ConReg: Analysis and Visualization of Conserved Regulatory Networks in Eukaryotes

Robert Pesch[1], Matthias Böck[2], and Ralf Zimmer[1]

1 **Institut für Informatik, Ludwig-Maximilians-Universität München, Amalienstr. 17, 80333 München,** {Robert.Pesch,Ralf.Zimmer}@bio.ifi.lmu.de
2 **Institut für Informatik / I12, Technische Universität München, Boltzmannstr. 3, 85748 Garching,** Matthias.Boeck@in.tum.de

Abstract

Transcription factors (TFs) play a fundamental role in cellular regulation by binding to promoter regions of target genes (TGs) in order to control their gene expression. TF-TG networks are widely used as representations of regulatory mechanisms, e.g. for modeling the cellular response to input signals and perturbations.

As the experimental identification of regulatory interactions is time consuming and expensive, one tries to use knowledge from related species when studying an organism of interest. Here, we present ConReg, an interactive web application to store regulatory relations for various species and to investigate their level of conservation in related species. Currently, ConReg contains data for eight model organisms. The regulatory relations stored in publicly available databases cover only a small fraction both of the actual interactions and also of the regulatory relations described in the scientific literature. Therefore, we included regulatory relations extracted from PubMed and PubMedCentral using sophisticated text-mining approaches and from binding site predictions into ConReg.

We applied ConReg for the investigation of conserved regulatory motifs in *D. melanogaster*. From the 471 regulatory relations in *REDfly* our system was able to identify 66 confirmed conserved regulations in at least one vertebrate model organism (*H. sapiens*, *M. musculus*, *R. norvegicus*, *D. rerio*). The conserved network consists among others of the well studied motifs for eye-development and the pan-bilaterian kernel for heart specification, which are well-known examples for conserved regulatory relations between different organisms.

ConReg is available at http://services.bio.ifi.lmu.de/ConReg/ and can be used to analyze and visualize regulatory networks and their conservation among eight model organisms. It also provides direct links to annotations including literature references to potentially conserved regulatory relations.

1998 ACM Subject Classification J.3 Life and Medical Sciences, H.3.5 Online Information Services, I.2.7 Natural Language Processing

Keywords and phrases web application, evolutionary biology, regulatory networks, text-mining

Digital Object Identifier 10.4230/OASIcs.GCB.2012.69

1 Introduction

The physical regulatory relationships of an organism can be described by gene regulatory networks (GRNs). Transcription factors (TFs) and their respective targets (TGs) define the majority of these regulations. GRNs can describe systems on the scale of a few genes, a particular pathway or even on the whole gene complement of an organism. The inference of these GRNs is generally done from experimental data sets, like gene expression data and additional prior information from available databases [11]. Even though more and more

German Conference on Bioinformatics 2012 (GCB'12).
Editors: S. Böcker, F. Hufsky, K. Scheubert, J. Schleicher, S. Schuster; pp. 69–81

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

high-throughput methods for the identification of TF-TG relations have been developed, data of experimentally validated relations is still sparse for higher multi-cellular organisms [22]. Therefore, transferring knowledge using orthologs from related species is typically done when studying an organism of interest. Several approaches have been proposed which are capable of transferring physical protein-protein interactions even between phylogenetically distant species, e.g. from *S. cerevisiae* to *A. thaliana* or *C. elegans* [38]. The conservation of a regulatory relation requires that at least the involved TF and the TG have to be conserved and that the TF binds to the promoter region of the TG in two or more organisms. Depending on the number of organisms in which the regulatory relation is conserved and the evolutionary distance, relations between different organisms can be transferred with a certain confidence [2, 27, 30].

Different methods have been proposed and used for the transfer of regulatory networks from one organism to another. A well-known example is KEGG, which transfers confirmed regulatory relations to (non-model) organisms based on ortholog definitions [13]. For bacteria more advanced approaches have successfully been applied, which additionally incorporate conserved information of the binding site [2] and subfamily classifications [27]. Similar approaches exist for eukaryotes, which also make use of conserved transcription factor binding sites [30]. Taher *et al.* claimed that 88 % of the orthologs between *H. sapiens* and *D. rerio* retain their regulatory mechanisms [30]. Nevertheless, the extend of regulatory relations that can be directly transferred between organisms remains controversial [2]. There are some well-known regulatory motifs, which appear to be conserved among a group of quite distant species, supporting the transfer of conserved regulatory relations. A famous example is the conservation of regulatory relations for the development of the eye in *D. melanogaster* and vertebrates. It was shown that in *M. musculus* and other vertebrates *Pax-6*, the ortholog of the *eyeless (ey)* gene - one of the central TFs controlling the eye development in *D. melanogaster* - shares an extensive sequence identity and is even capable of inducing ectopic eyes in *D. melanogaster* [36]. Also other motifs, like the pan-bilaterian kernel for heart specification [6] or regulation of apoptosis regulation in *D. melanogaster* and vertebrates [39] appear to be conserved. Studies revealing the similarity and the conservation of regulatory subnetworks have been conducted for different species as well, like *MAP kinase expression* in *C. elegans* and *H. sapiens* [15] or *Toll-like receptor 4* regulated genes [26].

In the following we present ConReg, an interactive web application to investigate regulatory relations. ConReg collects and visualizes evidences for the conservation of regulatory relations in other eukaryotic model organisms. For that purpose, we collected regulatory data for eight model organisms (*H. sapiens*, *M. musculus*, *R. norvegicus*, *D. rerio*, *D. melanogaster*, *C. elegans*, *A. thaliana* and *S. cerevisiae*). The data was obtained from general and species-specific regulatory databases, from text-mining approaches applied to PubMed abstracts and PubMedCentral full text publications and from transcription factor binding site predictions (TFBS).

2 Discovery of Conserved Regulatory Relations with ConReg

With ConReg conserved regulatory relations for a source species can be discovered in a target species if these relations can be found between the respective orthologs in both species (by default we do not require conservation of the transcription factor binding site in the two species). ConReg searches regulatory relations of a source species which were extracted from regulatory databases, in the specified target species. Several types of evidences for regulatory relations of the target species can be considered based on the user's selections.

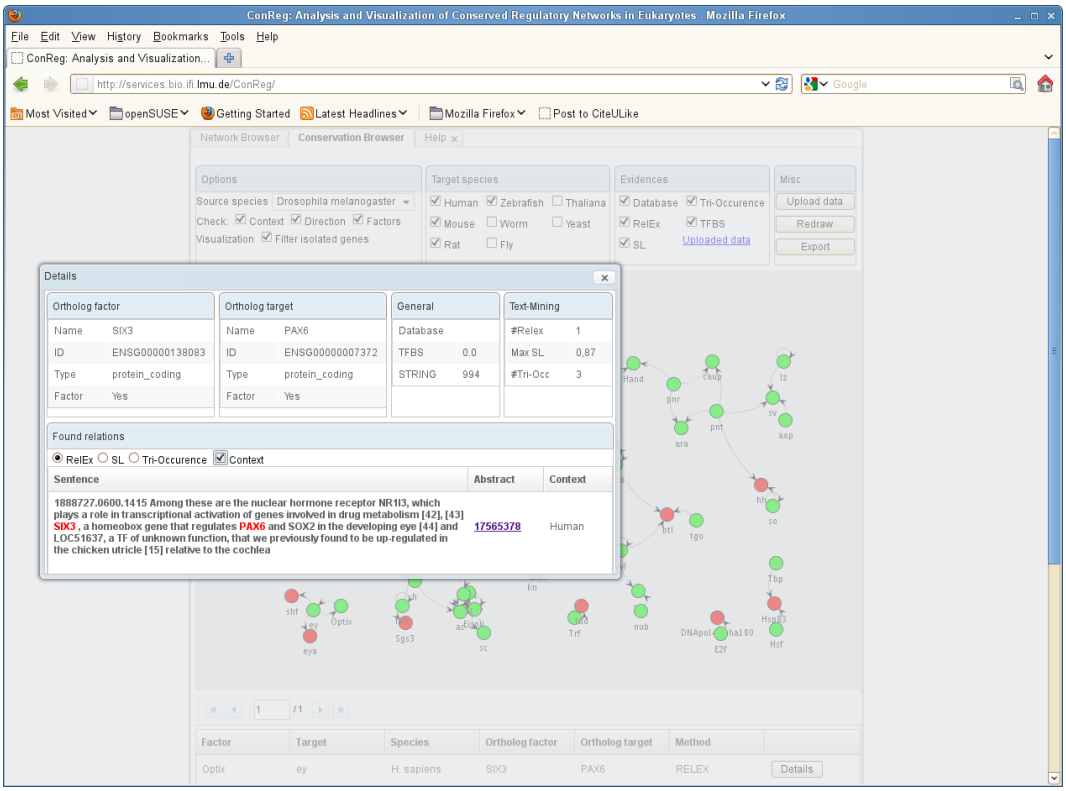

Figure 1 Screenshot of ConReg for the interactive discovery of conserved regulatory relations for a source species in user selected target species. Conservation of regulatory relations for a species can be interactively discovered using ConReg. The system allows searching for conservation in all provided species and with different prediction methods for the target species, whereas for the source species only reliable relations from databases are considered. For an identified conservation of a regulatory relation details such as text-mining results and binding site predictions can be visualized.

Currently, regulatory information from publicly available databases like TRANSFAC [18] or REDfly [10], relations found with text-mining approaches (RelEx [8], SL [9], Tri-occurrence) in PubMed and PubMedCentral and TFBS predictions can be selected (see Materials and Methods for details).

The *Conservation Browser*, where the entire predicted conserved network is shown and the *Motif Finder* with which the user can search for conservations in a defined subset of genes are the two main features of ConReg. For both features, the user can select the source species, regulatory data sources for the target species and further constraints for the text-mining approaches. Our system shows the conserved regulatory network as well as annotations for each found conserved relation. This includes information about the orthologs, the textual positions where our text-mining approach found regulatory relations in the literature, the TFBS predictions and the protein-protein interaction score from the STRING database (version 9) [29]. For further analysis the networks can be exported as tab separated files for use in advanced network analysis tools. An example can be seen in Figure 1, which shows a screenshot of the *Conservation Browser* with conserved relations for *D. melanogaster* as source species. The detail view window in the front shows information about the regulatory relations in the target species including an example of a regulatory relation which was discovered by RelEx between *Pax6* and *Six3* in *H. sapiens*.

Table 1 Overview of the model organisms from our database with the number of genes, number of predicted and known transcription factors and the number of factors with position weight matrices (PWMs). In addition, the number of regulatory relations collected from databases and relations which were extracted from the scientific literature by using different text-mining approaches (RelEx [8], SL [9] and Tri-occurrence) and from transcription factor binding site predictions (TFBS) are shown. Most regulatory relations could be found for *S. cerevisiae* and *A. thaliana* which originate mostly from genome-wide chromatin immunoprecipitation experiments. The numbers of found text-mining relations and of predicted binding sites is quite different for the model organisms.

Species	#Genes	#TF	#PWM	$\frac{100\times\#\text{TF}}{\#\text{Gene}}$	#Database	#RelEx	#SL	#Tri-Occ.	#TFBS
H. sapiens	21,673	1,416	300	6.5	3,230	20,391	29,422	103,511	220,245
M. musculus	23,497	1,431	276	6.1	932	10,682	15,616	51,729	130,456
R. norvegicus	22,503	1,181	20	5.2	321	5.950	8,905	33,857	3,050
D. rerio	21,322	1,081	0	5.9	0	2,930	4,322	16,219	0
D. melanogaster	14,076	570	139	4.0	471	2,433	3,802	11,635	6,054
C. elegans	19,992	688	6	3.2	128	102	149	385	1,308
A. thaliana	26,207	1,235	32	3.4	11,284	926	1,460	5,073	8,282
S. cerevisiae	5,884	233	170	4	29,716	812	1,446	4,036	6,075

2.1 Regulatory Data

For the source and target species, data from different data sources is available in our system. Table 1 gives an overview of the collected data in ConReg. For *S. cerevisiae* and *A. thaliana*, processed data from genome-wide chromatin immunoprecipitation (ChIP) experiments for some TFs was additionally available. This explains why quite many regulatory relations were found for these two species. For the other species only very few relations could be found which emphasizes the need of text-mining approaches to get a more complete view on the currently discovered regulatory networks. For instance, for *D. rerio* no relations were found in the databases, but 16,219 putative relations were found using text-mining. Nevertheless, we assume that data extracted from databases is reliable and use this data as source data for the discovery of conservations, whereas also the predicted relations are considered for the conservation search in the target species.

For our prediction methods, most relations were found with the Tri-occurrence text-mining approach and the TFBS predictions, but likely with a large number of false positives. Unfortunately, for some organisms the number of position weight matrices (PWM) for the search of TFBS is very limited. For *D. rerio* no PWMs were available and for *C. elegans* and *R. norvegicus* only six and 20 PWMs could be found in the public domain. This explains the comparably low number of binding site predictions for these three species. The Tri-occurrence approach was used as pre-filter for the more sophisticated relation extraction approaches RelEx and SL. By comparing the relations found with RelEx to known relations extracted from databases a small overlap can be observed. For example for *H. sapiens* 22 % of the database relations could also be found with RelEx. A similar consistency could be observed for *R. norvegicus*, *M. musculus* and *D. melanogaster*. For SL 21 % of the known relations from *H. sapiens* could be detected. By combining the two state-of-the-art relation extraction approaches RelEx and SL, this rate could be increased to 28 % for *H. sapiens*. For those species with regulatory data from ChIP experiments (*S. cerevisiae* and *A. thaliana*), this fraction is much lower as can be seen on the number of found regulatory text-mining relations. Furthermore, for *A. thaliana* and *C. elegans* only 34,729 and 31,325 species relevant abstracts

could be found, whereas for *H. sapiens* and *M. musculus* 13,053,996 and 1,121,698 abstracts could be used. This explains the quite small number of relations for *A. thaliana* and *C. elegans* extracted with our text-mining approaches.

Most of the TFBS predictions could neither be confirmed with databases knowledge nor with the text-mining results. For example for *S. cerevisiae* 84 % of the predictions were unique for this method. The number and accuracy of TFBS predictions strongly depends on the available PWMs and their quality. Short PWMs for example produce many hits, but only with low scores which were not considered for the predicted relations in ConReg.

The currently available regulatory data is distributed in many different databases and stems from different data sources such as manual literature curations or genome-wide ChIP experiments. In general, the overlaps and consistency between different sources are still quite small. The collection and integration of data from different sources and organisms is a difficult task and needs to be continued to make the most out of the available knowledge.

2.2 ConReg for the Discovery of Conserved Relations in *D. melanogaster*

We used *D. melanogaster* as source species to outline the usability of our system to find conserved regulatory relations for the 471 documented regulatory relations in REDfly[10]. We selected as target species the vertebrates *H. sapiens*, *M. musculus*, *R. norvegicus* and *D. rerio* and used all available data sources for these species. *D. melanogaster* is phylogenetically distant from the other species, but several conserved motifs are described in the literature as already mentioned in the introduction. We checked all conserved relations predicted by ConReg. We assume that relations extracted from databases are correct and manually checked the relations found with our Tri-occurrence approach by reading the literature reported as evidence for each found relation. The Tri-occurrence relations are a super set of the relations extracted with our other text-mining approaches so that the performance for these approaches could also be checked, whereas the TFBS predictions were compared to the relations found in the databases and with the text-mining approaches.

The entire predicted conserved network is shown in Figure 2. Manual annotations where we could confirm a conserved regulatory relation between the orthologs in at least one vertebrate are shown as red edges. The conserved *D. melanogaster* network also contains the well-studied motifs for eye-development (*Optix*, *ey*, *eya* and *shf*) and conservations for the pan-bilaterian kernel for heart specification, including the genes *Tin*, *Mef2* and *Mad*.

Only seven conserved relations, involving nine different genes could be identified with target relations extracted from databases. From these seven relations four were auto-regulations and the others were isolated edges. By using only the knowledge from databases, not even the well-studied conserved motifs between *D. melanogaster* and the other organisms could be rediscovered. With our Tri-occurrence approach 132 possible conservations could be found from which we could confirm 66 relations (50 %) in at least one species. We compared the different methods to each other with respect to the number of predicted and confirmed relations. Furthermore, we compared the intersections of the predicted conserved relations from the different approaches (see Figure 3). All of the 67 found conservations found with RelEx could be confirmed or were also found by SL or the binding site predictions. With SL six additional validated conservations could be found. In addition, 124 possible conserved relations were discovered with the TFBS predictions. 33 of these relations could be found with a different method including 25 confirmed Tri-occurrence relations. We note, that with RelEx the best relation extraction performance could be achieved with 57 out of 67 confirmed

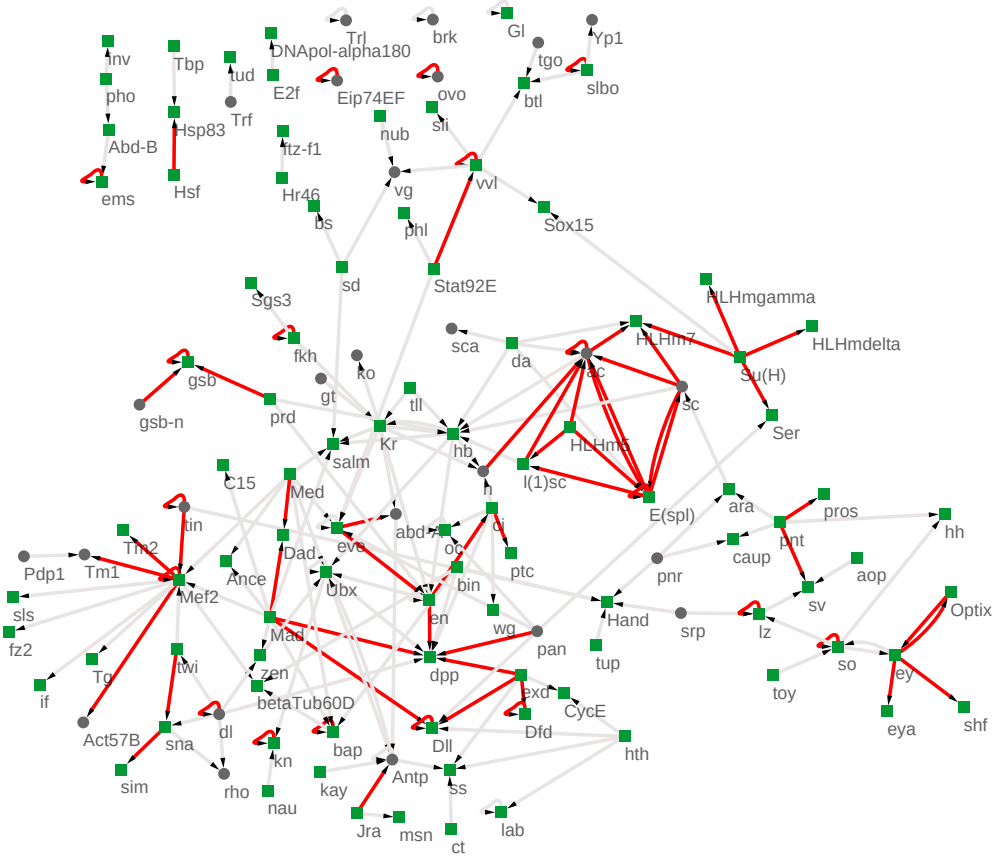

Figure 2 Network of conserved regulatory relations from the 471 documented regulatory relations in REDfly for *D. melanogaster* and in at least another vertebrate. The gray edges represent all relations where we could find a possible conservation. Red edges represent edges where we could confirm the relations between the orthologs in vertebrates using the literature references provided by ConReg. In green, we highlighted the nodes where at least two ortholog identification approaches found an ortholog mapping to another vertebrate for the respective gene. The conserved network contains among others, the well studied motifs for eye-development and the pan-bilaterian kernel for heart specification.

conserved relations (85 %). With SL a similar performance could be reached with 59 out of 76 confirmed conserved relations (78 %).

2.3 Comparison to alternative tools

There are several other tools which support the identification of conserved relations in eukaryotes. For example, with the UCSC Genome Browser [7] one can map TFBS (e.g. from ORegAnno [19]) and genome-wide chromatin immunoprecipitation experiments to the genome of interest together with the conservation of DNA sequences for different species. Another tool, the Genomatix suite[1], allows for uploading experimental data and for searching for conservations.

[1] http://www.genomatix.de

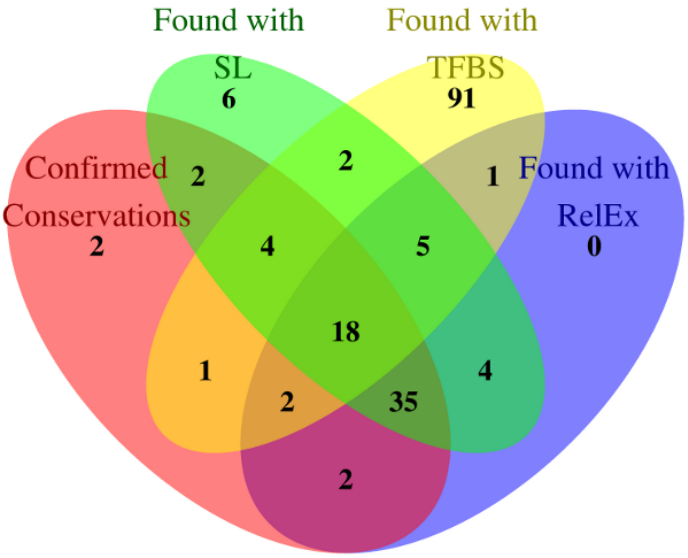

Figure 3 Venn-Diagram of found regulatory conservations between the 471 documented regulatory relations in REDfly for *D. melanogaster* and vertebrates. Confirmed conservations are regulatory relations which could be transferred from databases or which are correctly identified with our Tri-occurrence approach for at least one vertebrate (relations were manually checked by reading the corresponding literature). All relations found with RelEx could also be found with another method, whereas most of the TFBS predictions were not reported with our text-mining approaches.

Compared to prokaryotic genomes, eukaryotic genomes are rich in non-coding sequences of unknown functions and promoters can be several kilobases upstream from the transcription start site. Nevertheless, different approaches have been introduced to predict conserved binding sites [16, 4].

Furthermore, for microbial gene regulatory networks different platforms exist for the storage and web based analysis as reviewed by Baumbach *et al.* [3].

In comparison to these tools, ConReg focuses on eukaryotes only. ConReg provides detailed information of possible conservations. The main contribution of ConReg is the addition of conserved relations mined from the publicly available literature, only a small fraction of which is currently represented in databases. Moreover, TFBS predictions are also integrated. All these data can easily be selected via the web-interface of ConReg, via a Cytoscape [28] plug-in or downloaded from our server.

3 Conclusion

ConReg is a novel interactive online system for the discovery of conserved regulatory relations in currently eight eukaryotic model organisms. Our system allows searching for regulatory conservations among arbitrary user-definable sets of target species and outputs several annotations for predicted conserved relations.

We collected regulatory relations from databases, via text-mining from scientific text descriptions and from binding site predictions. We observed a severe incompleteness of regulatory relations in databases which are not even sufficient for the discovery of well-known conserved motifs. This slightly improves via the integration of information from state-of-the-art text-mining approaches and binding site predictions. E.g., several conserved motifs could be found using *D. melanogaster* as source species. For this showcase ConReg could identify

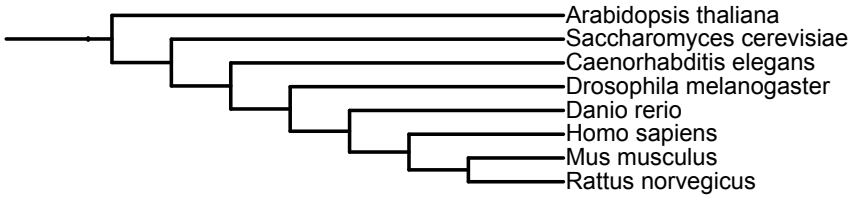

Figure 4 The eight species considered in our study and the associated phylogenetic tree as extracted from the NCBI taxonomy tree. Currently, ConReg contains six animal model organisms (*H. sapiens*, *M. musculus*, *R. norvegicus*, *D. rerio*, *D. melanogaster*, *C. elegans*) as well as *S. cerevisiae* and the model plant *A. thaliana*.

conserved regulations for 14 % (66 out of 461) of the relations from *REDfly* in at least one vertebrate. Still, it remains unknown to which extent regulatory relations are conserved since only few regulatory relations are experimentally confirmed for eukaryotes.

For our selected showcase we noticed that even with the simple Tri-occurrence text-mining approach 50 % of the identified regulatory relations are correctly identified if also experimentally validated regulatory relations between orthologs are known. Thus, with the integration of additional background knowledge the relation extraction could be significantly increased. We designed our system in such a manner, that other information sources can easily be added. In particular, we are planning to incorporate information from the increasing number of available ChIP experiments into ConReg.

Availability

The ConReg web interface is publicly available at `http://services.bio.ifi.lmu.de/ConReg/` and an interface is provided as Cytoscape plug-in to access the data for follow-up analyses. Furthermore, the extracted text-mining relations are provided for download on our website.

4 Materials and Methods

4.1 Data Sources

We collected regulatory relations for *H. sapiens*, *M. musculus*, *R. norvegicus*, *D. rerio*, *D. melanogaster*, *C. elegans*, *A. thaliana* and *S. cerevisiae* (see Figure 4 for a phylogenetic tree of these species). Regulatory relations were extracted from the multi-species curated databases TRANSFAC (Version 9.3) [18] and ORegAnno [19]. Species-specific relations were extracted from YEASTRACT [31], REDfly [10] and AtRegNet [20] and from curated pathways from Biocarta and NCI-Pathway [24]. Transcription factors were collected from these databases and the transcription factor prediction database [14]. For the transfer of relations, we used orthologs from InParanoid [21], EnsemblCompara [35] and OMA [25]. These databases were used due to the evaluation results in [1, 12] and the coverage of ortholog mappings for all considered species. All genes were mapped to Ensembl to obtain unique genomic locations and the associated annotations. Relations involving genes which could not be mapped were not considered.

4.2 Regulatory Relation Extraction from the Scientific Literature

Abstracts from PubMed (20,766,340 abstracts) and the corresponding full text publications from the PubMedCentral open access subset (389,322 documents) were used to search for regulatory relations in textual descriptions. In order to find relations in unstructured descriptions two tasks have to be accomplished: the named entity recognition (NER) of gene names and the correct identification of relations between genes. For example, consider the following sentence from [17]: *"There is evidence that the expression of Six3 is regulated by Pax6."* To infer a regulatory relation, the gene names *Six3* and *Pax6* need to be found and the regulatory relation between *Pax6* → *Six3* has to be identified. We used syngrep [5], a dictionary based NER tool, for the gene name recognition and the mapping of gene names to identifiers. Dictionaries were compiled by combining gene names, aliases and synonyms from UniProt, Ensembl, HGNC, MGI, RGD, Tair and FlyBase. Regulatory relations between genes were initially identified with a simple Tri-occurrence approach. For this approach, a relation was assumed between all pairs of genes which were found in a sentence, if a keyword indicating a regulatory relation was found and at least one gene is annotated as a TF. For this task, we defined a list of keywords, which are supposed to indicate regulatory interactions, like *regulates*, *represses*, or *downregulates*. Such a Tri-occurrence approach provides a good recall, but also implies many false positives. Therefore, we also used the following more sophisticated relation extraction approaches to filter the discovered relations found with the Tri-occurrence approach:

RelEx [8]: RelEx is a rule based relation extraction tool using dependency parse trees to find relations.

SL [9]: SL is a shallow linguistics SVM kernel for the identification of relations. Since no model was available for the identification of regulatory relations with this kernel, we used the simple margin active learning [34] approach to train a SVM. A set of 175 positive and negative relations was used to learn an initial model. This model was refined by applying the learned predictor to 10,000 randomly selected relations found by the Tri-occurrence approach. The 100 instances which were closest to the separating margin of the SVM were manually annotated and used for the next round of SVM training. The model was iteratively refined with this approach until no further performance improvement on a control set of 100 examples (including 33 positive regulatory relations) could be observed. An area under the receiver operating characteristic (ROC) of 0.85 and an area under the precision-recall curve of 0.72 could be achieved with the final model on the control set. Furthermore, with a standard probability threshold of 0.5 for the SVM, a precision of 0.56 and a recall of 0.75 were reached (see Figure 5 for the ROC curve of the final predictor).

We decided to use RelEx and SL due to their good performance on the task of identifying undirected protein-protein interactions on different corpora [33]. The Tri-occurrence and the SL kernel approach predict only undirected relations. We used our list of TFs to derive the direction from the transcription factor to the non-factor. In the case that both genes are non-factors or both are factors the relations were omitted by default in our system. Gene names between closely related species can highly overlap. Therefore, we identified the species context in each abstract using a pre-defined set of possible names for the different species.

4.3 Transcription Factor Binding Site Predictions

The promoter sequence for each gene was extracted using RSAT [32]. The same promoter size of 1 kilo base pairs upstream of the transcription start site was chosen for all species.

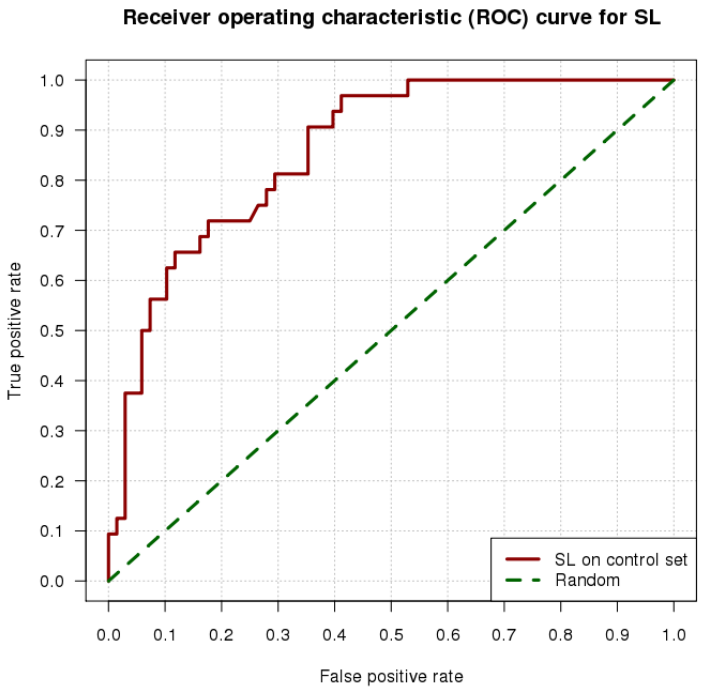

Figure 5 Receiver operating characteristic (ROC) curve for the final shallow linguistics (SL) SVM model which was used for the identification of regulatory relations. The evaluation set consists of 100 examples including 33 positive regulatory relations and 67 negative regulatory relations. On this control set the model reached an area under the ROC of 0.85 and an area under the precision-recall curve of 0.72. With a standard probability threshold of 0.5 for the SVM, a precision of 0.56 and a recall of 0.75 were reached.

Binding motifs for the different TFs were taken from TRANSFAC [18] and JASPAR [23]. The matching of these motifs to the promoter sequences was predicted with the R package cureos [37]. We used an empirically chosen threshold of 16 on the TFBS scores to filter out insignificant binding sites.

4.4 ConReg System Design

ConReg was developed as object oriented Java application using the open-source Ajax Web application framework ZK. The underlying data was unified in a structured MySQL database.

Funding

This work was supported by the DFG international research training group (IRTG) RECESS 1563 via scholarships to RP and MB.

Authors' contributions

RP designed and implemented ConReg and drafted the paper. MB provided TFBS predictions. RZ designed the study and helped with the evaluations. All authors contributed in reading and editing the paper.

References

1 Adrian M. Altenhoff and Christophe Dessimoz. Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods. *PLoS Comput Biol*, 5(1):e1000262, 2009.

2 Jan Baumbach. On the power and limits of evolutionary conservation–unraveling bacterial gene regulatory networks. *Nucleic Acids Res*, 38(22):7877–7884, Dec 2010.

3 Jan Baumbach, Andreas Tauch, and Sven Rahmann. Towards the integrated analysis, visualization and reconstruction of microbial gene regulatory networks. *Brief Bioinform*, 10(1):75–83, Jan 2009.

4 Eugene Berezikov, Victor Guryev, and Edwin Cuppen. CONREAL web server: identification and visualization of conserved transcription factor binding sites. *Nucleic Acids Res*, 33(Web Server issue):W447–W450, Jul 2005.

5 Gergely Csaba. syngrep - Fast synonym-based named entity recognition. Master's thesis, LMU-Munich, 2008.

6 Eric H. Davidson. *The Regulatory Genome. Gene Regulatory Networks In Development and Evolution.* Academic Press, Amsterdam, 2006.

7 Pauline A. Fujita, Brooke Rhead, Ann S. Zweig, Angie S. Hinrichs, Donna Karolchik, Melissa S. Cline, Mary Goldman, Galt P. Barber, Hiram Clawson, Antonio Coelho, Mark Diekhans, Timothy R. Dreszer, Belinda M. Giardine, Rachel A. Harte, Jennifer Hillman-Jackson, Fan Hsu, Vanessa Kirkup, Robert M. Kuhn, Katrina Learned, Chin H. Li, Laurence R. Meyer, Andy Pohl, Brian J. Raney, Kate R. Rosenbloom, Kayla E. Smith, David Haussler, and W James Kent. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*, 39(Database issue):D876–D882, Jan 2011.

8 Katrin Fundel, Robert Küffner, and Ralf Zimmer. RelEx–relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 2007.

9 Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. In *In Proc. EACL 2006*, 2006.

10 Marc S. Halfon, Steven M. Gallo, and Casey M. Bergman. REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. *Nucleic Acids Res*, 36(Database issue):D594–D598, 2008.

11 Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene van Someren, and Reinhard Guthke. Gene regulatory network inference: data integration in dynamic models – a review. *Biosystems*, 96(1):86–103, Apr 2009.

12 Tim Hulsen, Martijn A. Huynen, Jacob de Vlieg, and Peter M A. Groenen. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol*, 7(4):R31, 2006.

13 Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, 40(Database issue):D109–D114, 2012.

14 Sarah K. Kummerfeld and Sarah A. Teichmann. DBD: a transcription factor prediction database. *Nucleic Acids Res*, 34(Database issue):D74–D81, 2006.

15 Myon-Hee Lee, Brad Hook, Guangjin Pan, Aaron M. Kershner, Christopher Merritt, Geraldine Seydoux, James A. Thomson, Marvin Wickens, and Judith Kimble. Conserved regulation of MAP kinase expression by PUF RNA-binding proteins. *PLoS Genet*, 3(12):e233, 2007.

16 Gabriela G. Loots and Ivan Ovcharenko. rvista 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*, 32(Web Server issue):W217–W221, Jul 2004.

17 Martine Manuel, Thomas Pratt, Min Liu, Glen Jeffery, and David J. Price. Overexpression of Pax6 results in microphthalmia, retinal dysplasia and defective retinal ganglion cell axon guidance. *BMC Dev Biol*, 8:59, 2008.

18 V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue):D108–D110, 2006.

19 S. B. Montgomery, O. L. Griffith, M. C. Sleumer, C. M. Bergman, M. Bilenky, E. D. Pleasance, Y. Prychyna, X. Zhang, and S J M. Jones. ORegAnno: an open access database and curation system for literature-derive promoters, transcription factor binding sites and regulatory variation. *Bioinformatics*, 22(5):637–640, 2006.

20 Saranyan K. Palaniswamy, Stephen James, Hao Sun, Rebecca S. Lamb, Ramana V. Davuluri, and Erich Grotewold. AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiol*, 140(3):818–829, 2006.

21 M. Remm, C. E. Storm, and E. L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 314(5):1041–1052, 2001.

22 Richard Röttger, Ulrich Rückert, Jan Taubert, and Jan Baumbach. How Little Do We Actually Know? – On the Size of Gene Regulatory Networks. *IEEE/ACM Trans Comput Biol Bioinform*, 2012.

23 Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W. Wasserman, and Boris Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(Database issue):D91–D94, 2004.

24 Carl F. Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H. Buetow. Pathway Interaction Database. *Nucleic Acids Res*, 37:D674–9, 2009.

25 Adrian Schneider, Christophe Dessimoz, and Gaston H. Gonnet. OMA browser–exploring orthologous relations across 352 complete genomes. *Bioinformatics*, 23(16):2180–2182, 2007.

26 Kate Schroder, Katharine M. Irvine, Martin S. Taylor, Nilesh J. Bokil, Kim-Anh Le Cao, Kelly-Anne Masterman, Larisa I. Labzin, Colin A. Semple, Ronan Kapetanovic, Lynsey Fairbairn, Altuna Akalin, Geoffrey J. Faulkner, John Kenneth Baillie, Milena Gongora, Carsten O. Daub, Hideya Kawaji, Geoffrey J. McLachlan, Nick Goldman, Sean M. Grimmond, Piero Carninci, Harukazu Suzuki, Yoshihide Hayashizaki, Boris Lenhard, David A. Hume, and Matthew J. Sweet. Conservation and divergence in Toll-like receptor 4-regulated gene expression in primary human versus mouse macrophages. *Proc Natl Acad Sci U S A*, 109(16):E944–E953, 2012.

27 Rachita Sharma, Patricia A. Evans, and Virendrakumar C. Bhavsar. Regulatory link mapping between organisms. *BMC Syst Biol*, 5 Suppl 1:S4, 2011.

28 Michael E. Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, and Trey Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432, Feb 2011.

29 Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*, 39(Database issue):D561–D568, 2011.

30 Leila Taher, David M. McGaughey, Samantha Maragh, Ivy Aneas, Seneca L. Bessling, Webb Miller, Marcelo A. Nobrega, Andrew S. McCallion, and Ivan Ovcharenko. Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Res*, 21(7):1139–1149, 2011.

31 Miguel C. Teixeira, Pedro Monteiro, Pooja Jain, Sandra Tenreiro, Alexandra R. Fernandes, Nuno P. Mira, Marta Alenquer, Ana T. Freitas, Arlindo L. Oliveira, and Isabel Sá-Correia.

The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Res*, 34(Database issue):D446–D451, 2006.

32 Morgane Thomas-Chollier, Matthieu Defrance, Alejandra Medina-Rivera, Olivier Sand, Carl Herrmann, Denis Thieffry, and Jacques van Helden. RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res*, 39(Web Server issue):W86–W91, 2011.

33 Domonkos Tikk, Philippe Thomas, Peter Palaga, Jörg Hakenberg, and Ulf Leser. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Comput Biol*, 6:e1000837, 2010.

34 Simon Tong, Daphne Koller, and Pack Kaelbling. Support Vector Machine Active Learning with Applications to Text Classification. In *Journal of Machine Learning Research*, pages 999–1006, 2001.

35 Albert J. Vilella, Jessica Severin, Abel Ureta-Vidal, Li Heng, Richard Durbin, and Ewan Birney. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*, 19(2):327–335, Feb 2009.

36 S. Wawersik and R. L. Maas. Vertebrate eye development as modeled in Drosophila. *Hum Mol Genet*, 9(6):917–925, Apr 2000.

37 Frank Westermann, Daniel Muth, Axel Benner, Tobias Bauer, Kai-Oliver Henrich, André Oberthuer, Benedikt Brors, Tim Beissbarth, Jo Vandesompele, Filip Pattyn, Barbara Hero, Rainer König, Matthias Fischer, and Manfred Schwab. Distinct transcriptional MYCN/c-MYC activities are associated with spontaneous regression or malignant progression in neuroblastomas. *Genome Biol*, 9(10):R150, 2008.

38 Haiyuan Yu, Nicholas M. Luscombe, Hao Xin Lu, Xiaowei Zhu, Yu Xia, Jing-Dong J. Han, Nicolas Bertin, Sambath Chung, Marc Vidal, and Mark Gerstein. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res*, 14(6):1107–1118, Jun 2004.

39 Zongzhao Zhai, Nati Ha, Fani Papagiannouli, Anne Hamacher-Brady, Nathan Brady, Sebastian Sorge, Daniela Bezdan, and Ingrid Lohmann. Antagonistic regulation of apoptosis and differentiation by the Cut transcription factor represents a tumor-suppressing mechanism in Drosophila. *PLoS Genet*, 8(3):e1002582, Mar 2012.