Extensive Search for Discriminative Features of Alternative Splicing

H. Sakai and O. Maruyama

Pacific Symposium on Biocomputing 9:54-65(2004)

EXTENSIVE SEARCH FOR DISCRIMINATIVE FEATURES OF ALTERNATIVE SPLICING

H. SAKAI[a]
Graduate School of Mathematics, Kyushu University, Fukuoka 812-8581, Japan

O. MARUYAMA[b]
Faculty of Mathematics, Kyushu University, Fukuoka 812-8581, Japan

Abstract

Alternative pre-mRNA splicing events can be classified into various types, including cassette, mutually exclusive, alternative 3' splice site, alternative 5' splice site, retained intron. The detection of features of a particular type of alternative splicing events is an important and challenging problem in understanding the mechanism of alternative splicing. In this paper, we consider the problem of finding regulatory sequence patterns, which are specific to a particular type of alternative splicing events, on alternative exons and their flanking introns. For this problem, we have designed various pattern features and evaluated them on the alternative splicing data compiled in Lee's ASAP (Alternative Splicing Annotation Project) database. Through our work, we have succeeded in finding features with practically high accuracies.

1 Introduction

Nowadays one of the greatest challenging problems in biology is to elucidate the whole picture of alternative splicing because alternative splicing is a central mechanism to generate the functional complexity of proteome. It was assumed that for a long time that alternative splicing was an exceptional event and, in most cases, the sequence of exons unique to an ORF was spliced. The completion of large genomic sequencing projects, however, revealed that metazoan organisms abundantly use alternative splicing. For example, the draft sequences of the human genome published in 2001 led to the surprisingly low number of genes, about 30,000 $\sim$ 40,000 genes, as compared with a figure of over 100,000 which was previously estimated.

It is one of the very important subproblems to detect regulatory sequence elements for alternative pre-mRNA splicing events. For this issue, Brudno *et al.* [1] focused their attention on *tissue*, and detected candidate intron regulatory sequence elements for tissue-specific alternative splicing. Thanaraj and

[a] ma202033@math.kyushu-u.ac.jp
[b] om@math.kyushu-u.ac.jp

Stamm[2] summarized from the literature, regulatory elements at 5' splice sites and 3' splice sites, and exonic elements.

Alternative pre-mRNA splicing events can be classified into various types, including cassette, mutually exclusive, alternative 3' splice site, alternative 5' splice site, retained intron[2]. The detection of regulatory sequence elements closely related to such a particular type of alternative splicing events is also an important and challenging problem in understanding the mechanism of alternative splicing. However, it seems that it has not been given enough extensive computational analysis of examining whether there are candidate regulatory sequence elements characterizing types of alternative splicing events.

In this paper, we consider the problem of finding regulatory sequence patterns, which are specific to the types of alternative 5' splice site, alternative 3' splice site, and cassette, respectively, on their alternative exons and flanking introns. The data on alternative splicing which we use in this work is the product of Lee's ASAP (Alternative Splicing Annotation Project)[3].

The approach we take for this problem is based on various feature designs. In general, it is very important how to look at the raw data, i.e., designing and selecting appropriate models of features (or attributes) on the data in the process of knowledge discovery (see for example[4]), because it is necessary to detect features appropriate for explaining the data suitably in the process of discovering something new from the data. Since we have not had any deep insight into appropriate pattern models for regulatory sequence elements for alternative splicing events yet, we take the approach of designing and testing various features on sequences.

In this task, we consider, on DNA sequences, the various kinds of patterns: l-mers with some mismatches, strings over IUPAC nucleic acid codes, called degenerate patterns, and nucleic acid indexing, which is similar to amino acid indexing[5]. An alphabet indexing[6] is a classification of characters of an alphabet, by which an original sequence is transformed into a sequence over a smaller alphabet. On the sequences alphabet-indexed from DNA sequences, substring patterns are searched. Since all the patterns we use here are formulated as binary functions, we can deal with conjunctions and disjunctions of them easily. Such composite patterns are also evaluated. In finding discriminative sequence elements, it is also an important factor to locate search regions adequately. Through this approach, we have succeeded in finding discriminative features with practically high accuracies and reported the results.

This paper is organized as follows: In Section 2, we describe the materials and methods we use in this work. The results we have attained in our computational analysis are reported in Section 3. We describe concluding remarks in Section 4.

2 Materials and Methods

In this section, we describe data of alternative splicing, and sequence feature designs, including pattern modeling, pattern matcher specification and search region arrangements. The score function we use is described here.

2.1 Data

Lee *et al.* [3] have compiled information related to alternative splicing, and the result is available as an online database ASAP (Alternative Splicing Annotation Project). The text files of this database can be downloaded at the site, `http://www.bioinformatics.ucla.edu/HASDB/`. An entry of the database has a column indicating how much evidence we have for the alternative splicing event. The value "multiple" means that both splices have at least two ESTs or at least one mRNA observation. All other alternative splices are indicated by "single". The entries we use here are restricted to the ones where their evidences are labeled by "multiple". Through our computational experiments, all the constitutive exons are considered to be *negative examples.* On the other hand, all the alternative exons involved in the alternative splicing type in question, for example, the type of alternative 5' splice site, are used as *positive examples.* The sequences related to those exons are called *negative and positive sequences*, respectively.

In the definition of the alternative 5' and 3' splice site events, we use a strict version. It is required that the non-alternative splice sites of the two overlapped alternative exons should be located in the same position.

2.2 Designing Features on Sequences

Search Region Arrangements

It is mentioned in [7,2] that regulatory elements known as silencers or enhancers can be *intronic* or *exonic.* Reflecting this knowledge, we exhaustively search patterns on the regions of alternative exons and their flanking regions.

For two alternative exons e_1 and e_2 of alternative 5' splice sites, four kinds of search regions, which are called *upstream*, *overlapped exonic*, *non-overlapped right exonic*, and *downstream* regions, are defined in (A) of Fig. 1. The length l of the upstream and downstream regions is set at 100 nt, which is the same as the previous work on finding candidate intron regulatory sequence elements for tissue-specific alternative splicing[1]. In the same way, the four kinds of search regions for alternative exons of alternative 3' splice sites are defined (see (B)

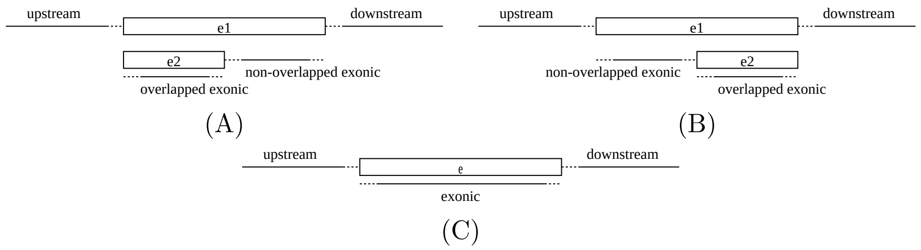

Figure 1: The search regions are shown as solid lines. Note that a search region does not contain any splice sites. It is apart from any splice sites at least 8 bp (shown as dotted lines), to avoid any influence of splice site consensus.

of Fig. 1). For cassette and constitutive exons e, the upstream, exonic and downstream regions are defined (see (C) of Fig. 1).

Pattern Models

Let Σ be a finite alphabet. The patterns we use in this work are mismatch patterns, degenerate patterns, numerical indexing patterns, and substring patterns over alphabet-indexed sequences. We describe the details of these patterns here.

- A *substring pattern* over Σ is a string p over Σ. The substring pattern matcher we use here returns, given a string t over Σ, `true` if there is at least one occurrence of p in t, and `false` otherwise.

- A *mismatch pattern* over Σ is a pair of a string p over Σ and a nonnegative integer k. The mismatch pattern matcher returns `true` if there is at least one substring of a given string t identical to p except at most k positions, and `false` otherwise.

- A *degenerate pattern* [8] over Σ is a sequence of subsets of Σ. For a degenerate pattern $p = p_1 p_2 \cdots p_n$ with $p_i \subseteq \Sigma$ for $i = 1, 2, \ldots, n$, the degenerate pattern matcher returns `true` if there is at least one substring $s = s_1 s_2 \cdots s_n$ $(s_i \in \Sigma)$ of a given string t such that s_i is included in p_i for each $i = 1, 2, \ldots, n$, and `false` otherwise. When Σ is set at the nucleotide set $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}, \mathrm{U}\}$, degenerate patterns are identical to strings over the IUPAC nucleic acid codes. The *degeneracy* of p is defined as the value of $\prod_{i=1}^{l} |p_i|$.

Notice that the occurrences of the above patterns conserve the order of the characters occurring in the patterns completely or incompletely. For example, a mismatch pattern p = ACGT with at most one mismatches matches the strings including as substrings, $*$CGT, A $*$ GT, AC $*$ T, or ACG$*$, where $*$ means any one symbol. Trivially, these substrings conserve the sequence of the characters of p partially, that is, A,C,G, and T. Then we also consider quite a different type of pattern models which do not have any constraint on the order of the characters of the substrings which those patterns match. We introduce a *numerical indexing*, which is a mapping from a finite alphabet Σ to a numerical value set V. This is a generalization of an *amino acid indexing*, a mapping from one amino acid to a numerical value [5].

Definition 1 (numerical indexing)
Let Σ be a finite alphabet, and V a set of numbers. For a given numerical index $I : \Sigma \to V$ and a string $s = s_1 s_2 \cdots s_n$ in Σ^* ($s_i \in \Sigma$ for $i = 1, 2, \ldots, n$), let $I(s)$ denote the homomorphism $(I(s_1); I(s_2); \cdots; I(s_n))$, where $(;)$ denotes a sequence of values. We will call $I(s)$ the *numerical-indexed* string.

A numerical indexing from Σ to V is called a *nucleic acid indexing* when Σ is the set of the nucleotides. A *numerical indexing pattern* is defined by a pair of a numerical indexing I and a threshold τ. The matcher of numerical indexing patterns we use here returns, given a string $t \in \Sigma$, `true` if the value of the function $\max \text{avg}_w$ is greater than or equal to τ, where $\max \text{avg}_w(I(t))$ is the average of a substring of size w in $I(t)$, which gives the maximum value (i.e. $\max\{I(t') \mid t = xwt'y, |t'| = w\}$). It returns `false` otherwise.

An *alphabet indexing* [6] is a classification of characters of an alphabet. This can be used as a preprocess of transforming original DNA sequences into degenerate sequences over a smaller alphabet. It is formally defined as follows:

Definition 2 (alphabet indexing[6])
An *alphabet indexing* Ψ is a mapping from one alphabet Σ to another alphabet Γ, where $|\Gamma| \leq |\Sigma|$. For $x = x_1 x_2 \cdots x_l$ in Σ^l, let $\Psi(x)$ denote the homomorphism $\Psi(x_1)\Psi(x_2) \cdots \Psi(x_l)$ in Γ^l. We will call $\Psi(x)$ the *alphabet-indexed* string.

On alphabet-indexed sequences, substring patterns are searched.

Notice that the returned values of the patterns mentioned in this section are binary. Thus, the conjunction (i.e., logical product) and disjunction (logical sum) of any two those patterns are defined and can be calculated.

2.3 Search Space

We here describe the search spaces of the patterns given in Section 2.2 and how to search them.

mismatch pattern: For a specified length l, all the substrings of l in the positive sequences are evaluated. For each of the strings, mismatch is allowed to be at most one. These mismatch patterns are evaluated in the following procedure.

1. Let P and N be sets of positive and negative sequences, respectively.
2. Let S be the set of all the substrings of length l in P.
3. For each $s \in S$, assign the set of the indexes (I, j) to s such that s is a substring of j-th sequence in I where I is either P or N. The set assigned to s is denoted by $assign(s)$.
4. For $s \in S$, let $L^l(s)$ be the set of strings t of length l such that s matches t with at most one mismatch, and calculate $\cup_{x \in L^l(s)} assign(x)$.

This procedure runs in $O(||P|| \cdot l \cdot (|P|+|N|)\log(|P|+|N|)+||N||)$, where for a set S of strings, $||S||$ denotes the sum of the lengths of the strings in S.

degenerate pattern: The length l of a pattern is also set at 4, 5 and 6. The degeneracy is set at 4. These patterns are calculated in a way similar to the mismatch patterns.

numerical indexing pattern: For the length parameter l, we consider a numerical indexing such that each character is assign an nonnegative integer in the range from 0 to l. A local search method is used to find high score patterns. However, it is still rather time-consuming, thus the threshold τ is fixed to be $l - 2$, and the length parameter l is restricted to be 6.

substring pattern on alphabet-indexed sequences: The alphabet indexing we consider here classifies the four nucleotides into two even-sized subcategories. For example, we would use $\Psi(A) = \Psi(C) = 0$ and $\Psi(G) = \Psi(T) = 1$ where $\Gamma = \{0, 1\}$. On the alphabet-indexed sequences over Γ, all the substrings of length 4, 5, and 6, which are extracted from the positive ones, are evaluated as substring patterns.

conjunction and disjunction: On each data set of alternative splicing types, the conjunctions and disjunctions of all the pairs of the top three patterns for a search region, a pattern model, a pattern length are evaluated.

2.4 Score Function

We here describe a score function F of patterns, whose value is called a *contrast* score. The contrast score based on the frequencies of occurrences of a pattern on a sequence is used in[1]. However, we use contrast score based on the binary values depending on whether there exists an occurrence of a specified pattern or not, which is defined as follows. Let p be a pattern, and let T be a set of strings. By $T(p)$ we denote the number of the strings t in T such that there is at least one occurrence of p in t. The contrast score function F returns, given a positive sequence set P and a negative sequence set N, the value $P(p)/|P| - N(p)/|N|$.

3 Results

In this section, we report results of computational experiments for searching for discriminative patterns on the search regions, characterizing the alternative splicing types of alternative 5' splicing, alternative 3' splicing, and cassette. The numbers of entries of alternative 5' splicing, alternative 3' splicing, and cassette in the ASAP database[3] are 227, 249, and 1299, respectively. As mentioned in Section 2.1, all the constitutive exons are used as negative examples, whose total number is 39,993. At first, our program is executed for all the combination of alternative splicing types, search regions, and pattern models. At the next stage, conjunctions and disjunctions derived from the patterns found in the previous stage are evaluated.

3.1 Alternative 5' Splice Site

The patterns found in alternative exons of alternative 5' splicing and their flanking introns are listed in Table 1.

Table 1: The patterns found in alternative exons of alternative 5' splicing and their flanking introns. The column labeled by R indicates a search region. U, O, N, and D denote the upstream, overlapped exonic, non-overlapped exonic, and downstream regions, respectively. The column labeled by C indicates a model of patterns. DP, MP, AI+SP, and NIP, denote the classes of degenerate patterns, mismatch patterns, substring patterns over alphabet-indexed sequences, and numerical indexing patterns, respectively. For a class C and a natural number l, Cl denotes a subset of C, $\{s \in C \mid \text{the length of } s \text{ is } l\}$. The columns labeled by $F, P(p)/|P|$ and $N(p)/|N|$ show, for the pattern in a row, the contrast scores, the ratio of true positives, and the ratio of false positives, respectively. In the last three rows, top three composite patterns, which are conjunction or disjunction of two single patterns, are listed.

R	C	pattern	F	$P(p)/\|P\|$	$N(p)/\|N\|$
U	DP	G[CT]C[CG]	16.44	68.28	51.83
		CC[CG][AC]	16.35	68.72	52.36
		CCC[AT][AT]	15.32	50.22	34.89
	MP	ACCCC	17.22	61.23	44.01
		CCCGT	16.02	60.35	44.32
		CGTCC	16.00	57.26	41.26
	AI+SP	000010(A,T=1)	14.35	80.17	68.72
		10000 (A,T=1)	12.58	80.17	67.59
		00001 (A,T=1)	12.47	80.17	67.69
	NIP6	{A:any,C:6,G:4,T:any} τ=4	11.80	89.42	77.62
O	DP	C[AC]C[AG]G	11.29	53.30	42.00
		CC[CT][CG]G	10.92	49.33	38.41
		C[AC]C[AC]G	10.49	49.33	38.41
	MP	CGCGG	10.16	48.01	37.85
		CGCGGA	9.28	26.87	17.58
		GACCA	9.11	81.93	72.82
	AI+SP	000100(A,T=1)	5.83	71.80	65.97
		001000(A,T=1)	4.81	70.92	66.10
		010000(A,T=1)	4.63	68.28	63.65
	NIP6	{A:any,C:4,G:6,T:any} τ=4	12.20	51.10	38.90
N	DP	GG[CG]TCC	5.61	12.77	7.15
		GG[CGT]TCC	5.05	14.09	9.04
		CGAG[CG][CG]	4.77	7.48	2.71
	MP	GCGCGG	-0.11	15.85	15.97
		GGCGCG	-0.25	14.97	15.97
		TAGGGT	-0.77	11.89	12.67
	AI+SP	000000(A,T=1)	-5.85	30.83	36.68
		000001(A,T=1)	-13.60	42.73	56.33
		00000(A,T=1)	-13.64	43.17	56.81
	NIP6	{A:any,C:4,G:5,T:any} τ=4	1.12	15.45	14.33
D	DP	G[AC]GG[AG]	17.01	49.33	32.31
		G[AG]GG[AG]	15.19	55.94	40.74
		G[CG][AG]GA	15.11	51.98	36.87
	MP	GCGGA	17.01	61.23	44.22
		GGAGGA	16.51	51.10	34.58
		GAGGAG	15.85	49.77	33.92
	AI+SP	001001(A,T=1)	12.14	78.85	66.70
		010000(A,T=1)	11.11	70.48	59.36

		001100(A,T=1)	10.98	69.60	58.62
	NIP6	{A:4,C:1,G:5,T:1} τ=4	12.84	50.74	38.90

(Pattern,R) × (Pattern,R)	F	$P(p)/\lvert P\rvert$	$N(p)/\lvert N\rvert$
(C[AC]CC[GT],U) OR (G[AG]GGA[AG],D)	21.41	64.75	43.34
([GT]CC[CG]C,U) OR (G[AG]GGA[AG],D)	21.30	62.55	41.24
([GT]CC[CG]C,U) OR (G[AC]GG[AG],D)	21.27	80.17	58.90

3.2 Alternative 3' Splice Site

The patterns found in alternative exons of alternative 3' splice sites and their flanking intronic regions are listed in Table 2. The top three patterns in this table are mismatch pattern CGGGG (22.51), CGGGGA (21.17), and degenerate pattern [CG][AG]GGG (20.34), which are patterns on downstream regions. Notice that these pattern share the string CGGGG at least. As for a G-rich element, it is known that intronic G triplets are frequently located adjacent to 5' splice sites, which would correspond to the left ends of the downstream search regions, and bind U1 snRNPs to enhance splicing and select 5' splice sites[9]. This knowledge would imply that those found patterns capture regulatory elements of alternative 3' splicing because of the fact that those patterns are also frequently occurred on the downstream search regions.

Table 2: The patterns found in alternative exons of alternative 3' splicing and their flanking introns.

R	C	pattern	F	$P(p)/\lvert P\rvert$	$N(p)/\lvert N\rvert$
U	DP	GG[ACT]C	18.96	73.89	54.92
		[CG]G[CT]C	18.59	69.47	50.88
		GG[CT][CG]	17.72	78.71	60.99
	MP	CCCCG	18.23	55.82	37.58
		GGTCG	17.08	54.61	37.52
		AGCCC	16.02	66.66	50.64
	AI+SP	000100(A,T=1)	16.89	72.69	55.79
		001000(A,T=1)	16.37	72.28	55.91
		100000(A,T=1)	14.29	61.04	46.75
N	DP	CGCCC	3.91	10.84	6.92
		CG[CT]CC	3.70	14.85	11.15
		CGC[CG]C	3.61	12.04	8.43
	MP	CCCGCG	1.79	19.67	17.88
		CCCCCG	1.02	22.48	21.46
		CGCCCG	0.73	17.67	16.93

	AI+SP	000000(A,T=1)	-8.57	28.11	36.69
		000001(A,T=1)	-16.17	40.16	56.33
		00000(A,T=1)	-16.65	40.16	56.81
O	DP	[CT][AC]CG	15.25	63.45	48.19
		C[AC][CG]G	14.11	77.10	62.99
		CG[ACG]G	14.08	55.02	40.93
	MP	CTCCGG	15.49	42.57	27.07
		GACACT	14.42	46.58	32.16
		CCGGAG	14.36	45.38	31.16
	AI+SP	000010(A,T=1)	12.74	76.70	63.96
		010000(A,T=1)	11.84	75.50	63.65
		100000(A,T=1)	9.87	66.26	56.38
D	DP	[CG][AG]GGG	20.34	61.04	40.69
		CCC[AC][CG]	19.93	56.62	36.69
		[CG][AC]GGG	19.72	58.63	38.90
	MP	CGGGG	22.51	66.26	43.75
		CGGGGA	21.17	44.57	23.39
		CGGGA	19.89	68.27	48.37
	AI+SP	000001(A,T=1)	17.77	69.47	51.70
		000100(A,T=1)	17.58	78.71	61.13
		00000(A,T=1)	17.50	69.47	51.97

(Pattern,R) × (Pattern,R)	F	$P(p)/\|P\|$	$N(p)/\|N\|$
(GG[ACT]C,U) AND (C[CG]C[AC],D)	24.30	61.04	36.74
(GG[ACT]C,U) AND (000100,D)	23.38	66.26	42.88
(GG[ACT]C,U) AND (CCC[AG],D)	22.93	55.82	32.88

3.3 Cassette

The patterns specific to the type of cassette are given in Table 3. Comparing with the other results, the contrast scores of these patterns are lower.

Table 3: The patterns found in alternative cassette exons and their flanking introns. The symbol 'E' in the column labeled by R denotes the exonic region.

R	C	pattern	F	$P(p)/\|P\|$	$N(p)/\|N\|$
	DP	[AG]TTT	10.03	76.52	66.48
		TTT[GT][AG]	9.99	60.43	50.43
		[AGT]TTT	9.43	80.90	71.47

U	MP	GTTTTT	8.88	64.97	56.08
		CGTTT	8.75	68.89	60.14
		TATTTT	8.56	63.43	54.86
	AI+SP	101111(A,T=1)	9.18	79.44	70.26
		111101(A,T=1)	9.16	79.52	70.35
		11111(A,T=1)	9.15	76.90	67.74
E	DP	[CT]TA[AG]	4.37	47.72	43.35
		T[CT]TCT	4.31	24.08	19.70
		[AT][CT]TCT	4.30	35.79	31.48
	MP	TCTTTT	3.99	35.)2	31.02
		CTTAGC	3.30	25.94	22.63
		TTTAGT	3.16	21.09	17.92
	AI+SP	1111(A,T=1)	2.17	75.90	73.72
		01111(A,T=1)	1.93	75.13	73.20
		101111(A,T=1)	1.75	59.58	57.82
D	DP	T[AT]T[AT]	10.33	73.97	63.64
		T[AT][AT]T	9.49	73.28	63.79
		[AT][AT]TT	9.26	74.67	65.41
	MP	TTAAT	8.71	70.13	61.41
		ATTTT	8.58	78.90	70.32
		TTTTA	8.57	76.52	67.94
	AI+SP	11111(A,T=1)	9.21	74.13	64.91
		111101(A,T=1)	9.18	76.44	67.25
		111110(A,T=1)	8.95	73.44	64.48

(Pattern,R) × (Pattern,R)	F	$P(p)/\|P\|$	$N(p)/\|N\|$
(TTT[GT][AG],U) OR ([AG]T[AT]TT,D)	12.01	75.57	63.66
(TTT[GT][AG],U) OR ([CT]TTTT[CG],D)	11.83	70.43	58.60
([AG]TT[AT],U) AND (111101,D)	11.32	68.36	57.03

4 Discussion

Through Table 1, 2, and 3, we can see that a higher score is obtained by composing single patterns. An interesting point is that, the two search regions of any composite pattern in the tables are a pair of upstream region (U) and downstream region (D).

As for the substring patterns on the alphabet-indexed sequences, high score patterns share the same alphabet indexing which separates A and T from C and G. Notice that this fact is not dependent on the types of alternative splicing.

As for a score function, we have also examined a function based on the frequencies of patterns p on a sequence t, instead of whether there is at least one occurrence of p on t, which is used in our score function F. Through our computational experiments, we have compared the two score functions, and F looks better than the frequency version (data is not shown).

Acknowledgments

The authors would like to thank anonymous referees for valuable comments. This study was supported from the Research for the Future Program of the Japan Society for the Promotion of Science.

References

1. M. Brudno, M.S. Gelfand, S. Spengler, M. Zorn, I. Dubchak, and J. G. Conboy. Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. *Nucl. Acids. Res.*, 29:2338–2348, 2001.
2. T.A. Thanaraj and S. Stamm. *Prediction and Statistical Analysis of Alternatively Spliced Exons*, pages 1–31. Progress in Molecular and Subcellular Biology **31**. Springer-Verlag, 2003.
3. C. Lee, L. Atanelov, B. Modrek, and Y. Xing. ASAP: the alternative splicing annotation project. *Nucl. Acids. Res.*, 31:101–105, 2003.
4. O. Maruyama and S. Miyano. Design aspects of discovery systems. *IEICE Transactions on Information and Systems*, E83-D:61–70, 2000.
5. H. Bannai, Y. Tamada, O. Maruyama, K. Nakai, and S. Miyano. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, 18:298–305, 2002.
6. S. Shimozono. Alphabet indexing for approximating features of symbols. *Theo. Comp. Sci.*, 210:245–260, 1999.
7. T.A. Cooper and W. Mattox. The regulation of splice-site selection, and its role in human disease. *Am J Hum Genet.*, 61:259–266, 1997.
8. D. Shinozaki, T. Akutsu, and O. Maruyama. Finding optimal degenerate patterns in dna sequences. In *Proc. European Conference on Computational Biology (ECCB 2003), Bioinformatics*, 2003. To appear.
9. A.J. Mccullough and S.M. Berget. An intronic splicing enhancer binds U1 snRNPs to enhance splicing and select 5' splice sites. *Molecular and Cellular Biology*, 20:9225–9235, 2000.