

## Automatic Term Management in Biology

Sophia Ananiadou, Goran Nenadic

NLP group, Computer Science  
University of Salford

<http://www.cs.salford.ac.uk/NLP.html>  
S.Ananiadou@salford.ac.uk, G.Nenadic@salford.ac.uk

S. Ananiadou, G. Nenadic

Date:18/02/02 1

## Outline

- Introduction: some problems of term management
- Our approach to Term Management in Biology: term recognition and clustering
- The ATRACT workbench: an integrated knowledge management system in Molecular Biology
- Conclusion and further research

S. Ananiadou, G. Nenadic

Date:18/02/02 2

## Term Management

- Difficult for domain experts to assimilate new knowledge without automated help
- Knowledge is encoded in textual documents, which are characterised by sets of specialized terms
- Our approach to **term management** is based on **automatic** term recognition and clustering
- Automatic term management - **ATM**

S. Ananiadou, G. Nenadic

Date:18/02/02 3

## Term recognition

- In biology, naming conventions are not always clear
  - there is a terminological confusion, although there are attempts to standardize terminology
- There is no formal criteria to distinguish terms/entity names from non-terms (e.g. *Bride of sevenless*)
- Frequency is not always a good indicator
- Chief motivation of term designation is classification

S. Ananiadou, G. Nenadic

Date:18/02/02 4

## Term variation

- Term variation
  - multiple forms for the same concept
  - multiple concepts for the same form
- orthographical
  - all-trans-retinoic acid      promyelocytic leukemia
  - all trans retinoic acid      promyelocytic leukaemia
- morphological
  - biochemical study      biochemical studies

S. Ananiadou, G. Nenadic

Date:18/02/02 5

## Term variation - problems (cont.)

- syntactic
  - human clones      clones of humans
- acronyms
  - NF-kappa B      nuclear factor-kappa B
  - NF-kB      nuclear factor-kappa B
  - NcoA      nuclear receptor coactivator
  - CBP      CREB binding protein
- semantic
  - eye surgery      ophthalmological surgery
  - carcinoma      cancer

S. Ananiadou, G. Nenadic

Date:18/02/02 6

## Term clustering/classification

- ♦ Recognizing terms is the first step in ATM
- ♦ Terms recognized should be related to each other and/or to existing knowledge
- ♦ Classification and clustering of terms and their variations
  - ♦ group semantically similar terms
- ♦ Experts often disagree on aspects of term similarity

S. Ananiadou, G. Nenadic

Date:18/02/02

7

## Our approach to ATR

- ♦ C/NC-value method (Ananiadou & Frantzi, 2000) for automatic extraction of **multiword** terms
- ♦ A hybrid approach incorporating:
  - ♦ linguistics
  - ♦ statistics
  - ♦ contextual information
- ♦ Assigns weights to candidate terms and ranks them

S. Ananiadou, G. Nenadic

Date:18/02/02

8

## Our approach to ATR (cont.)

- ♦ Extension of C/NC-value handles orthographic, morphological and some syntactic variations (Mima, Ananiadou & Nenadic, 2001)
- ♦ Variants are grouped prior to statistical calculation: a term candidate comprises all of its variations
- ♦ Acronym recognition is incorporated in term recognition process (Nenadic, Spasic & Ananiadou, 2002)

S. Ananiadou, G. Nenadic

Date:18/02/02

9

## Acronyms

- ♦ Acronyms frequently occurring in biology texts
- ♦ "Formation patterns"
- ♦ Ambiguities:
  - ♦ same acronym - different concepts  
GR glucocorticoid receptor  
glutathione reductase
  - ♦ same concept - different acronyms  
transcription intermediary factor-2 TIF-2  
TIF2

S. Ananiadou, G. Nenadic

Date:18/02/02

10

## Automatic acronym retrieval (AAR)

- ♦ Method based on
  - ♦ morphological features of acronym constituents -domain -specific dictionary of combining affixes
  - ♦ syntactic contextual features of acronym expanded form  
chloramphenicol acetyltransferase (CAT)
- ♦ Expanded forms are grouped to comprise all acronym variants

S. Ananiadou, G. Nenadic

Date:18/02/02

11

## AAR - sample results

RAR alpha	retinoic acid receptor alpha
RAR-alpha	retinoic-acid receptor-alpha
RARA	retinoic acid receptor alpha
-----	
APL	acute promyelocytic leukaemia acute promyelocytic leukemia
-----	
Amt	Ah receptor nuclear translocator
-----	
ATRA	all-trans retinoic acid all-trans-retinoic acid

S. Ananiadou, G. Nenadic

Date:18/02/02

12

## Automatic acronym retrieval (cont.)

- Disambiguating acronym occurrences in text:
  - we use the last introduced acronym in the document
- However, disambiguation is not important for term management purposes
- Precision: above 97%

S. Ananiadou, G. Nenadic

Date:18/02/02

13

## ATR - sample results

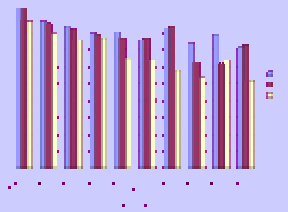
<u>COUP-TF II</u>	8.00
<u>retinoic acid receptor</u>	6.33
retinoic acid receptor	
retinoic acid receptors	
<u>nuclear receptor</u>	6.00
nuclear receptor	
nuclear receptors	
NR	
<u>nuclear receptor corepressor</u>	4.75
nuclear receptor corepressor	
NCoR	
<u>all-trans retinoic acid</u>	4.25
all trans retinoic acid	
all-trans-retinoic acids	
ATRA	

S. Ananiadou, G. Nenadic

Date:18/02/02

14

## ATR - overall precision



S. Ananiadou, G. Nenadic

Date:18/02/02

15

## Our approach to ATC

- Similar terms tend to appear in similar contexts
- Our approach is based on pattern mining: discovery of significant context patterns in which terms appear
- Our hybrid method incorporates:
  - context similarity (CS)
  - functional similarity (FS)
  - lexical similarity (LS)

S. Ananiadou, G. Nenadic

Date:18/02/02

16

## Our approach to ATC (cont.)

- Similarity between terms is defined as a linear combination of the three similarity measures:

$$\text{similarity}(T1, T2) = f(\text{CS}, \text{FS}, \text{LS})$$

- Once we have similarities, we can feed the results into any clustering method to single out clusters of terms
- We use AMI hierarchical clustering based on these similarities

S. Ananiadou, G. Nenadic

Date:18/02/02

17

## Term Similarities

- context similarity (CS)
- functional similarity (FS)
- lexical similarity (LS)

S. Ananiadou, G. Nenadic

Date:18/02/02

18

## ATC - context similarity

- Identify most important context patterns (CP) terms tend to appear in
- Although automatically collected, CPs are domain-specific!
- CPs are generalized morpho-syntactical REs representing left and right contexts of terms:

V:bind Term:rxr\_heterodimers PREP:with  
 <TERM> high\_affinity </TERM>  
 NP:value V:observe PREP:in NP:area

S. Ananiadou, G. Nenadic

Date:18/02/02

19

## ATC - context similarity (cont.)

- Context patterns consist of:
  - terms automatically recognized in text
  - significant sentence chunks (e.g. NP, V, PREP) recognized using the *lexie* tool (BioPATH)
- User can choose:
  - significant chunks
  - chunks to be instantiated (PREP vs. PREP:with)
- Normalized CPs contain only content words

S. Ananiadou, G. Nenadic

Date:18/02/02

20

## ATC - context similarity (cont.)

- Context patterns retrieved are ranked according to
  - their overall frequency in corpus  $f(p)$
  - pattern length  $|p|$
  - "nestedness"  $T(p)$

$$CS(p) = \left( \log \frac{f(p)}{f(\text{all terms})} \right) \cdot \left( \frac{1}{|p|} \right) \cdot \left( \frac{1}{T(p)} \right)$$

S. Ananiadou, G. Nenadic

Date:18/02/02

21

## ATC - context similarity (cont.)

- Sample of important left context patterns - only terms and most frequent verbs are instantiated

prep NP	272.65
prep NP prep	186.47
...	
prep NP V:stimulate	9.32
V:indicate NP	5.00
prep NP prep v:involve NP	4.64
prep Term:transcriptional_activity	4.47
V:require NP prep	4.38
prep Term:nuclear_receptor prep NP	4.00

S. Ananiadou, G. Nenadic

Date:18/02/02

22

## ATC - context similarity (cont.)

- Sample of important left context patterns - only prepositions are instantiated

prep:of NP	121.49
v NP	71.42
prep:of NP v	62.83
NP prep:of NP	59.72
prep:in NP	59.55
NP prep:of	43.37
prep:of NP v NP	37.64
prep:of Term	36.60

S. Ananiadou, G. Nenadic

Date:18/02/02

23

## ATC - context similarity (cont.)

- Context similarity

$$CS(T1, T2) = \frac{2 * (\# \text{ CPs shared})}{(\# T1 \text{ CPs}) + (\# T2 \text{ CPs})}$$

- Examples:

CS(retinoic acid receptor, nuclear receptor) = 0.58  
 CS(retinoid X receptor, nuclear receptor) = 0.60  
 CS(mutant receptor, nuclear receptor) = 0.20

S. Ananiadou, G. Nenadic

Date:18/02/02

24

## Term similarities

- \* context similarity (CS)
- \* functional similarity (FS)
- \* lexical similarity (LS)

## ATC - functional similarity

- \* Use of general **lexical patterns** that indicate functional similarity between terms
- \* Several types of lexical patterns are considered:
  - \* frozen expressions
  - \* coordination
  - \* apposition
  - \* anaphora

## ATC - functional similarity

- \* Frozen lexical expressions
  - \* Term *such as* (Term, )<sup>+</sup> *CONJ* Term
  - \* Term *like* (Term, )<sup>+</sup> *CONJ* Term
  - \* (Term, )<sup>+</sup> *CONJ* *other* Term
  - \* *both* Term *and* Term
  - \* *either* Term *or* Term
  - \* *neither* Term *nor* Term

## ATC - functional similarity (cont.)

- \* Coordination of terms
  - \* simple coordination (enumeration)  
including [HNF](#), [Ad4BP](#), [DAX-1](#), and [nur77/NGFIB](#),
  - \* argument coordination  
SMRT and Trip-1 mRNAs:  
[SMRT mRNA](#) and [Trip-1 mRNA](#)
  - \* head coordination  
adrenal glands and gonads  
[adrenal gland\(s\)](#) and [adrenal gonad\(s\)](#)

## ATC - functional similarity (cont.)

- \* appositions (in progress)  
... [CARM1](#), an previously unidentified [protein](#), ...
- \* anaphora resolution (in progress)  
... Murray and Towel recently found that cell nuclear extracts enhance the binding of [THR](#) to TRE. [This protein](#) has been designated ...

## ATC - functional similarity (cont.)

- \* Functional similarity FS(T1, T2) depends on the type of the lexical pattern the terms appear in
- \* More "reliable": frozen lexical patterns and non-simple coordination
- \* Examples  
FS( glucocorticoid receptor, estrogen receptor) = 1.00  
FS(SMRT mRNA, Trip1 mRNA) = 1.00  
FS( DAX1, NHF) = 0.70

## Term Similarities

- \* context similarity (CS)
- \* functional similarity (FS)
- \* lexical similarity (LS)

## ATC - lexical similarity

- \* Lexical similarities help identify similar terms
- \* Sharing the same head
  - receptor
  - progesteron receptor, estrogen receptor
- \* Sharing the arguments (specialised concepts)
  - nuclear receptor
  - hormone nuclear receptor
  - orphan nuclear receptor

## ATC - lexical similarity (cont.)

- \* Lexical similarities between terms T1 and T2:

$$LS(T1, T2) = a * shared\_head(T1, T2) + b * shared\_arguments(T1, T2),$$

$$a > b$$

- \* Examples

$$LS(\text{nuclear receptor, retinoid nuclear receptor}) = 0.85$$

$$LS(\text{nuclear receptor, retinoid X receptor}) = 0.70$$

## ATC - overall similarity (sample)

$$\text{similarity}(T1, T2) = \alpha * CS_{12} + \beta * FS_{12} + \gamma * LS_{12}$$

### glucocorticoid receptor

0.68 ( 0.42, 1.00, 0.50)	estrogen_receptor
0.66 ( 0.38, 1.00, 0.50)	steroid_receptor
0.55 ( 0.00, 1.00, 0.50)	progesterone_receptor
0.31 ( 0.52, 0.00, 0.50)	receptor_complex
0.30 ( 0.50, 0.00, 0.50)	nuclear_receptor
0.28 ( 0.43, 0.00, 0.50)	human_estrogen_receptor
0.23 ( 0.78, 0.00, 0.00)	estrogen_antagonist

## ATC - overall similarity (sample)

0.68	glucocorticoid receptor	estrogen receptor
steroid_receptor	0.66	0.64
progesterone_receptor	0.55	0.59
human_estrogen_receptor	0.28	0.37
retinoid_x_receptor	0.27	0.36
nuclear_receptor	0.30	0.33
receptor_complex	0.31	0.33
retinoic_acid_receptor	0.27	0.28
retinoid nuclear receptor	0.26	0.26

## ATC - evaluation

- \* Random samples of results have been evaluated by a domain expert from LION BioScience
- \* The measure introduced proves to be a good indicator of semantic similarity
- \* Consistent similarity measure: similar terms share the same "friends"

## ***ATRACT***

- ♦ ATRACT is an integrated knowledge management system in the domain of molecular biology
- ♦ ATRACT stands for  
**Automatic Term Recognition and Clustering of Terms**
- ♦ Knowledge mining is carried out by integrating
  - ♦ automatic term recognition (ATR)
  - ♦ automatic term clustering (ATC)

## ***ATRACT (cont.)***

- ♦ ATRACT provides user-friendly GUI
- ♦ **Adaptability**
  - ♦ tuning ATR and ATC
- ♦ **Real-time processing**
  - ♦ terms are recognised "on-line"

## ***ATRACT - demo***

## ***Conclusion***

- ♦ An efficient and reliable aid for term management
- ♦ Very good results for ATR and encouraging results for ATC
- ♦ Adaptability is important: tunable parameters allow customised approach
- ♦ Formal evaluation methods needed: *evaluation against an existing ontology*

## ***Further research***

- ♦ Self-tunable system (via supervised learning using manually evaluated results)
- ♦ Classification of terms (via supervised learning from a manually tagged corpus, e.g. GENIA)
- ♦ Populating & updating ontologies in a consistent manner (avoiding terminological confusion)

## ***Acknowledgement***

- ♦ LION BioScience  
D. Schuhmann, S. Albert, H. Kirsch
- ♦ Computer Science, University of Tokyo  
Prof. J. Tsujii, H. Mima
- ♦ NLP group, University of Salford  
I. Spasic, K. Manios