Vol.2011-MPS-82 No.11
2011/3/7

文献情報を用いたカーネル法による遺伝子機能アノテーション

マチュー・ブロンデル [†1]　関　和広[†1]　上原邦昭[†1]

本稿では，文献の情報を基に遺伝子機能を付与する問題に対して，カーネルを用いた機械学習の手法を提案する．生物医学文献の数は膨大であり，また遺伝子の種類も数多くある．そのため，各文献に記された遺伝子に対して遺伝子機能を手作業で付与（ラベル付け）するには，多大なコストが必要となる．その結果，機械学習を行うために必要な訓練データが，十分に集まらないことが多い．訓練データを必要としない手法として，従来，文字列一致が利用されてきたものの，この手法では，表記の揺れや未知語に対処できないという問題がある．本稿では，付加的な情報を容易かつ効果的に取り込むことができ，計算量的にも優れた性質を持つカーネルを用いることで，これらの問題に対処する．また，マルチラベル分類による遺伝子機能付与を行う際に，各クラスごとに正則化を行うことで，ラベル付きデータの数が特定のクラスに偏っているデータ（不均衡データ）の問題にも対処する．TREC ゲノムトラックのデータを用いた評価実験により，従来手法に対する提案手法の優位性を示す．

Literature-based gene function annotation with kernels

Mathieu Blondel ,[†1] Kazuhiro Seki [†1]
and Kuniaki Uehara[†1]

In recent years, a number of machine learning approaches for literature-based gene function annotation have been proposed. However, due to issues such as lack of labeled data, class imbalance and computational cost, they have usually been unable to surpass simpler approaches based on string-matching. In this paper, we propose a principled machine learning approach focusing on kernel classifiers. We show that kernels are computationally efficient and can address the task's inherent data scarcity by embedding additional knowledge. We also propose a simple and effective solution to deal with class imbalance. From experiments on the TREC Genomics Track data, our approach achieves better F_1-score than two existing approaches based on string-matching and cross-species information.

1. はじめに

ヒトゲノムプロジェクトの完了とともに，老化・病気といった人間の身体的機能への理解を目的として，遺伝子の役割を解明する研究が盛んに行われている．これらの研究によって生産される論文の数は膨大であり，その結果，研究者が特定の遺伝子に関する情報を直接論文から網羅的に収集することは，ますます困難になっている．

この問題を解決するため，多くの組織によって，Gene Ontology（GO）と呼ばれるオントロジーを用いた遺伝子機能情報の付与（GO アノテーション）が行われている．これにより，モデル生物の多くの遺伝子に関して，生物医学文献の内容に基づく遺伝子機能を容易に把握することが可能になった．現在 GO には，約 30,000 の GO タームが登録されており，これらの GO タームは，分子機能（MF）・細胞成分（CC）・生体内作用（BP）の三大分類のもと，有向非巡回グラフ（DAG）の構造で定義されている．一方，人的資源の限界や増加し続ける文献が原因となり，現在の手作業によるアノテーションだけでは，遺伝子機能のデータベースは永久に完了しないという報告がなされている[1)]．

そこで本稿では，機械学習の枠組みによって GO アノテーションの自動化を行う手法を提案する．機械学習を用いることで，文献中から遺伝子の機能情報を抽出し，アノテーションが未だ行われていない文献に対して，文献の内容に基づいた GO タームの予測を行う．従来用いられてきた機械学習の手法とは異なり，本稿では**カーネル分類器**を用いた方法を提案する．カーネルは，学習を行う際に，分野ごとの知識を付加的に取り込むことができ，さらに効率的な計算ができる性質を有している．また，識別問題においては，従来の機械学習が有効に機能しない不均衡データ（各クラスの事例数に大きな偏りがあるデータ）に対して対しても，簡単かつ効果的な対処が可能である．

以降，2 章で関連研究について述べ，3 章では提案手法の詳細について述べる．4 章では，提案手法で用いるカーネル分類器の基本的な考え方について紹介する．5 章では提案手法と従来手法との比較実験の結果を示す．6 章で，本研究のまとめと今後の課題について述べる．

2. 関連研究

遺伝子の機能を報告する文献の数は膨大であり，GO アノテーションを手動で行うには，生物医学の専門的な知識と多大な労力を要する．このような背景から，GO ドメインの分類や

†1 神戸大学大学院

GO タームのアノテーションを目的として，TREC 2004 ゲノムトラック[4] や BioCreative[2] といった評価型のワークショップが開催された．

ゲノムトラックでは，論文中に記述された遺伝子の機能を GO ドメインと呼ばれる上位のカテゴリ（MF，CC，BP）に分類する共通タスクが設定された．ワークショップの参加者には，マウスの遺伝子とその遺伝子について記述された文献の組が与えられ，文献の内容に基づいて各遺伝子に GO ドメインを付与する．このタスクでは，Seki と Mostafa[9] が，同義語辞書と遺伝子名の曖昧一致を用いて，遺伝子について述べた文を同定する方法を提案している．彼らは，特定の遺伝子について言及した文を抽出し，これをベクトル表現に変換，教師情報による語の重み付けを行い，k 近傍法を用いて分類を行った．

このように，ゲノムトラックではオントロジー最上位の GO ドメインだけを対象にしたのに対して，BioCreative では GO ターム（すべて）のアノテーションを目的とした．Chiang と Yu[3] は文の自動対応付けを行うことで，「*gene product* plays an important role in *function*」といった文の特徴を学習し，文中で GO タームに対応する遺伝子機能が述べられているかの判別を行った．

総じて，ゲノムトラックの参加者は教師付き学習に基づく分類手法の有効性を報告している．一方で，BioCreative の参加者は文字列一致の手法を主に用いている．これらの戦略の違いは，TREC ゲノムトラックが 3 つのクラス，すなわち GO ドメインしか考慮しない一方，BioCreative では約 30,000 種類に及ぶ GO タームを対象としなければならず，訓練データ不足から教師付き学習手法を効果的に適用しにくいことによる．

手法によらず，既存の GO タームのアノテーションの特徴として，精度が極めて低い点が挙げられる．この問題が困難である理由の 1 つは，自然言語の取扱いの難しさにある．GO タームには 1 つのタームに対して通常いくつかの同義語が存在し，また，文献中に現れる遺伝子機能を表す文字列に対して複数の GO タームが対応することがある．また，遺伝子機能を表す文字列が文中に現れていたとしても，遺伝子機能の存在を否定する内容であれば、GO タームを付与する必要はない．これらの理由から，既存の文字列一致の方法を用いると，偽陽性（誤った GO タームの付与）が多くなる傾向にあり，適合率が低下する．

もう一つの問題として，GO タームの種類の多さが挙げられる．そのため，ある GO タームが付与された文献は極めて少量しか存在せず，また他の GO タームが付与された文献は多数存在するという状況が生じやすい．これは，いわゆる**不均衡データ**の問題であり，このようなデータを用いて信頼性の高い分類器を作成するには特段の注意が必要である．

これらの問題に対処する試みとして，テキスト情報とアノテーションに有益なテキスト以外の情報を組み合わせる手法も提案されている．Stoica と Hearst[11] は与えられた遺伝子の相同分子種（共通祖先の種分化によって生じた機能的に類似の遺伝子）を用いて GO タームのアノテーション候補を制限する方法と，同じ文献に複数アノテーションされた GO タームの共起性を利用することで，不要なアノテーション候補を除去する方法を提案している．Seki ら[8] はこの考えを拡張して，相同分子種を考慮した訓練データをもとに，分類器を逐次的に作成する方法を提案している．一方，Si ら[10] は，テキスト，事前知識，生物学的配列の 3 つの情報源から 5 つのスコアを求め，ロジスティック回帰によりこれらのスコアを統合して GO タームの分類を行う方法を提案している．なお，このように複数の異なる知識源を組み合わせることは重要であるものの，相同分子種のような情報は全ての遺伝子や種が持っているわけではない．

これらの研究に対し，本稿では文献情報を用いたカーネル法に基づく学習・予測手法を提案する．カーネルは効率的な計算ができる性質を有しており，また有用な情報を付加的に取り込むことができる．TREC ゲノムトラックのデータを用いた実験から，我々の手法が，文字列一致と相同分子種の情報を用いた従来の手法よりも優れていることを示す．

3. 文献に基づく GO アノテーション

本研究の目的は，文献中の記述と遺伝子機能を表す GO タームとの関係を学習し，新たな文献に対して遺伝機能（GO ターム）のアノテーションを行うことである．これは，アノテーションされていない文献の中で，遺伝子 X に関して GO ターム Y に対応する機能情報が述べれられているかを判別することであると言い換えることが出来る．図 1 に本研究で提案する GO アノテーションシステムの概要を示す．提案システムは，大別して学習と予測の処理を行う．以下では，特に，前処理として情報抽出と単語集合（Bag-Of-Words）表現への変換について説明する．

情報抽出： 文献中から，1）曖昧文字列一致を用いて遺伝子について言及した部分（本文が取得可能な場合，Seki と Mostafa[9] の方法を用いて抽出する），2）文献のタイトルとアブストラクト，3）GO タームの定義文（学習時のみ使用）を抽出する．

前処理と単語集合表現： 続いて，抽出したテキスト情報を単語集合表現へと変換する．形式的には，各事例を $\mathbf{x}_i = (x_{i_1}, \ldots, x_{i_D}) \in \mathbf{R}^D$ という形で表現する．ここで，x_{i_j} は，文書 i にある j 番目の語彙の総数であり，D は語彙数（単語の数）を表す．

学習： 遺伝子機能のアノテーションはマルチラベル分類の問題として考えることができる．つまり，各文献を事例として考えた場合，対応する複数のラベルの組み合わせを予測す

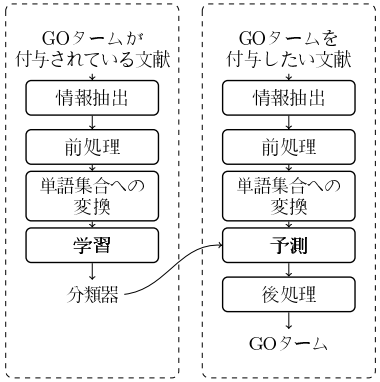

図 1 GO アノテーションシステムの概要：学習（左）と予測（右）．

ることに相当する（対応するラベルがない場合もある）．提案手法では，マルチラベル分類を行うための方法として，各クラス（GO ターム）1 つにつき 1 つの二値分類器を構築する one-vs-all の手法[7] を用いる．この手法は，モデルが簡潔であり，結果の解釈がし易く（どの特徴が GO タームの分類に役立ったかを確認できる），各分類器は独立であるために並列化が容易である．数式的には，文献 $\mathbf{x}_i$ に対して，$\mathbf{y}_i = (y_{i_1}, \cdots, y_{i_M}) \in \{-1, +1\}^M$ を割り当てることに対応する．ここで，M は GO タームの数であり，$y_{i_c} = +1$ は c 番目の GO タームが文献 $\mathbf{x}_i$ に付与されること，$y_{i_c} = -1$ は付与されないことを表す．

予測と後処理： 学習された分類器によって、GO タームの予測を行う．ただし，我々が使用する one-vs-all の手法では，二値分類器を逐次的に適用していくので，GO タームの付与中に矛盾する GO タームの組み合わせが生じる可能性がある．GO の構造である有向非巡回グラフの利点を生かした後処理を行えば，このような矛盾する GO タームの組み合わせを除去できるものと考えられる．具体的には，後処理として，GO タームが祖先と子孫の関係にあるとき，それらの内で尤もらしい方のみを付与する．

次章以降で，カーネル分類器を用いた学習と予測による GO アノテーションの詳細について説明する．

4. カーネル分類器

4.1 線形分類器の正則化

文献 $\mathbf{x}_*$ を所与としたとき，$s_c \colon \mathbf{R}^D \to \mathbf{R}$ を c 番目の GO タームのスコア関数とする．

$$s_c(\mathbf{x}*) = \boldsymbol{\beta}_c \cdot \phi(\mathbf{x}_*) \tag{1}$$

$\boldsymbol{\beta}_c = (\beta_{c_1}, \ldots, \beta_{c_D}) \in \mathbf{R}^D$ は c 番目の GO タームの重みベクトルである．直感的な解釈として， 強い正の値をとるとき（負の値をとるとき），重み β_{c_j} は j 番目の単語クラスに強く依存（クラスに依存していない）していることを示し，GO タームを付与する上で重要な手がかりとなる．ϕ は入力を高次元空間に写像するための関数であり，$\phi(\mathbf{x}) = \mathbf{x}$ のときは写像を行わない．

文献 $\mathbf{x}_*$ が c 番目の GO タームに属しているかどうかを決定するために，c 番目の GO タームの予測関数 $f_c \colon \mathbf{R}^D \to \{-1, +1\}$ を定義する．

$$f_c(\mathbf{x}_*) = sign(s_c(\mathbf{x}_*)) \tag{2}$$

ここで，$sign(a)$ は $a > 0$ のとき $+1$，それ以外の場合は -1 になる関数である．なお，f_c には閾値が存在しない．

本学習アルゴリズムの目的は，文献の集合 $\mathbf{x}_1, \ldots, \mathbf{x}_N$ と関連するラベルのベクトル $\mathbf{y}_1, \ldots, \mathbf{y}_N$ を与えたときに，以下の目的関数を最小化する重みベクトル $\boldsymbol{\beta}_c$ を学習することである．

$$L(\boldsymbol{\beta}_c) = C_c \sum_{i=1}^{N} \ell(y_{i_c}, s_c(\mathbf{x}_i)) + \frac{1}{2}||\boldsymbol{\beta}_c||^2 \tag{3}$$

C_c は c 番目の GO タームのハイパーパラメータであり，過学習の問題を防ぐために，モデルの複雑さを制御する役割を果たす．$C_c \to \infty$ は正則化しないことを示し，$C_c = 0$ は， 無限の正則化を行うことに対応する．ℓ は誤った予測を行った際の損失関数であり，ヒンジ損失関数，対数損失関数，二乗損失関数がそれぞれサポートベクトルマシン（SVM），ロジスティック回帰（LR），正則化最小二乗分類器（RLSC）に対応する．本稿では，これら 3 つのモデルを正則化線形モデルと見なす．

4.2 カーネル化

リプレゼンターの定理より，訓練データの事例を線形に組み合わせることで式 (3) を表現できる．

$$\boldsymbol{\beta}_c = \sum_{i=1}^{N} \alpha_{c_i} \phi(\mathbf{x}_i) \tag{4}$$

そのため，文献 $\mathbf{x}_*$ を与えた場合， c 番目の GO タームのスコアを $s_c(\mathbf{x}_*)$ と書き直すことができる．

$$s_c(\mathbf{x}_*) = \sum_{i=1}^{N} \alpha_{c_i} \phi(\mathbf{x}_i)\phi(\mathbf{x}_*) = \sum_{i=1}^{N} \alpha_{c_i} \mathcal{K}(\mathbf{x}_i, \mathbf{x}_*) \tag{5}$$

$\mathcal{K}$ はカーネル関数であり，$\mathbf{x}_i$ は $\alpha_{c_i} \neq 0$ のとき，c 番目の GO タームのサポートベクトルに対応する．

4.3 不均衡データの扱い

前述のように，GO タームの総数は約 3 万であり，限られた訓練データ中には，特定の GO タームが付与された論文数が多くなる一方，他の GO タームが付与された文献の数が少ないという状況が生じやすい．さらに，今回使用した one-vs-all の枠組みでは，個々のクラスに関して分類器を学習するため，負例の数が大きくなりやすい．GO アノテーションの従来研究では，不均衡データの問題はほとんど議論されていないものの，GO タームの効果的な予測を行うためには不均衡データへの適切な対処がきわめて重要である．

不均衡データの取扱いとして，Osuna ら[6] は，負例よりも強い重み付けを正例に対して行う手法を提案している．言い換えると，負例のクラスに対して正例のクラスよりも強い正則化を行う．この手法から着想を得て，本研究では，数多くある負例の影響を抑えつつ，事例を有効に活用する手法を新たに提案する．具体的には，重みの学習の際，目的関数である式 (3) の最小化を次式のように行う．

$$L(\boldsymbol{\beta}_c) = C_c \sum_{i=1}^{N} \mu_{i_c} \ell(y_{i_c}, g_c(\mathbf{x}_i)) + \frac{1}{2}||\boldsymbol{\beta}_c||^2 \tag{6}$$

$$\mu_{i_c} = \begin{cases} n_{\neg c}/n_c, & \text{if } y_{i_c} = +1 \\ 1, & \text{if } y_{i_c} = -1 \end{cases} \tag{7}$$

ここで μ_{i_c} は，事例 x_i が c 番目の GO を付与された場合，$n_{\neg c}$（c 番目の GO タームをラベルとして付与されていない事例の数）と n_c（c 番目の GO タームをラベルとして付与された事例の数の比率となり，それ以外のときは 1 となる．式で表現すれば，$n_c = |\{\mathbf{x}_i | y_{i_c} = +1\}|$，$n_{\neg c} = N - n_c$ となる．GO アノテーションの訓練データでは通常 $n_{\neg c} > n_c$ なので，負例よりも正例に強い重みが与えられることになる．我々の予備調査によると，この簡単なヒューリスティックが負例のダウンサンプリングよりも良い結果を示す．

4.4 ハイパーパラメータの調整

式 (6) の目的関数には，モデルの複雑さを制御するためのハイパーパラメータ C_c が存在し，より高性能の分類器を得るためには，その値を交差検定などで適切に決定することが重要である．また，前述の不均衡データの問題も考慮する必要がある．本研究では，以下のようにハイパーパラメータを調整する．まず，用意されたデータを検証データと訓練データに分割する．検証用のデータとして，正例 1 つに対して一定数の負例を用意する．負例の数は，ゲノムトラックの全訓練データにおける正負例の割合に従うように設定する．そして，残りを訓練データとして分類器を学習し，用意した検証データを用いて C_c に対する F_1 スコアを算出する．この作業を全ての正例が選択されるまで行い，スコアの平均を最大化する C_c をハイパーパラメータとして選択する．

5. 評 価 実 験

提案手法の枠組みを評価するため，TREC 2004 ゲノムトラックで構築されたデータを使用し，GO タームが付与された文献をテストデータとして実験を行った．このテストデータには 863 個[⋆1] の事例が存在し，各事例は，PubMed ID，文献中で述べられている遺伝子，GO タームの 3 組で表現されている．訓練データには，ゲノムトラックの訓練データセット（1,418 件の文献の全文データ）と MGD データベース（6,750 件の文献のアブストラクト）を用いた．データセットはマウスの遺伝子のみであり，前処理として，ストップワード・句読点・長い単語（50 文字以上）の除去，語形の変化を取り除き，小文字への変換を行った．

評価方法には，従来研究と直接的な比較を行うため，適合率，再現率，それらの調和平均である F_1 スコアを用いた．適合率は，正しく予測した GO タームの数を予測した GO の数で割った値であり，再現率は正しく予測できた GO タームの総数をテストデータ中にある GO タームの数で割った値である．

5.1 分類器毎の評価

本実験では，損失関数が異なる 3 つの分類器，すなわちサポートベクターマシン（SVM），ロジスティック回帰（LR），正則化最小二乗分類器（RLSC）を比較した．また，カーネルには式 (8) の線形カーネルを使用した．前述したように，これら 3 つの分類器は同一の予測方法を行い，最適な重みベクトル $\boldsymbol{\beta}_c$ の定義だけが異なる． 比較のため，ナイーブベイズ

⋆1 ゲノムトラックで構築されたデータには GO ドメインだけが付与されているため，本研究では，MGD データベースを基に，各レコードに付与された（複数の）GO タームの情報を復元した．そのため，ゲノムトラックよりも事例数が増加している．

表 1 ナイーブベイズ（NB）と線形カーネルを用いたサポートベクターマシン（SVM），ロジスティック回帰（LR），正則化最小二乗分類（RLSC）の分類性能の比較.

分類器	適合率	再現率	F_1 スコア
NB	0.21	0.14	0.17
SVM	0.36	**0.20**	**0.26**
LR	**0.39**	0.18	0.25
RLSC	0.34	0.18	0.24

分類器（NB）についても実験を行った．結果を表 1 に示す.

結果を見ると，SVM が最も良い性能を示していることが分かる．LR と RLSC は SVM にはわずかに劣るものの，ナイーブベイズ分類器よりも高い性能を示した．総じて，正則化線形モデルは GO アノテーションの問題に適しており，4.3 節と 4.4 節で述べたようなハイパーパラメータの調整と不均衡データの扱いを適切に行えば，比較的高い精度が得られることが判明した．なお，それらを考慮しなかった場合，SVM の F_1 スコアは 0.10 程度であった.

5.2 カーネル毎の比較

本節の実験では，サポートベクターマシンに 3 種類の異なるカーネルを適用し，分類器の性能比較を行う.

$$\mathcal{K}_{linear}(\mathbf{x}_i, \mathbf{x}_n) = \frac{\mathbf{x}_i \cdot \mathbf{x}_n}{||\mathbf{x}_i|| ||\mathbf{x}_n||} \tag{8}$$

$$\mathcal{K}_{poly}(\mathbf{x}_i, \mathbf{x}_n) = (1 + \mathcal{K}_{linear}(\mathbf{x}_i, \mathbf{x}_n))^d \tag{9}$$

$$\mathcal{K}_{plsa}(\mathbf{x}_i, \mathbf{x}_n) = \sum_{K \in \{8,16,32\}} \tilde{\mathcal{K}}_K(\mathbf{x}_i, \mathbf{x}_n) + \bar{\mathcal{K}}_K(\mathbf{x}_i, \mathbf{x}_n) \tag{10}$$

$\mathcal{K}_{linear}$ は線形カーネルであり，各事例は自身のノルムによって正規化される（つまりコサイン類似度）.

$\mathcal{K}_{poly}$ は d 次の多項式カーネルであり，順序を考慮しない 1-グラム，2-グラム，..., d-グラムの全組合わせに対応する.

$\mathcal{K}_{pLSA}$ は Hofmann[5] が提案した pLSA に基づくカーネルである．研究では，事例を表現するための特徴量として，遺伝子名によって抽出したテキストの断片，アブストラクトおよびタイトル，GO タームの定義文の 3 種類の情報を用いる．しかしながら，これらの情報には平均で 300 種以下の語彙しか含まれず，ベクトルの要素がほぼ 0 の疎なベクトルになるため，GO タームを特徴づける単語が少ない．そこで，pLSA を用いたカーネルの利用を提案する．pLSA モデルは全ての文献を用いて学習されるため，カーネルの中で単語の平滑化が行われる．少数の文献にしかアノテーションされていない GO タームを予測する上で，この手法は特に有効であると考えられる．$\tilde{\mathcal{K}}$ は文献 $\mathbf{x}_i$ と $\mathbf{x}_n$ のトピックの重なりに基づいた計算であると解釈できる．$\bar{\mathcal{K}}$ は，文献間の単語の経験分布を比較することで求められ，事後分布の重なりによる類似度によって重み付けされていると考えることができる.

表 2 サポートベクターマシーン（SVM）のカーネル毎の比較

カーネル	適合率	再現率	F_1 スコア
$\mathcal{K}_{linear}$	0.36	0.20	0.26
$\mathcal{K}_{poly}$ $(d = 2)$	0.35	0.19	0.25
$\mathcal{K}_{plsa}$	0.38	0.20	0.26
$\mathcal{K}_{plsa}$ (+U)	**0.39**	0.22	0.28
$\mathcal{K}_{plsa}$ (+U + T)	0.38	**0.24**	**0.29**

pLSA は教師なしのモデルであり，その副産物として，2 種類のカーネルを統合することで，ラベル付きのデータだけでなく，ラベルなしのデータの情報も利用することができる（半教師付き学習）．さらに，特定のテスト集合に対して分類器を適応させることも可能である（トランスダクティブ学習）．表記 K は pLSA モデルが K 個のトピックに従うことを表している.

表 2 において，$\mathcal{K}_{plsa}$ （+U） は MGD データベースから取得したラベルなしのアブストラクト 10,000 件を使用して，pLSA モデルを学習した結果である（半教師付き学習）．$\mathcal{K}_{plsa}$ （+U + T） は，テストデータについても pLSA を適用し，より分類対象に注目した学習の結果である（トランスダクティブ学習）．線形カーネルと比べた性能の向上は僅かであるものの，学習データ不足に悩まされる遺伝子アノテーションの問題において，これら生成モデルを用いた手法の結果は興味深いといえる．なお，単語の重要性の定量化方法としてよく使用される *tf-idf* についても実験を行ったが，線形カーネルよりも性能は低下した.

5.3 従来手法との比較

GO アノテーションにおいて，アルゴリズム毎の性能差を俯瞰するため ，Stoica と Hearst[11]，Seki ら[8] の手法との比較を行った．結果を表 3 に示す．Stoica と Hearst の結果は著者らの実装によるものであり，Seki らの結果は文献 8) に基づく．PROPOSED (+O) は，後述するように，相同分子種による制約を後処理として提案手法に加えた結果で

表 3 既存手法との比較

手法	適合率	再現率	F_1 スコア
PROPOSED	0.38	0.24	0.29
PROPOSED (+O)	**0.42**	0.23	**0.30**
STOICA & HEARST	0.19	**0.46**	0.27
SEKI ET AL.	0.26	0.27	0.26

ある．

結果，我々の手法（PROPOSED）が適合率と F_1 スコアにおいて最も良い結果を示した．一方，再現率については，Stoica と Hearst の手法が最良であった．この理由は，後者では，相同分子種の情報を使うことで一貫性のない候補を除去しているためだと考えられる（相同分子種は，共通の祖先の遺伝子を受け継いだ異なる種が持つ遺伝子のことであり，一般的に類似した遺伝的性質を持ち易い傾向にある）．なお，相同分子種に基づく制約は，後処理として提案手法でも利用可能であり，これを施した結果が PROPOSED (+O) である．具体的には，相同分子種（ラット）に付与されていない GO タームの予測を抑制することで，適合率の向上を図った．その結果，再現率が微減したものの，適合率が 0.38 から 0.42 に向上し，F1 スコアも 0.30 に向上した．マウスとラットはきわめて関連の高い種同士であるため，再現率の悪化を最小限に止めつつ適合率を上げることが出来たものと考えられる．t 検定を行ったところ，提案手法 PROPOSED (+O) と Stoica と Hearst の手法の差は，有意水準 0.05 で統計的に有意であった（$p = 0.03$）．

6. おわりに

本研究では，カーネルを用いた GO タームのアノテーション手法を提案した．TREC ゲノムトラックのデータによる実験から，提案手法により，文字列一致および異種間の情報を用いた従来手法よりも高い適合率と F_1 スコアが得られることが分かった．また，GO タームのアノテーションに起こりやすいラベル付きデータが不足する問題について，潜在トピックをカーネルに取り込むことで効果的に対処することが出来た．さらに，GO ターム毎（クラス毎）に正則化を行うことで，貴重な訓練データを活用しつつ，不均衡データのに対処する手法を提案した．

今後の発展として，カーネルに基づく 2 つの方法に取り組むことを考えている．1 つは，大規模データに適したカーネルを用いること，もう 1 つは，Gene Ontology の有向非巡回グラフの構造や他の知識源に対してそれぞれカーネルを定義し，マルチカーネルの統合・学習を行うことである．

参 考 文 献

1) Baumgartner, WilliamA., J., Cohen, K.B., Fox, L.M., Acquaah-Mensah, G. and Hunter, L.: Manual curation is not sufficient for annotation of genomic databases, *Bioinformatics*, Vol.23, No.13, pp.i41–48 (2007).
2) Blaschke, C., Leon, E., Krallinger, M. and Valencia, A.: Evaluation of BioCreAtIvE assessment of task 2, *BMC Bioinformatics*, Vol.6, No.Suppl 1, p.S16 (2005).
3) Chiang, J.-H. and Yu, H.-C.: MeKE: discovering the functions of gene products from biomedical literature via sentence alignment, *Bioinformatics*, Vol.19, No.11, pp.1417–1422 (2003).
4) Hersh, W., Bhuptiraju, R.T., Ross, L., Cohen, A.M. and Kraemer, D.F.: TREC 2004 Genomics Track Overview, *Proceedings of the 13th Text REtrieval Conference (TREC)* (2004).
5) Hofmann, T.: Learning the Similarity of Documents: An Information-Geometric Approach to Document Retrieval and Categorization, *Advances in Neural Information Processing Systems 12* (Solla, S.A., Leen, T.K. and Müller, K.-R., eds.), pp. 914–920 (1999).
6) Osuna, E.E., Freund, R. and Girosi, F.: Support Vector Machines: Training and Applications, Technical report (1997).
7) Rifkin, R. and Klautau, A.: In Defense of One-Vs-All Classification, *J. Mach. Learn. Res.*, Vol.5, pp.101–141 (2004).
8) Seki, K., Kino, Y. and Uehara, K.: Gene Functional Annotation with Dynamic Hierarchical Classification Guided by Orthologs, *Discovery Science* (Gama, J., Costa, V., Jorge, A. and Brazdil, P., eds.), Vol.5808, pp.425–432 (2009).
9) Seki, K. and Mostafa, J.: Gene Ontology Annotation as Text Categorization: An Empirical Study, *Information Processing & Management*, Vol.44, No.5, pp.1754–1770 (2008).
10) Si, L., Yu, D., Kihara, D. and Fang, Y.: Combining gene sequence similarity and textual information for gene function annotation in the literature, *Information Retrieval*, Vol.11, pp.389–404 (2008).
11) Stoica, E. and Hearst, M.: Predicting Gene Functions from Text Using a Cross-Species Approach, *Pacific Biocomputing Symposium*, Vol.11, pp.88–99 (2006).