Extracting protein-interactions from AIMed using a general-purpose knowledge extraction tool

Michael Gabilondo

CAP 6640, 2011

Table of Contents

Motivation

- Interest from biomedical community in creating structured databases from the knowledge in biomedical publications, both manually and automatically
  - Do two proteins/genes physically interact?
- " The **Fos** and **Jun** families of eukaryotic transcription factors *heterodimerize* to form complexes capable of binding 5 "-TGAGTCA-3" DNA elements . "
  - Output the unordered pair, (Fos, Jun)

Approach

- Construct a general purpose tool (SI) for extracting knowledge from specific domains
- Define predicates (verb senses) for only the relations we are interested in acquiring
  - PPI verbs: activate, bind, phosphosphorylate, cooperate, coprecipitate, etc.

1. Parse sentences with Stanford parser
2. Transform parser output to clause-level structure
3. Use verb-predicates to determine the the meaning of the verb and its arguments
4. Extract the protein-interactions from the semantically annotated clauses

Outline

MCR

- "A Minimal Reconstruction of Clause Structure from Constituent Parse Trees" (Millward, Gomez, 2010)
- The MCR outputs a list of **clauses** from the parse tree, each corresponding to a **main verb** of the sentence. The MCR tells us
  - whether the clause is in active voice or passive voice
  - the grammatical (pre-verbal) subject
  - the post-verbal relatives, PPs and and NP complements

MCR post-processing

- The verb predicate roles (in the SI) can override attachment decisions made by the parser, i.e., a PP attached to an NP can be made to be attached to the verb instead
  - For each NP, the post-processor finds all attached PPs and outputs them as a flat list, potentially attached to the NP
  - The indices to the parse tree for each PP and all other constituents are also output
- The MCR post-processor also takes in a list of verbs which can have nominalizations, and creates a clause structure for each potential nominalization
- **Next slide:** "The extracellular domain of the human **neurotrophin TRKB receptor** expressed in Chinese hamster ovary cells is a highly glycosylated protein , possessing binding ability for **brain-derived neurotrophic factor (BDNF)** . "

MCR Output

This clause can be read as "p1 possesses binding ability for p2"

```
(possessing-34
(VERB (MAIN-VERB possessing possess) (VERB-TYPE VERB)
      (VOICE ACTIVE) (TENSE VBG))
(PP
  (ID 38)
  (PREP (IN for))
  (NP
    (NN
      brain-derived_neurotrophic_factor_GP_25_26_LPAREN_BDNF_GP_27_28_RPAREN)))
(OBJECTS-0 (ID 35) (NP (NN binding) (NN ability)))
(SUBJECT-0 (ID 2) (NP (DT The) (JJ extracellular) (NN domain)))
(PP
  (ID 7)
  (PREP (IN of))
  (NP (DT the) (JJ human) (NN neurotrophin_TRKB_receptor_GP_6_9)))
(RELATIVE (ID 15) (CONJ ) (CLAUSE expressed-15))
```

Outline

Example Sentence and Predicate

The **Fos** and **Jun** families of eukaryotic transcription factors heterodimerize to form complexes capable of binding 5 "-TGAGTCA-3" DNA elements .

```
(interact
    (verbs interact)
    (theme (gr (subj)) (sr thing))
    (cotheme (gr (pp (prep with))) (sr thing))
    (parents action))

; To form a dimer from two different monomers
; form 1. THEME heterodimerize with COTHEME
(heterodimerize
    (verbs heterodimerize heterodimerizes)
    (parents interact))
```

Example SI Output

```
The Fos and Jun families of eukaryotic transcription factors heterodimerize to
form complexes capable of binding 5 "-TGAGTCA-3" DNA elements .

(SI
  (heterodimerize-16
    (pred verb-ont heterodimerize)
    (theme
      (np
        (AND
          (np (senses (protein (mod The) (head Fos_GP_1_2))))
          (np (senses (thing (mod eukaryotic transcription) (head factors))))
          (np (senses (thing (mod Jun_GP_3_4) (head families)))))))
    (purpose (rel (conj (TO to)) (clause form-21))))

  (form-21
    (pred verb-ont nil)
    (obj-0 (np (senses (complex (mod ) (head complexes)))))
    (obj-1 (np (senses (thing (mod ) (head capable)))))
    (subj
      (np
        (AND
          (np (senses (protein (mod The) (head Fos_GP_1_2))))
          (np (senses (thing (mod eukaryotic transcription) (head factors))))
          (np (senses (thing (mod Jun_GP_3_4) (head families))))))))))
```

Step 1: Disambiguate the verb

- For each MCR clause, the SI attempts to "instantiate" each *candidate* verb predicate with the clause:
  - For each **role** of the predicate, it finds a constituent in the MCR clause that matches the role's grammatical relation (GR) and selectional restriction (SR)
  - The predicate with the most roles satisfied is chosen as the meaning of the verb
  - This step also assigns senses to the head nouns of the arguments, since the SR constrains their meaning
- There may be more than one top predicate. The output is a set of instantiated predicates, each of which
  - represents a distinct verb meaning
  - has its own set of roles and NP head senses, as determined by the predicate

Step 2: Attach PP/Relatives to NPs of roles

- The constituents that have been mapped to roles from (Step 1) do not have PPs attached
- Some nouns in the ontology can subcategorize for a preposition
  - We look for such nouns in the instantiated predicate and find PPs to attach to them, further overriding attachment decisions
  - We do this under the assumption that the senses of head nouns chosen in (Step 1) have been disambiguated (may not be the case, depending on ontology and predicate definition)
  - The subcategorization definition also allows for an SR, which can choose (disambiguate) the sense of the head NP of the attaching constituent
- Next, we find the rest of the PPs/relatives that the parser attached to the role constituents

Step 3: Handle adjuncts and left-over constituents

- Adjuncts attach to the verb, but they are not arguments, since they can appear with almost any verb
  - "**In these cells** the adaptor protein Grb2 constitutively *binds* a substantial fraction of c-Cbl through the N-terminal SH3 domain . "
- The adjunct "roles" are chosen and attachments put back, by running (Step 1) and (Step 2) again
- Any remaining constituents are assigned grammatical roles, such as subj, obj and pp
  - **For clauses without matching predicates, this is the only thing we do**
  - This step automatically converts GRs for passive voice, without relying on predicates; e.g., the first by-PP is made the subj if the sentence is passive, but it is a pp if the sentence is active

Outline

Knowledge Extractor

- The knowledge extractor must extract the protein-interactions from the SI output
- The predicates are organized to make this easy
  - e.g., for "interact" predicates, the two proteins should be found in the theme and cotheme (currently 10 predicates of this type)
- The meaning of some clauses is determined by one of its arguments instead of the main verb (e.g., "**RbAP469** and **simian virus 40 T antigen** have homologous **Rb-binding** properties")
  - All such nouns (e.g., binding-property) are classified as interaction-property in ontology to facilitate the knowledge extraction

KE Challenges

- Current KE only gets interactions if they occur in *two different arguments*, and it does *not traverse relatives* (e.g., misses PPs attached to NOMs). Examples of things it misses:
  - "14-3-3_zeta negatively regulates raf-1 activity by interactions with the Raf-1 cysteine-rich domain . "
  - "A physical interaction between CDC37 and CDK4 ..."
  - "... of the IL-6D_GP_22_23_LPAREN_sIL-6R_GP_24_25_RPAREN 2 complex to couple with..."
- If a predicate is not defined, it relies on one of the arguments being an interaction-property, and tries to extract pairs from that argument and another argument, as determined by the rule

Evaluation: Corpora

- AIMed is a corpus of biomedical abstracts containing 1955 sentences and 1000 protein-interactions
- The verb predicates and the ontology were developed for AIMed
- The resulting system was also evaluated on AIMed
  - TP: 249, FP: 15, FN: 752
  - Precision (tp / (tp + fp)): **0.943**
  - Recall (tp / (tp + fn)): **0.249**

Sources of Error

- Analyzed 14 random sentences where an interaction was missed
- KE unable to extract if output is correct (6)
- MCR (or post-processor) errors (6)
  - Misses apposition (3)
  - Misses attached PP or relative (2)
  - Wrong subject (1)
- SI Errors (5)
  - No predicate defined where there should be (at least 2)
  - SI failed to attach PP to NP (1)
  - misses "its" which refers to something in front (1)
  - Unhandled constructions (1)
    - p1 and p2 receptors (p3 and p4, respectively)