

---

# Identifying Gene Specific Variations in Biomedical Text

---



**Fraunhofer** Institute  
Algorithms and  
Scientific Computing

**Roman Klinger**

Laura I. Furlong, Christoph M. Friedrich, Heinz Theodor Mevissen, Juliane Fluck,  
Ferran Sanz, and Martin Hofmann-Apitius  
25 September 2007

---

# Outline

---

- 1 Introduction
- 2 Machine Learning Method: Probabilistic Graphical Models
- 3 Named Entity Recognition of Single Nucleotide Polymorphisms
- 4 Summary

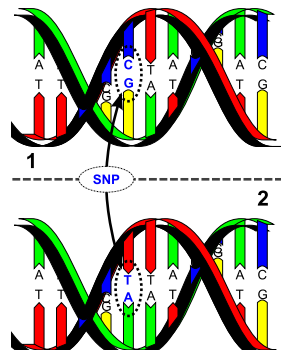


# Introduction

## Single Nucleotide Polymorphism (SNP)

Wikipedia<sup>1</sup>:

- “DNA sequence variation occurring when a single nucleotide—A, T, C, or G—in the genome differs between members of a species”
- “Successful” point mutations in a population:  $> 1\%$



<sup>1</sup>[http://de.wikipedia.org/wiki/Single\\_Nucleotide\\_Polymorphism](http://de.wikipedia.org/wiki/Single_Nucleotide_Polymorphism)

# Introduction

- Database with SNP information available
  - **dbSNP**<sup>1</sup>:  
51,312,474 variations for 43 different organisms
- Some Information of influence of **genetic variations** on **diseases** or **cellular processes** only in text
  - Interesting for diagnostics, pharmacogenomics, understanding molecular processes

---

<sup>1</sup><http://www.ncbi.nlm.nih.gov/projects/SNP/>



# Introduction

## Goal

- Find SNP mentions in text...

⇒ Search for articles with [variations] and [a special disease] or...

## Goal

- ...and related entries in dbSNP (normalization)

⇒ Find available information about that SNP in database  
**Combine information from text and database**



# Single Nucleotide Polymorphism—examples

**BACKGROUND AND PURPOSE:** The **collagen alpha2(I)** gene (**COL1A2**) on chromosome 7q22.1, a positional and functional candidate for intracranial aneurysm (IA), was extensively screened for susceptibility in Japanese IA patients. **METHODS:** Twenty-one **single nucleotide polymorphisms (SNPs)** of **COL1A2** were genotyped in genomic DNA from 260 IA patients (including 115 familial cases) (mean age, 59.9 years) and 293 controls (mean age, 61.6 years). Differences in allelic and genotypic frequencies between the patients and controls were evaluated with the  $\chi^2$  test. Circular dichroism spectrometry was monitored with collagen-related peptides that mimic triple-helical models of type I collagen with **Ala-459** and **Pro-459** to estimate the conformation and stability of alterations. **RESULTS:** Significant genotypic association in the dominant model was observed between an **exonic SNP** of **COL1A2** and familial IA patients ( $\chi^2=11.08$ ;  $df=1$ ;  $P=0.00087$ ; odds ratio=3.19; 95% CI, 2.22 to 6.50). This **SNP** induces **Ala** to **Pro substitution** at **amino acid 459**, located on a triple-helical domain. Circular dichroism spectra showed that the **Pro-459** peptide had a higher thermal stability than the **Ala-459** peptide. **CONCLUSIONS:** The variant of **COL1A2**

maps to rs number

states, locations, genes, types

PMID: 14739420, T. Yoneyama et. al: "Collagen type I alpha2 (COL1A2) is the susceptible gene for intracranial aneurysms".



# Methods for Text Mining

## ■ Dictionary based

- + High precision
- + Easy to normalize  
(mapping to a unique identifier)
- No "new" entities

## ■ Rule based

- + New entities findable
- Knowledge has to be stated explicitly

## ■ Non-Dictionary based

- + Better generalization

## ■ Machine Learning

- + New entities findable
- + Learning from examples



# Methods for Text Mining

---

- Neural Networks,  
Support Vector Machines
  - ⇒ Mainly for classification
- Methods especially for sequences exist
  - Conditional Random Fields used by 11 of 21 participants in BioCreative 2007 Challenge



# Machine Learning for Text Mining

## Goal

Map text sequence to label sequence:

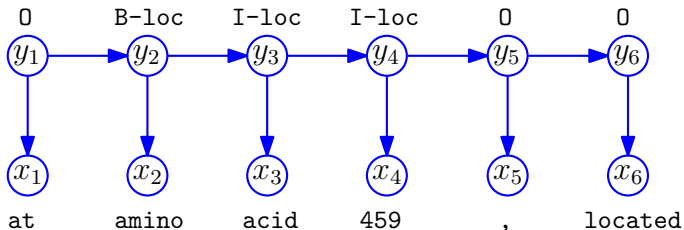
$x =$	at	amino	acid	459	,	located
$y =$	0	B-loc	I-loc	I-loc	0	0

- Learning: ⇐ How to learn that model?  
Given training set  $E_t = \{(\vec{x}_1, \vec{y}_1), \dots, (\vec{x}_n, \vec{y}_n)\}$
- Question: ⇐ How get that from the model?  
Given new examples  $E = \{(\vec{x}_1, ?), \dots, (\vec{x}_m, ?)\}$

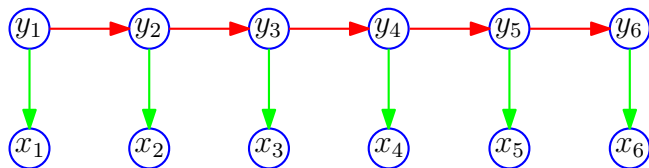


# Directed Graphical Models on Text

- Given text sequence:  $\vec{x} = (\text{at}, \text{amino}, \text{acid}, 459, ,, \text{located})$
- Label sequence:  $\vec{y} = (0, \text{B-loc}, \text{I-loc}, \text{I-loc}, 0, 0)$
- Hidden Markov Model (HMM):



# Directed Graphical Models on Text (HMM)



**a** Initial probability

**b** Transition probability

**c** Observation probability

$$p(\vec{y}, \vec{x}) = p(y_1) \cdot p(x_1|y_1)p(y_2|y_1) \cdots p(x_i|y_i)p(y_i|y_{i-1})$$

$$p(\vec{y}, \vec{x}) = \underbrace{p(y_1)}_a \prod_{i=1}^n \underbrace{p(y_i|y_{i-1})}_b \underbrace{p(x_i|y_i)}_c$$

## ■ Problems:

- No dependence in text, would be computational expensive
- Only dependence between directly succeeding labels

# Undirected Graphical Models

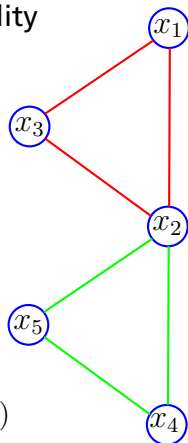
- Decomposition (**factorization**) of a given probability distribution:

$$p(\vec{x}) = \frac{1}{Z} \prod_C \Psi_C(\vec{x}_C)$$

- $C$ : maximal cliques in the independency graph
- $\Psi_C$ : potential functions
- $\vec{x}_C$ : subset of  $\vec{x}$  relevant for  $C$
- $Z$ : normalization given by  $Z = \sum_{\vec{x} \in \mathcal{X}} \prod_C \Psi_C(\vec{x}_C)$

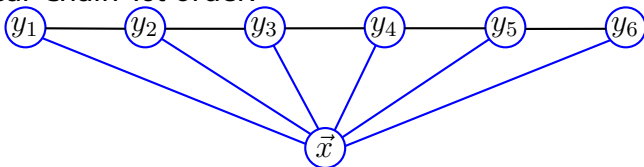
- Concrete example:

$$p(\vec{x}) = \frac{1}{Z} \cdot \Psi_1(x_1, x_2, x_3) \cdot \Psi_2(x_2, x_4, x_5)$$



# Conditional Random Fields

## ■ Linear Chain 1st order:



at amino acid 459 , located

$$p(\vec{y}|\vec{x}) = \frac{1}{Z(\vec{x})} \prod_{j=1}^n \Psi_j(y_j, y_{j-1}, \vec{x})$$

$$Z(\vec{x}) = \sum_{\vec{y}'} \prod_{j=1}^n \Psi_j(y'_j, y'_{j-1}, \vec{x})$$

$$\Psi_j(y_j, y_{j-1}, \vec{x}) = \exp \left( \sum_{i=1}^m \lambda_i f_i(y_{j-1}, y_j, \vec{x}, j) \right)$$

■ Training: Concave Optimization

■ Inference: Viterbi Algorithm



# Application: Named Entity Recognition of Mentions of Single Nucleotide Polymorphisms

Roman Klinger<sup>1</sup>, Laura I. Furlong<sup>2</sup>, Christoph M. Friedrich<sup>1</sup>, Heinz Theodor Mevissen<sup>1</sup>, Juliane Fluck<sup>1</sup>, Ferran Sanz<sup>2</sup>, and Martin Hofmann-Apitius<sup>1</sup>.

Identifying gene specific variations in biomedical text. *Journal of Bioinformatics and Computational Biology, Special Issue: Making Sense of Mutations requires Knowledge Management*, December 2007.

---

<sup>1</sup>Fraunhofer SCAI, Sankt Augustin, Germany

<sup>2</sup>Research Unit on Biomedical Informatics (GRIB) IMIM/UPF. Barcelona, Spain



# Single Nucleotide Polymorphism—other works

- **MuteXt, MutationGraB, MEMA, MutationFinder:**  
Regular Expressions
  - ⇒ limited due to the use of RegEx (standardized nomenclature is not always used), no mapping to database
- **VTag:** CRF-based approach
  - ⇒ in special context of cancer, no mapping to database
- **OSIRIS:** Query-expansion:  
for all SNPs of a found gene: PubMed query
  - ⇒ slow, limited to results of PubMed search engine



# Single Nucleotide Polymorphism—Classes

- States
  - locations
  - types
  - genes
  - rs numbers
- } Machine Learning (CRF)
- } Dictionary based (map. to EntrezGene, ProMiner)
- } Regular Expressions (ProMiner)

...ratio=3.19; 95% CI, 2.22 to 6.50). This **SNP** induces **Ala** to **Pro substitution** at **amino acid 459**, located on a triple-helical domain. Circular dichroism spectra showed that the **Pro-459** peptide had a higher thermal stability than the **Ala-459** peptide. **CONCLUSIONS:** The variant of **COL1A2** could be a genetic risk factor for IA patients with family history.

Annotation Guidelines from University of Pennsylvania:

<http://www ldc.upenn.edu/mamandel/itre/annotators/onco/definitions.html>



# Single Nucleotide Polymorphism—Corpus

## ■ Retrieve relevant documents from Medline

```
"Pathological Conditions, Signs and Symptoms"[MeSH] AND  
"Polymorphism, Single Nucleotide"[MeSH] AND "Humans"[MeSH]  
AND hasabstract[text] AND English[lang] AND  
("2004/01/01"[PDAT] : "2005/01/01"[PDAT]) AND "Chemicals  
and Drugs Category"[MeSH]
```

## ■ Manual annotation of 105 articles

## ■ Addition of 102 articles with negative examples (enhances precision)

- "14 between 22", "1968-1969", "21 ml", "56-year-old"...

⇒ Training set of 207 articles



# Single Nucleotide Polymorphism—Tokenization

- To get a sequence of tokens, **tokenize** the text:

## Tokenization

This | **SNP** | induces | **Ala** | to | **Pro** | **substitution** | at | **amino acid**  
**459** | , | located | on | a | triple | - | helical | domain | . | Circular |  
dichroism | spectra | showed | that | the | **Pro** | - | **459** | peptide |  
had | a | higher | thermal | stability | than | the | **Ala**  
| - | **459** | peptide | .

- Splitting in “words” necessary: “214|C|>|T”



# Single Nucleotide Polymorphism—Input Features

- All features are **boolean**

- Morphological

Initial capital letter

All capital

Contains Digit

Punctuation

Interleukine

IL

CD28

?

- Automatically generated

Suffix3

Prefix4

Bag-of-Words

phosphatase

homeotic

homeotic



# Single Nucleotide Polymorphism—Input Features

- Dictionary based:
  - Stop words
  - Units
  - Common Words





# Single Nucleotide Polymorphism—Input Features

## Regular Expressions for

### ■ Location:

`chr|chromosome [1-9] |1[0-9] |2[0-2] |X|Y`

`nuclotide [0-9]+`

`amino acid [0-9]+`

`condon [0-9]+`

`[0-9]+ ala|arg|asn|asp|cys|gln|glu|his|ile|leu|lys|met|phe`

`...`



# Single Nucleotide Polymorphism—Input Features

## Regular Expressions for

### ■ State:

ala|arg|asn|asp|cys|gln|glu|his|ile|leu|lys|met|phe|pro|se  
A|T|C|G

...

### ■ Type:

deletion

insertion

missense

duplication

inversion

point mutation

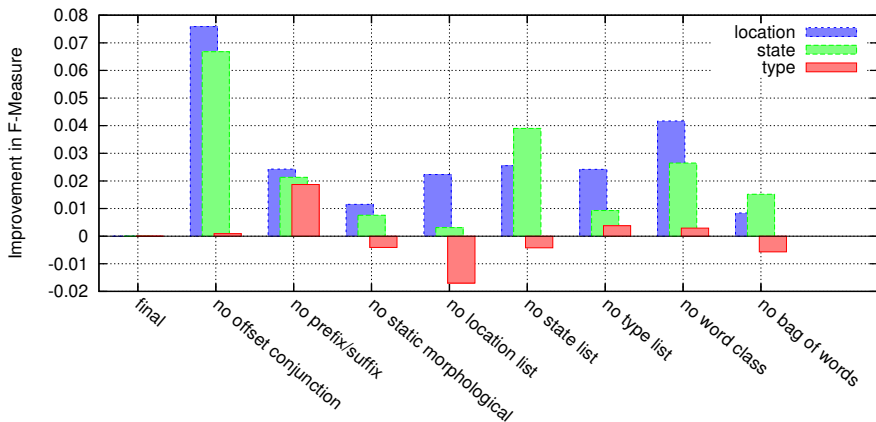
gain

loss

...



# SNP—Importance of the Features



# Single Nucleotide Polymorphism—Results

## ■ CRF Performance

Entity	precision (%)		recall (%)		$F_1$ -Score (%)	
Location	69.9	(0.0432)	67.2	(0.0417)	67.9	(0.0347)
Type	73.6	(0.0355)	51.2	(0.0395)	60.3	(0.0297)
State	78.0	(0.0275)	80.1	(0.0289)	79.2	(0.0205)

- Full medline: 16,848,632 articles, 8,975,073 abstracts
  - 312 CPUs with 2.6 GHz: **3.98** hours.



# SNP—RS numbers

## Finding RS numbers directly

### ■ First idea:

`[rR] [sS] [ ]*[0-9] [0-9]*`

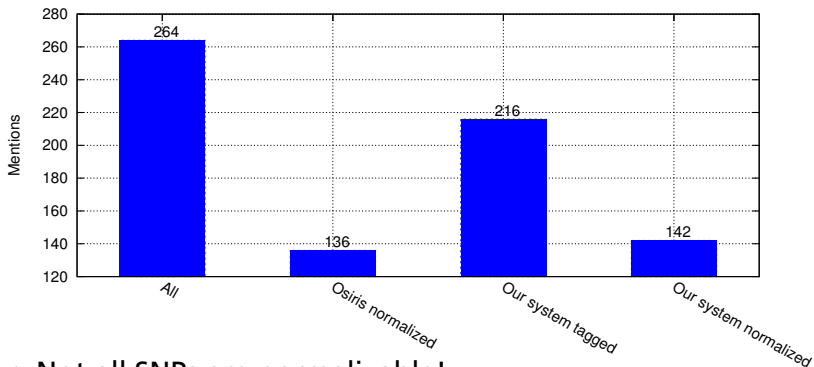
- Precision: 74%, Recall: 100%
- False Positives were:
  - cell-lines ("RS1")
  - computer names ("rs6000")
  - computer interfaces ("RS485")
  - indian rupees ("Rs1000")
  - chemical compounds ("RS61433")

- Use list of words that have to occur together with RS number: "SNP", "mutation" or "variation"...
- Use list of words that are not accepted as RS number (like "RS61433")
- Precision: 97%, Recall: 98%



# Single Nucleotide Polymorphism—Results

Analysis on independent test set of 100 abstracts



⇒ Not all SNPs are normalizable!

# Summary

---

- Conditional Random Fields are a state-of-the-art technique for labeling sequential data
- Advantages in comparison to dictionaries or rule based approaches
- Not possible to enumerate all SNPs in a dictionary  
⇒ works with CRFs
- Tagging full Medline possible
- Good results also with a small training corpus possible



Chemical occlusion of **vas** is quite effective in producing a block in the vas deferens of dogs.

In both cases, at the end of exposure the same level of **blood carboxyhemoglobin (COHb)** (about 50%) was reached.

**P** was induced by intraductal infusion of two different concentrations of glycodeoxycholic acid (GDOC 17 mmol and 34 mmol).

The UCR core sequence, CGCCATTTT, binds a ubiquitous nuclear factor and mediates negative regulation of **MuLV promoter** activity.

The predicted receptor structure includes a cysteine-rich extracellular domain, a single hydrophobic transmembrane domain, and a predicted **cytoplasmic serine/threonine kinase**.

*Saccharomyces cerevisiae* **UbcD1** encodes a highly conserved **ubiquitin**-conjugating enzyme involved in selective protein degradation.

The phenotypes of the **ICP0 nonsense mutants** were intermediate between those of the wild-type virus and 7134 in that the more **ICP0-coding sequence** expressed, the more the unphosphorylated form of **RNA polymerase II** is designated **IIA**, whereas the phosphorylated form is designated **IIO**.

Thank YOU for your attention!

These results indicate that an internal short element located at the very 5' terminal of L1 sequence and the nuclear factor binding to the element play a crucial role in the development of the virus.

The recombinant with a 5' end from **src** and a 3' end from **src** (recombinant **src**) transformed cells (CEF) to a spindle shape morphology, morphologically similar to that of **src**.

As was observed previously for **MATa cna1 cna2 double mutants**, **MATa cnb1 mutants** were defective in their ability to recover from **alpha-factor**-induced growth arrest.

**METHODS:** **IgG antibodies** vs **HHV-6 (anti-HHV-6-IgG)** were determined by indirect immunofluorescence in 100 IVDA (29 seronegative and 71 seropositive) and 100 healthy subjects of a similar age (control group).

The **hhl-1 mutation** caused a defect in synthesis of a 74-kD **heat shock protein**.

These findings suggest that the **MAP kinase activator/MAP kinase** system may be the downstream components of **ras** signal transduction pathways.

The subunit protein of **curli** was highly homologous at its amino terminus to **SEF-17**, the subunit protein of thin, aggregative fimbriae of *Salmonella enteritidis* 270b, a Gram-negative enterobacteria.

Questions?  
Remarks?

The distal portion of the **rat insulin I gene** 5'-flanking DNA contains **enhancer** elements, the Far and FLAT elements, that can function in combination, but not individually.

We have isolated and characterized a differentially-regulated **gene family** in the protozoan parasite *Leishmania major*.

Through Southern blot analyses of DNA from backcross and congenic mice, recombinant inbred strains, and somatic cell hybrids, the genetic loci that produce the **enhancer** elements were mapped on mouse chromosomes 5, 1, 17, 4, 14, 13, 7, X, and 8, respectively.

The **negatively supercoiled plasmid pUC19** did not compete, whereas an otherwise identical plasmid pUC19(CG), which contained a **(dG-dC)7 segment** in the Z-form of DNA.

These results lead us to hypothesize that a single **multisubunit TFIIID protein** supports transcriptional stimulation by diverse activation domains and from a TATA box.

From these results, **CBF-A** is a novel CArG box-, ssDNA- and RNA-binding protein, as well as a repressive transcriptional factor.

Surprisingly, the **Xenopus U7 gene** contains two adjacent octamer-binding motifs located only 12 and 24 bp upstream from the **PSE**, instead of the usual location.

Western blot analyses detect **anti-E**-specific immunoreactivity in affinity-purified extracts derived from the bacterial expression of a **truncated AMPD3 cDNA**.

Using either a **p50-** or **p65-**selected **kappa B motif**, which displayed differential binding with respect to the other protein, little to no binding was observed with the **anti-E**.

To define transcriptional control elements responsible for muscle-specific expression of the **human myoglobin gene**, we performed mutational analysis of upstream regulatory elements.

Linked to a **firefly luciferase gene**.

By using lambda gt11 expression cloning with oligonucleotides corresponding to the human immunodeficiency virus I TATA element, we report the identification of a **TATA element modulatory factor** (TMF).