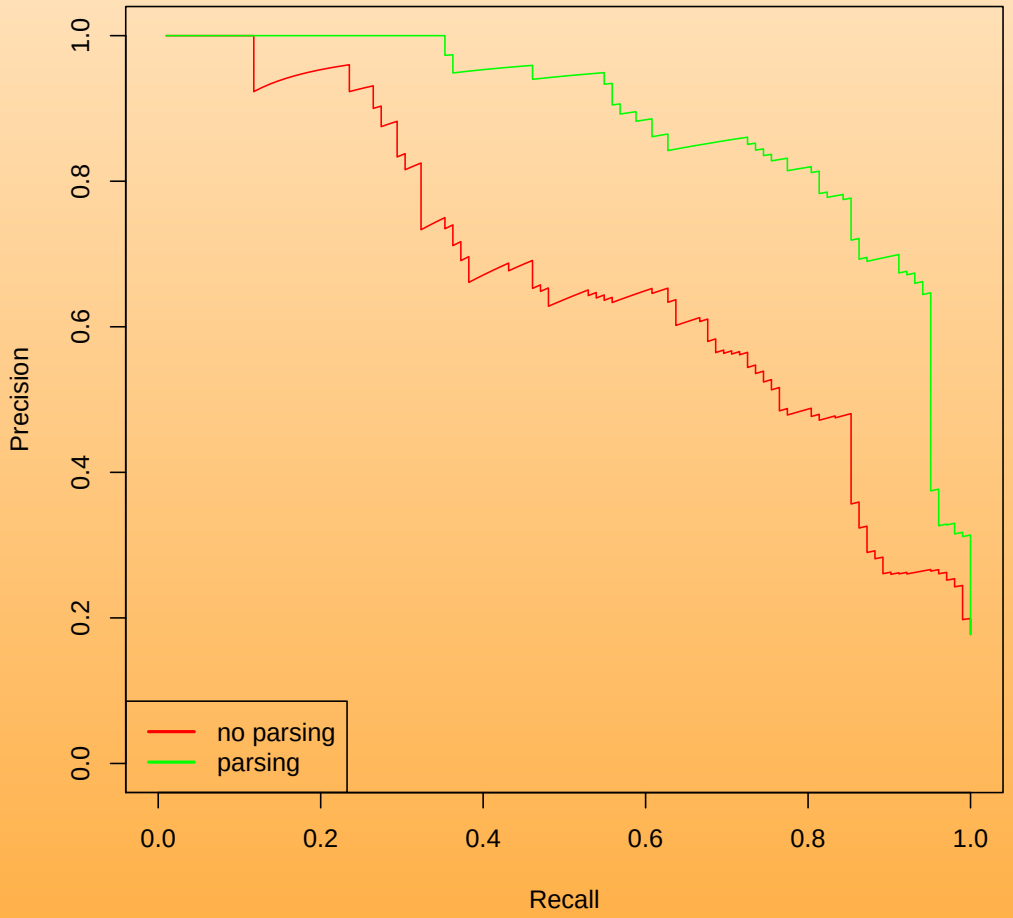
The rocG gene of B. subtilis requires for its expression RocR.

- the links form a forest
  $\Rightarrow$ there is at most one path between any two words

$$\text{rocG} \xleftarrow{\texttt{att:N-N}} \text{gene} \xrightarrow{\texttt{subj:V-N}} \text{requires} \xrightarrow{\texttt{obj:N-N}} \text{RocR}$$

- richer and more concise representation

Results with syntactic parsing



Kernel	Rec.	Prec.	F-meas.
no parsing	0.44	0.69	0.54
parsing	**0.77**	**0.88**	**0.82**

Refinement of the substitution model

- the substitution function is really poor: $\sigma(u, v) = \begin{cases} 1 \text{ if } u = v \\ g < 1 \text{ otherwise} \end{cases}$
- however, closer meaning should imply less/no substitution penalty, as in

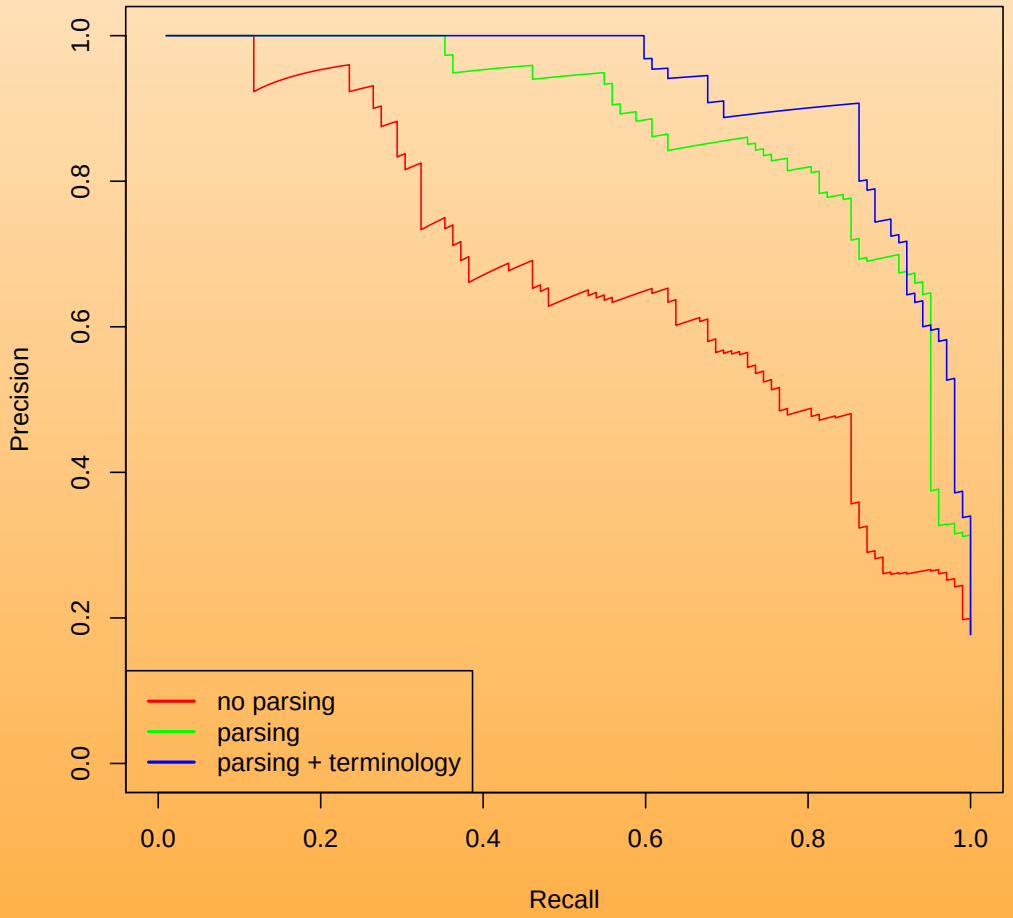
The expression of rocR is influenced by gerE.
rocR transcription is controlled by gerE.

- we manually built a terminology with synonym sets

transcription expression activity synthesis assembly ...	regulate control activate inhibit induce ...	protein factor kinase phosphatase enzyme ...	region promoter sequence element	require need	essential required needed necessary responsible sufficient	...

- 22 synonym sets, 84 words in total, 5 sets with more than 10 words

Results with terminology



Kernel	Rec.	Prec.	F-meas.
no parsing	0.44	0.69	0.54
parsing	0.77	0.88	0.82
parsing + termino	**0.85**	**0.88**	**0.87**

Conclusion

- Challenge submission
  - all previous results on the train dataset
  - on the test set: 0.69 recall, 0.88 precision, 0.77 F-measure
  - encouraging results, alas non comparable to literature (previously unreported errors in the test data)

Conclusion

- Challenge submission
  - all previous results on the train dataset
  - on the test set: 0.69 recall, 0.88 precision, 0.77 F-measure
  - encouraging results, alas non comparable to literature (previously unreported errors in the test data)
- Use case for linguistic techniques/resources in machine learning
  - syntactic parsing as a powerful abstraction for learning
  - (even small) structured terminology is worth building

Conclusion

- Challenge submission
  - all previous results on the train dataset
  - on the test set: 0.69 recall, 0.88 precision, 0.77 F-measure
  - encouraging results, alas non comparable to literature (previously unreported errors in the test data)
- Use case for linguistic techniques/resources in machine learning
  - syntactic parsing as a powerful abstraction for learning
  - (even small) structured terminology is worth building
- Next step
  - of course, using automatic syntactic dependency parsers like Link Grammar and Enju, and cope with alternative parsings
  - rely on less *ad hoc* terminological and conceptual ressources (notably hierarchies)