

BioCreAtIvE Task 2, user 15

---

## Protein function assignment using term-based SVMs

Simon Rice, Goran Nenadic, Ben Stapley

BioMolecular Sciences/Computation

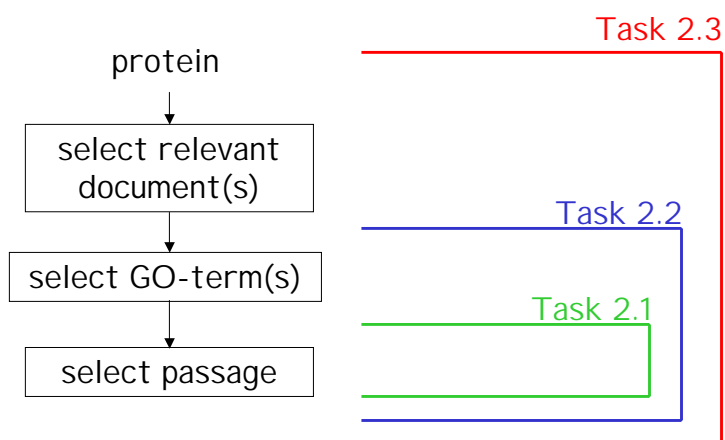
UMI ST, Manchester, UK



BioCreative Workshop, 30/3/2004, Granada, Spain

## Discussion

---



## Discussion

---

- task 2.3
  - realistic and challenging
- task 2.2
  - how papers suitable for annotation are selected?  
*"Finding suitable paper is not too difficult"*
  - is one document enough to predict annotations?  
*"more than 1 paper per annotation"*
- task 2.1
  - only for assignments derived from non-text methods
  - is a passage enough to "support" prediction?  
*(annotators provide PMIDs)*

## Approach

---

- entities that co-occur with proteins are indicative of their function(s)
- thus, proteins with similar co-occurrences will have related functions
- collect and make use of weak evidence
- learn informative co-occurring entities (features) for a given GO-term
- machine learning approach (SVMs)
- "closed" approach (as defined in Task 1)
  - only training data used for learning

## Method

---

1. feature selection and weighting
2. learning SVM for each GO-term
3. testing
  - each gene against SVMs
  - passage selection

## Method - feature selection

---

- what are good features?
- terminologically relevant sequences are more informative features
- GO-terms are not frequent in papers
- we use
  - single words (except standard stop-words)
  - automatically extracted terms as semantic indicators

## Method – feature selection

---

- recognise occurrences of biologically relevant (multi-word and nested) terms
- C/NC-value method
  - 9-cis retinoic acid*
  - retionic acid*
  - acute promyelocytic leukaemia*
  - RAR-alpha*
  - retinoic-acid receptor alpha*
- term variants are conflated and linked as they represent same features

## Method – feature unification

---

- a) typo-orthographic
  - leukaemia* and *leukemia*
  - amino-acid* and *amino acid*
- b) inflectional
  - nuclear receptors* and *nuclear receptors*
  - Down's syndrome* and *Down syndrome*
- c) acronyms and their variants
  - NFKB factor*, *NF-KB* and *nuclear factor kappa B*
  - RAR-alpha*, *RARA* and *retinoic acid receptor alpha*

## Method – feature selection

---

- features: **sets** of equivalent synonymous term variants (synterms)

*F1 = {all trans retionic acid, all-trans-retinoic acids, ATRA, at-RA}*

*F2 = {nuclear receptor, nuclear receptors, NR, NRs}*

*F3 = {9-c-RA, 9cRA, 9-cis-retinoic acid, 9-cis retinoic acid}*

...

- proteins represented as vectors of co-occurring (syn)terms (on document level)

## Method – feature selection

---

- feature weights

– *idf*-like measure

$$\log \frac{1 + \sum_{j \in R_g} f_j(w)}{N_w (1 + |R_g|)}$$

- collecting features

– use: abstract, “main” text and figure legends

– skip: *experimental*, *methods* and *references* sections

## Method – training phase

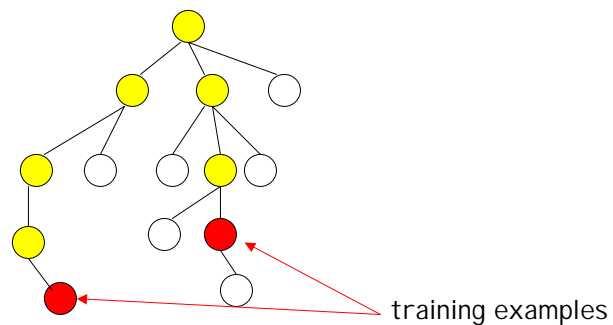
---

- for each GO term collect
  - positive examples:
    - “exact” examples from the training set
    - examples corresponding to their children
  - negative examples
    - from sibling GO-terms and their children
  - equal number of positive/negative examples
- train SVM for each GO term available

## Method – training phase

---

- GO-SVMs can only be trained for a subtree of the GO formed from root to GO terms occurring in the training data



## Method – prediction phase

---

- prediction of GO-term for a given protein
  - create feature vector for the protein from given/relevant document(s)
  - test the vector against all GO-classifiers, and select top-ranked
- prediction of relevant passage for a given protein-GO pair
  - create a vector for each *paragraph* (<p>) from given/relevant document(s)
  - test vectors against the corresponding GO-classifier, and select the top-ranked paragraph

## Task 2.1: GO + doc + protein → <p>

---

- we can select <p> only for GO-terms that appeared in training examples
  - ½ of test GO-terms do not appear in training set, covering 43% of test examples
- two submissions
  - if we do not have the corresponding GO-classifier
  - 1) skip the testing example
  - 2) use the a nearest neighbour GO-term (shortest path through ontology) to find <p>

## Task 2.1: results

GO code	protein	Submission 1 524 <sup>a</sup>		Submission 2 998 <sup>a</sup>	
		pairs	prec.	pairs	prec.
high	high	59	11%	125	12%
general	high	28	5%	69	7%
high	general	19	4%	38	4%
Total		106	20%	232	23%

<sup>a</sup>Total number of predictions.

recall:

10%

22%

2 X

## Task 2.2: doc + protein → GO, <p>

- make a vector for a given protein based only on the given doc, test the vector against all GO-SVMs and select top-ranked
- <p> selection as in Task 2.1
- two submissions
  - 1) use only training examples as provided (closed)
  - 2) use data from Task 2.1 as additional training data (closed++)

## Task 2.2: results

GO code	protein	Submission 1 502 <sup>a</sup>		Submission 2 485 <sup>a</sup>	
		pairs	prec.	pairs	prec.
high	high	3	1%	16	3%
general	high	8	2%	26	5%
high	general	2	0%	2	0%
Total		13	3%	44	9%

<sup>a</sup>Total number of predictions.

3 X

## Task 2.2: discussion

1. poor overlap between training and testing GO-terms: including Task 2.1 data significantly improved precision and recall
2. features are sparse if based on single document: more data is needed as we are not looking for "strong" evidence
  - submission 3 (not submitted)
    - use the whole corpus to make vectors, and then use the given doc only to select <p>

## Task 2.3: protein → doc, GO, <p>

---

- find the relevant docs
  - protein name variation is substantial: located docs for 81 of 138 genes (with 446 synonyms)
- ad-hock retrieval method
  - for a given protein, each document in the corpus is assessed using
    - all synonyms of the protein (DE field, Swiss-Prot/Trembl)
    - all single words from the respective DE and GN fields
  - scores based on *idf*-weighting
- two submissions as in Task 2.2

## Task 2.3: results

GO code	protein	Submission 1 36 <sup>a</sup>		Submission 2 52 <sup>a</sup>	
		pairs	prec.	pairs	prec.
high	high	11	31%	11	21%
general	high	7	19%	8	15%
high	general	0	0%	0	0%
<b>Total</b>		<b>18</b>	<b>50%</b>	<b>19</b>	<b>36%</b>

<sup>a</sup>Total number of predictions.

## Discussion

---

- two separate problems
  - a) function assignment
  - b) passage selection

### a) function assignment

---

- collecting substantial “weak” evidence
- more (relevant) text needed to learn correlations
  - assignments derived from many documents are more reliable than assignments based on single document
- consequently, better results for Task 2.3: 10 retrieved documents per protein
- relevant latent information can be inferred from weak evidence from many documents

## b) passage selection

---

- no training passages to learn from
- passages contain sparse features: cannot capture evidence, and, thus, cannot give accurate predictions
- we have relatively good assignment of GO-terms but non-relevant selection of paragraphs

## Example

---

- ✓ protein: *BARD1*
- ✓ GO-term: *DNA-directed RNA polymerase II, holoenzyme*

X paragraph:

**FIG. 6. ESI-TOF mass spectra for the binding of  $Zn^{2+}$  to the subunits of wild-type and mutant BC-112/BD-115 heterodimer complexes.** Spectra shown are for wild-type (A), C39A (B), C64A (C), the sample shown in C + 40  $\mu M$   $Zn^{2+}$  (D), the sample shown in C + 50 mM EDTA (E), and the sample shown in C + acetic acid (pH < 4.0) (F). The increased intensity of the **BRCA1** subunit relative to **BARD1** in panel A (asterisk) is primarily attributable to an excess of BRCA1 homodimer in this sample preparation.

# Example

- ✓ protein: *BARD1*
- ✓ GO-term: *DNA-directed RNA polymerase II, holoenzyme*

- “knowledge discovery” approach  
this (correct) association was learnt from weak evidence spread across several documents
- but, it was difficult to select relevant passage

# Example

- ✓ protein: *BARD1*
- ✓ GO-term: *DNA-directed RNA polymerase II, holoenzyme*

pubmed.ncbi.nlm.nih.gov/12102400/

## The BRCA1 and BARD1 Association with the RNA Polymerase II Holoenzyme

Satouka Chiba and Jeffrey D. Parvin<sup>1</sup>

Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115

### ABSTRACT

We have previously shown that endogenous BRCA1 and overexpressed epitope-tagged BRCA1 are present in the transcription complex called the RNA polymerase II holoenzyme (holo-poli). In this study, we further characterized BRCA1 association with the holo-poli by coexpressing deletion mutants of epitope-tagged BRCA1. We found that BRCA1-associated (SUN) domain protein (BARD1) is a component of the holo-poli complex. Deletion of the BRCA1 NH<sub>2</sub> terminus, which is bound by BARD1 as well as other proteins, eliminates ~99% of BRCA1 association with holo-poli. In contrast with earlier observations, we found that C-terminal truncations of BRCA1 did not affect holo-poli association. Immunoprecipitation analysis of BRCA1 showed that the NH<sub>2</sub> terminus of BRCA1 is required for nuclear dot formation in S-phase. BRCA1 NH<sub>2</sub> terminus is required for the association with holo-poli and for subnuclear localization in S-phase foci. Taken together, these data support a role for BRCA1 regulation of holo-poli function.

### INTRODUCTION

Subnuclear localization of BRCA1 changes are dynamically dependent on the cell cycle or DNA damage. In S-phase of cell cycle, BRCA1 localizes to discrete nuclear foci (dots) with BARD1 and Rad51 (22–24). The protein complexes in the nuclear foci are unknown. It is unknown which BRCA1 domains is important for the structure of nuclear dots and whether BRCA1 changes of subnuclear proteins contribute to holo-poli association. We found that BARD1 is a component of the holo-poli complex and deletion of the BRCA1 NH<sub>2</sub> terminus eliminates the association of the mutant BRCA1 protein with the holo-poli complex. This deletion mutant also fails to form nuclear foci, leading to the suggestion that the nuclear foci might contain holo-poli, although other protein complexes associated with the BRCA1 NH<sub>2</sub> terminus may be responsible for the BRCA1 nuclear dot formation. Our data suggest that BRCA1 is involved in the function of Pol II in several ways via association with the holo-poli and possibly via subnuclear localization.

not in the training/testing set

## Conclusions and questions

---

- SVM-based machine-learning approach with ATR terms as features
- sufficient data is needed both for training and for predictions
- we can infer (discover) some latent and relevant information from weak evidence
- poor performance on selection of short passages and on single documents

## Acknowledgments

---

- UK BBSRC-funded project  
“Protein functional classification using text data mining” (2002-2005)
- UK National centre for bio-text mining  
UMI ST, Manchester (2004-2007)
- annotators and organisers