

Julien Gobeill¹, Emilie Pasche², Douglas Teodoro², Anne-Lise Veuthey³, Patrick Ruch¹

¹ University of Applied Sciences, Information Sciences, Geneva

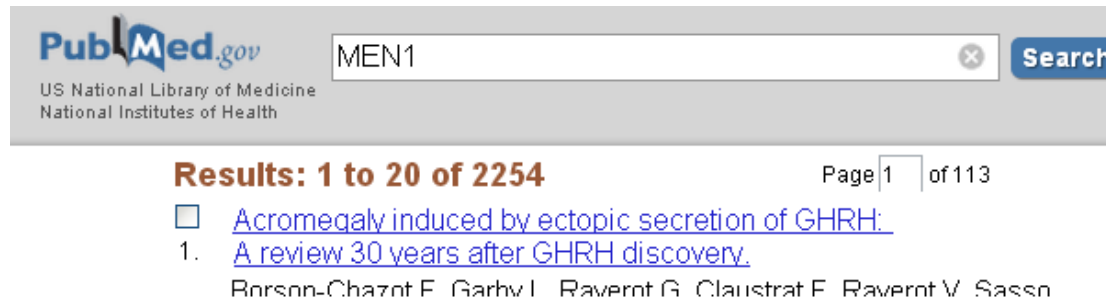
² Hospitals and University of Geneva, Geneva

³ Swiss-Prot group, Swiss Institute of Bioinformatics, Geneva

Answering Gene Ontology terms to proteomics questions by supervised macro reading in MEDLINE

Data deluge...

“What is the subcellular location of protein MEN1 ?”



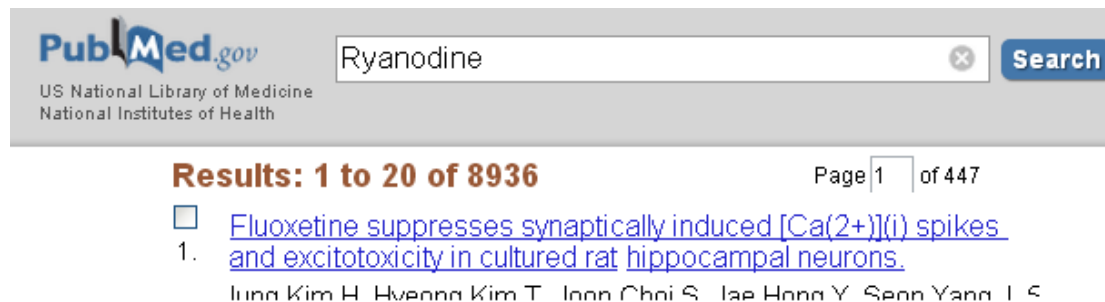
PubMed.gov
US National Library of Medicine
National Institutes of Health

MEN1

Results: 1 to 20 of 2254 Page 1 of 113

[Acromegaly induced by ectopic secretion of GHRH: A review 30 years after GHRH discovery.](#)
1. Borson-Chazot F, Garby L, Raverot G, Claustrat E, Raverot V, Sasson

“What molecular functions are affected by Ryanodine ?”



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

Ryanodine

Results: 1 to 20 of 8936 Page 1 of 447

[Fluoxetine suppresses synaptically induced \[Ca\(2+\)\]\(i\) spikes and excitotoxicity in cultured rat hippocampal neurons.](#)
1. Jung Kim H, Hyeong Kim T, Joon Choi S, Jae Hong Y, Seon Yang J, S

Ontology-based search engines



what

Top Terms

- ryanodine-sensitive calcium-release channel activity [7,554]
- sarcoplasmic reticulum membrane [3,373]
- sarcoplasmic reticulum lumen [3,385]
- sarcoplasm [3,503]
- sarcoplasmic reticulum [3,401]

Knowledge Base

- biological_process [7,215]
- cellular_component [7,473]
- molecular_function [8,181]
 - All of molecular_function [8,181]**
 - transporter activity [7,814]
 - catalytic activity [2,866]
 - transcription regulator activity [18]
 - transcription activator activity [10]
 - transcription cofactor activity [9]
 - transcription coactivator activity [8]
 - cAMP response element binding protein binding [7]
 - nutrient reservoir activity [14]
 - antioxidant activity [11]
 - motor activity [9]

Ryanodine

find

8,936 documents semantically analyzed

Fluoxetine suppresses synaptically induced [Ca(2+)]

Authors: Jung Kim, Hee, et.al.

Journal: Brain research, 2012

In addition, fluoxetine decreased the [Ca(2+)](i) responses induced by the metab

[PubMed 23131584](#) [Related Articles](#) [Read Full Text](#)

Calcium leak through ryanodine receptor is involved

Authors: Suzuki, Mari, et.al.

Journal: Biochemical and biophysical research communications, 2012

In this study, we examined involvement of ryanodine receptor (RyR), an endopl

[PubMed 23131566](#) [Related Articles](#) [Read Full Text](#)

Conduction Slowing Contributes to Spontaneous Ven

Authors: Zhang, Yanmin, et.al.

Journal: Journal of cardiovascular electrophysiology, 2012

It is associated with mutations involving the cardiac ryanodine receptor (RyR2).

[PubMed 23131176](#) [Related Articles](#) [Read Full Text](#)

Question Answering (EAGLi system)



eagl.unige.ch/EAGLi/jsp/result.jsp



What is the subcellular location of protein MEN1 ?

EAGLi PubMed

Search

Your question was : *MEN1*, reformulated as *MEN1*

- Possible answers are :
- nucleus (4 matches in 1 documents) ; cytosol (3/1)
 - membrane (2 matches in 1 documents) ; chromosome (2/2)
 - extracellular region (1 match in 1 documents) ; endoplasmic reticulum (1/1) ; T-tubule (1/1) ; limeric region (1/1) ; hydrogen:potassium-exchanging ATPase complex (1/1) ... [show all](#)

nucleus ▲



Expression and subcellular localization of menin in human cancer cells.

Ren F , Xu HW , Hu Y , Yan SH , Wang F , Su BW , Zhao Q
Exp Ther Med. 2012 Jun; 3(6): 1087-1091
Pmid : 22970022

... Western blotting was used to determine the quantity of menin in the **nucleus**, cytosol and membrane of the ...

cytosol ▲



Expression and subcellular localization of menin in human cancer cells.

Ren F , Xu HW , Hu Y , Yan SH , Wang F , Su BW , Zhao Q
Exp Ther Med. 2012 Jun; 3(6): 1087-1091
Pmid : 22970022

... Western blotting was used to determine the quantity of menin in the nucleus, **cytosol** and membrane of the ...

membrane ▲

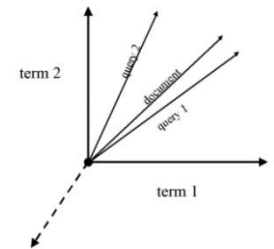
Redundancy hypothesis: The number of associated/co-occurring answers dominate other dimensions

Best way for extracting GO terms from a set of abstracts ? (1/3)

- Comparison based in two categorizers :

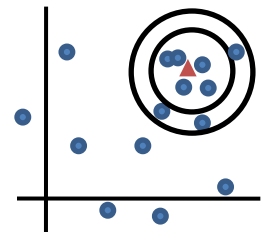
- Thesaurus-Based (EAGL)

- Competitive with MetaMap (Trieschnigg et al., 2009)
 - Compute lex. similarity between text and GO terms



- Machine Learning (GOCat)

- k -NN
 - Similarity between input text and already curated abstracts
 - KB derived from GOA : ~90'000 instances



Best way for extracting GO terms from a set of abstracts ? (2/3)

- Two tasks :
 - Classical categorization (micro reading ~ biocuration)



- Redundancy-based QA (macro reading)



Best way for extracting GO terms from a set of abstracts ? (3/3)

- One benchmark for micro reading evaluation
 - 1'000 abstracts and GO descriptors from GOA
- Two benchmarks for macro reading evaluation
 - 50 questions derived from a set of biological databases:
 - What molecular functions are affected by [chemical] ?
 - What cellular component is the location of [protein] ?



Results

	micro reading task		macro reading task			
Benchmark	1'000 abstracts		CTD		UniProt	
Metrics	P0	R10	P0	R100	P0	R10
EAGL (Thesaurus Based)	.23	.16	.34	.15	.33	.45
GOCat (k-NN)	.43 (+86%)	.47 (+193%)	.69 (+102%)	.33 (+120%)	.58 (+75%)	.73 (+62%)

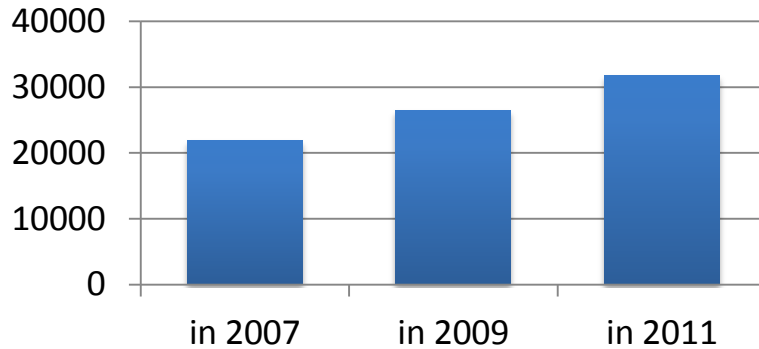
+ 75/120% for k-NN (sup. learning)

➔ Redundancy hypothesis insufficient

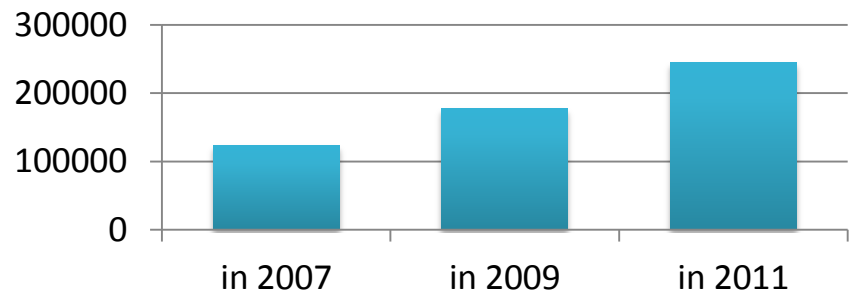
Why/Where is the power ? Size does or does not matter ?

Deluge is self-compensated 😊

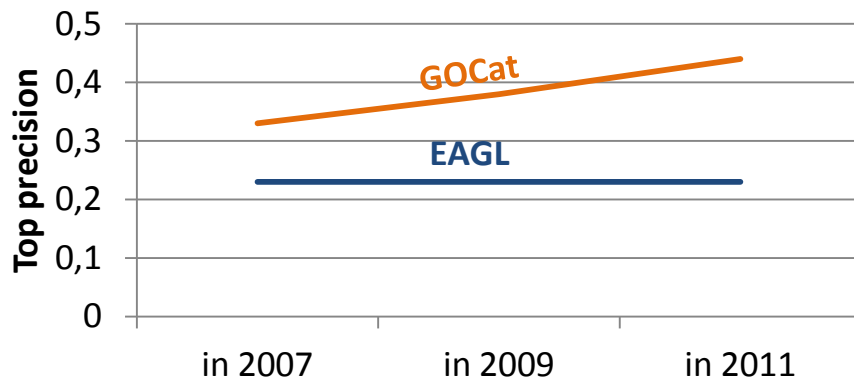
terms in GO: +150% / 2003



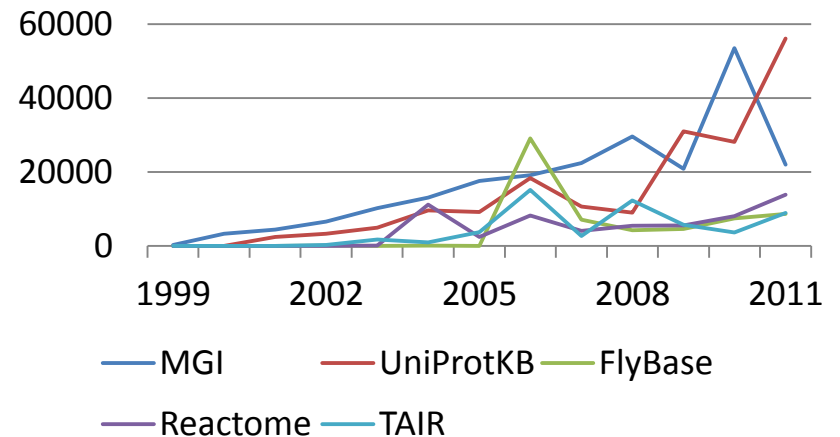
annotations with a PMID in GOA: + 100% / 2007



Performances of both categorizers across the time

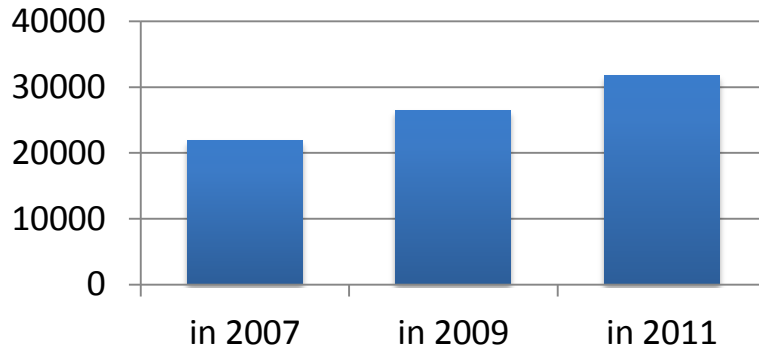


Annotations in GOA for the top 5 most contributing source

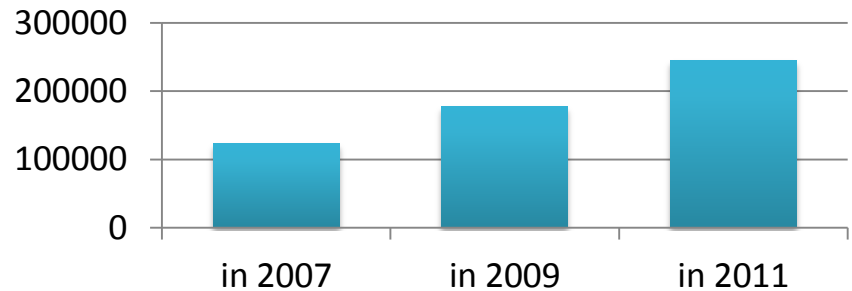


Deluge is self-compensated 😊

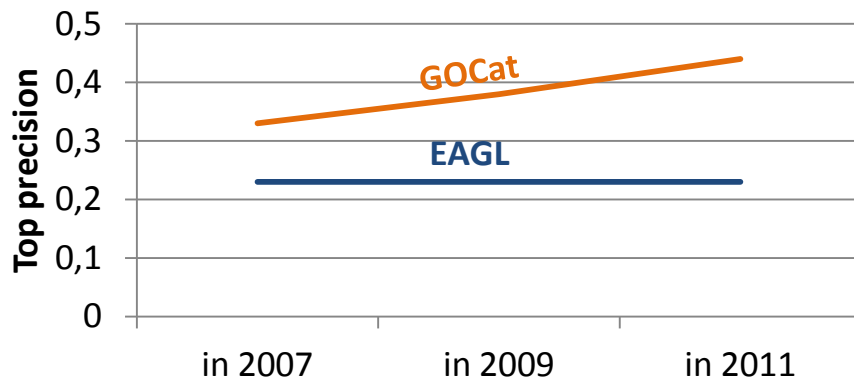
terms in GO: +150% / 2003



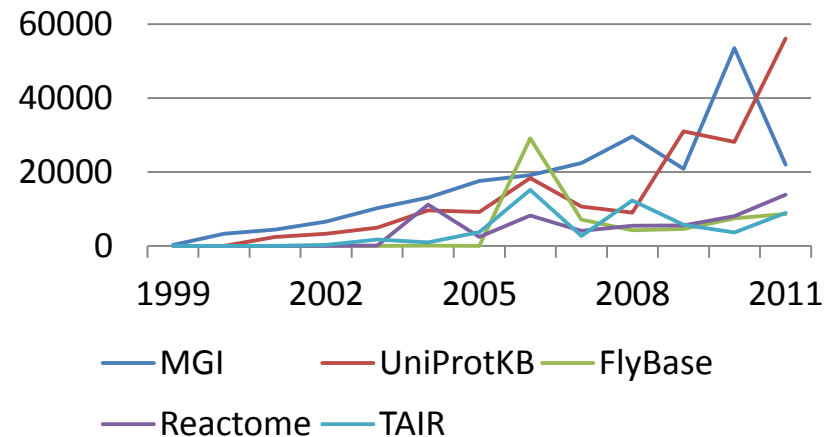
annotations with a PMID in GOA: + 100% / 2007



Categorization effectiveness moves faster than data



Annotations in GOA for the top 5 most contributing source



Magic !

The automatic categorization based on a PMID_{2007} performed in 2011 is of higher quality than a categorization on the same PMID_{2007} performed in 2007

No concept drift at all and even some improvement!

Example in toxicogenomics: CTD vs. GOCat

“What molecular functions are affected by Ryanodine ?”



Giving insight into how chemicals affect our health.

Ryanodine

GOCat

<u>GO Level</u>	<u>GO Term</u>	
9	GO0005219 : ryanodine-sensitive calcium-release channel activity	✓
7	GO0015279 : calcium-release channel activity	✓
7	GO0005262 : calcium channel activity	✓
6	GO0022834 : ligand-gated channel activity	
6	GO0015276 : ligand-gated ion channel activity	
3	GO0005516 : calmodulin binding	✓

<u>Rank</u>	<u>GO Term</u>	
1.	GO0005515 : protein binding	
✓ 2.	GO0005219 : ryanodine-sensitive calcium-release channel activity	
3.	GO0005245 : voltage-gated calcium channel activity	
4.	GO0005509 : calcium ion binding	
✓ 5.	GO 0005262 : calcium channel activity	
6.	GO0005102 : receptor binding	
✓ 7.	GO0005516 : calmodulin binding	
8.	GO0005388 calcium-transporting ATPase activity	
✓ 9.	GO0015279 : calcium-release channel activity	
10.	GO0005528 : FK506 binding	

Example in UniProt

“What is the subcellular location of protein MEN1 ?”

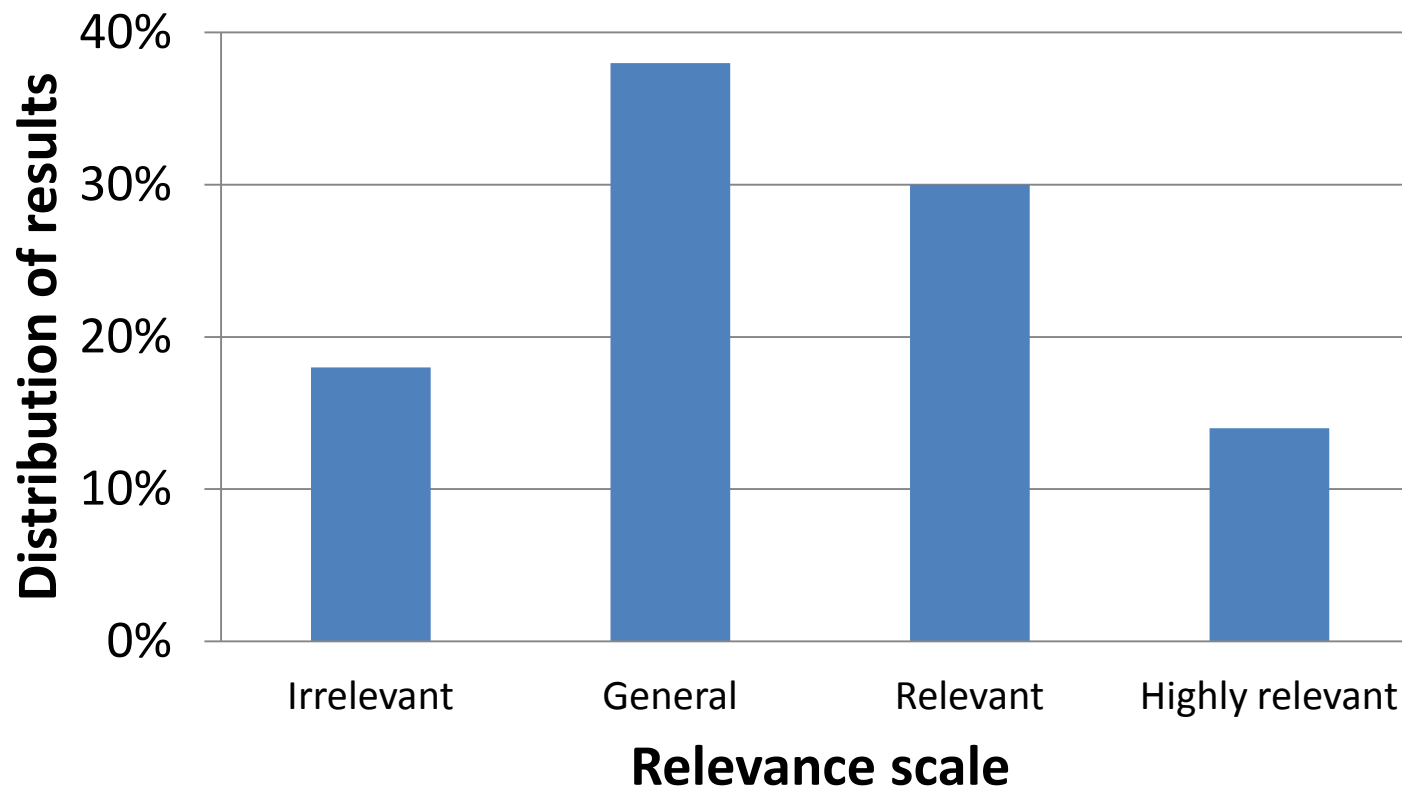
UniProt > UniProtKB	
Protein names	Recommended name: Menin
Gene names	Name: MEN1

GOCat

<u>GO Level</u>	<u>GO Term</u>	
6	GO0035097 : histone methyltransferase complex	✓
5	GO0000785 : chromatin	✓
5	GO0016363 : nuclear matrix	✓
4	GO0005829 : cytosol	✓
3	GO0032154 : cleavage furrow	

<u>Rank</u>	<u>GO Term</u>
1.	GO0005634 : nucleus
2.	GO0005737 : cytoplasm
3.	GO0005886 : plasma membrane
4.	GO0005615 : extracellular space
5.	GO0005887 : integral to plasma membrane
6.	GO0005739 : mitochondrion
✓ 7.	GO0005829 : cytosol
8.	GO0005576 : extracellular region
✓ 9.	GO0035097 : histone methyltransferase complex
✓ 10.	GO0000785 : chromatin
...	
✓ 15.	GO0016363 : nuclear matrix

Qualitative evaluation



Relevant vs irrelevant : 82% - 18%

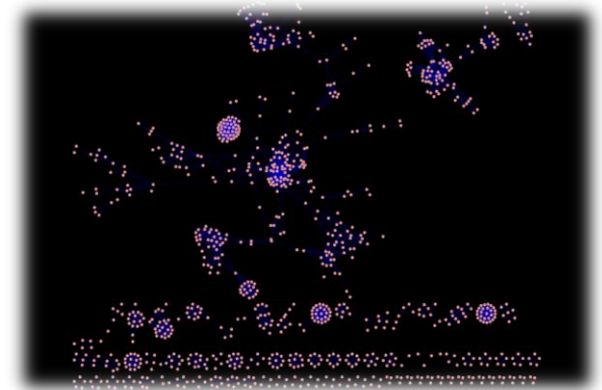
Conclusion and future work

- Automatic assignment of GO categories ~ 43%
[Camon et al 2003: GO kappa ~ 40%]
- Classification model improves faster than drift
[→ Consistency of annotation guidelines 😊]
- Next: Effective integration into the EAGLi' question-answering platform

Collaborations

- Automatic Functional Annotation of PubChem BioAssays

→ Generates semantic similarity clusters



- Automatically populating large protein datasets

COMBREX
COMputational BRidges to Experiments

76910213

2511394

780017

All genes in COMBREX (3302393 genes)

Genes with unvalidated predicted functions

Please visit EAGLi, the Bio-medical question answering engine <http://eagl.unige.ch/EAGLi/> !



EAGLi
Engine for question-Answering in Genomics Literature

Who are we ?

Search Engine examples | Question-Answering examples

Query:

Question-Answering examples:

- what diseases are associated with brca1 ?
- what diseases are important for boehringer ingelheim ?
- what proteins are associated with noonan syndrome ?
- what proteins can interact with cfr ?
- what drugs can prevent preterm birth delivery ?
- what cells are involved in heart looping ?
- what is retinoblastoma ?
- where is produced testosterone ?
- where is located the memory ?
- what protein can interact with fluorouracil ?

How to formulate a question

what proteins are associated with noonan syndrome ?

Search

Your question was : *what proteins are associated with noonan syndrome ?*, reformulated as *proteins associated noonan syndrome*

Possible answers are :

- ptpn11* (55 matches in 26 documents)
- raf1* (28 matches in 8 documents)
- sos1* (19 matches in 8 documents) ; *kras* (13/7) ; *erk1* (14/4) ; *gt* (7/5) ; *shoc2* (8/3) ; *ptp* (7/2) ; *nras* (6/3) ; *sh2* (5/2) ; *nf1* (5/1) ; *rap* (4/1) ; *sos* (10/1) ; *braf* (3/3) ; *hras* (3/2) ; *stat3* (2/1) ; *igf1* (2/1) ; *mek1* (2/2) ; *pip* (2/1) ... [show all](#)

ToxiCat on the selected articles (Beta) : 

ptpn11

PubMed  **Atrioventricular canal defect in patients with RASopathies.**
Diqilio MC , Romana Lepri F , Dentici ML , Henderson A , Baban A , Cristina Roberti M , Capolino R , Versacci P , Surace C , Anqioni A , Tartaglia M , Marino B , Dallapiccola B
Eur J Hum Genet. 2012 Jul
Pmid : 22781091
... diagnosed in 8/101 (8%) patients, including seven with a **PTPN11** gene mutation, and one single subject with ...

PubMed  **PTPN11-associated mutations in the heart: has LEOPARD changed Its RASpots?**
Lauriol J , Kontaridis MI
Trends Cardiovasc Med. 2011 May; 21(4): 97-104
Pmid : 22681964
... **PTPN11-associated** mutations in the heart: has LEOPARD changed Its RASpots ...

PubMed  **A rasopathy phenotype with severe congenital hypertrophic obstructive cardiomyopathy associated with a PTPN11 mutation and a novel variant in SOS1.**
Fahrner JA , Frazier A , Bachir S , Walsh MF , Applegate CD , Thompson R , Halushka MK , Murphy AM , Gunay-Aygun M
Am J Med Genet A. 2012 Jun; 158A(6): 1414-21
Pmid : 22585553
... A rasopathy phenotype with severe congenital hypertrophic obstructive cardiomyopathy associated with a **PTPN11** mutation and a novel variant ...

The Gene Ontology Categorizer: <http://eagl.unige.ch/GOCat/>



Out-of-date tutorial

PMID

Query

Browser model

Predictive model

Go to [EAGLi](#): a search engine for biology and medicine.

Other resources... TWINC (patent retrieval...)
<http://bitem.hesge.ch>



BITEM
Bibliomics and Text Mining Group

Home Research Publications Resources Partners People News

Last news

2012-11-11:
NETTAB in Como
Type: News
Patrick Ruch will present two papers next week at @

2012-11-08:
KART2
Type: Resource

BITEM Group

Who we are...

The BITEM Group, headed by Patrick Ruch, is part of the Information Sciences Department of the HEG (University of Applied Sciences, Geneva). It gathers a network of researchers (computer scientists, biologists, bioinformaticians...) affiliated to

Recent Publications

- Building a transnational biosurveillance network using semantic web technologies: requirements, design, and preliminary evaluation
- Pathogens and Gene Product Normalization in the Biomedical

Acknowledgments

- Swiss-prot group (SIB): Anne-Lise Veuthey, Yoannis Yenarios
- U. Indiana/SCRIPPS:
Rajarshi Guha / Stephan Schurer
- The COMBREX project: Martin Steffen
- NextProt: Pascale Gaudet

- SNF Grant: EAGL # 120758
- EU FP7: www.KHRESMOI.eu # 257528