6P−2

ブースティング法を応用したcDNA配列のコーディング領域予測*

清水 佳奈†
早稲田大学理工学研究科‡

足立淳§
理研GSC¶

村岡 洋一‖
早稲田大学理工学研究科‡

1 はじめに

ヒトやマウスなどのゲノムには何万もの遺伝子が存在し、それらはDNA配列上に遺伝子コーディング領域として点在している。遺伝子がタンパク質として発現するためには、まず、遺伝子領域を含むDNAの配列がmRNAとしてコピーされ、そのmRNAの配列に従ってアミノ酸が合成されてタンパク質が作られる。本研究が対象とするcDNAとは、mRNAを捕まえて人工的にmRNAのコピーを作ったDNA配列であり、その中には遺伝子領域が含まれていることが期待できる。よって、そのコーディング領域を予測するシステムは、タンパク質の機能分類・解析にとって重要な役割を果たす。

しかしながら、cDNA塩基配列を決定する精度が100%ではないために約1000塩基に対し1個の割合でフレームシフトまたは文字の置き換わりが起こり、コーディング領域の予測が困難である。また、確率モデルを用いた従来手法では単一規則による偏った予測になりがちであり、その推定精度に限界がある。

本稿ではブースティング法を用いて異なる複数の予測手法を結果に反映させることで、高い精度を得る手法を提案する。

2 提案する手法

AdaBoost.M2 [1] はブースティング法の一つであり、数多くの精度の低いルールを組み合わせて非常に精度の高い予測ルールを得るための汎用的かつ理論的な性能保証のある方式である。本節ではこれをcDNA配列のコーディング領域予測に応用する手法について述べる。

2.1 AdaBoost.M2を応用した手法

アルゴリズムを図1に示す。

- 従来の予測方法及び各生物学的規則によるN個の予測方法を弱仮説 $h_n(n = 1, ..., N)$ とする。
- 事例 (x_i, y_i) における x_i はcDNA配列、y_i を正解値、$h_t(x_i)$ は弱仮説がcDNA配列から予測する値とする。
- 予測値となりえる値の数は個々の配列によって異なるので、一律とするためにラベル集合Yは N個の弱仮説が返す予測値の集合とする。
- rはYに含まれる正解値の数とする。
- 弱学習アルゴリズムは h_n の中から、0以上N以下の任意数の弱仮説を選び $\tilde{h}_t$ の構成要素とする。(つまり $\tilde{h}_t$ は任意数の h_n がそれぞれかえす複数の予測値をまとまりとして持つことが出来る。)
- ラウンド t で得られた α_t を $\tilde{h}_t$ を構成している h_n の重み w_n に足していく。
- 最終仮説 $H_{final}(x)$ の予測結果は h_n の w_n による重み付多数決で得る。

2.2 弱仮説に使用したツール及び生物学的規則

以下に、弱仮説として使用したツール及び生物学的規則を示す。なお、kozak consensusについてはATGの前後で別々の規則とした。また、組成率については最もスコアの高いもの、2番目のもの、開始もしくは終止付近のものを別々の規則とした。

- DECODER
- kozak consensus
- コドンの組成率
- アミノ酸の組成率
- アミノ酸ペアの組成率

3 本システムの評価

本研究が比較対象とするのは福西らが開発したDECORDER [2] とGENSCAN [3] する。以下に述べるようなデータのもとで比較実験を行った。

*The application of boosting to identification of genes in full-length cDNA sequences with frame-shift error
†Kana Shimizu
‡Graduate School of Science and Engineering, Waseda University
§Jun Adachi
¶Genomic Science Center RIKEN (The Institute Of Physical and Chemical Research)
‖Yoichi Muraoka

Given: $(x_1, y_1), ..., (x_m, y_m)$ where $x_i \in \mathrm{X}$, $y_i \in Y = \{h_1(x_i), ..., h_N(x_i)\}$
Initialize $\tilde{D_1}(i,l) = \|l \neq y_i\|/(m(N-r))$
For $t = 1, ..., T$:
Train weak learner using pseudoloss defined by $\tilde{D_t}$.
Get $\tilde{h_t}$ from h_n
Let

$$\tilde{\varepsilon_t} = \frac{1}{2}\sum_{i=1}^{m}\sum_{l\in Y}\tilde{D_t}(i,l)\cdot(\|y_i \notin \tilde{h_t}(x_i)\|) + \|l \in \tilde{h_t}(x_i)\|).$$

Let $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\tilde{\varepsilon_t}}{\tilde{\varepsilon_t}}\right)$.
$w_n = w_n + \alpha_t$.
Update

$$\tilde{D_{t+1}}(i,l) = \frac{\tilde{D_t}\cdot\exp(\alpha_t(\|y_i \notin \tilde{h_t}(x_i)\| + \|l \in \tilde{h_t}(x_i)\|))}{Z_t}.$$

Out put final hypothesys:

$$H_{final}(x) = \sum_{n=1}^{N} w_n h_n(x).$$

$\diamond$ $\|\pi\| = 1$ if π is true,else $\|\pi\| = 0$.

図 1: AdaBoost.M2 を cDNA 配列のコーディング領域予測に応用したアルゴリズム.

3.1 使用データ

- 正解値が既知の理研マウスフルレングス cDNA 配列 1436 個。
- 上記のデータに対し、0〜2 の任意数のフレームシフトを人工的に起こした。(テストデータにおいてフレームシフトを含む配列数は 287 個)
- 半数を訓練データとしてブースティングに利用、その他をテストデータとして利用した。

3.2 各規則につけられた重みについて

ブースティングを行った結果、各規則につけられた重みについて、上位 6 つを表 1 に示す。

生物学的規則またはツール	重み
ATG より前の kozak consensus	1.051764
DECODER	0.751067
アミノ酸の組成率	0.427259
ATG より後の kozak consensus(A)	0.228820
ATG より後の kozak consensus(G)	0.214585
ATG 直後のアミノ酸ペアの組成率	0.182814

表 1: 各規則に付けられた重み

3.3 実験結果

表 2 に示す。表中の $\tilde{h_t} \leq n$ は $\tilde{h_t}$ を構成する弱仮説が n 個以下であることを示す。GENSCAN の正解率については予測が Syngle Exon でない場合は、Inital Exon の開始と Terminal Exon の終止を結果とした。なお、信頼度は (正解に含まれている領域/予測領域) × (予測領域に含まれている領域/正解領域) で定義した。

	正解率 [%]	信頼度 [%]
DECODER	66.43	91.35
GENSCAN	61.42	92.09
AdaBoost.M2($\tilde{h_t} \leq 2$)	71.17	93.49
AdaBoost.M2($\tilde{h_t} \leq 3$)	71.87	93.73

表 2: 実験結果

3.4 考察

実験結果より、最大で約 10.4%の精度向上が確かめられた。採用した生物学的規則にもとづく予測の半数以上が分布の難しい事例に対して 50%以上の精度を保てなかったことから、難しい事例に対して高い精度を保てるような生物学的規則が見つかりさえすれば、より高い精度の予測が期待できることがわかった。また、あらかじめ与えておく生物学的規則の組み合わせ次第で予測結果の変動がみられたので、適切に選択をする弱学習アルゴリズムを与えることでより高い精度の予測を行うことができる。

4 まとめ

本稿ではフレームシフトを含む cDNA 配列のコーディング領域予測に、AdaBoost.M2 を応用する手法について述べた。また、従来手法との比較実験を行った結果、精度の向上が確かめられた。難しい事例に対して高い精度を保てるような生物学的規則の発見と、ブースティングの各ラウンドごとに得られるパラメータから複数の生物学的規則を選択するような学習アルゴリズムを適用させることが今後の課題となる。

参考文献

[1] Yoav Freud,Robert Schapire *A decision-theoretic generalization of on-line learning and an application to boosting*,1997.

[2] Yoshifumi Fukunishi,Yoshihide Hayashizaki *Amino-acid translation program for full-length cDNA sequences with frame-shift error*, 2001 in press.

[3] http://genes.mit.edu/GENSCAN