Integrating Large-Scale Text Mining and Co-Expression Networks: Targeting NADP(H) Metabolism in E. coli with Event Extraction

Suwisa Kaewphan*, **Sanna Kreula**†, **Sofie Van Landeghem**‡,
Yves Van de Peer‡, **Patrik R. Jones**†, **Filip Ginter***

*Department of Information Technology, University of Turku
Joukahaisenkatu 3-5B, 20520 Turku, Finland
sukaew,figint@utu.fi

†Bioenergy group, University of Turku
Tykistökatu 6A, 6krs, 20520 Turku, Finland
sanmpe,patjon@utu.fi

‡Department of Plant Systems Biology, VIB,
Department of Plant Biotechnology and Bioinformatics, Ghent University
Technologiepark 927, 9052 Gent, Belgium
yves.vandepeer,sofie.vanlandeghem@psb.vib-ugent.be

Abstract

We present an application of EVEX, a literature-scale event extraction resource, in the concrete biological use case of NADP(H) metabolism regulation in *Escherichia coli*. We make extensive use of the EVEX event generalization based on gene family definitions in Ensembl Genomes, to extract cross-species candidate regulators. We manually evaluate the resulting network so as to only preserve correct events and facilitate its integration with microarray-based co-expression data. When analysing the combined network obtained from text mining and co-expression, we identify 41 candidate genes involved in triangular patterns involving both subnetworks. Several of these candidates are of particular interest, and we discuss their biological relevance further. This study is the first to present a real-world evaluation of the EVEX resource in particular and literature-scale application of the systems emerging from the BioNLP Shared Task series in general. We summarize the lessons learned from this use case in order to focus future development of EVEX and similar literature-scale resources.

Keywords: event extraction, EVEX, NADP(H), co-expression

1. Introduction

The field of natural language processing in the biomedical domain (BioNLP) aims at supporting life science research in dealing with the mass of available scientific literature in an efficient manner. Typical use cases for BioNLP include support for biological database curation, efficient retrieval of articles relevant to a particular biomedical molecule or process of interest, linking experimental data with available literature, and various other tasks which require aggregation of knowledge from a large number of scientific articles.

Among the main directions currently pursued within the BioNLP community is *event extraction*. This task involves the identification of biologically relevant events in scientific literature, covering both physical events involving genes and proteins as well as recursively defined regulatory events. Event arguments can have various semantic roles such as *cause* (effector) and *theme* (effectee). Event extraction was popularized through the BioNLP'09 and '11 Shared Tasks on Event Extraction (Kim et al., 2009; Kim et al., 2011), which allowed for a community-wide evaluation of numerous approaches to event extraction in a tightly controlled setting. The main advantage of the event representation is the relatively general and easily extensible definition of the task, as well as the level of detail provided for subsequent applications. An example of an event as defined in the Shared Tasks, is shown in Figure 1, illustrating *event nesting*, a crucial property of the event representation as well as the ability of events to abstract from the variation in natural language whereby a single event may represent a number of textually diverse statements. Additional details on event representation are given in the review of Ananiadou et al. (2010).

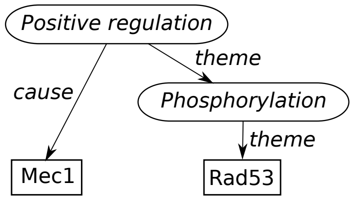

Figure 1: The event representation of the statements "*phosphorylation of Rad53 is controlled by Mec1*" (PMID:9315648), "*Mec1-dependent Rad53 phosphorylation*" (PMID:10449414), and "*. . . adaptor that enables Rad53 phosphorylation by Mec1*" (PMID:16085488).

As a follow-up to the BioNLP'09 Shared Task, the winning Turku Event Extraction System (TEES) (Björne et al., 2009) was applied to all PubMed abstracts, and the resulting set of 19 million extracted events was made publicly available for further research (Björne et al., 2010). This dataset was subsequently extended with event generalizations based on gene families and released as a relational database and web application[1] under the name *EVEX* by Van Landeghem et al. (2011; 2012).

The EVEX dataset addresses a fundamental shortcoming of the original event set: event extraction as defined in the Shared Tasks is purely text-based, i.e., the event extraction systems are not required to assign any biologically relevant identity (such as an Entrez Gene identifier) to the genes and proteins participating in the events. It is thus not possible to directly correlate the extracted events with other biological data, due to the numerous well-known issues caused by gene/protein name ambiguity (Chen et al., 2005). The EVEX dataset resolves this issue by assigning gene/protein mentions to their respective gene families; groups of homologous genes sharing sequence similarity. Gene families are retrieved from the publicly available resource Ensembl Genomes (Kersey et al., 2010) and every gene/protein mention in EVEX is assigned to at most one family. EVEX can thus define events with gene families as their arguments, rather than individual gene/protein mentions identified merely as character strings. Such events defined on top of entire families are referred to as *generalized*. For instance, the resulting family generalization of the event depicted in Figure 1 would have as its arguments the families *ATR* and *Rad53* with homologs in ca. 20 vertebrates, including human, and mouse.

The main advantage of the family generalizations is the fact that they rather straightforwardly support homology-based predictions, as sequence similarity often implies functional similarity. For example, if EVEX contains several regulatory events between pairs of genes that belong to families F_1 and F_2, this may be taken as supporting the hypothesis that other gene pairs belonging to F_1 and F_2 may exhibit a similar regulation pattern. Or, taken from a different perspective, given a pair of genes/proteins that are, based on experimental data, hypothesized to be involved in a regulation, the EVEX event generalizations can be used to straightforwardly access events among not only the given pair of genes/proteins, but also among their homologs.

In this study, we apply the EVEX dataset, as a literature-scale text mining resource, to the concrete BioNLP use case of identifying candidate regulators of NADP(H) metabolism in *Escherichia coli*. The purpose of this study is two-fold: First, we demonstrate the application of literature-scale event extraction to hypothesis generation in a real-life setting, driven by ongoing biological research on a specific molecule. Second, we aim at evaluating EVEX, and to some extent event extraction in general, to gain insight into its suitability for such hypothesis generation and to identify problem areas warranting future research.

2. Biological Motivation and Problem Setting

NADP(H) is a ubiquitous molecule that has a global role and the regulation of its metabolism is regarded as an ideal case-study in the well-studied model organism *E. coli*. NADP(H) is oxidized in more than 100 reactions while only three reactions contribute to the reduction of NADP$^+$, catalyzed by Zwf (Gdh), PntAB and Icd. The intracellular ratio of NADP(H) (reduced) to NADP$^+$ (oxidized) is tightly regulated under "normal" conditions (metabolic homeostasis) but is able to respond rapidly to changes in the intracellular environment, e.g. in the presence of reactive oxygen species (Ralser et al., 2007). NADP(H)-homeostasis is otherwise maintained at the border of thermodynamic limitations for whole cell-metabolism (Henry et al., 2007), with consequences for biotechnological applications (Walton and Stewart, 2004).

Whilst the regulation of the dynamic response is relatively well-established (*soxRS* regulon), there is currently no understanding of how NADP(H)-homeostasis is regulated in the absence of oxidative stress (Krapp et al., 2011). This study aims at identifying candidate regulators and other genes directly relevant to NADP(H)-homeostasis.

In a first step, genes that are known to influence NADP(H)-metabolism (typically enzymes or global regulators) were used to construct an initial list of *key genes* (KGs). This list was extended with *soxS/soxR* and *rob/marA*, well-studied genes that play a major role in the regulation of superoxide defense systems, as they are also known to influence the dynamic NADP(H)-response that is mediated by Zwf (Blanchard et al., 2007). Additional key genes were collected from EcoCYC (Keseler et al., 2011) and STRING databases (Jensen et al., 2009), leading to a final list of 14 key genes relevant to NADP(H)-metabolism that constitute the starting point for the text mining part of this study.

3. Related Work

The challenge of retrieving upstream regulators for any of the 14 key genes can be tackled by either querying *E. coli* specific knowledge bases, or by analyzing available literature.

3.1. *E. coli* Resources

PortEco (formerly EcoliHub) is a resource for laboratory strains of *E. coli*, providing a comprehensive summary on a queried gene by integrating data from EcoCYC (Keseler et al., 2011), EcoGene (Rudd, 2000), STRING (Jensen et al., 2009) and EcoliWiki (McIntosh et al., 2011). EcID (Andres Leon et al., 2009) further contains interactions extracted from KEGG (Kanehisa and Goto, 2000), MINT (Zanzoni et al., 2002) and IntAct (Hermjakob et al., 2004).

While these data sources provide valuable information on specific genes, the retrieved summaries sometimes lack pointers to experimental evidence, or merely link to full-text articles, preventing a quick manual validation of the results. Furthermore, the exponential growth of available experimental data in the life sciences prevents these resources from being fully up-to-date. Finally, organism-specific resources often exclude the retrieval of homology-based predictions. For these reasons, our aim was to track down candidate KG-regulators specifically from literature statements.

[1] http://www.evexdb.org

Search	**PubMed**		**Textpresso**	
	E. coli	**Any org.**	**EcoliWiki**	**EcoCyc**
NADPH	3,796	57,357	1,275	2,176
arcA	279	933	444	1,054
fnr	626	1,342	757	1,722
fruR	55	77	126	232
icd	50	16,005	132	388
marA	181	1,072	281	1,251
marR	139	1,905	310	1,213
pgi	103	2,069	165	414
pntA	8	13	25	75
pntB	8	11	25	37
rob	93	1,967	238	428
soxR	168	207	256	822
soxS	240	279	298	623
sthA	4	28	5	17
zwf	71	144	354	
Any KG	1,545	25,498	-	-
All articles	289,684	21,000,000	24,000	30,000

Table 1: Number of hits when searching for NADP(H) or the key genes in PubMed (with or without restricting the search to *E. coli*) or Textpresso (as implemented by EcoliWiki and EcoCyc).

3.2. Literature Search

Table 1 enumerates the number of articles retrieved from PubMed (Wheeler et al., 2007) when searching for one of the key genes in either *E. coli* or any organism. Further, it presents the results of querying the indexing framework Textpresso (Müller et al., 2004), as implemented by EcoliWiki or EcoCyc.

The large number of citations relevant to the key genes illustrates the necessity of fully automated text mining algorithms to manage the data abundance in the life sciences. For this purpose, many resources have previously been developed. For instance, iHOP allows fast retrieval of various relevant sentences for a certain gene, highlighting gene symbols, organism mentions and MeSH terms found within the same sentence (Hoffmann and Valencia, 2004). EBIMed covers Gene Ontology terms such as biological processes, as well as drugs and species names (Rebholz-Schuhmann et al., 2007).

The STRING database is a widely used resource containing protein-protein interactions predicted from text, amongst other resources (Jensen et al., 2009). The textual evidence is based on co-occurrence methods. PIE *the search* also searches PubMed for protein interaction data, using a classifier relying on word and syntactic features of whole articles (Kim et al., 2012).

For the case-study described in Section 2, we aim at retrieving more complex event structures, including various physical event types and regulatory events (cf. Figure 1). While the Medie search engine (Ohta et al., 2006) supports similar advanced queries, we focus specifically on the recently released EVEX resource (Van Landeghem et al., 2011), because its unique event family generalizations allow cross-species hypothesis generation, expanding the search domain also to homologs of the 14 key genes.

4. Methods and Resources

4.1. The EVEX Dataset

In this work, we use an extended version of the EVEX dataset, containing gene normalizations provided by the GenNorm system of Wei et al. (2011). The task of gene normalization is to disambiguate the gene and protein mentions in text to the biological object they represent, in our case by assigning them with a unique Entrez Gene identifier. The GenNorm system represents the state-of-the-art in gene normalization, having achieved first rank by several evaluation criteria in the BioCreative III Challenge (Lu et al., 2011). We used the Entrez Gene identifiers given by GenNorm to directly assign the gene and protein mentions to their corresponding Ensembl Genomes families. Where GenNorm does not assign an Entrez Gene identifier, the original algorithm of Van Landeghem et al. is used as a fallback.

Further, the dataset used in this study was also extended with events extracted from all full-text articles available in the Open Access subset of PubMed Central, substantially increasing the amount of literature available for text mining. The impact of this extension is separately evaluated in Section 5.1.

4.2. EVEX Event Preprocessing

As illustrated in Figure 1, events may constitute complex structures where an event may have as its argument another, recursively nested event. While these structures properly account for the semantics of the underlying natural language statements, they cannot be directly correlated with the vast majority of existing biological resources which generally take the form of networks of pairwise interactions between genes and proteins.

To this end, we have defined a rule-based procedure to decompose complex events into pairwise directed interactions. (Van Landeghem et al., 2012) This procedure assigns three interaction types: *regulation* (directed), *indirect regulation* (directed), and *binding* (undirected), stemming from the fact that only regulation and binding events may have more than one argument in the event scheme defined by the Shared Tasks and therefore can generate pairwise interactions. In this work, we further merge *regulation* and *indirect regulation* into a single *regulation* type. The result of applying this procedure to the common event structure of one gene regulating the interaction of two other genes, is shown in Figure 2. Since much of this study deals with generalized events, i.e., events defined on top of gene families, the result of the decomposition procedure will correspondingly be pairs of gene families.

The procedure may at first seem to be defeating the purpose of defining and extracting detailed event structures, since, as illustrated in Figure 2, the event representation captures the semantics of the underlying statement more accurately than the extracted pairs. However, it must be noted that in our current application setting, the underlying events are preserved: the pairwise interactions are used to identify events of interest, which are subsequently presented in full detail to the end-user. Therefore, rather than redefining events per se, we are merely defining a layer of simplified, pairwise interactions on top of the events. This layer serves as an interface between the events in EVEX and pairwise biological data, such as microarray co-expression studies as well as other existing and widely used resources.

4.3. EVEX Candidates

To search for novel regulators of the 14 given *key genes* from *Escherichia coli* strain *K-12* substrain *MG1655*, we first determine their families in Ensembl Genomes, resulting in 14 *key families* (KFs). Next, we extract all events from EVEX which involve at least one key family, regardless of its role in the event (cause or theme). The search is performed on the level of family pairs, as described in Section 4.2. We therefore obtain pairs of the following three types: *Binding(X,KF)*, *Regulation(X,KF)* and *Regulation(KF,X)*, where *X* is a *candidate family* of interest. Since the problem setting is specific to the aforementioned substrain of *E. coli*, we discard all events where the candidate family *X* does not contain a gene from our target organism. Subsequently, we manually evaluate all these extracted events, only preserving correctly extracted or otherwise biologically relevant events where both arguments are assigned to their correct family. The results of this manual evaluation are presented in detail in Section 5.1.

The final set of events that were evaluated as fully correct comprised 132 event occurrences of 81 unique Ensembl Genomes generalized events. These events linked 41 unique candidate gene families to 12 of the initial key gene families. For two key gene families, no EVEX events were found. From this final set of events, we constructed an *E. coli*-specific gene network (referred in further text as the EVEX network) by selecting the *E. coli* gene member in each family. The network is shown in Figure 3.

4.4. Microarray Data

Microarray data was collected from the Affymetrix chip [Ecoli_Asv2] (Affymetrix *E. coli* Antisense Genome Array). Specific microarray data were selected based on their expected relevance for NADP(H)-metabolism in *E. coli*. In order to ensure consistency across all treatments, all data was extracted from only two extensive series of microarray analyzes carried out by Covert et al. (2004), focusing on oxidative stress, and Dong et al. (2008), focusing on the global regulator RpoS. The transcriptome data were extracted from Gene Expression Omnibus (GEO) with accession number GPL199 (Barrett et al., 2011). The microarray analysis platform Chipster was used to generate normalized expression values and p-values. Networks were constructed, analyzed and visualized with the freely available software Cytoscape (Shannon et al., 2003). Plug-in ExpressionCorrelation[2] was employed to construct networks using expression and significance values. A similarity network strength threshold of 0.65 was selected to calculate a similarity matrix using the Pearson correlation coefficient.

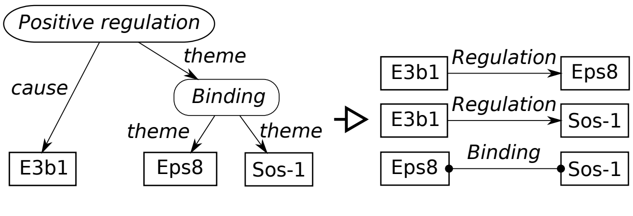

Figure 2: Pairwise decomposition of an event with recursive nesting from the statement "*E3b1 (. . .) plays a critical role (. . .) by facilitating the interaction of Eps8 with Sos-1*" (PMID:15178460).

The co-expression based gene network (referred to as CoEx) thus obtained was then overlaid with the EVEX network, as shown in Figure 3.

5. Results and Discussion

In the following section, we analyze and discuss the results from two perspectives: evaluation of the text mining methods and resources used, and the biological relevance of the findings.

5.1. Text Mining Findings

The set of initial candidate events extracted from EVEX comprises of 348 unique generalized events aggregated from 461 individual event occurrences in text. Each of these events, by definition, has at least one key family as an argument. In total, these events involve 152 unique families. In the following, we manually evaluate the 461 candidate event occurrences based on two criteria: correctness of the extracted event, i.e., whether the event reflects the statement from which it was extracted and, as a second criterion, the correctness of the assignment of the gene and protein mentions to their respective families. This second criterion is particularly crucial when using the events for family-based hypothesis generation.

There were 243 (53%) correctly extracted event occurrences comprising 169 unique generalized events, well in line with the precision figures reported for the Turku Event Extraction System in the official BioNLP'09 Shared Task evaluation for multiple-argument events (50% for bindings and 46% for regulations) (Kim et al., 2009). In addition to genuine false positives, we found among the remaining 218 events two classes of events which, although considered false positives from the strict event definition point of view, were judged biologically relevant and were thus considered for further evaluation. First, these include 36 relevant events extracted with incorrect type, either through label substitution of regulation vs. binding, or constituting a relationship which does not have an appropriate type defined in the event representation. Secondly, three events were found encoding regulation in the opposite direction (i.e. rather than the correct *Regulation(X,KF)*, the false positive *Regulation(KF,X)* is extracted).

Since we are interested in events which recover upstream regulators and binding partners of the key families, we disregard from further evaluation all events of the type *Regulation(KF,X)* (naturally, still preserving *Regulation(KF,KF)*). The remaining 183 events (representing 118 unique generalized events and 76 unique families), comprising both true positives and corrected relevant false positives, were evaluated for the correctness of gene family assignment. Of these, 132 (72%) events were such that both arguments were resolved to their correct Ensembl Genomes family in EVEX. These fully-correct occurrences constitute 81

[2] http://chianti.ucsd.edu

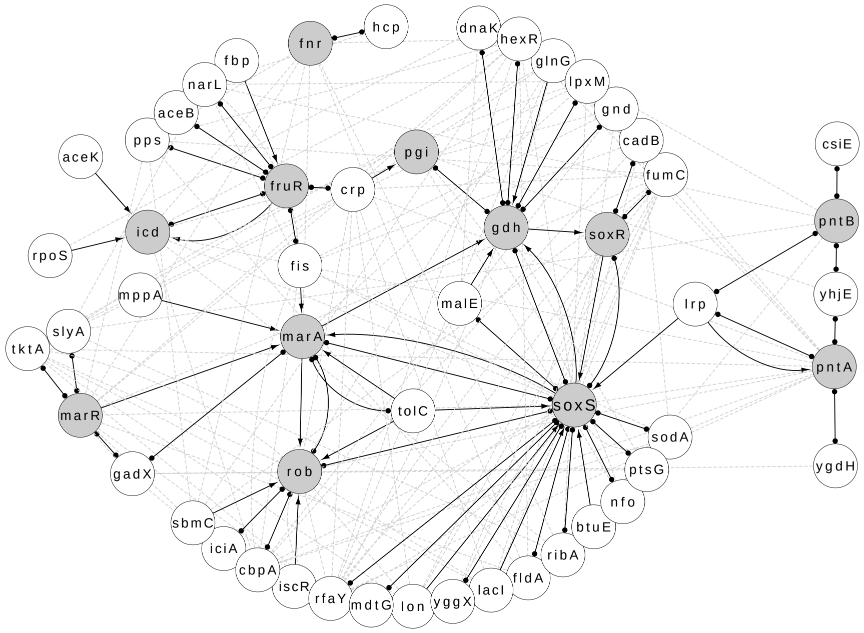

Figure 3: The complete network obtained from EVEX (solid lines) and microarray-based co-expression analysis (dashed lines). In the EVEX network, circle-terminated connections indicate binding and arrows indicate regulation. The key genes are highlighted in gray; 2 key genes are not present since no EVEX events were extracted for them. Note that only events involving at least one key gene are extracted, therefore no events between candidate genes are present.

unique generalized events and 53 unique families (12 key families and 41 candidate families).

To summarize, the precision of the two key components is 53% for event extraction and 72% for gene family assignment of both arguments. However, since the errors are cumulative, the overall precision of even state-of-the-art systems leaves room for improvement. In addition, it is also important to note that a number of false positives among the events do bear biological significance and were deemed relevant for the current study. Naturally, this is highly use case specific and should not be interpreted as an attempt to artificially boost the precision figures.

Manually evaluating the initial set of events to construct the EVEX network amounted to a little less than three days of work of one person. Of the two validation steps (event correctness and family assignment), evaluating the correctness of the family assignments was clearly the more labor-intensive one, as it often required careful identification of the species, strain, and sub-strain involved — information rarely present in the abstract. However, in order to be able to rely on the integration of the EVEX and CoEx networks, we consider the manual evaluation step of great importance and not excessively labor-intensive, particularly compared to the effort that would be necessary to build such a network without any text mining support.

Finally, we discuss the issue of event extraction from full-text articles versus abstracts. The need for text mining in full text articles, in addition to PubMed abstracts, is becoming broadly recognized in the BioNLP community. The EVEX dataset, as used in this study, was extended with full-text articles from the PubMed Central Open-Access (PMC-OA) section and we can thus evaluate the impact of full-text mining on this real-world use case. We find that 18 out of the 41 candidate families, i.e. nearly half, were identified only from a body of a full-text article. This figure clearly demonstrates the added value of full-text articles for text mining and, consequently, the importance of opening full-text articles for automated access.

5.2. Biological Findings

In order to analyze the EVEX network relative to the CoEx network, we initially focus on three important patterns in which these two networks can support each other, illustrated in Figure 4.

Direct support for EVEX-identified relationships by the co-expression network (pattern A) was found in two cases: *sodA-soxS* and *soxS-rfaY*. Both were determined to be genuine positives based on detailed experimental evidence including chromatin immunoprecipitation, DNA-binding and co-expression. Further, 49 triangular clusters (24 of type B, 23 of type C, and 2 which can be classified as either B or C) were identified.

On one hand, such triangular relationships may indicate

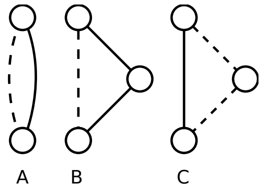

Figure 4: Three patterns of particular interest when referencing the EVEX (solid lines) and CoEx (dashed lines) networks. The two networks may fully support each other (A), the CoEx network may provide further support for an indirect relation from the EVEX network (B), or, finally, the EVEX network may provide further explanation for an indirect relation in the CoEx network (C).

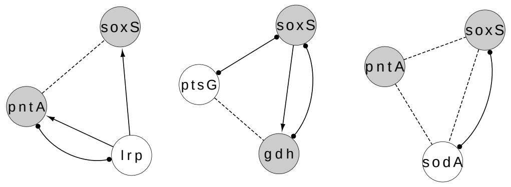

Figure 5: Two type B patterns and one type A/C pattern with varying biological explanations.

that apparent indirect interactions may in fact be direct. One such example is the type B pattern between *gdh*, *soxS*, and the gene encoding for a component of the central glucose-uptake system *ptsG*, shown in Figure 5 (middle). The pattern suggests that a link between *NADP(H)-metabolism* and glucose uptake may exist which warrants further experimental investigation.

On the other hand, a type B pattern may also indicate the very opposite: relationships which, even though appearing direct in one network, are in fact shown to be indirect, in light of the relationships present in the other network. Consider, for instance, the type B pattern observed between *pntA*, *soxS*, and *lrp* in Figure 5 (left). The *pntA–soxS* co-expression (a direct relationship in the CoEx network) is most likely an indirect relationship caused by co-regulation, since both *pntA* and *soxS* are members of the *lrp* regulon. This becomes apparent from the EVEX network, and the actual statements underlying the EVEX events.

An example of a type C pattern (which also contains a type A sub-pattern) is illustrated in Figure 5 (right). One EVEX binding event was identified between the transcriptional regulator *soxS* and *sodA* encoding superoxide dismutase. Both of these two genes are known to respond and contribute to alleviate oxidative stress, whilst *pntA* until now is not known to be involved in such a metabolic manner. A connection between all three genes was identified in CoEx, supporting the EVEX event and also interestingly linking PntAB to dynamic stress conditions which until now it has not been described to be involved in.

These three examples serve to illustrate the diversity of hypotheses obtained from an initial analysis of simple triangular patterns in the combined EVEX/CoEx networks.

The ability to support homology-based function prediction has been presented as one of the primary motivations for the family-based generalization in EVEX. Therefore, candidate genes identified from organisms other than *E. coli* warrant a closer inspection. Of the 41 candidate genes, only five originated entirely from non-*E. coli* studies and further three originated both from *E. coli* and non-*E. coli* literature. In total, about 20% of candidate genes were thus identified through the generalization. This relatively low number is likely due to *E. coli* serving as a model organism in studies of prokaryote central carbon metabolism — the effect of the generalization would likely be more pronounced if the target organism was less studied, or the gene families more coarsely defined, in which case candidates from a related model organism would be identified through the generalization. The five candidates extracted uniquely from non-*E. coli* literature are summarized in Table 2, together with the number of co-expression associations with the key genes. *hexR* is clearly the most “interconnected” candidate and we thus select it for further discussion. *hexR* has not previously been studied in *E. coli* but has been shown to act as a transcriptional regulator of several genes encoding enzymes in central carbon metabolism of *Pseudomonas putida*, including the NADP$^+$-reducing glucose-6-phosphate dehydrogenase (gdh) (Daddaoua et al., 2009). Interestingly, *hexR* is located adjacent to *gdh* in the genome of *E. coli* and they both appear to share the promoter-region with binding-motifs for SoxS, MarA and Rob. There is no direct co-expression between *gdh* and *hexR*, however, triangular type B patterns are observed with both *gnd* and *soxS* (Figure 6). The microarray-analysis also shows co-expression with both *icd* and *pntA*, closely linking *hexR* with all three NADP$^+$-reducing enzymes. *hexR* therefore represents a highly interesting candidate to study further.

EVEX event	# of co-expressed KGs
hexR - gdh	4
glnG - gdh	2
cadB - gdh	2
lpxM - gdh	2
slyA - marR	1

Table 2: The number of key genes co-expressed with the candidate genes identified by EVEX in organisms other than *E. coli*.

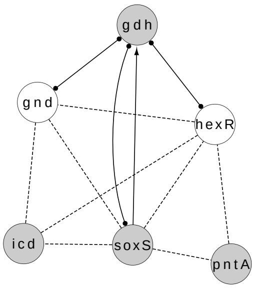

Figure 6: *HexR*-related sub-network in EVEX and in CoEx.

In summary, several new relationships were uncovered by the combined analysis generating several testable and potentially interesting hypotheses, in particular the notion that NADP$^+$-reduction is subject to coordinated regulation by the transcriptional regulators SoxS and HexR (Figure 6). Importantly, even though not previously studied in the target host organism, *hexR* could be identified by EVEX-analysis alone and further supported by triangular relationships involving also CoEx. The relative lack of EVEX-events for the critical transhydrogenases, despite a wealth of edges in CoEx, supports our conclusion that regulatory interactions influencing non-dynamic NADP(H)-homeostasis still remain to be explored in prokaryotes.

6. Conclusions and Future Work

We have demonstrated the application of EVEX, a literature-scale event extraction resource to a real-world biological use case, with an encouraging result. With a reasonable manual effort, we were able to extract a network of candidate genes related to the metabolism of NADP(H) in *E. coli*, starting with 14 key genes, and to integrate the network with microarray-based co-expression data. Integrating the two networks and using them as mutually supporting resources was a crucial step, and we were able to identify several candidate genes of particular interest, warranting further experimental evaluation. This study was only possible because the predictions of the event extraction system could be, via the gene family assignment procedure implemented in EVEX, directly related to available experimental data, focusing specifically on genes from the target organism, or their homologs.

Our evaluation has shown that, even when state-of-the-art event extraction and gene normalization systems are employed, automatically extracted text mining results need further manual validation to enable meaningful integration with experimental data in similar focused use cases. However, we expect that after this initial case study the manual effort involved in the process can be further decreased by developing tools specifically supporting such applications, for instance focusing on the labor-intensive task of gene family assignment evaluation.

Since NADP(H) is a metabolite, and not a gene/protein, it falls out of scope in the majority of BioNLP studies. Metabolites are of great relevance and it is important to focus on incorporating events pertaining to metabolites into the EVEX dataset. This can be supported by the methods developed for the BioNLP'11 Shared Task (Kim et al., 2011) that involved metabolites in the ID sub-task. A further challenge is presented by the fact that metabolites cannot be assigned into a gene family, which is a strict requirement of the current event generalization procedure in EVEX. A more relaxed criterion will therefore need to be implemented so as to account for events among different classes of bio-entities, most importantly between proteins and metabolites, without losing the benefit obtained from the family-based generalization.

Further future work can be charted in several directions. First, the current network can be expanded by extracting and verifying events among the currently identified candidate families, as well as including events directly involving NADP(H). Then, the network can be expanded by binding partners and regulators of the current candidate families, essentially adding a layer of 2^{nd} degree regulators. Since the network is expected to grow substantially and manual evaluation of all 2^{nd} degree regulators may not be feasible, it will be important to investigate external resources as well as internal statistics which can be used to rank the new candidates and focus the exploration of the network to the most promising areas. Finally, since the use case presented in this study is an example of what we expect to be a commonly faced problem, we will consider developing novel tools to support and automatize building the network without requiring extensive data-processing skills.

7. Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Grant Agreement n. [260661] (P.R.J.) and Academy of Finland projects no. 253269 (S.Kr.) and 138796 (S.Ka.). S.VL. and Y.vdP. further acknowledge the support of Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks"). S.VL. also thanks the FWO (research foundation Flanders) for funding her research.

8. References

S. Ananiadou, S. Pyysalo, J. Tsujii, and D.B. Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, 28(7):381–390.

E. Andres Leon, I. Ezkurdia, B. García, A. Valencia, and D. Juan. 2009. EcID. a database for the inference of functional interactions in E. coli. *Nucleic Acids Research*, 37(suppl 1):D629–D635.

T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, R.N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva. 2011. NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic acids research*, 39(suppl 1):D1005.

J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski. 2009. Extracting complex biological events with rich graph-based feature sets. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 10–18, Morristown, NJ, USA. Association for Computational Linguistics.

J. Björne, F. Ginter, S. Pyysalo, J. Tsujii, and T. Salakoski. 2010. Scaling up biomedical event extraction to the entire PubMed. In *Proceedings of the BioNLP 2010 Workshop*, pages 28–36. Association for Computational Linguistics.

J.L. Blanchard, W.Y. Wholey, E.M. Conlon, and P.J. Pomposiello. 2007. Rapid changes in gene expression dynamics in response to superoxide reveal SoxRS-dependent and independent transcriptional networks. *PLoS ONE*, 2(11):e1186, 11.

L. Chen, H. Liu, and C. Friedman. 2005. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21:248–256.

M.W. Covert, E.M. Knight, J.L. Reed, M.J. Herrgard, and B.O. Palsson. 2004. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429(6987):92–96.

A. Daddaoua, T. Krell, and J.L. Ramos. 2009. Regulation of glucose metabolism in Pseudomonas: The phosphorylative branch and Entner-Doudoroff enzymes are regulated by a repressor containing a sugar isomerase domain. *Journal of Biological Chemistry*, 284(32):21360.

T. Dong, M.G. Kirchhof, and H.E. Schellhorn. 2008. Rpos regulation of gene expression during exponential growth of Escherichia coli K12. *Molecular Genetics and Genomics*, 279:267–277. 10.1007/s00438-007-0311-4.

C.S. Henry, L.J. Broadbelt, and V. Hatzimanikatis. 2007. Thermodynamics-based metabolic flux analysis. *Biophysical journal*, 92(5):1792–1805.

H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler. 2004. IntAct: an open source molecular interaction database. *Nucleic Acids Research*, 32(suppl 1):D452–D455.

R. Hoffmann and A. Valencia. 2004. A gene network for navigating the literature. *Nat Genet*, 36(7):664, Jul.

L.J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering. 2009. STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(suppl 1):D412–D416.

M. Kanehisa and S. Goto. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 28:27–30.

P.J. Kersey, D. Lawson, E. Birney, P.S. Derwent, M. Haimel, J. Herrero, S. Keenan, A. Kerhornou, G. Koscielny, A. Kähäri, R. J. Kinsella, E. Kulesha, U. Maheswari, K. Megy, M. Nuhn, G. Proctor, D. Staines, F. Valentin, A.J. Vilella, and A. Yates. 2010. Ensembl Genomes: Extending Ensembl across the taxonomic space. *Nucleic Acids Research*, 38(suppl 1):D563–D569.

I.M. Keseler, J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muñiz Rascado, C. Bonavides-Martinez, S. Paley, M. Krummenacker, T. Altman, P. Kaipa, A. Spaulding, J. Pacheco, M. Latendresse, C. Fulcher, M. Sarker, A.G. Shearer, A. Mackie, I. Paulsen, R.P. Gunsalus, and P.D. Karp. 2011. EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Research*, 39(suppl 1):D583.

J.D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 1–9, Morristown, NJ, USA. Association for Computational Linguistics.

J.D. Kim, S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen, and J. Tsujii. 2011. Overview of BioNLP Shared Task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 1–6. Association for Computational Linguistics.

S. Kim, D. Kwon, S.Y. Shin, and W.J. Wilbur. 2012. PIE the search: searching PubMed literature for protein interaction information. *Bioinformatics*, 28(4):597–598.

A.R. Krapp, M.V. Humbert, and N. Carrillo. 2011. The soxRS response of Escherichia coli can be induced in the absence of oxidative stress and oxygen by modulation of NADPH content. *Microbiology*, 157(4):957.

Z. Lu, H.Y. Kao, C.H. Wei, M. Huang, J. Liu, C.J. Kuo, C.N. Hsu, R.T. Tsai, H.J. Dai, N. Okazaki, H.C. Cho, M. Gerner, I. Solt, S. Agarwal, F. Liu, D. Vishnyakova, P. Ruch, M. Romacker, F. Rinaldi, S. Bhattacharya, P. Srinivasan, H. Liu, M. Torii, S. Matos, D. Campos, K. Verspoor, K.M. Livingston, and W.J. Wilbur. 2011. The gene normalization task in BioCreative III. *BMC Bioinformatics*, 12(Suppl 8):S2+.

B.K. McIntosh, D.P. Renfro, G.S. Knapp, C.R. Lairikyengbam, N.M. Liles, L. Niu, A.M. Supak, A. Venkatraman, A.E. Zweifel, D.A. Siegele, and J.C. Hu. 2011. EcoliWiki: a wiki-based community resource for Escherichia coli. *Nucleic Acids Research*.

H.M. Müller, E.E. Kenny, and P.W. Sternberg. 2004. Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11):e309, 09.

T. Ohta, Y. Miyao, T. Ninomiya, Y. Tsuruoka, A. Yakushiji, K. Masuda, J. Takeuchi, K. Yoshida, T. Hara, J.D. Kim, Y. Tateisi, and J. Tsujii. 2006. An intelligent search engine and GUI-based efficient MEDLINE search tool based on deep syntactic parsing. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 17–20. Association for Computational Linguistics.

M. Ralser, M.M. Wamelink, A. Kowald, B. Gerisch, G. Heeren, E.A. Struys, E. Klipp, C. Jakobs, M. Breitenbach, H. Lehrach, and S. Krobitsch. 2007. Dynamic rerouting of the carbohydrate flux is key to counteracting oxidative stress. *Journal of biology*, 6(4):10.

D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, and P. Stoehr. 2007. EBIMed–text crunching to gather facts for proteins from Medline. *Bioinformatics*, 23(2):e237–e244.

K.E. Rudd. 2000. EcoGene: a genome sequence database for Escherichia coli K-12. *Nucleic Acids Research*, 28(1):60–64.

P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.

S. Van Landeghem, F. Ginter, Y. Van de Peer, and T. Salakoski. 2011. EVEX: a PubMed-scale resource for homology-based generalization of text mining predictions. In *Proceedings of the BioNLP 2011 Workshop*, pages 28–37. Association for Computational Linguistics.

S. Van Landeghem, K. Hakala, S. Rönnqvist, T. Salakoski, Y. Van de Peer, and F. Ginter. 2012. Exploring biomolecular literature with EVEX: Connecting genes through events, homology and indirect associations. *Advances in Bioinformatics*. To appear.

A.Z. Walton and J.D. Stewart. 2004. Understanding and improving NADPH-dependent reactions by nongrowing Escherichia coli cells. *Biotechnology Progress*, 20(2):403–411.

C.H. Wei and H.Y. Kao. 2011. Cross-species gene normalization by species inference. *BMC bioinformatics*, 12(Suppl 8):S5.

D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, L.Y. Geer, Y. Kapustin, O. Khovayko, D. Landsman, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, V. Miller, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R.L. Tatusov, T.A. Tatusova, L. Wagner, and E. Yaschenko. 2007. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 35(suppl 1):D5–D12.

A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni. 2002. MINT: a Molecular INTeraction database. *FEBS Letters*, 513(1):135 – 140.