

Publikációgyűjtemény tudásbázisának építése természetes nyelven



HORNYÁK ZSUZSANNA, MÉSZÁROS TAMÁS
BME MIT

Miért szükséges az intelligens tudásbázis?



„breast cancer gene”

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

[RSS](#) [Save search](#)

[Show additional filters](#)

Display Settings: Summary, 20 per page, So

Results: 1 to 20 of 48135 << First <

[Ethanollic Neem \(Azadirachta indica\) Leaf and HeLa Cells and Potentiates the The](#)

Article types
Clinical Trial
Review
More ...

„breast cancer ccna2”

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

[RSS](#) [Save search](#)

[Show additional filters](#)

Display Settings: Summary, 20 per page, So

See 257 articles about **CCNA2** gene function
See also: [CCNA2 cyclin A2](#) in the Gene datab
[ccna2](#) in [Homo sapiens](#) | [Mus musculus](#) | [Ratt](#)

Results: 1 to 20 of 50 << First

[CCNA2 Is a Prognostic Biomarker for ER Resistance.](#)

1. [Gao T, Han Y, Yu L, An S, Li Z, Li J](#)

Article types
Clinical Trial
Review
More ...

Text availability
Abstract available
Free full text available

- Mire kíváncsi valójában?
 - Kulcsszavak tárolása nem elég segítség

Intelligens tudásbázis



- A cikk tartalma
- Gépek által értelmezhető módon
- Ki készíti el?
 - Technológiai és tárgyterületi ismeret kell
 - Nyelvi elemzés nehéz
 - Adatbányászat: felülvizsgálás szükséges

```
<rdf:RDF xmlns:rdf="
http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="
http://www.rdfabout.com/">
    <dc:title>Germline mutation in the RAD51B
gene confers predisposition to breast
cancer</dc:title>
    ...
  </rdf:Description>
  <sp:Statements>
    <sp:Statement>
      <sp:GeneRole>
        <sp:Gene>RAD51C</sp:Gene>
        <sp:Disease>Breast cancer
</sp:Disease>
        ...
      </sp:GeneRole>
    </sp:Statement>
    ...
  </sp:Statements>
</rdf:RDF>
```

Kontrollált természetes nyelvű absztraktok



Megoldás tudásbázis építésére

- **Természetes nyelvű állítások**
 - Cikk tartalmi összefoglalása
- **Kontrollált nyelv**
 - Kötött nyelvi struktúrák
 - Javaslatokkal támogatott szerkesztői felület
- **Absztraktokból egyértelmű szemantikus reprezentáció generálása**

Példa imlementáció



**AZ ELKÉSZÍTETT ZOTERO PROTOTÍPUS
BEMUTATÁSA**

Zotero



Firefox

breast cancer ccna2 mutation - Pub... x Regulation of progesterone receptor ... x

www.ncbi.nlm.nih.gov/pubmed/21550420

NCBI Resources How To Sign in to NCBI

PubMed US National Library of Medicine National Institutes of Health

Advanced Search Help

Display Settings: Abstract Send to: ELSEVIER FULL-TEXT ARTICLE Free in PMC free full-text archive

Int J Biochem Cell Biol. 2011 Aug;43(8):1157-67. doi: 10.1016/j.biocel.2011.04.009. Epub 2011 Apr 21.

Regulation of progesterone receptor activity by cyclin dependent kinases 1 and 2 occurs in part by phosphorylation of the SRC-1 carboxyl-terminus.

Moore NL¹, Weigel NL.

Author information

Everything

Title	Creator
Germline mutation in the RAD...	Golmard et al.
Guidelines for establishing loc...	Vihinen et al.
How to catch all those mutatio...	Kohonen-Cori...
Improving sequence variant de...	Wildeman et al.
Kiegészítőkezelő	
Locus-specific database domai...	Mitropoulou e...
LOVD BRCA1	
LOVD v.2.0: the next generatio...	Fokkema et al.
Mutation nomenclature extens...	den Dunnen a...
Recommendations for locus-s...	Cotton et al.
Sharing data between LSDBs a...	den Dunnen et...

Info Notes Tags Related

Item Type: Journal Article

Title: Germline mutation in the RAD51B gene confers predisposition to breast cancer

- Author: Golmard, Lisa
- Author: Caux-Moncoutier, Virginie
- Author: Davy, Grégoire
- Author: Al Ageeli, Essam
- Author: Poirot, Brigitte
- Author: Tirapo, Carole
- Author: Michaux, Dorothée
- Author: Barbaroux, Catherine

zotero

Zotero



- **Publikációk gyűjtése**
- **Automatikus metaadat kinyerés**
- **Többféle formátum**
- **Kollaboráció**
- **Open source**
- **Javascript/Server API**
 - Adatok lekérése és bővítése
 - Új keresőfunkció

Nyelvtan készítése



- Mik a fontos tartalmi állítások?
 - Mi alapján keresnek?
 - Statisztika
- Tárgyterületi ontológia
 - Taxonómia
 - Kapcsolódó információk
- Formális reprezentáció tárolása
 - XML, RDF,..?
- sentence => [patient], [had], genechange, [of], \$gene.
- \$change => [overexpression].
- \$change => [mutation].
- genechange => \$change.
- genechange => \$change, [and], \$change.
- „Patient had mutation of BRCA1”

Dinamikus nyelvtan



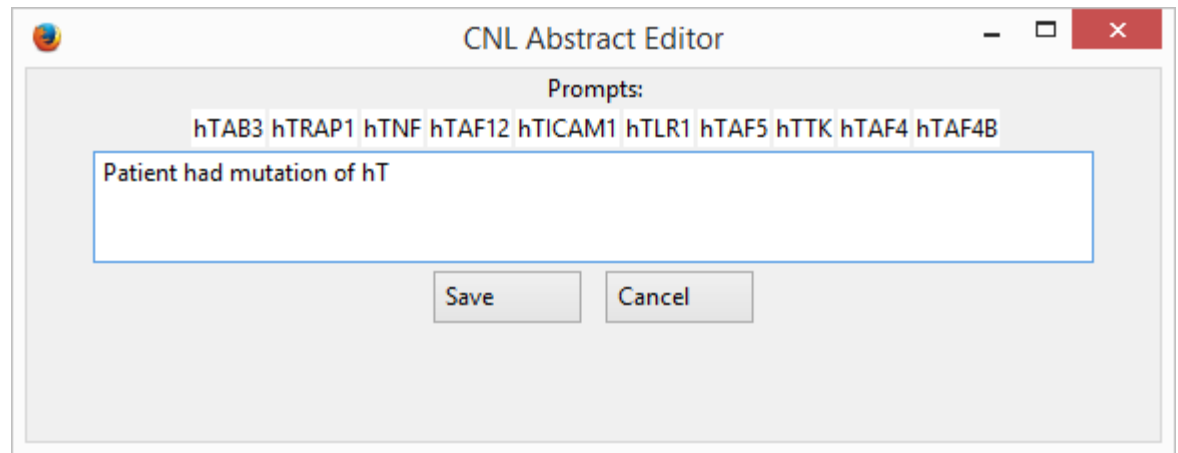
- **Ontológia betöltése**
 - SPARQL lekérdezés, alkalmazásfüggő
- **Szó szintű strukturális szabályok**
 - Reguláris kifejezés
 - Elemzési időben



Otthonossági probléma

Betűnkénti frissítés

AJAX kommunikáció



Kontrollált nyelvű szerkesztői felület

Nyelvi elemzés



- **Kontrollált nyelvi szabályok**
 - Egyértelmű
- **Inkrementális nyelvi elemző**
 - Minden szó után eldönthető, van-e részleges illeszkedés
 - Folytatások generálása
- **Kész absztrakt mondatonként elemzése**
 - Elemzési fa
 - Formális reprezentáció előállítása

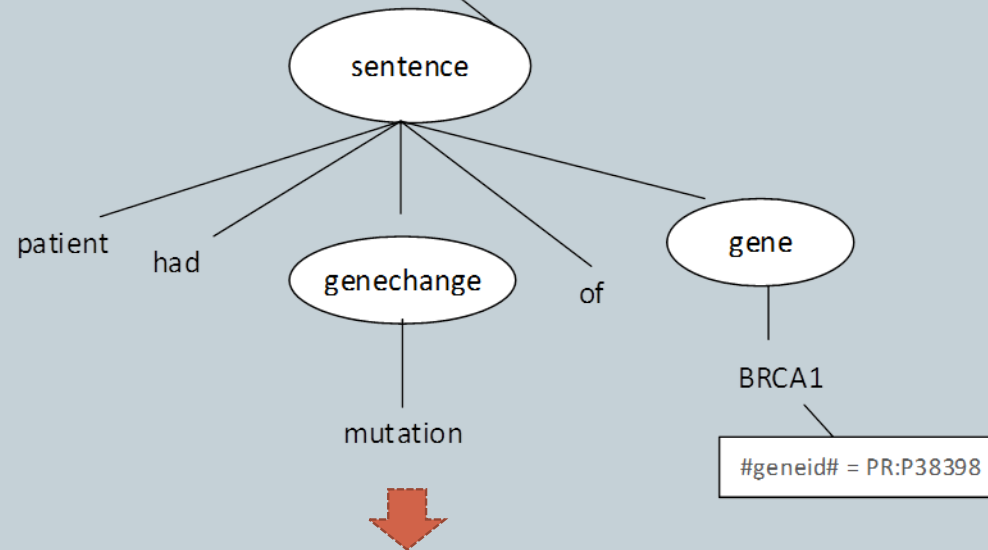
Elemzési fa



Patient had mutation of BRCA1



patient(#pubid#,#genechange#,#geneid#)



patient(245689,"mutation",PR:P38398).

Intelligens keresés



- Mely cikkekben foglalkoztak BRCA1 mutációval rendelkező betegekkel?
- ? - `patient(ID,mutation, PR:P38398)`

Összefoglalás



- Egyszerű kiegészítése digitális könyvtárnak
- Természetes nyelvű szemantikus absztraktok
- Egyértelmű formális reprezentációra fordítás
- Intelligens tudásbázis
- Komplex, tartalomcentrikus keresés