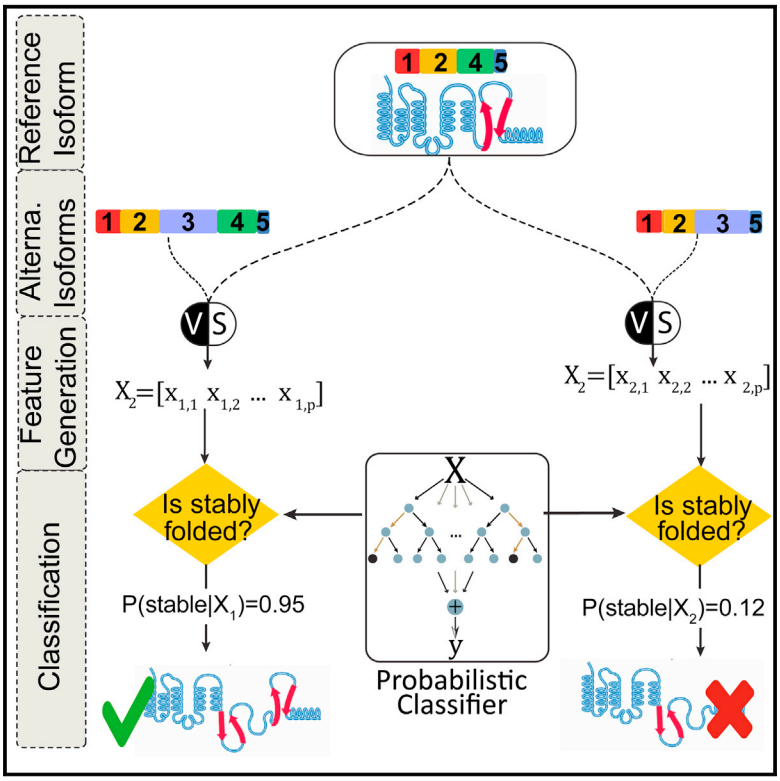
Report

Cell Reports

Semi-supervised Learning Predicts Approximately One Third of the Alternative Splicing Isoforms as Functional Proteins

Graphical Abstract



Authors

Yanqi Hao, Recep Colak, Joan Teyra, ..., Daisuke Kaida, Thomas Kislinger, Philip M. Kim

Correspondence

pi@kimlab.org

In Brief

Here, Hao et al. present PULSE, a novel machine-learning method that predicts which alternative splicing isoforms generate stably folded, viable proteins. They predict roughly one-third of isoforms as functional, many of which have considerable variation within their folded domains.

Highlights

- Predicting which isoforms produce stably folded proteins has been an open problem
- Recent advancements in theoretical machine learning enable us solve this problem
- PULSE, our proposed algorithm, predicts ~32% of isoforms as stably folded
- Probability of stably folding varies significantly across functional gene categories

Hao et al., 2015, Cell Reports *12*, 183–189
July 14, 2015 ©2015 The Authors
http://dx.doi.org/10.1016/j.celrep.2015.06.031

Cell Reports
Report

Semi-supervised Learning Predicts Approximately One Third of the Alternative Splicing Isoforms as Functional Proteins

Yanqi Hao,[1,2,10] Recep Colak,[1,2,10] Joan Teyra,[1,10] Carles Corbi-Verge,[1] Alexander Ignatchenko,[3] Hannes Hahne,[4] Mathias Wilhelm,[4] Bernhard Kuster,[4,5] Pascal Braun,[6] Daisuke Kaida,[7] Thomas Kislinger,[3,8] and Philip M. Kim[1,2,9,*]
[1]Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON M5S 1AS, Canada
[2]Department of Computer Science, University of Toronto, Toronto, ON M5S 3G4, Canada
[3]Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada
[4]Chair for Proteomics and Bioanalytics, TU Muenchen, Freising 85354, Germany
[5]German Cancer Consortium (DKTK), Munich, Germany; German Cancer Research Center (DKFZ), Heidelberg, Germany; Center for Integrated Protein Science Munich, Munich, Germany; Bavarian Biomolecular Mass Spectrometry Center, Technische Universität München, Freising, Germany
[6]Lehrstuhl fuer Systembiologie der Pflanzen, TU Muenchen, Munich, Germany
[7]Frontier Research Core for Life Sciences, University of Toyama, Toyama 930-8555, Japan
[8]Princess Margaret Cancer Center, University Health Network, Toronto, ON M5T 2M9, Canada
[9]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1AS, Canada
[10]Co-first author
*Correspondence: pi@kimlab.org
http://dx.doi.org/10.1016/j.celrep.2015.06.031

SUMMARY

Alternative splicing acts on transcripts from almost all human multi-exon genes. Notwithstanding its ubiquity, fundamental ramifications of splicing on protein expression remain unresolved. The number and identity of spliced transcripts that form stably folded proteins remain the sources of considerable debate, due largely to low coverage of experimental methods and the resulting absence of negative data. We circumvent this issue by developing a semi-supervised learning algorithm, positive unlabeled learning for splicing elucidation (PULSE; http://www.kimlab.org/software/pulse), which uses 48 features spanning various categories. We validated its accuracy on sets of bona fide protein isoforms and directly on mass spectrometry (MS) spectra for an overall AU-ROC of 0.85. We predict that around 32% of "exon skipping" alternative splicing events produce stable proteins, suggesting that the process engenders a significant number of previously uncharacterized proteins. We also provide insights into the distribution of positive isoforms in various functional classes and into the structural effects of alternative splicing.

INTRODUCTION

Alternative splicing (AS) can modify important molecular aspects of a large number of proteins (Stamm et al., 2005). AS can result in complete loss of function or the acquisition of new function, but the majority of cases reported predict subtle functional modulations. Evidence for AS almost exclusively comes from transcriptional profiling, and isoform databases have exploded in the last couple of years, achieving substantial coverage (Wang et al., 2008).Yet, the number of isoforms that are known to produce functional proteins is still considerably limited, and the lack of proteomic evidence for most isoforms raises uncertainty about their expression as proteins and physiological activity. In particular, the fraction of splice isoforms that produce stably folded proteins is an intensely debated topic (Melamud and Moult, 2009a; Tress et al., 2007, 2008a). We will refer to this issue for the remainder of the article as the "stably folded isoform discovery" (SFID) problem; our working definition of a "functional" protein is simply stably folded, viable, and not degraded.

Unfortunately, this debate is unlikely to be settled in the near future through experimental means. Mass spectrometry (MS) is the main high-throughput technique for detecting protein variants in cell lysates. Its chief limitation is that the false-negative rate is unknown, meaning that a particular isoform may not be expressed in a given time, tissue, or experimental setup or peptides that would distinguish it from the canonical isoform may be difficult to detect via MS (Blakeley et al., 2010). Therefore, whereas we have reasonably sized positive sets of bona fide expressed and stable protein isoforms, we do not have a reliable negative set, hindering the use of traditional computational approaches, such as supervised machine-learning algorithms, to predict which isoforms lead to a stably folded protein. These algorithms require both positive and negative training data.

Instead of predictive models, computational work has thus far focused on the exploratory analysis of bio-molecular properties that characteristically differentiate functional from non-functional isoforms. Many groups focused on structural attributes

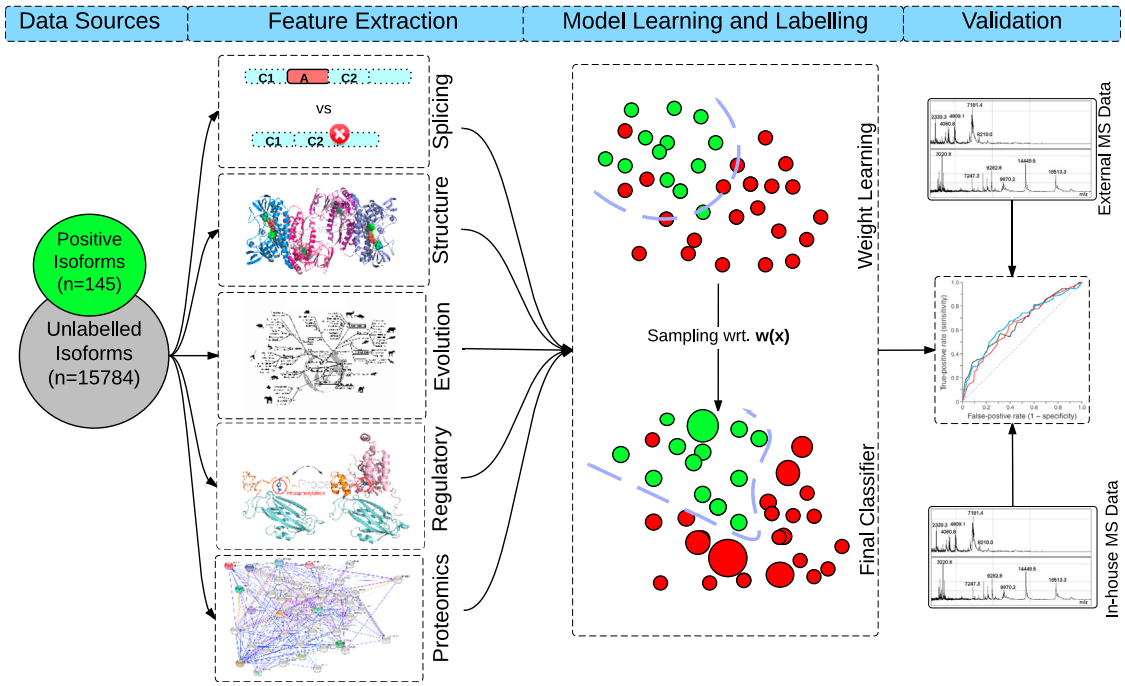


Figure 1. Overview of PULSE Framework

The PULSE framework consists of four main steps: (1) mapping splicing events to their protein counterparts and the canonical and identification of alternative isoforms; (2) comparison of the reference and the alternative isoforms across 48 features coming from five major categories of predictive features; (3) generation of a machine-learning algorithm-based prediction using the features; and (4) validation of the predictions using independent data. See also Table S1.

such as domains, folding (Birzele et al., 2008), and disordered regions (Tress et al., 2008a), whereas others developed generative models that include features such as expression levels, number of exons (Melamud and Moult, 2009a, 2009b), signal peptides, conservation, and exonic structure (Rodriguez et al., 2013). These studies illuminate important properties but draw diverse conclusions with respect to the percentage of functional isoforms. The limited—often non-overlapping—features that are used for analysis by different groups have thus far precluded consensus, and no single feature is a strong predictor of functionality in its own right (Birzele et al., 2008; Ezkurdia et al., 2012; Floris et al., 2011; Leoni et al., 2011; Melamud and Moult, 2009a; Rodriguez et al., 2013; Tress et al., 2008b). For instance, the recently developed APPRIS pipeline uses a subset of the aforementioned features to assign one particular isoform as the principal isoform, which is a related but different problem. Whereas a highly useful exploratory analysis tool, the scope of APPRIS does not extend to predicting which alternative isoforms will form stably folded proteins. Therefore, there is a need for a practical, comprehensive, unbiased, and robust computational model that can reliably identify functional (or stably folded) isoform products at a genome-wide scale.

RESULTS AND DISCUSSION

Overcoming the Missing Negative Data Problem: PULSE Algorithm

Supervised learning algorithms have been used in various biological problems with tremendous success. The input to a traditional binary classifier usually consists of a positive and a negative set. Unfortunately for the SFID problem, negative sets are unobtainable due to experimental limitations in proving the non-existence of an isoform's protein expression. As such, traditional supervised learning approaches can't be applied to the SFID problem, whereas unsupervised learning approaches are suboptimal as they cannot make use of the positive set. Here, we sidestep the problem of missing negative data by formulating SFID as an instance of a positive unlabeled learning (Elkan and Keith, 2008) problem. As a semi-supervised algorithm, it does not require a negative set for training and thereby overcomes the biggest hurdle in creating a predictive model. Given both a positive and unlabeled set, it can train a model that infers labels for the unlabeled set and also can predict labels for future data points. To do this, PULSE first trains a binary classifier wherein unlabeled data points are treated as negatives. Using the classifier's predictions for each data point, PULSE generates a modified (re-weighted) version of the data points, on which the second and the final classifier are built (see Figure 1 and Section S3a for algorithmic and implementation details).

PULSE leverages 48 predictive features spanning five categories: splicing, evolutionary, regulatory, proteomic, and structural features (see Section S1 for detailed descriptions). These features represent a unification and expansion of several feature sets that were individually but never jointly analyzed in previous studies (Blakeley et al., 2010; Ezkurdia et al., 2012; Floris et al., 2011; Hegyi et al., 2011; Leoni et al., 2011; Melamud and Moult, 2009a; Rodriguez et al., 2013; Severing et al., 2011; Tress et al., 2008b; Table S1; Section S1).

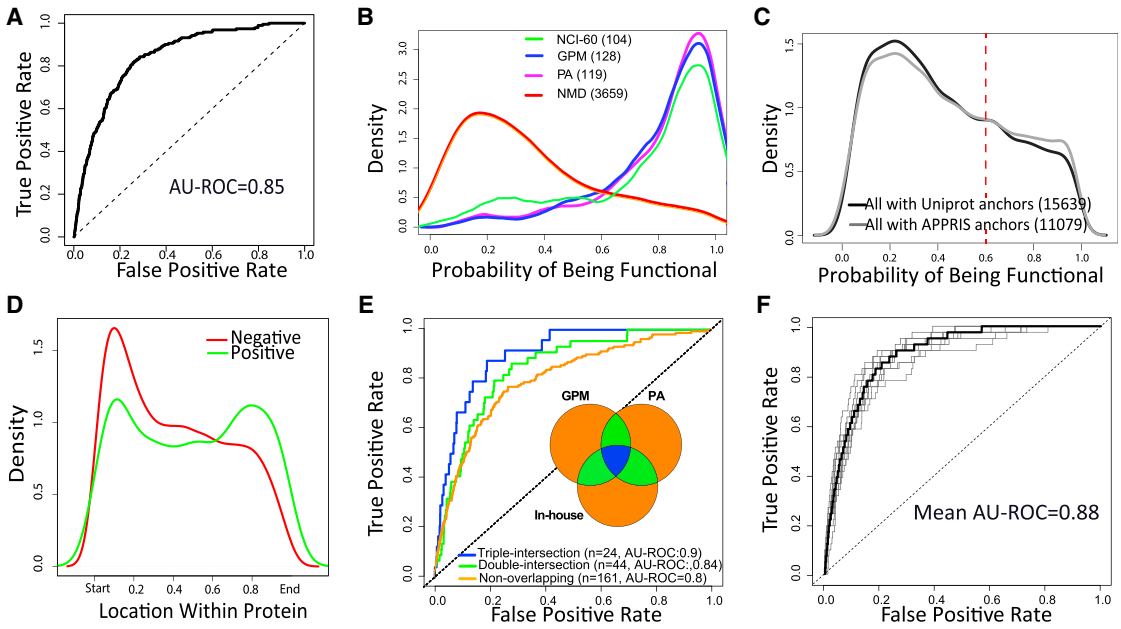


Figure 2. Performance Analysis

(A) AU-ROC analysis of PULSE prediction performance. The Hegyi set is used as positive training set, and MS-based sets are used for validation (see also Figure S1A).

(B) Distribution of PULSE scores, i.e., probability of being stably folded, for mass-spectrometry based isoforms from GPM (Global Proteome Machine, blue), PA (PeptideAtlas, magenta), NCI60 (green), and for NMD targets (red).

(C) Distribution of PULSE scores for the full isoform set using Uniprot (black) and APPRIS (gray) canonical isoforms for anchoring. Predictive feature values used for training, as well as amino acid sequence, identifier annotation, and the PULSE scores for the full isoform set can be found in Table S2.

(D) Density plot of relative location of splicing events causing NMD. Positive predicted isoforms tend to happen closer to the 3′ end, which is expected as the closer the event to the end, the shorter the length of the affected downstream regions.

(E) AU-ROC analysis with MS-based isoforms with varying confidence levels (the three-way intersection, three different two-way interactions, and the union of the GPM/PA/NCI60 sets).

(F) AU-ROC analysis of PULSE predictions, where MS-based isoforms are used as positive isoforms for both training and validation in a 10-fold cross-validation experiment.

See also Tables S2, S3, and S4.

Performance Evaluation

We undertook a comprehensive performance evaluation to ensure robustness of PULSE (see Figure S1A for schematic visualization of validation experiments). We trained PULSE using the set named as "verified" from Hegyi et al. (2011) as the positive set (we also evaluated performance when training our algorithm with other data sets; see Section S3b). This set consists of manually curated isoforms that have been detected on the protein level—usually from western-blot-based expression studies. Next, we separated the unlabeled set into training and held-out validation sets (Figure S1A). The training unlabeled set along with the Hegyi set is fed into PULSE algorithm, and the obtained model is used for labeling of the non-overlapping validation sets: held-out unlabeled set (negative) and the MS-based isoform sets (positive). This process is repeated ten times, each time with a different subset of the unlabeled set used as held-out set. For validation, we extracted isoforms from MS peptide databases (PeptideAtlas [PA]; Desiere et al., 2006; and Global Proteome Machine [GPM]; Craig et al., 2004) to independently assess the accuracy of our predictions. These isoforms are discovered by matching MS peptides to splice junctions of our isoforms (Section S2) and are non-overlapping and independent from our training data. Assuming these experimentally validated isoforms are positives and that the algorithm's performance is roughly the same in the positives of the unlabeled set, we are able to approximate the true-positive and false-positive rates of PULSE. Our results are summarized in the receiver operating characteristic (ROC) curve shown in Figure 2A. PULSE achieves an AU-ROC value of 0.85, suggesting that it outputs high probability values for functional isoforms.

Next, we had an in-depth look at the distribution of predicted scores for the three categories of isoforms (Figure 2B). As expected, both MS-verified isoform sets, namely GPM (blue) and PA (magenta) have a uni-modal distribution peaking at a score of 0.9, reaffirming our conclusions from Figure 2A. On the other hand, isoforms annotated to undergo nonsense-mediated decay (NMD) have a distribution peaking at 0.2, consistent with the expectation that most of these isoforms would not produce stably folded proteins.

To minimize the risk of bias from external data, we tested PULSE's performance on three distinct additional sets of MS-validated isoforms, based on proteome analysis of the NCI-60 cell lines (NCI-60), the Human Proteome Map (HPM) (Kim et al., 2014), and a draft human proteome ("Munich set"; Wilhelm

Table 1. Top 15 Predictive Features

Feature	Category	RI Score (%)	Positive Mean	Negative Mean
lengthDelta	splicing	100	64.57	301.0
lengthDeltaNormalized	splicing	80	0.10	0.49
disorderRateCanonical	structure	48	0.36	0.16
lengthAfter	splicing	45	315.4	99.2
disorderRateC2	structure	45	0.43	0.11
isFrameShift	splicing	39	0.24	0.69
seqConAve	evolution	39	0.71	0.62
disorderRateA	structure	37	0.41	0.12
elmRateCanonical	regulatory	33	0.16	0.09
seqConMin	evolution	31	0.51	0.34
coreRateA	structure	26	0.10	0.23
coreRateCanonical	structure	24	0.14	0.22
coreRateC2	structure	23	0.11	0.24
disorderRateC1	structure	22	0.40	0.13
ptmRateCanonical	regulatory	21	0.026	0.02

Category, relative importance score (RI score), and mean value in the positive set (positive mean) and in the negative set (negative mean) for the top 15 predictive features.

et al., 2014; see Section S2 for details). PULSE gives high scores for MS-validated isoforms (Figures 2B [green], S2D, and S2E), again confirming the high accuracy of our predictor. In addition, to account for the imperfect threshold selection and unknown false-positive rate in identification of spectra, we performed additional validation experiments to gauge PULSE's performance with validation sets of various confidence levels (Figure 2E). Isoforms that exist in the intersection of GPM, PA, and NCI-60 set (triple intersection) are the most reliable, followed by isoforms overlapping in only two data sets (double intersection), and finally by isoforms observed only in one set (no intersection). PULSE achieved AU-ROC values of 0.8, 0.84, and 0.9 for no-, double-, and triple-intersection sets, respectively, thus achieving better accuracy when using more-reliable validation sets (see also Section S3 and Figures S2C–S2F for FDR-based robustness analyses).

Feature Predictive Importance

Having confirmed PULSE's ability to correctly differentiate functional isoforms, we next analyzed the relative predictive power of features used (Section S4a).

Table 1 summarizes the 15 most-important features (see Figure S3A for the complete distribution) normalized against the most important feature. Almost all the features contribute to predictive power, and the top predictive features are from a diverse set of categories. Structural and splicing features constitute 42% and 33%, respectively, of the total importance. Notably, isoform length difference (i.e., the length of the AS exon) is the most important feature, consistent with the intuition that a longer insertion/deletion in a protein is more difficult to accommodate (Table 1). Likewise, AS exons that are disordered or flanked by disordered regions are easier to accommodate, whereas those that map to the protein core or a domain are not (Table 1). As expected, frameshifts are also not well tolerated. Most positive frameshift events arise from frame shifting near the end of the protein, which does not significantly affect the alternative isoform. We found, however, that a number of frame-shifted isoforms are likely to be functional (see Structural Characterization of Predicted Protein Variants). Interestingly, more positive isoforms arise when the alternative isoform is longer in length due to frameshift (Figure S3C). The remaining are regulatory (13% of importance), evolutionary (10%), and proteomic (2%) features. This finding implies that the problem at hand involves interactions of tens of features, and hence, any conclusion based on analysis of subsets of these features will be—to a large extent—suboptimal and nonconclusive (see Section S4 for additional details and analyses).

Predicted Functional Proteins and Characterization of Highly Spliced Genes

Given the relatively high accuracy of our predictions, we ran PULSE to label 15,639 unlabeled transcripts from the BodyMap data set (Cabili et al., 2011; Colak et al., 2013; see Figure S1B for the experimental setup). At a 90% true-positive rate, we predict that about 32% of isoforms are functional (Figure 2C), roughly consistent with one school of thought, which estimate around 20%–30% to be functional (Floris et al., 2011; Rodriguez et al., 2013). Thus, AS leads to a sizeable number of previously uncharacterized proteins (a total of 5,023 in this data set alone). To prevent any biases/errors that might be caused by using Uniprot canonical isoforms for anchoring (the mechanism by which we quantify how much an alternative isoform differs from its canonical pair—see Experimental Procedures), we repeated the analysis using APPRIS principal isoforms for anchoring. The score distribution remains qualitatively unchanged (Figure 2C).

Interestingly, we find that our predicted negative isoforms are enriched ($p < 0.03$) in poorly expressed genes, supporting the notion that some of these events may just be noisy splicing (Melamud and Moult, 2009a, b). The adaptability of protein isoforms to tolerate larger insertions and deletions is also better than previously predicted. For example, some isoforms maintain functionality despite splicing events affecting globular domains (see below).

Next, we analyzed high-level functional characteristics of highly spliced genes, which are genes ($n = 1,898$) that have at least three scored isoforms (Section S5a). We split these genes into two categories based on the ratio of positive predicted isoforms. We designate the first group as "highly positive" (HP), containing 384 genes that have a ratio of more than 0.5. The second group, termed "highly negative" (HN), contains 1,514 genes that have more than half of their isoforms negative (ratio ≤ 0.5). According to Gene Ontology enrichment analysis performed at biological process (level 2), the HP group has greater regulatory and developmental functions compared to the HN group, which tend to be involved in localization and metabolic processes. Compared to the HN group, the HP group also tends to be enriched with essential ($p < 4.8e{-}7$), non-house-keeping ($p < 7e{-}4$), and highly expressed ($p < 0.03$) genes (Figure S4).

Structural Characterization of Predicted Protein Variants

Most human proteins are composed of multiple domains that carry out different molecular processes required for prescribed

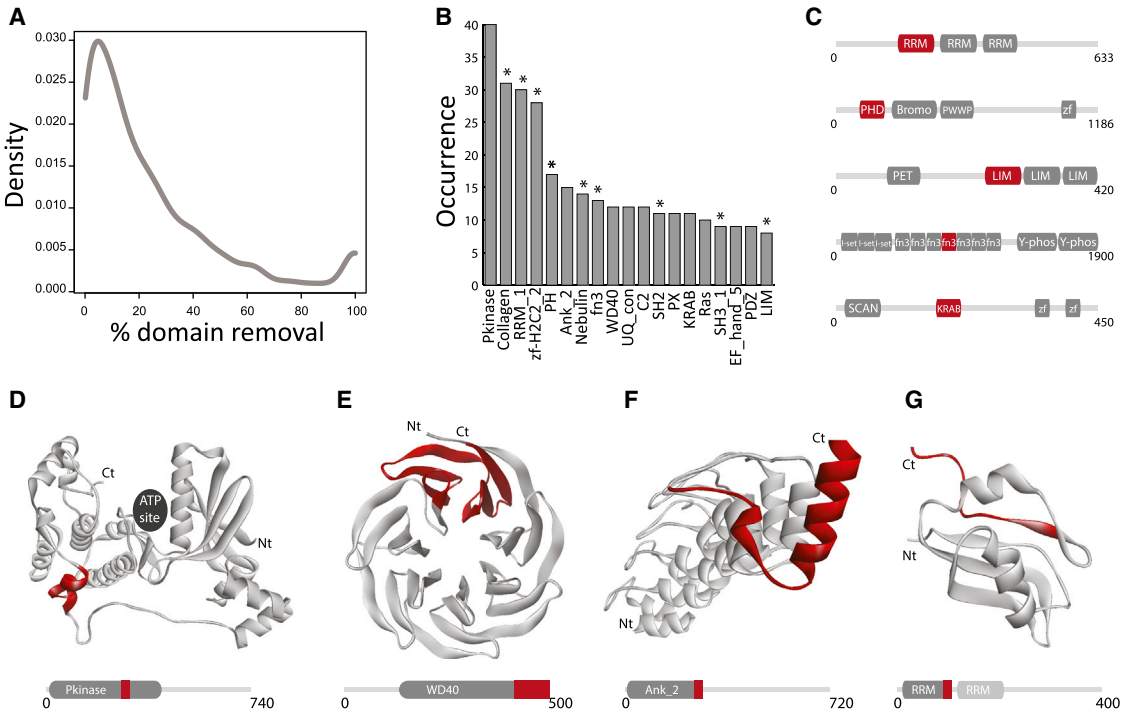


Figure 3. Structural Analysis of the Predicted Viable Protein Isoforms

(A) Distribution of spliced domains based on the relative percentage of fold removal. Negative numbers show domain growth by insertions.

(B) Most-abundant domain families within the predicted functional protein isoforms. Stars indicate significant enrichment compared to background ($p < 0.01$; chi-square test).

(C) Selection of multidomain containing protein isoforms in which a domain is fully removed. Uniprot and Isoform identifiers of the wild type protein are shown from top to bottom, where the domain deleted in the isoform is shown in red (O43390:IHs.373763.1, Q9ULU4:IHs.446240.4, Q9UGI8:IHs.592286.1, P10586:IHs.272062.1, Q9NWS9:IHs.697107.2).

(D) PKinase domain of BRSK2 protein isoform with region 261–271 spliced out.

(E) WD40 domain of Pleiotropic regulator 1 protein isoform with region 431–496 spliced out.

(F) Ankyrin repeat domain of ANKRD6 protein isoform with region 296–322 spliced out.

(G) First RNA-binding domain from DAZ-associated protein 1 isoform with region 80–102 spliced out.

Domain structures are represented in white ribbons, and deleted regions are shown in red. Structural mapping was done using PDB structures 1ZMU, 3CFS, 3B7B, 2DH8 as templates for (D), (E), (F), and (G), respectively.

biological function. Of the set of events that are predicted to generate functional isoforms, 36% affect the structural integrity of globular domains. In total, 3,089 positive events happen at non-structured regions of the canonical proteins and 67.4% of these affect disordered regions—intuitively easier to accommodate. This overrepresentation was previously observed by us and others (Colak et al., 2013; Romero et al., 2006). Interestingly, AS provides a mechanism to remove or add domains in protein isoforms in order to modulate biological processes. For instance, in tenascin C, AS determines the number of fibronectin type III domains that regulate binding to fibronectin (Chiquet-Ehrismann et al., 1991; Figure 3C). We find that 91% of the predicted isoforms with full domain removal (proteins with >90% domain removal in Figure 3A) contain more than one domain (versus 58% in the whole data set; $p < 0.01$; chi-square test). In fact, 54% of these cases contain multiple copies of the same domain family (see Figure 3C for examples). In addition, we observe some enrichment in our data for several modular and promiscuous domains in proteins, such as SH3, PDZ, LIM, zinc fingers, SH2, and PH domains (Figure 3B).

Perhaps surprisingly, our results show that 36% (1,820 in total) of our predicted functional isoforms have the alternatively spliced exon within a domain. It has been structurally shown that certain domain folds can accommodate relatively small modifications and still become soluble folded units (Birzele et al., 2008; Hegyi et al., 2009, 2011). Subtle structural changes have also been shown to affect the function of a protein by changing the affinity, specificity, or catalytic activity to native protein partners. Interestingly, the distribution of our predicted domain-affecting splice events by percentage of domain spliced is bimodal (Figure 3A). The major peak (80% of all events) represents domain folds with short truncations, where AS may influence the molecular properties of proteins, whereas a minor peak (6.1% of events) represents complete domain removal, presumably leading to a major functional change of the isoform (see Section S5b for examples and discussion of splicing tolerant domains).

Conclusions

In this paper, we present the first algorithm to predict the likelihood of yielding stably folded protein isoforms from AS events.

We solve the problem of a missing negative set by leveraging advancements in theoretical machine learning. To our best knowledge, this is the first analysis/algorithm that incorporates all existing—and some new—features, surpassing the depth and breadth of previous work. We estimate that around 32% of alternatively spliced isoforms lead to functional proteins, and we obtain a large list of putative and previously unobserved proteins that can be prioritized in future experimental studies. Ongoing sequencing efforts will continue to identify new splicing isoforms, and knowledge of those that will produce functional proteins is crucial. In this study, we only applied our methods to the most-simple form of AS events, that of exon skipping. It would, however, be straightforward to extend the approach to more-complex cases. Understanding the effects of AS at the protein level will be one of the major challenges in the near future. Having here made an important first step toward this goal, we expect PULSE's performance to further improve with ever increasing ground truth available for training.

EXPERIMENTAL PROCEDURES

Data Sets

An overview of our methodology is given in Figure 1. We use Uniprot (version 2013_04; 20,253 sequences) as our main source of canonical (principal/reference) isoforms, against which we compare the alternative isoform to derive comparative statistics. Note that, as Uniprot principal isoforms are not always correct, we also used alternative APPRIS canonical isoforms (Figure 2C). We observed almost no difference in performance; hence, we used Uniprot as source of canonical isoforms. In addition, it is important to once more underline that the principal isoforms are not treated as positive isoforms in PULSE, they are merely used for anchoring, the process through which we measure how much an alternative isoform deviates from the representative isoform of the gene, irrespective of whether or not the canonical one is functional.

We then used 27,240 AS events from Illumina's Human BodyMap 2.0 project (Cabili et al., 2011), which can be accessed at ArrayExpress (E-MTAB-513). These splicing events are represented by the alternative exon A and flanking exons C1 and C2. Of the 27,240 distinct human cassette exon AS events from RNA-seq data corresponding to a pair of isoforms each (C1-A-C2 and C1-C2), we were able to map 15,784 pairs of them such that one isoform from each pair maps to a canonical Uniprot protein. During mapping, we required a 95% identity match and absolute length difference of the matched C1-A-C2 peptide segment to be less than or equal to three in order to allow for sequencing errors and imperfect mappings. Out of the mapped pairs, we identified 12,015 had inclusion events (C1-A-C2) and 3,769 exclusion ones (C1-C2). We used the canonical pair as an anchor to derive predictive features in comparison to the alternative isoform. To account for the fact that we may be incorrect in our assignment of the canonical isoform, we also "inverted" all our predictions, i.e., we treated each inclusion event as an exclusion event and vice versa. Our prediction accuracy remained almost unchanged (Figure S2A). Isoforms are then mapped to Ensembl transcript data to obtain the isoforms at a transcript sequence level, which we used for translation. A subset of these isoforms containing 145 experimentally validated alternative isoforms (Hegyi set; Hegyi et al., 2011) is treated as the positive set to train our model. Unfortunately, a significant portion of the Hegyi set (~350) did not overlap with our gene set and/or the above mapping criteria was not satisfied. For validation, we used five distinct MS-validated isoform sets, which are described in detail in Section S2.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Results and Discussion, Supplemental Experimental Procedures, four figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2015.06.031.

AUTHOR CONTRIBUTIONS

Y.H., R.C., J.T., and C.C.-V. performed the experiments and analyzed the data; A.I., B.K., H.H., M.W., and T.K. provided mass spectra data and analyses; T.K., P.B., and D.K. provided critical reading and editing of the manuscript; and Y.H., R.C., J.T., and P.M.K. conceived the research project and wrote the manuscript.

ACKNOWLEDGMENTS

We thank Kevin Yuk Lap Yip, Ben Blencowe, and Manuel Irimia for valuable discussions. P.M.K., P.B., and D.K. acknowledge support from a Human Frontiers Science Program Young Investigator Award (RGY0080/2013).

Received: March 27, 2014
Revised: February 18, 2015
Accepted: June 9, 2015
Published: July 2, 2015

REFERENCES

Birzele, F., Csaba, G., and Zimmer, R. (2008). Alternative splicing and protein structure evolution. Nucleic Acids Res. *36*, 550–558.

Blakeley, P., Siepen, J.A., Lawless, C., and Hubbard, S.J. (2010). Investigating protein isoforms via proteomics: a feasibility study. Proteomics *10*, 1127–1140.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Chiquet-Ehrismann, R., Matsuoka, Y., Hofer, U., Spring, J., Bernasconi, C., and Chiquet, M. (1991). Tenascin variants: differential binding to fibronectin and distinct distribution in cell cultures and tissues. Cell Regul. *2*, 927–938.

Colak, R., Kim, T., Michaut, M., Sun, M., Irimia, M., Bellay, J., Myers, C.L., Blencowe, B.J., and Kim, P.M. (2013). Distinct types of disorder in the human proteome: functional implications for alternative splicing. PLoS Comput. Biol. *9*, e1003030.

Craig, R., Cortens, J.P., and Beavis, R.C. (2004). Open source system for analyzing, validating, and storing protein identification data. J. Proteome Res. *3*, 1234–1242.

Desiere, F., Deutsch, E.W., King, N.L., Nesvizhskii, A.I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S.N., and Aebersold, R. (2006). The PeptideAtlas project. Nucleic Acids Res. *34*, D655–D658.

Elkan, C., and Keith, N. 2008. Learning classifiers from only positive and unlabeled data. Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (KDD-2008) 1–8.

Ezkurdia, I., del Pozo, A., Frankish, A., Rodriguez, J.M., Harrow, J., Ashman, K., Valencia, A., and Tress, M.L. (2012). Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. Mol. Biol. Evol. *29*, 2265–2283.

Floris, M., Raimondo, D., Leoni, G., Orsini, M., Marcatili, P., and Tramontano, A. (2011). MAISTAS: a tool for automatic structural evaluation of alternative splicing products. Bioinformatics *27*, 1625–1629.

Hegyi, H., Buday, L., and Tompa, P. (2009). Intrinsic structural disorder confers cellular viability on oncogenic fusion proteins. PLoS Comput. Biol. *5*, e1000552.

Hegyi, H., Kalmar, L., Horvath, T., and Tompa, P. (2011). Verification of alternative splicing variants based on domain integrity, truncation length and intrinsic protein disorder. Nucleic Acids Res. *39*, 1208–1219.

Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft of the human proteome. Nature *509*, 575–581.

Leoni, G., Le Pera, L., Ferrè, F., Raimondo, D., and Tramontano, A. (2011). Coding potential of the products of alternative splicing in human. Genome Biol. *12*, R9.

Melamud, E., and Moult, J. (2009a). Stochastic noise in splicing machinery. Nucleic Acids Res. *37*, 4873–4886.

Melamud, E., and Moult, J. (2009b). Structural implication of splicing stochastics. Nucleic Acids Res. *37*, 4862–4872.

Rodriguez, J.M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.J., Lopez, G., Valencia, A., and Tress, M.L. (2013). APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res. *41*, D110–D117.

Romero, P.R., Zaidi, S., Fang, Y.Y., Uversky, V.N., Radivojac, P., Oldfield, C.J., Cortese, M.S., Sickmeier, M., LeGall, T., Obradovic, Z., and Dunker, A.K. (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc. Natl. Acad. Sci. USA *103*, 8390–8395.

Severing, E.I., van Dijk, A.D., and van Ham, R.C. (2011). Assessing the contribution of alternative splicing to proteome diversity in Arabidopsis thaliana using proteomics data. BMC Plant Biol. *11*, 82.

Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T.A., and Soreq, H. (2005). Function of alternative splicing. Gene *344*, 1–20.

Tress, M.L., Martelli, P.L., Frankish, A., Reeves, G.A., Wesselink, J.-J., Yeats, C., Olason, P.I., Albrecht, M., Hegyi, H., Giorgetti, A., et al. (2007). The implications of alternative splicing in the ENCODE protein complement. Proc. Natl. Acad. Sci. USA *104*, 5495–5500.

Tress, M.L., Bodenmiller, B., Aebersold, R., and Valencia, A. (2008a). Proteomics studies confirm the presence of alternative protein isoforms on a large scale. Genome Biol. *9*, R162.

Tress, M.L., Wesselink, J.-J., Frankish, A., López, G., Goldman, N., Löytynoja, A., Massingham, T., Pardi, F., Whelan, S., Harrow, J., and Valencia, A. (2008b). Determination and validation of principal gene products. Bioinformatics *24*, 11–17.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., et al. (2014). Mass-spectrometry-based draft of the human proteome. Nature *509*, 582–587.