White Paper Gene Regulation Knowledge Commons
An opinion paper from
the Gene Regulation Consortium - GRECO
Initiated by the COST Action CA15205 - GREEKC

Content:

What is the Gene Regulation Knowledge Commons
-Brief description of the GRKC and why this was taken up as a mission by GRECO
-Description of the interdisciplinarity necessary to develop a well constructed knowledge commons, and how COST and other projects help us to achieve that.

Biocuration
Summary of the ISB paper, setting the stage of Biocuration

Where does GRECO aim to extend on this, and why
Map out the specific additions that GRECO strives for:

- enabling systems biology / computational help in hypothesis assessment
- setting the stage for a resource for computational analysis and model building, with a focus on Gene regulation, which is a fundamental subdomain of biology
- assessing potential of text mining to increase efficacy of curation process
- assure interoperability and sharing

Description of the areas that underpin the Gene Regulation Knowledge Commons
Four working groups + their state of the art, gaps and opportunities

The way forward
Community engagement, communication and defining common objectives

What GRECO is about:
Biological knowledge discovery is becoming increasingly dependent on computational modelling and simulation. Model building requires comprehensive knowledge bases describing biological entities and their interactions. Dedicated action is needed to enter such knowledge in knowledge bases, as scientific results cannot be effectively shared with the community through publications alone: their information content needs to be carefully checked, or curated, and archived in standardised formats in public resources, to become broadly available for computational integration and analysis.
Many of the established Knowledge Commons resources service biological research from an ‘entity-based perspective’: Experimental data can be checked against database repositories as individual entities or sets, but from a ‘static perspective’. Analysis from a dynamic perspective, let’s call it a systems perspective, allows interpretation of dynamics/kinetics and how a system responds to perturbations in a time perspective. All this calls for a model-based approach where components are integrated through a mathematical framework that facilitates causality to propagate component changes through a causally interlinked system.

If we observe the database ecosystem with information about gene regulation, much of this information is significantly fragmented, only has limited coverage, sometimes is not compliant with existing data standards, and often lacks documented quality control procedures. Most initiatives for standardising the description, recording and exchange of biological data have been shaped by needs arising from specific molecule- or data types, and not by the challenge to cover all subdomains of a complete biological process domain. GRECO (Gene Regulation Consortium, www.theGRECO.org), specifically targets the domain of gene regulation: transcription factors interacting with the genome and RNA synthesis machinery, orchestrated by a complex web of signal transduction molecules, thus crucial to fully comprehend cellular control mechanisms at the systems level. GRECO aims to establish communication and foster coordination of the efforts of all groups who are stakeholders of the Gene Regulation Knowledge Commons: Biocurators, database managers, developers of standards and ontologies, computational tool builders and users, life scientists in general, but also SMEs active in biological discovery, publishers of scientific papers, science policy makers and funding agencies. By bringing all stakeholders together in a quest for a common understanding and consensus about how to annotate and share computable knowledge, GRECO strives to set the stage for the development of an integrated knowledge management framework for this key area of molecular biology.

The GRKC may serve as an example of what can be achieved through a concerted effort to create a particular data infrastructure that fulfils both the needs of bench biologists to access detailed information on their gene of interest, and the computational biologists who need an abundance of computationally accessible information resources. This requires that the content of the GRKC is both 'human readable' and browsable through a web interface, and available through an API or web service. For both uses the information needs to be enhanced by the 'richer' expressions of molecular entities' functions, the relations between entities, the 'emergent' effect of their interactions, as well as experimental evidence and biological context so as to underpin and enhance the use of this information in regulatory network building and computational analysis.

Biocuration - ISB paper

In the recent white paper *Biocuration: Distilling data into knowledge* (2017), the International Society for Biocuration (ISB) stresses the importance of high-quality curated digital biological knowledge resources for the entire research cycle of the life science community. As the ISB notes, "information that is not easily computationally accessible does not fundamentally advance scientific research" (ref). The ISB paper emphasizes the pivotal role of expert manual curation in enabling sharing, by adhering to accepted standards and guidelines, integrating information from different sources, and keeping track of provenance. Further, highly skilled biocurators effectively re-appraise the findings of the original author(s) and ensure the quality of the information.

To be incorporated:

- Biocuration is the process of creation and maintaining the Knowledge Commons: the ecosystem of databases and knowledge bases that describe biological function aspects of genes, proteins and other components of biological interest.
- This Knowledge Commons represents a freely available resource that the Life Sciences can rely on for biological data analysis and interpretation.

- Biocuration is done by experts, because it requires training and dedication
- They create value, with many-fold return
- The process depends on standards that should be imposed early in the research chain
- These standards include ontologies and controlled vocabularies: a formalised and hierarchical structure that describes aspects of biological reality.
- Without computational accessibility knowledge does not exist
- Innovation in curation is dearly needed, also incentives
- Recognition and support for the domain is needed
- Community curation is taking off
- End with describing what *knowledge commons* means in the ISB context - and what we want to highlight more in the current document.

Where does GRECO aim to extend on this, and why

Currently, comprehensive knowledge resources targeting the needs of systems biology and computational modeling are only beginning to become available, and a much broader and detailed coverage of their content is needed to fully exploit the additional power that computational modelling and simulation can provide in the analysis of biological systems. In order to satisfy the needs of computational biologists, additional efforts should be made to foster a dialogue between the different scientific communities, both generating and eventually using these resources. Gap-analyses need to be performed to identify data that are currently lacking in systems biology (for example a comprehensive characterization of all proteins known and hypothesized to function as transcription factors (DNA-binding and co-transcription factors)), and in addition, new experimental techniques may need to be developed and applied in comprehensive approaches to fill those gaps. The data life-cycle then needs to be completed, with researchers feeding new experimental observations back into the databases.

Information needs to be managed and stored in ways that make it (re)useable for different purposes. It has been noted that the main challenge in achieving infrastructural interoperability lies in the human dimension (Palfrey et al 2012), and the key to making biological knowledge interoperable lies in bringing people together in order to steer the innovation processes towards shared goals. The aim of GRECO is to target the subfield of gene regulation knowledge and foster dialogue and synergies between data providers, data managers and users. Examples of pragmatic actions which can be taken by members of GRECO could include contributing to the improvement of the Gene Ontology in the branches relating to gene regulation, improving data interoperability by recommending a restricted set of controlled vocabulary/ontology terms to be used in the annotation process, and potentially also developing a new standard/checklist for the capture of data enabling causal reasoning across gene regulation networks: Minimum Information for a CAusality STatement (MICAST).

Capturing Gene Regulation knowledge

To fully describe the process of gene regulation, the following pieces of information need to be captured in a systematic manner:

1. A parts list of components - an unambiguous identification of the transcription factor proteins, regulatory non-coding RNAs and complexes and the transcribed genes they regulate.
2. The positional information - the coordinates of the gene regulatory regions where the protein(s) and RNAs bind, the domains of the proteins required to bind to nucleic acids mapped to underlying reference sequences and updated in line with changes to these sequences.
3. The chromatin conformation information - The 'state' of the regulatory regions in a specific cell, as defined by methylation events (DNA or protein), chromatin accessibility, other.
4. The regulatory information - the interconnectivity of the upstream signaling network which transduces the information from cell surface to nuclear protein and the directionality and flux of the information flow.
5. The meta-data - the details of the causative environmental factors such as e.g. the cell-/tissue type and state, concentration of an agonist to which a particular cell type is exposed for a stated length of time to elicit a measured response in terms of up- and down-regulation of a defined set of genes.

Description of the areas that underpin the Gene Regulation Knowledge Commons

Standards:

As outlined above, a comprehensive and meticulous description of the regulatory network (components, relations, processes) needs to be analysed, described, curated and annotated, and made part of the KC. The meta-data associated with every experiment needs to be systematically captured in order to fully understand the context (cell/tissue type, external environment) in which a gene regulation observation was made. The use of ontologies/controlled vocabularies and common data formats then become critical to ensure that data are comparable across resources, are 'interoperable' and can be merged and analysed with software tools that follow established interoperability guidelines. The use of data standards has the additional benefit of ensuring data sustainability - many databases exist for only a limited period of time [ref exists] and the valuable information they contain, and the resources expended in capturing that data are lost unless the information can readily be merged into a more stable resource.The IMEx Consortium serves as an example of this - the data originally curated by now defunct MPIDB has been transferred to, and maintained in, the IntAct database, as has more recently the data from MINT, allowing the latter group to concentrate its resource purely on curation of new data [PMID:24234451, 22453911]. There are also gains in that the tools required to support curation activities, such as editors and validation software can be developed once only and shared by multiple groups, eliminating the need for expensive redundant software development. Examples of this working in practise include the Protein2GO tool used by many disparate groups that contribute to the Gene Ontology Consortium and the IntAct molecular interaction editorial tool shared by the members of the IMEx Consortium [PMID:25776020].

Curation:

A **biocurator** collects, annotates, and validates information and enters it into a database or equivalent resource. A biocurator will generally work using a defined set of rules, or guidelines, developed by the resource, ideally in consultation with the user community and in collaboration with other related resources. Biocurators are also often responsible for the

developed of the ontologies/controlled vocabularies used to annotate the data and play a major role in the development of standards in that field. Any initiative such as GRECO needs to bridge any gap between the data collector and the data user to ensure that the information is being captured with the depth of detail required by any downstream analysis, for example by contributing to the development of ontologies/CVs or identifying the metadata which needs to be captured.

Curation capacity:
In order to obtain the needed coverage of biological knowledge in the KC, strategies are needed to increase biocuration throughput. Such strategies must at the same time nurture continuous focus on high quality of the curated content. GRECO efforts can contribute to this by development of systems module templates (e.g. GO-Noctua), text-mining assisted triage and information extraction, and of biocuration tools/interfaces tailored for the domain of Gene Regulation. In parallel, development of tools to enable productive, quality-focussed collaboration between professional and community curators can be pursued, for example pre-loading curation tools with information extracted from the literature for subsequent expert evaluation.. Such tools can be considered to be developed at data/knowledge bases themselves (example PomBase?) and/or at institutions like PubMed (e.g. tools for web-annotation (https://europepmc.org/Annotations). Although community annotation projects have had limited success, the development of more user-friendly interfaces and the provision of scientific rewards to volunteer curators may elicit more activity.

Sharing and use:
The success of a coordinated effort to assemble an enhanced Gene Regulation Knowledge Commons with ultimately be measured by the efficacy of sharing its content. This requires the development of formats and exchange mechanisms that are equally satisfying human and computational consumption. A close interaction with computational biologists and tool builders is required for gap analysis concerning data exchange languages and web service protocols, and consensus is needed to ensure interoperability between resources and tools.

The way forward
As the ISB white paper argues, the development of ontologies, standards and formats cannot be left to isolated groups, but must be coordinated across larger fields in order to satisfy the requirements of those who will be using the knowledge resources in the end. Detailed knowledge about the end-users and their requirements thus becomes crucial for the endeavor to develop a gene regulation knowledge commons. This is why the GREEKC Action (www.greekc.org) is soliciting resource producer and user feedback, by way of surveys, use case definitions, and contributions to topical workshops dedicated to bring together representatives that have a stake in the production and use of this Knowledge Commons.