CENG 734
Advanced Topics in Bioinformatics

Week 9
Text Mining for Bioinformatics: BioCreative II.5

Fall 2010-2011
1. Draw the decompressed graph for the following graph summary:

![Graph Diagram]

- $A(w) = \{d, e, f\}$
- $A(x) = \{a, b\}$
- $A(y) = \{c\}$
- $A(z) = \{g, h\}$

2. Describe the two criteria that is used in the SPICi algorithm for stopping clustering expansion.
Overview of BioCreative II.5

• Tasks
  – Rank articles on how likely they contain PPIs
  – Identify interacting proteins in articles that contain PPIs
  – Identify interacting protein pairs

• Evaluation test set: 595 full text articles (61 PPI related)

• Performance measures:
  – Area under the precision-recall curve (favors recall)
  – Balanced F-measure (favors precision)
Overview of BioCreative II.5

• 15 teams participated
• Best AUC for article classification was 0.70
• For interacting proteins, the best system achieved good macroaveraged recall (0.73) and AUC (0.58)
• After filtering incorrect species and mapping homonymous orthologs; for interacting protein pairs, the top (filtered, mapped) recall was 0.42 and AUC was 0.29.
• Ensemble systems improved performance for the interacting protein task.
Curation of PPIs from Literature

• Databases:
  – MINT
  – IntAct
  – BioGRID

• Manual process
FEBS Letters Experiment

• In 2007 the FEBS Letters journal asked the authors to submit structured, i.e., machine readable, information along with their manuscript.
• They have found that this slowed down publication process.
• They now have MINT curators to extract structured information and ask approval of authors.
• Benchmark dataset for automated tools.
Tasks: in detail

• **Article categorization task (ACT):** Binary classification and ranking of articles (document classification) as relevant for curation, i.e., for extracting PPI annotations

• **Interactor normalization task (INT):** Ranked lists of identifiers of proteins for which the article reports evidence for an interaction with experimental evidence

• **Interaction pair task (IPT):** Lists of binary interaction pairs with experimental evidence as protein identifier pairs per article
The BioCreative Meta Server

![Diagram of the BioCreative Meta Server]

10 min time limit
Performance measures: detailed

- Balanced F measure
  \[ F_\beta = \frac{(1 + \beta^2) \times (p \times r)}{\beta^2 \times (p + r)} \]

- Interpolated AUC
  \[ A(f_{pr}) = \sum_{j=1}^{n} (p_{ij} \times (r_j - r_{j-1})) \]
  \[ p_i(r) = \max_{r' \geq r} p(r') \]

- For ACT task
  \[ Acc = \frac{TP + TN}{TP + FP + FN + TN} \]
  \[ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \]
Difficulty of the challenge

• Even the human curators do not always agree
• 81% agreement between curators of different database
• 93% agreement between curators of the same database
Availability

• All data and results are available at

• BioCreative III was held this year. Workshop was on September 2010
ACT results

Positive Articles

1/3 articles with low agreement

17 “easy” articles

13 difficult articles

Negative Articles

1/30 articles with low agreement

mRNA interaction

gene/protein regulation

>36 easy articles

2 errors in gold standard
Different approaches

• For the most successful teams in the IPT (teams 18, 37, and 42), the approaches seem to be technically quite variable: Team 42 only used pattern matching, Team 18 employed conjunction handling including shallow parsing, while Team 37 employed a very wide range of NLP techniques, even deep parsing of the grammatical structures of the text
Team 42

- Efficient Extraction of Protein-Protein Interactions from Full-Text Articles
- Jörg Hakenberg, Robert Leaman, Nguyen Ha Vo, Siddhartha Jonnalagadda, Ryan Sullivan, Christopher Miller, Luis Tari, Chitta Baral, and Graciela Gonzalez
- Most of the authors are from ASU
Team 42

• Submitted results for 21 of 61 relevant articles for the IPT task

• Two goals:
  – Protein named entity recognition, including normalization
  – Extraction of protein-protein interactions from full text

• Can process a full text article from 10 seconds to 2 minutes based on different configurations
Summary

• Propose strategies to transfer document-level annotations to the sentence-level, which allows for the creation of a more fine-grained training corpus.

• Use this corpus to automatically derive around 5,000 patterns.

• Rank sentences by relevance to the task of finding novel interactions with physical evidence, using a sentence classifier built from this training corpus.
A remarkable early work

• Fundel et al. found around 150,000 distinct PPIs of human genes/proteins within one million PubMed abstracts from 1990 to 2007.

Text mining needed for

- identify abstracts/full-text articles that contain relevant data,
- spot relevant passages in a given article (as opposed to established background information),
- recognize mentions of relevant biomedical entities (proteins and organisms),
- map each entity mention to a database identifier (such as a UniProt ID for each protein),
- extract relationships between entities (such as mapping a protein to an organism or finding PPIs),
- extract additional information on the experiments described (interaction-detection method, clone library, antibodies, etc.).
Named Entity Recognition (NER)

**Steps**

- machine-learning-based NER of protein names,
  - BANNER NER system is used which is based on Conditional Random Fields. Trained on BioCreative II data
- dictionary-based recognition of species names,
- generating candidate IDs for each recognized mention based on dictionary matching,
- filtering of IDs by species, and
- ranking of candidate IDs.
Relationship Extraction

• Use a pattern based approach
  – Common patterns of sentences that describe PPIs
• Generate task specific patterns by clustering annotated training data
• Conceptualizations of tokens
  – lists of words referring to protein interactions
• Use the OpenDMAP framework for maintaining and matching patterns
Generation of Sentence Level Training Data

• Given: Full text and pairs of interacting proteins in that full text
• Use NER to predict locations of the identifiers in full text
• Remove sentences which do not contain interaction related terms
• At the end: 784 positive out of 14,844 sentences in total are identified and used for training
Pattern Generation

• Convert each sentence to a pair of protein ids and a keyword that describes the interaction.
  – A interacts with B
  – A binds B
## Concept Groups

<table>
<thead>
<tr>
<th>Term</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>regulator (noun, singular)</td>
<td>accelerator, acceptor, activator, agent, anchor, catalysator, catalyzyator, deconjugator, down-regulator, downregulator, effector, immunoreactant, inhibitor, interactor, ligand, ligase, modifier, promoter, promotor, reactant, receptor, regulator, repressor, stimulator, suppressor, target, up-regulator, upregulator</td>
</tr>
<tr>
<td>regulation (noun, singular)</td>
<td>acetylation, activation, degradation, demethylation, dephosphorylation, depletion, destabilization, destruction, disruption, down-regulation, downregulation, elevation, expression, hyperexpression, induction, inhibition, methylation, modification, modulation, mono-ubiquitination, monoubiquitination, multi-ubiquitination, multiubiquitination, mutation, obstruction, over-expression, overexpression, phosphorylation, poly-ubiquitination, polyubiquitination, stabilization, stimulation, transacetylation, transactivation, transcription, up-regulation, upregulation</td>
</tr>
<tr>
<td>activity (noun, singular)</td>
<td>abolishment, abrogation, acceleration, accumulation, activity, addition, affection, amplification, augmentation, augmentation, conjugation, control, conversion, cross-reactivity, deconjugation, expansion, exposition, immunoreactivity, inactivation, infection, ligation, mediation, participation, precipitation, prevention, production, promotion, reduction, regulation, sequestration, substitution, sumoylation, suppression, transduction, ubiquitination</td>
</tr>
<tr>
<td>interaction (verb, past tense)</td>
<td>abolished, abrogated, absorbed, accelerated, accepted, accumulated, acetylated, activated, added, affected, amplified, anchored, antagonized, arrested, assembled, associated, attached, augmented, blocked, bound, broke, catalysed, catalysed, cleaved, co-eluted, co-eluted, co-immunoprecipitated, coimmunoprecipitated, cofected, complexed, conjugated, contacted, controlled, controled, co-purified, copurified, cross-linked, crosslinked, deacetylated, deconjugated, decreased, degraded, demethylated, depended, dephosphorylated, depleted, derived, destabilized, destructed, detached, dimerized, disassembled, disassociated, discharged, disrupted, down-regulated, downregulated, elevated, encoded, encompassed, enhanced, evoked, exhibited, formed, fused, fastened, immunoblotted, immunoprecipitated, immunoreacted, impaired, inactivated, incited, increased, induced, infected, influenced, inhibited, initaited, injected, interacted, interfered, interplayed, ligated, linked, methylated, modified, modulated, mono-ubiquitinated, multi-ubiquitinated, oxidised, oxidized, participated, phosphorylated, poly-ubiquitinated, polyubiquitinated, potentiared, precipitated, prevented, produced, reacted, recognised, recognized, recruited, reduced, regulated, related, removed, repaired, replaced, repressed, required, responded, restricted, severed, stabilized, stained, stimulated, substituted, sumoylated, suppressed, synthesised, synthesized, targeted, tethered, transacetylated, transactivated, transcribed, transduced, transferred, transformed, treated, ubiquitinated, up-regulated, upregulated</td>
</tr>
</tbody>
</table>
Example Patterns

... association between MAP1S and SOCS3 ...
  ⇒ \{w-interact-attachment\} \{w-preposition-of\} [interactor1] and [interactor2]

... binding activity between TSC-22 and fortilin ...
  ⇒ \{w-interact-activity-noun-s\} \{w-preposition-of\} [interactor1] and [interactor2]

... a novel role for MafG in HIF-1alpha accumulation ...
  ⇒ [interactor1] \{w-in\} [interactor2] \{w-interact-activity-noun-s\}

... localization of dysferlin, a binding partner of affixin, in ...
  ⇒ [interactor1], \{w-determiner\}? \{w-interact-attachment\} partner \{w-preposition-of\} [interactor2]

... modification of Prox1 by SUMO-1 ...
  ⇒ \{w-interact-regulation-noun-s\} \{w-preposition-of\} [interactor1] \{w-preposition-by\} [interactor2]
OpenDMAP

• The generated patterns are input to OpenDMAP which is an ontology-driven, rule-based concept analysis and information extraction system.
Relevance Ranking of Sentences

• Train an SVM model on 9750 sentences using libSVM
• Features extracted for each sentence:

1. section: in which major section does the sentence occur—Abstract, Introduction, Results, ...
2. position: where in a paragraph can a sentence be found—mapped to a value between 0 (⇒ section heading) and 1 (⇒ last sentence),
3. is the sentence a (sub-) section heading?
4. does the sentence occur in a) a figure caption or b) a table caption?
5. does the sentence contain a reference to
   - a. a figure,
   - b. a table,
   - c. a cited paper, or
   - d. supplementary information (see Appendix)?
6. bag-of-words, lemmatized; drawn from current and previous sentence (treated as separate features)
7. number of proteins mentioned in the sentence.
Ranking of Extracted Interactions

• Based on
  – confidence scores obtained for each protein mention
  – the disambiguation to find a UniProt ID per protein
  – the relevance score of the ranked sentence
  – the number of evidences found for each pair in the overall article.
## Results

### TABLE 1

(a) Interaction Pair Normalization Raw Results and (b) Ortholog-Mapped Results

<table>
<thead>
<tr>
<th>Result set</th>
<th>Results</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>AUC iP/R</th>
<th>Result set</th>
<th>Results</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>AUC iP/R</th>
</tr>
</thead>
<tbody>
<tr>
<td>42 s01 o</td>
<td>100</td>
<td>43.4</td>
<td>48.2</td>
<td>42.9</td>
<td>38.6</td>
<td>42 s01 o</td>
<td>100</td>
<td>67.3</td>
<td>52.2</td>
<td>55.1</td>
<td>49.1</td>
</tr>
<tr>
<td>18 run4 f</td>
<td>519</td>
<td>23.6</td>
<td>44.1</td>
<td>28.6</td>
<td>25.1</td>
<td>31 s18 o</td>
<td>340</td>
<td>46.7</td>
<td>52.6</td>
<td>44.5</td>
<td>37.5</td>
</tr>
<tr>
<td>14 run1 f</td>
<td>199</td>
<td>30.4</td>
<td>29.1</td>
<td>27.7</td>
<td>24.7</td>
<td>37 run3 f</td>
<td>584</td>
<td>44.1</td>
<td>54.7</td>
<td>41.8</td>
<td>44.0</td>
</tr>
<tr>
<td>37 run3 f</td>
<td>584</td>
<td>22.1</td>
<td>50.3</td>
<td>25.7</td>
<td>30.8</td>
<td>18 run3 f</td>
<td>519</td>
<td>41.1</td>
<td>51.9</td>
<td>40.9</td>
<td>38.0</td>
</tr>
<tr>
<td>42 s19 o</td>
<td>269</td>
<td>19.5</td>
<td>33.3</td>
<td>22.6</td>
<td>22.7</td>
<td>14 run1 f</td>
<td>199</td>
<td>55.6</td>
<td>36.4</td>
<td>40.5</td>
<td>33.3</td>
</tr>
<tr>
<td>42 s02 o</td>
<td>703</td>
<td>16.7</td>
<td>47.5</td>
<td>21.9</td>
<td>30.3</td>
<td>42 s19 o</td>
<td>269</td>
<td>47.5</td>
<td>42.5</td>
<td>39.6</td>
<td>37.8</td>
</tr>
<tr>
<td>31 s18 o</td>
<td>340</td>
<td>16.2</td>
<td>39.7</td>
<td>21.8</td>
<td>18.1</td>
<td>42 s02 o</td>
<td>703</td>
<td>35.3</td>
<td>52.5</td>
<td>35.5</td>
<td>44.5</td>
</tr>
<tr>
<td>42 s03 o</td>
<td>777</td>
<td>14.4</td>
<td>45.1</td>
<td>20.0</td>
<td>25.8</td>
<td>42 s03 o</td>
<td>777</td>
<td>33.7</td>
<td>50.7</td>
<td>34.7</td>
<td>39.6</td>
</tr>
<tr>
<td>42 s20 o</td>
<td>1362</td>
<td>10.5</td>
<td>53.5</td>
<td>16.7</td>
<td>33.3</td>
<td>10 s09 o</td>
<td>1784</td>
<td>25.5</td>
<td>71.8</td>
<td>34.0</td>
<td>60.1</td>
</tr>
<tr>
<td>10 s09 o</td>
<td>1784</td>
<td>8.7</td>
<td>59.1</td>
<td>14.5</td>
<td>42.8</td>
<td>32 l f</td>
<td>1697</td>
<td>28.3</td>
<td>55.9</td>
<td>32.5</td>
<td>39.3</td>
</tr>
<tr>
<td>32 s12 o</td>
<td>4328</td>
<td>3.1</td>
<td>57.9</td>
<td>5.8</td>
<td>29.2</td>
<td>42 s20 o</td>
<td>1362</td>
<td>27.4</td>
<td>63.8</td>
<td>32.1</td>
<td>52.6</td>
</tr>
<tr>
<td>22 run1 f</td>
<td>5133</td>
<td>20.4</td>
<td>69.7</td>
<td>26.6</td>
<td>56.3</td>
<td>22 run1 f</td>
<td>5133</td>
<td>20.4</td>
<td>69.7</td>
<td>26.6</td>
<td>56.3</td>
</tr>
</tbody>
</table>
# Results

## TABLE 2

(a) Interaction Pair Extraction Raw Results and (b) Ortholog-Mapped Results

<table>
<thead>
<tr>
<th>Result set</th>
<th>Results</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>AUC iP/R</th>
<th>Results</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>AUC iP/R</th>
</tr>
</thead>
<tbody>
<tr>
<td>18 run5</td>
<td>f</td>
<td>612</td>
<td>29.0</td>
<td>23.6</td>
<td>22.2</td>
<td>110</td>
<td>30.9</td>
<td>23.6</td>
<td>23.2</td>
<td>19.1</td>
</tr>
<tr>
<td>42 s01</td>
<td>o</td>
<td>407</td>
<td>21.3</td>
<td>29.6</td>
<td>22.1</td>
<td>128</td>
<td>38.0</td>
<td>29.6</td>
<td>30.1</td>
<td>25.7</td>
</tr>
<tr>
<td>37 run7s</td>
<td>f</td>
<td>2068</td>
<td>11.5</td>
<td>34.7</td>
<td>12.3</td>
<td>22.2</td>
<td>504</td>
<td>18.2</td>
<td>39.2</td>
<td>19.1</td>
</tr>
<tr>
<td>14 run4</td>
<td>f</td>
<td>572</td>
<td>12.0</td>
<td>14.5</td>
<td>12.1</td>
<td>13.4</td>
<td>192</td>
<td>20.9</td>
<td>17.4</td>
<td>16.2</td>
</tr>
<tr>
<td>31 UWMFull</td>
<td>f</td>
<td>86</td>
<td>18.0</td>
<td>10.8</td>
<td>11.6</td>
<td>9.6</td>
<td>49</td>
<td>37.2</td>
<td>21.0</td>
<td>23.6</td>
</tr>
<tr>
<td>32 s07</td>
<td>o</td>
<td>219</td>
<td>12.3</td>
<td>10.1</td>
<td>10.3</td>
<td>9.1</td>
<td>63</td>
<td>23.6</td>
<td>18.1</td>
<td>19.6</td>
</tr>
<tr>
<td>42 s20</td>
<td>o</td>
<td>27106</td>
<td>2.7</td>
<td>32.5</td>
<td>4.6</td>
<td>7.9</td>
<td>3552</td>
<td>11.1</td>
<td>51.5</td>
<td>14.6</td>
</tr>
<tr>
<td>42 s19</td>
<td>o</td>
<td>1306</td>
<td>2.7</td>
<td>9.8</td>
<td>3.7</td>
<td>3.3</td>
<td>432</td>
<td>13.5</td>
<td>16.6</td>
<td>11.9</td>
</tr>
<tr>
<td>42 s02</td>
<td>o</td>
<td>10488</td>
<td>2.2</td>
<td>22.6</td>
<td>3.1</td>
<td>6.1</td>
<td>1915</td>
<td>14.8</td>
<td>39.6</td>
<td>16.1</td>
</tr>
<tr>
<td>42 s03</td>
<td>o</td>
<td>10082</td>
<td>1.9</td>
<td>17.3</td>
<td>3.0</td>
<td>5.3</td>
<td>1453</td>
<td>14.5</td>
<td>32.5</td>
<td>15.9</td>
</tr>
<tr>
<td>51 precision2</td>
<td>f</td>
<td>25692</td>
<td>1.8</td>
<td>12.6</td>
<td>2.7</td>
<td>3.4</td>
<td>3233</td>
<td>7.5</td>
<td>19.6</td>
<td>8.4</td>
</tr>
<tr>
<td>22 s12</td>
<td>o</td>
<td>131243</td>
<td>0.4</td>
<td>33.3</td>
<td>0.8</td>
<td>6.9</td>
<td>(run1, f)</td>
<td>36317</td>
<td>10.6</td>
<td>64.2</td>
</tr>
</tbody>
</table>
Annotation Server Configurations

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>s01</td>
<td>pattern-based</td>
<td>2m 11s</td>
<td>22.1</td>
<td>30.1</td>
<td>42.9</td>
<td>55.1</td>
</tr>
<tr>
<td>s19</td>
<td>pattern-based, w/o relevance ranking</td>
<td>45s</td>
<td>3.7</td>
<td>11.9</td>
<td>22.6</td>
<td>39.6</td>
</tr>
<tr>
<td>s02</td>
<td>sentence-level co-occurrence</td>
<td>2m 12s</td>
<td>3.1</td>
<td>16.1</td>
<td>21.9</td>
<td>35.5</td>
</tr>
<tr>
<td>s03</td>
<td>sentence &amp; figure caption co-occurrence</td>
<td>1m 16s</td>
<td>3.0</td>
<td>15.9</td>
<td>20.0</td>
<td>34.7</td>
</tr>
<tr>
<td>s20</td>
<td>sentence &amp; figure caption co-occurrence, low confidence</td>
<td>11s</td>
<td>4.6</td>
<td>14.6</td>
<td>16.7</td>
<td>32.1</td>
</tr>
</tbody>
</table>
# Patterns with Highest Support

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>{w-interact-attachment} {w-prep-of} [interactor1] and [interactor2]</td>
<td>75</td>
</tr>
<tr>
<td>[interactor1] DASH [interactor2] {w-interact-attachment}</td>
<td>25</td>
</tr>
<tr>
<td>[interactor1] {w-interact-verb-s} {w-prep-with} [interactor2]</td>
<td>21</td>
</tr>
<tr>
<td>[interactor1] {w-interact-verb-s} [interactor2]</td>
<td>20</td>
</tr>
<tr>
<td>[interactor1] {w-is} {w-interact-verb-d} {w-prep-for} [interactor2]</td>
<td>18</td>
</tr>
<tr>
<td>{w-interact-attachment} {w-prep-of} [interactor1] {w-prep-with} [interactor2]</td>
<td>14</td>
</tr>
<tr>
<td>[interactor1] {w-is} {w-interact-verb-d} {w-prep-with} [interactor2]</td>
<td>14</td>
</tr>
<tr>
<td>[interactor1] {w-interact-verb-d} {w-prep-with} [interactor2]</td>
<td>12</td>
</tr>
<tr>
<td>[interactor1] and [interactor2] {w-interact-verb-i}</td>
<td>10</td>
</tr>
<tr>
<td>[interactor1] {w-interact-attachment} {w-prep-with} [interactor2]</td>
<td>8</td>
</tr>
<tr>
<td>{w-interact-attachment} {w-prep-of} [interactor1] and {w-determiner}? [interactor2]</td>
<td>8</td>
</tr>
<tr>
<td>{w-interact-verb-s} [interactor2] {w-interact-adjective-dashed}</td>
<td>6</td>
</tr>
<tr>
<td>{w-interact-verb-i} {w-prep-of} [interactor1] and [interactor2]</td>
<td>6</td>
</tr>
<tr>
<td>[interactor1] deficient {c-cell} COMMA {w-interact-regulation-noun-s} {w-prep-of} [interactor2]</td>
<td>6</td>
</tr>
<tr>
<td>{w-interact-verb-d} [interactor1] DASH [interactor2]</td>
<td>6</td>
</tr>
<tr>
<td>[interactor1] as {w-determiner}? {w-adjective}? [interactor2] {w-interact-attachment}</td>
<td>5</td>
</tr>
<tr>
<td>[interactor1] {w-interact-verb-s} {w-prep-with} {w-determiner}? [interactor2]</td>
<td>5</td>
</tr>
<tr>
<td>[interactor1] {w-in} [interactor2] {w-interact-adjective-dashed}</td>
<td>5</td>
</tr>
<tr>
<td>[interactor1] {w-aux} {w-interact-verb-i} [interactor2]</td>
<td>4</td>
</tr>
<tr>
<td>[interactor1] {w-is} {w-interact-verb-d} {w-prep-for} [interactor2]</td>
<td>4</td>
</tr>
<tr>
<td>[interactor1] {w-interact-activity-noun-s}</td>
<td>4</td>
</tr>
<tr>
<td>[interactor1] and [interactor2] {w-interact-verb-s}</td>
<td>4</td>
</tr>
<tr>
<td>[interactor1] {w-interact-verb-s} {w-prep-with} {w-determiner}? {w-adjective}? domain {w-prep-of} [interactor2]</td>
<td>4</td>
</tr>
<tr>
<td>[interactor1] mediates {w-interact-attachment} {w-prep-with} [interactor2]</td>
<td>4</td>
</tr>
</tbody>
</table>
Positions of Interactions

![Heatmap showing positions of interactions in various sections of a document.](image)
Next Week

• 1 paper to read about reconstruction of signaling networks: