

Knowledge and Natural Language Processing in Molecular Biology

Kevin Bretonnel Cohen

Center for Computational Pharmacology,
University of Colorado School of Medicine

Thanks to George Acquah-Mensah, Andrew Dolbey,
Jens Eberlein, Aaron Gabow, Larry Hunter, Imran
Shah, and Ron Taylor

What could we do with knowledge bases?

- annotate/understand high-throughput data (Stein 2001)
- priors for Bayesian analysis
- data mining
-

What should knowledge look like?

- canonical
- support inference
- work synergistically with NLP to build itself

How to build knowledge bases?

- Fukuda et al. 1998
- Sekimizu et al. 1998
- Blaschke et al. 1999
- Rindflesch et al. 1999, 2000
- Humphreys et al. 2000
- Hatzivassiloglou et al. 2001
- Marcotte et al. 2001
- Ono et al. 2001
- Ray and Craven 2001
- Ding et al. 2002

Where are we now?

- information extraction
- syntactically:
 - not robust
 - doesn't necessarily help

Key problem: ambiguity

- lexical
- syntactic (structural)
- referential

Ways to resolve ambiguity:

- statistically (Collins 1996, 1999; others)
- knowledge about entities and relationships

Advantages of statistical approach

- know how to do it
- easy to incorporate new data

Disadvantages of statistical approach

- not good at lexicalizing grammar
- requires lots of data
- training data doesn't exist (yet)
- training data will be expensive

Advantages of knowledge-based approach

- know how to do it
 - semantic grammars
 - constraint satisfaction
- already have some of knowledge and some of the representational structure
 - KEGG
 - Rzhetsky
 - EcoCyc
 - GO
 - UMLS
- can use the knowledge for other purposes

Disadvantages of knowledge-based approach

- early decisions, if wrong, are costly

(How) can knowledge help in our domain?

- semantics: form \rightarrow meaning
- meaning $==$ knowledge
- therefore, semantics \approx knowledge
- semantics helps us:
 - resolve ambiguities
 - constrain search space

Two kinds of knowledge:

- contextual

- general

- We measured the concentration of this posttranslational form with a mass spectrometer.
- posttranslational forms are proteins
- proteins don't have mass spectrometers

Two kinds of general knowledge:

- knowledge about subclass and partonymy relations
- knowledge about particular instances
- (linguistic: synonym sets, thematic roles, etc.)

Two ways to disambiguate with semantics:

- selectional restrictions
- semantic grammar

ComplexKinasePhrase →
KinasePhrase *and* KinasePhrase

KinasePhrase → “ERK1”

KinasePhrase → “MAPKKK”

[CKP [KP ERK1 KP] and [KP MAPKKK KP]
CKP]

ComplexNounPhrase → NounPhrase
and NounPhrase

NounPhrase → ERK1

NounPhrase → embryos

The *Hypocrea jecorina* HAP 2/3/5 protein complex binds to the inverted CCAAT-box (ATTGG) within the *cbh2* (cellobiohydrolase II-gene) activating element.

binds [to the inverted...box] [within the ...activating element]

binds [to the inverted...box within the...activating element]

$\exists e, x, y$ Bound(x) && Location(y) && Binding(e, x, y)

$\exists x, y$ Box(x) && ActivatingElement(y) && Has-a(y, x)

---(TAACC)---┌-----*cbh2*-----

MAPK/fus3 activity is very low, whereas MAP/kss1 exhibits high activity, in the absence of mating pheromones.

(pheromones that are mating)

(a particular kind of pheromone)

The hypothesis was considered that direct regulation by tyrosine phosphorylation catalyzed by the insulin receptor would regulate the activity of serine-threonine protein kinases.

regulation [by tyrosine phosphorylation][catalyzed by the insulin receptor]

regulation [by tyrosine phosphorylation [catalyzed by the insulin receptor]]

Activity of the MAP kinase ERK2 is controlled by a flexible surface loop.
[flexible surface] [loop]
[flexible] [surface loop]

This study examined the transduction pathways activated by epinephrine in the pacemaker region of the toad heart.

(what's in the pacemaker region of the toad heart—the transduction pathways, or the epinephrine?)

KNOWLEDGE HAS A ROLE IN BOTH

- selectional restrictions/constraint satisfaction: impossible without it
- semantic grammar: intrinsic relationship between *well-represented* KB and the language used to represent concepts drawn from it

A well-represented knowledge base:

- helps us get more out of what we know
- is itself knowledge
- licenses inference

Why a DMAP solution?

- supports semantic rules nicely
- inference engine for constraint satisfaction
- memory-efficient

www.inetmi.com

Conclusion

- knowledge-based NLP can get us to the next level in molbio text data mining
- to do it, we need:
 - rich representations
 - conceptually tagged data