Original Research Article

POWER-LAW BEHAVIOR OF ALTERNATIVE SPLICING OF EXONS IN HUMAN TRANSCRIPTOME

.

ABSTRACT

Aims: To establish the common rule of combinatorics of exons during RNA splicing.

Study design: Inferring of the plausible statistical model of combinatorics of exons during RNA splicing from annotated models of human genes.

Place and Duration of Study: Department of Genetics (Belarusian State University), Proteome and Genome Research Unit (Luxembourg Institute of Health), Department of Genetics (Lomonosov Moscow State University) and Moscow Center of Experimental Embryology and Reproductive Biotechnologies, between January 2017 and July 2019.

Methodology: We used human mRNA and EST sequences from GenBank and linear models of the human genes from Ensembl, AceView, ECgene, NCBI RefSeq, UCSC Genome Browser and VEGA to calculate combinatorial index of human exons. We inferred the most plausible statistical model describing the distribution of combinatorial index of human exons using Clauset's mathematical formalism. Predictors of the value of combinatorial index and the functional outcomes of predefined behavior of exons during splicing were also determined.

Results: Power-law is the most plausible statistical model describing the combinatorics of exons during RNA splicing. The combinatorial index of human exons is defined more than 90% by the 138 features that have different importance. The most important of these features are abundance of exon in transcripts, strength of splice sites, rank of exon in transcripts and type of exon. Analysis of the marginal effects shows that the same feature, but its different values have an unequal influence on the combinatorial index of human exons. Power-law behavior of exons during RNA splicing pre-determines structural diversity of transcripts, low sensitivity of splicing process to random perturbations and its high vulnerability to manipulation with highly combinative exons.

Conclusion: Exons widely involved in alternative splicing are part of the common power-law phenomenon in human cells. The power-law behavior of exons during RNA splicing gives the unique characteristics to human genes.

Keywords: human exons, RNA splicing, combinatorics, statistical modeling, power-law, predictors, functional outcomes.

1. INTRODUCTION

Alternative splicing is a unique process of unzipping genetic information archived in the nucleotide sequence of the gene. This is widespread phenomenon in human cells and tissues. It was estimated that 92-94% of human genes produce appreciable levels of two or more distinct populations of RNA isoforms [1]. Alternative isoforms of transcripts may appear at the level of single cells or population of cells of the same type [2,3], different tissues of the same individual or the same tissue but in different individuals [1,4] and at different stages of human development [5,6].

The main outcome of the alternative splicing is significant expansion of the complexity of the transcriptome with limited starting material. Through this process, a single gene can produce a wide variety of RNA molecules. These molecules may be translated into a variety of structurally and functionally distinct proteins [7]. Moreover, some of these molecules can be noncoding and play regulatory role [8]. The set of such diverse products of the same gene often forms a sub-network, which is tightly integrated into global cellular regulatory network and provides flexibility in adaptation and functioning of the cell [9,10].

Results of studies in recent years, obtained by high-throughput OMICS-technologies, have substantially expanded our understanding of alternative splicing and its biological role. These results also suggest that we see yet only the tip of the iceberg of the entire transcriptome complexity of a cell. The ever-growing set of empirical data in this area requires the elucidation of common principles or rules by which the transcriptome of a cell is formed and functions. We hope that through the knowledge of such rules the further progress in this area will be achieved. Over the last decade there have already been some successes in this direction. In particular, some basic properties of the “splicing code” were disclosed [11,12]. However, this knowledge is not enough for a full understanding of the rules of the combinatorics in exons and systemic factors that drive and control the splicing process.

Analysis of the human transcriptome shows that the number of various splicing events involving exon may vary significantly for different exons. There exists a large set of exons that are involved only in single splicing events. On the other hand, the human transcriptome contains a limited number of exons, which take part in many different splicing events. In this regard, we set a goal to figure out whether there are any principles of local combinatorics of exons and if so, where and how it is predestined.

In the context of this article, the term “local combinatorics of exons” refers to pairwise splicing events involving given exon during formation of different RNA isoforms. In contrast, the term “extended combinatorics of exons” refers to splicing of different exons during formation of given RNA isoform. For the purposes of this article, we also use term “combinatorial properties of exons” which means a set of properties of exons which predetermine the diversity of their alternative splicing. Moreover, we introduced the “exon’s combinatorial index” (ECI) which is simple equivalent to a topological index “node degree” of graph theory. In further analytical work we used both a total-degree of exon (“total” exon’s combinatorial index or simply total-ECI) as well as its decomposed variant (separately ingoing and outgoing degrees, or in-ECI and out-ECI, respectively). Herewith total-ECI means the sum of all (ingoing as well as outgoing) unique splicing events that involved exon. The terms in- and out-ECI refer to the sum of all ingoing or outgoing unique splicing events that involved exon, respectively.

2. MATERIAL AND METHODS / EXPERIMENTAL DETAILS / METHODOLOGY

For the purposes of this paper, human mRNA and ESTs sequences deposited in GenBank were downloaded via FTP-server of the UCSC Genome Browser. These sequences were aligned by BLAT [13] against GRCh38/hg38 reference assembly of the human genome and were subjected to four levels of filtration: records with only one aligned block, mismatches, exons and/or introns length below the 5th quantile of distribution (23 and 88 nucleotides in length for exons and introns, respectively) were deleted. The resulting collection of sequences we called Dataset 1 with 1093522 records. Additionally, we trimmed terminal exons of sequences from Dataset 1 and formed Dataset 2 with 627733 records.

Statistical modeling and statistical analysis of the above mentioned datasets was carried out using R programming language. The key steps of this analysis are described in the relevant sections of RESULTS.

3. RESULTS AND DISCUSSION

3.1 RESULTS

3.1.1 Power-law behavior of local combinatorics of human exons

Our work was based on the analysis of the seven data sets: full list of human mRNA and EST sequences from GenBank [14] and linear models of the human genes from Ensembl [15], AceView [16], ECgene [17], NCBI RefSeq [18], UCSC Genome Browser [13] and VEGA [19]. For a more compact representation and future use these data were converted into exon graphs. Each of exon graphs is presented by a set of exons (vertices or nodes of a graph) connected to each other via set of splicing events (edges or links of a graph) [20]. Such a graph is directed acyclic graph in the sense that the exons present in any mature transcript of gene are retained in the correct 5’ to 3’ linear order and reverse edges are prohibited.

The results of topological analysis of reconstructed exon graphs suggest that values of the ECI follow a power-law distribution with heavy right tail: the vast majority of exons have low value of the ECI, while a small part of the exons are characterized by very high value of this index (Fig. 1A). However, power-law is only one of the members of broad family of distributions with heavy right tails [21]. In addition, the selection of the correct statistical model for that kind of data is not a trivial task because of the incompleteness of the empirical biological data and their high variability (especially in the area of heavy tail). Therefore, we had to use three-step approach based on the mathematical formalism developed by Clauset A. et al. [22,23] to find appropriate statistical model and to check our preliminary supposition.

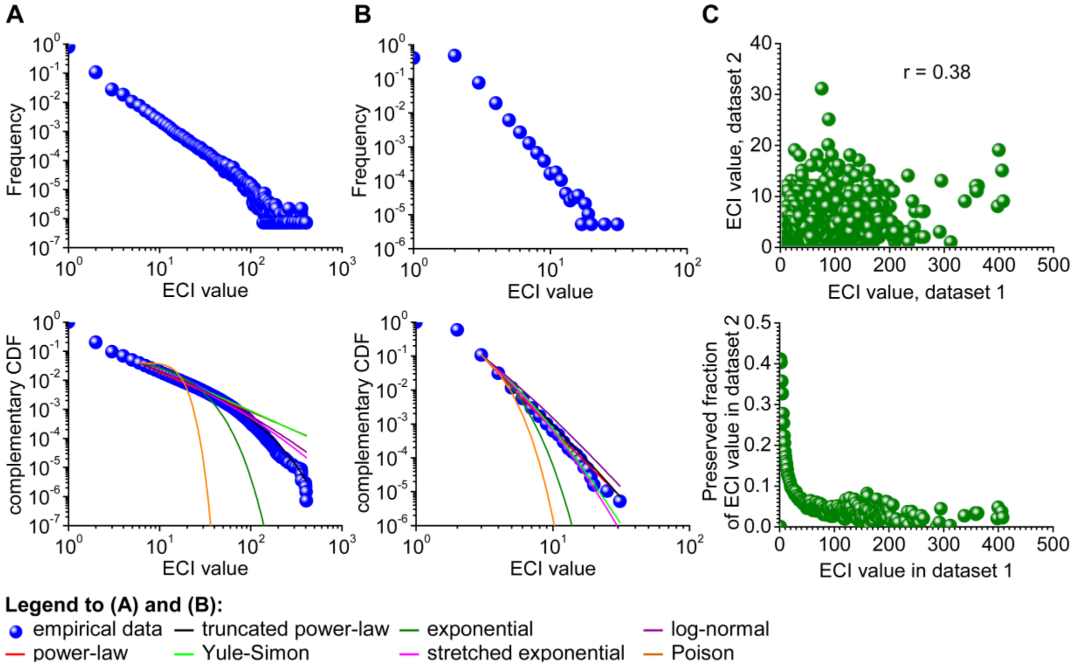

Fig. 1. Removing of uncertainty with the terminal exons leads to a clear manifestation of the power-law component in the human transcriptome. (A) Frequency (upper panel) and complementary CDF (lower panel) plots of the ECI values distribution from whole set of human transcripts and exons. Truncated power-law with an exponential cut-off is the best statistical model for this empirical distribution among the set of competitive models of distributions with heavy right tail. (B) Frequency (upper panel) and complementary CDF (lower panel) plots of the ECI values distribution after removing of the terminal exons from transcripts and data reanalyzing. For these transformed data, there exists a clear superiority of the power-law model as compared to other statistical models. (C) Removing of the terminal exons from transcripts leads to significant change in ECI values of exons (upper panel). Herewith the exons with initially high values of the ECI underwent the most profound changes (lower panel).

First, we rejected those statistical models that clearly did not fit the empirical distributions and chose five closest models: power-law distribution, truncated power-law distribution (or power-law with exponential cut-off), exponential distribution, stretched exponential distribution (or complementary cumulative Weibull distribution) and log-normal distribution. Next, selected statistical models were fitted to the empirical distributions according to “x_{min}” paradigm [22]: only heavy tail of empirical distribution was the subject of our attention because it contains the most outstanding sub-set of values of the distribution. Finally, Kolmogorov-Smirnov test and log-likelihood ratio test were used to assessment of the plausibility of the statistical hypothesis and for direct comparison of alternative statistical models [22,24,25].

The above mentioned approach allowed us to detect several features of empirical distributions. First, the results of our statistical modeling permit to postulate that values of the ECI of human exons follow a truncated form of the power-law with exponential cut-off (Fig. 1B, Table 1). Herewith exponential component of distribution may be substantially reduced by filtering out of the 5’- and 3’-terminal exons (with in-ECI = 0 and out-ECI = 0, respectively) together with edges and first neighbours from exon graphs (data not shown). However, it should be noted that for the three data sets (gene models from AceView, NCBI RefSeq and VEGA) there remains uncertainty when truncated power-law is comparing with stretched exponential or log-normal models: log-likelihood ratio test does not favor one model over the other and only Kolmogorov-Smirnov test gives a slight preference for the power-law with exponential cut-off. Second, the beginning of the heavy tail (lower bound) for different data sets ranges from 5 to 15. Third, in the frame of a truncated power-law model the scaling parameter α lies within the range from 2.378 to 7.248 and rate parameter λ falls into the broad range from 5.713×10^{-9} to 1.77×10^{-1} for different data sets.

Table 1. Log-likelihood ration test (A) and statistical tests on plausibility (B) confirm the presence of a power-law component in the human transcriptome

A

Dataset	Basic model	LLR test	Competing statistical model						
			Power-law	Truncated power-law	Yule-Simon	Exponential	Stretched exponential	Log-normal	Poison
Dataset 1	Power-law	R	–	-446.6	-20.6	41.22	-17.7	-19.6	51.03
		p	–	3.0e-196	2.3e-94	0.0	2.8e-70	3.5e-85	0.0
	Truncated power-law	R	446.6	–	30.5	47.34	20.6	30.9	51.23
		p	1	–	0.0	0.0	0.0	0.0	0.0
Dataset 2	Power-law	R	–	-0.8	4.2	11.12	3.6	-0.6	13.64
		p	–	0.2	2.3e-05	0.0	3.4e-04	0.5	0.0
	Truncated power-law	R	0.8	–	4.8	11.37	4.0	0.8	13.76
		p	1	–	2.0e-06	0.0	5.2e-05	0.4	0.0

B

Dataset	Test	Competing statistical model						
		Power-law	Truncated power-law	Yule-Simon	Exponential	Stretched exponential	Log-normal	Poison
Dataset 1	AIC	345205.6	344314.5	345304.3	363155.6	344595.3	344593.7	897718.6
	BIC	345214.5	344332.3	345313.2	363164.5	344613.1	344611.5	897727.5
	KS distance	0.01715	0.02909	0.01553	0.17865	0.03159	0.03015	0.23341
Dataset 2	AIC	39064.8	39065.2	39158.1	40201.8	39140.2	39065.6	42408.7
	BIC	39072.7	39081.1	39166.0	40209.7	39156.1	39081.4	42416.6
	KS distance	0.00188	0.04049	0.01658	0.05476	0.04989	0.05232	0.08236

AIC - Akaike information criterion; BIC - Schwarz Bayesian criterion; KS distance - Kolmogorov-Smirnov distance; LLR - log-likelihood ration.

In fact, observed empirical distributions with power-law component cannot be produced by random attachment of exons during splicing step of gene expression (Fig. 1C). On the other hand, this class of distributions can be easy generated by preferential attachment process [26]. In a frame of preferential attachment model, the different exons have different attractiveness to connect other exons. The results of our modeling indicate that artificial data can be fitted to any of our empirical data set by varying of power parameter α (Fig. 1C). However, exact nature of the observed difference of exon attractiveness is not taken into account yet. This question is investigated below.

<u>3.1.2 Predictors of the value of ECIs in human transcriptome</u>

It seems to be true that almost all (if not all) processes in the cell are controlled by complex multilevel mechanisms [27,28] and splicing process is unlikely to be the exception. So, we hypothesized that there isn't one or some features (predictors) which determine the value of ECI but this index is determined by multiple predictors. To find out of such predictors, we assembled a compendium of 22114 features of the target exons (ECIs of which are the objects of analysis) and their upstream and downstream first neighboring exons and adjacent introns. This compendium included seven classes of features: abundance of exons in transcripts, connectivity of first neighboring exons, sequence's features of target exons, sequence's features of first neighboring exons, length of adjacent introns, rank of exons in transcripts and functional type of target exons (Fig. 2). We took the Ensembl data set as the main object of research, and the other sets of data were used for cross-validation of the results of the study. The relationship between a feature and ECI as well as contribution of feature to the value of the ECI was assessed by pairwise Spearman's rank correlation coefficient and by data mining. Data mining was based on machine leaning by regression random forests. This learning algorithm was chosen by comparison with the two other algorithms (lasso regression and generalized boosting regression) as the most suitable for the task.

In fact, for any gene of interest the presence of multiple exons, wide spectrum of produced transcripts and different abundance of exons in these transcripts are minimal prerequisites to development of the non-equivalency of ECI value of exons. Most interesting is the situation with the abundance of exons in transcripts. On the one hand, there exists a high positive correlation between the value of ECIs and abundance of target exons in transcripts and data mining revealed the high importance of this predictor.

However the converse statement is not true: not all exons widely represented in transcripts are characterized by high combinatorial capacity. Moreover, profiles of the marginal effects point out to the unequal importance of the different values of this feature in the determination of the ECI. On the other hand, the abundance itself is determined by multiple factors as was shown early [11].

Second class of features (connectivity of first neighboring exons) has a moderate or little effect on the value of ECI as was revealed by correlation analysis and data mining by random forests. In general, exon graphs are disassortative (with value of disassortativity index up to -0.247): the exons with high values of ECI prefer to attach the exons with low value and vice versa.

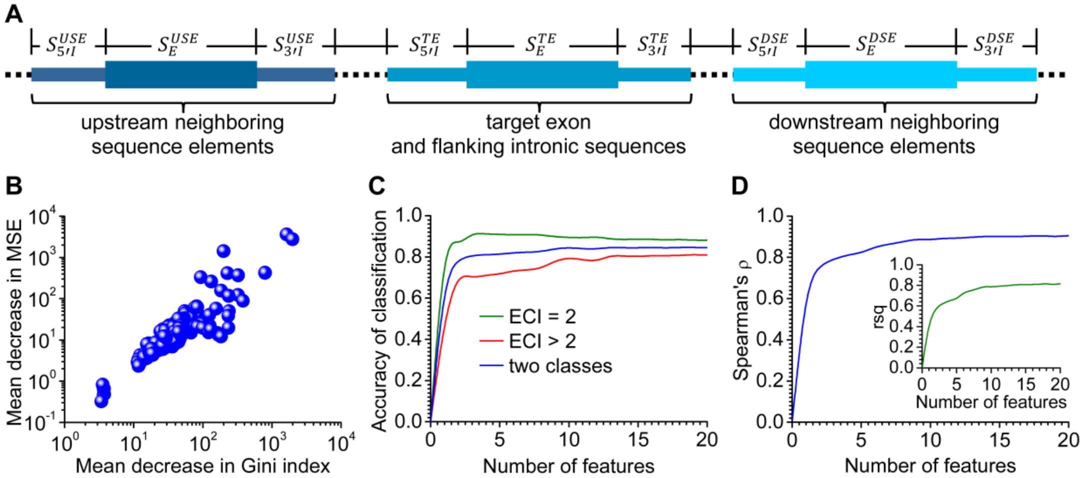

Fig. 2. Small subset of features from exons and flanking introns may determine the value of the ECI in tissue-independent fashion. (A) Features of the five different classes (sequence features, sequence-related features, functional features, epigenetic features and features related to structure of gene) were extracted from three classes of the genomic/RNA elements: target exon and its flanking intronic sequences, upstream first neighboring exons and their flanking intronic sequences and downstream first neighboring exons and corresponding flanking intronic sequences. (B) Importance of these features for the prediction of the ECI value was inferred from random forest classification by Gini index and from random forest-based multiple nonlinear regression by mean square error. (C) Random forest-based classification shows that the maximum accuracy in prediction of the ECI achieved by using no more than top-20 features. (D) Top-20 features permits to achieve a maximum in explanation of the ECI variance and high accuracy in prediction of the ECI values by random forest-based multiple nonlinear regression.

In our compendium, classes sequence’s features of target and first neighboring exons include length of exons, linear density of minimal free energy of exons, strength of the 5‘ and 3’ splice sites, regional counts of the short motifs (1-3 nucleotides), frequency of the known splicing enhancer and silencer motifs and count of the new predicted motifs associated with high combinatorial exons (MAHCE). One of our approaches in identification of MAHCE was correlation-based approach. This approach allowed us to cluster a plenty of sequences discarding irrelevant oligomers. In turn, the decrease in the number of the unique sequences led to a reduction in the dimension of the space of variables and allowed to use of machine learning algorithms for the identification of important predictors.

Despite the large number of studied structural features as well as cis-elements, carefully carried out analysis and the desire to identify strong predictors we didn't find such. The most significant was only a relationship between the strength of the splice sites and the value of the ECI. It should be noted that there is a clear cross-relationship between the value of the in- or out-ECI and the strength of the respective splice site as well as increased correlation at use of the total (overall) score of the strength of splice sites. Herewith the profiles of marginal effects of the strength of splice sites of the target and first neighboring exons are completely different. The remaining cis-elements have a moderate or little effect on the value of ECI (wherein both the total-ECI and the in- and out-ECI) as was revealed by correlation analysis and data mining by random forests. A similar situation was observed in the case of parameters such as length of exons and stability of their secondary structure.

From next class of features, we just studied the minimal, maximal and mean length of the adjacent introns in relation to value of the ECI. These predictors were successfully selected by feature selection algorithm as important for prediction of the ECI, but with low contribution (no more than 9.33% in the average increase in squared out-of-bag residuals), that is agree with results of correlation analysis.

One of the most informative and important features in prediction of the value of ECI was position of exon in transcripts. We used two different approaches to find out of this metric: averaged short (total-, in- and/or out-) distance of exon to another exons in exon graph and direct determination of averaged exon position (exon rank) relative to start (5’-end) and/or to end (3’-end) in the transcripts of gene of interest. Simple ratio of these two parameters gives us the position of the exon relative to center of transcripts (the centrality of exon position in transcripts). The centrality of exon position was expressed either in relative units (where 1 is a relative center of transcripts, which include the exon of interest) or in absolute distance (measured in the number of exons) from the center of transcripts. In last case, zero position means the center of transcripts which include the exon of interest. If exon located left of center (closer to the 5'-end of transcripts), its position has a negative sign, otherwise a positive sign.

All exons with high values of total-ECI show nonrandom positional distribution and tend to occupy central position in transcripts (Fig. 3). Permutation of the values of predictors from this class of features increases more than 30% of squared out-of-bag residuals in a case of target exons, and more than 20% in a case of first neighboring exons. Again, as was mentioned for other features, profiles of the marginal effects point out to the unequal importance of the different values of this predictor in the determination of the ECI with clear transition point near to the central position.

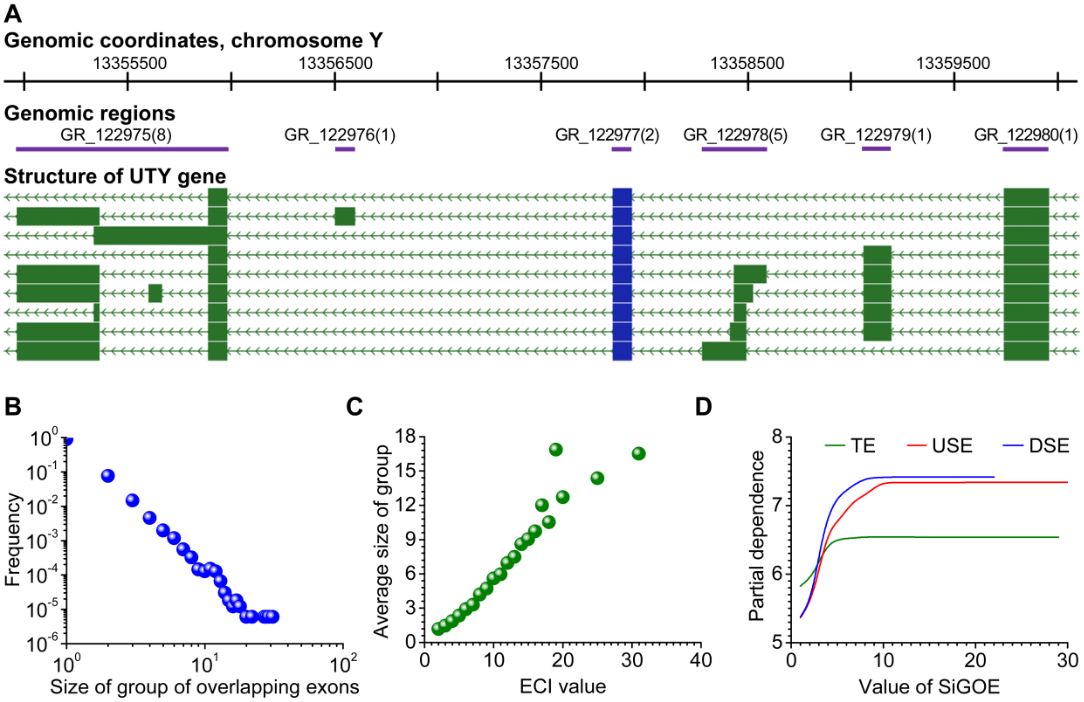

Fig. 3. The most important predictor of the ECI value of target exon is a structure of its upstream and/or downstream neighboring exon-coding genomic regions. (A) Genomic structure of the small part of UTY gene is used as example. Exon-coding genomic regions that belong to this part of gene are depicted as GR. The number of exons originated from each genomic region (size of group of overlapping exons, or SiGOE) is indicated in parentheses. Target exon is colored in dark blue. This exon has ECI = 10 and 70% of its splicing events happening with exons from two genomic regions GR_122975 and GR_122978. (B) Empirical distribution of sizes of groups of overlapping exons from human genome follows a power-law function. (C) The relationship between the ECI value of target exon and the average size of its neighboring upstream and/or downstream groups of overlapping exons. This relationship is close to the linear. (D) Marginal effects of size of neighboring upstream (USE) and downstream (DSE) groups of overlapping exons on the ECI value of target exon. The effect of size of overlapping group that owns the exon (TE) is also shown.

Finally, it was studied the possible influence of evolutionary conservatism of exon and its functional type on the value of the ECI. We found that group of exons with high values of ECIs (≥10) characterized by an average level of conservatism. However there observed the low positive correlation between the conservatism of exon and value of its ECI and this variable was not selected as important by feature selection algorithm. As for functional type of exon, ANOVA and data mining by random forests confirm importance of this feature in determination of the value of ECI. In particular, there is a clear link between the multifunctionality (when exon is annotated as a multitype exon) of exon and high value of ECI and permutation of the values of this feature increases more than 30% of squared out-of-bag residuals.

In summary, as it was originally supposed among the features we studied is no single predictor or a small group of predictors that would entirely determine the value of the ECI of human exons. On the contrary, the value of the ECI is defined more than 90% by the multidimensional space of predictors (138 features in a case of total-ECI of Ensembl exons) that have different importance. The most important of these predictors are abundance of exon in transcripts, strength of splice sites, rank of exon in transcripts and type of exon. Furthermore, analysis of the marginal effects shows that even the same predictor, but its different values have an unequal influence on the ECI.

3.1.3 Exon graphs with power-law structure: the functional outcomes

To establish the biological significance of power-law structure of the exon graphs, we compared this type of graphs with artificial full exon graphs on three criteria: diversity of the generated transcripts, flexibility of the alternative splicing and robustness to the random perturbations. In our modeling, we used sub-set of top-100 exon graphs (in terms of the number of vertices) from Ensembl-based human transcriptome exon graph. These empirical exon graphs (which we called power-law exon graphs) have topology with power-law component while our artificial full exon graphs do not have this component in the distribution of the ECI.

As was expected, full exon graphs capable to produce a significantly more diverse transcripts than exon graphs with power-law component. For example, directed walk along the tree of exon graphs shows that full exon graphs generate in 6.6 fold more different transcripts than power-law exon graphs and in 210.4 fold more than it was experimentally verified ($p = 1\times10^{-16}$). However, the length of such transcripts is clearly smaller than the length of transcripts generated by power-law exon graphs as well as empirical transcripts. In addition, a full crawl of the power-law exon graphs shows that they have great hidden potential to generate a variety of transcripts, a superior variety of known transcripts in 1472.2 fold. Herewith this potential can be seen not only in the structural diversity, but in the ability to generate long transcripts: there are clearly visible two distant peaks compared with the empirical data.

Second one in our interest was a flexibility of the alternative splicing with different types of exon graphs. We modeled the situation when any fraction of the ranked exons purposefully skipped or included in mature transcripts by splicing system. Power-law exon graphs are extremely sensitive to manipulation with top-ranked exons: active involving or skipping even a small fraction of these exons into splicing process may substantially change the possibilities for the formation of a variety of transcripts. At the same time, full exon graphs do not have such flexibility. And this applies to both the structural diversity of transcripts, and a variety of lengths of transcripts.

Finally, we modeled the effect of random perturbations on the different types of exon graphs and tested their ability to withstand such perturbations. Exact physical nature of random perturbations may be different, for example, it can be accidental loss of exon(-s) because of deletion at genomic DNA level or inclusion failure of exon(-s) into mature RNA because of mutations of the splicing cis-regulatory elements. The results of our modeling indicate that full exon graphs are significantly more robust to random attacks than power-law exon graphs. This difference is most clearly seen in the case where the robustness is estimated to change the length of the generated transcripts.

3.2 DISCUSSION

Power-law distributions appear in an enormous variety of fundamentally different complex systems: from engineering to biological and social systems [29,21]. Biological systems as the most complex systems are particularly rich in this phenomenon which is manifested at all levels of the organization of living organisms, from the molecular to the ecosystem level [30,29,31]. Therefore, we did not have a very surprise that the combinatorial properties of human exons are subject to the same law. More exciting here is the question why this phenomenon is so common in biological systems. Perhaps this is due to the fact which unique properties acquire of the system, when the distribution of some of its parameters obeys power-law.

The first of these properties is scale-free of distribution [21]. We have seen this property in our data sets. For example, random sampling of GenBank data didn't change the ratio between the numbers of exons with high and low values of ECI or randomly sampled sub-set of Ensembl data didn't change the form of distribution (data not shown). The main outcome of scale-free is the scalability of the system without losing its characteristics. So, we may speculate that power-law component permits to adaptively scale up or scale down of the transcriptome in individual human cells in response to environmental conditions without change the critical system parameters.

Second one is adaptive flexibility of system with power-law component. It was shown in the model and experimental studies that power-law distribution of any systemic parameter is a sign that the system is in the vicinity of phase transition or critical point [21,32]. Being close to the critical point, the system can be quickly reconstructed and can adapt to changing environmental conditions [33]. Our results are consistent with these ideas. In particular, our modeling shows that actively involving into splicing process or skipping even of small fraction of the exons with high value of ECI may substantially to change the possibilities for the formation of a variety of transcripts. Moreover, the theoretically possible diversity calculated on the basis of the structure of exon graphs is much greater than the experimentally confirmed diversity of the transcripts in human transcriptome. Consequently, human genes have significant hidden potential to produce different RNA molecules. Of course, it may be also due to incomplete of empirical data or the existence of obscure limits on the formation of some variants of the transcripts and this can be the topic of a special study.

The third property of systems with power-law component is their sensitivity to accidental damages. Despite the fact that the complex biological systems can stably operate in various conditions they yet are fragile [34]. We see this in our model studies, the results of which show that the exon graphs are sensitive to random perturbations. To improve the robustness of such fragile living systems, the nature has taken the path of increasing diversity and complexity of regulatory mechanisms [27,28]. Splicing process is controlled by a variety of mechanisms based on a redundant inner diversity of the cis- and trans-regulatory factors. These factors are organized in a spatially and functionally distributed intracellular network with multiple positive and negative forward and feedback reverse regulatory circuits (Braunschweig U. et al., 2013). We believe that because of this, we could not find one or more predictors which completely would determine the value of the ECI. Instead of that, we found more than one hundred features which are involved in determining the value of ECI. A similar situation exists with other properties of splicing, for example, there was identified more than two hundreds predictors that determine the inclusion or exclusion of exon in/from the mature transcripts in the different type of human tissues [11].

In light of the problem of predictors, the most interesting and unexpected for the ECI were such features as the position of the exon in transcripts and the functional type of exon that requires special attention. In fact, the high-combinatorial exons are not only “hot points” of alternative splicing but they prefer to be located near the alternative transcription start and termination sites (but these exons usually are not 5’- or 3’-terminal exons). Exon 8b of human RUNX1T1 gene is a typical example. And our metric “centrality of exon position in transcripts” reflects only average position of exon in transcripts, which include the exon of interest. However, this feature is highly informative. Moreover, we also believe that highly combinative exons due to the specific of their location are usually multifunctional (these exons can be 5’UTR, 5’UTR/CDS, CDS, CDS/3’UTR and/or 3’UTR exon depends on transcript) and characterized by a middle level of conservatism.

4. CONCLUSION

In general, our results confirm the existence of the “exons-switches” of alternative splicing [1]. But we have obtained substantial refinements in this concept. In particularly, we showed that the “exons-switches” are part of common power-law phenomenon in human cells. We also found that the combinatorial properties of human exons are defined more than 90% by the multidimensional space of predictors that have different importance and different profiles of the marginal effects. Finally, we found that the power-law component gives the unique characteristics of the human genes.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456 (7221):470-476. doi:10.1038/nature07509
2. Gurskaya NG, Staroverov DB, Zhang L, Fradkov AF, Markina NM, Pereverzev AP, Lukyanov KA (2012) Analysis of alternative splicing of cassette exons at single-cell level using two fluorescent proteins. Nucleic acids research 40 (8):e57. doi:10.1093/nar/gkr1314
3. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, Wold BJ (2014) From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome research 24 (3):496-510. doi:10.1101/gr.161034.113
4. Castle JC, Zhang C, Shah JK, Kulkarni AV, Kalsotra A, Cooper TA, Johnson JM (2008) Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nature genetics 40 (12):1416-1425. doi:10.1038/ng.264
5. Gamazon ER, Stranger BE (2014) Genomics of alternative splicing: evolution, development and pathophysiology. Human genetics 133 (6):679-687. doi:10.1007/s00439-013-1411-3
6. Mazin P, Xiong J, Liu X, Yan Z, Zhang X, Li M, He L, Somel M, Yuan Y, Phoebe Chen YP, Li N, Hu Y, Fu N, Ning Z, Zeng R, Yang H, Chen W, Gelfand M, Khaitovich P (2013) Widespread splicing changes in human brain development and aging. Molecular systems biology 9:633. doi:10.1038/msb.2012.67
7. Nilsen TW, Graveley BR (2010) Expansion of the eukaryotic proteome by alternative splicing. Nature 463 (7280):457-463. doi:10.1038/nature08909

8. Rosa A, Brivanlou AH (2013) Regulatory non-coding RNAs in pluripotent stem cells. International journal of molecular sciences 14 (7):14346-14373. doi:10.3390/ijms140714346

9. Biamonti G, Caceres JF (2009) Cellular stress and RNA splicing. Trends in biochemical sciences 34 (3):146-153. doi:10.1016/j.tibs.2008.11.004

10. Chandler DS, Singh RK, Caldwell LC, Bitler JL, Lozano G (2006) Genotoxic stress induces coordinately regulated alternative splicing of the p53 modulators MDM2 and MDM4. Cancer research 66 (19):9502-9508. doi:10.1158/0008-5472.CAN-05-4271

11. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ (2010) Deciphering the splicing code. Nature 465 (7294):53-59. doi:10.1038/nature09000

12. Matlin AJ, Clark F, Smith CW (2005) Understanding alternative splicing: towards a cellular code. Nature reviews Molecular cell biology 6 (5):386-398. doi:10.1038/nrm1645

13. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hinrichs AS, Learned K, Lee BT, Li CH, Raney BJ, Rhead B, Rosenbloom KR, Sloan CA, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ (2014) The UCSC Genome Browser database: 2014 update. Nucleic acids research 42 (Database issue):D764-770. doi:10.1093/nar/gkt1168

14. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2014) GenBank. Nucleic acids research 42 (Database issue):D32-37. doi:10.1093/nar/gkt1030

15. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Giron CG, Gordon L, Hourlier T, Hunt S, Johnson N, Juettemann T, Kahari AK, Keenan S, Kulesha E, Martin FJ, Maurel T, McLaren WM, Murphy DN, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ruffier M, Sheppard D, Taylor K, Thormann A, Trevanion SJ, Vullo A, Wilder SP, Wilson M, Zadissa A, Aken BL, Birney E, Cunningham F, Harrow J, Herrero J, Hubbard TJ, Kinsella R, Muffato M, Parker A, Spudich G, Yates A, Zerbino DR, Searle SM (2014) Ensembl 2014. Nucleic acids research 42 (Database issue):D749-755. doi:10.1093/nar/gkt1196

16. Thierry-Mieg D, Thierry-Mieg J (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome biology 7 Suppl 1:S12 11-14. doi:10.1186/gb-2006-7-s1-s12

17. Kim P, Kim N, Lee Y, Kim B, Shin Y, Lee S (2005) ECgene: genome annotation for alternative splicing. Nucleic acids research 33 (Database issue):D75-79. doi:10.1093/nar/gki118

18. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA, Pujar S, Rajput B, Rangwala SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, Wallin C, Webb D, Weber J, Wu W, DiCuccio M, Kitts P, Maglott DR, Murphy TD, Ostell JM (2014) RefSeq: an update on mammalian reference sequences. Nucleic acids research 42 (Database issue):D756-763. doi:10.1093/nar/gkt1114

19. Harrow JL, Steward CA, Frankish A, Gilbert JG, Gonzalez JM, Loveland JE, Mudge J, Sheppard D, Thomas M, Trevanion S, Wilming LG (2014) The Vertebrate Genome Annotation browser 10 years on. Nucleic acids research 42 (Database issue):D771-779. doi:10.1093/nar/gkt1241

20. Heber S, Alekseyev M, Sze SH, Tang H, Pevzner PA (2002) Splicing graphs and EST assembly problem. Bioinformatics 18 Suppl 1:S181-188. doi:10.1093/bioinformatics/18.suppl_1.s181

21. Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46 (5):323-351. doi:10.1080/00107510500052444

22. Clauset A, Shalizi CR, Newman MEJ (2009) Power-Law Distributions in Empirical Data. SIAM Review 51 (4):661-703. doi:10.1137/070710111

23. Virkar Y, Clauset A (2014) Power-law distributions in binned empirical data. The Annals of Applied Statistics 8 (1):89-119. doi:10.1214/13-aoas710

24. Klaus A, Yu S, Plenz D (2011) Statistical analyses support power law distributions found in neuronal avalanches. PloS one 6 (5):e19779. doi:10.1371/journal.pone.0019779

25. Vuong QH (1989) Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica 57 (2):307. doi:10.2307/1912557

26. Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Reviews of Modern Physics 74 (1):47-97. doi:10.1103/RevModPhys.74.47

27. Grinev VV, Ramanouskaya TV, Gloushen SV (2013) Multidimensional control of cell structural robustness. Cell Biology International 37 (10):1023-1037. doi:10.1002/cbin.10128

28. Stelling J, Sauer U, Szallasi Z, Doyle FJ, 3rd, Doyle J (2004) Robustness of cellular functions. Cell 118 (6):675-685. doi:10.1016/j.cell.2004.09.008

29. Koonin EV, Wolf YI, Karev GP (2006) Power laws, scale-free networks and genome biology. Molecular biology intelligence unit. Landes Bioscience/Eurekah.com ;
Springer Science+Business Media, Georgetown, Tex.
New York, N.Y.

30. Kaneko K, Furusawa C (2008) Consistency principle in biological dynamical systems. Theory in biosciences = Theorie in den Biowissenschaften 127 (2):195-204. doi:10.1007/s12064-008-0034-z

31. Marquet PA, Quinones RA, Abades S, Labra F, Tognelli M, Arim M, Rivadeneira M (2005) Scaling and power-laws in ecological systems. The Journal of experimental biology 208 (Pt 9):1749-1769. doi:10.1242/jeb.01588

32. Stauffer D (2009) Phase Transitions on Fractals and Networks.6783-6789. doi:10.1007/978-0-387-30440-3_406

33. Nykter M, Price ND, Larjo A, Aho T, Kauffman SA, Yli-Harja O, Shmulevich I (2008) Critical networks exhibit maximal information diversity in structure-dynamics relationships. Physical review letters 100 (5):058702. doi:10.1103/PhysRevLett.100.058702

34. Carlson JM, Doyle J (2002) Complexity and robustness. Proceedings of the National Academy of Sciences of the United States of America 99 Suppl 1:2538-2545. doi:10.1073/pnas.012582499