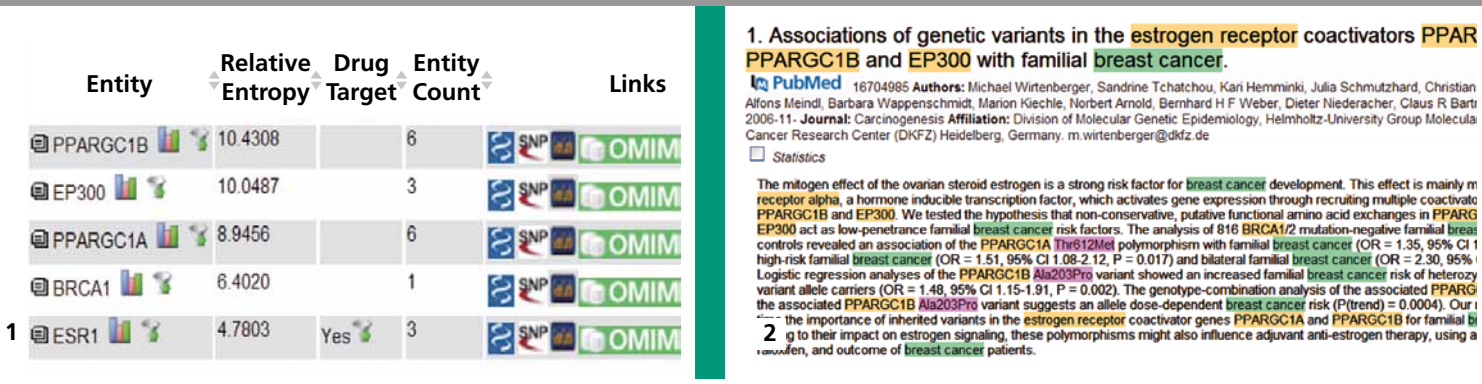
Fraunhofer
SCAI

FRAUNHOFER INSTITUTE FOR ALGORITHMS AND SCIENTIFIC COMPUTING SCAI



Entity	Relative Entropy	Drug Target	Entity Count	Links
PPARGC1B	10.4308		6	SNP OMIM
EP300	10.0487		3	SNP OMIM
PPARGC1A	8.9456		6	SNP OMIM
BRCA1	6.4020		1	SNP OMIM
ESR1	4.7803	Yes	3	SNP OMIM

1

1. Associations of genetic variants in the estrogen receptor coactivators PPAR PPARGC1B and EP300 with familial breast cancer.

PubMed 16704985 **Authors:** Michael Wirtenberger, Sandrine Tchatchou, Kari Hemminki, Julia Schmutzhard, Christian Alfons Meindl, Barbara Wappenschmidt, Marion Kiechle, Norbert Arnold, Bernhard H F Weber, Dieter Niederacher, Claus R Bartr 2006-11- **Journal:** Carcinogenesis **Affiliation:** Division of Molecular Genetic Epidemiology, Helmholtz-University Group Molecula Cancer Research Center (DKFZ) Heidelberg, Germany. m.wirtenberger@dkfz.de

Statistics

The mitogen effect of the ovarian steroid estrogen is a strong risk factor for breast cancer development. This effect is mainly m receptor alpha, a hormone inducible transcription factor, which activates gene expression through recruiting multiple coactivato PPARGC1B and EP300. We tested the hypothesis that non-conservative, putative functional amino acid exchanges in PPARG EP300 act as low-penetrance familial breast cancer risk factors. The analysis of 816 BRCA1/2 mutation-negative familial breas controls revealed an association of the PPARGC1A Thr612Met polymorphism with familial breast cancer (OR = 1.35, 95% CI 1 high-risk familial breast cancer (OR = 1.51, 95% CI 1.08-2.12, P = 0.017) and bilateral familial breast cancer (OR = 2.30, 95% Logistic regression analyses of the PPARGC1B Ala203Pro variant showed an increased familial breast cancer risk of heterozy variant allele carriers (OR = 1.48, 95% CI 1.15-1.91, P = 0.002). The genotype-combination analysis of the associated PPARG the associated PPARGC1B Ala203Pro variant suggests an allele dose-dependent breast cancer risk (P(trend) = 0.0004). Our the importance of inherited variants in the estrogen receptor coactivator genes PPARGC1A and PPARGC1B for familial bi g to their impact on estrogen signaling, these polymorphisms might also influence adjuvant anti-estrogen therapy, using a fen, and outcome of breast cancer patients.

2

1 *Found entities could be indexed, ranked and linked to other data.*

2 *Highlighting in text corresponding to the entity classes.*

ProMiner®: RECOGNITION AND NORMALIZATION OF NAMED ENTITIES IN SCIENTIFIC TEXT

Fraunhofer Institute for Algorithms and Scientific Computing SCAI

Schloss Birlinghoven
53754 Sankt Augustin
Germany

Contact
Dr. Juliane Fluck
phone +49 2241 14-2188
juliane.fluck@scai.fraunhofer.de
www.scai.fraunhofer.de/prominer

Distribution
scapos AG
phone +49 2241 14-2820
www.scapos.com

ProMiner
Recognition and Normalization of Named Entities in Scientific Text

Challenges

Scientific publications found in abstract databases, full text journals or patents are the main and most up-to-date information source, but the amount of text is overwhelming for most Life Science areas.

Recognition of Life Science terminology is a key prerequisite for performing automatic information retrieval and information extraction. Huge and complex terminologies with high numbers of synonymous expressions, ambiguous terminology and numerous generations of new names and classes present named entity recognition with a real challenge.

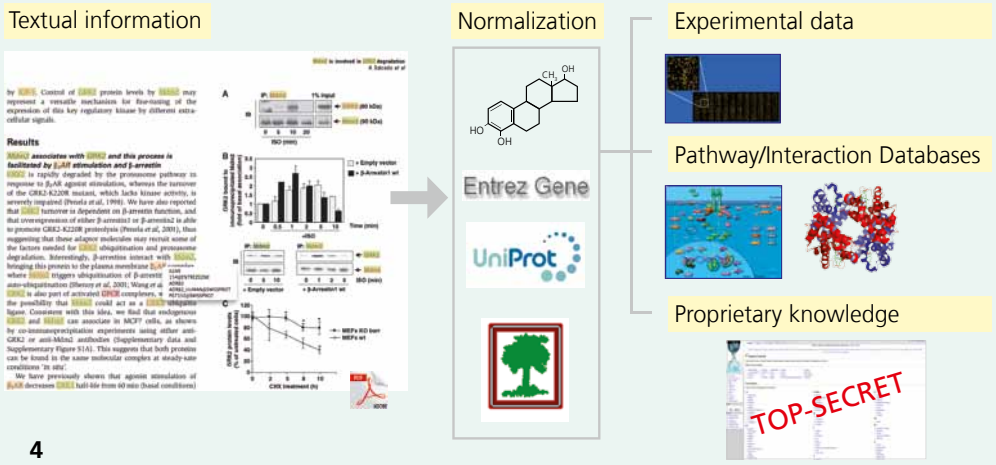
ProMiner is a tool for specific terminology recognition and addresses several fundamental issues in name entity recognition in the field of Life Sciences:

- ProMiner can handle voluminous dictionaries, complex thesauri and large controlled vocabularies derived from ontologies
- Regularly updated dictionaries through automatic curation followed by a manual evaluation process
- Mapping of synonyms to reference names and data sources
- Context dependent disambiguation of biomedical termini and resolution of acronyms
- Specific handling of common English word synonyms
- Spelling variants of expressions in the source dictionary can be recognized
- High speed tagging and parallel workflow for multiple dictionaries
- Incorporation of regular expressions (e.g. for the recognition of SNP rs numbers)
- Full text annotation in XML, HTML or PDF format
- Patent annotation

BioCreAtIvE evaluation

Organism (evaluation)	ProMiner®	Best performance
Mouse (BioCreAtIvE04)	0,79	0,79
Fly (BioCreAtIvE04)	0,82	0,82
Yeast (BioCreAtIvE04)	0,90	0,92
Human (BioCreAtIvE07)	0,80	0,81

3



4

Proof of Performance

The performance of ProMiner recognition of gene and protein names was tested in the international "critical assessment of text mining in biology" (BioCreAtIvE I and BioCreAtIvE II). ProMiner was benchmarked against other industrial and academic named entity recognition tools (cf. 3). Updated and new generated dictionaries are continually evaluated in industrial applications.

Fields of Application

Indexing machinery for fast indexing of huge document resources

Customer feedback on ProMiner:

- "We are amazed about its speed and ability to work with large input files".
- "…impressive to get the combination of information in text with enriched background and experimental data".

Module for named entity recognition in a larger workflow for information extraction

- Java module with defined input and output streams
- An annotator service for named entities in the Unstructured Information Management Architecture (UIMA) framework
- Already integrated in the TEMIS - BER information extraction environment software

Content generation for the interpretation of large scale experimental data

- Simple output file to fill/supplement database content
- Linkage to other data is easily possible through the provided mapping to databases or controlled vocabulary (cf. 4)

Available dictionaries

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

Dictionary independent recognition

While parts of the Life Science terminology could be found with the help of dictionaries in some entity classes, it is not possible to enumerate all names. Examples are IUPAC names or SNP rs numbers. Here, we offer other techniques integrated as plugin into ProMiner:

- Machine learning based
  - IUPAC recognition
  - SNP recognition
- Regular expression based
  - rs number recognition
  - chromosomal location

Technical Specification and Parameter settings

- ProMiner is available for UNIX/Linux and Microsoft Windows
- A scheduler for cluster runs could be provided

Text format

- ProMiner supports ASCII text, MEDLINE format, XML, HTML and PDF full text
- Output format as: Meta-information, XML tagged text and HTML output

Annotation of PDF

The increasing number of electronically available full text publications offers the ability to process these documents and annotate the knowledge stored in them. Integrated in the ProMiner software, we offer a special PDF plugin for the annotation in PDF documents. Here the annotations are directly written into the PDF output format.

For semantic search and visualization, we offer the semantic search engine SCAIView. Further information about SCAIView: www.scai.fraunhofer.de/scaiview.

3 *Results in the international "critical assessment of text mining in biology" (BioCreAtIvE I and II).*

4 *Relation to experimental data, interaction data bases or proprietary data through the provided mapping.*