

Functional Analysis of Cystic Fibrosis Associated Point Mutations By Text Mining

Lawrence C. Lee^{1,3*}, Fred E. Cohen^{1,2}

¹Department of Cellular and Molecular Pharmacology, ²Department of Biophysics, ³Program in Biological and Medical Informatics, University of California, San Francisco. *lawrence.lee@cmpharm.ucsf.edu

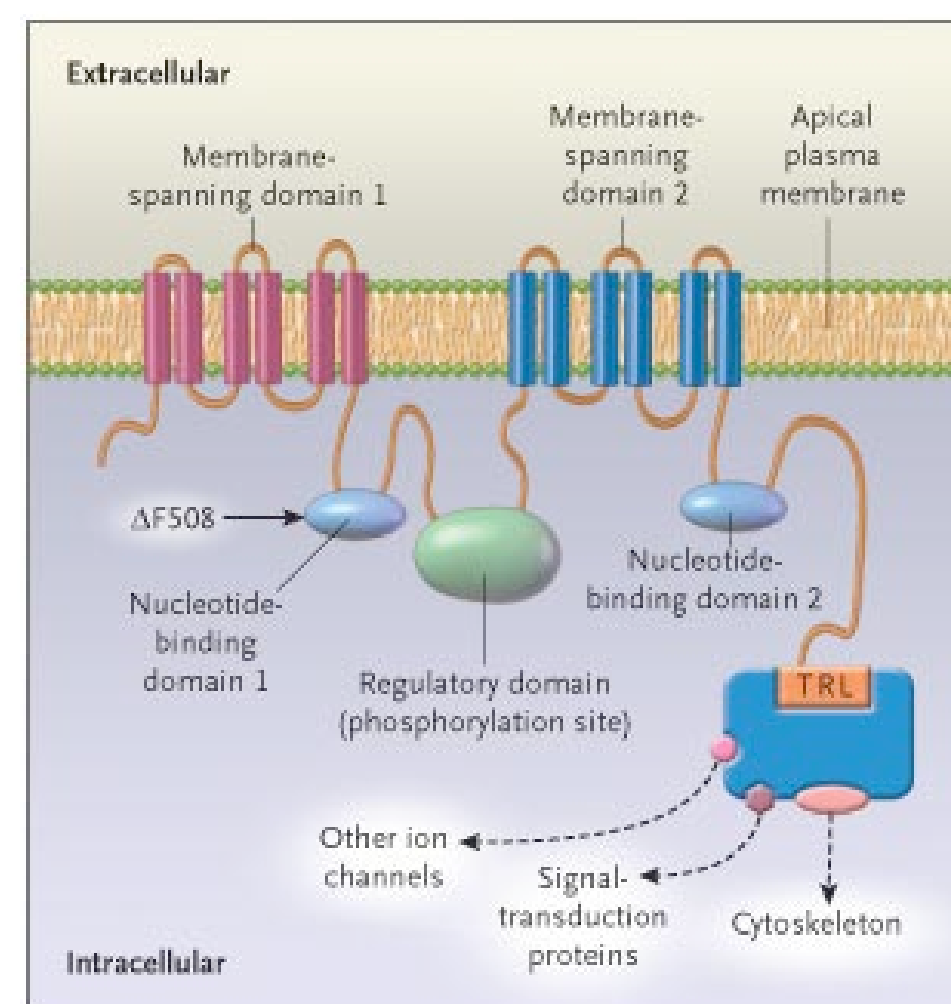


Introduction

A wealth of protein point mutation information is contained exclusively within the scientific literature and not in any curated database. Text mining can be used to extract such information without manually reading all the articles. We have previously developed a program, Mutation GraB, which identifies and extracts point mutations from journal literature. Our goal is to understand the underlying mechanisms of heritable disorders by analyzing the effects of protein point mutations. We first downloaded articles related to cystic fibrosis mutations and extracted the point mutations within. Then we searched the point mutation-containing sentences for verb phrase patterns to analyze the effects of the point mutations.

Background

Cystic Fibrosis is a disease caused by mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) protein. It affects approximately 30,000 people in the United States, and 10 million other Americans are carriers for the defective gene. The defect in the CFTR protein, which is a Cl⁻ channel, manifests itself in many different disease phenotypes, from mucus buildup in the lungs to decreased sperm production in males. While 70% of CFTR mutations in the population are ΔF508, over 1400 mutations have been found. It would benefit the CF community to be able to characterize the functional effects of the point mutations.

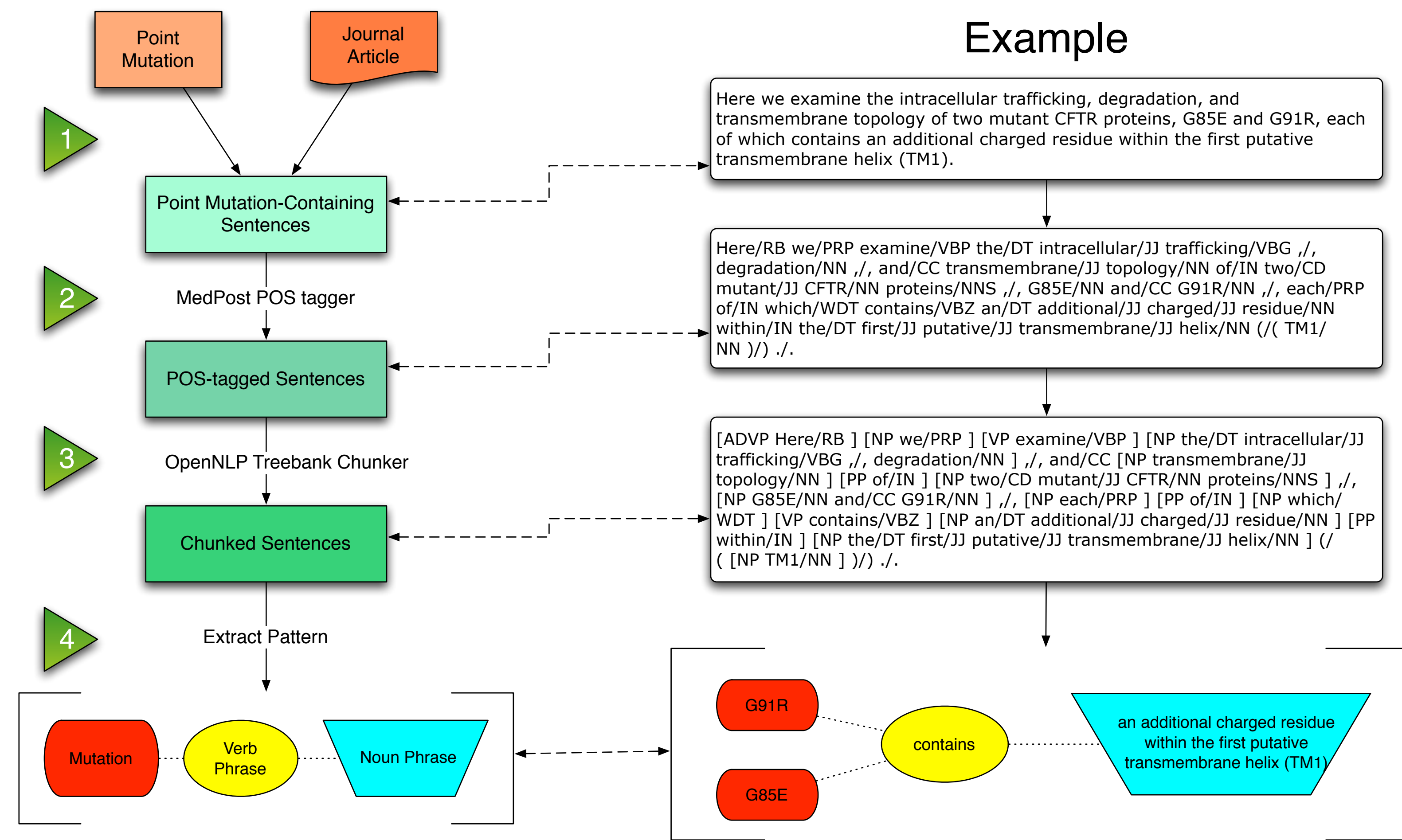


Rowe et al., 2005

Objectives

- 1 Identify and extract point mutations from CF related articles retrieved from PubMed.
- 2 Compare text-mined mutations to those contained in manually curated databases.
- 3 Extract the functional effects of the point mutations.

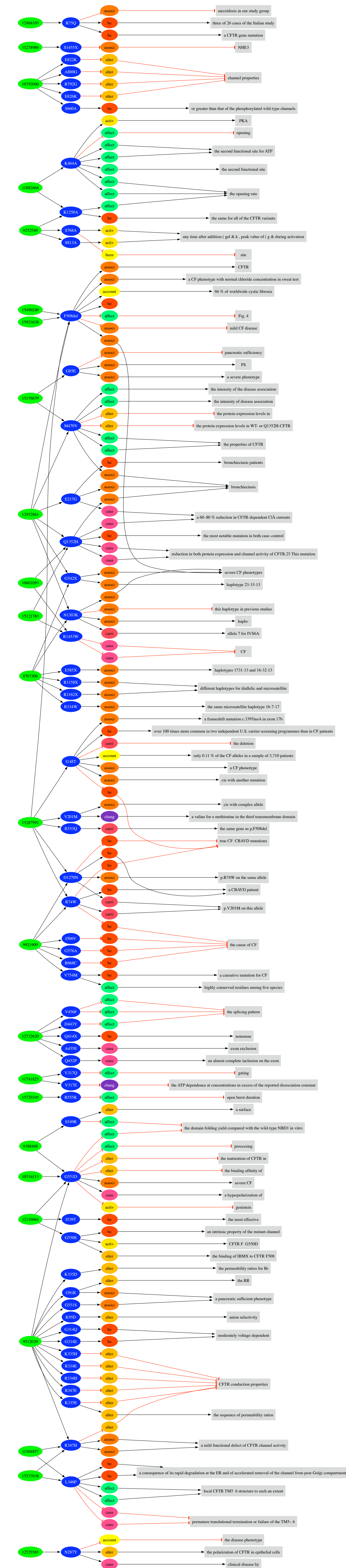
Methods



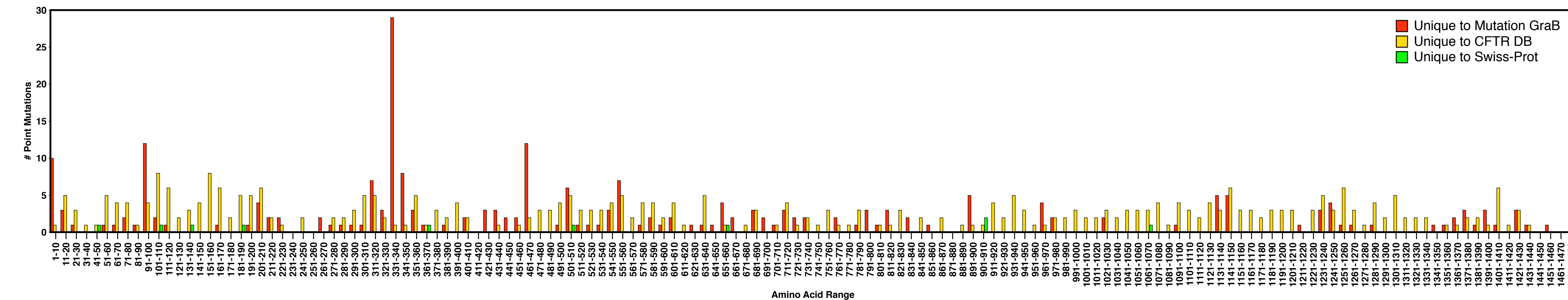
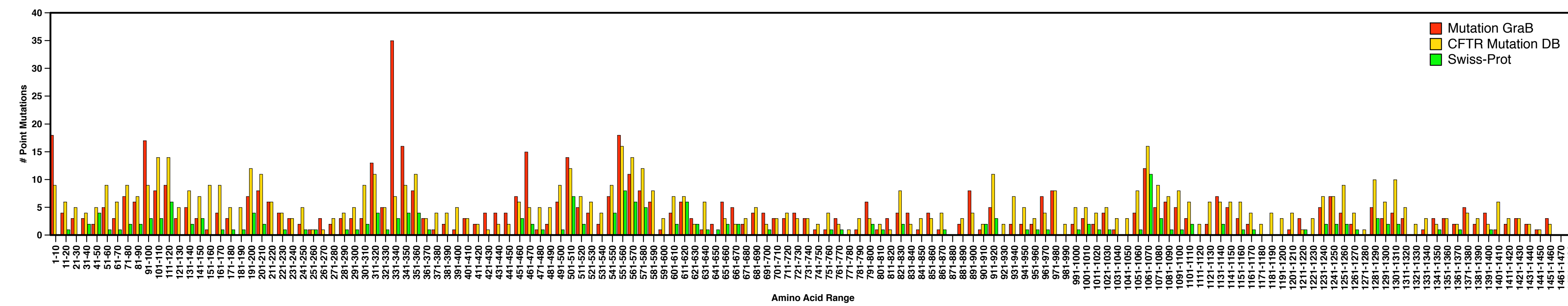
- 1 Identify sentences which contain point mutations.
- 2 Use the MedPost POS tagger to tag part-of-speech for each word.
- 3 Use the OpenNLP Treebank Chunker to chunk words into phrases.
- 4 Identify <MP><VP><NP> pattern where MP is the noun phrase containing the point mutation

Article→Mutation→Verb→Noun Relationship

A graph showing the relationships between the articles, point mutations, verbs, and noun phrases. The leftmost column represents the article PMIDs, the second column shows the point mutations, third column shows the verbs, and the right-most column shows the noun phrases. The black lines represent positive verb associations (i.e. "increase") while the red lines represent negative ones (i.e. "does not increase").



CFTR Protein Point Mutations by Amino Acid Position



Most Frequent Verbs and Nouns

Verb	Count	Noun	Count
is	209	the mutation rate	36
was	184	unknown total	26
associate	131	28 % of frameshift mutations	25
are	131	an elucigena cf20 kit	24
show	116	snp typing with a masscode system	22
found	107	this study	22
were	107	the portuguese population	20
use	103	the cystic fibrosis genetic analysis consortium	19
be	100	it	16
had	94	disease-associated microsatellite t5-7	14
exhibit	82	genistein	14
increase	81	they	14
have	76	ivs8-6t cfrdele2	12
occur	73	that	12
cause	73	cf	11
detect	66	congenital bilateral absence of the vas deferens	11
describe	65	reverse dot blot strips	11
suggest	58	the instructions of highsmith et al.	11
reduce	57	variant ivs8-5t	11
report	54	alone	10

Results and Future Work

We downloaded 536 CFTR mutation articles from PubMed and ran Mutation GraB on the set. We were able to retrieve 3792 protein point mutation terms representing 593 unique point mutations. Examining these point mutations against those in the CFTR Mutation Database and Swiss-Prot shows that many unique point mutations are not within the curated databases. Our preliminary mutation effect extraction efforts show that verb and noun phrases related to the point mutation can be identified. Future work involves developing a method to find the most relevant verb and noun phrases.

References

- Lee LC, Horn F, Cohen FE (2007) Automatic extraction of protein point mutations using a graph bigram association. PLoS Comput Biol 3(2): e16. doi:10.1371/journal.pcbi.0030016
- Cystic Fibrosis Mutation Database: <http://www.genet.sickkids.on.ca/cftr/app>
- Open NLP Tools: <http://opennlp.sourceforge.net/>
- L. Smith , T. Rindfleisch , and W. J. Wilbur MedPost: a part-of-speech tagger for biomedical text, Bioinformatics Advance Access published on September 22, 2004, DOI 10.1093/bioinformatics/bth227. Bioinformatics 20: 2320-2321.