Modelling gene expression

J. Saric, E. Ratsch, I. Rojas, R. Kania, U. Wittig
European Media Lab Research gGmbH, Heidelberg
http://www.eml-research.de

A. Gangemi
Laboratory for Applied Ontology, ISTC-CNR, Rome
http://www.loa-cnr.it
Overview

1. Introduction and Motivation
2. Preliminary information extraction work
4. Formalization and foundation of the ontology
6. Future work and open issues
Motivation

- Huge effort in the bioinformatics community to build large knowledge bases
- Types of entities recorded in KBs are heterogeneous syntactically, linguistically and conceptually
- Gene Ontology
- Static vs. dynamic knowledge assumption
- Conferences (e.g. PSB Biomedical Ontologies)
- Projects (e.g. Semantic Mining FP6 NoE)
- Use of ontologies for
  - information extraction from text
  - categorization and integration of information in/from different sources
  - inference of facts from available (structured) data
A short introduction to gene expression

1. Transcription: The process where DNA is transcribed into mRNA.
2. Translation: The process where mRNA is translated into a protein.

Key terms:
- TF: Transcription Factor
- DNA
- mRNA
- Protein
- Promoter
First steps (an IE experiment)

- Information extraction of gene regulation networks (details in Saric04, ACL proceedings).
- Case study organism: Yeast.

The system had to answer the questions:

- Which proteins (transcription factors) regulate the expression of which genes?
- Which type of regulation is mentioned (i.e. up-regulation, down-regulation, underspecified)?
- Which is the organism that this takes place in?

Methods:

- Shallow NLP techniques
- Hand-crafted rules detecting linguistic patterns
... the putative gene for *Saccharomyces cerevisiae* riboflavin synthase beta chain ...
Characteristics of the system

- Medline Corpus (MeSH terms)
- Tokenisation and multi-word detection
- Part-of-speech tagging
- Semantic labeling
  - Gene and protein names
  - Cue words for entity recognition
  - Verbs for relation extraction
- Named entity chunking
  - [nxgene The GAL4 gene]

- Relation chunking
  - [nxexpr The expression of [nxgene the cytochrome genes [nxpg CYC1 and CYC7]] is controlled by [nxpg HAP1]]
NLP needs knowledge
Term boundary recognition needs semantics

What are the borders of the following term?
And, how can we re-construct the nested (compositional) structure?

Eg.

5. *Nuclear factor NF-kappa-B p50 subunit* ....
⇒ Need for a terminological dictionary of proteins and protein families with associated protein functions.

8. *Endotoxin increased NF-kappaB p50/p65 heterodimer binding.*
⇒ *heterodimer* presupposes existence of A and B with $A \neq B$:
  a. $A = \text{NF-kappaB}$ and $B = \text{p50/p65}$
  b. $A = \text{p50}$ and $B = \text{p65}$

The a-reading is false, we need to know that p50 and p65 are proteins being part of the complex NF-kappaB.
The built-in *informal* schema
Results overview

- The **precision** of our method is very good
  - 83-90% on relation extraction
  - 97% on named entity recognition
- Evaluating the **recall** is difficult, estimate:
  - ~30% (looking through 250 of 44,354 sentences that contain at least two gene/protein names)

⇒ The quality of our results are not so bad, but …
... some drawbacks

1. Recognising terminology within a text:
   • What is a technical term?
   • What are the boundaries of the term?

2. Categorisation of recognised terms:
   • What is/are the correct semantic category/ies for a recognised term?
   • The categorisation of the terms cannot be easily done in a compositional way (nestedness & scalability)?
   • Although the template (and pattern) construction reflects an underlying ontology on gene expression, it is hard-wired (implicit).

3. Scalability: although we used rules for related questions (i.e. protein interaction), the scalability of the system is limited.
In order to overcome these drawbacks: create a more detailed and complete ontology that acts as a backbone for the NLP system -- and also for database design, population, and integration --
Basic types and rationale

- DOLCE axiomatic theory (*Descriptive Ontology for Linguistic and Cognitive Engineering*): [http://www.loa-cnr.it](http://www.loa-cnr.it)
- ≈10 basic types, ≈20 basic relations, ≈200 axioms
- Wide-range application: Law, Fishery, Finance, Anatomy, ...
- *Very preliminary* application in biology
- Foundational types use from DOLCE: *Substance, Process, Collection*
- Foundational (formal) relations used from DOLCE+: *(Proper)Part, Component, Member, Participation, Connection, Succession*
- Substance types are considered: dna and rna sequence, gene, peptide, protein, nucleotide, aminoacid, etc.
- 3 process types are considered: transcription, RNA processing, translation
Some axioms. Sequences, parts and collections

- \( \text{Sequence}(x) = \text{Substance}(x) \land \forall y,z. (\text{Part}(x,y) \land \text{Part}(x,z)) \rightarrow \text{TransitiveConnection}(y,z) \land \exists j,k. \text{Part}(x,j) \land \text{Part}(x,k) \land \text{StrongConnection}(j,k) \land \text{DirectSuccessor}(j,k) \)

- \( \text{Sequence}(x) \rightarrow \forall y,z. (\text{Part}(x,y) \land \text{Part}(x,z)) \rightarrow \neg (\text{Successor}(y,z) \land \text{Successor}(z,y)) \)

- \( \text{dnaSequence}(x) \rightarrow \forall y. \text{PartOf}(y,x) \rightarrow \text{Deoxyrybosenucleotide}(y) \)

- \( \text{Gene}(x) \rightarrow \forall y. \text{PartOf}(y,x) \rightarrow (\text{dnaSequence}(y) \lor \text{Deoxyrybosenucleotide}(y)) \)

- \( \text{Gene}(x) \rightarrow \exists c,n,o. \text{CodingSequence}(c) \land \text{NonCodingSequence}(n) \land (\neg (c \oplus n), x) \land \text{Organism}(o) \land \text{in}(x,o) \land \neg \exists z. \text{ComponentOf}(z,c) \land \text{ComponentOf}(z,n) \)
Other axioms. Processes, time, roles.

- Transcription\((x) \rightarrow \text{ChemicalReaction}(x) \land \exists g,o,prom,ts,gt,enz,tf,compl.\)
  Gene\((g) \land \text{in}(g,o) \land \text{Substrate}(x,g) \land \text{Promoter}(prom) \land \text{Substrate}(x,prom) \land \text{TerminationSequence}(ts) \land \text{Substrate}(x,ts) \land \text{Transcript}(gt) \land \text{Product}(x,gt) \land \text{rnaPolymerase}(enz) \land \text{Catalyzer}(x,enz) \land \text{TranscriptionFactor}(tf) \land \text{Regulator}(x,tf)\)

- Translation\((x) \rightarrow \text{ChemicalReaction}(x) \land \exists mr,tr,rib,pep. \text{mRNA}(mr) \land \text{TemplateFor}(mr,x) \land \text{tRNA}(tr) \land \text{Substrate}(x,tr) \land \text{Ribosome}(rib) \land \text{Catalyzer}(x,rib) \land \text{Peptide}(pep) \land \text{Product}(x,pep)\)

- TemplateFor\((x,y) \rightarrow \text{mRNA}(x) \rightarrow \forall z,w,pep. [\text{Codon}(z) \land \text{Component}(x,z) \land \text{Aminoacid}(w) \land \text{Peptide}(pep) \land \text{Component}(pep,w) \land \text{Product}(y,pep)] \rightarrow \text{Maps}(w,z)\)

- Meets\((x,y) \rightarrow \exists t_1,t_2. \text{Loc}(x,t_1) \land \text{Loc}(y,t_2) \land t_1<t_2\)

- Translation\((x) \rightarrow \exists y. \text{Transcription}(y) \land \text{Meets}(x,y)\)
Foundational issues

- Gene as a “knowledge object”: functional collection, what unity criterion? (Inferred from transcript results? Characters? Evolutionary constraints?)
- Gene for an organism: type or token? What is the prototypical gene, given individual variability? Similarly for genome:
  \[ \text{Genome}(x) \rightarrow \exists y. \text{Organism}[\text{type}](y) \land \forall z. \text{Gene}(z) \land \text{in}[\ast](z,y) \rightarrow \text{Member}(x,z) \]
- **Formal vs. material relations:** e.g. connection vs. covalent binding
  - Two different layers in the ontology?
  - Sequences are at the functional or at the substantial layer?
- How to formalize interaction btw different layers/systems?
  - E.g. membrane topology and gene processes
  - E.g. gene functional sequences and protein biochemical structure
- Should we be engaged in these issues?
Further work: Ontology design patterns for functional ontologies