Fine-Grained Protein Mutation Extraction from Biological Literature

Rui Wang  
Computational Linguistics  
Saarland University  
Germany

Shirley W.I. Siu & Rainer A. Böckmann  
Theoretical & Computational Membrane Biology  
Saarland University  
Germany
Motivation

Figure 1. Growth in the Biomedical Literature, 1986–2005

(Hunter and Cohen, 2006)
NLP → BioNLP

- Natural Language Processing (NLP)
  - Named-Entity (NE) extraction
  - NE relation extraction
  - ontology construction
  - ...

- NLP in the Biological domain (BioNLP)
  - Protein-Protein Interaction
  - Gene ontology construction
  - Mutation Extraction
  - ...
Protein Mutation

(Benedix et al., 2009)
Various Expressions

- ARG23ALA or R23A
- His-230 and His-309 were mutated to phenylalanine
- Ser172 were selected and mutated to Phe and Ala
- The replacement of Ile209 with an Ala residue
- D27 in ZMPDC was altered to alanine
- Asn-Gly pairs were changed into Leu (Asn244, Asn255, Asn437) or Ala (Asn276)
- Each of the seven Cys residues of rrSE were individually mutated to Ala.
Related Work

- Mutation Finder (Caporaso et al., 2007)
  - Mutation Extraction based on regular expressions

- Mutation Miner (Witte and Baker, 2005)
  - Relations between proteins and mutations
An Example from MedLine

- CcP (E290K) has a charge-reversal mutation in the tight-binding domain, which should weaken binding, and it weakens the 1:1 complex; $K_1$ decreases 20-fold at 18 mM ionic strength.

![Diagram](image_url)  

Figure 1. Binary relation for the MutationExperiment object created from Ex1.
Template

• Example
  • CcP (E290K) has a charge-reversal mutation in the tight-binding domain, which should weaken binding, and it weakens the 1:1 complex; $K_1$ decreases 20-fold at 18 mM ionic strength.

• MutationExperiment
  • <List<protein,List<mutation>>,List<condition>,List<result>>
Extraction Approach
Linguistic Preprocessing

• Tokenization
  • E.g. \( (5S,6E,8Z,11Z,14Z)\)-5-hydroperoxy-6,8,11,14-eicosatetraenoic acid (5S-HpETE)
  • ABNER

• POS Tagging
  • LingPipe trained on GENIA corpus

• Dependency Parsing
  • MSTParser
An Example

His-230 and His-309 were mutated to phenylalanine.
Information Extraction

• Object Recognition
  • Protein Name

• Mutation
  • Wild-type, Position, and Mutant

• Experimental conditions
  • The temperature and the pH value

• Results
  • Experimental results, DeltaG, DeltaDeltaG, K(cat), etc.
Information Extraction (cont.)

• Relation Extraction
  • Find the dependency path between A and B*;
  • Find all the common ancestor verbs for A and B;

*where A and B are objects extracted before

• Example
  • MBP-H213A and H216A TfdA have elevated $K(m)$ values for 2,4-D, and the former showed a decreased $k(cat)$, suggesting these residues may affect substrate binding or catalysis.
Experiments

- **Data**
  - MedLine: 922 abstracts with the keyword *mutagenesis*

- **Gold standard**
  - Protein Mutant Database (PMD)
    - over 30 years manually mutation extraction from 45239 publications

- **Baseline**
  - MutationFinder

- **Metrics**
  - Precision, Relative Recall, and F-Score
Results

- 3818 mutations
- A large increase in recall, a drop in precision

Table 1. Results of mutation extraction

<table>
<thead>
<tr>
<th></th>
<th>MF (Baseline)</th>
<th>MF+ME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>94.3</td>
<td>89.4</td>
</tr>
<tr>
<td>Relative Recall</td>
<td>88.3</td>
<td>100.0</td>
</tr>
<tr>
<td>F-Score</td>
<td><strong>91.2</strong></td>
<td><strong>94.4</strong></td>
</tr>
</tbody>
</table>
Results (cont.)

Table 2. Results of relation extraction

<table>
<thead>
<tr>
<th></th>
<th>Exp. Conditions</th>
<th>Exp. Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unlabeled Precision</td>
<td>69.6</td>
<td>88.5</td>
</tr>
<tr>
<td>Labeled Precision</td>
<td>/</td>
<td>84.6</td>
</tr>
<tr>
<td>Labeled Accuracy</td>
<td>/</td>
<td>92.3</td>
</tr>
</tbody>
</table>

- Manually read about 15% of the data
- Unlabeled vs. Labeled
- Mutation-condition : mutation-result ~ 13.5%
Qualitative Analysis

- E.g. We *mutated* Ala137 of *T. brucei* glycerol kinase *into* a serine
  - Ala137Ser

- E.g. Mutants of *tyrosine hydroxylase* with alanine substituted for Phe300
  - Although tyrosine hydroxylase is a protein, tyrosine itself is a residue, thus Phe300Tyr was wrongly reported.
Qualitative Analysis (cont.)

- E.g. Asn-185 of CitS was mutated to Val and Glu-194 was mutated to Gln
  - Parsing errors

- E.g. Glu112, Ser113 and Ser115 that … replaced by Pro, Gly and Glu, respectively

- E.g. MBP-H213A and H216A TfdA have elevated K(m) values for 2,4-D, and the former showed a decreased k(cat) …
Conclusion

- Improved the mutation extraction through combining linguistic processing with a regular-expression-based system
- Explored the extraction of relations between the mutations and the experimental measurements
Future Work

- More sophisticated linguistic analysis
  - Deep language processing
  - Cross sentence

- From abstracts to full texts
  - More information
  - More complex relations
Acknowledgements

- Alexander Benedix
- Rui Wang is supported by PIRE scholarship PhD program
- Shirley Siu and Rainer Böckmann are supported Graduate School 1276/1 and by DFG BIZ 4/1
References


References (cont.)


Thank you!

Questions?