GREEKC MEETING MALTA

Setting the stage for the Gene Regulation Knowledge Commons

Time	Subject	Speaker
April 3, 2017:		
08:30 -08:40	Introduction	
	Session 1: Survey of what is available	
08:40	Capturing Gene Regulation Knowledge: Where we are/where we want to be	Sandra Orchard (SO)
09.10	RegulonDB and the challenge of encoding knowledge	Julio Collado Vides (JCV)
	Next: 5 lightning talks: 10 minute presentations of current resources + 2 minute questions	
09:40	SIGNOR: A signaling resource	Livia Perfetto (LP)
09:52	HTRIdb: A collection of experimentally validated human transcriptional regulatory interactions	Marcio Acencio (MLA)
10.04	Yeastract: Predicting gene and genomic regulation in yeasts	Miguel Teixeira (MT)
10.16	ReMap : Integrative analysis of public ChIP-seq experiments	Benoît Ballester (BB)
10:28	MINT: The Molecular INTeraction database	Luana Licata (LL)
10.40 - 11.00	**Break**	
	Session 2: User stories	
11:00	The landscape of literature curated signaling pathways	Dénes Türei (DT)
11:30	iRegulon and i-cisTarget: Reconstructing Regulatory Networks Using Motif and Track Enrichment	Stein Aerts (SA)
12:00	Regulatory networks reconstruction using literature based knowledge	Julien Dorier (JD)
12:30 - 15.30	**Lunch**	
	Session 3: Systems level representations	
15:30	Tactical formalization with OWL	Michel Dumontier (MD)
16:00	The Gene Regulation Ontology: Scope, Requirements and Design Principles	Stefan Schulz (SS)
16:30	Representing complex models of biology using the Gene Ontology	Paul Thomas (PT)
	Session 4: Network component level representations	
17:00	Using Gene Ontology to describe transcription factors	Ruth Lovering (RL)

April 4, 2017:

	Session 5: Nucleotide level annotations	
08:30	The PSI-MI format: Capturing molecular interaction data	Sandra Orchard (SO)
09.00	A functional annotation resource for microRNAs	Rachael Huntley (RH)
09:30	Capturing the structure of genomic functional elements using Apollo	Suzanne Lewis (SL)
10:00	GENCODE gene annotation and regulation	Daniel Zerbino (DZ) Adam Frankish (AF)
10:30 – 10:50	**Break**	
	Session 6: Text mining	
10:50	BioCreative: Lessons learned from the user interactive task and beyond	Cecilia Arighi (CA)
11:20	Text mining technologies and assisted curation	Fabio Rinaldi (FR)
11.50	BeCalm, PITL and OpenMinted initiatives on integration, annotation and gold standards for text mining	Martin Krallinger (MKr)
12:05	Introduction to breakout sessions: Discussion topics and targets	Martin Kuiper (MK)
12:15 – 15:30	**Lunch**	
	Session 6: Breakout discussions	
15:30 -17:15	Breakout sessions – Integrated Gap and Solution analysis. Topics will be decided during the workshop and assigned to the Working Groups (ontologies, curation, text mining, data sharing). A sign-up sheet will be available.	WG leaders
17:15 – 18:00	Breakout reports, conclusions and synthesis of the workshop	WG leaders

Workshop report

Introduction: MK - The GREEKC COST Action is an initiative to bring people together who ‘have a stake’ in the Gene Regulation Knowledge Commons, meaning that they are involved in the production and maintenance of the Knowledge Commons or that they are users of the KC resource for computational analysis of regulatory systems. The COST Action is the result of GRECO: the Gene Regulation Consortium, a global initiative to organize and structure the way knowledge about gene regulation is curated, annotated, stored and shared. The GREEKC Action has 4 working groups dedicated to 1) Ontologies, 2) Curation, 3) Text mining and 4) Data sharing. Representatives of all areas are together to discuss the state of the art of the Gene Regulation Knowledge Commons (GRKC), current gaps and hurdles, and blue-sky scenarios that may guide the further development of the KC.

Session 1: Resources

SO presented a first analysis of a survey to catalogue existing resources. The prime reason for doing so is to know what resources are active in the domain of gene regulation. It is important to know what type of data these resources contain, how curation is organised, what ontologies and CVs are used, as a first step to assess whether their content might be worth preserving and made widely shareable by for instance common web services. One such web service is PSICQUIC, a project within the HUPO Proteomics Standard Initiative, with the aim to standardise programmatic access to molecular interaction databases. Blue-sky: more widespread sharing of molecular interaction data through PSICQUIC.

JCV presented a special example of a resource: RegulonDB. This database contains curated information generated in a high-throughput way, assisted by text mining, on E. coli K12. RegulonDB contains a huge amount of information concerning mechanisms of regulation. The information is qualitative, and defines Gensor units: the whole chain between a Sensor, signaling pathways, TFs and target genes, and the proteins / enzymatic pathways encoded. Although RegulonDB focuses on eukaryotes, its design may constitute a model for what the GRKC should enable. Blue-sky: developing REST services for computational access to RegulonDB.

LP presented the SIGNOR database, a resource with molecular interactions relevant to understand signaling pathways and mechanisms down to the gene regulation level. Interactions are described as binary, causal interactions (meaning the effect and direction of the regulation is recorded). Blue-sky: further development as a platform that allows browsing of integrated gene regulation data, also taking into account the spatiotemporal context.

MLA presented the HTRIdb, a database that contains information about human transcription regulation, including a largen number of TF-TG interactions based on ChIP data, among others from large-scale efforts. The database is manually curated, but not maintained anymore. Blue-sky: it would be nice to have more TF-TG interactions in the database, and sharing the resource through a PSICQUIC web service.

MT presented the Yeastract database, a resource with transcription regulation information on yeast: transcription factors, binding sites, target genes and regulatory effect. The database is a resource to study gene expression data and also predict gene regulation events. Blue sky: include ChIP analysis, enable better GO classifications and orthology predictions.

BB presented ReMap, a platform that enables the integrated analysis of public ChIP-seq data. The Blue-sky scenario is to convert it into a platform (Unimap) that will be portal for distributed mass annotation of gene regulatory elements. Experts who curate a paper can do so aided by an automatically generated scaffold of TF-DNA interactions generated by the ReMap software.

LL presented the MINT resource: the Molecular Interaction database. MINT is an example of a database that was committed rather early to the agreed standards for molecular interaction data, and the data sharing and further curation efforts are closely coordinated with the IntAct team, making sure essentially that its valuable content would ‘persist’ and remain available as part of the Knowledge Commons. In this sense Blue-sky: mapping the entire human genome, or at least: more coverage.

Session 2: User stories

DT presented the Omnipath resource, developed to facilitate modelling of regulatory processes. Omnipath collects information from a wide repertoire of resources, with confidence levels varying from high to somewhat lower. 30% of the human proteome is covered in directed interactions, and 60% in undirected interactions. Pathways are represented as sets of components (proteins/genes) plus a topology. The resource can be used as input for logical modelling. Blue-sky: establish confidence weighting system and include drug target interactions to enable integration of drug-target with target-target interaction information

SA talked about using position weight matrix collections, including motif discovery in a set of co-expressed genes. Some 20.000 PWM have been pre-processed and filtered and can be linked to TFs (motif2TF). He pointed to opportunities arising from combining PWM-based analysis with ChIP-based ranking for identification of TF-binding. Blue sky: TF-binding data to feed into comprehensive TF-TG regulatory network resources to be used e.g. to rank likelihood of gain or loss of function of mutations in cancer

JD described how qualitative regulatory networks can be constructed using curated knowledge. The work demonstrated the benefits of having modellers work closely together with curators. This allowed a good definition of what needs to be recorded in the curation process. JD and his team record and use both direct molecular interactions and indirect causal interactions. Blue-sky: an integrated resource that has both causal interaction information and the context in which these statements are true.

Session 3: Systems level representations

MD provided an introductory presentation of OWL: the web ontology language, and he showed examples of how it can be used to model biological processes and events, including gene regulatory events.

SS described the Gene Regulation Ontology, a conceptual model developed for gene regulation and published in 2008. GRO could be developed further, and subjected to competency questions to see how it can mature into a useful artifact for describing and querying gene regulation events. Blue-sky: Develop GRO into a computable, ontology-based framework.

PT discussed how the Gene Ontology could be used to represent complex biological models, in a form that would deliver computable models. The Noctua annotation tool was presented as a new direction for gene annotation, focusing on how molecular functions of genes/proteins (their activities) regulate and participate in biological processes. The tool is for people who are experts in GO, but other user interfaces could make information entry easier. As the curated information is stored in OWL, a reasoner can do checking of entries in real-time, offering quality assurance. Blue-sky: develop and increase expressivity of Noctua annotations e.g. by developing ‘signal integrator’ functionalities that can integrate a wide variety of regulatory input (like e.g. an array of gene regulatory elements and their interaction with transcription factors in determining the outcome of gene regulatory signals).

Session 4: Network component level representation

RL presented protein centric approaches, focusing on the annotation of transcription factors. Annotation extensions have been invented to be able to cope with additional annotation details for instance … The curation process makes use of Protein2GO, an efficient annotation tool for entry of protein/gene IDs and GO term annotations. In a comparison of Protein2GO to Noctua the differences were likened to those between a racehorse (Protein2GO: fast and light) and an elephant (slow but good for heavy loads).

SO presented a historical perspective of the HUPO PSI-MI standard, and the emergence of the iMEX common standard for molecular interactions. The standards initially covered protein-protein interactions, but it was clear that also other complexes should be considered (protein/DNA, protein/RNA, RNA/RNA etc.). Several PSI-MI XML versions have been released with increased coverage of molecule types, cell and tissue types, experimental conditions and interaction types. Lately attention has been directed to the representation of causality and the need to capture directionality. This has given rise to CausalTab, an extension of the MITab exchange format definition that includes causal interactions. Blue-sky: All data connected through web services / pipelines (PSICQUIC, other) that connect to the Knowledge Commons (interactions protein-protein, protein-genome, directionality, effects, metadata/context), and aligned with the most recent Ensembl builds (particularly with regards to regulatory elements).

Session 5: Nucleotide level annotation

RH described the use of GO for the annotation of non-coding RNAs. Interactions are annotated with annotation extensions containing Relation Ontology terms terms. Large interaction datasets involving ncRNAs (high proportion miRNA) can be retrieved via PSICQUIC. Ontology for ncRNA (NCRO) is available, but needs rework/update. Blue-sky: Build ncRNA resources annotated with functional aspects including causal interactions, biological context (cell/tissue type etc), and with info on transcription factors regulating ncRNA gene expression; possibly achieved via community-driven annotation efforts.

SL presented the webtool Apollo, which allows users to add annotations to gene sequences that are not generated automatically. These may include for instance natural sequence variants. Blue-sky: New ways to visualize curation results, and regulatory sites and domains annotated on the genome, for easy interpretation?

DZ/AF presented ENSEMBL, focusing on regulatory track display. This involves a merging of HAVANA and Genebuilds and is embedded in the GENCODE consortium. The HAVANA efforts use manual curation, based on annotation guidelines. Blue-sky: Efficient workflows to enable combination of many data types to underpin manual gene curation. High quality annotation of the non-coding genome for the Gene Regulation Knowledge Commons.

Session 6: Text mining

CA gave an overview of text mining: what works and what doesn't. There are a number of gaps that need attention, for instance problems related to the pdf format (difficult to access the text), and the lack of proper IDs in text, but most importantly the lack of links between curation and text mining. Blue-sky: Repository that offers all available text mining tools. Unlimited access to all full length papers.

FR talked about the different tasks needed to mature text mining pipelines. These include document classification (yes/no interest, 'triage'), entity recognition and normalization, and finding relationships between entities. Blue-sky: Text-mining-based pipelines that support digital and semi-automated annotation of components, functions, processes, biological context.

MKr talked about text-mining efforts and gold standards needed to improve text mining approaches by tailoring them to the needs of biologists and integrating them into knowledge management systems. Biologists should provide lists of rules, or guidelines of what text miners should extract from literature. Advanced text mining approaches can make the literature more accessible to biologists. Blue-sky: Text mining tools made available via web services/other are technically interoperable, accessible for performance comparisons and linked to rich resources of Gold standard corpora, curation guidelines, and error analysis approaches.

Breakouts:

Discussions centered around the thematics of the four working groups identified a number of action points that the WGs should focus on for the next Workshop to be held in Lisbon.

WG1: The GRO was discussed. Possibly the idea behind it could be revived, so for the next meeting an architecture of this GRO could be proposed, including how current ontologies are sufficient to supply the necessary terms and concepts for GRO. However, it is clear that GRO would just serve as an 'under the hood' engine for gene regulation knowledge, and as such also links with Noctua should be investigated.

WG2: This group focused on how causality could be better covered by MITAB, which was originally developed to describe physical interactions, and not directionality or effects on processes: descriptions of activation cannot be expressed as a change in an affected protein, but rather in a change of what the protein does (similar to how Noctua describes causality). The role of the Relation Ontology in describing causality will be assessed, and likewise the role of the Sequence Ontology for describing gene features essential for gene regulation. Examples of directed/causal interactions will be mapped to the main representation schemas available today (Noctua, BEL, SciCura/VSM, CausalTab). Display and visualization of gene regulation knowledge in genome browsers will be discussed, to see how this information can be more easily 'consumed' both by humans and computers.

WG3: This group discussed how text mining could be better integrated into the curation workflow. Text mining tools need to become easier to find, for instance through a portal. It was concluded that direct interactions between text miners and curators are essential for this, as off-the-shelf tools often need careful configuration to provide acceptable results. Before the next meeting an assessment should be done of a subset of text mining approaches giving the maximum impact on curation.

WG4: This group discussed user needs, and how to take care of all the details and intricacies of gene regulation in databases. One of the conclusions was that this all starts with the building of conceptual models of regulatory processes and events, and that a more intense interaction with 'users' is needed (at conferences, or through use cases that are submitted to the GREEKC website). The idea was discussed to organize a user/developer workshop, including a practical session where users and developers interact in a hackathon-mode to identify how resources and tools can create a better interoperability.