塩基列の長さと重複率に基づくSVMを用いたエクソン領域抽出精度の向上法

冨成 裕輔[1)] 山森 一人[2)] 吉原 郁夫[3)] 藤田 祐二[4)]

Accurate Identification of Exon-Intron Boundaries by SVM based on Three-tuple Encoding and Base-repetition Ratio

Yusuke TOMINARI Kunihito YAMAMORI Ikuo YOSHIHARA Yuji FUJITA

ABSTRACT

In this report, we evaluate the base-encoding methods to identify Exon-Intron boundaries by Support Vector Machine. In conventional experiments, 2-bits for 4-bits binary encoding was employed to describe bases. We focuced on that only the euclidean distance is used in Support Vector Machine, then propose three-tuple encoding which has the same euclidean distance among four encoded bases with small dimensions. In addition, we also show a new method how to decide the shortest sequence of bases to identify boundaries. Experiments showed that our method could achieve good identification ratio.

Key Words:
Support Vector Machine, Genome, Exon-Intron, Encoding

1 はじめに

生物では，タンパク質が生体機能の発現や調整に重要な役割を担っている．したがって，DNA 塩基配列中からタンパク質合成の情報を保持しているエクソン領域のみを高精度で抽出することは，遺伝子の持つ機能を解明する上で非常に重要な課題である．

塩基配列の解析には，いままで多くの手法が提案されている．例えば，上舞ら[1)]は複数のニューラルネットワーク（NN）を組み合わせることで，単体のNNよりも境界識別精度を向上させるマルチモーダルニューラルネットワークを用いた境界識別法を提案している．太田ら[2)]や藤田ら[3)]は遺伝的プログラミングを，Yuanら[4)]はサポート・ベクタ・マシン (SVM)[5)] を用いた境界識別手法を提案している．これらの中でも，SVMは未知データに対する汎化能力の高さから極めて有望な手段の1つとされている．

本報告では，SVMによりエクソン-イントロン境界を高精度に識別することを目的とする．そのため，各塩基の数値コード化に，コード長が短く，各塩基コード間のユークリッド距離が均一となる3変数表現を採用する．また，切り出す塩基配列長を決定するため，境界から上流側，下流側それぞれについて切り出す長さを変化させ，切り出された各塩基配列が十分に異なる最短の塩基配列長を用いる手法を提案し，遺伝子データベースから取り出したヒトゲノム情報を用いて提案手法の評価を行った．

[1)]工学研究科情報工学専攻，現在セイコーエプソン（株）
[2)]情報システム工学科助教授
[3)]情報システム工学科教授
[4)]工学研究科情報工学専攻，現在三菱スペースソフトウェア（株）

2 サポート・ベクタ・マシン

2.1 SVMによる線形分離可能なサンプルの分類

n 個のサンプル集合 $\boldsymbol{x}_i(i=1,\cdots,n)$ が線形分離可能な場合，SVMでは（1）式に基づいて各サンプルを2つのクラス C_1，C_2 に分類する．

$$f(x) = sign(g(\boldsymbol{x})) = sign(\boldsymbol{w}^t\boldsymbol{x} + b). \qquad (1)$$

ここで $\boldsymbol{w}$，b はそれぞれ識別関数を決定するパラメータである．

サンプル $\boldsymbol{x}_i(i=1,\cdots,n)$ を（1）式に入力することにより1，-1の出力が得られ，$\boldsymbol{x}_i$ は（2）式，(3) 式に基づいてクラス C_1，C_2 に分けられる．ここで $y_i(i=1,\cdots,n)$

は x_i の正解クラスラベルである．

$$\forall i, g(x_i) = w^t x_i + b \begin{cases} \geq & 1, \quad x_i \in C_1, \\ \leq & -1, \quad x_i \in C_2, \end{cases} \tag{2}$$

$$y_i = \begin{cases} 1, & x_i \in C_1, \\ -1, & x_i \in C_2. \end{cases} \tag{3}$$

(2) 式, (3) 式より, (4) 式が得られる．

$$\forall i, \quad y_i \cdot (w^t x_i - b) - 1 \geq 0. \tag{4}$$

(4) 式は，図1で示したように，K 次元平面 $H1: (w^t x - b) = 1$ と $H2: (w^t x - b) = -1$ の間にはサンプルが存在しないことを示し，全てのサンプルは完全に分離されることを意味している．このとき，K 次元平面 $H1, H2$ 間の距離はマージンと呼ばれ，その距離は $\frac{2}{\|w\|}$ となる．マージンを最大にするためには，$\|w\|$ を最小化すればよい．問題を簡単にするため，$\frac{1}{2}\|w\|^2$ を最小化することとし，これを目的関数とする．SVM では，一般にはラグランジュ未定乗数法を使い，最小化問題を双対問題に帰着させる．

ラグランジェ未定乗数 $\lambda(\geq 0)$ を導入し, (4) 式の制約条件の下で目的関数を (5) 式に変形する．

$$L(w, b, \lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \lambda_i y_i (w^t x_i - b) - 1. \tag{5}$$

$L(w, b, \lambda) = 0$ でパラメータ w および b に関する偏微分を行い, (6) 式と (7) 式を得る．

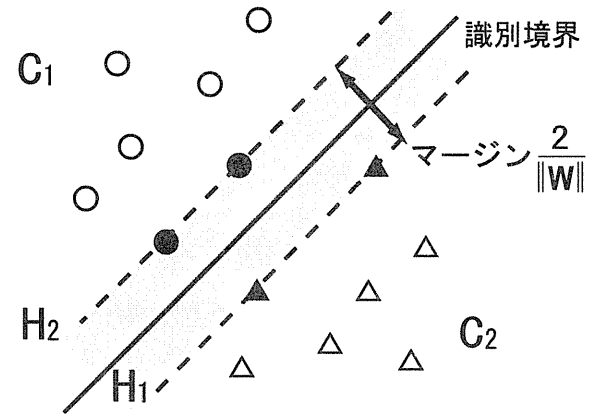
$$w - \sum_{i=1}^{n} \lambda_i y_i x_i = 0, \tag{6}$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0. \tag{7}$$

(6) 式, (7) 式を目的関数である (5) 式に代入すると，制約条件 $\sum_{i=1}^{n} \lambda_i y_i = 0 \quad (0 \leq \lambda_i,\ i = 1, \cdots, N)$ の下で，

$$L(\lambda) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j x_i^t x_j, \tag{8}$$

を最大化する問題が得られる．(8) 式を最大化する λ_i を λ_i^* とする．このとき，λ_i^* が 0 でない，すなわち $\lambda_i^* > 0$ となっているサンプル x_i はサポートベクタ (SV) と呼ばれ，図1のように2つの K 次元平面 $H1$, $H2$ のどちらかの上に存在している．通常，SV はもとのサンプル数に比べると極めて少ない．つまり，多くのサンプルから少数の SV を見つけ出すことで, (1) 式のパラメータ w, b を決定することができる．実際，



○:クラス C_1 の学習パターン　● ▲:サポートベクタ
△:クラス C_2 の学習パターン　　:マージン領域

図. 1　線形識別が可能な場合

双対問題の最適解 λ_i^*，および (6) 式, (7) 式から，パラメータ w の最適値 w^* は，

$$w^* = \sum_{i=1}^{n} \lambda_i^* y_i x_i, \tag{9}$$

となる．また，パラメータ b の最適値 b^* は SV が K 次元平面 $H1$, $H2$ の上に存在しているという関係を利用して (10) 式により求められる．

$$b^* = w^{*t} x_s - y_s. \tag{10}$$

ここで，S は SV の集合であり，x_s は任意の SV である．最終的に線形 SVM による識別関数は (11) 式のように表現される．

$$\begin{aligned} f(x) &= sign(w^{*t} x - b*), \\ &= sign(\sum_{i \in S} \lambda_i^* y_i x_i^t x - b^*). \end{aligned} \tag{11}$$

すなわち，SVM ではマージンの最大化という基準から，$\lambda_i^* = 0$ となる識別境界付近の少数の SV が自動的に選ばれて識別関数が構成される．

2.2 SVM による線形分離不可能なパターンの分類

サンプル集合が線形分離不可能な場合, (4) 式を満たす w は存在しない．そこで，制約を緩めて学習パターンの誤識別をある程度許すようにするソフトマージンと呼ばれる手法を用いる．ソフトマージンを用いた時の，識別境界と SV の様子を図2に示す．具体的には，パラメータ ξ を用いて制約条件を (12) 式で示したように緩和する．

$$\forall i, g(x_i) = w^t x_i - b \begin{cases} \geq & 1 - \xi_i \quad x_i \in C_1, \\ \leq & -1 + \xi_i \quad x_i \in C_2. \end{cases} \tag{12}$$

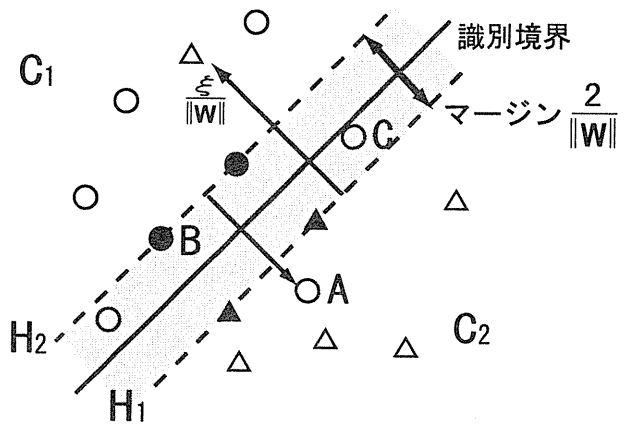


図. 2　線形識別が不可能な場合

このとき，目的関数は（13）式のように定められる．

$$f(\boldsymbol{w},\boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + c\sum_{i=1}^{n}\xi_i. \tag{13}$$

（13）式の第一項は2.2節で述べたようにマージンを最大化するための項である．第二項はマージン領域をはみだして存在しているサンプルに対するペナルティ項である．係数 c は第一項と第二項のバランスを定める定数である．つまり，c が小さい程，マージンはより大きくなる．最適な c の値は経験的に決定される．（13）式の $\boldsymbol{w}$ を最小化するには，基本的には節で述べた順と同様となるので結果のみを示す．ラグランジュ未定乗数 λ，γ を導入して，

$$\begin{aligned} L(\boldsymbol{w},b,\boldsymbol{\lambda},\boldsymbol{\gamma}) \;=\; & \frac{1}{2}\|\boldsymbol{w}\|^2 - c\sum_{i=1}^{n}\gamma_i \\ & -\sum_{i=1}^{n}\lambda_i y_i(\boldsymbol{w}^t\boldsymbol{x}_i - b) - (1-\xi_i) \\ & -\sum_{i=1}^{n}\gamma_i\xi_i. \end{aligned} \tag{14}$$

$L(\boldsymbol{w},b,\boldsymbol{\lambda},\boldsymbol{\gamma}) = 0$ として（14）式を w，b，γ_i で各々偏微分すると，

$$w - \sum_{i=1}^{n}\lambda_i y_i \boldsymbol{x}_i = 0, \tag{15}$$

$$\sum_{i=1}^{n}\lambda_i y_i = 0, \tag{16}$$

$$\lambda_i - c + \gamma_i = 0, \tag{17}$$

という関係が成り立つ．これらを目的関数である（14）式に代入すると，制約条件 $\sum_{i=1}^{n}\lambda_i y_i = 0\,(0 \le \lambda_i \le \gamma, i = 1,\cdots,N)$ の下で，

$$L(\boldsymbol{\lambda}) = \sum_{i=1}^{n}\lambda_i - \frac{1}{2}\sum_{i,j=1}^{n}\lambda_i\lambda_j y_i y_j \boldsymbol{x}_i^t\boldsymbol{x}_j, \tag{18}$$

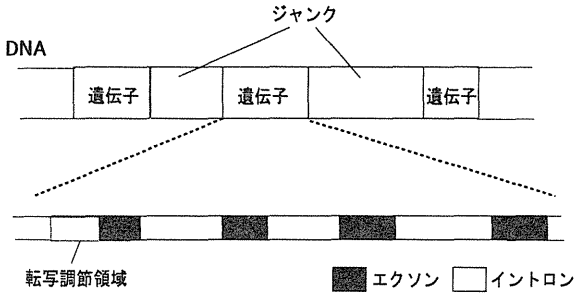


図. 3　遺伝子内部の模式図

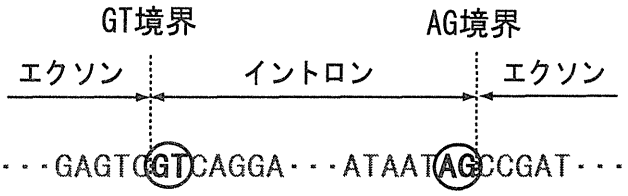


図. 4　イントロンでの GT - AG 則

となり，（18）式を最大とする双対問題が得られる．このとき，最適解 λ_i^* の値により，線形分離可能な時と同様にサンプルを分類することができる．$\lambda_i^* = 0$ ならば，図2中のサンプルAのように，マージン領域の外側，すなわち $H1$ または $H2$ の外側に存在している．$0 < \lambda_i^* < c$ の時，対応するサンプルは図2のサンプルBのように，$H1$, $H2$ のどちらかの上に存在している．$\lambda_i^* = c$ の時，図2のサンプルCのように $H1$，$H2$ より内側に存在している．

3　エクソン - イントロン境界識別

3.1　エクソン - イントロン境界

DNA 塩基配列はアデニン（A: Adenine），グアニン（G: Guanine），シトシン（C: Cytosine），チミン（T: Thymine）の4塩基から構成されており，その総量は塩基数にして約30億あるとされている．遺伝子は機能によりエクソン，イントロン，転写調節領域の3つに分けられ，図3のように転写調節領域の後にエクソンとイントロンが交互に出現する構造となっている．遺伝子の中で実際にタンパク質に翻訳される領域がエクソンであり，イントロンは転写が行われる際に取り除かれ，タンパク質に翻訳されない領域である．転写調節領域は転写の活性調節を行なう領域である．

遺伝子中のエクソンとイントロンの境界では，イントロン先端の塩基が GT，末端の塩基が AG となっており，これらを識別することによりエクソンを抽出できる．しかし，GT および AG という塩基配列は，エクソンとイントロンの内部（非境界）にも多数存在する．このことから，塩基配列のデータは着目する GT，AG が実際に境界である境界データと，それ以外の非境界データに分けられる．本報告では，塩基配列中で

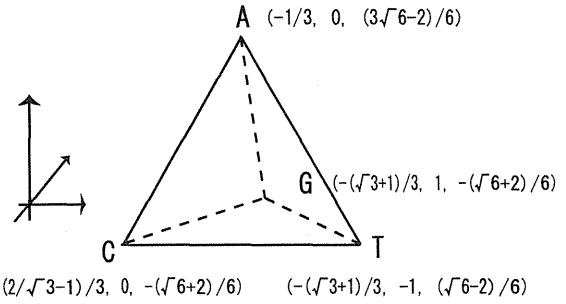


図. 5 正立方体の頂点に対応する塩基

着目する GT，AG の前後の塩基配列から，その GT，AG が境界か非境界かを SVM によって判定する．

3.2 塩基の数値コード化

SVM は，サンプルの識別をサンプル間のユークリッド距離に基づいて行うため，数値コード化された塩基間のユークリッド距離が均一で，コード長が短い方がよい[4]．

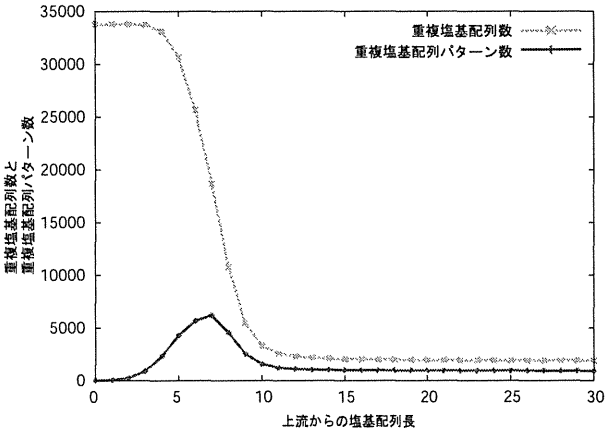
従来研究では，塩基の数値コード化に4変数表現，もしくは2変数表現が用いられてきた．2変数表現の場合は，各塩基に $A=(00)$，$C=(01)$，$G=(10)$，$T=(11)$ のようにコードを割り当てる．そのため，コード長が短くなるが，各塩基コード間のユークリッド距離は均一でないという問題がある．4変数表現の場合では，各塩基に $A=(0001)$，$C=(0010)$，$G=(0100)$，$T=(1000)$ のようにコードを割り当てる．このとき，4次元座標系において数値コード化された各塩基間のユークリッド距離は等しいが，コード長が長くなり2変数表現時の2倍必要となる．

本報告では，塩基のコード長を3に減らすと同時に，数値コード化された各塩基コード間のユークリッド距離が均一となる3変数表現を採用する．具体的には，図5に示すように各頂点の距離が均一である正四面体の頂点座標を各塩基に当てはめる．

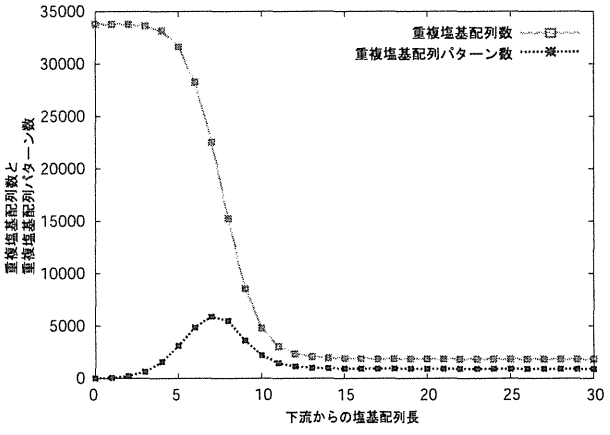
3.3 識別に適した塩基配列の抽出

SVM で境界判定を行う前に，判定するのに必要十分な塩基配列長を検討する必要がある．第1章でも述べたように，識別に用いる塩基配列は境界近傍で現れる特徴を含み，かつできるだけ短いことが望ましい．

そこで，境界である塩基配列の集合から塩基配列パターンが同一であるものを抽出する．抽出された塩基配列パターンに一致する塩基配列数を重複数と定義し，重複数の和を「重複塩基配列数」と定義する．なお，ユニークな塩基配列の重複数は0とした．さらに，重複数が0より大きい（境界である塩基配列集合内に2つ以上出現する塩基配列）塩基配列パターン数の和を



(a) GT 塩基配列上流



(b) GT 塩基配列下流

図. 6 GT 塩基配列の重複数と重複塩基配列パターン数

「重複塩基配列パターン数」と定義する．本報告では，境界から上流，下流それぞれで切り出す長さを変化させ，切り出された塩基配列の重複塩基配列数と重複塩基配列パターン数を調べ，これらが共に小さくなる長さで切り出すこととした．

詳細な切り出し範囲を決定するため，塩基配列を上流と下流に分けて重複塩基配列数と同パターン数について調査した．図6から，GT 塩基配列上流，下流の重複塩基配列数と同パターン数は上流15，下流15の位置で，ほとんど変化が見られなくなることが分かる．

AG 塩基配列についても同様に調査を行い，上流18，下流18から重複塩基配列数と同パターン数の変化が極めて小さくなることが分かった．

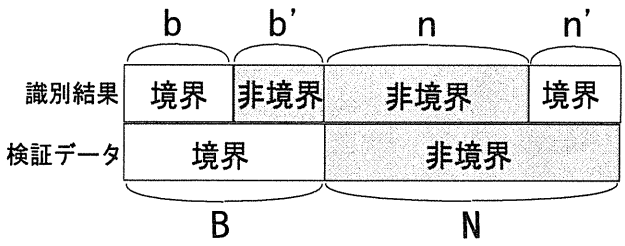


図. 7 識別尺度の基準

4 境界識別実験

4.1 実験条件

4.1.1 塩基データ

実験で用いた塩基配列はNCBI[6]のヒトゲノムデータベースから入手した．なお，実験で用いたSVMにはlibSVM[7]を採用した．入手したデータ総数は，GT塩基配列では境界データが33,792，非境界データが10,513,172，AG塩基では境界データが22,633，非境界データが8,481,368である．実験に使用するGT塩基配列数およびAG塩基配列数は，収集したデータから以下の数だけランダムに選びデータセットとした．

- GT塩基データ
  - 学習用境界データ： 15000
  - 学習用非境界データ： 15000
  - 検証用境界データ： 5000
  - 検証用非境界データ： 5000
- AG塩基データ
  - 学習用境界データ： 15000
  - 学習用非境界データ： 15000
  - 検証用境界データ： 5000
  - 検証用非境界データ： 5000

こうしたデータセットを10組作成し，これらでのSn, Sp, Sn', Sp'の平均を総合的な評価対象とした．

4.1.2 境界識別精度の評価尺度

エクソン-イントロン境界の識別評価の尺度として，$Sensitivity(Sn)$と$Specificity(Sp)$の2つが一般に用いられる．図7に示したパラメータを用いると，Sn, Spはそれぞれ（19）式,（20）式で定義される．

$$S_n = \frac{b}{B} \times 100(\%), \tag{19}$$

$$S_p = \frac{b}{b+n'} \times 100(\%). \tag{20}$$

S_nは実際の境界データを正しく識別した割合であり，S_pは境界データと識別されたデータがどの程度正確であるかを測る尺度である．S_nが小さい場合，境界データの見落としが多くなっていることを示している．また，S_pが小さい場合，非境界データを境界データであると誤識別していることが多いことを示している．

実験で使用する塩基配列において，境界データと非境界データの数の割合は4.3節に示したようにおよそ1:250である．境界データの稀少性から，境界データの取りこぼしは致命的であり，S_nの精度は重要となる．さらに，非境界データに対する識別精度を評価するため，境界のS_n, S_pに相当するSn', Sp'を以下のように定義した．

$$S'_n = \frac{n}{N} \times 100(\%) \tag{21}$$

$$S'_p = \frac{n}{n+b'} \times 100(\%) \tag{22}$$

本報告では，これら4つの識別尺度を用いて境界識別の比較・検証を行う．

4.2 数値コード化手法と識別精度

ヒトゲノムのGT塩基配列，AG塩基配列ともに，2変数表現，3変数表現，4変数表現で数値コード化を行い，境界識別実験を行った．この実験では，塩基の数値コード化別による識別尺度の違いを比較・検証することが目的であるので，実験データの塩基配列の切り出しは従来と同じくGT境界では上流10，下流10，AG境界では上流30，下流7とした．

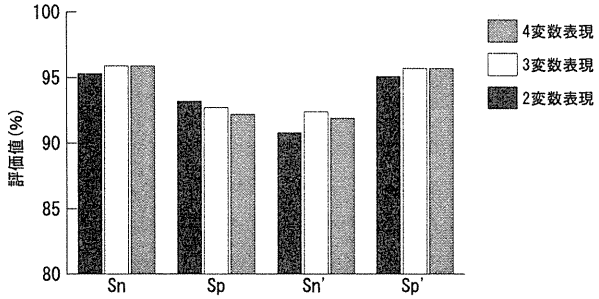
図8から，GT塩基配列では3変数表現や4変数表現での評価値はほぼ同等であることが分かる．一方，2変数表現でのSn, Sn', Sp'は，3変数表現や4変数表現より約0.5(%)低下するが，Spについては最も高い値を示した．AG塩基配列では3変数表現が全ての評価尺度の値において最も高いことが分かる．

4.3 塩基配列長と識別精度

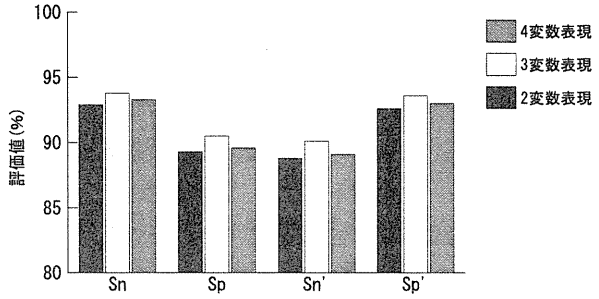
3.3節で示した切り出し塩基配列長の妥当性を評価するため，上流10下流10および上流20下流20で切り出した場合との比較を行った．塩基の数値コード化には4.2節で最も識別尺度が高かった3変数表現を用いた．

実験に使用するGT塩基配列およびAG塩基配列は節で作成したデータセットで，塩基配列長のみを変化させたものである．

図9(a)から，GT塩基配列では各評価尺度の評価値は切り出した塩基配列長によらずほぼ同等であり，切

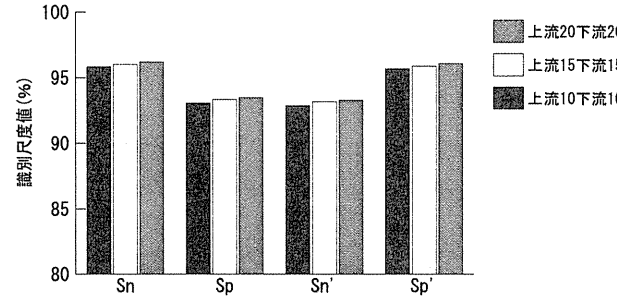


(a) GT 塩基に対する各塩基表現での評価値

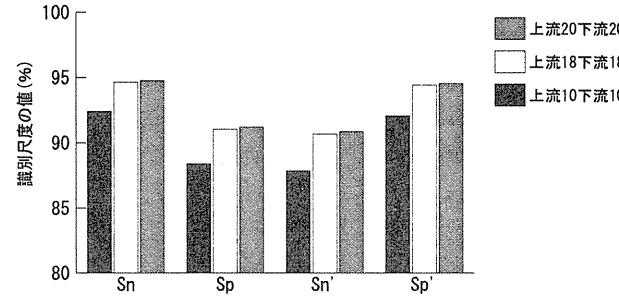


(b) AG 塩基に対する各塩基表現での評価値

図. 8 各塩基表現での評価値



(a) GT 塩基配列での塩基配列長別による評価値



(b) AG 塩基配列での塩基配列長別による評価値

図. 9 各境界における塩基配列の塩基配列長別による評価値

り出し長は上流・下流ともに10塩基で十分であることが分かる．一方，AG塩基配列では上流10下流10での評価尺度は，他の切り出し長での結果と比べて約3(%)ほど低く，上流・下流18塩基以上ではほとんど変化していない．これらより，3.3節で求めた境界識別のための塩基配列長が妥当であることが示された．

5 おわりに

本報告では，DNA塩基配列中からエクソン領域を抽出するため，エクソンとイントロンの境界近傍に含まれる情報からSVMによって境界識別判定を行った．まず，塩基数値コード化では2変数表現，3変数表現，4変数表現について定量的な比較を行った．識別実験の結果，3変数表現では2変数表現と比べて評価値は向上し，4変数表現と同等の高い評価値を得ることが可能であることを示した．

つぎに，識別に最適な塩基配列の長さを求めるため，重複塩基配列数と重複塩基配列数パターン数を用いた手法を提案し，SVMでの評価を通じて提案手法の妥当性を示した．

今後の課題として，非境界データの重複塩基配列数とその重複塩基配列パターン数を考慮した手法について検討することが挙げられる．

参考文献

[1] Y. KAMIMAI, I. YOSHIHARA, K. YAMAMORI and M. YASUNAGA, "Methods to Extract Exon Regions from DNA sequence -Development of Multi-modal Neural Network-", Fuzzy, Artificial Intelligence, Neural Networks and Computational Intelligence (FAN'02), pp. 385–390 (2002).

[2] T. OHTA, I. YOSHIHARA, K. YAMAMORI and M. YASUNAGA, "GP-based Method for Identifying Exon Region in DNA Sequences", Proc. 4th Asia-Pacific Conference on Simulated Evolution And Learning 2002 (SEAL'02) (2002), cr1333(CD-ROM).

[3] Y. FUJITA, K. YAMAMORI, I. YOSHIHARA and M. AIKAWA, "Performance of Parallel Island Model with Dividing Population and Training-Set", Genome Informatics 2001, Universal Academy Press, Inc., pp. 420–422 (2001).

[4] Y. X, Buckles.B.P. and Zhang.J, "Comparison Study Descision Tree and SVM to Classify Gene Sequence", Electrical Engineering and Computer Science Department, Tulane University (2003).

[5] 栗田多喜夫, "サポートベクターマシン入門", 産業技術総合研究所 脳神経情報研究部門 (2003).

[6] http://www.ncbi.nlm.nih.gov/.

[7] http://www.csie.ntu.edu.tw/ cjlin/libsvm/.