Improving Text Mining with Controlled Natural Language: A Case Study for Protein Interactions

Tobias Kuhn (speaker)
Loïc Royer
Norbert E. Fuchs
Michael Schroeder

DILS'06, Hinxton (UK)
21 July 2006

Cooperation of

University of Zurich

(Norbert E. Fuchs, Tobias Kuhn)

and

TU Dresden

(Loïc Royer, Michael Schroeder)

Introduction

- Biomedical literature is growing at a tremendous pace
- *PubMed* contains 16 million articles and grows by over 600'000 articles per year
- Computational support is needed!

Today's Solution

The β-adaptin clathrin adaptor interacts with the mitotic checkpoint kinase BubR1

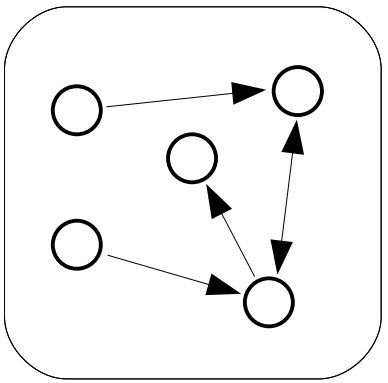
Corinne Cayrol, Céline Cougoule, Michel Wright

Abstract

The adaptor AP2 is a heterotetrameric complex that associates with clathrin and regulatory proteins to mediate rapid endocytosis from the plasma membrane. Here, we report the identification of the mitotic checkpoint kinase BubR1 as a novel binding partner of beta2-adaptin, one of the AP2 large subunits. Using two-hybrid experiments and in vitro binding assays, we show that beta2-adaptin binds to BubR1 through its amino-terminal beta2-'trunk' domain, while the beta2-binding region of BubR1 maps to the carboxy-terminal kinase domain. Subcellular immunolocalization studies suggest that the interaction between BubR1 and beta2-adaptin could take place in the cytosol at any time during the cell cycle. In addition, we found that BubR1 and the BubR1-related kinase, Bub1, also bind to beta-adaptins of other AP complexes. Together, these results support a model in which the mitotic checkpoint kinases BubR1 and BuB1, by binding to beta-adaptins, may play novel roles in the regulation of vesicular intracellular traffic.

Keywords: Protein interactions; Two-hybrid; Vesicular traffic; Adaptor protein; Protein kinase; Mitotic checkpoint.

NLP, manual annotation



Our Approach

- Let the researchers express their own results in a formal language
- Perfect processing of scientific results by computers
- This formal language has to be ...
  - easy to learn and understand
  - expressive enough to express even complicated scientific results

Knowledge Representation Languages

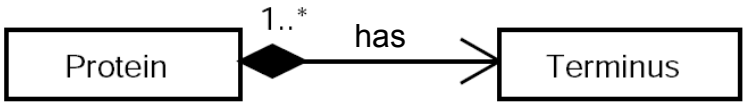
OWL with RDF/XML

```
<owl:Class rdf:ID="Protein">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#has"/>
      <owl:someValuesFrom rdf:resource="#Terminus"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

ACE

Every protein has a terminus.

UML



Description Logics

$$Protein \sqsubseteq \exists has.Terminus$$

first-order logic

$$\forall X(protein(X) \rightarrow \exists Y(terminus(Y) \wedge has(X,Y)))$$

Attempto Controlled English (ACE)

- Formal language that looks like natural English
- Unambiguously translatable into first-order logic
- Restricted grammar
- Unlimited vocabulary
- `www.ifi.unizh.ch/attempto`

Formal Summaries

The β2-adaptin clathrin adaptor interacts with the mitotic checkpoint kinase BubR1

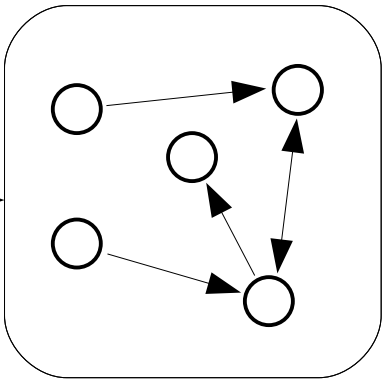
Corinne Cayrol, Céline Cougoule, Michel Wright

Abstract

The adaptor AP2 is a heterotetrameric complex that associates with clathrin and regulatory proteins to mediate rapid endocytosis from the plasma membrane. Here, we report the identification of ...

Keywords: Protein interactions; Two-hybrid; Vesicular traffic; Adaptor protein; Protein kinase; Mitotic checkpoint.

ACE Summary: Beta2-Adaptin binds BubR1 in Yeast-Two-Hybrid. A trunk-domain of Beta2-Adaptin interacts-with BubR1. Bub1 interacts-with the trunk-domain of Beta2-Adaptin. Bub1 interacts-with every beta-sheet of AP and BubR1 interacts-with every beta-sheet of AP.



Formal Summaries

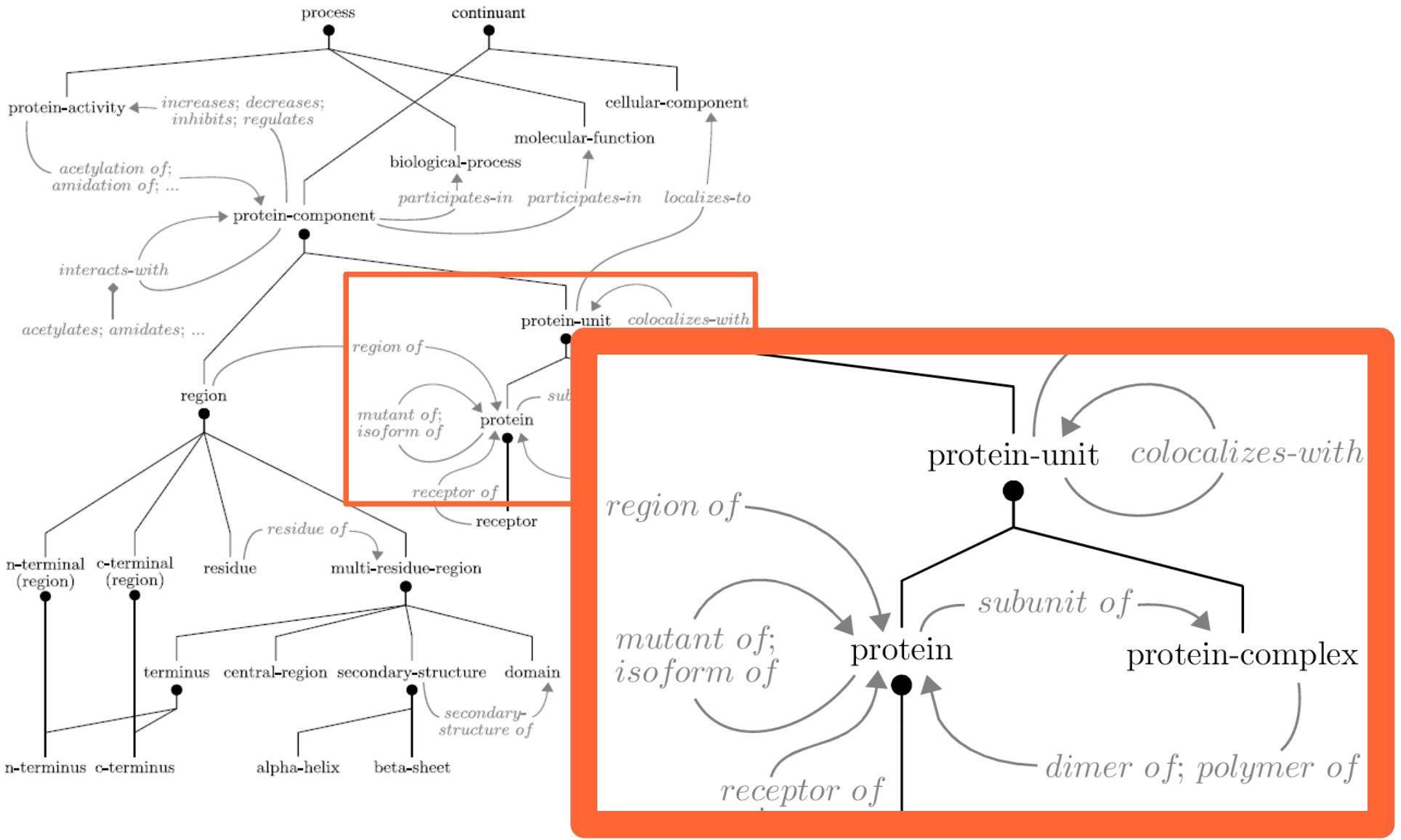
ACE text

BubR1 interacts-with a trunk-domain of Beta2-Adaptin.

Logical representation (DRS)

```
[A, B, C, D]
named(A, BubR1)-1
object(A, atomic, named_entity, object, cardinality, count_unit, eq, 1)-1
named(B, Beta2-Adaptin)-1
object(B, atomic, named_entity, object, cardinality, count_unit, eq, 1)-1
object(C, atomic, trunk-domain, unspecified, cardinality, count_unit, eq, 1)-1
relation(C, trunk-domain, of, B)-1
predicate(D, unspecified, interact_with, A, C)-1
```

Ontology for Protein Interactions

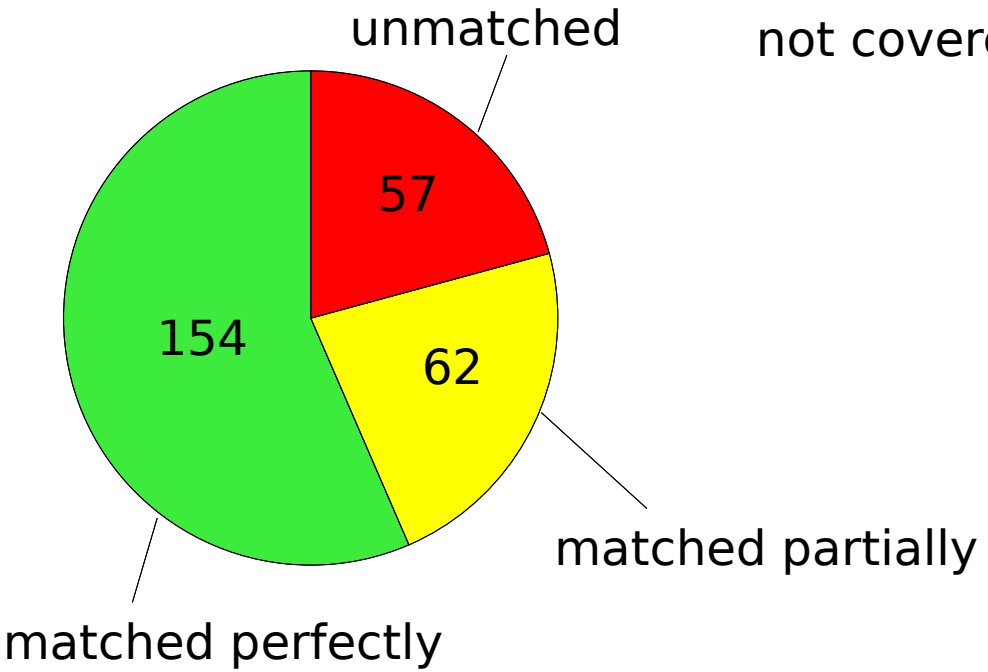


Empirical Study

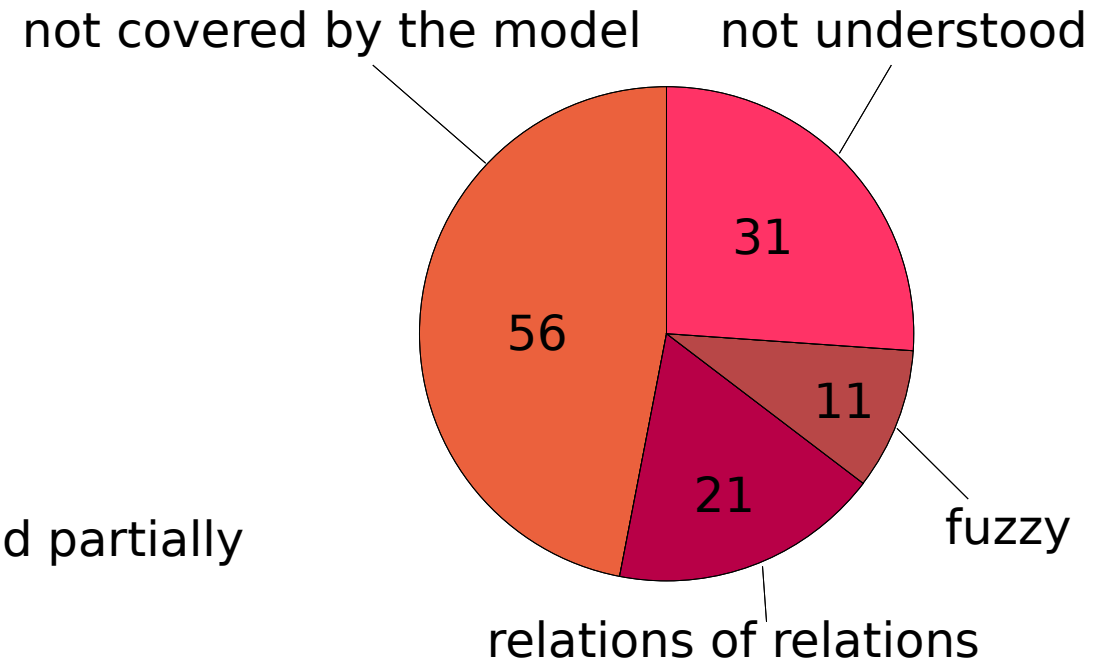
- "How suitable is ACE together with our ontology to express scientific results of protein interactions?"
- Manual translation of 273 facts about protein interactions
- These facts are subheadings of the "Results"-sections of 89 articles (journals by *Elsevier*)

Empirical Study

Total:



Non-perfect:

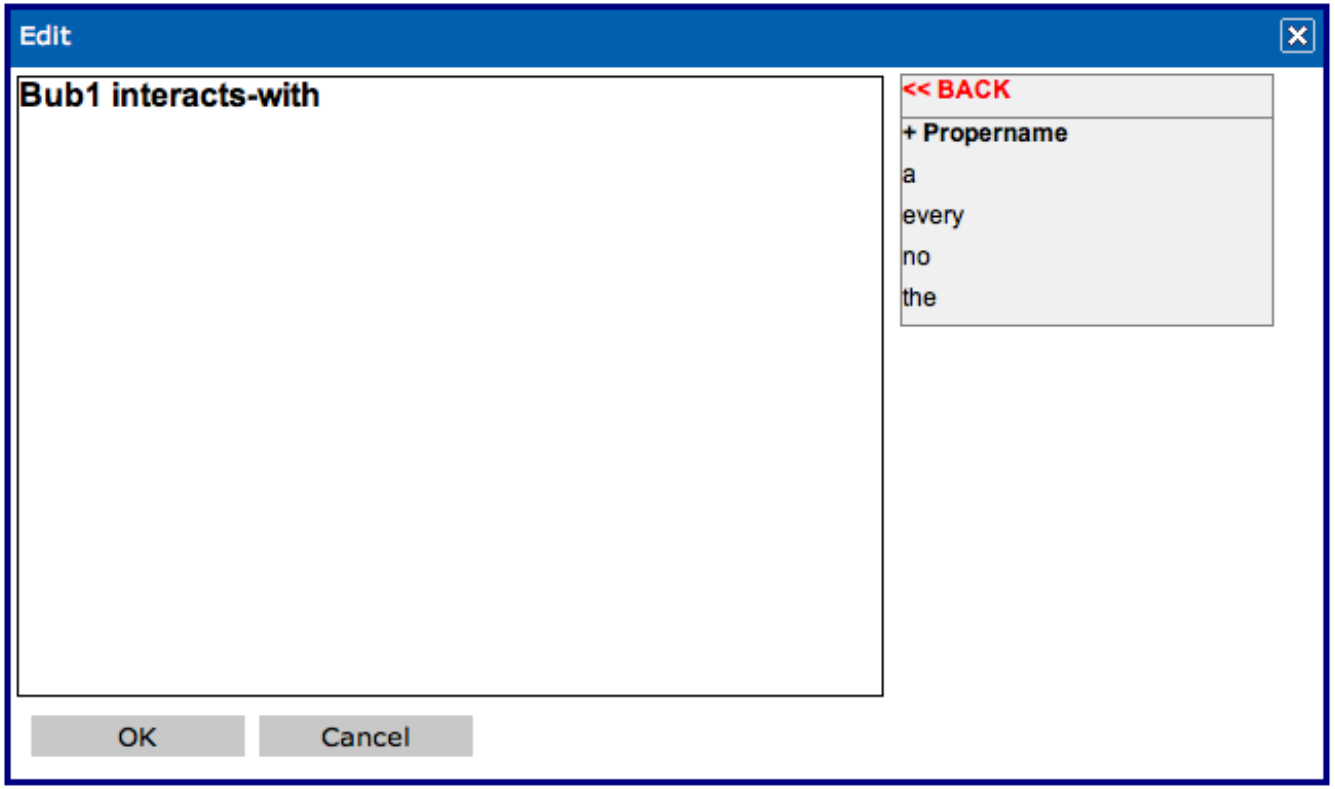


Authoring tool

- Helps writing ACE sentences
- Shows step by step the possible continuations of the sentence
- New words can be created on-the-fly
- Awareness of the underlying ontology
- The users do not need to know the details of the ACE syntax and of the underlying ontology

Authoring tool: Prototype demo

http://gopubmed.biotec.tu-dresden.de/AceWiki/



Benefits of our Approach

- Consistency / redundancy checks
  - “Is there a paper that contradicts my results?”
  - “Is there a paper that comes to the same or similar results?”
- Answer extraction
  - “Which proteins interact with a certain domain of protein X?”
- Automatically updated knowledge bases
  - “Give me an overview of the relations of a protein X to other proteins!”

Conclusions

- Formal summaries for scientific articles can make text mining easier and more powerful
- ACE combines the power of ontologies with the convenience of natural language
- Let the researchers formalize their own results!

Thank you for your attention!

Questions
&
Discussion

Subheadings: Example

2. Results

2.1. YETI binds specifically to both KHC and KLC in the yeast two-hybrid system

To identify proteins that may help kinesin perform its intracellular functions, yeast two hybrid screens were performed to identify *Drosophila* proteins that bind to the kinesin tail domain. Amino acids 675-975 of the kinesin heavy chain and the tetratricopeptide (TPR) motifs of the kinesin light chain were subcloned into the LexA DNA-binding domain plasmid pEG202 and used to screen the RELY1

Degree of Matching: Examples

- Matched perfectly:
  - *Interaction of Act1 with TRAF6*
  - → *Act1 interacts-with TRAF6.*
- Matched partially:
  - *The mtFabD protein is part of the core of the FAS-II complex*
  - → *MtFabD is a subunit of FAS-II.*
- Unmatched:
  - *Cav1 interacts differentially with distinct Dyn2 forms*

Reasons for Non-perfect Matching: Examples

- Not covered by the model:
  - *Daxx Potentiates Fas-Mediated Apoptosis*
- Relations of relations:
  - *Kal-GEF1 activation of Pak does not require GEF activity*
- Fuzzy:
  - *ANKRD1 contains potential CASQ2 binding sequences located in both its NT- and CT-regions*
- Not understood:
  - *hSrb7 does not interact with other nuclear receptors*