Automated Extraction of Information on Protein-Protein Interactions From The Biological Literature

Toshihide Ono, Haretsugu Hishigaki, Akira Tanigami and Toshihisa Takagi

勤沛儒
Bioinformatics Program
Institute of Genetics
National Yang Ming University

Importance of Studying Protein-Protein Interaction

- Understanding in biological process

Importance of Studying Protein-Protein Interaction

- DNA replication
  - DnaB-DnaC complex
- Transcription
  - Transcription Factor
- Metabolic pathway
  - a-ketoglutarate dehydrogenase complex
- Signaling pathway
  - Insulin-IRa
- Cell cycle control
  - Cyclin-CDK

Database of Protein-Protein Interaction

- DIP(Database of Interacting Proteins)
- FlyNets
  - *Drosophila*
- MIPS
  - *Saccaromyces*
- EcoCyc
  - Metabolic pathway of *E. coli*
- KEGG
  - Map of metabolic pathway
- All assembled manually!

Difficulties of Extracting Information From Scientific Literature

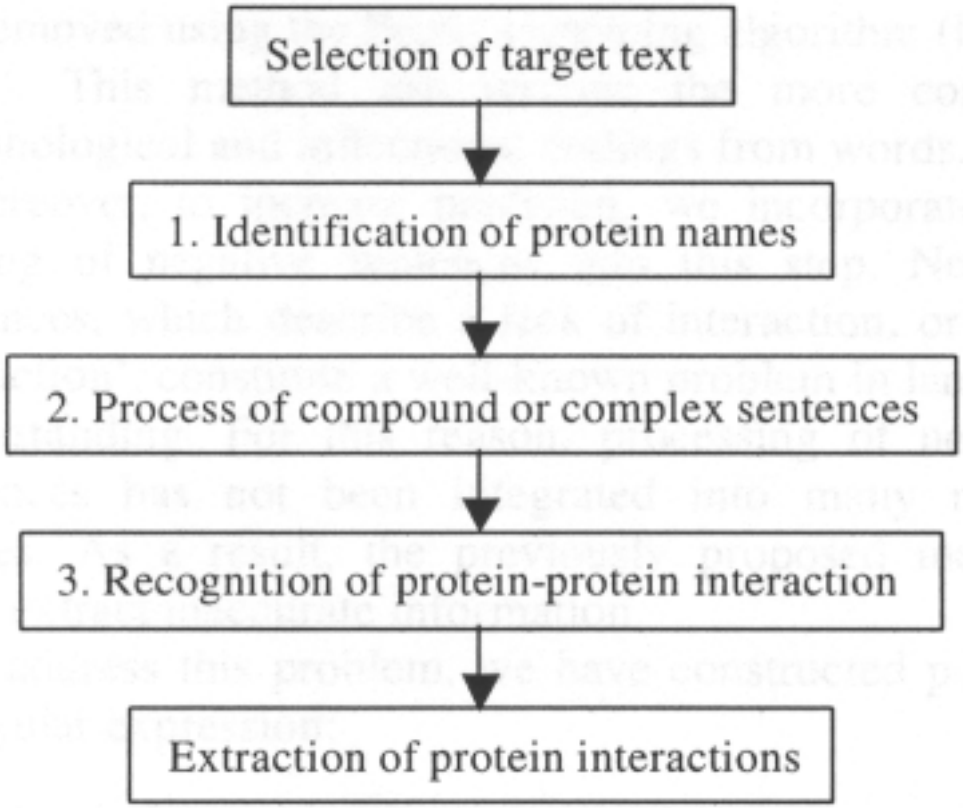
- Written by natural language
  - Collection data manually will take too much time and labor
- Extraction of information by computer
  - Artificial Intelligence (AI)
  - Natural Language Process (NLP) technique
    - Semantic and discourse analysis
  - Too complex to handle!

Previous study on automated protein-protein interacted-information extraction

- Sekimizu *et al.* (1998)
  - Determine candidate noun phrases in the surrounding text
  - Precision rate: 67.8%~83.3%
- Blaschke *et al.* (1999)
  - Simple match
- Tomas *et al.* (2000)
  - HighLight, a general-purpose information extraction engine
  - Precision rate: 77%

The methodology developed by authors to extracting information efficiently

- Part-of-speech rule
  - Grammar analysis
- Pattern match
  - Identification of protein name and keyword


Selection of target text
1. Identification of protein names
2. Process of compound or complex sentences
3. Recognition of protein-protein interaction
Extraction of protein interactions

Step 1. Identification of protein names

- Creating dictionary manually
  - Contain protein name entries
  - Yeast protein name was derived from SGD
    - 6084 molecules and 16,772 synonyms
  - *E. coli* protein name was constructed using K-12 data
    - 4405 entries
- Pattern match method
  - Match with entries in the dictionary

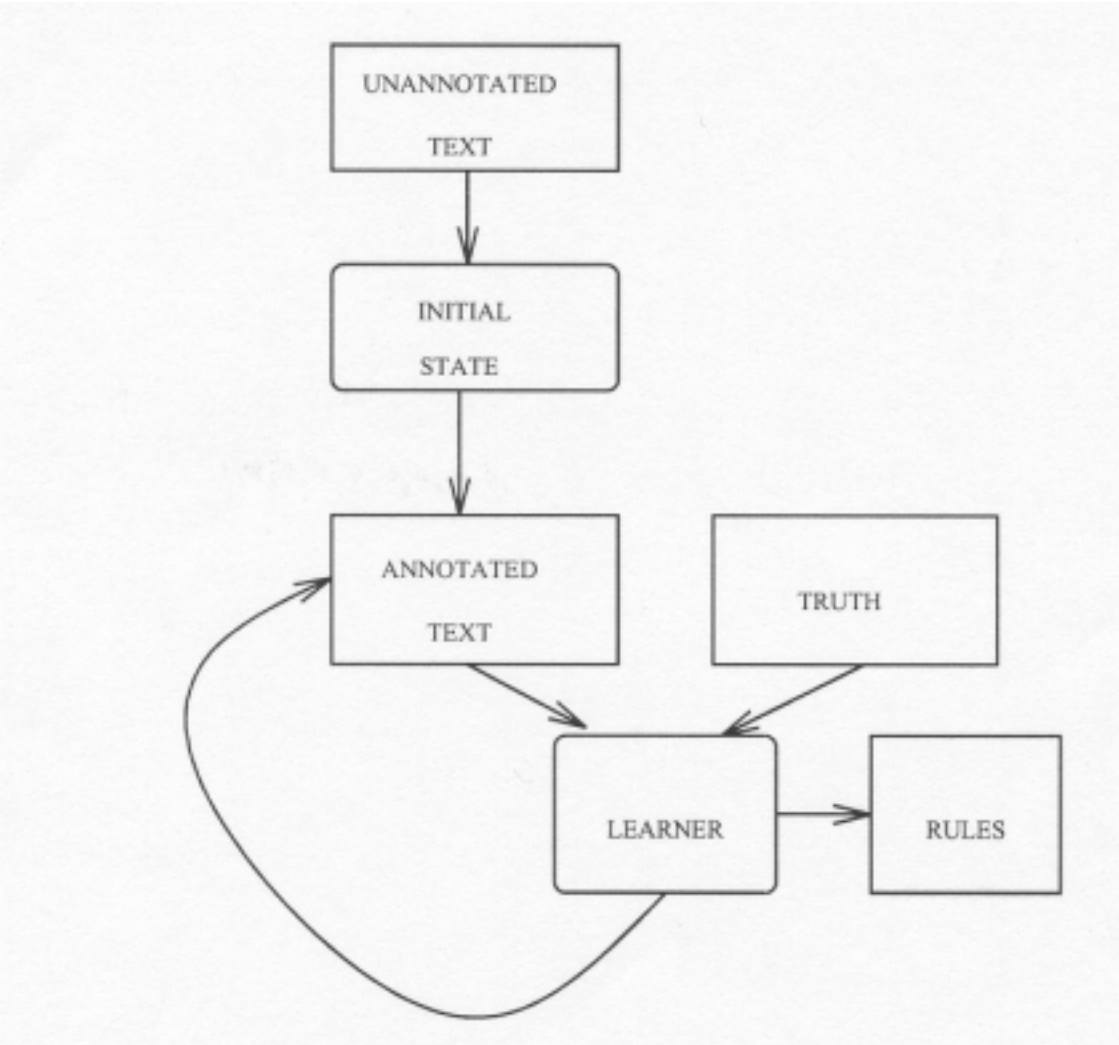
•The gap1 mutant blocked stable association of Ste4p with the plasma membrane, and the ste18 mutant blocked stable association of Ste4p with both plasma membranes and internal membranes.

•The gap1 mutant blocked stable association of Ste4p with the plasma membrane, and the ste18 mutant blocked stable association of Ste4p with both plasma membranes and internal membranes.

Step 2. Processing compound or complex sentences

- Simple part-of speech rules
  - Brill POS tagger package
  - Analysis of sentence structure

How does Brill POS tagger package work?



Let’s reflash the simple grammar

- CC (coordinating conjunction)
  - “and” “or” “but” “nor” “so”
- DT(determiner)
  - “a” “the” “this” “that” “some” “each”
- IN(preposition)
  - “in” “at” “on”
- JJ(adjective)
  - “beautiful” “useful”
- NN(noun)
  - “apple”

Let’s reflash the simple grammar

- NNP(proper noun)
  - “lysosome” “multitask”
- NNS(noun, plural)
  - “apples”
- IN(subordinating conjugation)
  - “when” “if” “after”
- VB(verb) and VBN(verb,past participle)
- P(1/2)(phrase)
- P(3/4/5)(phrase without verb)

- The gap1 mutant blocked stable association of Ste4p with the plasma membrane, and the ste18 mutant blocked stable association of Ste4p with both plasma membranes and internal membranes.

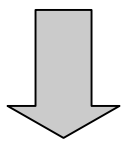
- The/DT gap1/NNP mutant/JJ blocked/VBN stable/JJ association/NN of/IN Ste4p/NNP with/IN the/DT plasma/NN membrane/NN,/, and/CC the/DT ste18/JJ mutant/JJ blocked/VBN stable/JJ association/NN of/IN Ste4p/NNP with/IN both/DT plasma/NN membranes/NNS and/CC internal/JJ membranes/NNS./.

Rules of part-of-speech

- Rule 1.
  - P1 [(,CC DT) | (,IN) | : | ;] P2 can be separated to P1 and P2
- Rule 2.
  - P3 VB1 P4 VB2 CC P5 can be separated to
    - P3 VB1 P4
    - P3 VB2 P5

Example of Rule 1

The/DT gap1/NNP mutant/JJ blocked/VBN stable/JJ association/NN of/IN Ste4p/NNP with/IN the/DT plasma/NN membrane/NN,/, and/CC the/DT ste18/JJ mutant/JJ blocked/VBN stable/JJ association/NN of/IN Ste4p/NNP with/IN both/DT plasma/NN membranes/NNS and/CC internal/JJ membranes/NNS./.

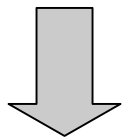


P1 ⇨ The/DT gap1/NNP mutant/JJ blocked/VBN stable/JJ association/NN of/IN Ste4p/NNP with/IN the/DT plasma/NN membrane/NN

P2 ⇨ ste18/JJ mutant/JJ blocked/VBN stable/JJ association/NN of/IN Ste4p/NNP with/IN both/DT plasma/NN membranes/NNS and/CC internal/JJ membranes/NNS./.

Example of Rule 2

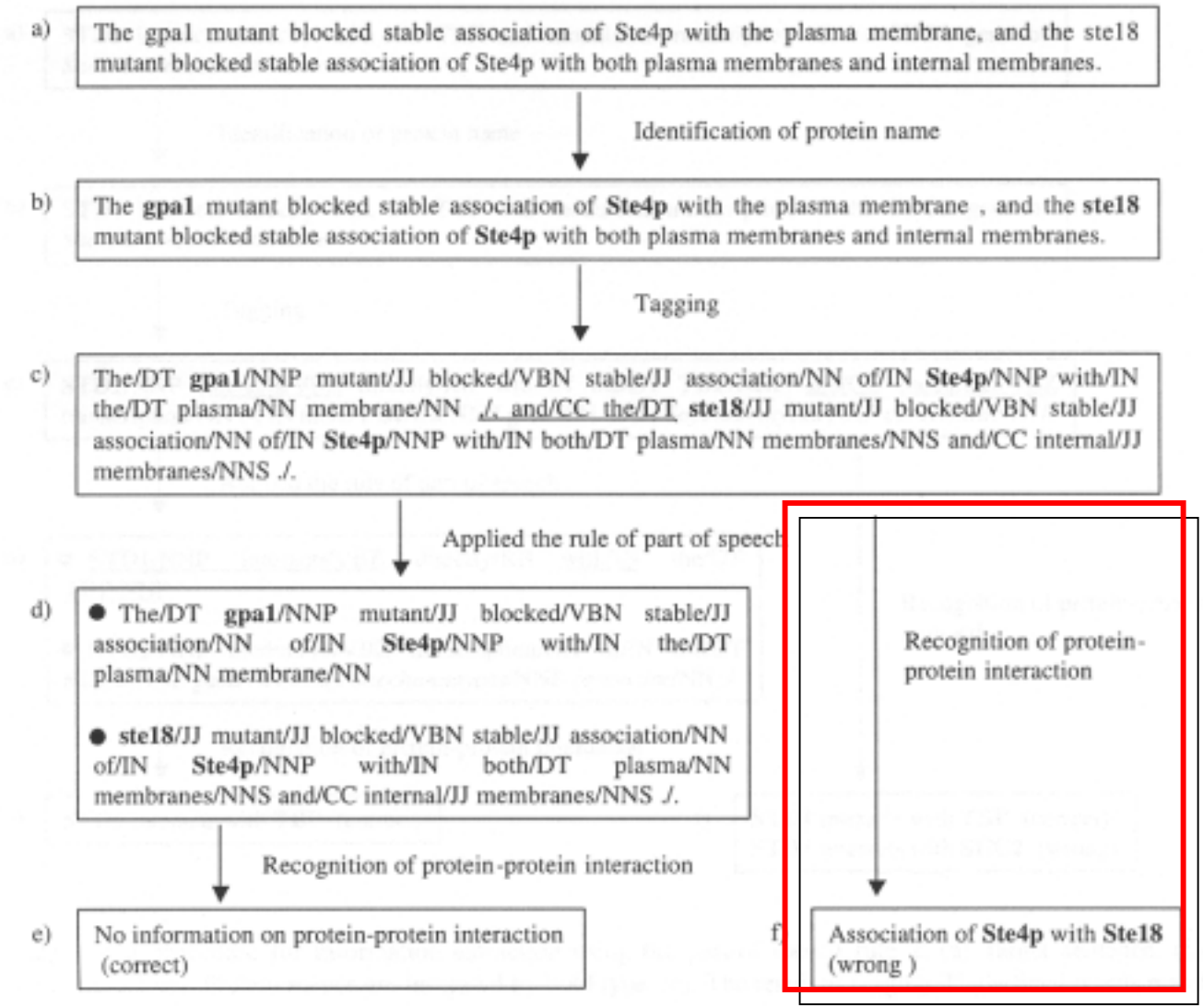
STD1/NNP interacts/VBZ directly/RB with/IN the/DT TBP/NNP and/CC modulates/VBZ transcription/NN of/IN the/DT SUC2/NNP gene/NN of/IN *Saccharomyces*/NNP *cerevisiae*/NN./.



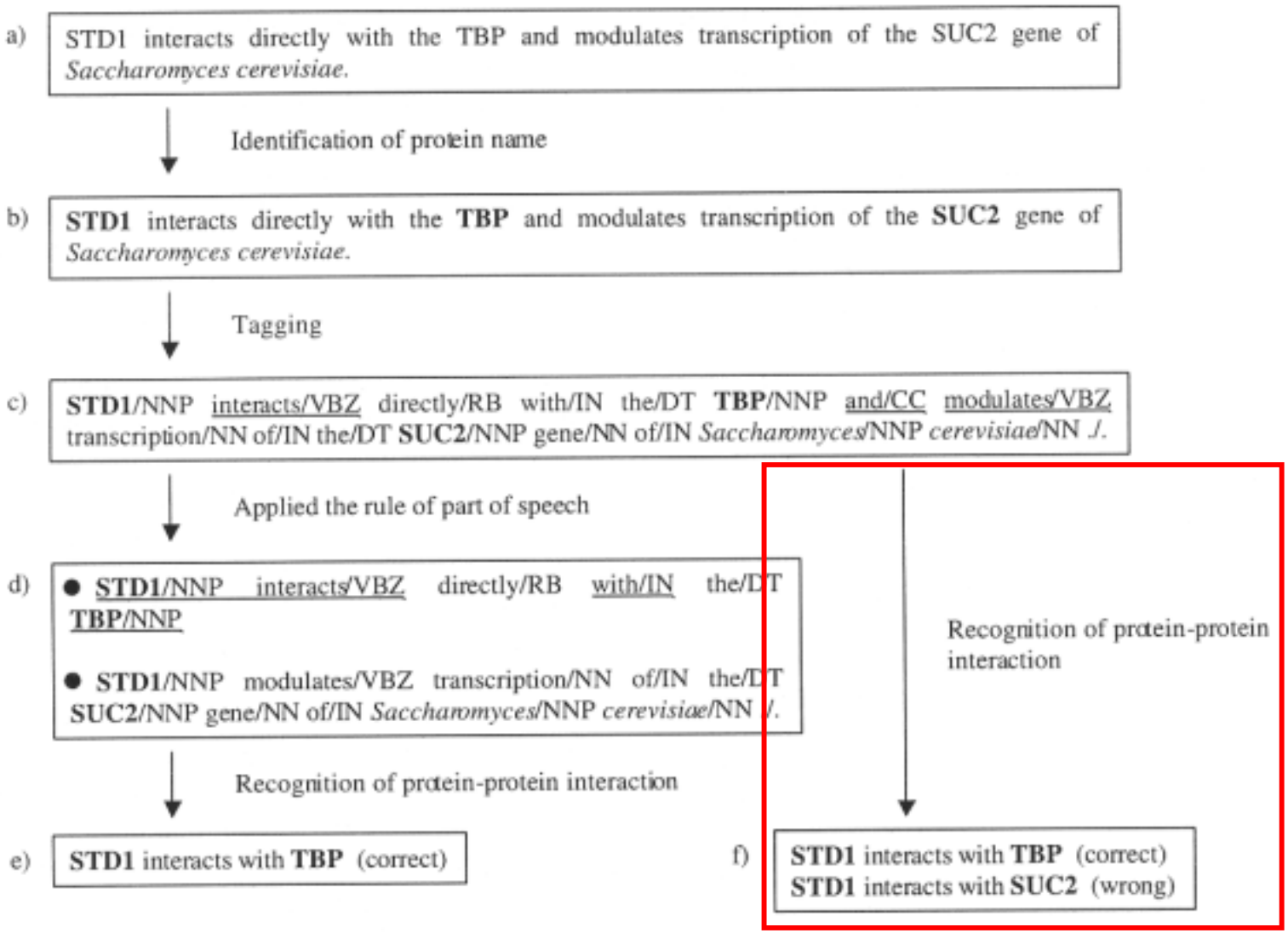
STD1/NNP interacts/VBZdirectly/RB with/IN the/DT TBP/NNP

STD1/NNPmodulates/VBZ transcription/NN of/IN the/DT SUC2/NNP gene/NN of/IN *Saccharomyces*/NNP *cerevisiae*/NN./.

Without applying part-of-speech rules



Without appling part-of-speech rules



Step 3. Recognition of the protein-protein interaction

- Keyword match
  - “interact” “associate” “bind” “complex”
- Negative sentence
  - “not interact” “not associate”
  - To increase precision
- Suffixes removing
  - To remove the inflection of keyword
  - Porter stemming algorithm (1980)

Keyword match

Table 2. A set of word patterns for recognition of protein–protein interaction. *A* and *B* indicate the protein name

Keyword	Pattern	Example of sentence
Interact	*A* interact with *B*	*Spc97p interacts with spc98* and *Tub4* in the two-hybrid system.
	interaction of *A* (with|and) *B*	The *interaction of Cet1 with Ceg1* elicits...
	interaction (between|among) *A* and *B*	Functional and physical *interaction* between *Rad24* and *Rfc5*...
	A–*B* interaction	These data suggest that the *Cert1-Ceg1 interaction* is...
	A and *B* interact	*Stn1 and Cdc13* proteins displayed a physical *interaction* by...
Associate	*A* associate with *B*	*Atx1* also *associated* directly *with* the cytosolic domains of *Ccc2*.
	association between *A* and *B*	Physical *association between GCN5 and ADA2*.
	association of *A* (with|and) *B*	*Association of Vma12p* with *Vph1p*.
	A and *B* association with each other	The *SET4 and STE18* gene products *associated with each other*.
Bind	*A* bind to *B*	*GCN binds to ADA2*...
	bind of *A* to *B*	The *binding of Met28* to *DNA*.
	A and *B* bind	*Cdc24p and Bem1p bind* to each other
	bind between *A* and *B*	*Binding between TIF34* and *TIF35* in_vitro.
	A bind *B*	the N-terminal of *SIN1* is suffisient to *bind SAP1*.
Complex	*A*(-|/)*B* complex	*Pc11, 2-Pho85* kinase *complexes* become essential...
	A and *B* complex	*Cdc46p* and *Cdc47p* ... *complex* with each other.
	complex *A* and *B*	*Pol1* and *Pob3* may form a *complex*...
	A complex with *B*	*GCG20* was... *complex* formation *with GCN1*.
	A complex... contain *B*	*Boi1p* is part of a larger *complex* that *contains Cdc42p*.
	A complex *B*	*Ste11 complexed* to *Ste7*...

Negative sentence

- Pattern 1.
  - Protein 1 . * not (interact|associate|bind|complex) . * Protein 2
  - Dmc1 does not interact in the two-hybrid assay with Rad52p or Rad54p.
- Pattern 2.
  - Protein 1 . * Pattern.* but not Protein 2
  - Bnr1p interacts with another Rho family member, Rho4p, but not with Rho1p.

Suffixes removing

- Inflection of keyword will decrease the precision of information extraction
- Porter stemming algorithm
  - Connected
  - Connecting
  - Connection
  - Connections

“Connect”

How does Porter stemming algorithm work?

- The concept of ‘consonant’ and ‘vowel’
  - Consonant: other than A, E, I, O, U and other than Y preceded by a consonant
  - Vowel: A, E, I, O, U, or Y
  - TOY -> Consonants are T and Y
  - SYZYGY -> Consonants are S,Z and G, vowel is Y
- Grouping
  - C : A list of ccc…. of length greater than 0
  - V: A list of vvv…. of length greater than 0

- Any word has one of the four forms…
  - CVCV…C
  - CVCV…V
  - VCVC…C
  - VCVC…V
- All be represented by the single form…
  - [C]VCVC…[V] or $[C](VC)^m[V]$
  - m=0 TR(C), EE(V), TREE(CV)
  - m=1 TROUBLE(CVCV), OATS(VC)
  - m=2 TROUBLES(CVCVC), PRIVATE(CVCVCV)

- Dealing with plurals and past participles (Step 1.)
  - SSES -> SS caresses -> caress
  - IES -> I ponies -> poni
  - SS -> SS caress -> caress
  - S -> cats -> cat
  - (m>0) EED -> EE agreed -> agree
    feed -> feed
  - (*v*) ED -> plastered -> plaster
  - (*v*) ING -> motoring -> motor
    sing -> sing

- AT -> ATE conflat(ed) -> conflate
- BL -> BLE troubl(ed) -> trouble
- IZ -> IZE siz(ed) -> size
- (*d and not (*L or *S or *Z)) -> single letter
  hopping -> hop
- (m=1 and *o) -> E fil(ing) -> file
  - *o: stem ends CVC and 2nd C is not W, X or Y
- (*v*) Y -> I happy -> happi
  sky -> sky

- Dealing with noun, adjective (Step 2,3)
  - (m>0) TIONAL -> ATE relational -> relate
  - (m>0) FULNESS -> FUL hopefulness -> hopeful
  - (m>0) FUL -> hopeful -> hope
- Dealing with noun, adjective which m>1 (Step 4.)
  - (m>1) ANCE -> allowance -> allow
  - (m>1 and (*S or *T)) ION -> adoption-> adopt

- Dealing with remains (Step 5.)
  - (m>1) E -> probate -> probat
  - (m=1 and not *o) E -> cease -> ceas
  - (m>1 and *d and *L) -> single letter
    controll -> control
    roll -> roll

Efficiency of Porter stemming algorithm

Suffix stripping of a vocabulary of 10,000 words

Number of words reduced in Step 1. : 3597

Number of words reduced in Step 2. : 766

Number of words reduced in Step 3. : 327

Number of words reduced in Step 4. : 2424

Number of words reduced in Step 5. : 1373

Number of words not reduced : 1513

Evaluation of information extraction

- Extraction for yeast and *E. coli* proteins
  - Yeast protein name was derived from SGD
    - 6084 molecules and 16,772 synonyms
  - *E. coli* protein name was constructed using K-12 data
    - 4405 entries
- The way to obtain target sentence
  - Using keyword like ‘protein binding’, ‘yeast’, ‘E coli’, ‘protein’ and ‘interaction’
  - Containing at least two protein names and one keyword
  - 834 and 752 sentences for yeast and *E. coli* respectively

Table 3. Results of information extraction. (a) The value of recall and precision for yeast proteins. (b) The value of recall and precision for of *E.coli* proteins

Key word	TP	$TP + TN$	$TP + FP$	Recall (%)	Precision (%)
			(a)		
Interact	198	222	206	89.1	96.1
Associate	55	68	61	80.9	90.2
Bind	103	119	108	86.6	95.3
Complex	152	176	164	86.4	92.7
Total	508	585	539	86.8	94.5
			(b)		
Interact	173	208	180	83.2	96.1
Associate	34	44	38	77.3	89.4
Bind	133	166	139	80.1	95.7
Complex	155	182	172	85.2	90.1
Total	495	600	529	82.5	93.5

Recall(Sensitivity):

TP/(TP+TN)

Precision(Specificity):

TP/(TP+FP)

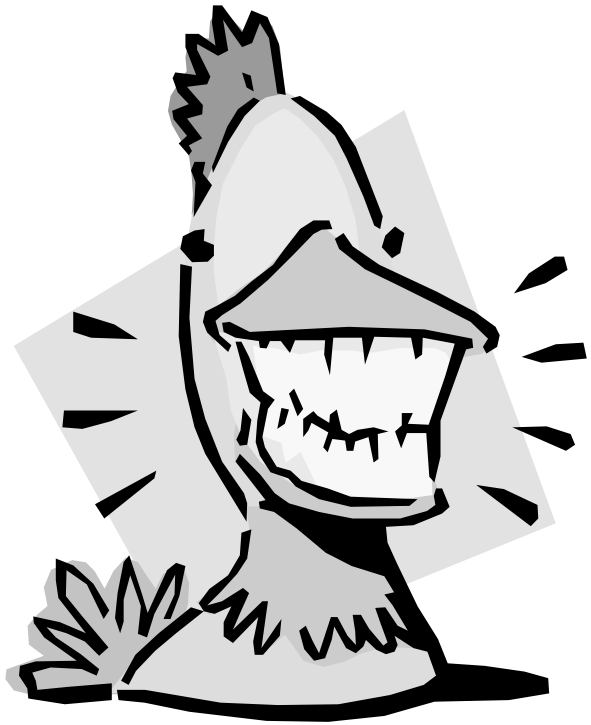
TP= number of sentences extracted correctly by this method

TP+TN=number of sentences containing information on protein-protein interactions

TP+FP= number of sentences retrieved by this method

Discussion

- Using negative sentence and extraction information about non-interaction
- Species-independent with proper dictionary
- Some errors arise from semantic differences and anaphoric terms
  - *These findings suggest that Msp1p is a component of the secretary vesicle docking complex whose 'function' is closely associated with that of Dec1p*
  - *'They' form a complex even in the absence of cross-linker*
  - Pronominal anaphora resolution algorithm



Questions or comments?