OntoGene www.ontogene.org

ODIN: An Advanced Interface for the Curation of Biomedical Literature

Fabio Rinaldi, Simon Clematide, Gerold Schneider, Martin Romacker, Thérèse Vachon

Introduction

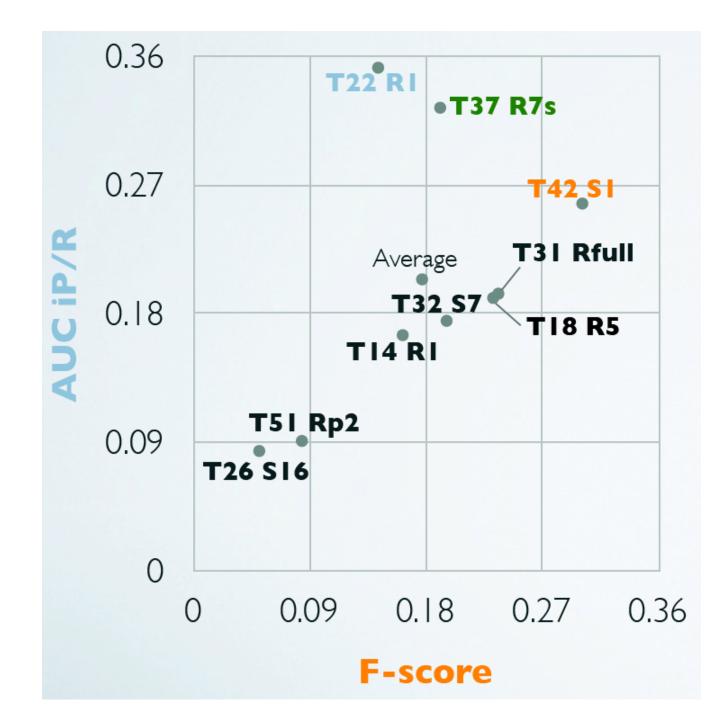
ODIN (Ontogene Document INspector) is a system for interactive curation of biomedical literature, developed within the scope of the **SASEBio** project (Semi-Automated Semantic Enrichment of the Literature), as a collaboration between the OntoGene group at the University of Zurich and the NITAS/TMS group of Novartis Pharma AG. The purpose of the system is to allow a human annotator/curator to leverage upon the results of an advanced text mining system in order to enhance the speed and effectiveness of the annotation process.

The OntoGene system takes as input a document (e.g a full paper from PubMed Central) and processes it with a custom NLP pipeline, which includes Named Entity recognition and relation extraction. Entities which are currently supported include proteins, genes, experimental methods, cell lines, species. Entities detected in the input document are disambiguated with respect to a reference database (UniProt, EntrezGene, NCBI taxonomy, PSI-MI ontology). The annotated documents are handed back to the ODIN interface, which allows multiple display modalities.

Evaluation: BioCreative competitions

In our initial application [1], relation mining was based on manually constructed cascading rules, which were organized modularly in order to support increasingly abstract types of queries. This approach has been validated through our participation in the BioCreative competitive evaluations of biomedical text mining systems. Our results in BioCreative II (2006), were among the best reported [2].

Later we developed a new approach based on a combination of machine learning and linguistic insight, which learns the syntactic paths expressing protein-protein interactions [3]. At BioCreative II.5 (2009) this approach served as the basis of our contribution. This resulted in the best run for the detection of protein-protein interactions (according to the 'raw' AUC iP/R metric). Our system was overall considered the best balanced system and among the best three [4] (look for system T37 in the graph to the right). In 2010, we have participated to the BioCreative III text mining challenge, achieving very competitive results in all of the tasks [5].

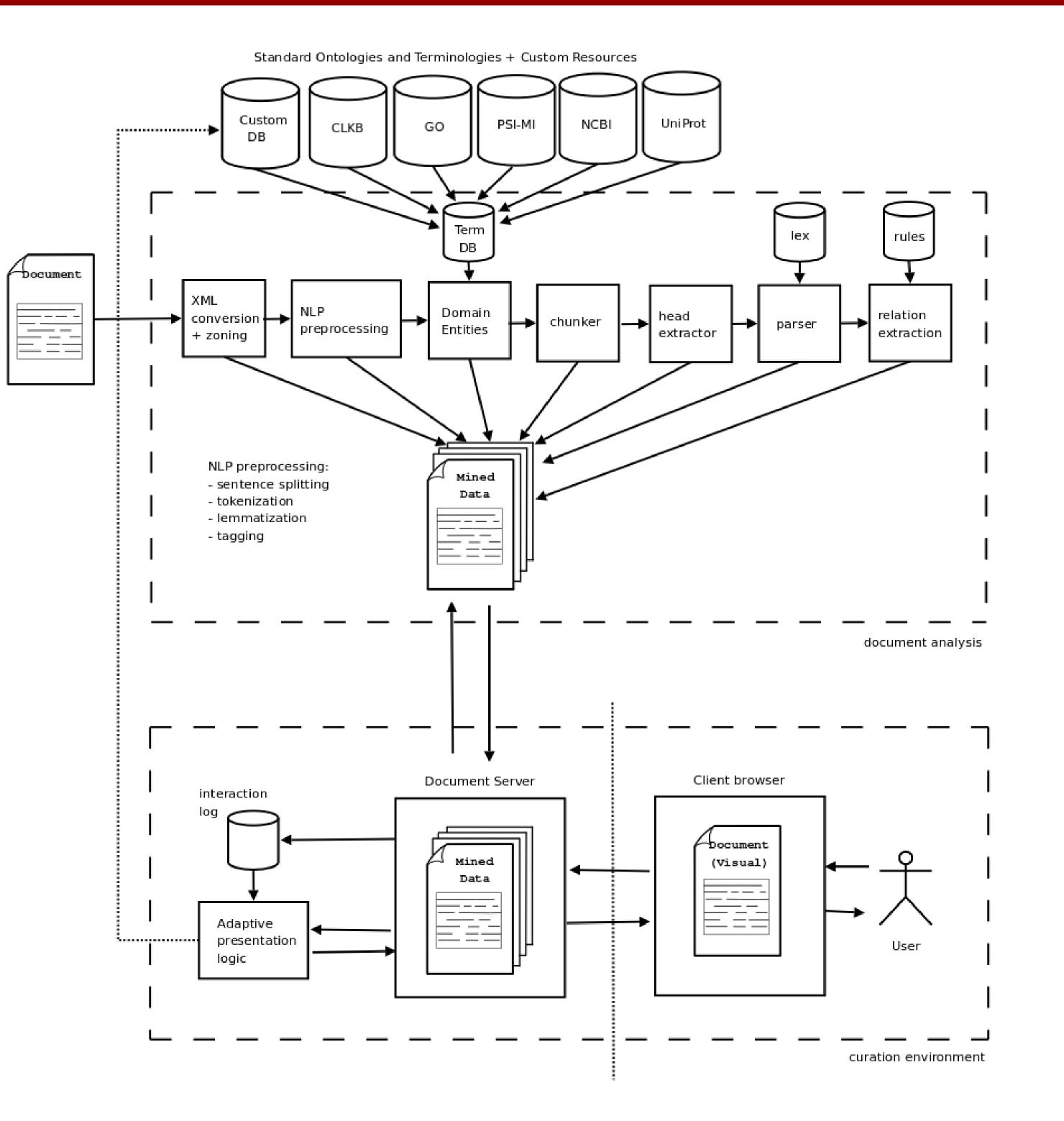


Inspectors	Comment PMC 1226314			Annotation				
Term Inspector	Show PubmedCentral			Genes/Protein	s Tem	ns Gro	uped Gene	es/Protei
Save Cancel Remove Term Show	GATA-1 mediates auto-regulation of Gfi-1B transcrip	tion in K562 cells		😨 Refresh 🛛 🥥 S	elected	\ominus Unsel	ected	🛒 Expor
Term:				i Concept	Sco	ore Freq	Туре	Zone
Gfi-1B		erythroid-specific transcription factor , whose expression plays an essential role in		∃ 📝 100041004	MOI 0	2	GEN	text
		n <mark>Gfi-1B</mark> promoter region and shown that <u>GATA-1</u> mediates erythroid-specific <u>Gfi-1B</u> B promoter , here we report that (i) Gfi - 1B transcription is negatively regulated by i	_	Browse Entrez	-			
erm Type:		ly to the Gfi-1 - like sites in the Gfi-1B promoter and (iii) Gfi-1B suppresses GATA-1 -	3	window.	Conc only	, 1000 1100	<u>• </u>	caro
GEN		teraction . These results not only demonstrate that interaction of GATA-1 and Gfi-1B		∃ 1026_HUM	AN 0	2	GEN	text
Concept Values:		expression of the Gfi-1B gene , but also provide the first evidence that Gfi-1B can exe		Browse Entrez		v 1026 in s	senarate w	vindow.
14582_MOUSE 8328_HUMAN		tion without direct binding to the Gfi-1 site of the target genes . Based on these data ,				/ <u>1020</u> III 3		
		rtant in restricting the expression level of Gfi-1B , thus optimizing its function in		10661_HU		1	GEN	text
	erythroid cells .	<u> </u>		∃ 10993_HUI		4	GEN	text
				H 11262_HU	MAN 0	1	GEN	text
comment:		an erythroid-specific Gfi - family transcriptional factor , which was identified by low		∃ 11770_MO	USE 0	2	GEN	text
		be (1). Both Gfi-1 and Gfi-1B contain a SNAG domain that mediates transcriptional	_	∃ 12575_MO	USE 0	2	GEN	text
		NA binding to the TAAATCAC (A/T) \underline{GCA} recognition sequence (1 - 3). Expression		∃ 12700_MO	USE 0	1	GEN	text
		in human (4,5), whereas Gfi-1 is more abundant in the lymphopoietic thymus (6 - 8 $\frac{1}{100}$ $\frac{1}{1$).	H 12702 MO	USE 0	2	GEN	text
Search Databases		arget genes of $Gfi-1B$ - mediated transcriptional repression (1,9). Since p21 is a cell		∃ 12703_MO		2	GEN	text
		s cytokine signaling , the functional role of <u>Gfi-1B</u> is considered to be important in		∃ 1387 HUM		1	GEN	text
Search Terms		ells . Its importance in erythropoiesis has been further highlighted by gene targeting yonic lethality due to loss of red blood cell formation (10). Enforced expression						
🗹 EMBL 🗹 UniProtKB		-1B induces a drastic expansion of erythroblast independent of its SNAG repression				1	GEN	text
		required for proliferation of erythroblasts (5) . On the other hand , a recent study has				1	GEN	text
		through its transcription repression function (11). Likely, the function of Gfi-1B in		∃ 14460_MO	USE 0	128	GEN	title
		ce context of its targeted gene promoter . Despite the differential roles of Gfi-1B in			USE 0	1	GEN	text
		e that elevation of Gfi-1B level alters the program of normal erythropoiesis (5,11) .		∃ 14581_MO	USE 0	25	GEN	abstra
		in erythroid cells and whether there is a direct relationship between Gfi-1B and other		∃ 14582_MO	USE 0	230	GEN	title
		ssion of many eukaryotic transcription factors has been shown to be auto-regulated		∃ 🔲 15452_MO	USE 0	1	GEN	text
	positively and negatively (12 – 16) . In most auto-regulatory cas	es , a given factor binds to its own promoter and either activates or represses		🗄 🔲 161882 HU	0 1AM	1	GEN	text
	transcription . In this study , we observed negative auto-regulation	on of <u>Gfi-1B</u> in <u>K562</u> cells . By analyzing the sequence of <u>human <u>Gfi-1B</u> gene promote</u>		∃ 16909 MO	USE 0	1	GEN	text
	region (17) , we found the presence of two tandem repeats of G	fi-1 - like sites located at $-59/-56$ and $-47/-44$ relative to its transcription start sit				2	GEN	text
	Very recently , a report has demonstrated that mouse <u>Gfi-1B</u> dire	ectly binds to the <u>Gfi-1</u> binding sites near the mRNA transcription start site of the <u>mou</u>	e				GEN	
	<u>Gfi-1B</u> promoter and is able to auto-repress its own expression (18) . However , here we showed that mutations in these two <u>Gfi-1</u> - like sites reduced		∃ 17863_MO				text
	the promoter activity of the human <u>Gfi-1B</u> promoter in K562 cells	s , indicating that these sites mediate transcriptional activation rather than silencing .	y	∃ 18045_MO		1	GEN	text
		d of <u>Gfi-1B</u> , is the main transcription factor preferentially binding to these non-typical			USE 0	1	GEN	text
		complex with GATA-1 , by which GATA-1 - mediated transcription is repressed by Gfi-	B	∃ 19165_MO	USE 0	2	GEN	text
		s a complex with GATA-1 and associates with the myc and myb promoters in mouse		🗄 📄 199699_HU	0 1AMI	2	GEN	text
		on of <u>Gfi-1B</u> in erythroid progenitors induces growth arrest and that expression of my		∃ 199_HUMA	N 0	2	GEN	text
		sized that GATA-1 / Gfi-1B is a repressive complex that suppresses transcription of m	C	± 20683_MO	USE 0	2	GEN	text
		he first direct evidence that transcriptional repression function of <u>Gfi-1B</u> can work		∃ 21784 MO		2	GEN	text
	through its interaction with GATA-1 independent of its direct DN	A binding to the gene promoter . Since our previous study has shown that <u>GATA-1</u> is	_			-	GEN	

ODIN: screenshot illustrating the inspection and editing functionalities.

The curator/annotator can view the whole document with in-line annotations highlighted, or can browse the extracted entities and be pointed back to the mentions of the entities

Architecture



within the original document. All entity mentions are entirely editable: the curator can easily add or delete any of them, and also change their extent (i.e. add/remove words to its right or left) with a simple click of the mouse. Different entity views are supported, with sorting capabilities according to different criteria (entity type, entity mention, confidence score, etc.). Selective highlighting of text units (e.g. sentences containing desired entities) is supported. Additionally, extensive logging functionalities are provided. All documents and entities are fully interlinked to reference databases, for the purpose of simplified inspection. Entities can be grouped in classes (e.g. by species) and actions can be applied to whole classes, for selective editing or removal.

File - View - Search - Mode -							Help +
Inspectors Comment PMC 2713427				Annotation			
Term Inspector	Show PubmedCentral Show Pubmed	Entry		Grouped Genes/Proteins			
Save Cancel Remove Term >>>		us [12] – [15] . The locus encodes two	n i	🔹 Refresh			
Term:		<u>4ARF</u> (<u>p19Arf</u> in <u>mice</u>), that activate the <u>ssor</u> pathways, respectively (reviewed in		Term	Concept	Туре	
CBX7	[10] , [19] . Both are implicated in			🗏 🍓 23492_HUM/	JMAN (Gene) 53 Terms		
Term Type:	when primary cells are exposed to various		CBX7	23492_HUMAN	GEN		
GEN		gnaling , telomere erosion or oxidative ears to be context dependent with <u>INK4a</u>		CBX7		GEN	- 11
Concept Values:	playing the predominant role in huma			CBX7	23492_HUMAN	GEN	
23492_HUMAN 52609_MOUSE		9]. Two strands of evidence point to an		CBX7	23492_HUMAN	GEN	_
362962_RAT	important role for Polycomb complexe	es in regulating the locus . Several mouse		Cbx7	23492_HUMAN	GEN	
	PcG gene knockouts have been show		CBX7	23492_HUMAN	GEN		
Comment:		erepression of Ink4a / Arf [20] - [25] .		CBX7	23492_HUMAN	GEN	
		1, Cbx7, and Cbx8 has been shown to		Cbx7	23492_HUMAN	GEN	
	delay senescence in both mouse and			Cbx7	23492_HUMAN	GEN	
		oader effort to characterize the entire		CBX7	23492_HUMAN	GEN	
Search Databases	range of <u>PRC1</u> complexes that regulate <u>INK4a</u> in human cells , we have identified <u>MEL18</u> as a direct repressor of the locus in partnership with either <u>CBX7</u> or <u>CBX8</u> . ChIP analyses revealed that <u>CBX7</u> , <u>CBX8</u> , <u>MEL18</u> and <u>BMI1</u> are CBX7						
Search Term Text							
A same A line and supp	all present at the promoter and first e			CBX7	23492_HUMAN	GEN	
Entrez VuniProt EMBL	A	egulation of p16INK4a Importantly , we		CBX7	23492_HUMAN	GEN	
Term Properties Type: GEN		articipate in multiple , distinct PRC1		CBX7	23492_HUMAN	GEN	
Value: 23492_HUMAN 52609_M	OUSE 362962_RAT	xes bind simultaneously to the INK4a		CBX7	23492_HUMAN	GEN	
	locus , and that their binding appears	to be interdependent . To our knowledge ,	Y	CBX7	23492_HUMAN	GEN	-
	this is the first indication that PcG-me	diated repression requires cooperation	*	CBX7	23492_HUMAN	GEN	Ŧ

References

[1] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Michael Hess, and Martin Romacker, An Environment for Relation Mining

Another view of ODIN

Funding

OntoGene is partially supported by the **Swiss National Science Foundation** (grants 100014 - 118396/1 and $105315_{130558/1}$). Additional support is provided by NI-TAS/TMS, Text Mining Services, **Novartis Pharma AG**, Basel, Switzerland.

over Richly Annotated Corpora: the case of GENIA, BMC Bioinformatics, 7(Suppl 3):S3, 2006.

- [2] Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon, Ontogene in Biocreative II, Genome Biology, 9:S13, 2008.
- [3] Gerold Schneider, Kaarel Kaljurand, Thomas Kappeler, Fabio Rinaldi. Detecting protein-protein interactions in biomedical texts using a parser and linguistic resources. CICLING 2009.
- [4] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Simon Clematide, Therese Vachon, Martin Romacker, "OntoGene in BioCreative II.5," IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7(3), pp. 472-480, 2010.
- [5] Fabio Rinaldi, Gerold Schneider, Simon Clematide, Silvan Jegen, Pierre Parisot, Martin Romacker and Therese Vachon. OntoGene (Team 65): preliminary analysis of participation in BioCreative III. BioCreative III workshop, Bethesda, Maryland, September 13-15, 2010.



University of Zurich Institute of Computational Linguistics



Fonds national suisse Schweizerischer Nationalfonds Fondo nazionale svizzero Swiss National Science Foundation



Contact: Fabio Rinaldi rinaldi@cl.uzh.ch fabio@ontogene.org