Prediction of Gene Structures from RNA-seq Data Using Dual Decomposition

TATSUMU INATSUKI¹ KENGO SATO^{1,a)} YASUBUMI SAKAKIBARA^{1,b)}

Received: May 21, 2015, Accepted: July 25, 2015

Abstract: Numerous computational algorithms for predicting protein-coding genes from genomic sequences have been developed, and hidden Markov models (HMMs) have frequently been used to model gene structures. For eukaryotes, more complex gene structures such as introns make gene prediction much harder due to isoforms of transcripts by alternative splicing machinery. We develop a novel gene prediction method for eukaryote genomes that extends the traditional HMM-based gene prediction model by incorporating comprehensive evidence of transcripts by using RNA sequencing (RNA-seq) technology. We formulate gene prediction as an integer programming problem, and solve it by the dual decomposition technique. To confirm the utility of the proposed algorithm, computational experiments on benchmark datasets were conducted. The results show that our algorithm efficiently and effectively employs RNA-seq data in gene structure prediction.

Keywords: gene structure prediction, RNA-seq, hidden Markov models, Lagrangian relaxation, dual decomposition

1. Introduction

Numerous computational algorithms for predicting protein-coding genes from genomic sequences have been developed, and hidden Markov models (HMMs) have frequently been used to model gene structures. Protein-coding genes can successfully be predicted for prokaryotes since no intronic sequences are included. In contrast, for eukaryotes, more complex gene structures (such as introns) make gene prediction much harder due to isoforms of transcripts by alternative splicing machinery. Because of this, adequate accuracy of gene prediction has not yet been realized for higher organisms such as human that have complex gene structures.

Existing methods for computational prediction of gene structures can be roughly classified into three groups. The first group is called the *ab initio* approach, which predicts gene structures from genomic sequences alone. HMMs are frequently employed for this approach. GENSCAN [5] and AUGUSTUS [22] fall into this group. The second group additionally uses aligned DNA sequences from other genomes to observe conservation patterns that help in making accurate predictions because protein-coding sequences show typical conservation. N-SCAN [11] and CONTRAST [12] implement this approach. The final group, including CONTRAST [12] and AUGUSTUS+ [20], [21] (an extension of AUGUSTUS), uses transcriptional evidence such as proteins, expressed sequence tags (ESTs), and RNA-sequencing (RNA-seq) reads, which support the gene structures, to make predictions. The model proposed in this paper takes this approach.

In this study, we develop a novel gene prediction method for eukaryote genomes. The method extends the traditional HMM-based gene prediction model by incorporating comprehensive evidence of transcripts by RNA-seq technology. We formulate gene prediction as an integer programming problem whose objective function is the sum of the HMM-based score for gene structures and the number of RNA-seq reads that support the gene structure. The algorithm calculates an optimal gene structure that maximizes the objective function subject to several constraints that should be satisfied by the predicted gene structure and the observed RNA-seq reads. In contrast to traditional HMM-based gene prediction algorithms, dynamic programming techniques cannot be applied to this optimization problem because of the additional constraints imposed for RNA-seq reads. We therefore use dual decomposition, which iterates over the following steps. (1) The constraints on RNA-seq reads are relaxed by Lagrangian relaxation. As a result, the original optimization problem is decomposed into two independent sub-problems: HMM-based gene structure prediction and supported read-maximization. These sub-problems can efficiently be solved by the Viterbi algorithm and coefficient comparison. (2) The consistency of RNA-seq constraints is maintained by imposing score penalties on inconsistent predictions. To confirm the proposed algorithm, we conducted computational experiments on benchmark datasets. The results show that our algorithm with RNA-seq data predicts gene structures significantly more accurately than do other methods.

2. Methods

2.1 Preliminaries

As several existing works do, we construct a gene model as an HMM in which emission symbols correspond to nucleotides and hidden states correspond to internal gene structures such as ex-

¹ Department of Biosciences and Informatics, Keio University, Yokohama, Kanagawa 223–8522, Japan

^{a)} satoken@bio.keio.ac.jp

^{b)} yasu@bio.keio.ac.jp

Fig. 1 The binary valued matrix $y \in \mathcal{Y}(x)$ that represents the assignments of gene structures for each nucleotide. Each gray circle represents $y_{il} = 1$, that is, x_i is assigned g_l .

ons and introns. We employ m kinds of gene structures, denoted $G = \{g_1, g_2, \dots, g_m\}$. Let G_{exon} and G_{intron} be the subsets of G that contain gene structures corresponding to exons and introns, respectively. See Section A.1 in the Appendix for details.

Let Σ be the set of four DNA bases (adenine (A), cytosine (C), guanine (G), and thymine (T)) and let Σ^* denote the set of all finite DNA sequences, which consist of bases in Σ . Given a DNA sequence $x = x_1, \dots, x_n \in \Sigma^*$ consisting of n bases, let $\mathcal{Y}(x)$ be the space of all possible gene structures of x . An element $y \in \mathcal{Y}(x)$ is represented as an $n \times m$ binary-valued matrix, where $y_{il} = 1$ indicates that the base x_i is assigned to the gene structure $g_l \in G$ as shown in Fig. 1. We define the problem of gene structure prediction as follows: given a DNA sequence x , predict a gene structure $y \in \mathcal{Y}(x)$.

2.2 Scoring Model

A scoring function f is a function that assigns real-valued scores to pairs consisting of a DNA sequence x and a gene structure $y \in \mathcal{Y}(x)$. Our aim is to find a gene structure $y \in \mathcal{Y}(x)$ that maximizes the scoring function $f(x, y)$ for a given DNA sequence x . The scoring function f consists of two parts: a gene-model-based scoring function f_{gene} and an evidence-based scoring function f_{evidence} .

The gene-model-based scoring function f_{gene} is defined as

$$f_{\text{gene}}(x, y) = \sum_{i=1}^n \sum_{l=1}^m \mu_{il} y_{il} + \sum_{i=2}^n \sum_{l=1}^m \sum_{k=1}^m \nu_{lk} y_{i-1l} y_{ik}, \quad (1)$$

where μ_{il} is a parameter of the preference of the gene structure g_l at x_i , and ν_{lk} is a parameter of the preference of transition of the gene structures from g_l to g_k . Since each base is assigned to exactly one gene structure as shown in Fig. 1, the following constraint must be satisfied:

$$\sum_{l=1}^m y_{il} = 1 \quad (\text{for } 1 \leq i \leq n). \quad (2)$$

To calculate μ_{il} , we employ three types of 5-mers for local features around x_i : those centered at x_i (i.e., $x_{i-2} \dots x_{i+2}$), those centered at x_{i-2} (i.e., $x_{i-4} \dots x_i$), and those centered at x_{i-4} (i.e., $x_{i-6} \dots x_{i-2}$). We define the corresponding parameters of the local features on a gene structure g_l as $\mu_{x_{i-2} \dots x_{i+2}, l}^{(0)}$, $\mu_{x_{i-4} \dots x_i, l}^{(-2)}$, and

$\mu_{x_{i-6} \dots x_{i-2}, l}^{(-4)}$, respectively, and use these to calculate μ_{il} as follows:

$$\mu_{il} = \mu_{x_{i-2} \dots x_{i+2}, l}^{(0)} + \mu_{x_{i-4} \dots x_i, l}^{(-2)} + \mu_{x_{i-6} \dots x_{i-2}, l}^{(-4)}. \quad (3)$$

If transcriptional evidence such as RNA-seq reads is available, we can use them to improve the accuracy of predicting gene structures. To that end, we define an evidence-based scoring function f_{evidence} that evaluates the degree of fitness of a gene structure $y \in \mathcal{Y}(x)$ against observed transcripts. We consider two types of gene structure evidence from observed transcripts. The first type is split positions of mapped transcripts, as obtained by a spliced aligner such as TopHat [15] or STAR [8], which is represented by an $n \times n$ triangular matrix $S = (S_{ij})_{i < j}$, where $S_{ij} = 1$ if at least one transcript mapped onto x is split at x_i and x_j , and $S_{ij} = 0$ otherwise. The second type of evidence is mapped bases of transcripts, represented by an n -dimensional vector $R = (R_i)$, where $R_i = 1$ if at least one transcript mapped onto x covers x_i , and $R_i = 0$ otherwise. We define f_{evidence} as the weighted sum of the number of evidence items that support a gene structure $y \in \mathcal{Y}(x)$:

$$f_{\text{evidence}}(x, y, s, r) = \alpha \sum_{1 \leq i < j \leq n} \sigma_{ij} s_{ij} + \beta \sum_{1 \leq i \leq n} \rho_i r_i, \quad (4)$$

where s_{ij} is a binary-valued variable that indicates whether the split transcripts at x_i and x_j support the gene structure y . Similarly, r_i is a binary-valued variable that indicates whether the transcripts that cover x_i support the gene structure y . Here, $\alpha \geq 0$ and $\beta \geq 0$ mean the contribution ratio of the two types of evidence, and σ_{ij} and ρ_i specify the weight for each evidence. We used $\alpha = 30$, $\beta = 0.1$, $\sigma_{ij} = 1.0$, and $\rho_i = 1.0$ for all i, j in our experiments.

To make the variables s and r consistent with the gene structure y , the following constraints must be satisfied:

$$s_{ij} \leq S_{ij} \quad (\text{for } 1 \leq i < j \leq n) \quad (5)$$

$$s_{ij} \leq \sum_{l: g_l \in G_{\text{intron}}} y_{kl} \quad (\text{for } 1 \leq i < j \leq n) \quad (6)$$

$$s_{ij} \leq \sum_{l: g_l \in G_{\text{exon}}} y_{il} \quad (\text{for } 1 \leq i < j \leq n) \quad (7)$$

$$s_{ij} \leq \sum_{l: g_l \in G_{\text{exon}}} y_{jl} \quad (\text{for } 1 \leq i < j \leq n) \quad (8)$$

$$r_i \leq R_i \quad (\text{for } 1 \leq i \leq n) \quad (9)$$

$$r_i \leq \sum_{l: g_l \in G_{\text{exon}}} y_{il} \quad (\text{for } 1 \leq i \leq n) \quad (10)$$

Here, the constraints (5) and (9) indicate that supporting transcripts can be enabled only for observed transcripts. In other words, not all the observed transcripts have to support the gene structure. The constraints (6)–(8) represent a condition of supporting transcripts where one of the intron states must be assigned into spliced regions from the $(i + 1)$ th base to the $(j - 1)$ th base, and that one of the exon states must be assigned to both ends, the i th and j th bases (Fig. 2). The constraint (10) assigns one of the exon states to the base covered by the supporting transcripts (Fig. 3).

2.3 Dual Decomposition

Our aim is to find a gene structure $y \in \mathcal{Y}(x)$ that maxi-

Fig. 2 The constraints (6)–(8) mean that if s_{ij} supports the gene structure y , i.e., $s_{ij} = 1$, then x_k for $i < k < j$ must be assigned to one of the intron states and both ends x_i and x_j must be assigned to one of the exon states.

Fig. 3 The constraint (10) means that if R_i supports the gene structure y , i.e., $r_i = 1$, then x_i must be assigned to one of the exon states.

mizes $f(x, y, s, r) = f_{\text{gene}}(x, y) + f_{\text{evidence}}(x, y, s, r)$ under the constraints (2) and (5)–(10). This is an integer programming problem for which no efficient algorithm is yet known. If we drop the constraints (6)–(8) and (10), maximization of f_{gene} and f_{evidence} can be solved separately and efficiently. This means that the constraints (6)–(8) and (10) make the problem of gene structure prediction with transcriptional evidence extremely complex. To circumvent this difficulty, we deal with these constraints using Lagrangian relaxation [16], which is an efficient technique to solve large and complex problems as appeared in the field of bioinformatics [1], [3], [17], [19]. First, we move the constraints (6)–(8) and (10) to the objective function $f(x, y, s, r)$ by applying Lagrangian multipliers:

$$L(\lambda) = \max_{y, s, r} \left\{ f_{\text{gene}}(x, y) + f_{\text{evidence}}(x, y, s, r) + \sum_{i < j} \sum_{i < k < j} \lambda_{ijk}^{(1)} \left(\sum_{l: g_l \in G_{\text{intron}}} y_{kl} - s_{ij} \right) + \sum_{i < j} \lambda_{ij}^{(2)} \left(\sum_{l: g_l \in G_{\text{exon}}} y_{il} - s_{ij} \right) + \sum_{i < j} \lambda_{ij}^{(3)} \left(\sum_{l: g_l \in G_{\text{exon}}} y_{jl} - s_{ij} \right) + \sum_i \lambda_i^{(4)} \left(\sum_{l: g_l \in G_{\text{exon}}} y_{il} - r_i \right) \right\}. \quad (11)$$

Here, $\lambda_{ijk}^{(1)} \geq 0$ ($1 \leq i < k < j \leq n$), $\lambda_{ij}^{(2)} \geq 0$ ($1 \leq i < j \leq n$), $\lambda_{ij}^{(3)} \geq 0$ ($1 \leq i < j \leq n$), and $\lambda_i^{(4)} \geq 0$ ($1 \leq i \leq n$) are Lagrangian multipliers. Equation (11) can then be transformed as:

$$L(\lambda) = \max_y \left\{ \sum_{i=1}^n \sum_{l=1}^m \mu'_{il} y_{il} + \sum_{i=2}^n \sum_{l=1}^m \sum_{k=1}^m v_{lk} y_{i-1l} y_{ik} \right\} + \max_s \sum_{1 \leq i < j \leq n} \sigma'_{ij} s_{ij} + \max_r \sum_{1 \leq i \leq n} \rho'_i r_i \quad (12)$$

where

```

1: Set  $\lambda_{ijk}^{(1)} = 0$ ,  $\lambda_{ij}^{(2)} = 0$ ,  $\lambda_{ij}^{(3)} = 0$  and  $\lambda_i^{(4)} = 0$ .
2: for  $t = 1$  to  $T$  do
3:    $\hat{y} \leftarrow \arg \max_{y \in \mathcal{Y}(x)} \sum_{i=1}^n \sum_{l=1}^m \mu'_{il} y_{il} + \sum_{i=2}^n \sum_{l=1}^m \sum_{k=1}^m v_{lk} y_{i-1l} y_{ik}$ 
4:    $\hat{s} \leftarrow \arg \max_s \sum_{1 \leq i < j \leq n} \sigma'_{ij} s_{ij}$ 
5:    $\hat{r} \leftarrow \arg \max_r \sum_{1 \leq i \leq n} \rho'_i r_i$ 
6:   if  $\hat{y}, \hat{s}, \hat{r}$  satisfy the constraints (6)–(8) and (10) then
7:     return  $\hat{y}, \hat{s}, \hat{r}$ 
8:   end if
9:    $\lambda_{ijk}^{(1)} \leftarrow \lambda_{ijk}^{(1)} - \eta_t \left( \sum_{l: g_l \in G_{\text{intron}}} y_{kl} - s_{ij} \right)$ 
10:   $\lambda_{ij}^{(2)} \leftarrow \lambda_{ij}^{(2)} - \eta_t \left( \sum_{l: g_l \in G_{\text{exon}}} y_{il} - s_{ij} \right)$ 
11:   $\lambda_{ij}^{(3)} \leftarrow \lambda_{ij}^{(3)} - \eta_t \left( \sum_{l: g_l \in G_{\text{exon}}} y_{jl} - s_{ij} \right)$ 
12:   $\lambda_i^{(4)} \leftarrow \lambda_i^{(4)} - \eta_t \left( \sum_{l: g_l \in G_{\text{exon}}} y_{il} - r_i \right)$ 
13: end for
14: return  $\hat{y}, \hat{s}, \hat{r}$ 
    
```

Fig. 4 The algorithm for predicting gene structures using dual decomposition. T is the maximum number of iterations, set at 100.

$$\mu'_{il} = \begin{cases} \mu_{il} + \sum_{j' < i < k'} \lambda_{j'ki'}^{(1)} & (\text{for } g_l \in G_{\text{intron}}) \\ \mu_{il} + \sum_{j' > i} \lambda_{ij'}^{(2)} + \sum_{j' < i} \lambda_{ji'}^{(3)} + \lambda_i^{(4)} & (\text{for } g_l \in G_{\text{exon}}) \\ \mu_{il} & (\text{otherwise}) \end{cases} \quad (13)$$

$$\sigma'_{ij} = \alpha \sigma_{ij} - \sum_{i < k < j} \lambda_{ijk}^{(1)} - \lambda_{ij}^{(2)} - \lambda_{ij}^{(3)} \quad (14)$$

$$\rho'_i = \beta \rho_i - \lambda_i^{(4)}. \quad (15)$$

We can easily see that each term of Eq. (12) can be calculated independently and efficiently. We call this technique the dual decomposition. The first term of Eq. (12) can be efficiently computed using dynamic programming techniques like the Viterbi algorithm for HMM using the following recursion:

$$V(i, l) = \begin{cases} \mu'_{il} & (\text{for } i = 1) \\ \mu'_{il} + \max_{1 \leq k \leq m} [V(i-1, k) + v_{kl}] & (\text{for } i > 1) \end{cases} \quad (16)$$

We can obtain the first term of Eq. (12) as $\max_{1 \leq k \leq m} V(n, k)$, and the optimal gene structure $\hat{y}$ can be recovered by traceback from the n th base. The second and last terms of Eq. (12) can be computed simply by finding positive coefficients for s_{ij} and r_i under the constraints (5) and (9).

The dual objective function $L(\lambda)$ gives an upper bound of the primal objective function $f(x, y, s, r)$. Therefore, we can obtain a better upper bound by minimizing Eq. (12) with respect to the multipliers. It is known that the Lagrangian function $L(\lambda)$ is convex, but not differentiable. Thus, we can apply the subgradient optimization to minimize the dual objective function (12). The Lagrangian multipliers $\lambda^{(1)}$, $\lambda^{(2)}$, $\lambda^{(3)}$ and $\lambda^{(4)}$ are iteratively updated using their subgradients, $\sum_{l: g_l \in G_{\text{intron}}} y_{kl} - s_{ij}$, $\sum_{l: g_l \in G_{\text{exon}}} y_{il} - s_{ij}$, $\sum_{l: g_l \in G_{\text{exon}}} y_{jl} - s_{ij}$, and $\sum_{l: g_l \in G_{\text{exon}}} y_{il} - r_i$, respectively. Finally, **Fig. 4** shows our algorithm for predicting gene structures. Here, $\eta_t > 0$ is a step size for each update. It is known that if $\lim_{t \rightarrow \infty} \eta_t = 0$ and $\sum_{t=1}^{\infty} \eta_t = \infty$, then the Lagrangian dual $L(\lambda)$ always converges to the optimal value. The multipliers are iteratively updated until a solution is found or the number of iterations reaches a sufficiently large maximum number of iterations T (here, we use $T = 100$).

The Lagrangian multipliers work as penalty scores against inconsistency between the gene-model-based scores and the evidence-based scores for the constraints (6)–(8) and (10).

2.4 Learning Algorithm

To optimize the feature parameters μ and ν , we employ a max-margin framework called structured support vector machines [23]. Given a training dataset $\mathcal{D} = \{(x^{(k)}, y^{(k)})\}_{k=1}^K$, where $x^{(k)} \in \Sigma^*$ and $y^{(k)} \in \mathcal{Y}(x^{(k)})$ are respectively DNA sequences and their corresponding gene structures, we aim to find μ and ν that minimize the objective function

$$\mathcal{L}(\mu, \nu) = \sum_{(x,y) \in \mathcal{D}} \left(\max_{\hat{y}, \hat{s}, \hat{r}} [f(x, \hat{y}, \hat{s}, \hat{r}; \mu, \nu) + \Delta(y, \hat{y})] - \max_{s,r} f(x, y, s, r; \mu, \nu) + C(\|\mu\|_1 + \|\nu\|_1) \right), \quad (17)$$

where $\|\cdot\|_1$ is the ℓ_1 norm and C is a weight for the ℓ_1 regularization term to avoid over-fitting to the training data. Here, $\Delta(y, \hat{y})$ is a loss function of $\hat{y}$ for y defined as

$$\begin{aligned} \Delta(y, \hat{y}) = & \delta^{\text{FN}} \sum_{i=1}^n \sum_{l=1}^m I(y_{il} = 1) I(\hat{y}_{il} = 0) \\ & + \delta^{\text{FP}} \sum_{i=1}^n \sum_{l=1}^m I(y_{il} = 0) I(\hat{y}_{il} = 1) \\ & + \delta^{\text{FN}} \sum_{i=2}^n \sum_{l=1}^m \sum_{k=1}^m I(y_{i-1l} y_{ik} = 1) I(\hat{y}_{i-1l} \hat{y}_{ik} = 0) \\ & + \delta^{\text{FP}} \sum_{i=2}^n \sum_{l=1}^m \sum_{k=1}^m I(y_{i-1l} y_{ik} = 0) I(\hat{y}_{i-1l} \hat{y}_{ik} = 1), \end{aligned} \quad (18)$$

where $I(\text{condition})$ is an indicator function that takes a value of 1 or 0 depending on whether *condition* is true or false. Here, δ^{FN} and δ^{FP} are hyperparameters to control the trade-off between sensitivity and specificity for learning the parameters. In this case, we can calculate the first term of Eq. (17) by replacing scores μ_{ij} and ν_{kl} in Eq. (12) as

$$\begin{aligned} \bar{\mu}_{ij} &= \begin{cases} \mu_{ij} - \delta^{\text{FN}} & (\text{if } y_{ij} = 1) \\ \mu_{ij} + \delta^{\text{FP}} & (\text{if } y_{ij} = 0) \end{cases} \\ \bar{\nu}_{kl} &= \begin{cases} \nu_{kl} - \delta^{\text{FN}} & (\text{if } y_{i-1l} y_{ik} = 1) \\ \nu_{kl} + \delta^{\text{FP}} & (\text{if } y_{i-1l} y_{ik} = 0) \end{cases} \end{aligned}$$

We used $C = 1.0$, $\delta^{\text{FN}} = 1.0$, and $\delta^{\text{FP}} = 1.0$ in our experiments.

To minimize the objective function (17), we can apply stochastic subgradient descent (Fig. 5) or forward-backward splitting [10].

3. Results

To verify the model described in Section 2, we developed a preliminary implementation of our model. We conducted computational experiments using EGASP [13] as training and test datasets. Our model was trained using two regions from the EGASP training dataset, ENm004 and ENm006. The model was then evaluated on 31 regions from the EGASP test dataset. Gene annotations were obtained from GENCODE [14]. We used two replicates of RNA-seq data from liver hepatocellular carcinoma cell line HepG2 [7] as transcriptional evidence. We mapped RNA-seq reads into training and test regions using STAR [8]. **Table 1** shows a summary of the datasets.

We evaluated the accuracy of predicting gene structures

```

1: Set all targeted parameters to 0.
2: repeat
3:   for all  $(x, y) \in \mathcal{D}$  do
4:      $\hat{y}, \hat{s}, \hat{r} \leftarrow \arg \max_{\hat{y}, \hat{s}, \hat{r}} [f(x, \hat{y}, \hat{s}, \hat{r}; \mu, \nu) + \Delta(y, \hat{y})]$ 
5:     for all  $s \in \Sigma^S, l \in \{1, \dots, m\}, k \in \{0, -2, -4\}$  do
6:        $\mu_{s,l}^{(k)} \leftarrow \mu_{s,l}^{(k)} - \zeta C \text{sgn } \mu_{s,l}^{(k)}$ 
7:     end for
8:     for all  $i \in \{1, \dots, n\}, l \in \{1, \dots, m\}, k \in \{0, -2, -4\}$  do
9:        $\mu_{x_{i+k-2} \dots x_{i+k+2}, l}^{(k)} \leftarrow \mu_{x_{i+k-2} \dots x_{i+k+2}, l}^{(k)} - \zeta (\hat{y}_{il} - y_{il})$ 
10:    end for
11:    for all  $l, k \in \{1, \dots, m\}$  do
12:       $\nu_{lk} \leftarrow \nu_{lk} - \zeta C \text{sgn } \nu_{lk}$ 
13:    end for
14:    for all  $i \in \{2, \dots, n\}, l, k \in \{1, \dots, m\}$  do
15:       $\nu_{lk} \leftarrow \nu_{lk} - \zeta (\hat{y}_{i-1l} \hat{y}_{ik} - y_{i-1l} y_{ik})$ 
16:    end for
17:  end for
18: until all parameters have converged
    
```

Fig. 5 Stochastic subgradient descent algorithm for structured support vector machines. sgn is the sign function. $\zeta > 0$ is the predefined learning rate.

Table 1 Summary of datasets.

	# regions	# of nt	# genes	# mapped reads
Training	2	3038447	511	561676
Test	31	21459613	1111	3144765

Table 2 The accuracy of our model compared with AUGUSTUS.

	PPV	SEN	F
Our model ($f_{\text{gene}} + f_{\text{evidence}}$)	15.7	31.1	20.9
Baseline (f_{gene})	14.1	7.1	9.4
AUGUSTUS	55.9	17.8	27.0

through nucleotide-level measurements as defined by Ref. [6]. The nucleotide-level accuracy is assessed by positive predictive value (PPV) and sensitivity (SEN), defined as

$$PPV = \frac{TP}{TP + FP}, \quad SEN = \frac{TP}{TP + FN},$$

where TP is the number of nucleotides at which exon states are correctly predicted (true positives), FP is the number of nucleotides at which exon states are incorrectly predicted (false positives), and FN is the number of nucleotides in the true exon states that were not predicted (false negatives). We also used the F-value as a balanced measure between PPV and SEN, which is defined as their harmonic mean:

$$F = \frac{2 \times PPV \times SEN}{PPV + SEN}.$$

Table 2 shows the accuracy of our model (i.e., $f_{\text{gene}} + f_{\text{evidence}}$) in comparison to our model without the evidence-based scores (i.e., only f_{gene}) as a baseline method, and suggests that significant improvement, especially in sensitivity, can be observed by using the evidence-based scores. We also compared our method with one of the *ab initio* method, AUGUSTUS [22], with default options, which was trained using the same regions as in our setting.

4. Discussion

The *ab initio* approach for predicting gene structures only from genomic sequences has been actively developed [2], [5], [22], and successfully models gene structures using HMMs. Note

that our model without the evidence-based scoring model (only f_{gene}) is regarded as this approach. To improve the prediction accuracy, the *ab initio* approach has been extended by incorporating various source information. In particular, observation of mRNAs using cDNA technology such as ESTs and RNA-seq is known to successfully help predict gene structures. Such an approach has been implemented by numerous methods including AUGUSTUS+ [20], [21], CONTRAST [12], and our model with the evidence-based scoring mode ($f_{\text{gene}} + f_{\text{evidence}}$). As shown in Section 3, experimental results suggest that mRNA information contributes to significant improvement of prediction accuracy in our model. The main difference between our model and previous methods is that our model uses mRNA information not only for the scoring function, but also as explicit constraints for gene structures (Eqs.(5)–(10)); optimal gene structures can be efficiently predicted by maximizing the scoring function under these constraints by a sophisticated technique called dual decomposition. This enables us to take into account long-range consistency between splice donors and acceptors by constraints (6)–(8). In contrast, previous methods such as AUGUSTUS+ and CONTRAST consider mRNA information only for scoring functions, which can be expected to implicitly lead to supported gene structures. However, these methods cannot account for long-range consistency of the spliced sites, mainly due to computational complexity.

Although our model is improved by the use of evidence-based scoring, it is not yet sufficiently accurate compared with existing methods. This is because the preliminary implementation of our model cannot perform the learning algorithm using a large dataset, despite using dual decomposition for efficiency. To realize large-scale training, we need to implement an improved stochastic subgradient descent algorithm such as AdaGrad [9], which would allow faster convergence for the dual decomposition in Fig. 4 and the learning algorithm in Fig. 5. A parallelized subgradient descent algorithm would also enable us to train our model on a large dataset using a supercomputing system. We also need to further optimize hyperparameters including α and β for the weights of transcriptional evidences, C for the ℓ_1 regularization, and δ^{FN} and δ^{FP} to control the trade-off between sensitivity and specificity.

5. Concluding Remarks

We develop a novel gene prediction method for eukaryote genomes that extends the traditional HMM-based gene prediction model by incorporating comprehensive evidence of transcripts using RNA sequencing (RNA-seq) technology. We formulated gene prediction as an integer programming problem, and solved it using the dual decomposition technique. To confirm the proposed algorithm, we conducted computational experiments on benchmark datasets. The results showed that our algorithm with RNA-seq data works efficiently and effectively.

One area for future research is isoform identification and quantification to predict alternative splicing of mRNAs using RNA-seq data. Recently, network flow algorithms have been applied to solve this problem [4], [18], so we plan to develop a novel model that combines network flow algorithms for modeling al-

ternative splicing of mRNAs with HMMs for evaluating protein-coding potential, and to solve the problem by separating it into two sub-problems using the dual decomposition technique.

Acknowledgments This work was supported in part by a Grant-in-Aid for Scientific Research (A) (KAKENHI) (No.23241066) from the Japan Society for the Promotion of Science (JSPS) to Y.S. and K.S., a Grant-in-Aid for Scientific Research on Innovative Areas (KAKENHI) (No.221S0002) from the Ministry of Education, Culture, Sports, Science and Technology to Y.S., and a Grant-in-Aid for Scientific Research (C) (KAKENHI) (No.25330348) from the JSPS to K.S.

References

- [1] Andreotti, S., Klau, G.W. and Reinert, K.: Antilope—a Lagrangian relaxation approach to the de novo peptide sequencing problem, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, Vol.9, No.2, pp.385–394 (2011).
- [2] Asai, K., Itou, K., Ueno, Y. and Yada, T.: Recognition of human genes by stochastic parsing, *Pac. Symp. Biocomput.*, pp.228–239 (1998).
- [3] Bauer, M., Klau, G.W. and Reinert, K.: Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization, *BMC Bioinformatics*, Vol.8, p. 271 (2007).
- [4] Bernard, E., Jacob, L., Mairal, J. and Vert, J.P.: Efficient RNA isoform identification and quantification from RNA-Seq data with network flows, *Bioinformatics*, Vol.30, No.17, pp.2447–2455 (2014).
- [5] Burge, C. and Karlin, S.: Prediction of complete gene structures in human genomic DNA, *J. Mol. Biol.*, Vol.268, No.1, pp.78–94 (1997).
- [6] Burset, M. and Guigo, R.: Evaluation of gene structure prediction programs, *Genomics*, Vol.34, No.3, pp.353–367 (1996).
- [7] Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Rozowsky, J., Roder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Bar, N.S., Batut, P., Bell, K., Bell, I., Chakraborty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E., Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigo, R. and Gingeras, T.R.: Landscape of transcription in human cells, *Nature*, Vol.489, No.7414, pp.101–108 (2012).
- [8] Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R.: STAR: ultrafast universal RNA-seq aligner, *Bioinformatics*, Vol.29, No.1, pp.15–21 (2013).
- [9] Duchi, J., Hazan, E. and Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization, *The Journal of Machine Learning Research*, Vol.12, pp.2121–2159 (2011).
- [10] Duchi, J. and Singer, Y.: Efficient online and batch learning using forward backward splitting, *Journal of Machine Learning Research*, Vol.10, pp.2899–2934 (2009).
- [11] Gross, S.S. and Brent, M.R.: Using multiple alignments to improve gene prediction, *J. Comput. Biol.*, Vol.13, No.2, pp.379–393 (2006).
- [12] Gross, S.S., Do, C.B., Sirota, M. and Batzoglou, S.: CONTRAST: A discriminative, phylogeny-free approach to multiple informant de novo gene prediction, *Genome Biol.*, Vol.8, No.12, p. R269 (2007).
- [13] Guigo, R., Flicek, P., Abril, J.F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V.B., Birney, E., Castelo, R., Eyas, E., Ucla, C., Gingeras, T.R., Harrow, J., Hubbard, T., Lewis, S.E. and Reese, M.G.: EGASP: The human ENCODE Genome Annotation Assessment Project, *Genome Biol.*, Vol.7, Suppl 1, pp.1–31 (2006).
- [14] Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J.M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigo, R. and Hubbard, T.J.: GENCODE:

The reference human genome annotation for The ENCODE Project, *Genome Res.*, Vol.22, No.9, pp.1760–1774 (2012).

- [15] Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L.: TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions, *Genome Biol.*, Vol.14, No.4, p.R36 (2013).
- [16] Korte, B. and Vygen, J.: *Combinatorial Optimization: Theory and Algorithms*, Springer Verlag, Berlin, Germany (2008).
- [17] Kumozaki, S., Sato, K. and Sakakibara, Y.: A machine learning based approach to de novo sequencing of glycans from tandem mass spectrometry spectrum, *IEEE/ACM Trans. Comput. Biol. Bioinform.* (in press).
- [18] Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T. and Salzberg, S.L.: StringTie enables improved reconstruction of a transcriptome from RNA-seq reads, *Nat. Biotechnol.*, Vol.33, No.3, pp.290–295 (2015).
- [19] Sato, K., Kato, Y., Akutsu, T., Asai, K. and Sakakibara, Y.: DAFS: Simultaneous aligning and folding of RNA sequences via dual decomposition, *Bioinformatics*, Vol.28, No.24, pp.3218–3224 (2012).
- [20] Stanke, M., Diekhans, M., Baertsch, R. and Haussler, D.: Using native and syntenically mapped cDNA alignments to improve de novo gene finding, *Bioinformatics*, Vol.24, No.5, pp.637–644 (2008).
- [21] Stanke, M., Schoffmann, O., Morgenstern, B. and Waack, S.: Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources, *BMC Bioinformatics*, Vol.7, p.62 (2006).
- [22] Stanke, M. and Waack, S.: Gene prediction with a hidden Markov model and a new intron submodel, *Bioinformatics*, Vol.19, Suppl 2, pp.i215–225 (2003).
- [23] Tsochantaridis, I., Joachims, T., Hofmann, T. and Altun, Y.: Large Margin Methods for Structured and Interdependent Output Variables, *J. Mach. Learn. Res.*, Vol.6, pp.1453–1484 (2005).

Appendix

A.1 Gene structure model

In this section, we show the model of gene structures corresponding to hidden states of hidden Markov models, defined as

$$G = G_{\text{exon}} \cup G_{\text{intron}} \cup G_{\text{intergenic}}$$

$$G_{\text{exon}} = G_{\text{CDS}} \cup G_{\text{UTR}}$$

$$G_{\text{CDS}} = \{\text{SingleCDS}_0, \text{SingleCDS}_1, \text{SingleCDS}_2, \\ \text{StartCodon}_0, \text{StartCodon}_1, \text{StartCodon}_2, \\ \text{StopCodon}_0, \text{StopCodon}_1, \text{StopCodon}_2, \\ \text{InitialCDS}_0, \text{InitialCDS}_1, \text{InitialCDS}_2, \\ \text{InternalCDS}_0, \text{InternalCDS}_1, \text{InternalCDS}_2, \\ \text{TerminalCDS}_0, \text{TerminalCDS}_1, \text{TerminalCDS}_2\}$$

$$G_{\text{UTR}} = \{5\text{UTR}, 3\text{UTR}\}$$

$$G_{\text{intron}} = \{\text{Intron}_0, \text{Intron}_1, \text{Intron}_2, \\ \text{SpliceDonor}_0, \text{SpliceDonor}_1, \text{SpliceDonor}_2, \\ \text{SpliceAcceptor}_0, \text{SpliceAcceptor}_1, \text{SpliceAcceptor}_2, \\ 5\text{UTRSpliceDonor}, 5\text{UTRSpliceAcceptor}, \\ 3\text{UTRSpliceDonor}, 3\text{UTRSpliceAcceptor}\}$$

$$G_{\text{intergenic}} = \{\text{Intergenic}\}.$$

To represent the codon frame at the states for coding sequences (CDS), the state transition from a CDS state to another CDS state follows a cyclic order of 0, 1, and 2 for their subscripts that represents the codon frame number (e.g. $\text{InternalCDS}_0 \rightarrow \text{InternalCDS}_1 \rightarrow \text{InternalCDS}_2 \rightarrow \text{InternalCDS}_0 \rightarrow \dots$). The states for introns also have subscripts for the codon frame number to remember the codon frame from which the next exon restarts.

Figure A-1 shows the state transitions between exons, introns,

Fig. A-1 State transitions of our gene structure model. States with black boxes can have self-transitions of variable length. States with red boxes can have self-transitions of fixed length. Directed edges colored red and green respectively indicate state transitions incoming to and outgoing from exons.

and intergenic regions of our gene structure model.

Tatsumu Inatsuki received his B.S. and M.S. in Bioinformatics in 2013 and 2015 from Keio University, Japan. He currently works for a software company in Japan.

Kengo Sato is an Assistant Professor at the Department of Biosciences and Informatics at Keio University, Japan. He received his Ph.D. in Computer Science from Keio University, Japan, in 2003. He was awarded the Oxford Journals JSBi Prize in 2008, and IPSJ Yamashita SIG Research Award in 2012. His research interests include bioinformatics, computational linguistics, and machine learning.

Yasubumi Sakakibara is a Full Professor at the Department of Biosciences and Informatics at Keio University, Japan. He received his Doctor of Science degree from Tokyo Institute of Technology in 1991. He spent one year as a postdoctoral researcher at the University of California, Santa Cruz and collaborated with Professor David Haussler in a project on stochastic context-free grammars for modeling RNAs. He also worked for Fujitsu Laboratory. His research interests include bioinformatics, comparative genomics, DNA computers, and computer science.

(Communicated by Susumu Goto)