# **Biomedical literature mining** (and why we *really* need Open Access)



Lars Juhl Jensen EMBL Heidelberg

## why biomedicine?

## why literature mining?

why open access?

#### MEDLINE

#### 17 million citations



Jensen et al., Nature Reviews Genetics, 2006

#### too much to read

## literature mining

#### open access

### information retrieval

## finding the papers

#### ad hoc retrieval

S NCBI	Publed National Library of Medicine NLM
All Databases	PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books
Search   PubMed	Go Clear
	Limits Preview/Index History Clipboard Details
About Entrez	• To get started, enter one or more search terms.
Text Version	• Search terms may be topics, authors or journals.
Entrez PubMed Overview Help   FAQ Tutorial New/Noteworthy E-Utilities PubMed Services Journals Database MeSH Database Single Citation Matcher Batch Citation Matcher Clinical Queries Special Queries LinkOut My NCBI (Cubby) Related Resources Order Documents NLM Mobile NLM Catalog NLM Gateway TOXNET Consumer Health Clinical Alerts	Willingson       News flash: deliver PubMed search results directly to your desktop with an RSS feed.         To set up an RSS feed:       In Run your search in PubMed.         (2) Select RSS Feed from the Send to menu.       Control Create Feed and copy the XML icon into your RSS Reader.         Read the PubMed Help to explore other options for automated e-mail updates using My NCBI.         PubMed is a service of the National Library of Medicine that includes over 15 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s. PubMed includes links to full text articles and other related resources.
PubMed Central	
	<u>Write to the Help Desk</u> <u>NCBI   NLM   NIH</u> <u>Department of Health &amp; Human Services</u> <u>Privacy Statement   Freedom of Information Act   Disclaimer</u>

users-specified query

"yeast AND cell cycle"

## stemming

## yeast / yeasts

dynamic query expansion

## yeast / S. cerevisiae

S NCBI	Publed National Library of Medicine NLM [Sign In] [Register]
All Databases Search PubMed	PubMed     Nucleotide     Protein     Genome     Structure     OMIM     PMC     Journals     Books       Image: The second
	Limits Preview/Index History Clipboard Details
About Entrez	Display Summary Show 20 Sort by Send to
Text Version	All: 13713 Review: 1460 🛠
Entrez PubMed	Items 1 - 20 of 13713 Page 1 of 686 Next
Help   FAQ Tutorial	1: Homma MK, Wada I, Suzuki T, Yamaki J, Krebs EG, Homma Y.
New/Noteworthy E-Utilities	CK2 phosphorylation of eukaryotic translation initiation factor 5 potentiates cell cycle progression. Proc Natl Acad Sci U S A. 2005 Oct 14; [Epub ahead of print] PMID: 16227438 [PubMed - as supplied by publisher]
PubMed Services	2: Leevers SJ, McNeill H. Links
MeSH Database Single Citation Matcher	Controlling the size of organs and organisms. Curr Opin Cell Biol. 2005 Oct 11; [Epub ahead of print] PMID: 16226450 [PubMed - as supplied by publisher]
Batch Citation Matcher Clinical Queries	T 3: <u>Wu JQ, Pollard TD.</u> Related Articles, Links
Special Queries LinkOut My NCBI (Cubby)	Counting cytokinesis proteins globally and locally in fission yeast. Science. 2005 Oct 14;310(5746):310-4. PMID: 16224022 [PubMed - in process]
Related Resources	T 4: David-Pfeuty T. Related Articles, Links
Order Documents NLM Mobile NLM Catalog NLM Gateway TOXNET	The flexible evolutionary anchorage-dependent Pardee's restriction point of mammalian cells. How its deregulation may lead to cancer. Biochim Biophys Acta. 2005 Sep 20; [Epub ahead of print] PMID: 16219425 [PubMed - as supplied by publisher]
Consumer Health Clinical Alerts	T 5: <u>Anekonda TS, Reddy PH.</u> Related Articles, Links
ClinicalTrials.gov PubMed Central	Neuronal protection by sirtuins in Alzheimer's disease. J Neurochem. 2005 Oct 7; [Epub ahead of print] PMID: 16219030 [PubMed - as supplied by publisher]
	6: Heinisch JJ. Related Articles, Links
	Baker's yeast as a tool for the development of antifungal kinase inhibitors-targeting protein kinase C and the cell integrity pathway. Biochim Biophys Acta. 2005 Sep 12; [Epub ahead of print]

#### MEDLINE

#### abstracts

#### complete papers

Mitotic cyclin (Clb2)-bound Cdc28 (Cdk1 homolog) directly phosphorylated Swe1 and this modification served as a priming step to promote subsequent Cdc5-dependent Swe1 hyperphosphorylation and degradation



cell cycle?

entity recognition

## identifying the substance(s)

Mitotic cyclin (Clb2)-bound Cdc28 (Cdk1 homolog) directly phosphorylated Swe1 and this modification served as a priming step to promote subsequent Cdc5-dependent Swe1 hyperphosphorylation and degradation

## $Cdc28 \Rightarrow yeast$

## $Cdc28 \Rightarrow cell cycle$

good synonyms list

#### manual curation

## orthographic variation

## CDC28
### Cdc28p

## disambiguation

# hairy

#### SDS

#### Cdc2



Information hyperlinked Over Proteins

#### Search Gene

Gene Model Developer's Zone new Contact Help

ceceen



site powered by pdg

Concept & Implementation by Robert Hoffmann



Hoffmann, R., Valencia, A. A Gene Network for Navigating the Literature. Nature Genetics 36, 664 (2004)

#### Search for a gene synonym or accession number... (Click here for an example: SNF1)

all fields	💌 in	all organism	-
Construction of the constr			

[SEARCH]

Symbol Name		Synonyms	Organism
CDC28 Cell divis	ion control protein 28	CDK1, HSL5, SRM5, YBR1211, YBR160W	Saccharomyces cerevisiae
UniProt IntAct NCBI Gene NCBI RefSeq NCBI Accession	P00546 P00546 852457 NP_009718 CAA25065, CAA56509, CAA85119		
Homologues of CI	DC28 new		
Definitions for CD	C28 📑		
Enhanced PubMee	d/Google query new		
WARNING: Please keep	in mind that gene detection is done automatically and can exhibit a ce	ertain error. Read more.	
			Find in this Page
Furthermore, SW14	associates with CLB2 protein and is a substrate for the	CLB2-associated CDC28 kin	nase in vitro. 🛛 📔 圭 👗
Furthermore, the C	ks1 protein was shown to be physically associated with	active forms of the Cdc28 pr	otein kinase. 🛛 📓 圭 👗
The cyclin-depende cerevisiae.	ent kinase Cdc28p associates with the cyclin Clb2p to in	nduce mitosis in the yeast Sa	ccharomyces 🛛 📓 圭 👗
We find that G1 arro	est in the cdc37-1 mutant is accompanied by a decrease	e in the Cdc28 activity associ	iated with the G1 📔 圭 👗
We found that Hct1 of the cyclin-depend	was <b>phosphorylated</b> in vivo at multiple CDK consensu dent kinase Cdc28 is high and APC activity is low.	s sites during cell cycle stage	es when activity 🛛 📓 🛃
It is likely, therefore specific multimeric	e, that Cks1 mediates a more specialized function of the complexes or to localize properly in cellular compartmer	e Cdc28 kinase such as its al nts.	bility to form 🛛 📄 ± 👗
Cdc37 promotes th	ne stability of protein kinases Cdc28 and Cak1.		🔳 🛨 🕌
In addition, Cdc37	promotes the production of Cak1, but not that of Cdc28	, when coexpressed in insect	cells.
The B-type cyclins	CIb5 and CIb6 are the primary activators of the S phase	e function of the budding yea	ast CDK Cdc28. 🛛 📔 圭 👗
All three cak1 muta	ints displayed significant synthetic interactions with loss	-of-function mutations in CDC	C28 and <u>KIN28</u> . 📓 🛃 📕

#### abstracts

#### complete papers

#### information extraction

### formalizing the facts



#### co-mentioning

#### statistical methods

# Natural Language Processing

Gene and protein names Cue words for entity recognition Verbs for relation extraction

[nxexpr The expression of [nxgene the cytochrome genes [nxpg CYC1 and CYC7]]] is controlled by [nxpg HAP1] <u>Mitotic cyclin (Clb2)-bound Cdc28 (Cdk1</u> <u>homolog) directly phosphorylated Swe1</u> and this modification served as a priming step to promote subsequent <u>Cdc5-dependent Swe1</u> <u>hyperphosphorylation</u> and degradation



#### new discoveries

### text mining



#### WELCOME: This site is open during construction.





#### abstracts

#### complete papers

#### temporal trends



Jensen et al., Nature Reviews Genetics, 2006

#### buzzwords



Jensen et al., Nature Reviews Genetics, 2006

### grant applications

integration of text and data

#### Home · Download · Help/Info

#### 🌸 STRING





Genomic neighborhood



Species co-occurrence



Gene fusions



Experimental interaction data



Microarray expression data



Database imports



Literature mining

#### genotype to phenotype

	Rickettsia (2 species)
	Mesorhizobium loti
	Brucella melitensis –
	Bradyrhizobium japonicum –
	Rhizobiales subgroup (3 species)
	Caulobacter crescentus —
	Betaproteobacteria (2 species)
	Escherichia coli (4 species)
	Shigella flexneri – 🦳 🛛 🚺
	Salmonella (2 species)
	Yersinia (2 species) —
	Vibrio cholerae —
	Shewanella oneidensis — _ h H
	Buchnera aphidicola (2 species)
	Pasteurellaceae (2 species)
	Pseudomonas (2 species)
	Ralstonia solanacearum
	Xylella fastidiosa —
	Xanthomonas campestris
	Xanthomonas axonopodis
	Helicobacter pylori
	Campylobacter jejuni -
	Bacillus subtilis
	Bacillus halodurans-
	Listeria (2 species) —
	Staphylococci (4 species) — h
	Streptococcaceae (8 species)
1	Mollicutes (4 species)
	Clostridium acetobutylicum
	Clostridium perfringens
	Streptomyces coelicolor
	Mycobacteria (3 species) —
	Corynebacterium glutamicum
	Corynebacterium efficiens –
RS1 - <u>14</u> 12	Bifidobacterium longum —
	Leptospira interrogans
	Borrelia burgdorferi
	Treponema pallidum -
	Chlamydiales (5 species)
	Nostoc sp. PCC /120
	Synechocystis sp. PCC 6803
	Synechococcus elongatus -
	Deinesseus maiedurans
	Europodorium puoloolum
	Pusobactenum nucreatum
	Aquilex aeolicus

-0.19	-0.29	-0.37
0.99	0.82	0.39
-0.33	0.19	-0.34
-0.22	-0.35	-032
0.35	0.88	-0.77
2.03	3.07	-0.28
-1.65	-2.09	-1.69
0.12	0.08	0.52
-0.39	-0.77	-0.02
1.64	1.26	1.18
-0.78	-0.40	-0.1
-0.06	1.60	2.21
-0.10	0.53	-0.11
-0.05	-0.09	-0.06
-1.40	-1.43	-1.42
0.35	0.78	0.69
0.76	0.58	0.35
-0.57	-0.82	0.11
0.30	0.11	-0.40
-0.57	-0.82	0.11
0.71	0.73	-0.90
3.15	2.10	1.54
0.93	0.41	-0.72
0.10	0.73	0.27
0.76	-0.18	0.98
0.76	-0.18	0.98 -0.86
0.76 -1.97 -1.38	-0.18 -2.07 -1.95	0.98 -0.86 -2.02
0.76 -1.97 -1.38 -1.19	-0.18 -2.07 -1.95 -1.20	0.98 -0.86 -2.02 -0.66
0.76 -1.97 -1.38 -1.19 0.59	-0.18 -2.07 -1.95 -1.20 0.87	0.98 -0.86 -2.02 -0.66 0.79
0.76 -1.97 -1.38 -1.19 0.59 -1.29	-0.18 -2.07 -1.95 -1.20 0.87 -1.69	0.98 -0.86 -2.02 -0.66 0.79 1.52
0.76 -1.97 -1.38 -1.19 0.59 -1.29 -2.24	-0.18 -2.07 -1.95 -1.20 0.87 -1.69 -2.72	0.98 -0.86 -2.02 -0.66 0.79 1.52 -2.27
0.76 -1.97 -1.38 -1.19 0.59 -1.29 -2.24 -2.11	-0.18 -2.07 -1.95 -1.20 0.87 -1.69 -2.72 -2.57	0.98 -0.86 -2.02 -0.66 0.79 1.52 -2.27 -1.67
0.76 -1.97 -1.38 -1.19 0.59 -1.29 -2.24 -2.11 -1.03	-0.18 -2.07 -1.95 -1.20 0.87 -1.69 -2.72 -2.57 -1.07	0.98 -0.86 -2.02 -0.66 0.79 1.52 -2.27 -1.67 0.39
0.76 -1.97 -1.38 -1.19 0.59 -1.29 -2.24 -2.11 -1.03 -0.50	-0.18 -2.07 -1.95 -1.20 0.87 -1.69 -2.72 -2.57 -1.07 -0.48	0.98 -0.86 -2.02 -0.66 0.79 1.52 -2.27 -1.67 0.39 -0.14
0.76 -1.97 -1.38 -1.19 0.59 -1.29 -2.24 -2.11 -1.03 -0.50 -0.82	-0.18 -2.07 -1.95 -1.20 0.87 -1.69 -2.72 -2.57 -1.07 -0.48 -1.14	0.98 -0.86 -2.02 -0.66 0.79 1.52 -2.27 -1.67 0.39 -0.14 1.36
0.76 -1.97 -1.38 -1.19 0.59 -1.29 -2.24 -2.11 -1.03 -0.50 -0.82 0.48	-0.18 -2.07 -1.95 -1.20 0.87 -1.69 -2.72 -2.57 -1.07 -0.48 -1.14 0.90	0.98 -0.86 -2.02 -0.66 0.79 1.52 -2.27 -1.67 0.39 -0.14 1.36 0.75
0.76 -1.97 -1.38 -1.19 0.59 -1.29 -2.24 -2.11 -1.03 -0.50 -0.82 0.48 1,93	-0.18 -2.07 -1.95 -1.20 0.87 -1.69 -2.72 -2.57 -1.07 -0.48 -1.14 0.90 0.82	0.98 -0.86 -2.02 -0.66 0.79 1.52 -2.27 -1.67 0.39 -0.14 1.36 0.75 -0.34
0.76 -1.97 -1.38 -1.19 0.59 -1.29 -2.24 -2.11 -1.03 -0.50 -0.82 0.48 1,93 0.47	-0.18 -2.07 -1.95 -1.20 0.87 -1.69 -2.72 -2.57 -1.07 -0.48 -1.14 0.90 0.82 1.22	0.98 -0.86 -2.02 -0.66 0.79 1.52 -2.27 -1.67 0.39 -0.14 1.36 0.75 -0.34 -0.94
0.76 -1.97 -1.38 -1.19 0.59 -1.29 -2.24 -2.11 -1.03 -0.50 -0.82 0.48 1.93 0.47 -1.04	-0.18 -2.07 -1.95 -1.20 0.87 -1.69 -2.72 -2.57 -1.07 -0.48 -1.14 0.90 0.82 1.22 -1.34	0.98 -0.86 -2.02 -0.66 0.79 1.52 -2.27 -1.67 0.39 -0.14 1.36 0.75 -0.34 -0.94 -0.94 -1.05
0.76 -1.97 -1.38 -1.19 0.59 -1.29 -2.24 -2.11 -1.03 -0.50 -0.82 0.48 1.93 0.47 -1.04 -0.56	-0.18 -2.07 -1.95 -1.20 0.87 -1.69 -2.72 -2.57 -1.07 -0.48 -1.14 0.90 0.82 1.22 -1.34 -0.13	0.98 -0.86 -2.02 -0.66 0.79 1.52 -2.27 -1.67 0.39 -0.14 1.36 0.75 -0.34 -0.94 -1.05 -0.56
0.76 -1.97 -1.38 -1.19 0.59 -1.29 -2.24 -2.11 -1.03 -0.50 -0.82 0.48 1.93 0.47 -1.04 -0.56 -0.89	-0.18 -2.07 -1.95 -1.20 0.87 -1.69 -2.72 -2.57 -1.07 -0.48 -1.14 0.90 0.82 1.22 -1.34 -0.13 -0.53	0.98 -0.86 -2.02 -0.66 0.79 1.52 -2.27 -1.67 0.39 -0.14 1.36 0.75 -0.34 -0.94 -1.05 -0.56 -0.91
0.76 -1.97 -1.38 -1.19 0.59 -1.29 -2.24 -2.11 -1.03 -0.50 -0.82 0.48 1.93 0.47 -1.04 -0.56 -0.89 0.01	-0.18 -2.07 -1.95 -1.20 0.87 -1.69 -2.72 -2.57 -1.07 -0.48 -1.14 0.90 0.82 1.22 -1.34 -0.13 -0.53 0.82	0.98 -0.86 -2.02 -0.66 0.79 1.52 -2.27 -1.67 0.39 -0.14 1.36 0.75 -0.34 -0.94 -1.05 -0.56 -0.91 -0.70
0.76 -1.97 -1.38 -1.19 0.59 -1.29 -2.24 -2.11 -1.03 -0.50 -0.82 0.48 1.93 0.47 -1.04 -0.56 -0.89 0.01 -0.22	-0.18 -2.07 -1.95 -1.20 0.87 -1.69 -2.72 -2.57 -1.07 -0.48 -1.14 0.90 0.82 1.22 -1.34 -0.13 -0.53 0.82 0.77	0.98 -0.86 -2.02 -0.66 0.79 1.52 -2.27 -1.67 0.39 -0.14 1.36 0.75 -0.34 -0.94 -1.05 -0.56 -0.91 -0.70 -0.22
0.76 -1.97 -1.38 -1.19 0.59 -1.29 -2.24 -2.11 -1.03 -0.50 -0.82 0.48 1.93 0.47 -1.04 -0.56 -0.89 0.01 -0.22 -0.53	-0.18 -2.07 -1.95 -1.20 0.87 -1.69 -2.72 -2.57 -1.07 -0.48 -1.14 0.90 0.82 1.22 -1.34 -0.13 -0.53 0.82 0.77 -0.78	0.98 -0.86 -2.02 -0.66 0.79 1.52 -2.27 -1.67 0.39 -0.14 1.36 0.75 -0.34 -0.94 -1.05 -0.56 -0.91 -0.70 -0.22 -0.55
0.76 -1.97 -1.38 -1.19 0.59 -1.29 -2.24 -2.11 -1.03 -0.50 -0.82 0.48 1.93 0.47 -1.04 -0.56 -0.89 0.01 -0.22 -0.53 -0.51	-0.18 -2.07 -1.95 -1.20 0.87 -1.69 -2.72 -2.57 -1.07 -0.48 -1.14 0.90 0.82 1.22 -1.34 -0.13 -0.53 0.82 0.77 -0.78 -0.75	0.98 -0.86 -2.02 -0.66 0.79 1.52 -2.27 -1.67 0.39 -0.14 1.36 0.75 -0.34 -0.94 -1.05 -0.56 -0.91 -0.70 -0.22 -0.55 -0.52

Korbel et al., PLoS Biology, 2005



Korbel et al., PLoS Biology, 2005



#### where are we now?


Jensen et al., Nature Reviews Genetics, 2006

# abstracts

# complete papers

## restricted access

### open access

## the tools are there

### now we need the text!

#### Acknowledgments

Jasmin Saric Rossitza Ouzounova Michael Kuhn Jan Korbel Tobias Doerks Isabel Rojas Miguel Andrade Peer Bork