MethodsX 12 (2024) 102697

Contents lists available at ScienceDirect

MethodsX

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

CaTCH: Calculating transcript complexity of human genes ☆

Koushiki Basu, Anubha Dey, Manjari Kiran*

Department of Systems and Computational Biology, School of Life Sciences, University of Hyderabad, Hyderabad, Telangana 500 046, India

ARTICLE INFO

Method name:
CaTCH

Keywords:
Transcript complexit
Alternative splicing
Linear Regression
Random Forest

ABSTRACT

The findings based on whole transcriptome sequencing suggest that alternative splicing occurs in approximately 95% of human multi-exon genes, thus, playing a crucial role in promoting proteome diversity. According to the latest GENCODE annotations, most genes have less than four transcripts, positively correlating with the number of exons. Thus, it is more accurate to measure the splice variant efficiency of a gene with respect to the number of exons, which is a measure of Transcript Complexity (TC). In addition to that, the theoretical number of transcripts is substantially higher than the actual number of transcripts produced by Alternative Splicing Events, and the features restricting this phenomenon need to be explored.

In this method, we have extracted the data of various features contributing to TC from different databases. Linear regression is used to identify the determinant features and to train and test the model of TC. The results indicate that exon length is the determining feature of TC, followed by coding potential, presence of chromatin signature, and 5′ splice site dinucleotide, all of which negatively affect a gene's TC, except exon length. To further classify the genes based on TC, random forest is used to identify the determinant features.

- The splicing efficiency of a gene can be inferred by the transcript complexity, which is the number of transcripts per exon.
- CaTCH is a linear regression-based model to calculate the transcript complexity of human genes, which can be calculated from the exon length, coding potentiality, presence of chromatin signature/s, and 5' splice site dinucleotide.

Specifications table

Subject area:	RNA biology
More specific subject area:	Transcription
Name of your method:	CaTCH
Name and reference of original method:	Machine learning
Resource availability:	Koushiki Basu, Anubha Dey & Manjari Kiran (2023) Inefficient splicing of long non-coding RNAs is associated with higher transcript complexity in human and mouse, RNA Biology, 20:1, 563–572, DOI: 10.1080/15,476,286.2023.2242649

☆ **Related research article**: Koushiki Basu, Anubha Dey & Manjari Kiran (2023) Inefficient splicing of long non-coding RNAs is associated with higher transcript complexity in human and mouse, RNA Biology, 20:1, 563-572, DOI: HYPERLINK 10.1080/15476286.2023.2242649.

* Corresponding author.
E-mail addresses: koushiki.basu26@gmail.com (K. Basu), manjari.hcu@uohyd.ac.in (M. Kiran).
Social media: @KoushikiBasu26 (K. Basu), @manjkira564 (M. Kiran)

https://doi.org/10.1016/j.mex.2024.102697
Received 4 December 2023; Accepted 3 April 2024
Available online 4 April 2024

Background

The DNA sequence of a gene is converted to pre-RNA through transcription, which is further processed into mature RNA, also called transcript. During post-transcriptional regulation, transcripts are processed by splicing to include only exonic sequences [1]. Some exons can be excluded or retained alternatively during the splicing process, resulting in distinct mature RNAs from the same pre-RNA by a process known as alternative splicing [2]. Thus, alternative splicing events have an indispensable role in increasing transcriptomic diversity. The errors in alternative splicing, including genetic changes and altered expression of trans-acting factors and pre-RNAs, lead to various diseases and syndromes such as cancer [3–5].

Few studies also mention structural and organizational differences in genes, such as the average length of the intronic region or the average number of exons, which may explain the lower number of splice variants [6–8]. However, whether a low/high number of exons or shorter/longer intronic length explains the lower efficiency of splice variants is unclear. Furthermore, different intron-specific features such as 5' and 3' splice sites dinucleotides and their strength, count of the consensus sequence of branch points and their distance from the 3' splice site, or interaction with certain splicing factors reduce the rate of alternative splicing [9,10]. Again, which 5' (AT, GC, or GT) and 3' (AC or AG) splice site dinucleotides and which branch points explain the lower efficiency of splice variants is yet to be clarified. In this study, it is investigated how different features regulate alternative splicing of a gene and contribute to its Transcript Complexity (TC). TC measures the number of alternative spliced product of a gene with respect to the number of exons. Recently, the number of RNA-seq reads mapped to spliced versus unspliced transcripts has been used to measure splicing efficiency [11]. Few studies have also found that high coding potential, high epigenetic regulation (e.g., H3K9me3 histone modification), lower interaction with splicing factors (such as U2AF65 binding), fewer SR protein binding sites, and the lack of RNA polymerase II phosphorylation at the 5' splice site reduce alternative splicing efficiency [8,11–15].

The method proposed in the present study uses gene annotation from the GENCODE [13] project to find the determinant features of Transcript Complexity (TC). The linear regression equation is derived to calculate the TC of genes, and the random forest is used to classify any gene based on TC. In addition, the model is validated on other annotation data. In this study, the TC of a gene is explored with different genomic and splicing features. It is observed that genes with high TC are associated with (i) longer exons and shorter introns, (ii) introns having GC as a 5' splice site dinucleotide, (iii) lower splice site strength, (iv) long distance between the 3' splice site and branch point, (v) less fraction of transcripts having chromatin signature and (vi) low conservation scores.

Rationale

Alternative splicing is thought to affect more than half of all human genes, and recent studies are exploring its biological impact on a large scale [16]. Aside from having a role in generating proteome diversity, alternative splicing can also regulate gene expression by splicing RNAs into unproductive transcripts, which are targeted for degradation. From the annotation data extracted from the GENCODE dataset and plotting the distribution of the number of transcripts for each gene, it is observed that most of the genes have less than 4 transcripts, and no study has been reported to date on how many exons contribute to this number of transcripts (Fig. 1A).

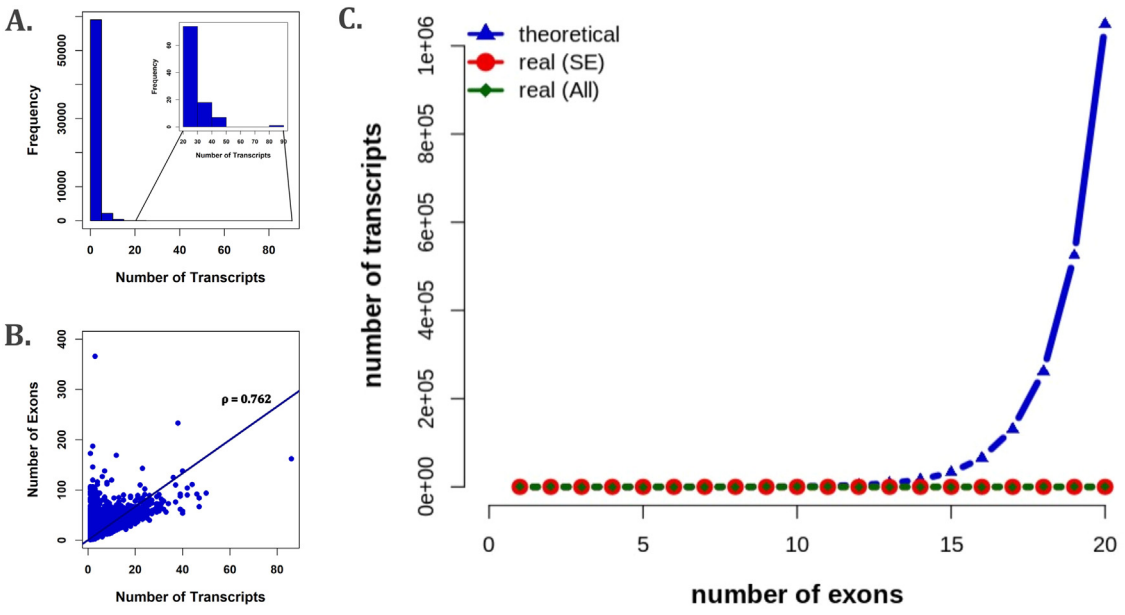

Fig. 1. (A) Distribution of number of transcripts in genes (B) Correlation between the number of transcripts and the number of exons in all genes [$\rho = 0.762$ & p-value < 2.2e-16] (C) Distribution of theoretical and real number of transcripts produced by a particular number of exons.

Furthermore, there is a positive correlation between the number of exons and the number of transcripts [ρ = 0.762 & p-value < 2.2e-16]; thus, assessing a gene's splice variant efficiency in relation to the number of exons is critical (Fig. 1B). Another gap in the literature is the large difference between the theoretical number of transcripts produced by Skipping Exon (SE) and the actual number of transcripts produced by either SE or considering all the Alternative Splicing Events (Fig. 1C). And there must be some undiscovered gene-related attributes that limit alternative splicing. Despite numerous studies mentioning various regulatory mechanisms that limit the number of alternatively spliced transcripts of each expressed gene, no model has been reported to predict the TC of a gene.

Material and methods

Data collection

For model building, the data from the release v41 annotated on the human genome sequence GRCh38.p13 is downloaded from GENCODE (gencode.v41.basic.annotation.gtf.gz) [17]. Additionally, for testing the model, various annotation data are downloaded from the GENCODE [17] (version 38 - gencode.v38.basic.annotation.gtf.gz), the Ensembl database [18] (version 38 - Homo_sapiens.GRCh38.109.gtf), and the RefSeq database at NCBI [19] (T2T consortia data – GCF_009914755.1_T2T_CHM13v2.0_genomic.gtf). The exon lengths are extracted from exon sequences, and intron lengths, 5'/3' splice site dinucleotide, splice site strength, and the distance between branchpoint and 3' splice site are retrieved from the intron sequences from the UCSC table browser. For coding potential calculation, the FASTA sequences of transcripts are downloaded from GENCODE (gencode.v38.transcripts.fa). The chromatin signature data is downloaded from the ENCODE project GEO database [20].

Transcript complexity calculation

As mentioned previously [21], the number of transcripts per exon is used to compute the TC of a gene. The number of transcripts and exons for each gene is obtained from basic annotation datasets in order to calculate the TC for each gene.

$$Transcript\ Complexity\ (TC) = \frac{Number\ of\ Transcripts}{Number\ of\ Exons}$$

Alternative splicing events analysis

The GTF files of basic annotations from the GENCODE database and the SUPPA tool [22] are used to characterize the alternative splicing event for each gene. The SUPPA tool is used to identify alternative splicing events involving the GT-AG and GC-AG introns. Different types of alternative splicing events are categorized as follows: SE, alternative 5'/3' splice site (A5/A3 or SS if both are considered together), retained intron (RI), mutually exclusive exons (MX), and alternative first/last exon (AF/AL or FL if both are considered together).

Coding potential calculation

From GENCODE, the FASTA sequence of transcripts (gencode.v38.transcripts.fa) is downloaded and used as the input file of the Coding Potential Calculator (CPC2), and the coding potential of each transcript is calculated [23]. CPC2 is the upgraded version of CPC and, like CPC, uses SVM to construct classifiers without alignment. The main four features are the longest ORF length, Fickett score, ORF integrity, and pI value. The pI's characteristic is obtained by translating the longest ORF into an amino acid sequence and then computing the physicochemical property of the pI of amino acids. Consequently, the pI feature performed well in the CPC2 model.

Exon and intron analysis

Human GRCh38 sequences are used to obtain exon and intron sequences from the Table Browser tool at UCSC [24]. Since single-exon genes do not undergo splicing, they are all excluded from this analysis. The total exon length and intron length for each transcript are determined by counting the number of bases. The 5' and 3' splice site dinucleotides are also extracted from each intron sequence. The splice site strength, which refers to the efficiency with which a particular sequence motif is recognized as a splice site during the process of pre-mRNA splicing, is calculated using the MaxEntScan web tool based on the "Maximum Entropy Principle" and generalizes most prior probabilistic models of sequence motifs [25]. The MaxEntScan web tool predicts the strength of the splicing sequences by considering both adjacent and non-adjacent dependencies between positions.

Each 5' and 3' splice site contains consensus sequences recognized by the spliceosome, a complex of RNA and protein molecules facilitating the splicing process. Splice site strength is determined by how closely its sequence aligns with the consensus sequence. In calculating splice site strength, the 9 and 23 nucleotide sequence motifs from the 5' and 3' splice sites, respectively are extracted from exon and intron sequences. The 5' splice site is scored by MaxEntScan::score5ss using the input FASTA file containing sequence motifs of 9 nucleotides (3 bases in exon and 6 bases in intron). The 3' splice site is scored similarly using MaxEntScan::score3ss, and each sequence motif in the FASTA file is 23 nucleotides long (20 bases in intron and 3 bases in exon). The maximum entropy model (MAXENT) is used to calculate the strength of the 5' and 3' splice sites. The "GencoDymo" R package version 0.2.1 is used to obtain the FASTA file of the sequence motifs used as input for the MaxEntScan function [9].

Also, from each intron sequence, the position of occurrence of the consensus sequence of branching point (yUnAy) [26] and the position of the 3' splice site are extracted. The distance between the positions of the 3' splice site and branch points is calculated, and the median value of the distance of all introns is mapped against the gene.

Chromatin signature analysis

The narrowPeak files from publicly available ChIP-seq data are downloaded from ENCODE [20] for *Homo sapiens* GM12878 cell lines, and each experiment's all replicates are chosen. For each histone modification of GM12878 cell line - H3K4me3 (GSE95899), H3K9me3 (GSM733664), and H3K36me3 (GSM733679), annotatePeaks.pl command of the HOMER package [27] is used to annotate each peak.

Splice site conservation analysis

Using phastCons data from UCSC [24], the conservation score of the 5' and 3' splice sites for each intron is assessed. The PhastCons score for 99 vertebrate genomes' multiple sequence alignments to the human genome (hg38.phastCons100way.bw) is downloaded. The positions of the 5' and 3' splice sites are retrieved from the intron sequence data (from UCSC) for each intron sequence, and the phastCons conservation score is assigned to each position.

Statistical analysis

R version 4.2.1 is used for statistical and data analysis. Spearman's rank correlation test is used to calculate the correlation between the number of transcripts and the number of exons of the gene. A p-value < 0.05 is considered significant for all statistical tests.

Model building and testing

Exon length, intron length, coding potential, 5' and 3' splice site dinucleotide, the strength of 5' and 3' splice site, the distance between branchpoint and 3' splice site, presence or absence of chromatin signature and conservation score of 5' and 3' splice sites are included in the data for the model building. Since the PhastCons score data from UCSC [24] for most genes are missing, they are omitted from the final dataset. 70% (4024) of the dataset is used for training, whereas the remaining 30% (1725) is reserved for testing.

The caret R package [28] and linear regression with the base R function lm are used to predict the equation for calculating TC. A Random Forest classifier is also built using the randomForest R package [29] to predict the gene's TC. Based on the median value of TC (0.333), the genes are divided into two major categories/classes, that is, "high" and "low." The high TC corresponds to those of which the TC value is greater than 0.333, and the low TC corresponds to those with a TC value less than or equal to 0.333. Each feature and various combinations of features are tested against TC to calculate the area under the curve (AUC) values using the Metrics R package [30]. The auc function is used to determine the model's accuracy. The best model is trained and further evaluated on the test dataset using AUC obtained with the auc function. For AUC, 1000 and 500 permutations of splitting the data into training and testing datasets are performed for linear regression and random forest, respectively. The roc function of the pROC R package is used to plot the Receiver Operating Characteristic (ROC) curve for each test dataset [31].

Method validation

Positive correlation between the number of transcripts and the number of exons

Any gene composed of exons produces multiple numbers of transcripts, and it is observed that the number of transcripts is positively correlated to the number of exons (Spearman correlation test, $\rho = 0.769$ and p-value $< 2.2e\text{-}16$) (Fig. 1B). Most reports have focused on genes with 1–4 transcripts because most genes have 1 to 4 transcripts. However, given the number of exons, some genes might generate more transcripts. Splice variant efficiency can be calculated by measuring the number of transcripts reported for a gene with respect to the number of exons. This is called a "Transcript Complexity" (TC) score for the gene.

Determinant features of transcript complexity

To identify the determinant feature of TC for the dataset of all genes, 37 linear models are constructed and assessed. The model accuracy of the first eight models corresponds to one of the single features. Utilizing exon length, coding potential, or presence or absence of chromatin signature, an AUC of 0.6–0.7 is achieved. It is observed that the AUC of any linear model constructed by combining different features increases if at least two of the features in the model are either exon length, coding potential, or presence of chromatin signature. No significant increase is observed in AUC for the last few models, which had more than four features. Eventually, data from exon length, coding potential, presence of chromatin signature, and 5' splice site dinucleotide is combined for all genes, which resulted in an AUC rise to 0.75 (Fig. 2A). In conclusion, exon length, coding potential, presence of chromatin signature, and 5' splice site dinucleotides are the features that can influence TC.

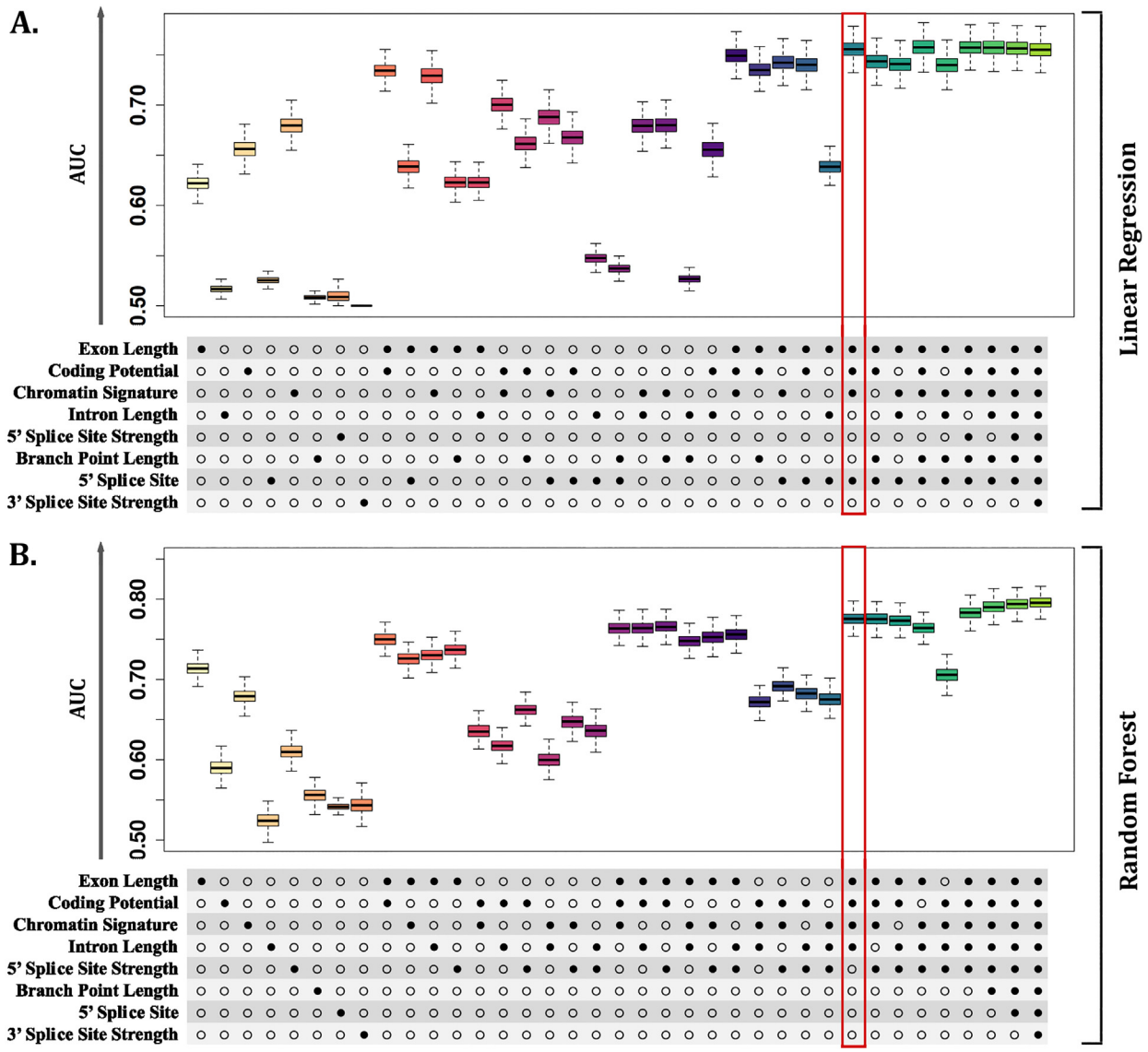

Fig. 2. Distribution of AUC for each model taking different features of transcript complexity for all genes (A) calculating TC of a gene using linear regression, (B) classifying a gene based on TC using Random Forest classifier (It is calculated based on 1000 [linear regression] and 500 [random forest] permutations of splitting data into training and testing sets. The red box shows the best model and the determinant features of TC for all genes).

Validation of the transcript complexity model

The best linear model for predicting TC with an AUC of 0.78 includes features such as exon length, coding potential, presence of chromatin signature, and 5' splice site dinucleotide. The model is trained on 70% of the GENCODE dataset. (release v41 annotated on the human genome sequence GRCh38.p13). Additionally, the equation to predict TC is also derived. The equation is as follows:

$$TC = 0.3455 + (0.0002 * Exon\ Length) - (0.0853 * Coding\ Potential) - (0.0283 * Chromatin\ Signature) - (0.0178 * 5'Splice\ Site)$$

Thus, the TC of any gene can be calculated just from the exon length, coding potential, presence of chromatin signature, and 5' splice site dinucleotide data. And from the equation, it is observed that longer exons tend to increase the TC where as high coding potentiality, presence of chromatin signature and AT as 5' splice site dinucleotide decreases the TC of a gene. The model is further validated with 30% of the GENCODE dataset (release v41) and with other annotation data as well - the GENCODE project (version 38 - gencode.v38.basic.annotation.gtf.gz), the Ensembl database (version 38 - Homo_sapiens.GRCh38.109.gtf), and the RefSeq database at NCBI (T2T consortia data – GCF_009914755.1_T2T_CHM13v2.0_genomic.gtf).

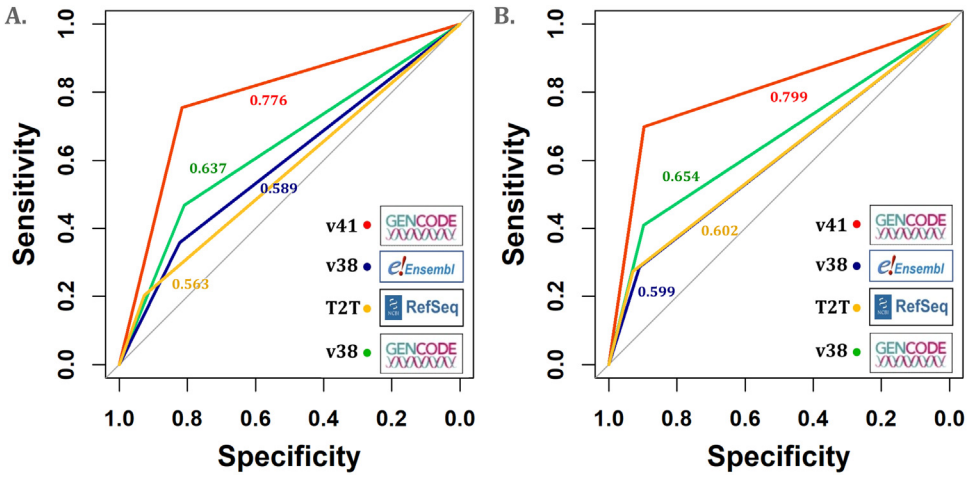

Fig. 3. Plot of Receiver Operating Characteristic (ROC) curve of the best model for (A) calculating TC of a gene using linear regression, (B) classifying a gene based on TC using random forest classifier for different test datasets (release v41 GENCODE dataset, release v38 GENCODE dataset, release v38 Ensembl dataset, and T2T consortia data from the RefSeq database).

Based on the median value of TC (0.333), genes are divided into two groups, i.e., high and low TC, and are denoted by 1 and 0, respectively. Based on the observed and predicted TC, the area under the curve (AUC) is determined, and the ROC curve is plotted (Fig. 3A). The 30% of the GENCODE dataset (release v41) has a maximum AUC of 0.7764, followed by the AUCs of the release v38 GENCODE dataset (0.637), the release v38 Ensembl dataset (0.589), and the T2T consortia data from the RefSeq database (0.563). The inference is that the model performance is adequate in different validation datasets.

Random forest classifier based on transcript complexity

Like linear regression, 37 random forest classifiers are constructed and assessed (Fig. 2B). The classifier consists of 11 features, out of which "Exon Length," "Coding Potential," "presence of Chromatin Signatures," and "Intron Length" turn out to be the discriminatory ones from the first eight models. This suggests that permutation in the actual values of these features drops the model accuracy to a great extent. An accuracy of 79.9% is obtained utilizing version 41 of the GENCODE data. As this dataset consists of 2937 "low" and 2812 "high" TC, no oversampling/undersampling is carried out to balance the data further. Other datasets such as "GENCODE (version 38)", "Ensembl," and "T2T consortium data from RefSeq database" are employed for model validation. Accuracy of 65.4%, 59.9%, and 60.2% are respectively received for the above-mentioned datasets (Fig. 3B). The accuracy and predictive power of random forest servers the purpose of selecting it against other classifiers.

Conclusion and future works

Although this study identifies discriminatory features, it raises new questions for unexplored features. Are there any other features affecting TC besides the mentioned features, and what are they? Is the polypyrimidine tract, which promotes the binding of spliceosomes, responsible for a gene's higher TC? Is the enrichment of transcripts affected by splice site recognition and the spliceosome assembling proteins such as specific enhancers, silencers, or SRSFs? What is the mechanism in which the strength of the 5' splice site sequence affects alternative splicing of a gene, in turn affecting the TC? Does subcellular localization or GT as a 5' splice site influence the TC of a gene?

Although there are many unanswered questions, this study is the first to report a model for predicting the TC of a gene. With the help of this model, the possible number of transcripts can be calculated for any novel gene by just using the determinant features. This study also, for the first time, reports the model that classifies the gene based on TC. Further studies on finding mechanisms and linking splicing efficiency to TC would help to understand alternative splicing in different genes.

The current work calculates a gene's TC by considering its exon count. Most prior studies assessed the effectiveness of gene splicing by counting transcripts or estimating the ratio of reads mapped to spliced to unspliced forms. TC refers to the efficiency of a gene's splice variants, measured based on the number of transcripts per exon. Previous research focused on the transcript's splicing efficiency and found that splicing kinetics strongly depended on the nature and position of 5' splice site flanking sequences. According to our findings, longer exons, less coding potential, absence of chromatin signature, and weak 5' splice site lead to inefficient splicing, giving higher TC.

Ethics statements

No animal or human subjects were used in the present work.

Funding

KB is a registered Integrated Systems Biology master's student and AD is a registered PhD student at the University of Hyderabad. A part of the work is funded by a Start-up Research Grant (SRG/2020/002146) awarded to MK from the Science and Engineering Research Board, Department of Science and Technology (SERB, DST), Government of India. AD and MK also acknowledge funding support from UoH-IoE Grant (UoH-IoE-RC2–21–012).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Koushiki Basu: Conceptualization, Methodology, Software, Data curation, Formal analysis, Writing – review & editing. **Anubha Dey:** Methodology. **Manjari Kiran:** Conceptualization, Methodology, Writing – review & editing, Validation, Supervision.

Data availability

All the processed data and codes are available on https://github.com/Koushiki26/CaTCH.git.

Acknowledgments

We thank the members of MKlab for the helpful discussions.

References

[1] P. Papasaikas, J. Valcárcel, The spliceosome: the ultimate RNa chaperone and sculptor, Trends Biochem. Sci. 41 (2016) 33–45, doi:10.1016/j.tibs.2015.11.003.
[2] Q. Pan, O. Shai, L.J. Lee, B.J. Frey, B.J. Blencowe, Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing, Nat. Genet. 40 (2008) 1413–1415, doi:10.1038/ng.259.
[3] Y. Zhang, L. Yan, J. Zeng, H. Zhou, H. Liu, G. Yu, W. Yao, K. Chen, Z. Ye, H. Xu, Pan-cancer analysis of clinical relevance of alternative splicing events in 31 human cancers, Oncogene 38 (2019) 6678–6695, doi:10.1038/s41388-019-0910-7.
[4] M.M. Scotti, M.S. Swanson, RNA mis-splicing in disease, Nat. Rev. Genet. 17 (2016) 19–32, doi:10.1038/nrg.2015.3.
[5] R.K. Singh, T.A. Cooper, Pre-mRNA splicing in disease and therapeutics, Trends Mol. Med. 18 (2012) 472–482, doi:10.1016/j.molmed.2012.06.006.
[6] J.J. Quinn, H.Y. Chang, Unique features of long non-coding RNA biogenesis and function, Nat. Rev. Genet. 17 (2016) 47–62, doi:10.1038/nrg.2015.10.
[7] E.T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, C.B. Burge, Alternative isoform regulation in human tissue transcriptomes, Nature 456 (2008) 470–476, doi:10.1038/nature07509.
[8] M.R. Khan, R.J. Wellinger, B. Laurent, Exploring the Alternative Splicing of Long Noncoding RNAs, Trends Genet. 37 (2021) 695–698, doi:10.1016/j.tig.2021.03.010.
[9] M. Abou Alezz, L. Celli, G. Belotti, A. Lisa, S. Bione, GC-AG introns features in long non-coding and protein-coding genes suggest their role in gene expression regulation, Front. Genet. 11 (2020), doi:10.3389/fgene.2020.00488.
[10] G. Roberts, Co-transcriptional commitment to alternative splice site selection, Nucleic Acids Res. 26 (1998) 5568–5572, doi:10.1093/nar/26.24.5568.
[11] M. Melé, K. Mattioli, W. Mallard, D.M. Shechner, C. Gerhardinger, J.L. Rinn, Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs, Genome Res. 27 (2017) 27–37, doi:10.1101/gr.214205.116.
[12] G. Leoni, L.L. Pera, F. Ferrè, D. Raimondo, A. Tramontano, Coding potential of the products of alternative splicing in human, Genome Biol. 12 (2011) R9, doi:10.1186/gb-2011-12-1-r9.
[13] Z. Krchňáková, P.K. Thakur, M. Krausová, N. Bieberstein, N. Haberman, M. Müller-McNicoll, D. Staněk, Splicing of long non-coding RNAs primarily depends on polypyrimidine tract and 5' splice-site sequences due to weak interactions with SR proteins, Nucleic Acids Res. 47 (2019) 911–928, doi:10.1093/nar/gky1147.
[14] I. Gonzalez, R. Munita, E. Agirre, T.A. Dittmer, K. Gysling, T. Misteli, R.F. Luco, A lncRNA regulates alternative splicing via establishment of a splicing-specific chromatin signature, Nat. Struct. Mol. Biol. 22 (2015) 370–376, doi:10.1038/nsmb.3005.
[15] T.V. Ramanouskaya, V.V. Grinev, The determinants of alternative RNA splicing in human cells, Mol. Genet. Genom. 292 (2017) 1175–1195, doi:10.1007/s00438-017-1350-0.
[16] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E.J. Sodergren, K.C. Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R.S. Kucherlapati, D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D.R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H.M. Lee, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R.W. Davis, N.A. Federspiel, A.P. Abola, M.J. Proctor, B.A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W.R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J.A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D.G. Brown, C.B. Burge, L. Cerutti, H.C. Chen, D. Church, M. Clamp, R.R. Copley, T. Doerks, S.R. Eddy, E.E. Eichler, T.S. Furey, J. Galagan, J.G.R. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L.S. Johnson, T.A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W.J. Kent, P. Kitts, E.V. Koonin, I. Korf, D. Kulp, D. Lancet, T.M. Lowe, A. McLysaght, T. Mikkelsen, J.V. Moran, N. Mulder, V.J. Pollara, C.P. Ponting, G. Schuler, J. Schultz, G. Slater, A.F.A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf, K.H. Wolfe, S.P. Yang, R.F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A. Felsenfeld, K.A. Wetterstrand, R.M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D.R. Cox, M.V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, A. Patrinos, M.J. Morgan, Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921, doi:10.1038/35057062.

[17] T. Derrien, R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D.G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J.B. Brown, L. Lipovich, J.M. Gonzalez, M. Thomas, C.A. Davis, R. Shiekhattar, T.R. Gingeras, T.J. Hubbard, C. Notredame, J. Harrow, R. Guigó, The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression, Genome Res. 22 (2012) 1775–1789, doi:10.1101/gr.132159.111.

[18] F. Cunningham, J.E. Allen, J. Allen, J. Alvarez-Jarreta, M.R. Amode, I.M. Armean, O. Austine-Orimoloye, A.G. Azov, I. Barnes, R. Bennett, A. Berry, J. Bhai, A. Bignell, K. Billis, S. Boddu, L. Brooks, M. Charkhchi, C. Cummins, L.Da Rin Fioretto, C. Davidson, K. Dodiya, S. Donaldson, B.El Houdaigui, T.El Naboulsi, R. Fatima, C.G. Giron, T. Genez, J.G. Martinez, C. Guijarro-Clarke, A. Gymer, M. Hardy, Z. Hollis, T. Hourlier, T. Hunt, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J.C. Marugán, S. Mohanan, A. Mushtaq, M. Naven, D.N. Ogeh, A. Parker, A. Parton, M. Perry, I. Piližota, I. Prosovetskaia, M.P. Sakthivel, A.I.A. Salam, B.M. Schmitt, H. Schuilenburg, D. Sheppard, J.G. Pérez-Silva, W. Stark, E. Steed, K. Sutinen, R. Sukumaran, D. Sumathipala, M.M. Suner, M. Szpak, A. Thormann, F.F. Tricomi, D. Urbina-Gómez, A. Veidenberg, T.A. Walsh, B. Walts, N. Willhoft, A. Winterbottom, E. Wass, M. Chakiachvili, B. Flint, A. Frankish, S. Giorgetti, L. Haggerty, S.E. Hunt, G.R. IIsley, J.E. Loveland, F.J. Martin, B. Moore, J.M. Mudge, M. Muffato, E. Perry, M. Ruffier, J. Tate, D. Thybert, S.J. Trevanion, S. Dyer, P.W. Harrison, K.L. Howe, A.D. Yates, D.R. Zerbino, P. Flicek, Ensembl 2022, Nucleic Acids Res. 50 (2022) D988–D995, doi:10.1093/nar/gkab1049.

[19] N.A. O'Leary, M.W. Wright, J.R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C.M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V.S. Joardar, V.K. Kodali, W. Li, D. Maglott, P. Masterson, K.M. McGarvey, M.R. Murphy, K. O'Neill, S. Pujar, S.H. Rangwala, D. Rausch, L.D. Riddick, C. Schoch, A. Shkeda, S.S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R.E. Tully, A.R. Vatsan, C. Wallin, D. Webb, W. Wu, M.J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T.D. Murphy, K.D. Pruitt, Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, Nucleic Acids Res. 44 (2016) D733–D745, doi:10.1093/nar/gkv1189.

[20] ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (2012) 57–74, doi:10.1038/nature11247.

[21] K. Basu, A. Dey, M. Kiran, Inefficient splicing of long non-coding RNAs is associated with higher transcript complexity in human and mouse, RNA Biol. 20 (2023) 563–572, doi:10.1080/15476286.2023.2242649.

[22] J.L. Trincado, J.C. Entizne, G. Hysenaj, B. Singh, M. Skalic, D.J. Elliott, E. Eyras, SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions, Genome Biol. 19 (2018) 40, doi:10.1186/s13059-018-1417-1.

[23] Y.J. Kang, D.C. Yang, L. Kong, M. Hou, Y.Q. Meng, L. Wei, G. Gao, CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features, Nucleic Acids Res. 45 (2017) W12–W16, doi:10.1093/nar/gkx428.

[24] D. Karolchik, The UCSC Table Browser data retrieval tool, Nucleic Acids Res. 32 (2004) 493D–4496, doi:10.1093/nar/gkh103.

[25] G. Yeo, C.B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals, in: Proceedings of the Seventh Annual International Conference on Computational Molecular Biology - RECOMB '03, New York, New York, USA, ACM Press, 2003, pp. 322–331, doi:10.1145/640075.640118.

[26] K. Gao, A. Masuda, T. Matsuura, K. Ohno, Human branch point consensus sequence is yUnAy, Nucleic Acids Res. 36 (2008) 2257–2267, doi:10.1093/nar/gkn073.

[27] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y.C. Lin, P. Laslo, J.X. Cheng, C. Murre, H. Singh, C.K. Glass, Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities, Mol Cell 38 (2010) 576–589, doi:10.1016/j.molcel.2010.05.004.

[28] M. Kuhn, Building predictive models in *R* using the caret package, J Stat Softw 28 (2008), doi:10.18637/jss.v028.i05.

[29] A. Liaw, M. Wiener, Classification and regression by randomForest, R News 2 (2002) 18–22.

[30] B. Hamner, M. Frasco, E. LeDell, Package 'Metrics', R Foundation for Statistical Computing, 2018.

[31] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.C. Sanchez, M. Müller, pROC: an open-source package for R and S+ to analyze and compare ROC curves, BMC Bioinform. 12 (2011) 77, doi:10.1186/1471-2105-12-77.