



SIXTH FRAMEWORK PROGRAMME

Specific Targeted Research Projects

PRIORITY 1

LIFE SCIENCES, GENOMICS AND BIOTECHNOLOGY FOR HEALTH

Contract no: LSHG-CT-2004-512143

DIAMONDS

Dedicated Integration And Modelling Of Novel Data and prior knowledge to enable Systems biology

EU Deliverable

D4.2

Selection of crucial cell cycle candidate genes and feedback on experimental setup

Dissemination Level : PU

(Public)

Due Date: 1st January 2007

Delivery Date: 15th November 2008

Version 1

Partner responsible: Alfonso Valencia (CNIO)

With help from

Martin Krallinger, Carlos Rodriguez-Penagos, Michael Tress, and Michal Linial (HUJI)



Deliverable 4.2 Selection of crucial cell cycle candidate genes and feedback on experimental setup

Part 1: Text mining inspired approaches

Executive summary:

A support vector machine (SVM) based strategy for generating a ranked selection of cell cycle candidate genes for the model organism *A. thaliana* has been developed. The SVM text classifier was trained on a collection of cell cycle relevant abstracts and non-relevant abstracts and then applied to a literature collection of abstracts and full text articles mentioning *A. thaliana* genes. The tool allows queries such as protein names, identifiers or keyword searches and is available at: <http://zope.bioinfo.cnio.es/aratreg/>

Contributors:

CNIO (Martin Krallinger, Carlos Rodriguez-Penagos, Michael Tress, Alfonso Valencia)

Background:

Genes and gene products may be implicated in the cell cycle process at various levels, ranging from the transcriptional regulation of genes crucial for cell division to entire structural assemblies that control the correct segregation of chromosomes. Experimental and bioinformatics approaches to studying the implication of genes in the cell cycle have a tendency to consider narrow aspects, for example gene expression profiles consider time constraints, and protein interaction experimental techniques and sub-cellular localization studies consider only spatial constraints. These experiments on their own can only provide a partial view of our current understanding of what is a dynamic biological process. The cell cycle requires a coordinated control, not only of the regulatory processes but also of the physical protein assemblies.

Most bioinformatics efforts to detect relevant cell cycle genes rely on information provided by annotation databases that in turn are based on manual literature curation by experts. Manual literature curation is a labor-intensive task, information contained in unstructured text (scientific articles) is transformed into structured database records. This relies extensively on the use of controlled vocabularies as well as human inference^[1]. Biological annotation databases contain only a fraction of the currently available (published) functional gene product characterizations and generally do not provide a straightforward way to trace back biological evidence supporting each annotation.

Recently, efforts have been made to enable a more systematic access to relevant information of genes and proteins hidden in large literature repositories using text mining and information extraction technologies^[2]. These approaches included systems like iHOP, which provides direct links between bio-entities and bio-entity interactions and the corresponding references in abstracts^[3], Mscanner, which classifies PubMed abstracts^[4], and G2D, which prioritizes candidate genes for certain disease types^[5].

In the framework of the DIAMONDS project we have implemented a Support Vector Machine (SVM) based tool for selecting cell cycle relevant *Arabidopsis thaliana* genes, from abstracts and full text passages using a bag of words as basic system features^[6]. The tool provides a comprehensive approach to selecting and ranking genes and gene products relevant to a specific biological process for a given model organism.

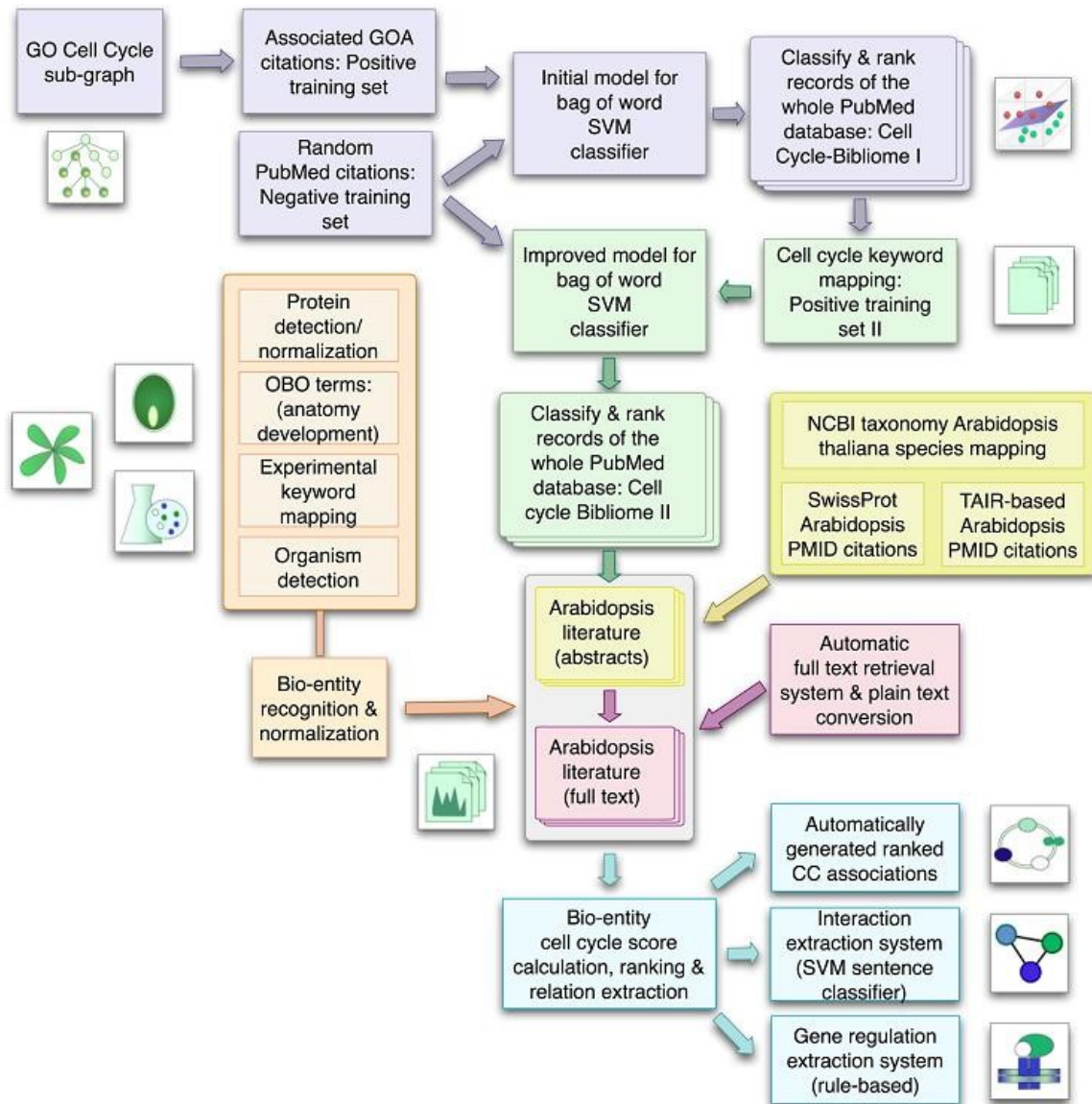
Methods: 206

The prior assumption underlying our approach to the detection cell cycle relevant proteins is the definition of the biological context as the local textual context (within scientific articles). This implies that if a protein is described in articles containing many cell cycle relevant terms it should also show some relationship to this biological process. The textual context is defined here from PubMed abstracts and full text sentence windows (text passages of 5 or 7 sentences).

The system integrates three main modules: (1) the cell cycle text classifier, (2) the bio-entity module and the (3) relation extraction modules. Three cell cycle classifier models were generated, one for PubMed abstracts and two for full text passages (of 5 and 7 sentences). These models were obtained through a supervised learning method, based on support vector machines (SVMs), and trained on a collection of 5000 positive items (cell cycle relevant abstracts) and 5000 negative non-relevant items (random PubMed non-cell cycle relevant abstracts and full text passages).

For the bio-entity normalization, i.e. the linking of text to gene/protein records, a dictionary look-up system was used. The original gene name dictionary extracted from the TAIR database was automatically extended to generate common typographical variants of the gene symbols. Highly ambiguous names were filtered to avoid false positive mappings.

Two types of relations were extracted for the mentioned bio-entities, gene regulation associations were detected using a high precision rule-based system, and protein interaction relations were extracted using a sentence classifier approach based on a training set of manually classified interaction sentences.



Cell cycle gene selection using text mining: The overall system pipeline for the detection of cell cycle relevant genes for the model organism *Arabidopsis thaliana* is illustrated in this figure.

Results: 320

The resulting classifier for PubMed abstracts used a set of 19,427 words as features, and obtained a performance of 88.20 for precision and 89.12 for recall using a radial basis kernel function and leave-one-out cross validation.

Each of the 16,436 *Arabidopsis*-related PubMed abstracts was classified according to their cell division relevance and the mean cell division score for each of the mentioned *Arabidopsis* genes was calculated based on their associated abstracts. This resulted in a total of 138 cell cycle relevant proteins out of which 33 showed a very strong association to the cell cycle. An extended set of proteins, including also those that were associated to at least one high scoring cell cycle abstract consisted of 203 proteins.

Using an in-house PDF retrieval and plain text conversion system, an additional

collection of 7,608 full text articles was assembled. The full text passage classifier models were applied to classify and score each of the *Arabidopsis* full text sentence passages using a sliding window approach, resulting in a collection of cell cycle-scored windows of 2,987,342 (5 sentences) and 2,971,840 (7 sentences) passages.

Also *Arabidopsis* gene and protein mentions were associated to full text sentences, obtaining a total of 420,511 protein to sentence mappings.

A

TAIR-Id	CC-SVM sum	CC-SVM mean	CC-SVM pos. sum	Nr. abstracts	Mentions	PA	GR	KW	EV
AT1G272390	4.482702	4.482702	4.482702	1					
AT1G272320	3.3084781	3.3084781	3.3084781	1					
AT2G472980	2.8264443	2.8264443	2.8264443	1					
AT3G270000	2.7108122	2.7108122	2.7108122	1					
AT3G521115	2.6906195	2.6906195	2.6906195	1					
AT3G061500	5.1173723	2.55868615	5.1173723	2					
AT3G604600	4.8360692	2.4180346	4.8360692	2					
AT3G115200	2.1334821	2.1334821	2.1334821	1					
AT2G267600	2.1334821	2.1334821	2.1334821	1					
AT3G047400	1.9907867	1.9907867	1.9907867	1					
AT5G054900	11.57761703	1.92960283833	12.3889299	6					

B

Protein-Id	Protein name	Term-Id	Term/concept	PMID	CC Score
AT5G054900	SYN1 # syn1	GO:0051323	metaphase	10072401	2.88435
AT5G054900	SYN1 # syn1	GO:0005695	chromatid	10072401	2.88435
AT5G054900	SYN1 # syn1	GO:0007062	sister chromatid cohesion	10072401	2.88435
AT5G054900	SYN1 # syn1	GO:0007126	meiosis	10072401	2.88435
AT5G054900	SYN1 # syn1	GO:0030261	chromosome condensation	10072401	2.88435
AT5G054900	SYN1 # syn1	GO:0051323	metaphase	12783989	2.82644
AT5G054900	SYN1 # syn1	GO:0005695	chromatid	12783989	2.82644
AT5G054900	SYN1 # syn1	GO:0007067	mitosis	12783989	2.82644
AT5G054900	SYN1 # syn1	GO:0051322	anaphase	12783989	2.82644
AT5G054900	SYN1 # syn1	GO:0005711	meiotic chromosome	12783989	2.82644
AT5G054900	SYN1 # syn1	GO:0051325	interphase	12783989	2.82644
AT5G054900	SYN1 # syn1	GO:0005698	centromere	12783989	2.82644
AT5G054900	SYN1 # syn1	GO:0007126	meiosis	12783989	2.82644
AT5G054900	DFI1 # dif1	GO:0007067	mitosis	10504568	2.70303
AT5G054900	DFI1 # dif1	GO:0007059	chromosome segregation	10504568	2.70303

C

Protein-Id	Protein name	Exp-Id	Experimental keyword	PMID	CC Score
AT5G054900	SYN1 # syn1	MT:742	hybridization	12783989	2.82644
AT5G054900	SYN1 # syn1	MT:743	in situ	12783989	2.82644
AT5G054900	SYN1 # syn1	MT:807	fluorescence in situ hybridization	12783989	2.82644
AT5G054900	SYN1 # syn1	MT:581	in situ hybridization	12783989	2.82644
AT5G054900	SYN1 # syn1	MT:755	localize	12783989	2.82644
AT5G054900	DFI1 # dif1	MT:856	homology	10504568	2.70303
AT5G054900	DFI1 # dif1	MT:264	mutagenesis	10504568	2.70303
AT5G054900	SYN1 # syn1	MT:676	immunodetection	16897472	2.1629
AT5G054900	dif1	MT:800	bromodeoxyuridine	12783800	2.12286
AT5G054900	SYN1 # syn1	MT:777	fractionation	15972315	2.08132
AT5G054900	SYN1 # syn1	MT:610	cell fractionation	15972315	2.08132
AT5G054900	SYN1 # syn1	MT:618	co-localize	15972315	2.08132
AT5G054900	SYN1 # syn1	MT:755	localize	15972315	2.08132
AT5G054900	syn1	MT:149	fluorescence microscopy	9161029	1.85224
AT5G054900	syn1	MT:255	microscopy	9161029	1.85224
AT5G054900	SYN1	MT:756	rnai	16582011	0.348051

Arabidopsis cell cycle genes: A. Part of the online summary table of mean cell cycle score sorted *Arabidopsis* genes. B. Co-mentioned gene ontology terms are shown for each of the proteins, sorted by their cell cycle context, as well as C. co-occurring experimental techniques

Co-mentioned Gene Ontology terms and plant anatomy and development terms were labeled in this context. Finally experimental keywords have been automatically tagged to account for experimental information described in the literature.

Conclusions:

A text mining and information extraction system for selecting cell cycle relevant Arabidopsis genes and documents has been implemented within the framework of the DIAMONDS project. The text mining application will also be integrated with the Arabidopsis cell cycle gene database that has been developed as part of work package 3.

The extraction system is available at: <http://zope.bioinfo.cnio.es/aratreg/>

Perspectives:

This text mining system allows not only retrieval of topic specific articles (cell cycle documents), but also the selection of a subset of cell cycle relevant proteins together with experimental keywords. The present system can serve as the base both for more efficient topic specific information retrieval as well as to select and generate a reference set of cell cycle relevant proteins. Variations of the system are being developed for other areas of application such as biodegradation pathways, spindle body proteins and cancer mutations.

Part 2: Classification based approaches

Cell cycle - Recovering Orphan Sequences

As cell cycle (CC) is one of the most critical process in any eukaryotic cell and it is under a tight control, it is feasible view that there will be a great deal of conservation in multi cellular organisms (Human, plants) and many of them will be also conserved in function and structure in unicellular organism as the yeast.

Indeed, CC related proteins composed of very diverse sequences. Surprisingly there are numerous sequences that were associated with CC that can be considered orphans due to the minimal number of sequence similarity. These sequences may results of:

- (i) mistake in the annotation as CC due to fault inference
- (ii) lack of sequence similarity but still structural similarity is evident
- (iii) newly evolved gene that is specialized and thus, it is not detected throughout phylogenetic tree.
- (iv) evidence for CC related function is fully dependent of experimental evidence and minimal information can be retrieved from computational and prediction tools.

We investigated these options and confirmed that data-mining and additional experimental support from multiple resources is fundamental for recovery of many of the 'orphan' sequences. For most of them (85%), strong functional relevance to CC is confirmed.

To test our hypothesis we focused on the human proteome as a test case and activated the following protocol for searching the classification performance:

1. Identify the "Cell Cycle" resource classification knowledge.
2. Select only human representatives
3. Classify them by any of the public ally available classification systems
4. Identify 'orphans' (not belong to any known classification, with minimal sequence similarity).
5. Use Clustering by ProtoNet 5.1 for inference and for family assignment.
6. In case of failure, expand the search for other resources (with a minimal use of data mining)
7. Provide a final evaluation of the CC functional relevance.

In the following paragraph we illustrate the success of 'orphan recovery protocol' for human CC proteome. We detected all together 668 proteins that are CC related (following filtration of all 'fragments'). There are 44 proteins with fewer than 50 homologues (at a threshold of Blast e-score of $<e^{-5}$, searching over 5 millions sequences). Among them 15 sequences showed no homologues (<50) even with the most relaxed Blast e-score <10). A survey of these 15 sequences is shown. All these sequences can be considered Orphans. No assignment is available by InterPro or PIRSF, the two very prominent family classification resources that cover a large fraction of all known sequences.

Protein ID	Protein Name	aa	E<=-5	E<=10	Functiona l Family ^a
1. MORN_HUMAN	Morphogenetic neuropeptide	11	0	0	-
2. DEC1_HUMAN	Deleted in esophageal cancer 1	70	1	1	Seq
3. LEU1_HUMAN	Leukemia-associated protein 1	78	3	4	Integr
4. LEU2_HUMAN	Leukemia-associated protein 2	84	3	6	Integr
5. ST20_HUMAN	Suppressor of tumorigenicity protein 20	79	10	13	-
6. RPRM_HUMAN	Protein reprimo	109	16	20	Interg
7. CDC26_HUMAN	Anaphase-promoting complex subunit CDC26	85	9	28	Seq+Integr
8. APC13_HUMAN	Anaphase-promoting complex subunit 13	74	11	28	Seq
9. TUSC2_HUMAN	Tumor suppressor candidate 2	110	15	30	Seq
10. ZWILC_HUMAN	Protein zwilch homolog	591	14	36	Integr
11. AVPI1_HUMAN	Arginine vasopressin-induced protein 1	147	11	41	Seq
12. SPDYA_HUMAN	Speedy protein A	313	35	42	Seq
13. NSL1_HUMAN	Kinetochore-associated protein NSL1 homolog	281	13	43	Integr
14. GML_HUMAN	Glycosyl-phosphatidylinositol-anchored molecule-like protein precursor	158	7	46	Seq+Integr
15. SPDYC_HUMAN	Speedy protein C	293	34	46	Seq+Integr

Table 1: Orphan human proteins related to Cell cycle for which no homologues are known (<50 homologues in Balst e-score of <10 searching 4 million sequences of UniProtKB).

^aSeq, information gained by ProtoNet and EVEREST resources Integr - information gained by expert experimental and computational data from protein-protein interactions, from disease models and other resources

Mining for these 15 proteins (relaxed threshold of Blast e-score $e < 10$) revealed the following biological insights (see Table 1):

1. Additional sequence and structure family assignment methods (SMART, CATH etc) failed to assign family definitions. InterPro, Pfam or PIRSF have not addressed these sequences.
2. Activating search using ProtoNet 5.1 beta and EVEREST 2.1 revealed potential interesting relationship to other sequences and raised hypotheses based on connectivity to functional informative families. We identified additional valuable information for third of the instances. All revealed hidden classification and additional information.
3. For 35% of the proteins extra information was based on sequence alone (Table 1, Seq) and on 65% the integration of data from parallel resources became essential (Table 1, Integr) for gaining novel biological insights.

Several examples illustrate the potential of global clustering protocols:

DEC1: DEC1_HUMAN a very short 70 amino acids (aa) protein with no Blast similarity. We identified 30 family members that are all enriched with short proteins (average of 168 aa). The family consists of several Archaeal strongly conserved proteins whose genes are associated with CRISPRs (Clustered, Regularly Interspaced Short Palindromic Repeats). The function of these proteins has not been experimentally determined, but computational analysis has suggested that they may function as nucleases in DNA repair, similar to RecB. The similarity to endonuclease and to the repair system

suggests that the archeal-eukaryotic connection that was not revealed by other methods. No structural support could be validated.

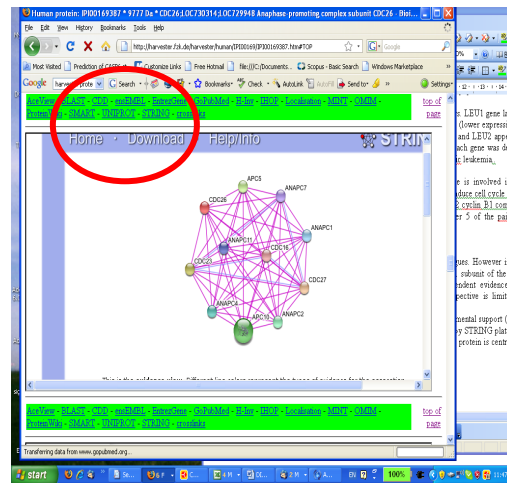
LEU1 and LEU2. These are 2 short proteins (~80 aa) that are named as “Deleted in lymphocytic leukemia 1”. The immediate sequence based information is minimal. However integration of data from multiple resources became informative. LEU1 binds HTATIP which is the Catalytic subunit of the histone acetyltransferase complex which is involved in transcriptional activation. Specifically, with acetylation of nucleosomal histone H4 /H2A, this modification may (i) alter nucleosome-DNA interactions and (ii) promote interaction of the modified histones with other proteins which positively regulate transcription. This complex may be required for the activation of transcriptional programs associated with oncogene and proto-oncogene mediated growth induction, tumor suppressor mediated growth arrest and replicative senescence, apoptosis, and DNA repair.

This interaction suggests that these proteins act as tumor suppressors. LEU1 gene lacks a TATA box. It is strongly expressed in testis, thymus, and intestine (lower expression in prostate, spleen, peripheral blood lymphocytes, and ovary). LEU1 and LEU2 appear as pairs that are transcribed in opposite directions. The first exon of each gene was deleted in all cases of chromosomal 13q14 loss in B-cell chronic lymphocytic leukemia..

RMRM: based on similarity it was suggested that the sequence is involved in the regulation of p53-dependent G2 arrest of the cell cycle. It seems to induce cell cycle arrest by inhibiting CDC2 activity and nuclear translocation of the CDC2 cyclin B1 complex. Sequence and structural based became non informative. However 5 of the pairwise interacting partners are main cell-cycle partners.

CDC26 and APC13 are very short proteins with a minimal number of homologues. For CDC26 and APC13, taxonomical support can be suggested. CDC26 is supported by 15 different taxonomical groups there is a strong evidence for relatedness by sequence alone. Surprisingly, the relatedness from taxonomical perspective is limited to primate, rodents and few but not all insects. For APC13 there are 35 taxonomical derived elements of support, despite a lack of additional experimental evidence.

CDC26 has the strongest evidence for APC related complex is based on experimental support (at the protein level, see Figure). Using the connectivity map (as presented by STRING platform) we gained evidence for the APC complex relatedness ensuring this protein is central for CC arrest and CC control. Co occurrence on the gene expression level is an additional support (blue edges, Figure). CDC26 is one of the nine or so subunits identified within APC but its exact function is not known (based on Pfam). For the APC13, minimal experiments available but clearly,



as this gene is highly expressed (with over 120 independent evidences for expression) its presence in CC related phenomena is unquestionable.

Similar analysis was completed for the other sequences in Table 1. The assessment of 13/15 to CC was validated and the role of advances sequence based classification became evident by the application of ProtoNet, EVEREST, PANDORA and additional integration based tools.

Interestingly despite minimal sequence similarity for the Orphan proteins, for some, direct evidence on phosphorylation as well as other post translational modifications (PTM) ensures their functional role in the control of CC sequence of events. For example, TUSC2_HUMAN is based on a myristoylation that is required for tumor suppressor activity.

References

- ¹ Camon EB, Barrell DG, Dimmer EC, Lee V, Magrane M, Maslen J, Binns D, Apweiler R. 2005. An evaluation of GO annotation retrieval for BioCreAtlvE and GOA. BMC Bioinformatics 6 Suppl 1:S17.
- ² Krallinger M, Valencia A, Hirschman L. 2008. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biology 2008, 9:S8 [accepted]
- ³ Fernández JM, Hoffmann R, Valencia A. 2007. iHOP web services. Nucleic Acids Res. W21-26
- ⁴ Poulter GL, Rubin DL, Altman RB, Seoighe C. 2008. MScanner: a classifier for retrieving Medline citations. BMC Bioinformatics, 9:108.
- ⁵ Perez-Iratxeta C, Wjst M, Bork P, Andrade MA. 2005. G2D: a tool for mining genes associated with disease. BMC Genet., 6:45.
- ⁶ Krallinger M, Rojas AM, Valencia A. 2008. Creating reference datasets for Systems Biology applications using text mining. Annals of the New York Academy of Sciences [accepted]