Journal of Biomedical Informatics 60 (2016) 77–83

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A knowledgebase of the human Alu repetitive elements

Izaskun Mallona *, Mireia Jordà, Miguel A. Peinado

Institute of Predictive and Personalized Medicine of Cancer (IMPPC) and Health Research Institute Germans Trias i Pujol (IGTP), Can Ruti Campus. Ctra. de Can Ruti, camí de les escoles, s/n, 08916 Badalona, Spain

ARTICLE INFO

Article history:
Received 10 August 2015
Revised 20 January 2016
Accepted 22 January 2016
Available online 28 January 2016

Keywords:
Alu
Repetitive element
Knowledgebase
Ontology

ABSTRACT

Alu elements are the most abundant retrotransposons in the human genome with more than one million copies. Alu repeats have been reported to participate in multiple processes related with genome regulation and compartmentalization. Moreover, they have been involved in the facilitation of pathological mutations in many diseases, including cancer. The contribution of Alus and other repeats in genomic regulation is often overlooked because their study poses technical and analytical challenges hardly attainable with conventional strategies. Here we propose the integration of ontology-based semantic methods to query a knowledgebase for the human Alus.

The knowledgebase for the human Alus leverages Sequence (SO) and Gene Ontologies (GO) and is devoted to address functional and genetic information in the genomic context of the Alus. For each Alu element, the closest gene and transcript are stored, as well their functional annotation according to GO, the state of the chromatin and the transcription factors binding sites inside the Alu. The model uses Web Ontology Language (OWL) and Semantic Web Rule Language (SWRL). As a case of use and to illustrate the utility of the tool, we have evaluated the epigenetic states of Alu repeats associated with gene promoters according to their transcriptional activity.

The ontology is easily extendable, offering a scaffold for the inclusion of new experimental data. The RDF/XML formalization is freely available at http://aluontology.sourceforge.net/.

1. Introduction

A striking feature of most eukaryote genomes is the abundance of repetitive elements, reaching up to 80% of the total DNA content in some plants. In humans, about half of the genome is derived from repetitive elements, whereas protein coding sequences represent less than the 2%. The study of repetitive elements provides important clues on evolution and the underlying genetic mechanisms, but their functional impact on genome structure and regulation is still a matter of controversy. Moreover, genome-scale studies often overlook these elements, as they are intrinsically difficult to sequence and map. As an example, the ENCODE consortium flag paper claiming that 80% of the human genome was functional [9] disregarded the contribution of the repeats. Excellent reviews on the classification, structure and function of repeat elements have been published [35,31,15,5,3,26].

Alu is the most frequent repeat element in the human genome with more than one million copies per haploid genome. Alu elements are members of short interspersed repetitive elements (SINE). Being non autonomous retrotransposons, they produce RNA species during their life cycle and rely on other repeats to be retroprocessed. They are small ($\simeq$ 300 bp) and carry a PolIII promoter in their 5'. They harbor polyA elements and CpG domains. Their origin is the 7SL tRNA. They are flanked by short direct repeats [40].

Interestingly, Alus are not randomly distributed within the human genome, as they tend to accumulate in GC-rich regions [26] and participate in the architecture of the genome by delimiting the active/inactive domains and the epigenetic landscape [6] and gene regulation at different levels [3,4].

The advent of next generation sequencing technologies and their application to profile the genome and the epigenome of literally thousands of experimental settings offers a new opportunity to explore the structural and functional properties of Alus. Here we propose the use of ontologies to address this issue. Ontologies model the knowledge of realm while defining it in a formal manner [17] by providing a controlled vocabulary to refer explicitly to its subjects [1]. Indeed, they describe the inner properties of the system, such as the relationships between subjects. The annotated data can be stored in different exchangeable formats allowing semantically rich queries [21]. Finally, ontologies can be used for hypothesis evaluation [37].

* Corresponding author.
E-mail address: imallona@imppc.org (I. Mallona).

http://dx.doi.org/10.1016/j.jbi.2016.01.010
1532-0464/

The usage of ontologies is an emerging field in biology and biomedicine, although some controlled vocabularies are widely used. For instance, Sequence Ontology (SO) offers a hierarchy of concepts and relationships to be used to annotate genomic data; and Gene Ontology (GO) provide a set of terms to describe molecular functions, biological processes and cellular locations of genes and gene products. In this paper we report a biological ontology of human Alu repetitive elements covering their physical characteristics, their epigenetic status and the functional annotation of their nearby elements.

2. Materials and methods

The UCSC repository was queried through its MySQL public interface for Ensembl Genes, Repeat Masker, Gene Ontology, and Chromatin State Segmentation using Hidden Markov Model (HMM) from ENCODE/Broad [10]. Methylation data was retrieved from the Lister's whole genome bisulphite sequencing (WGBS) data [29]. A summary of the data origins is available as Fig. 1 and supplementary files S1 and S2.

GO Slim generic as provided by Open Biomedical Ontologies (OBO) [45] was downloaded from Gene Ontology Consortium [14]. Sequence Ontology [8] version 2.5.1 was retrieved from Sequence Ontology Consortium [42].

The OWL/RDF ontology was modelled with Protégé v4.0.2 [27] and populated with a set of custom-made bash and python scripts. Source code is available at Mallona, I. [30] under the GPL v2 terms.

DL Queries were run in a Fedora Core 14 Linux Workstation with Intel Xeon at 2.40 GHz processor and 16 GB of memory.

Statistical analysis were performed under the R environment v3.1.1. Genome-wide validations were performed using BedTools v2.19.1.

3. Approach

3.1. Ontology scope

During primates evolution, Alu elements were inserted at different evolutionary time frames. The majority of human Alus were incorporated before the divergence of human and non-human primates and are said to belong to old subfamilies [41], but others are restricted to the human lineage and some are still being amplified. As retrotransposition events are a source of genomic variation with enormous impact, we hypothesize that insertion permissiveness might be related to the genomic landscape as shaped by genetic elements and the epigenetic code. This trait is difficult to model as many features might shape the Alu insertion and selection dynamics. In our opinion, an integrative Alu knowledgebase may help to elucidate the functional implications of the Alu distribution along the genome. With this purpose in mind, we gathered structural and epigenetic properties of the human Alus.

Chromatin functional status has been described as the result of the crosstalk of epigenetic modifications (mainly histone modifications) [10], so we took the chromatin states of each Alu as a proxy to its putative functional properties. Given that Alu elements harbor active, protein-recruiting domain, we introduced sequence-based predictions of transcription factor binding sites (TFBS). Methylation status was also included as it is associated with functional repression. We took the Ensembl gene set, which includes protein-coding genes, non-coding RNAs and automatically-annotated pseudogenes, and assigned the closest one to each Alu regardless of the distance between them, as Alu tend to accumulate in GC-rich regions [26]. Finally, we also recovered from Ensembl the Gene Ontology annotation of these genes and gene products.

3.2. Modeling and formalization

Ontologies are commonly represented by using the Web Ontology Language (OWL). OWL uses formal semantics and represents them using RDF/XML-based schemata. The World Wide Web Consortium (W3C) endorses OWL and is a standard for ontology dissemination [33]; along with the Open Biomedical Ontologies (OBO) format, OWL is widely used in the biomedical field [19]. As a result, the ontology formalization produces a text file with a machine-readable syntax; moreover, the semantics is stated in such a manner that is also readable by computers.

As the repetitive elements are really abundant, the data annotated by the ontology, even regarding only the Alu elements, involve over a million instances. This fact implies that the usage of ontology viewers and reasoners might require a noticeable amount of computational resources for genome-wide queries. On the other hand, Alu elements are not randomly distributed along the genome, showing a noticeable heterogeneity between chromosomes [22]. Therefore, we splitted the Alu Ontology as a set of 24 OWL ontologies serialized in XML/RFDL, one for each canonical chromosome. As indicated at supplementary file S3, the resulting subontologies range from a couple of hundreds (chrY) to nearly half a million individuals (chr1), covering the Alu elements, genes, chromatin colors, etc. The ontologies are fully compatible with ontology viewers and formal reasoners like Pellet [20]. Splitting the information into chromosome-centered ontologies does not undermine whole-genome analysis, as the viewers and reasoners can serialize multiple ontologies at once; and allows the user to focus on a subset of chromosomes matching further requirements, such as gene content or autosomal nature.

Finally, we took advantage of the OBO initiative [20] mature ontologies to describe sequences and functional annotations,

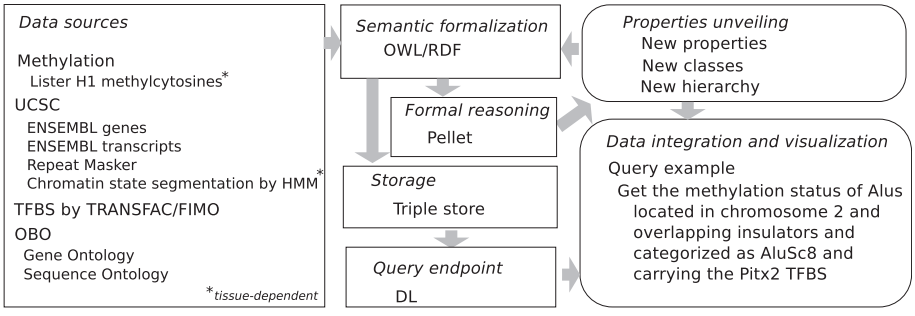

Fig. 1. The data sources include: UCSC's information on Ensembl Genes, Gene Ontology, RepeatMasker, Broad ChromHMM; Lister's data DNA methylomes at base resolution; and TRANSFAC-based FIMO transcription factor binding sites predictions. H1 hESC data was retrieved from Lister's and ChromHMM sources since methylation and chromatin states are cell-type-dependent. To perform the semantic formalization, OWL/RDF and SWRL (Semantic Web Rule Language) have been used.