



Fraunhofer Institute
Algorithms and
Scientific Computing
Bioinformatics



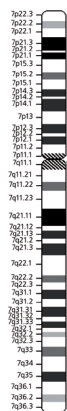
Recognition and
Normalization of
Named Entities in
Scientific Text

ProMiner™ is a tool for biological name recognition which has been developed at the Fraunhofer Institute for Algorithms and Scientific Computing SCAI in a collaboration project with Aventis Pharma AG. ProMiner™ addresses several fundamental issues in name entity recognition in the field of life sciences:

- recognition of biomedical entities and their spelling variants in text
- mapping of synonyms to reference names and data sources
- context-dependent disambiguation of biomedical termini and resolution of acronyms



UMLS
CUI: C0007766
Concept: Intracranial Aneurysm
Semantic Type: Disease or Syndrome



Genomwide-linkage and haplotype-association studies map intracranial aneurysm to chromosome 7q11.19.

Rupture of intracranial aneurysms (IAs) causes subarachnoid hemorrhage, a devastating condition with high morbidity and mortality. Angiographic and autopsy studies show that IA is a common disorder, with a prevalence of 3%-6%. Although IA has a substantial genetic component, little attention has been given to the genetic determinants. We report here a genomwide linkage study of IA in 104 Japanese affected sib pairs in which positive evidence of linkage on chromosomes 5q22-31 (maximum LOD score [MLS] 2.24), 7q11 (MLS 3.22), and 14q22 (MLS 2.31) were found. The best evidence of linkage is detected at D7S2472, in the vicinity of the elastin gene (ELN), a candidate gene for IA. Fourteen distinct single-nucleotide polymorphisms (SNPs) were identified in ELN, and no obvious allelic association between IA and each SNP was observed. The haplotype between the intron-20/intron polymorphism of ELN is strongly associated with IA ($P=3.81 \times 10^{-6}$), and homozygous patients are at high risk ($P=.002$), with an odds ratio of 4.39. These findings suggest that a genetic locus for IA lies within or close to the ELN locus on chromosome 7.

Entrez Gene

GeneID: 2006
Official Symbol: ELN
Name: elastin



Accession number: P15502
Protein Name: elastin

Contact

Dr. Juliane Fluck
Fraunhofer Institute SCAI
Schloss Birlinghoven
D-53754 Sankt Augustin
Germany

phone: +49 (0) 2241-14-2188
fax: +49 (0) 2241-14-2656
juliane.fluck@scai.fraunhofer.de

www.scai.fraunhofer.de/bio.html

Identification of biological entities in ProMiner™ is based on a dictionary approach. These regularly updated dictionaries are generated and curated automatically followed by a manual evaluation process. Furthermore, a classification according to their ambiguous usage in scientific texts assures high specificity. ProMiner™ can work with voluminous dictionaries,

complex thesauri and large controlled vocabularies derived from ontologies. The proprietary ProMiner™ dictionary developed by SCAI for Homo sapiens has more than 32 000 entries and contains about 400 000 synonyms. This dictionary covers the vast majority of all human gene and protein names and thus allows for efficient identification of human gene and protein names in unstructured text.

Proof of Performance

The performance of ProMiner™ recognition of gene and protein names was tested in the »critical assessment of text mining in biology« (BioCreAtIvE I and II). ProMiner™, was benchmarked against other industrial and academic name entity recognition tools and scored in BioCreAtIvE I highest in text dealing with the two multicellular organisms (cf. Table 1).

Available dictionaries

- Gene and protein name dictionaries for various organisms
 - Human
 - Mouse
 - Rat
 - On request: Yeast, Fly, Arabidopsis, Bacterial organisms
- Gene ontology dictionary
- Mesh term dictionary
- Organism name dictionary
- Disease name dictionary
- Drug name dictionary

Technical Performance

ProMiner™ is available for UNIX™/Linux and the Microsoft Windows™ operating system. To reduce the time required to process huge text corpora, ProMiner can run distributed in a grid enabled hardware. The software is already integrated in the IBM-UIMA framework and can be combined with other text processing software. ProMiner™ has been successfully integrated in the Temis BER Skill Cartridge, and can easily be adapted as pre-processing for other semantic analysis environments by tagging entity references in texts. ProMiner searches named entities in a single pass through the text corpus, and takes about 1 - 2 days (depending on the dictionary size) on a single processor machine and ½ - 3h in a grid enabled hardware environment (30 workstations) to scan the full Medline™ database (~ 16 000 000 entries of 24 GByte data).

Performance

BioCreAtIvE evaluation

	Mouse		Fly		Yeast		Human	
BioCreAtIvE	I (2004)		I (2004)		I (2004)		II (2006)	
best automatic system		ProMiner™ system	best automatic system	ProMiner™ system	best automatic system	ProMiner™ system	best automatic system	ProMiner™ system
F-Measure	0,79	0,79	0,82	0,82	0,92	0,9	0,81	0,8

Table 1: Result of the BioCreAtIvE assessment. ProMiner showed the best F-measures and set the benchmark for the best automated system in two out of three application scenarios.

¹ Hanisch, D.; Fluck, J.; Mevissen, H.-T.; Zimmer, R.: *Playing biology's name game: identifying protein names in scientific text*. Pac Symp Biocomput. 2003. 403-414.

² Hanisch, D.; Fundel, K.; Mevissen, H.-T.; Zimmer, R.; Fluck, J.: *ProMiner: rule-based protein and gene entity recognition*. BMC Bioinformatics 2005, 6 (Suppl 1), p 14v

³ Fluck, J.; Mevissen, H.-T.; Dach, H.; Oster, M.; Hofmann-Apitius, M.: *ProMiner: Recognition of Human Gene and Protein Names using regularly updated Dictionaries*. 2nd BioCreAtIvE Challenge Workshop 2006, Critical Assessment of Information Extraction in Molecular Biology, Madrid Spain.