Maximum Entropy Model for Alternative Splicing Starting Pattern

Qianyun Guo
Advisor: Qian Yu, Jotun Hein

October 25, 2013

Abstract

RNA splicing is an essential and precisely regulated post-transcriptional process that occurs prior to mRNA translation. In eukaryotes, alternative splicing occurs frequently to increase the biodiversity of proteins. With the development of high throughput sequencing method, there are more and more annotated genomes for bioinfomatics research. The Human-transcriptome DataBase for Alternative Splicing (H-DBAS) is a specialized database of alternatively spliced human transcripts.

In this work, we aim to use a maximum entropy model (MEM) to find alternative splicing related motifs in the upstream of a skipped exon, which is not conservative in all the transcripts. We test our model in prediction of skipped exons in human genome.

keywords alternative splicing, maximum entropy model, sequence motif, iterative scaling

BACKGROUND

Alternative splicing (AS) is a widespread mechanism for generating protein diversity and regulate protein expression in eukaryotes. Human genome produces around 150,000 different proteins from around 30,000 genes. An estimated 95% of transcripts from multiexon genes undergo alternative splicing, with a number of pre-mRNA transcripts spliced in a tissue-specific manner[1]. Abnormal variations in splicing may cause severe genetic disorders[2].

The process of splicing is regulated by trans-acting proteins (repressors and activators) and corresponding cis-acting regulatory sites silencers and enhancers) in the pre-mRNA[3]. These elements and factors governs how splicing will occur under different. In general, the determinants of splicing work in an inter-dependent manner that depends on context[4].

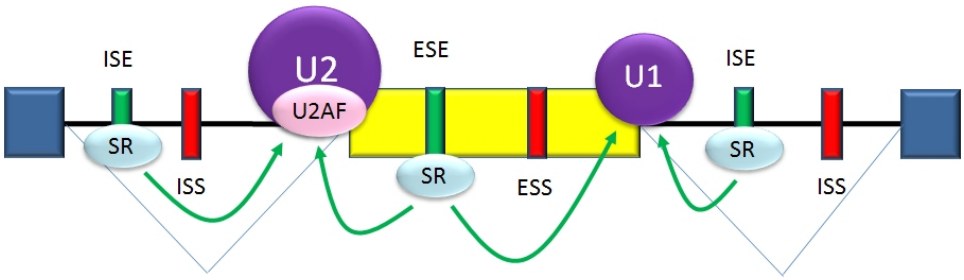

Figure 1: Splicing activator proteins bind to splicing enhancers in exons (ESE) and introns (ISE). They assist in the binding of U1 snRNP to the donor site and of U2AFs and U2 snRNP to the acceptor site and branch point.Alternative splicing occurs[5]

.

Motifs regarding splice enhancing and silencing exist within and without exons. Motif sequences which enhance splicing in exon are called ESE (Exonic splicing enhancer) and those which silence splicing in exon are called ESS (Exonic splicing silencer). Motif sequences which enhance splicing in intron are called ISE (intronic splicing enhancer) and those which silence splicing in

intron are called ISS (intronic splicing silencer). By the trans-acting splicing factors bound specificity with these motifs operate to snRNPs, alternative splicing occurs (Figure 1).

There are different patterns for alternative splicing (Figure 2). Cassette exon or skipping exon is the most commom mode in mammalian pre-mRNAs[6].

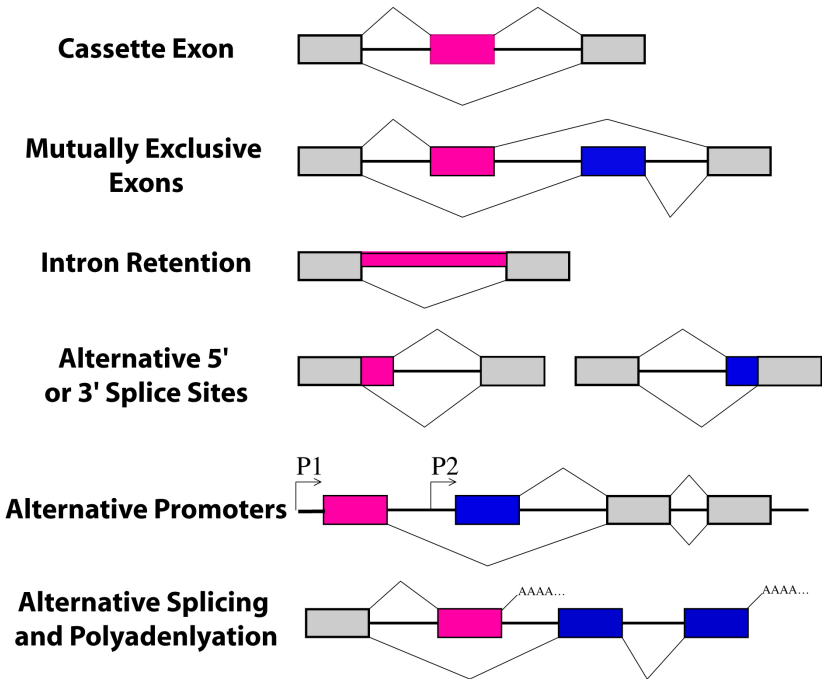

Figure 2: Types of alternative splicing. Boxes represent exons and lines represent introns. Colored exon regions are included in mature mRNA and gray regions are spliced out. Arrows indicate promoters and multiple A is the polyadenylation site[7]

When we look into the different transcripts of one gene, some exons may be kept in every transcripts while others may show up only in some of the transcripts. In this work, we mainly focus on the cassette skipped exons and strictly conservative exons. We are interested in the signals before conser-

vative exons and skipped exons. In order to find the signal motif, we use maximum entropy method to estimate the distribution of nucleotides in the upstream region of exons. We then evaluate our model with test set.

METHODS AND DATA

Maximum entropy model

Information theory provides a constructive criterion for setting up probability distributions on the basis of partial knowledge. The maximum entropy estimate is the least biased estimate possible on the given information, that is, the maximally noncommittal with regard to missing information[8].

The concept of maximum entropy can be traced back to an early age. However, only when computational methods are powerful enough, it can be widely applied to real world problems. A classical application of maximum entropy model is the pattern recognition in natural language processing. Recently, maximum entropy models are also used in bioinfomatic field, detecting signal patterns in genome sequence[9].

Let X be a sequence of λ random variables $X = \{X_1, X_2, ...X_\lambda\}$, where $X_i \in \{A, C, T, G\}$. Let $x = \{x_1, x_2, ...x_\lambda\}$ be a specific DNA sequence. Let $p(X)$ be the joint probability distribution $p(X_1 = x_1, X_2 = x_2, ...X_\lambda = x_\lambda)$ and $P(X = x)$ be the probability of a state in this distribution.

According to the principle of maximum entropy, among all the possible distributions in the hypothesis space that satisfy all the prior constraints, the distribution which is the best approximation of the true distribution is the one with the largest Shannon entropy,

$$H(\hat{p}) = -\sum \hat{p}(x) log_2(\hat{p}(x)) \tag{1}$$

where the sum is taken over all possible sequence, x. As there are four nucleotides, we use logarithms to base 2, so entropy can be measured in bits. Shannon entropy describes the "uniformity" of a random variable X. Intuitively, the principle is simple: model all that is known and assume nothing about that is unknown. In other words, given a collection of constraints i.e,

type	test set	training set	gene
skip	20427	20408	5636
conservative	26363	26429	8852

Table 1: Numbers of sequences in training set and test set and numbers of genes different exons located on

facts, choose a model consistent with all the constraints, but otherwise as uniform as possible.

After defining the concept of best model, we use an iterative scaling method to approach the maximum entropy model in our work. In each step, we apply a constraint in the distribution we get from last step, that is, putting information into the model. The result of adding information is the decrease in Shannon entropy.

Transcript data

To study the signal patterns, we observe the behavior of motifs in the upstream of skipped exons comparing with conservative exons. We download the transcript data RASV_human_flcdna which contains 95,160 transcripts in 27,193 locus from H-DBAS[10]. We then map the transcript sequence to UCSC Homo_sapiens.GRCh37.73.dna.chromosome to identify conservative exons and skipped exons (Figure 3). We classify the exons using a straightforward method. If an exon appears in all the transcripts and the coordinates are strictly the same, it is considered as a conservative exon. In other word, we ruled out the possibility of intron retention and alternative 5' or 3' splice site. If an exon with explicit boundaries only appears in some of the trancripts, it is considered as skipped, or cassette exon.

We divide the exons into training set and test set (Table 1) . We choose sequences at position (-6 to 0) of the upstream of exons. A pattern of AG consensus follows this region indicating the beginning of splicing.

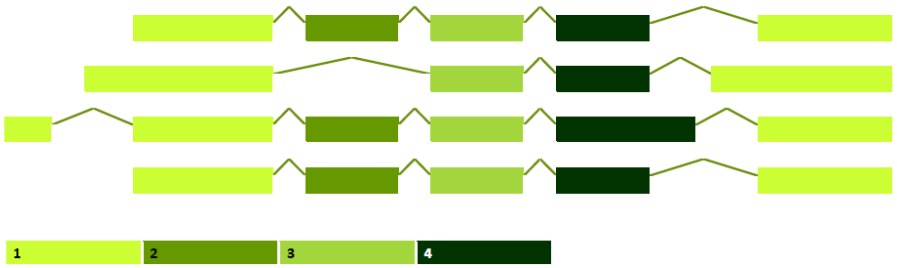

Figure 3: The gene has four observed transcripts. We choose type 2 exon as skipped exons and type 3 as conservative exons. Most part of type 4 exon conserved, but in transcript 3 there exists a retention 5' sites, so we don't consider this type as neither skipped nor conserved.

Constraints

In our work, we consider constraints carrying information of both position dependency and nucleotide frequency.

Let S_X be the set of all marginal distribution of the full distribution, $p(X = \{X_1, X_2, ...X_\lambda\})$. A marginal distribution is a joint distribution over a proper subset of X. For example, for $\lambda = 3$,

$$S_X = \{p(X_1), p(X_2), p(X_3), p(X_1, X_2), p(X_2, X_3), p(X_1, X_3)\} \tag{2}$$

Let $S_s^m \subseteq S_X$,m refers to the marginal order and s refers to skips in position.We divide all the constraints into there categories according to different patterns of motifs.

a. The first order constraints, $S_0^1 = \{p(X_1), p(X_2), ...p(X_3\lambda\}$, are the empirical frequencies of each nucleotide at each position. If only the first order constraints are applied in modeling, the distribution is the weight matrix model.In 2, the first three subsets are this kind of constraints.

b. The second order constraints, S_s^m, the subscript$s$ indicates the length of

skip in positions, the superscript m indicates the maximum length of skip in positions. In 2, $p(X_1, X_2), p(X_2, X_3)$ are the constraints with 0 skip ($s = 0$), and $p(X_1, X_3)$ is the constraints with 1 skip ($s = 1$). When we take m into consideration, S_s^m is a union of several subsets. For example, for $\lambda = 3$,

$$\begin{aligned} S_0^1 &= \{p(X_1), p(X_2), p(X_3)\} \\ S_0^2 &= \left\{S_0^1, p(X_1, X_2), p(X_2, X_3)\right\} \\ S_1^2 &= \left\{S_0^1, p(X_1, X_3)\right\} \end{aligned}$$

c. The higher order constraints focus on non-skip motifs. If a pattern is consisted of more than two nucleotides, we only consider the situation without skip, i.e, $s = 0$. For example, for $\lambda = 4$,

$$S_4 = \{p(X_1, X_2, X_3), p(X_2, X_3, X_4)\}$$

For all the constraints mentioned above, the observed frequency values for a particular member of constraints are added. For example, $p(X_1)$ contains 4 elements responding to $\{p_1(A), p_1(G), p_1(C), p_1(T)\}$, where $p_1(A)$ is the observed frequency of A in position 1.

Iterating scaling method

We use an iterating scaling method to approach the maximum entropy model. In each step of iteration, adding constraints into the model, the approximation to the maximum entropy estimate improves, using 1 as a measure of closeness of the approximating distribution to the true distribution.It may happen when inconsistent constraints are applied in the process of iteration. However, the set of all the constraints is the subset of the empirical distribution and therefore be consistent. The convergence of Shannon en-

tropy and the uniqueness of maximum entropy distribution can be rigorously proved[11].

We begin with a uniform distribution with $P^0(X) = 4^{-\lambda}$. Next, we add constraints Q_i to the distribution and update the distribution using,

$$P^j = P^{j-1}\frac{Q_i}{\hat{Q}_i^{j-1}}\hat{Q}_i^{j-1} = \sum_{x \in S_X - Q_i} P^{j-1}(X = x) \tag{3}$$

where P^{j-1}, P^j is the distribution at $j-1$ and j step of iteration, Q_i is the ith nucleotide relating constraint in a particular position constraint, and $\hat{Q}_i^{j-1}$ is the value of distribution corresponding to the ith constraint determined from p at the $j-1$ step. In other word, each step is a rearrangement of probabilities for all the $4^{-\lambda}$ possibilities responding to a new constraint.

The rate of convergence depend on how good the constraints are. Therefor we can control the rate of convergence by changing the order in which the constraints are applied[9]. As there are too many constraints, we apply a greedy strategy in choosing constraints in each step. We calculate the reduction in H relative to the distribution determined by previous step and choose the constraint with the largest ΔH for this step. The iteration stops when ΔH is small enough$|\Delta H| < 10^{-10}$.

The rank of constraints depend on the constraints ranked before. For example, if we consider the constraint set of

$$\{p(X_1 = A), p(X_2 = G), p(X_1 = A, X_2 = G)\}$$

suppose the pattern of AG is the true determining factor, i.e, with biological significance. In the case where constraints are chosen randomly, if we applied the first two constraints, the nucleotide bias in separated positions, before the third one, the significance of AG pattern may be obscured by the former ones. While in the ranked case, the constraint of AG pattern is the first

choice. After AG, the constraints are reordered. The ranking of A* (the first position is A) and *G (the second position is G) decline after updating.

Prediction and test

We applied the iteration over the upstream sequences of skipped exons and conservative exons, and get the two distribution $p^{skip}(X), p^{cons}(X)$. Given a new sequence in test set, the maximum entropy model can be used to distinguish skipped exons from conservative exons based on the likelihood ratio, L,

$$L(X = x) = \frac{P^{skip}(X = x)}{P^{cons}(X = x)} \tag{4}$$

where $P^{skip}(X = x)$ and $P^{cons}(X = x)$ are the probability of occurrence of sequence x from the distributions of skipped exon(skip) and conservative exon(cons).

RESULTS AND DISCUSSION

The greedy strategy works quite well(Figure 4). With the constraints ranked, information content $(12 - H)$ increases with a higher rate than that with random constraints. The top 20 constraints with highest ΔH are listed in Table2.

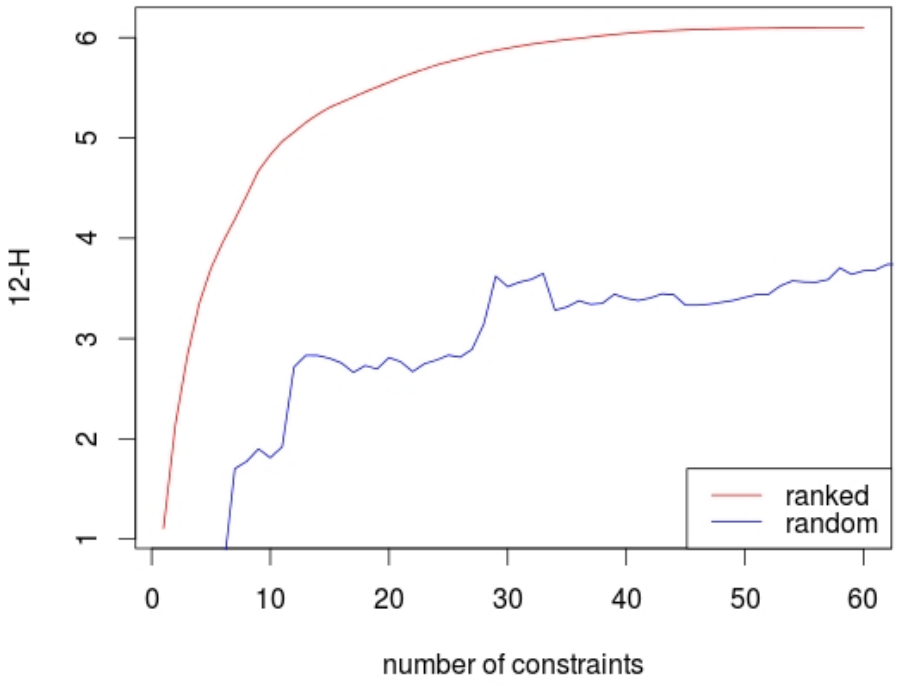

Figure 4: Add S_2^4 constraints in ranked and random strategy. With ranked constraints, the information content converges quickly while with random constraints, the information content converges very slowly.

The maximum entropy models for skipped exons and conservative exons are quite deifferent (Figure 5), which inspires us to test the models in real data. However, the result is not as good as we have expected(Figure 6).

conservative exon		skipped exon	
pattern	ΔH	pattern	ΔH
A	1.6081	*T*A**	1.7234
T***G*	0.9421	***AG*	1.1829
*TACC	0.5868	CTTACC	0.9067
CAG	0.4190	***ACC	0.7882
***CC	0.3618	G*****	0.5323
*GA**	0.3294	TTCAGG	0.2413
*A***	0.2890	T***G*	0.4333
T*A**	0.2813	C****C	0.4169
G**A**	0.1607	****GC	0.2129
C****	0.1567	*****T	0.1382
ACAGA	0.1333	*TA***	0.1056
AGTACC	0.1207	**CA**	0.0562
CTCACC	0.1101	T**A**	0.0337
TTCAGG	0.0885	C**A**	0.0392
**AGG	0.0732	TTCAGA	0.0284
A***C*	0.0704	**TA**	0.0527
T***C*	0.0546	TTTAGG	0.0216
*C**T	0.0590	T*T***	0.0466
TTTAGT	0.0472	*TTAGG	0.0262
T**C*	0.0485	C*****	0.0342

Table 2: The top 20 informative patterns in skipped and conservative exons.

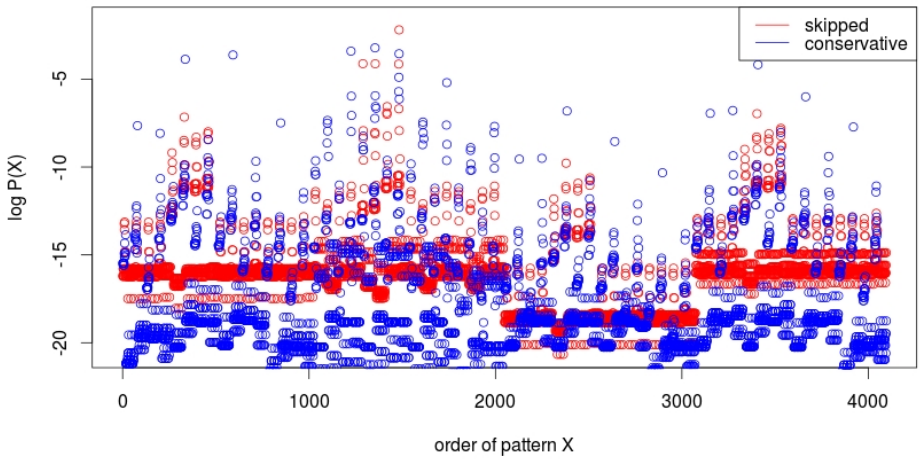

Figure 5: The distribute of skipped model and conservative model.The two models are very different.

The value of likelihood $L(X = x)$ which indicates the occurrence of skipping isn't significantly large in skipped exon sequences or significantly small in conservative ones. We discussed over the results and made some assumptions. One possible reason may be that, the mechanism of alternative splicing is so complicated that the 6 base-pair upstream sequence only provide very little information. The convergence of information content is not because of our model approaching the real distribution, but of using up all the information the constraints carries. The iteration stops at a local optimal points rather than a global one. This can be inferred from the estimation. Even though ΔH is very small, the convergent curve is very nice, there are still many components of the estimated distribution with same value. The information is not enough to distinguish between the components so they stay uniform with each other.

Another possible reason lies in the strategy we choose exons. The skipped exons can be spliced into transcripts sometimes. The signal sequence, if there

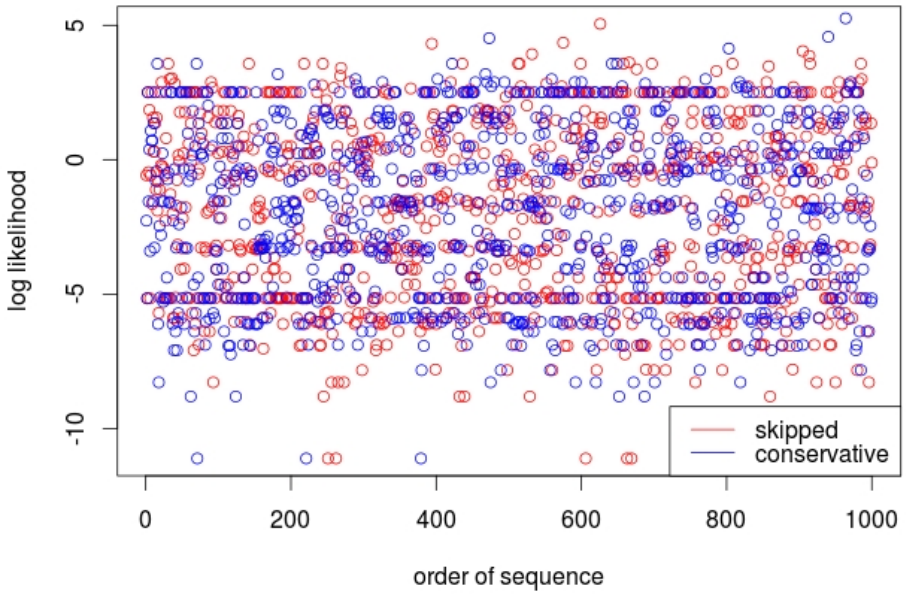

Figure 6: The likelihood corresponding to different models of the first 1000 sequences in test data.The differences are not significant enough to tell skipped exons from conservative ones.

is any, should contain both splicing and alternative splicing signals. The splice signal may be so strong that it covers the alternative splicing signals. In Burge's work, the short sequences matching the patterns of ***GT**** and **********AGN***, instead of real conservative site sequences, are used as decoy 5' and 3' site background. This strategy relies on the empirical knowledge of splicing mechanism, which is not easy for the alternative splicing situation.

Back to the biological data, it's not easy to pre-assume which pattern of motif determines or influence the process of alternative splicing and where the motif with significance locates in the genome. The important motif may be far away from the splicing cite and the secondary structure of sequence should be taken into consideration. Next, we may update the model in many aspects, for example, extracting a slightly longer sequence in both upstream and downstream, constructing the background data using Burge's strategy and adding high order constraints with skips in positions.

REFERENCE

[1] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, 40(12):1413–1415, Dec 2008.

[2] A. J. Matlin, F. Clark, and C. W. Smith. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.*, 6(5):386–398, May 2005.

[3] K. H. Lim, L. Ferraris, M. E. Filloux, B. J. Raphael, and W. G. Fairbrother. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc. Natl. Acad. Sci. U.S.A.*, 108(27):11093–11098, Jul 2011.

[4] Y. Barash, J. A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B. J. Blencowe, and B. J. Frey. Deciphering the splicing code. *Nature*, 465(7294):53–59, May 2010.

[5] Z. Wang and C. B. Burge. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, 14(5):802–813, May 2008.

[6] M. Sammeth, S. Foissac, and R. Guigo. A general definition and nomenclature for alternative splicing events. *PLoS Comput. Biol.*, 4(8):e1000147, 2008.

[7] A. M. Zahler. Pre-mRNA splicing and its regulation in Caenorhabditis elegans. *WormBook*, pages 1–21, 2012.

[8] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957.

[9] G. Yeo and C. B. Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, 11(2-3):377–394, 2004.

[10] J. Takeda, Y. Suzuki, R. Sakate, Y. Sato, T. Gojobori, T. Imanishi, and S. Sugano. H-DBAS: human-transcriptome database for alternative splicing: update 2010. *Nucleic Acids Res.*, 38(Database issue):86–90, Jan 2010.

[11] C Terrance Ireland and Solomon Kullback. Contingency tables with given marginals. *Biometrika*, 55(1):179–188, 1968.