Computational and Structural Biotechnology Journal 14 (2016) 69–77

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

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

A Graph Based Framework to Model Virus Integration Sites

Raffaele Fronza [a,*], Alessandro Vasciaveo [a,b], Alfredo Benso [b], Manfred Schmidt [a]

[a] Department of Translational Oncology, National Center for Tumor Diseases and German Cancer Research Center, Im Neuenheimer Feld 581, 69120 Heidelberg, Germany
[b] Department of Control and Computer Engineering, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

ARTICLE INFO

Article history:
Received 27 April 2015
Received in revised form 20 October 2015
Accepted 23 October 2015
Available online 30 November 2015

Keywords:
Gene therapy
Systems biology
Genomics
Insertional mutagenesis

ABSTRACT

With next generation sequencing thousands of virus and viral vector integration genome targets are now under investigation to uncover specific integration preferences and to define clusters of integration, termed common integration sites (CIS), that may allow to assess gene therapy safety or to detect disease related genomic features such as oncogenes.

Here, we addressed the challenge to: 1) define the notion of CIS on graph models, 2) demonstrate that the structure of CIS enters in the category of scale-free networks and 3) show that our network approach analyzes CIS dynamically in an integrated systems biology framework using the Retroviral Transposon Tagged Cancer Gene Database (RTCGD) as a testing dataset.

© 2016 Fronza et al.. Published by Elsevier B.V. on behalf of the Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Viral vector integration is a process exploited in gene therapy (GT) to correct defective cells of an individual and to drive the health status from the pathological condition to a normal one [1–6]. As consequence of this perturbation, i.e. if vectors integrate into cellular genome positions where the expression of an important gene is dysregulated, the affected cell may step from the primary illness state to a secondary state. Thus, insertional mutagenesis is a potential risk that may accompany vector integration events [7–11].

Therefore, large insertional mutagenesis screenings are used to assess the safety of the treatment in clinical GT, to design safer GT protocols and to discover new disease (i.e. cancer) candidate genes [12–16].

The central role in integration site (IS) analyses is given in the assessment of the genome-wide integration profile and the identification of integration clusters that could alter gene expression. The definition of these clusters or common integration sites (CIS) is not standardized and usually based on accumulation of IS that are unlikely to occur by chance and statistically significant different compared to a random *in silico* control. A regular interpretation (Standard Windows Method, SWM) is founded on the number of integrations in a predefined genomic window, that classifies CIS as follows: a) 2 IS are within 30 Kb or b) 3 IS within 50 Kb or c) 4 IS within 100 Kb or d) ≥5 IS within 200 Kb [17,18]. It is obvious that this historical definition can be used only as a first approximation for discovery of biologically (and clinically) relevant CIS, because the results are highly dependent from the size of the IS dataset [19]. Even if methods not constrained on a predefined set of fixed windows were released [20–23], the data importation together with CIS generation and analysis remain tedious tasks. The static tabular and text oriented nature of the CIS representation requires extensive processing steps that can involve custom programming, format exchanging and manual interpretation of the results. These computational difficulties reduce the analysis capability of standard life science labs and strictly rely on elevated bioinformatics skills.

We hypothesized that in the next generation sequencing area only systems biology approaches may be able to dissect biologically (and clinically) relevant CIS. Here, we developed a new CIS construction framework using an approach based on graphs. This approach has numerous advantages: 1) the resulting CIS are represented by networks, 2) graph theory can be used to infer characteristics and properties of the integration process (i.e. the node degree distribution of the networks can be linked to the randomness of the integration process, and 3) a large repertoire of IS can be imported and parsed without any prior constraint (except for the maximal distance between two IS).

More in detail, the graph model allows an easy structural organization of the annotations in the Gene Atmosphere (GA) (e.g. protein coding atmosphere, post-transcriptional atmosphere, etc.) by using different layers of node categories while performing the enrichment as described in Paragraph 2.6. The implementation of this model in software tools, which allow networks visualization, provides a broader overview of the data at a glance. An example of this feature is depicted in Fig. 3 where the CIS network is disposed on a plane by applying a force-directed layout. Furthermore, with the availability of Hi-C data, this

[*] Corresponding author.
E-mail address: raffaele.fronza@nct-heidelberg.de (R. Fronza).

http://dx.doi.org/10.1016/j.csbj.2015.10.006
2001-0370/© 2016 Fronza et al.. Published by Elsevier B.V. on behalf of the Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

framework is ready to embed relations among CIS in the spatial organization of the genome, enriching the modeling to a multidimensional level [24] (i.e. moving from a linear genomic vicinity modeling to a topological genomic modeling). Another interesting feature which emerges using this graph model is the capability of assessing biological properties by exploiting topological characteristics of the network. For example, the scale-free distribution of a set of CIS can be used to establish if the dataset under analysis contains genomic regions enriched in IS, an observation that is a prerequisite in order to properly recognize CIS.

Recent approaches to the identification of hot-spots try to take into account the size of the IS's dataset and the prior knowledge about vector integration preferences [23]. The implementation of this graph model on a normal computer machine could be greedy of computational resources when dealing with very huge IS datasets (i.e. millions of nodes). To our knowledge, there are no IS datasets in literature big enough that cannot be easily represented by our model. Enhanced with annotated genomic data, this model could be easily extended in order to drive the identification of CIS exploiting the information contained in the annotations. It is our intention to evaluate this possibility in future. One of the main differences between our model and the statistical frameworks used to the identification of CIS is that here the statistical method is applied after the CIS identification leaving to the user of the model the ability to give a biological meaning to the CIS (e.g. the CIS is not excluded a priori by the statistical method). With the complex annotation feature, as described in Paragraph 2.6, our model is able to perform a many-to-many mapping against genomic features (i.e. genes) and integration sites while other methods just perform a one-to-one mapping (i.e. one gene, one IS). In this way, our model provides a more refined granularity, when it is enhanced with complex annotation, allowing the simultaneous representation of different genomic features in the model (i.e. transcriptional elements, protein coding genes, etc.), that other models do not allow.

A Cytoscape [25] draft prototype plugin was developed to test the framework of this paper.

2. Results and Discussion

2.1. CIS Definition

First of all we define what a common integration site is. A set of n IS in the database is represented as the set of n vertices V of the graph G. Then, for each couple of vertices v_i and v_j ($i, j = 1, 2, \ldots, n\ i \neq j$) we add an edge e_{ij} if the distance between the corresponding IS is below a threshold T_H of 50 Kbp. A weight w_{ij} is associated with the edge e_{ij} and represents the distance between the corresponding IS. The default value of 50 Kbp was selected using the maximal influence window size where a causal relation is found between an insertion event and gene expression [22]. De Jong showed that the presence of viral integration is correlated with the local amount of gene expression and that 50 Kbp is an upper bound on which the presence of IS can be linked with gene expression. At the end of this process, we obtain the undirected weighted graph $G = (V, e)$ as abstract representation of all the distance relations in the IS dataset. The graph G is composed by a set of unconnected subgraphs (Connected Components, CC). Each CC is the natural graph representation of a CIS in which the order is represented by the number of vertices.

2.2. Integration Process and Node Degree Distribution

The non-random character of virus and viral vector integration suggests the existence of sub genomic regions that are preferentially targeted. As many complex biological systems where many components interact together, also the viral integration process derives from intricate functional interactions that involve viral and host proteins/DNA. The behaviors of complex systems are captured by a characteristic of the network that is called scale-free property [26–28]. This property depends on the distribution of the nodes degree. The node degree is the number of edges that connect a node with the neighbors. The degree distributions of several networks follow a power law, precisely defined with the functional $d(k) = ak^{-\gamma}$, where $d(k)$ is the degree distribution, k= 0,1,2,… is the node degree, a is the normalization constant and γ is the degree exponent. In scale-free network the exponent is usually less than three ($\gamma < 3$), whereas in random networks $\gamma \geq 3$.

To prove that the mechanism of viral CIS or hot-spot (HS) formation is embedded via scale-free property into the network representation, we developed a series of synthetic transfection experiments that consisted of placing a fixed number of integrations on human genome carrying a random number of artificial hot-spots. The integrations were divided in two subsets: 1) IS placed on a simulated genome with hot-spots (IS_{SYN}) and 2) IS randomly placed on a genome without hot-spots (IS_{RAND}). The scale-free property of CIS networks found in IS_{RAND} and IS_{SYN} was then verified using the Cytoscape “Network Analysis” plugin.

We further verified the presence of a HS driven mechanism on six datasets: five in which we expected a scale free behavior (LV [1], HIV [29], GV1 [2], GV2 [16] and RTCGD [12]); and one from an adeno-associated viral (AAV) vector study [30] where we expected a random integration profile. In Fig. 1 the degree distributions of the groups of the experimental IS sets are plotted. The richness in integration sites of the datasets is: ~1000 IS_{SYN} (g), ~15,000 IS_{RAND} (e), ~4000 IS_{LV1} (b), ~2000 IS_{AAV} (a), ~35,000 IS_{HIV} (d), ~15,000 IS_{GV1} (h), ~800 IS_{GV2} (c), and ~8800 IS_{RTCGD} (f). All the experimental and synthetic sets, except for the AAV and RAND set, have a log–log degree distribution that follows a power law with gamma exponent $\gamma < 3$. Only two datasets, the random dataset IS_{RAND} ($\gamma = 3.6$) and IS_{AAV} ($\gamma = 4.8$) have no scale-free degree distribution. This last finding is in line with our and other published studies that did not attribute to AAV any HS driven integration pattern [30,31]. From a practical point of view and as a first result of our graph modeling, the node degree distribution in a network that represent integration events indicate the presence of an accumulation process driven by genomic hotspots.

Réka [26] demonstrated that complex systems that display a high degree of error tolerance (robustness) are represented by scale-free networks. An incomplete IS dataset can be seen as the result of a process that remove IS from the complete basin of integrations present in a sample, due to unavoidable experimental subsampling. Recalling the robustness property for scale free networks we can prove that genomic hot-spots are identified even within an incomplete set of experimental IS.

2.3. General Structure of the CIS Pool and RTCGD Dataset

The Retroviral Transposon tagged Cancer Gene Database (RTCGD; http://variation.osu.edu/rtcgd/, [12]) was used as test case for our graph model.

The RTCGD dataset has been first analyzed in order to compare two general CIS properties, the order and the dimension of the 10 biggest CIS identified by our framework and the SWM. Fig. 2(A) shows the general structure and shape of all the CIS with order bigger than 9 as they appear analyzing RTCGD integration.

5110 IS are selected by the CIS construction tool as belonging to reputed CIS and 4035 compose CIS with p-value < 0.05 (see Appendix A Table 1 in [41]). How the p-value is computed per CIS is explained in the Paragraph 3.6. The CIS order goes from 2 to 82 (in Fig. 2(A) CIS from order 2 to order 8 are not shown). RTCGD data contains 2910 IS and the CIS falls in the same range order. No statistical model is applied in order to test the CIS significance.

In the 10 biggest CIS the order and dimension of 3 of them (*myc*, *ahi* and *rasgrp1*) were returned identical by the analysis performed using our model and RTCGD and other 4 CIS (*gif1*, *lvis1*, *pim1* and *notch1*) were comparable (difference in the order is less than 10; see Table 1).

Download English Version:

https://daneshyari.com/en/article/2079087

Download Persian Version:

https://daneshyari.com/article/2079087

Daneshyari.com