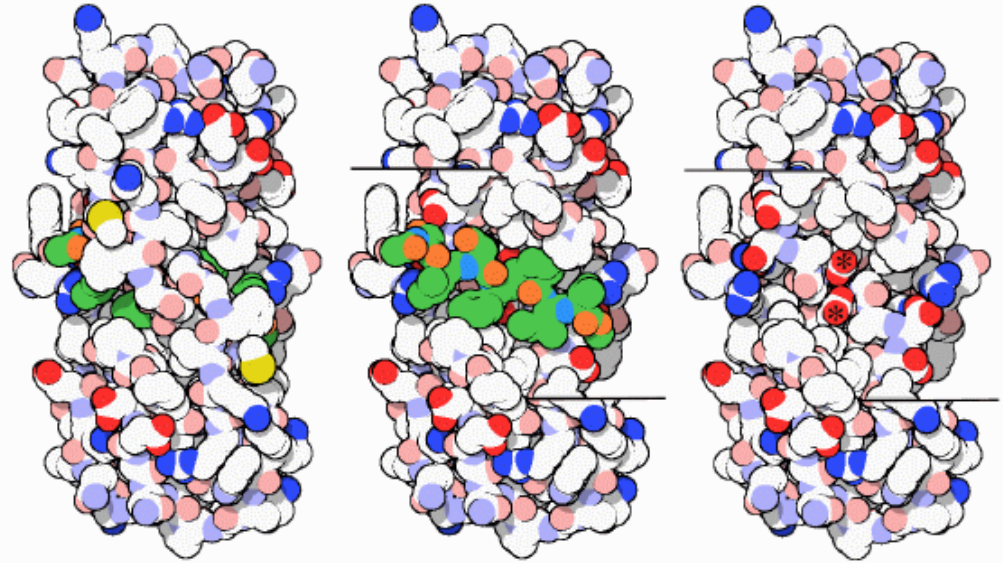


A Sentence
Recognizer for
Mutant Protein
Structure Studies:
Toward Intelligent
Systems for the
Management of
Structural Biology Data



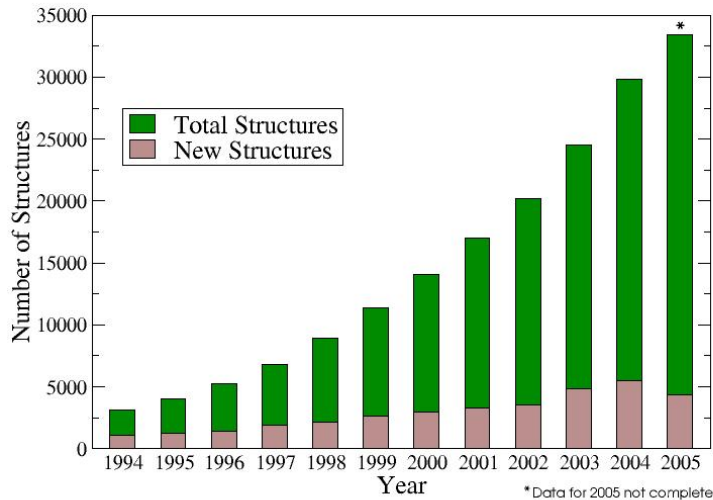
J. Gregory Caporaso^{1,2}, K. Bretonnel Cohen¹, Lawrence Hunter¹

University of Colorado Health Sciences Center

¹Center for Computational Pharmacology

²Program in Biomolecular Structure

Growth of the Protein Data Bank (PDB)



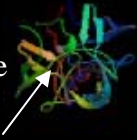
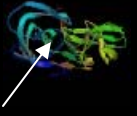
- 50 – 100 new structures per week
- Two-dimensional growth

- 213 HIV-1 protease structures
 - 32% mutant studies
- Difficult to compile results of related studies
 - ~136 journal articles

Number of PDB Hits for some common proteins		
PDB Query	Hits	
	Apr '05	Nov '05
Alcohol Dehydrogenase	106	119
HIV-1 Protease	206	213
Carbonic Anhydrase	207	220
p53	99	105
Actin	315	344
Anthrax Toxin	15	16
Estrogen Receptor	44	49
Rubisco	36	39
Ubiquitin	210	262

Machine learning techniques to identify/extract site-specific mutagenesis results



Wild-type Protein Information		
PDB ID: HIV1	HIV-1 Protease	
Source: Human Immunodeficiency Virus-1		
Mutated Variant(s)		
1AXA	A28S	modified active site H-bonds 
Relevant PMIDs: 9521105 9884625		
1DAZ	Q7K, L33I, L63I, C67A	disrupted hydrophobic surface 
Relevant PMIDs: 10429209 10987654		

- First step in information extraction is identification
- Sentence classifiers trained to identify sentences mapping protein mutations to structural changes

✓ “The G55A mutation escalated the strain in the second beta turn...”

✗ “The crystal structure of the phosphorylated, activated form of the insulin receptor tyrosine kinase has been determined at 1.9 Å resolution.”

✗ “The mutations targeted key residues L212Glu and L213Asp of this transmembrane protein-cofactor complex.”

✓ “The alanine substitutions caused an expansion of the cavity rather than its collapse.”

Naïve Bayes and SVM classifiers identify relevant sentences with precision and recall approaching that of humans.

