Developing a linked-data knowledgebase and organizing information about gene regulatory apparatus

The past decade has seen a continuous influx of large-scale genomic data becoming readily available providing a rich and fertile medium for the study of gene regulation. In this light the creation of an easily accessible ontology for compiling the information related to the gene regulatory apparatus is particularly valuable.

We have extensive experience in developing tools that enable the linking of heterogeneous data into comprehensive databases. For example we have built a knowledgebase of pseudogenes \cite{17099229,18957444,22951037,25157146} and described the relationships between pseudogenes and segmental duplications by extending the existing Sequence Ontology and creating logical rules for querying using Semantic Web Rule Language \cite{20615899,20529940}. This resource integrates genomics data allowing us to carry out comparative analysis of pseudogenes over multiple species \cite{25157146}.

Extracting and standardizing information from literature

While the genome sequence is an information rich resource, the true value of a genome is only as good as its annotation \cite{11433356} and one of the best ways to annotate it is by extending the capabilities of relevant biological literature \cite{10858136}. Following closely the explosion of genomics data, the number of scientific publications in the biomedical sector has seen an exponential increase \cite{20739925}. This highlighted the necessity for tools that would automatically process and extract the information recorded in journal articles, using a standardized vocabulary with ontological relationships \cite{17495904,18328823}. In this direction, we have developed a number of tools for extracting and analyzing information from literature \cite{16168087,20739925,18328823,17495904,17923450}. We created PubNet \cite{16168087}, a web-based application used to extract and integrate information from PubMed providing a graphical visualization of complex networks in order to infer functional similarities. We have also made extensive contributions in the area of digitalization of journal articles as well as promoted the development and use of structured digital abstracts \cite{20739925,18328823,17495904,17923450}.

Studying trends in the literature

Leveraging on our previous knowledge of integrative analysis we are interested in examining trends and patterns in the information recorded by journal articles. In this direction we have already started by looking at the rate at which scientific information is spread online. \cite{19649304,18614002,21603617,17465677} For this we analyzed the web statistics of PLOS article level metrics \cite{21603617}. We noticed two distinct phases: 1) the “short-term fame” of a paper and the “long-term” citation statistics. Similarily, we used PubNet to examine the collaborations and publishing patterns in the field of RNAi, indicating a “social phase transition” \cite{17465677}.