UNIVERSITAS SCIENTIARUM SZEGEDIENSIS

Research Group on Artificial Intelligence

Approximate dictionary matching for biomedical information extraction

György Móra

gymora@inf.u-szeged.hu

Department of Informatics
University of Szeged

Biomedical Information extraction

Motivation

- Help life scientist to get information
- Automatic pathway search
- Searching for patterns in medical data

Tasks

- Named Entity Recognition (NER)
- Relation extraction
- Event extraction

Bio NER

NE types

- Species and organism names
- Gene and protein names
- Drug and chemical names
- Cell type
- Diseases

Methods

- Rule/dictionary-based (species)
- Machine learning (genes/proteins)
- Hybrid

Dictionary tagging

Dictionary: finite set of Named Entities

- Find the mentions of the entities in the text

Problems

- Linguistic variability
- Roman/Arabic numerals
- Greek letters
- Word order
- Abbreviation

Examples

Nuclear Factor Kappa Beta

- NF-κβ
- NF-kappa β
- NF-kappa B

Olfactory receptor, family 4, subfamily D, member 6

- Member 6 of subfamily D of olfactory receptor family 4
- OR4D6

Normalization

B → b or beta

β → beta

IV → iv or 4

receptors → receptor

were → be

Olfactory receptor, family 4, subfamily D, member 6

→ 4 6 d family member olfactory receptor

Normalized matching

Naive algorithm

- Generate all phrases of token length 1 to N
- Normalize phrases
- Search in the database

Problems

- Phrases can be as long as 25 words (intractable number of sub-phrases in a long sentence)
- Slow normalization method

Solution

Database

- Tokenized entity names
- Normalised form of each token
- Indexed with Lucen
- Index on “real token” length
- Punctuations, stopwords, etc. are not considered as real words

Matching method

1. Break up the text into sentences, words and sub-tokens
   - KappaB → kappa B
2. True/False search in the index for each sub-token
3. Determine for each token the longest possible sequence
4. Search for the possible entities (with length limit)

True/False search

0 - 0 1 -

However the precise function of

1 - 1 - 1 1 - 1

DNA - PK in DNA double - strand

0 1 - 1 0

break repair is not known .

Longest possible term

0 0 0 6 -
However the precise function of

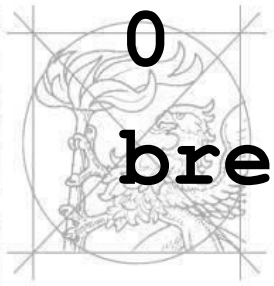
6 - 6 - 6 6 - 6
DNA - PK in DNA double - strand

0 2 - 2 0
break repair is not known .

Search for entities

0 0 0 A(2) -
However the precise function of
B(2) B(2)
C(3)- B(2)- C(3)C(3) - C(3)
DNA - PK in DNA double - strand

0 E(2) - F(2)0
break repair is not known .



Collect matches

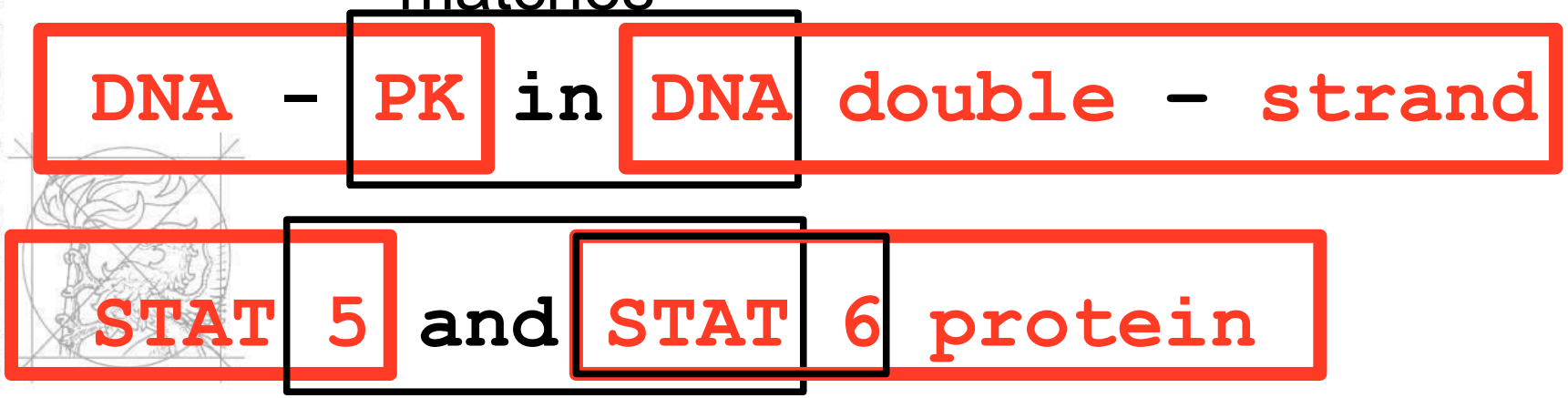
0 0 0 ~~A(2)~~ -
However the precise function of
B(2) B(2)
C(3)- B(2)- C(3)C(3) - C(3)
DNA - PK in DNA double - strand

0 ~~E(2)~~ - ~~F(2)~~0
break repair is not known .

Nested matches

"There can be only one!"

- Pick the longest/best matching by modified edit distance based on the word-form
- Remove all directly intersecting matches
- Repeat until there are no intersecting matches



alpha plasmids, these operons are transcribed from a bidirectional promoter region consisting of trfAp for the [26] trfA operon and trbAp and trbBp for the trb operon. Comparison of the encoded ERCC3Dm protein with the homologous proteins of mouse and man shows a strong amino acid conservation (71% identity), especially in the postulated [20] DNA binding region and seven 'helicase' domains. Genetic analysis places [20] CWH43 upstream of the [2] BCK2 branch of the [12] PKC1 signalling pathway, since [20] cwh43 mutations were synthetic lethal [12] with [12] pkc1 deletion, whereas the [20] cwh43 defects could be rescued by overexpression of [2] BCK2 and not by high-copy-number expression of genes encoding downstream proteins [12] of [12] the [12] PKC1 pathway However, unlike [2] BCK2, whose disruption in a [40] cln3 mutant resulted in growth arrest in G(1), no growth defect was observed in a double [20] cwh43 [40] cln3 mutants. Expression of the E2 protein resulted in rapid repression of HPV E6 and E7 expression, followed approximately [2] 12 h later by profound inhibition of cellular [20] DNA synthesis. Peak reactive hyperemia (mL.min-1.100 [10] mL-[2] 1) was determined in the calf and forearm immediately before and after 12 weeks of training. The hematopoietic form of [22] PTPN6 transcript is initiated at a downstream promoter separated by 7 kb from the first. Reversal of biliary sphincter spasm with low dose glucagon during operative cholangiography. In ten other endotoxin-albumin-treated pigs PGE1 infusion (0.25 micrograms X kg-1 X min-[2] 1) was begun after established pulmonary and cardiovascular dysfunction, for closer mimicking of clinical use. In particular, MF males receiving either [2] a [4] 5.0-mg/kg [6] CDP dose [2] or [2] a [18] 3.0-mg/kg RO dose explored the object more often than MM males.(ABSTRACT TRUNCATED AT 250 WORDS) A single HD session using cellulose triacetate or polysulfone membrane significantly increased water content both at forearm and lower leg ($p<0.05$). In the yeast two-hybrid assays, PNRC interacted with the orphan receptors SF1 and ERRalpha1 in a ligand-

Conclusions

"Fast" and flexible dictionary matching

~1 document / second

Handles reordering, Greek letters, Roman numerals

Arbitrary normalisation can be used

Not depends on a particular database format

Can be used for other tasks too

Thank you!

Acknowledgement

Nemzeti Fejlesztési Ügynökség
www.ujszechenyiterv.gov.hu
06 40 638 638

MAGYARORSZÁG MEGÚJUL

A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

ÚJ SZÉCHENYI TERV

The publication/presentation is supported by the European Union and co-funded by the European Social Fund.

Project title: "Broadening the knowledge base and supporting the long term professional sustainability of the Research University Centre of Excellence at the University of Szeged by ensuring the rising generation of excellent scientists."

Project number: TÁMOP-4.2.2/B-10/1-2010-0012