

# Applications of Text and Data Mining of biomedical databases

Miguel Andrade Computational Biology & Data Mining group Max Delbrück Center for Molecular Medicine miguel.andrade@mdc-berlin.de Gene structures

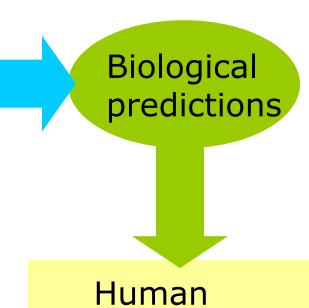
Gene expression

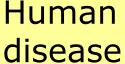
Protein sequences

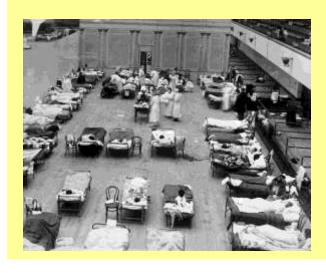
Protein databases

Literature databases

AGCTGGTACGAAGATGTCTCGCA
MLVPIEKAEVPRYILKTEFRKAILTS
In a phosphorylation dependent
001001001001001010111101:







## Molecular Biology databases

Protein and nucleotide sequences (UniProt, Entrez), Protein domains (PFAM, SMART), Structures (PDB), Diseases (OMIM), Gene expression (GEO), Bibliography (records, MEDLINE) (full text, PubMed Central)

## Molecular Biology databases

Bibliography (records, MEDLINE) (full text, PubMed Central)

Compressed PubMed in XML: 17GB

23M items (exhaustive back to 1966, oldest from 1809)

PubMed Central open access subset **26GB** of raw XML files (text only), compressed 8GB.

2.6M items

## Molecular Biology databases

Bibliography (records, MEDLINE) (full text, PubMed Central)

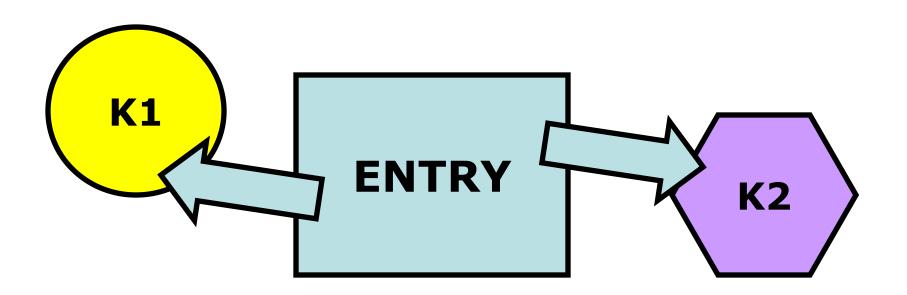
Compressed PubMed in XML: 17GB

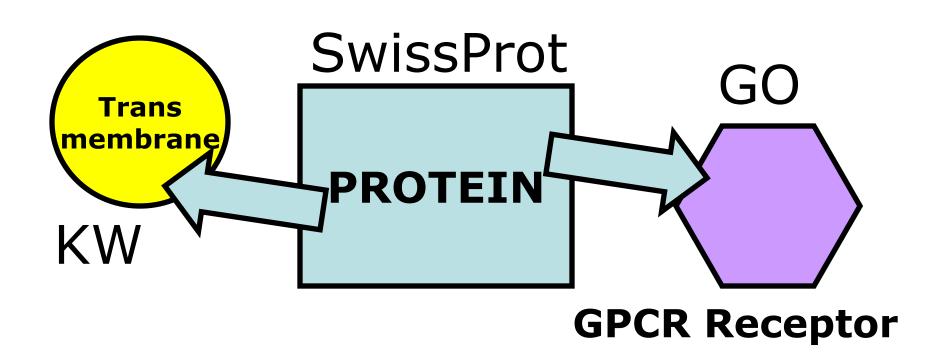
23M items (exhaustive back to 1966, oldest from 1809)

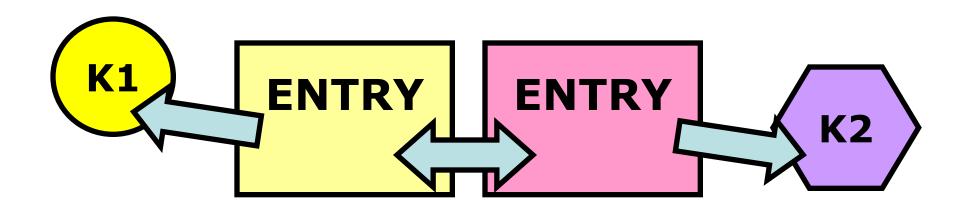
PubMed Central open access subset **26GB** of raw XML files (text only), compressed 8GB.

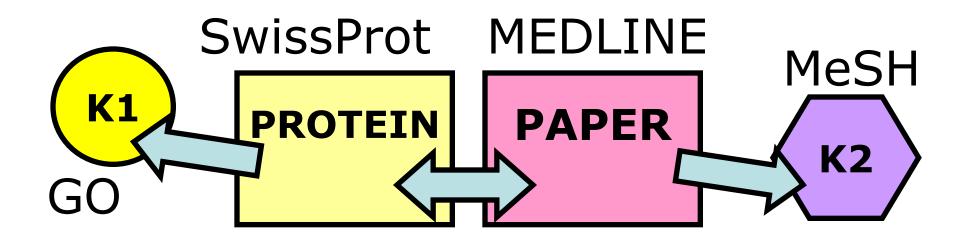
2.6M items

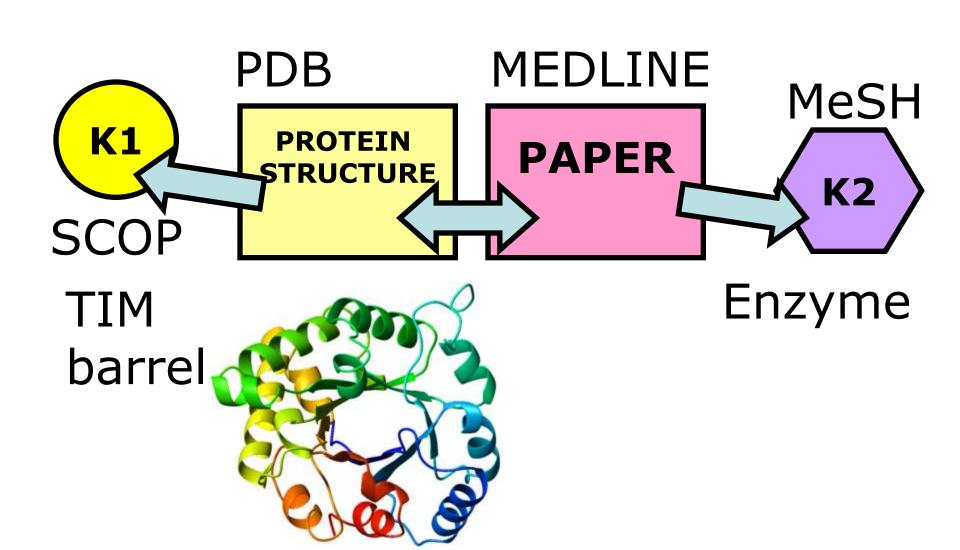
1 Human Genome 320GB



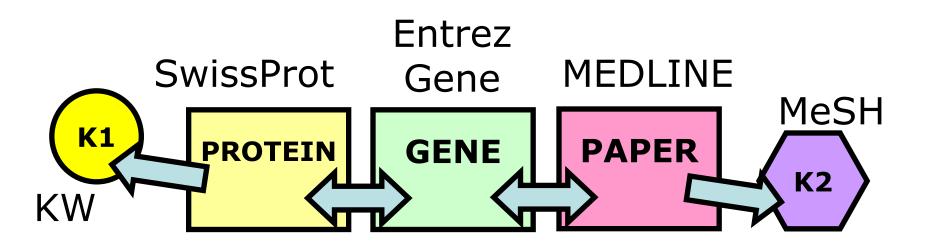


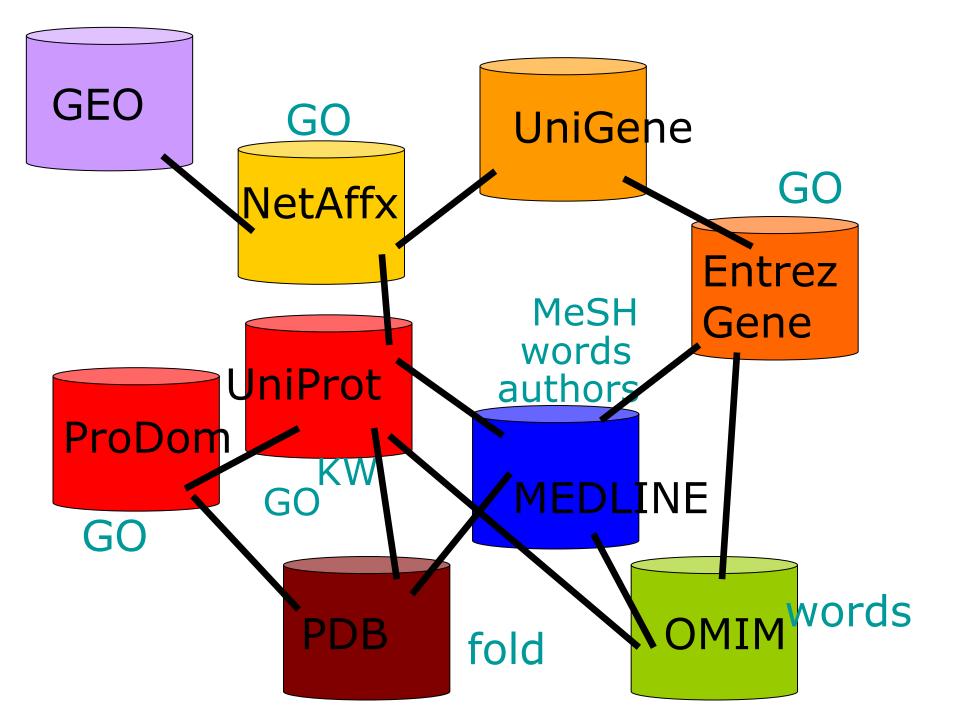




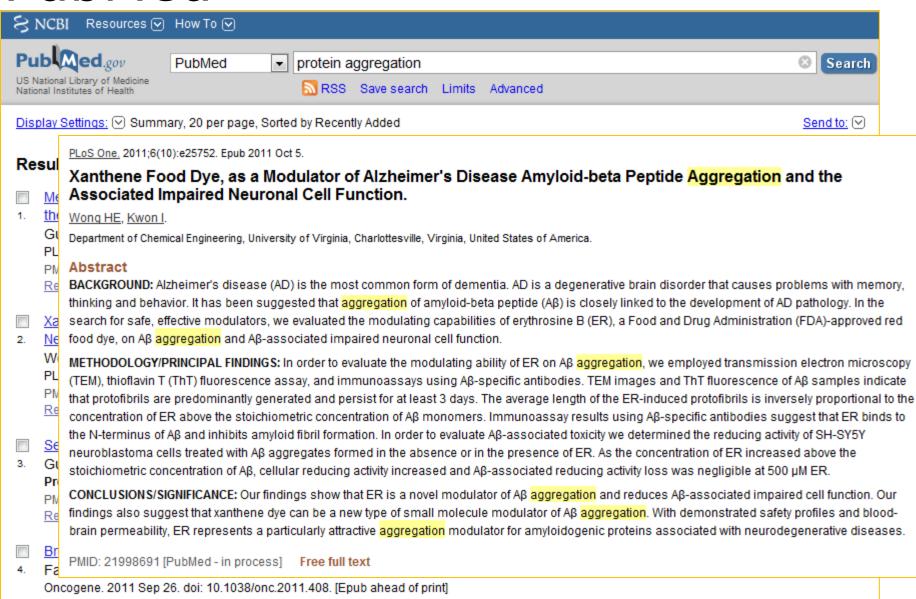


## Iterate!





### PubMed



Oncogene. 2011 Sep 26. doi: 10.1038/onc.2011.408. [Epub ahead of pri PMID: 21996731 [PubMed - as supplied by publisher] Related citations

#### PubMed



NCBI Resources 
 How To

Oncogene, 2011 Sep 26, doi: 10.1038/onc.2011.408. [Epub ahead of print]

Breast cancer genome-wide association studies: there is strength in numbers.

Fanale D, Amodeo V, Corsini LR, Rizzo S, Bazan V, Russo A.

Department of Surgical and Oncological Sciences, Section of Medical Oncology, University of Palermo, Palermo, Italy.

PLoS One. 2011;6(10):e25984. Epub 2011 Oct 6.

Membrane-bound steel factor maintains a high local concentration for mouse primordial germ cell motility, and defines the region of their migration.

Gu Y, Runyan C, Shoemaker A, Surani MA, Wylie C.

Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, United States of America.

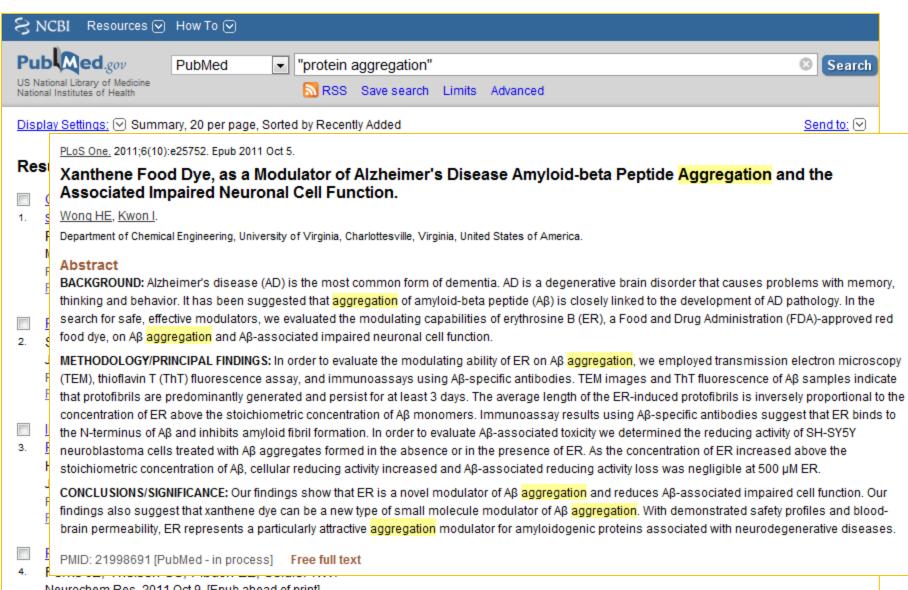
#### Abstract

Steel factor, the protein product of the Steel locus in the mouse, is a multifunctional signal for the primordial germ cell population. We have shown previously that its expression accompanies the germ cells during migration to the gonads, forming a "travelling niche" that controls their survival, motility, and proliferation. Here we show that these functions are distributed between the alternatively spliced membrane-bound and soluble forms of Steel factor. The germ cells normally migrate as individuals from E7.5 to E11.5, when they aggregate together in the embryonic gonads. Movie analysis of Steel-dickie mutant embryos, which make only the soluble form, at E7.5, showed that the germ cells fail to migrate normally, and undergo "premature aggregation" in the base of the allantois. Survival and directionality of movement is not affected. Addition of excess soluble Steel factor to Steel-dickie embryos rescued germ cell motility, and addition of Steel factor to germ cells in vitro showed that a fourfold higher dose was required to increase motility, compared to survival. These data show that soluble Steel factor is sufficient for germ cell survival, and suggest that the membrane-bound form provides a higher local concentration of Steel factor that controls the balance between germ cell motility and aggregation. This hypothesis was tested by addition of excess soluble Steel factor to slice cultures of E11.5 embryos, when migration usually ceases, and the germ cells aggregate. This reversed the aggregation process, and caused increased motility of the germ cells. We conclude that the two forms of Steel factor control different aspects of germ cell behavior, and that membrane-bound Steel factor controls germ cell motility within a "motility niche" that moves through the embryo with the germ cells. Escape from this niche causes cessation of motility and death by apoptosis of the ectopic germ cells.

PMID: 21998739 [PubMed - in process] Free full text

Fanale D, Amodeo V, Corsini LR, Rizzo S, Bazan V, Russo A. Oncogene. 2011 Sep 26. doi: 10.1038/onc.2011.408. [Epub ahead of print] PMID: 21996731 [PubMed - as supplied by publisher] Related citations

#### PubMed



Neurochem Res. 2011 Oct 9. [Epub ahead of print]
PMID: 21984199 [PubMed - as supplied by publisher]
Related citations

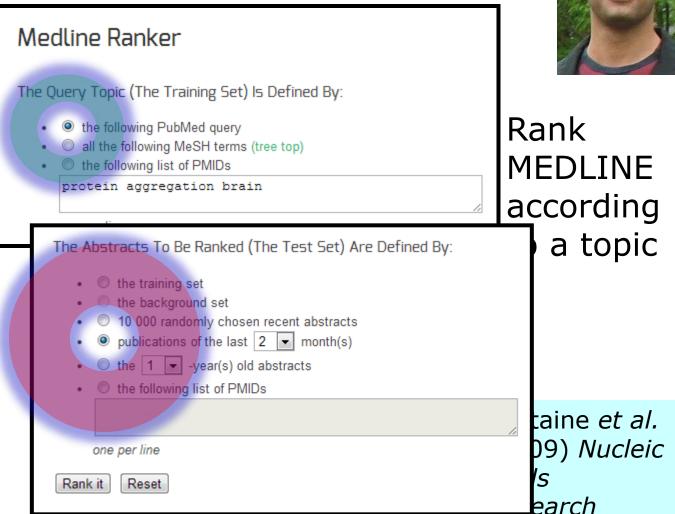
#### MedlineRanker

Jean-Fred Fontaine



#### Discriminative words

rank	word	weight
1	aggregation	7.13
2	synuclein	6.26
3	misfolding	5.59
4	1ewy	5.33
5	tangle	5.07
6	alzheimer	4.98
7	tau	4.92
8	neurodegeneration	4.89
9	amyloid	4.79
10	aggregate	4.75
11	app	4.59
12	huntington	4.58
13	fibril	4.49
14	prion	4.47
15	oligomer	4.44
16	ad	4.26
17	hallmark	4.26

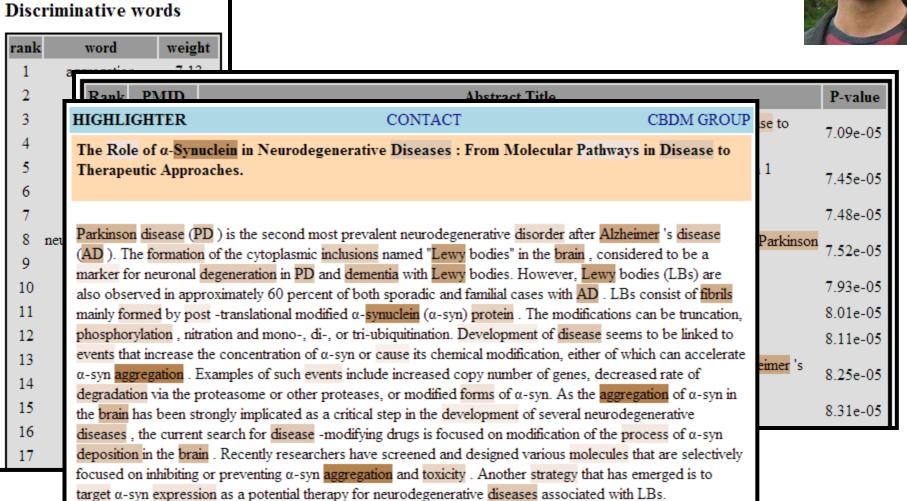


http://cbdm.mdc-berlin.de/tools/medlineranker/

### MedlineRanker

Jean-Fred Fontaine





http://cbdm.mdc-berlin.de/tools/medlineranker/

## Génie

#### Discriminative words

rank	word	weight
1	aggregation	7.13
2	synuclein	6.26
3	misfolding	5.59
4	lewy	5.33
5	tangle	5.07
6	alzheimer	4.98
7	tau	4.92
8	neurodegeneration	4.89
9	amyloid	4.79
10	aggregate	4.75
11	app	4.59
12	huntington	4.58
13	fibril	4.49
14	prion	4.47
15	oligomer	4.44
16	ad	4.26
17	hallmark	4.26

Define Your Topic Of Interest:

■ articles matching the following PubMed query
■ all articles associated with the following MeSH terms (tree top)
■ only the following PMIDs

protein aggregation brain

Ranks a set of genes from a whole genome according to a topic

2. Sel	ect The Genes To Be Ranked:
:	<ul> <li>all genes from this species (a taxonomic ID or scientific name)</li> <li>only the following NCBI Entrez Gene IDs</li> </ul>
	Human
	one per line
	Examples of species: human, chimpanzee, cattle, dog, mouse, rat, chick

Fontaine *et al.* (2011) *Nucleic Acids Research* 

http://cbdm.mdc-berlin.de/tools/genie/

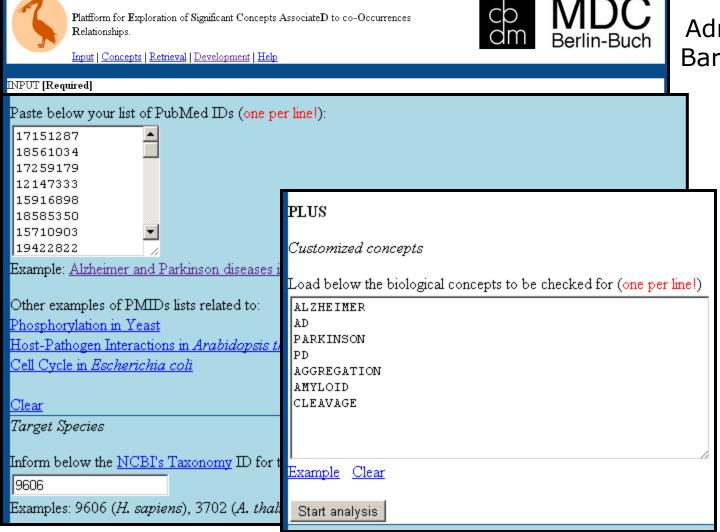
# Génie

Rank	GeneID	Symbol	Homologs PMIDs Hits FDF	Top 10 abstracts	
				21613474 (p=7.09e-05), 16651889 (p=7.17e-05), 15180968 (	p=7.35e-
1	6622	SNCA	HIGHLIGHTER	CONTACT CBDM GR	OUP
			Induction of intracellular tau a	ggregation is promoted by α-synuclein seeds and provides novel	
			insights into the hyperphospho	rylation of tau .	бе-
	0.51	4 DD			
2	351	APP			
				clusions, primarily composed of tau or α-synuclein (α-syn), are	
			_	of Alzheimer's disease (AD) and Parkinson's disease (PD),	)e-
3	4137	MAPT		stence of these pathological <mark>aggregates</mark> is identified in many luding spectrum disorders of AD and PD. Whereas α-syn can	
	4157	141111	_	yloidogenic fibrils, in vitro, tau polymerization requires an inducing age	nt I
				an-derived cellular model, in which recombinant, preformed α-syn fibri	_
			cross-seed intracellular tau to pro	mote the formation of neurofibrillary tangle -like aggregates . These	5e-
4	23621	BACE1	aggregates were hyperphosphory	lated, Triton insoluble, and thioflavin-S positive, either comingling with	
				gregates or induced by only exogenously applied recombinant α-syn fit	orils
				idogenic tau took over the cellular soma, displacing the nucleus and	
				likely preventing cellular function. Although a significant proportion of v	v <b>ild</b> - Be-
5	3064	HTT		usions, the P301L mutation in tau increased aggregation propensity r 50% of total tau protein. The role of phosphorylation on the develope	
			<u> </u>	r 30% of total <b>tau protein</b> . The role of phosphorylation on the developt tigated by coexpressing glycogen synthase kinase 3 β or microtubule-	hent
				finity-regulating kinase 2. Expression of either kinase inhibited the formation	ation
				Analyses of phosphorylation sites suggest that multiple complex factors	) C -
6	5663	PSEN1		and that Triton-soluble versus Triton-insoluble tau may be independent	
				work not only provides an exceptional cellular model of tau pathology ,	
				inclusion formation and provides novel insights into hyperphosphorylatio	n
			observed in disease.		

http://cbdm.mdc-berlin.de/tools/genie/

## PESCADOR

PESCADOR







Extract interactions and filter by concepts

Barbosa-Silva et al. (2010) BMC Bioinformatics

Barbosa-Silva et al. (2011) BMC Bioinformatics

http://cbdm.mdc-berlin.de/tools/pescador/

### PESCADOR

#### PESCADOR

Platt Target Organism: Homo sapiens (TAXID: 9606)

Click on a co-occurrence type to filter the type of co-occurrences displayed.

Selected PubMed ID: <u>10764738</u>

<u>Sentence 1</u>: Microtubule-associated protein 1B is a component of cortical Lewy bodies and <mark>binds</mark> alpha-synuclein filaments.

Sentence 2: Lewy bodies, neuropathological hallmarks of Parkinson's disease and dementia with Lewy bodies, comprise alpha-synuclein filaments and other less defined proteins.

Sentence 3: Characterization of Lewy body proteins that interact with alpha-synuclein may provide insight into the mechanism of Lewy body formation.

Sentence 4: Double immunofluorescence labeling and confocal microscopy revealed approximately 80% of cortical Lewy bodies contained microtubule-associated protein 1B (MAP-1B) that overlapped with alpha-synuclein.

Sentence 5: Lewy bodies were isolated using an immunomagnetic technique from brain tissue of patients dying with dementia with Lewy bodies.

Sentence 6: Lewy body proteins were resolved by polyacrylamide gel electrophoresis.

Sentence 7: Immunoblotting confirmed the presence of MAP-1B and alpha-synuclein in purified Lewy bodies.

Sentence 8: Direct binding studies revealed a high affinity interaction (IC(50) approximately 20 nm) between MAP-1B and alpha-synuclein.

Sentence 9: The MAP-1B-binding sites were mapped to the last 45 amino acids of the alpha-synuclein C terminus.

Sentence 10: MAP-1B also bound in vitro assembled alpha-synuclein fibrils.

Sentence 11: Thus, MAP-1B may be involved in the pathogenesis of Lewy bodies via its interaction with monomeric and fibrillar alphasvnuclein.

## PESCADOR

Type 1

MAP-1B also bound in vitro assembled alpha-synuclein fibrils.

Term + [Biointeraction] + Term

Type 2

Direct binding studies revealed a high affinity interaction (IC(50) approximately 20 nm) between MAP-1B and alpha-synuclein.

[Biointeraction] +Term + Term + [Biointeraction]

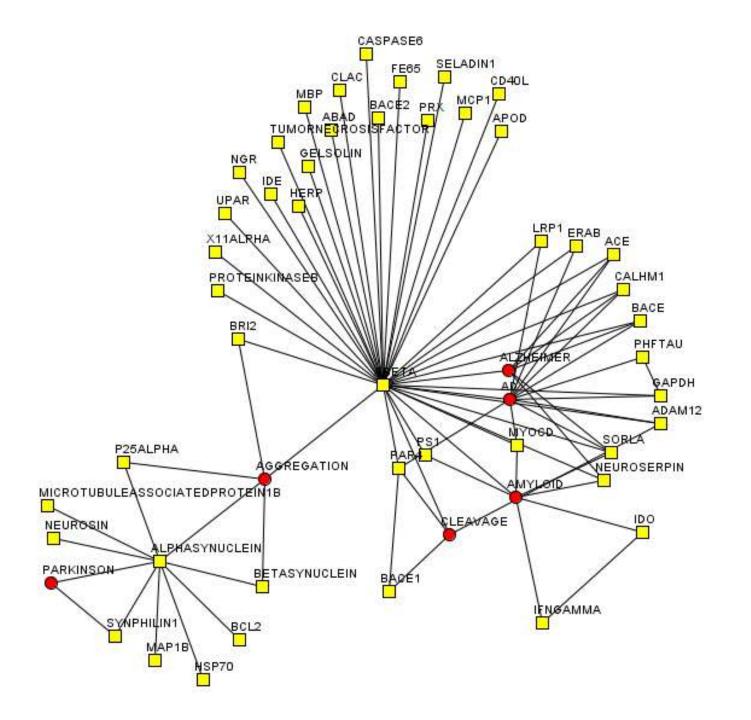
Type 3

Immunoblotting confirmed the presence of MAP-1B and alpha-synuclein in purified Lewy bodies.

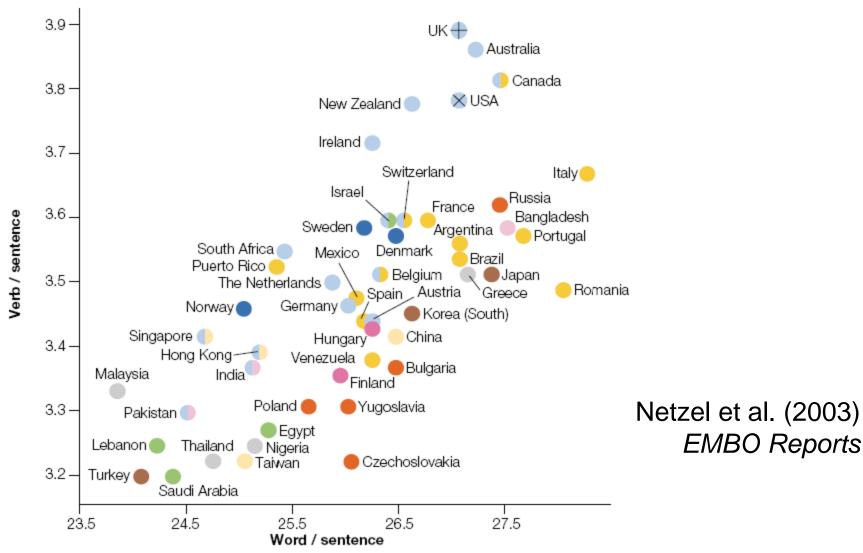
Term + Term

Type 4

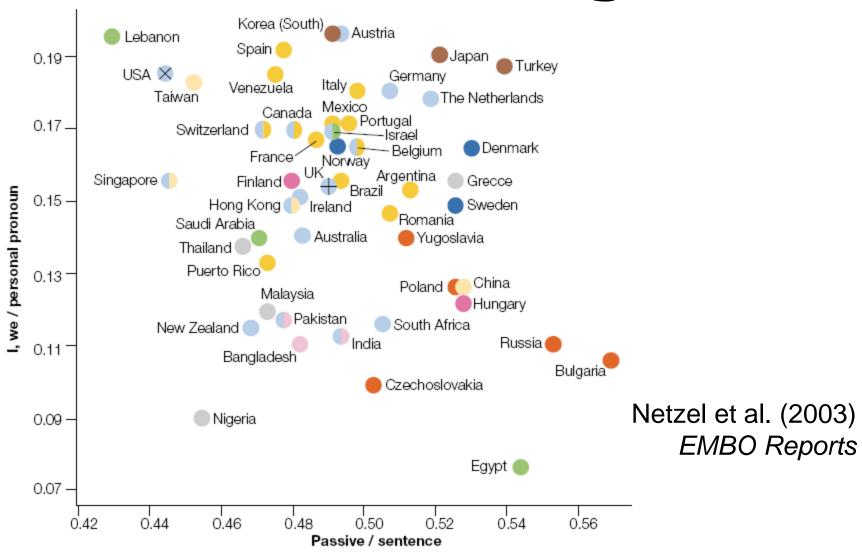
co-occurrence in abstract



# Country-specific variations of English

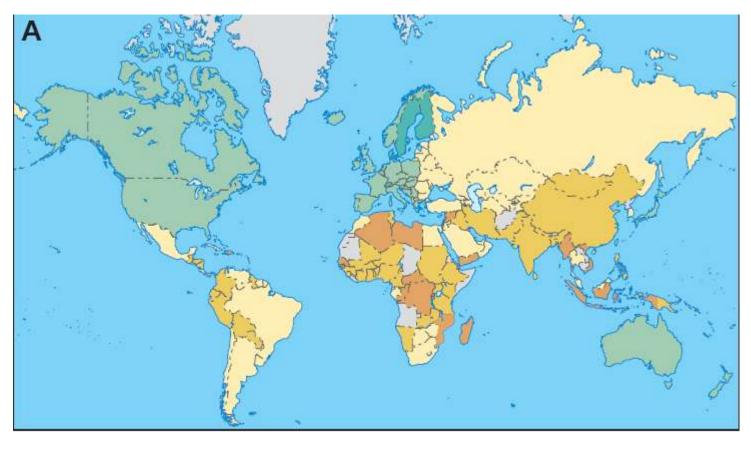


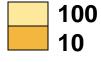
# Country-specific variations of English



#### Worldwide scientific publishing activity

Approximate amount of publications for the years 1996–2001 per million inhabitants by country:

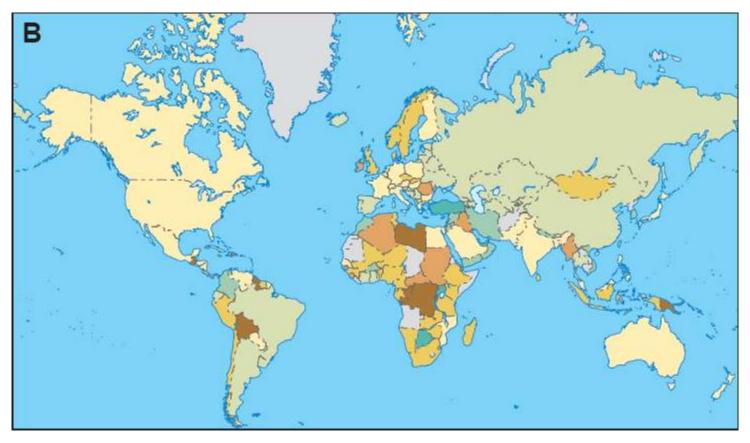






#### Worldwide scientific publishing activity

#### Ratio publications for 1996–2001 / 1989–95









Perez-Iratxeta and Andrade (2002) Science

# peer2ref

#### Find referees

http://www.ogic.ca/peer2ref/

#### peer2ref

Get suggested referees for your paper

about supplement contact us

#### Paste title + abstract from your manuscript help example

DNA methylation is a dynamic epigenetic mark that undergoes extensive changes during differentiation of self-renewing stem cells. However, whether these changes are the cause or consequence of stem cell fate remains unknown. Here, we show that alternative functional programs of hematopoietic stem cells (HSCs) are governed by gradual differences in methylation levels. Constitutive methylation is essential for HSC self-renewal but dispensable for homing, cell cycle control and suppression of apoptosis. Notably, HSCs from mice with reduced DNA methyltransferase 1 activity cannot suppress key myeloerythroid regulators and thus can differentiate into myeloerythroid, but not lymphoid, progeny. A similar methylation dosage effect controls stem cell function in leukemia. These data identify DNA methylation as an essential epigenetic mechanism to protect

Reset

run with advanced options



Carolina Perez-Iratxeta (OHRI-Ottawa)

# peer2ref

#### Find referees

http://www.ogic.ca/peer2ref/

Blood. 2005 Apr 1;105(7):2717-23. Epub 2004 Nov 30.

Related citations

PMID: 15572596 [PubMed - indexed for MEDLINE] Free Article



	NCBI Resources ☑ How To ☑					
	bli€d.gov PubMed ▼					
	ational Library of Medicine Advanced Advanced					
nome about	<u>Display Settings:</u>					
suppl conta	Results: 5					
	IFN-gamma negatively modulates self-renewal of repopulating human hemopoietic stem cells.  Yang L, Dybedal I, Bryder D, Nilsson L, Sitnicka E, Sasaki Y, Jacobsen SE.  J Immunol. 2005 Jan 15;174(2):752-7.  PMID: 15634895 [PubMed - indexed for MEDLINE] Free Article  Related citations					
	Molecular evidence for hierarchical transcriptional lineage priming in fetal and adult stem cells and multipotent progenitors.  Månsson R, Hultquist A, Luc S, Yang L, Anderson K, Kharazi S, Al-Hashmi S, Liuba K, Thorén L, Adolfsson J, Buza-Vidas N, Qian Soneji S, Enver T, Sigvardsson M, Jacobsen SE.  Immunity. 2007 Apr;26(4):407-19. Epub 2007 Apr 12.  PMID: 17433729 [PubMed - indexed for MEDLINE]  Related citations	Н,				
	Identification of Lin(-)Sca1(+)kit(+)CD34(+)Flt3- short-term hematopoietic stem cells capable of rapidly reconstituting and rescuing myeloablated transplant recipients.  Yang L, Bryder D, Adolfsson J, Nygren J, Månsson R, Sigvardsson M, Jacobsen SE.	g				

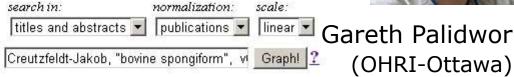
# **MLTrends**

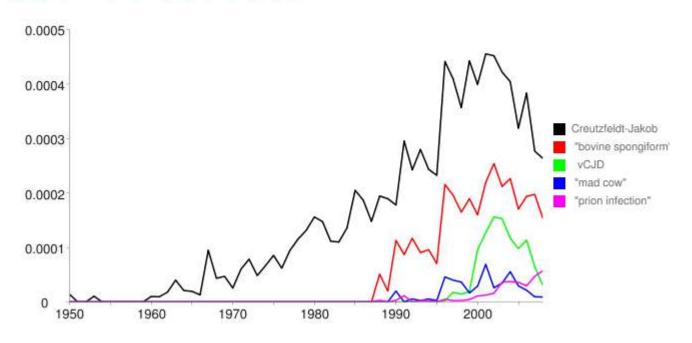
Graph historical term usage in MEDLINE

http://www.ogic.ca/mltrends/









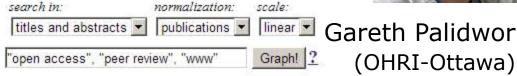
# **MLTrends**

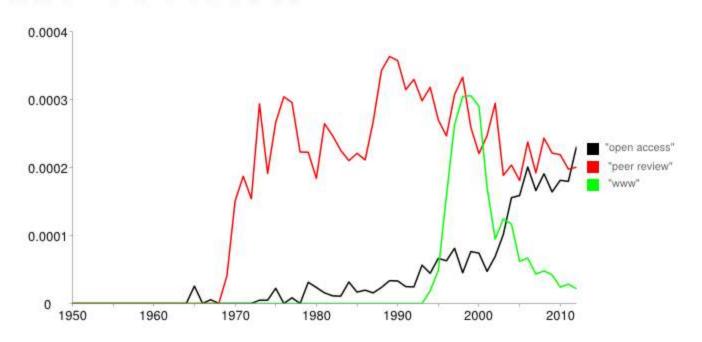
Graph historical term usage in MEDLINE

http://www.ogic.ca/mltrends/

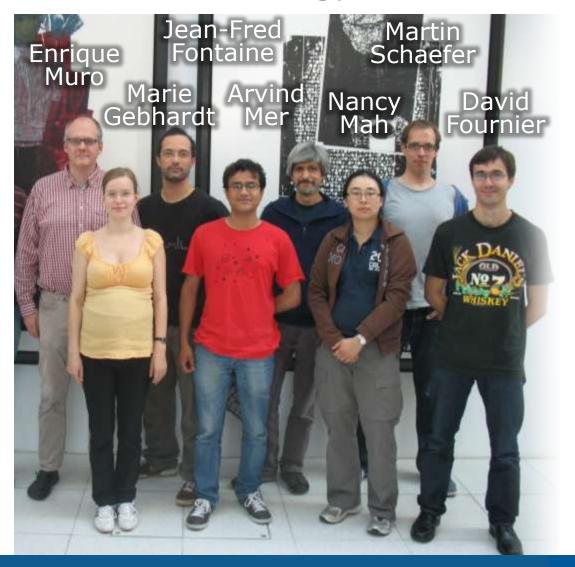








#### **Computational Biology and Data Mining group**





Alliance on Systems Biology







http://cbdm.mdc-berlin.de/

