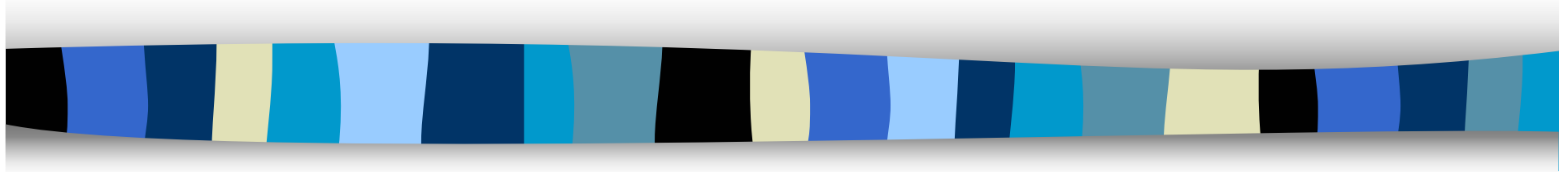


# Fine-Grained Protein Mutation Extraction from Biological Literature



*Rui Wang*

*Computational Linguistics*

*Saarland University*

*Germany*

*Shirley W.I. Siu & Rainer A. Böckmann*

*Theoretical & Computational Membrane Biology*

*Saarland University*

*Germany*

# Motivation

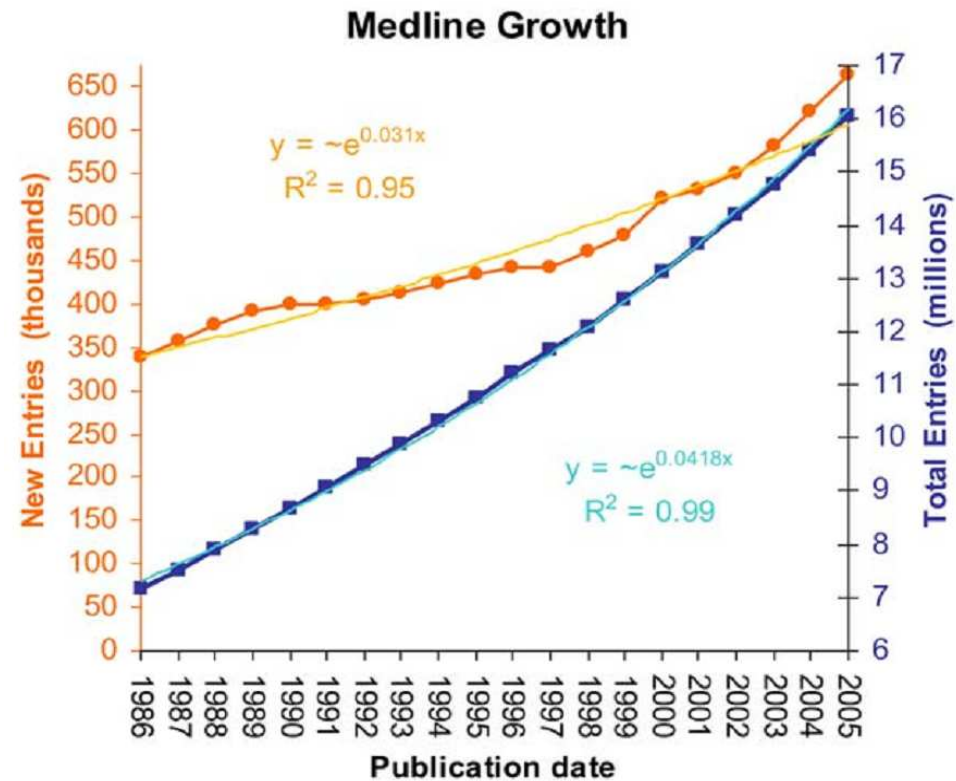


Figure 1. Growth in the Biomedical Literature, 1986–2005

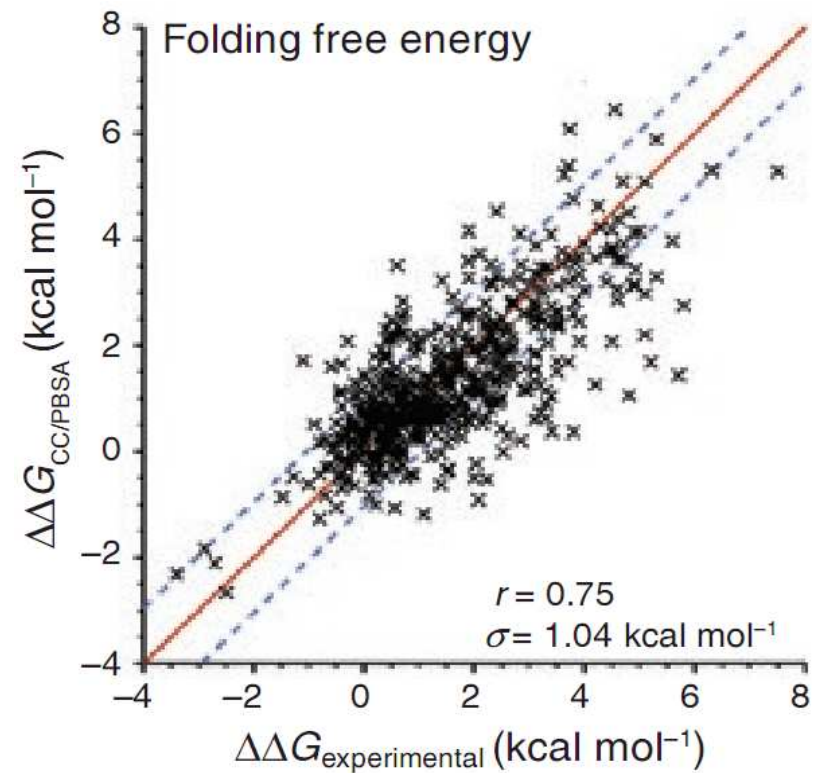
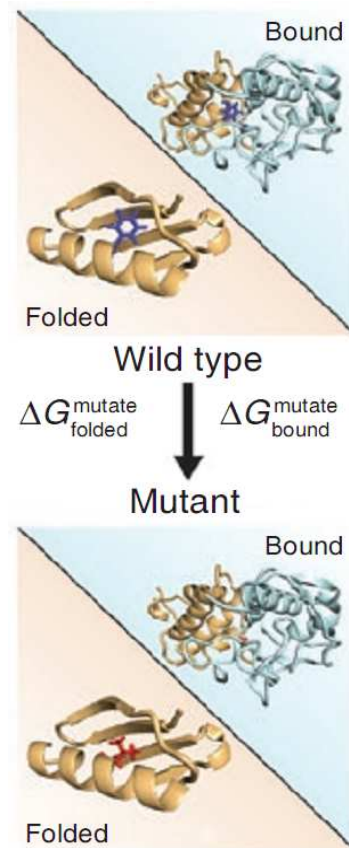
(Hunter and Cohen, 2006)



# 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.

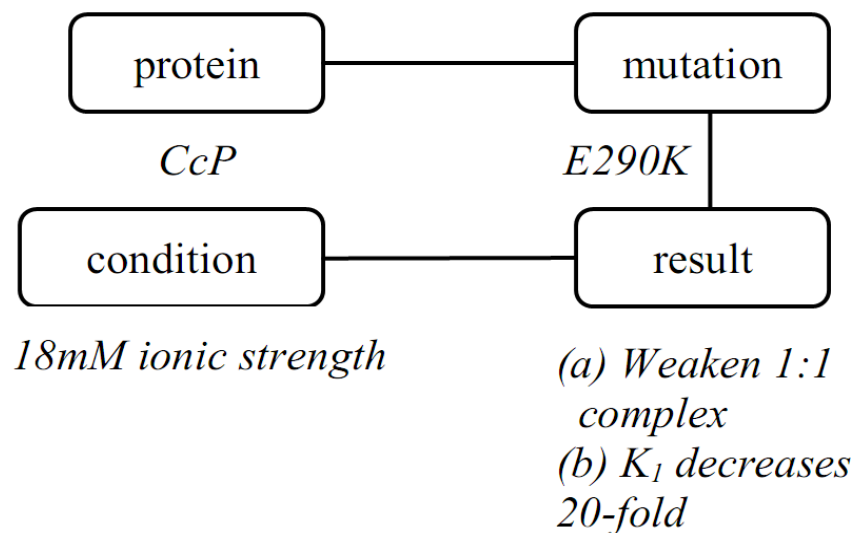


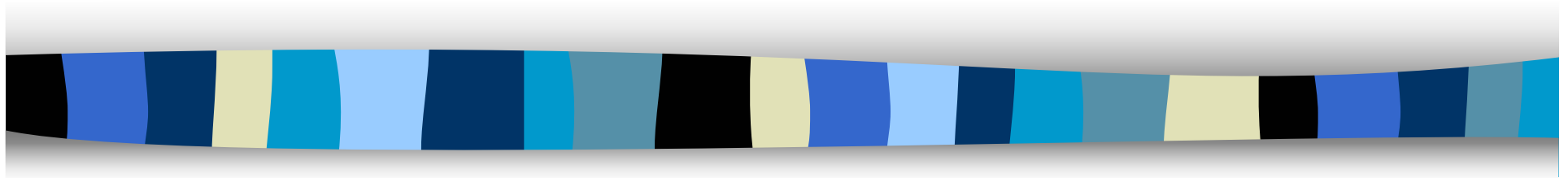
Figure 1. Binary relation for the MutationExperiment object created from Ex1.



# 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; **K1 decreases 20-fold** at **18 mM ionic strength**.
- MutationExperiment
  - <List<protein,List<mutation>>,List<condition>>  
List<result>>

# Extraction Approach

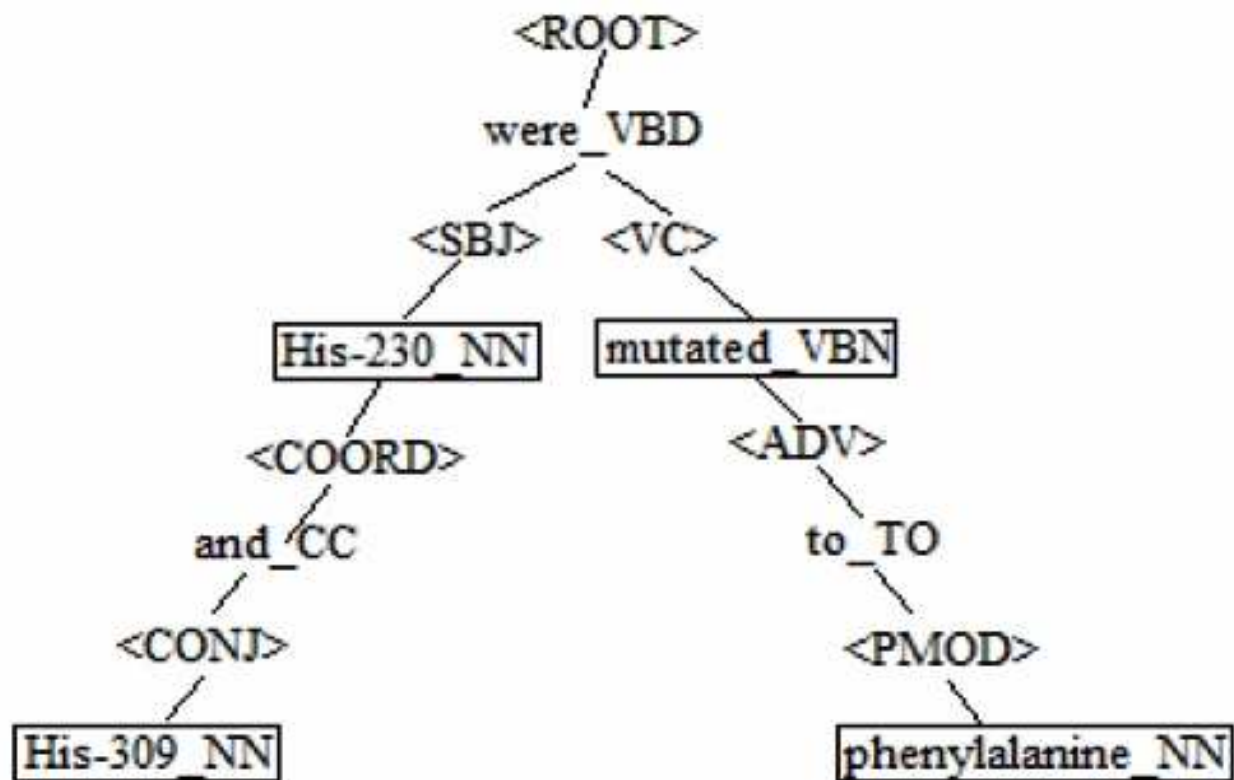




# Linguistic Preprocessing

- Tokenization
  - E.g. *(5S,6E,8Z,11Z,14Z)-5-hydroperoxy-6,8,11,14-eicos atetraenoic 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

**Table 1. Results of mutation extraction**

	MF (Baseline)	MF+ME
Precision	94.3	89.4
Relative Recall	88.3	100.0
F-Score	<b>91.2</b>	<b>94.4</b>

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



## Results (cont.)

**Table 2. Results of relation extraction**

	Exp. Conditions	Exp. Results
Unlabeled Precision	<b>69.6</b>	<b>88.5</b>
Labeled Precision	/	84.6
Labeled Accuracy	/	92.3

- 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

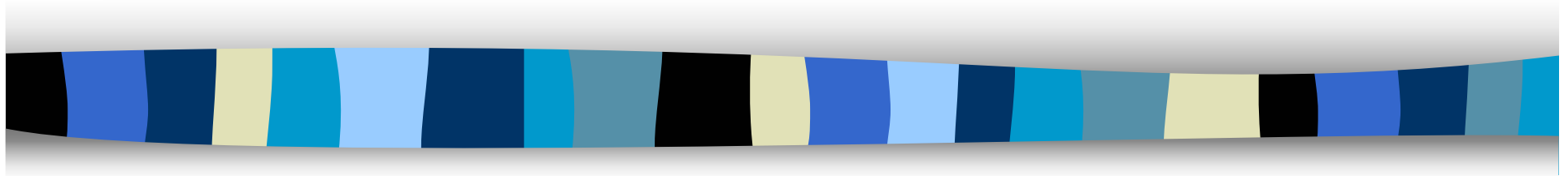
- Lawrence Hunter and K. Bretonnel Cohen. 2006. Biomedical Language Processing: What's Beyond PubMed? *Molecular Cell*, Volume 21, Issue 5, 3 March 2006, Pages 589-594.
- Alexander Benedix, Caroline M. Becker, Bert L. de Groot, Amedeo Caflisch, Rainer A. Böckmann. Predicting Free Energy Changes Using Structural Ensembles. *Nature Methods* 6 (2009) 3-4
- T. Kawabata, M. Ota, and K. Nishikawa. The protein mutant database. *Nucleic Acids Res.* 27(1), 1999, pp. 355-7.
- R. Witte and C.J.O. Baker. Combining biological databases and text mining to support new bioinformatics applications. *NLDB 2005*, LNCS 3513, pp. 310–321, 2005.
- J.G. Caporaso, W.A. Baumgartner Jr, D.A. Randolph, K.B. Cohen, and L. Hunter. MutationFinder: a high-performance system for extracting point mutation mentions from text. *Bioinformatics*, 23(14), 2007, pp. 1862-1865.



## References (cont.)

- B. Settles. ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics*, 21(14), 2005, pp. 3191-3192.
- Alias-i, LingPipe 3.6.0., <http://alias-i.com/lingpipe>, 2008.
- McDonald, Ryan, Fernando Pereira, Kiril Ribarov, and Jan Hajic. 2005. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of HLT-EMNLP 2005*, pages 523–530, Vancouver, Canada.
- R. Wang and G. Neumann. 2007a. Recognizing Textual Entailment Using a Subsequence Kernel Method. In *Proceedings of AAAI-2007*, Vancouver.

Thank you!



Questions?