社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DBS－146 (35)
2008／9／22

相同分子種を利用した多階層分類による遺伝子機能アノテーション

木野 嘉祐† 関 和広†† 上原 邦昭 †††

†, ††† 神戸大学大学院工学研究科 〒6507-8501 兵庫県神戸市灘区 1-1
†† 神戸大学自然科学系先端融合研究環 〒6507-8501 兵庫県神戸市灘区 1-1
E-mail: †kino@ai.cs.scitec.kobe-u.ac.jp, ††seki@cs.kobe-u.ac.jp, †††uehara@kobe-u.ac.jp

あらまし 本研究では，共通祖先からの種分化によって生じた遺伝子（相同分子種）を利用し，遺伝子機能の階層構造を考慮した多階層分類による遺伝子機能アノテーションの手法を提案する．遺伝子機能とは，当該遺伝子（の生成物）が持つ性質であり，FlyBase や MGI など既存のモデル生物データベースにおいて各遺伝子の主要な情報として付与されている．これらの遺伝子機能の記述は，複数のモデル生物データベースに対する横断的なアクセスを可能にするため，一種の統制語彙である Gene Ontology（GO）に基づいて行われている．GO は類似の性質ごとに階層関係を生成し，無閉路有向グラフ（DAG）として体系化されている．提案手法は，所与の遺伝子とその相同遺伝子とのマッピングに基づき，当該相同遺伝子に既に付与されている遺伝子機能を制約とし，この制約上で利用可能な訓練事例から動的に分類器を作成することで高精度な分類を行う．先行研究との比較により，提案手法の有効性を示す．
キーワード オントロジー，多階層分類，遺伝子機能，相同分子種，アノテーション

Gene Functional Annotation by Ortholog-based Hierarchical Classification

Yoshihiro KINO†, Kazuhiro SEKI††, and Kuniaki UEHARA†††

†, ††† Graduate School of Engineering, Kobe University Nada 1-1, Kobe 6507-8501, Japan
†† Organization of Advanced Science and Technology, Kobe University Nada 1-1, Kobe 6507-8501, Japan
E-mail: †kino@ai.cs.scitec.kobe-u.ac.jp, ††seki@cs.kobe-u.ac.jp, †††uehara@kobe-u.ac.jp

Abstract This paper proposes a novel method for gene functional annotation in the framework of hierarchical classification that uses as constraints the known (already annotated) functions of genes orthologous to a given gene. A gene function is a biological property of a gene or the product it encodes, and is given as main information on each gene in the model organism databases, such as FlyBase and MGI. These gene functions are annotated based on Gene Ontology (GO) that is a kind of controlled vocabulary structured as Directed Acyclic Graph (DAG) to enable uniform access to multiple model organisms databases. Our proposed approach uses gene functions of orthologous gene as constraints, dynamically creating classifiers from the training data available under the constraints. The effectiveness of the proposed approach is demonstrated in comparison with the related work.
Key words Ontology, Hierarchical Classifier, Gene Function, Orthlogous gene, Annotation

1. まえがき

ヒトゲノム計画の完了以降，分子生物学の重要な課題の一つとして，個々の遺伝子の機能同定に関する研究が活発に行われている．マイクロアレイなど高速な遺伝子解析技術の登場にも後押しされ，生物医学分野の学術論文は近年ますます増大し，MEDLINE が現在索引を提供する論文数は，1800 万件にも達する．しかしながら，これら大量の論文は自然言語で記述されているため，所望の情報を網羅的に収集・利用することは容易ではない．これらの文章中に埋もれた有用な情報を整理・構造化し，効率的なアクセスを可能にするため，現在多くの研究がなされている．遺伝子（の生成物）が持つ性質を表す遺伝子機能は，FlyBase や MGI など既存のモデル生物データベースにおいて各遺伝子の主要な情報として付与されている．これらの遺伝子機能の記述は，一種の統制語彙である Gene Ontology（GO）に基づいて行われている．Ontology とは，データ間の論理的な関係性を記述した構造で，知識を共有するために必要なデータの分類体系の枠組みである．GO は対象を遺伝子機能として，類似の性質ごとに階層関係を生成し，無閉路有向グラフ（DAG）として体系化されている（図 1）．

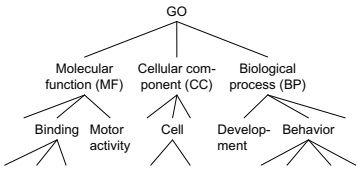

図 1　Gene Ontology の基本構造

図 1 のように，GO には第一階層の Biological Process（BP），Cellular Component（CC），Molecular Function（MF）の三つの遺伝子機能（GO ドメイン）が存在し，第一階層以下すべての遺伝子機能（GO ターム）はそれぞれ類似の性質で細分化されている．

GO が定義する遺伝子機能数は 2008 年 8 月現在 26651 件で，Mouse や Human，Rat など種々のモデル生物種に関して，遺伝子機能の記述に利用されている．これより，生物医学分野における各種生物種の情報を整備することができる有用な統制語彙となる．本研究では，遺伝子機能付与の対象である遺伝子とは別の生物種の相同遺伝子に既に付与された遺伝子機能を制約として利用し，さらに，GO の階層構造を考慮したテキスト分類の手法を用いた多階層分類を行うことで，文献への高精度な遺伝子機能アノテーション（遺伝子の記述を含む医学文献に対して遺伝子機能を付与すること）を目指す．

以下，2 節で遺伝子機能アノテーションの関連研究についてまとめる．続いて，3 節では提案手法について詳述し，4 節で提案モデルの評価実験について報告する．5 節で本論文のまとめと今後の課題について述べる．

2. 関 連 研 究

遺伝子機能アノテーションは，アノテーション（注釈付け）の対象となる遺伝子数，遺伝子機能数，文献数の膨大さ，さらには，内容の専門性の高さゆえに，大変な労力に加え，分野固有の広範な知識が必要とされるため，専門家によって手作業で行われている．よって，この遺伝子機能アノテーションを自動化，あるいは半自動化できれば，人間の負担を大きく軽減することができる．このような背景から，遺伝子機能アノテーションに関する研究が Text Retrieval Confemrence（TREC） 2004 の Genomics トラック [1]，および，BioCreative 2003 [2] で行われた．これら二つの会議で取り組まれたタスクとその参加者，および関連研究の手法に関して説明する．

- TREC Genomics Track

2004 年に開催された Mouse の遺伝子を対象とした遺伝子機能アノテーションタスクであり，遺伝子の記述を含む文献に対して付与される GO ドメインの推定を行う．

Settle と Craven [3] は TREC のタスクに対してベイズ（NB）分類器と最大エントロピー（ME）モデルを多段的に用いた枠組みを提案した．その中で，論文の論理構造（セクション）に注目し，あらかじめ定義した 6 つのセクション型のそれぞれについて NB 分類器を構築し，その結果を ME モデルを用いて統合した．

また，関とモスタファ [9] は χ 二乗統計値と複数の重み付け法を用いた kNN 分類器による分類手法を提案した．具体的には，最初に論文中で言及されている個々の遺伝子について記述されたテキスト断片の抽出を行った．次に，分類において重要な素性となる単語を χ 二乗統計値に基づいて選択（素性選択）し，語の集合で表されているテキスト断片を単語ベクトルへと変換し，TFIDF 法など計三種類による素性の重み付けを行った．そして，kNN によって作成した分類器によって分類することで GO ドメインの推定を行った．

- BioCreative

2003 年に開催された生物医学分野における Human の遺伝子を対象とした遺伝子機能アノテーションタスクであり，遺伝子の記述を含む文献に対して付与される GO ターム の推定を行う．

Ray と Craven [4] は BioCreative のタスクに対して 4 つのアノテーションデータベースから，各 GO ターム の統計的に学習する手法を提案した．その中で，GO ターム を含むパラグラフ内で一致する可能性の高い単語を抽出し，その単語のパターンマッチによる結果をランク付けすることで各カテゴリーにおけるナイブベイズ分類を行った．

また，Stoica と Hearst [5] は，相同分子種に付与されている遺伝子機能を制約として利用し，二種類の文字列のマッチング手法を提案した．相同分子種とは，共通祖先からの種分化によって生じた遺伝子の組であり，元々は同じ遺伝子であったため類似の機能を持つことが多い．ここでは，Human の遺伝子に対して，相同分子種となる Mouse の遺伝子に付与されている遺伝子機能を文献に付与する遺伝子機能の制約として設けた．その制約の中で，一つ目のマッチング手法として，相同分子種に付与されている遺伝子機能を文献内で探索し，その遺伝子機能の記述の 75%以上の単語が存在した場合にその文献は当該遺伝子機能の記述を含むと考え遺伝子機能を付与した（Cross Species Match（CSM））．二つ目として，すべての遺伝子機能を文献内で探索する，各遺伝子機能内の単語と完全一致した遺伝子機能と相同分子種に付与されている遺伝子機能との関連性を χ 二乗統計値を用いることでスコア化し，その値が閾値を越えた場合に該当遺伝子機能として文献に付与した（Cross Spicies Correlation（CSC））．これは，相同分子種に付与されている遺伝子機能が各文献に付与される該当遺伝子機能となる可能性が高いことにより，文献内での記述の存在による関連性の高低を図った方法である．この CSM と CSC によって付与された遺伝子機能を統合することで GO ターム の推定を行った．

TREC と BioCreative の手法を比較した場合，TREC では分類手法がいくらか見られるのに対して，BioCreative では主に文字列や単語列マッチング手法が多く見られる．マッチング手法は，遺伝子や遺伝子機能の記述には別名や省略形，シンボル名などの存在に加え，空白や記号の挿入，削除による表記のゆれがあり，文書内の遺伝子や遺伝子機能などの記述を検出する

ことは容易ではない．遺伝子の記述例として，membrane associated transporter protein は，underwhite, Matp, uw, Dbr, bls, Aim1 などの別名が存在し，表記のゆれの例として，遺伝子機能を 3'-5'-exoribonuclease activity と記述する文献もあれば，3' to 5' exoribonuclease activity や，3' → 5'exoribonuclease activity と記述する文献も存在するため，完全一致で検出することは難しい．さらに，一つの遺伝子に対して一般に複数の遺伝子機能が付与されているため，各文献内で検出した遺伝子機能の記述が当該遺伝子機能の記述と必ずしも同定できない．それに対して，分類手法は訓練事例から個々の当該遺伝子機能に特徴的な記述（語彙）を取得することができるため，遺伝子機能の同定に適していると考えられる．しかし，分類を行う中で，問題となるのが，遺伝子機能の総数と正しい記述を含む文献数（訓練事例数）である．遺伝子機能数が増加するほど分類対象が多くなるため，有用性は失われる．そこで，分類手法をより有効に適用するために，遺伝子機能数に制約を設けることができれば望ましい．本研究では，Stoica と Hearst [5] らが用いた相同分子種に着目し，付与対象となる遺伝子機能に対して制約を設けることでより効果的な分類を実現する．さらに，GO の多階層構造が遺伝子機能の性質ごとに構成されていることを利用して，より多くの訓練事例を獲得し，分類手法による高性能な遺伝子機能アノテーションを目指す．

3. 提案手法

3.1 あらまし

本研究では，Mouse の遺伝子が記述された文献に付与する遺伝子機能に制約を設け，GO の多階層構造を考慮した分類手法によって遺伝子機能アノテーションを実現する．制約として設ける遺伝子機能（制約遺伝子機能）は，共通祖先からの種分化によって生じた遺伝子（相同遺伝子）を利用する．遺伝子機能は Mouse の相同分子種となる生物種にも同様に付与されているため，その情報をもとに制約として設ける．分類手法において必要な訓練事例を，性質ごとに構成された GO の階層構造を考慮することで，訓練事例の収集を行い，可能な限り多くの訓練事例を取得する．その訓練事例数と制約遺伝子機能数がある条件を満たした場合，動的に分類器を作成し，文献に付与される遺伝子機能を同定する．

3.2 相同分子種による制約

約 3 万件の遺伝子機能に制約を設けることで，遺伝子機能の母数を減らし効率的かつ効果的な分類を行う．そのため，MGI [6] や Gene Ontology Annotation（GOA） [7] などの既存の各生物種情報を記載したモデル生物データベースより，利用可能な相同分子種を選択し，候補となる遺伝子機能を制約として各文献に付加する．

本研究で対象としている Mouse の遺伝子の相同分子種として Human，Rat，Chimpanzee，Dog など様々な生物種が存在する．これらの相同分子種の各遺伝子に対して GO の遺伝子機能が付与されているため，文献内で発見された Mouse の遺伝子と相同遺伝子との対応をとり，相同遺伝子に付与されている遺伝子機能を候補となる遺伝子機能（制約遺伝子機能）として Mouse の当該遺伝子を含む各文献に付与する．例えば，Mouse の遺伝子 Sox21 に対して，Human の相同遺伝子として SOX21 が存在する．この遺伝子には，GO:0003700，GO:0003702，GO:0006325，GO:0006350，GO:0006357，GO:0005634 などの遺伝子機能が付与されている．これらを Mouse の遺伝子 Sox21 の記述を含む文献に付与する遺伝子機能候補とし，これらの中から付与すべき遺伝子機能を推定する．

3.3 多階層分類

相同分子種によって制約遺伝子機能を設けた文献に対して，GO の階層構造を考慮した多階層分類を行う．その中で各種データベースより相同分子種のデータや GO における階層構造の抽出・評価データの生成を行う．その各処理について説明し，最後に本研究の目的となる GO タームの推定方法に関して詳述する．

3.3.1 データ抽出

生物医学分野で様々な情報を含有したデータベースが存在する．本研究では MGI と GOA の二つのモデル生物データベースを利用して様々な情報を抽出する．また，Gene Ontology（GO）データベース [8] に記載された GO の詳細情報より階層関係の取得を行う．以下，本研究で利用する各種情報資源について述べる．

- MGI データベース

主に，各生物種のデータベース ID 情報や生物種間のマッピング表などが存在する．これにより，Mouse の相同分子種である Human や Rat，Chimpanzee，Dog などとの遺伝子間の対応を取得することができる．

- GOA データベース

主に，各生物種の詳細情報が存在する．これにより，遺伝子名や同義語，略称，付与されている遺伝子機能などの制約遺伝子機能の付与を行う上で必要なデータを取得することができる．

- Gene Ontology データベース

主に，遺伝子機能に関する情報が存在する．これにより，遺伝子機能同士の階層関係や，その遺伝子機能の持つ性質，遺伝子機能の名称，ID，同義語など GO における詳細情報を抽出することができる．図 2 に GO のレコード例を示す．ここで，*id* タグが GOID，*name* タグが GO ターム，*namespace* が GO ドメイン，*is_a* が当該遺伝子機能の 1 階層上の遺伝子機能を表している．

GO タームの推定を行う際には，分類を行う階層の探索や訓練事例の同定のため，遺伝子機能の階層構造を抽出する必要がある．分類の対象となる階層を探索するためには各遺伝子機能の階層と位置を把握しておかなければならない．また，訓練事例数が遺伝子機能数に比べ大幅に少ないため，階層関係を考慮することで，極力多くの訓練事例を収集する．これらの理由により，階層構造の抽出を行う．

3.3.2 GO ターム推定

制約遺伝子機能が付与された各文献に対して，制約遺伝子機能数と訓練事例数を変数として多階層構造を考慮することによって動的に分類し，GO タームの推定を行う．GO ターム推定の流れを図 3 に表す．

```
[Term]
id:  GO:0000001
name:  mitochondrion inheritance
namespace:  biological_process
def:"The distribution of mitochondria, including the
mitochondrial genome, into daughter cells after mitosis or
meiosis, mediated by interactions between mitochondria and
the cytoskeleton."
[GOC:mcc, PMID:10873824, PMID:11389764]
synonym:  "mitochondrial inheritance"
is_a:  GO:0048308 !  organelle inheritance
is_a:  GO:0048311 !  mitochondrion distribution
```

図 2　Gene Ontology のレコード例

1: **Input:** *Test_Instance*
2: **Output:** *GO_Term* = {};
3: **for** *every Test_Instance* **do**
4: **Get** *Restrict_Gene_Function (RestrictGF)* .
5: **if** *(RestrictGF >= 2)*
6: **Get** *Common_Gene_Function (CommonGF)* .
7: **while** *(CommonGF >= 2)*
8: **if** *(Training_Instance > n)*
9: **Make Classifier.**
10: **add** *Classfied_Gene_Function* **to** *GO_Term.*
11: CommonGF - -;
12: **else if** *(RestrictGF == 1)*
13: **add** *RestrictGF* **to** *GO_Term.*

図 3　GO ターム推定

- 訓練事例

分類における GO ターム推定で最も重要な変数は訓練事例数である．限られた訓練事例を用いて効果的に分類を行うため，訓練事例の閾値 n の違いによる分類精度の変化を観測する．

- 制約遺伝子機能

制約遺伝子機能数は文書間で必ずしも同じであるとは限らない．各文献に付与される制約遺伝子機能数が 0 個や 1 個の場合もあれば，多数の場合もあり，該当する制約遺伝子機能の数には大きな差がある．各文献に対して制約遺伝子機能数に応じた処理を行う必要がある．制約遺伝子機能を持たない場合には，分類するための制約遺伝子機能が存在しないため，付与は行わないこととし，1 個の場合は，唯一の制約遺伝子機能しか存在しないため，その制約遺伝子機能を当該遺伝子機能として付与を行うこととする．2 個以上の場合は分類可能なため，制約遺伝子機能の全てが訓練事例数の閾値を満たすとき分類を行う．

図 3 について説明すると，最初に各文献に付与された制約遺伝子機能数によって分類・付与の判断を行う．そして，分類できると判断された場合，各制約遺伝子機能に共存する遺伝子機能 共通遺伝子機能） の探索を行う．これは性質によって分かれる一つ上の階層である親の遺伝子機能を探索するためである．そこで，その共通遺伝子機能が最も多く含まれる階層の一階層下の階層を分類対象として分類器の生成を行う．生成された分類器によって，各文献の分類を行い，遺伝子機能の付与を行う．この処理を共通遺伝子機能が 2 個以上の制約遺伝子機能の階層の中に存在する場合，その数を 1 ずつ減少させることで深い階層まで分類する処理を繰り返し，多階層分類を行う．

表 1　相同分子種により付加した制約遺伝子機能の正解率

Ortholog	正解件数	正解率
Human	287	47.5%
Rat	431	71.3%
HumanRat	489	80.9%

4. 評価実験

4.1 データセット

本研究では，TREC2004 で配布されたデータを用いる．GO ドメインの推定を行う評価データ数が 495 件に対して，訓練事例となる文献数が 872 件である．また，評価データは以下のように文献 ID，データベース ID，遺伝子名，GO ドメイン，Evidence コードを各評価事例が保有している，このデータを用いて GO ドメイン推定の評価を行ことができる．

```
12411446 MGI:2448704 Afmid MF TAS
12411446 MGI:2448704 Afmid CC IDA
```

しかし，このデータは GO ドメイン推定を意図したデータであるため，GO ターム推定実験を行うにはデータが不十分である．そこで，これらのデータに新たな情報を付加し，GO ターム推定実験に利用する．例が示す通り，各文献に付与されている遺伝子機能は最上位の GO ドメインのみで記述されているため，どの GO タームが付与されているのか同定することができない．そのため，GO ドメイン以外の情報より，評価データと一対一対応する GO タームを MGI データベースより抽出する．その結果，以下のような形式に書き換え，GO ターム推定における評価データ（604 件） として使用する．

```
12411446 MGI:2448704 Afmid MF TAS GO:0004061
12411446 MGI:2448704 Afmid CC IDA GO:0005737
```

ここで生成したデータを用いて相同分子種の妥当性や GO ターム推定の評価を行う．

4.2 相同分子種の妥当性

本研究では Mouse の相同分子種として，Human，Rat を試験的に利用し，これらの生物種の遺伝子に付与されている遺伝子機能を制約として利用する．まず，これらの相同分子種の制約としての妥当性に関して検証する．

4.1 節で示した形式の 604 件の評価データを利用することで，相同分子種の妥当性を検証する．まず，遺伝子の記述を含む各文献に対して，制約遺伝子機能をそのまま付与した場合の正解率を調べた．結果を表 1 に示す（表 1 の正解率は以下で示す Recall と同等である）．

表 2　GO ドメイン推定結果

	Precision	Recall	F-Score
TREC（BEST）	0.441	0.769	0.561
TREC（Worst）	0.169	0.133	0.149
TREC（Mean）	0.360	0.581	0.382
先行研究	0.549	0.642	0.592
追試	0.378	0.782	0.510

4.2.1　実験結果と考察

表 1 の結果より，Mouse の相同分子種として Human より Rat の方が優れているといえる．これは，Human より Rat の方が生物種として Mouse との類似性が高いためであると考えられる．また，Human と Rat の両方を付与した場合が最も制約遺伝子機能を多く含んでいるため，制約遺伝子機能として付与する場合には，一つの生物種を用いるより統合して用いた方が有用であると考えられる．

4.3　多階層分類の評価実験

制約遺伝子機能や訓練事例に閾値を設け，様々な条件の中で評価実験をおこなった．最初に本研究で用いた評価指標について，その後，多階層分類における GO ドメイン推定，GO ターム推定の評価実験に関して詳述する．

4.3.1　評 価 指 標

TREC や BioCreative などの参加者の実験報告との比較を容易にするため，同一の評価指標である F 値を用いた．F 値は次のように，再現率（R）と適合率（P）の調和平均として定義されている．

$$P = \frac{\text{システムが付与した正しいクラス数}\ (tp)}{\text{システムが付与したクラス総数}\ (fp)} \quad (1)$$

$$R = \frac{\text{システムが付与した正しいクラス数}\ (tp)}{MGI\ \text{によって付与されたクラス総数}\ (fn)} \quad (2)$$

$$F = \frac{2 \times P \times R}{P + R} \quad (3)$$

4.3.2　GO ドメイン推定

関とモスタファ [9] の手法と同様に，テキスト断片の抽出，χ 二乗統計値を用いた素性選択，TFIDF 法を用いた素性の重み付けを行い，SVM を用いて分類を行った結果を表 2 に示す．表 2 より，TREC のベストスコアや先行研究の結果に劣るものの，TREC 参加者の平均に比べ，顕著に高い評価値を得ることができた．先行研究と同じ結果が得られなかった原因として，素性の重みや利用した分類器等の相違が考えられる．

4.3.3　GO ターム推定

制約遺伝子機能を用いて多階層分類における GO タームの推定を行う．ここで，Rat の制約遺伝子機能を用いて訓練事例数の閾値を変化させながら分類実験を行う．その結果を表 3 に示す．

分類結果より，訓練事例の閾値が 1 のとき最も良い評価値を示していることがわかる．

さらに，ここで，文献の中で唯一の制約遺伝子機能のみを持つ事例に関して説明する．分類するためには，各文献に対して 2 個以上の制約遺伝子機能が存在しなければならない．そのため，一つの制約遺伝子機能しか持たない文献は分類の対象とすることができない．しかし，文献に対して制約遺伝子機能が一つしか存在しないため，その制約遺伝子機能は当該遺伝子機能として付与される可能性が高いと考え，付与を行う．

表 3　訓練事例の閾値を変化させた時の GO ターム推定結果

訓練事例数	Precision	Recall	F-Score
1	0.105	0.139	0.119
2	0.100	0.098	0.099
3	0.096	0.058	0.096
5	0.125	0.043	0.064

表 4　他の手法との比較実験

	Precision	Recall	F-Score
Stoica and Hearst[5]	0.168	0.121	0.140
Chiang and Yu[10]	0.332	0.051	0.089
提案手法（Human）	0.108	0.161	0.129
提案手法（Rat）	0.128	0.207	0.158
提案手法（Human と Rat）	0.159	0.182	0.168

これを，Human，Rat，Human と Rat の三種類の制約遺伝子機能に対して付与し，評価実験を行う．この結果を同じ MGI データベースを用いた Stoica と Hearst [5] や Chiang と Yu [10][11] と比較する（データ量は異なる）．結果を表 4 に示す．表 4 より，提案手法は，Precision の値が関連研究に対して劣るものの，Recall の値は相同分子種の利用により高いことがわかる．結果の比較により，相同分子種を用いたマッチング手法による Stoica らの研究よりも，分類手法による提案手法の方が評価値が高かったことから，相同分子種を利用した多階層分類による遺伝子機能アノテーションは有用であると考えられる．また，提案手法の中でも，相同分子種の妥当性における予備実験同様，Human と Rat の両方を制約遺伝子機能として用いたとき最も高い評価値を示した．

4.4　追 加 実 験

本研究の分類における重要な要因は訓練事例数であることは前述した．一般的に，訓練事例が多いほど文献の分類に有用であると考えられるのに反して，本研究の多階層分類では表 3 より訓練事例数の閾値が 1 のときに最も高い評価値を示した．これまでの実験では，訓練事例の条件を満たさない場合，分類・付与は行わなかった．そのため，本来付与されるべき制約遺伝子機能が除去され，評価値が低下した可能性がある．そこで，閾値の条件緩和による追加実験を行う．

追加実験方法は，Rat の制約遺伝子機能に対して，訓練事例の条件を満たさない制約遺伝子機能数が n 個以下の場合は，その階層の制約遺伝子機能をすべて付与する．その結果に分類結果と唯一制約遺伝子機能の付与を組み合わせ，評価値の向上を図る．ここでは，経験的に $n = 1$ とする．

実験結果より，訓練事例の閾値を 8 とした場合に最も良い評価値を示し，先述の結果よりも+20%増しの結果となった．追加実験として，訓練事例の閾値の条件緩和を行ったことにより，多階層分類において訓練事例の使用方法によって結果は大きく異なり，効率良く訓練事例を用いることで多階層分類は有用で

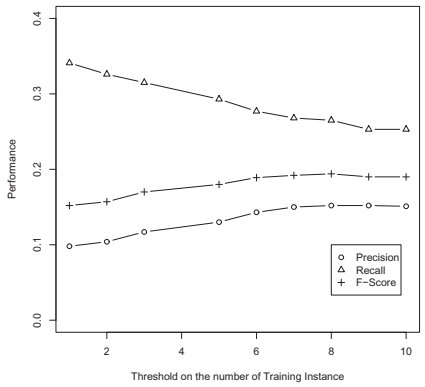

図 4　訓練事例数の閾値による実験

あることが確かめられた．

5. あとがき

本研究では，相同分子種を利用し，さらに多階層構造を考慮することで遺伝子機能アノテーションをおこなった．まず，Mouse の相同分子種である Human と Rat において，各遺伝子に付与されている遺伝子機能を Mouse の制約遺伝子機能として用いることは有用であると考えられる．特に，Rat は Mouse の相同分子種として，遺伝子機能アノテーションにおいて大きな役割を示す．また，制約遺伝子機能として 2 種類の相同分子種の遺伝子機能を統合することも有用であるといえる．

次に，先行研究との比較より，GO の階層構造を考慮して分類を行うことの有用性を示すことができた．しかし，訓練事例の使用方法によって結果は大きく異なると考えられる．そのため，今後，他のデータを用いて多階層分類の有用性を図っていきたいと考えている．

今後の課題として，まず多階層分類において，訓練事例の使用方法の改善を行い，より少ない訓練事例の中で高精度な分類が行えるような手法を考える．また，現在，他のアプローチ方法として，編集距離を用いたマッチング手法や，ベクトル空間モデルを用いた検索手法 [12] に関して取り組んでいる．分類手法のみでは，訓練事例を持たない場合に遺伝子機能を付与することが不可能となるため，マッチング手法における各遺伝子機能と文献の編集距離の計算，文献内の遺伝子機能の記述の検出，検索手法におけるベクトル空間モデルによる遺伝子機能の定義と遺伝子機能の記述との距離の計算，最も近い遺伝子機能の検索などを統合することで，全ての遺伝子機能付与の処理を網羅できるような，より高性能な遺伝子機能アノテーション手法を考えていく．

文　献

[1] W. Hersh, R. Bhuptiraju, L. Ross, P. Johnson, A. Cohen and D. Kraemer, TREC 2004 Genomics Track Overview, In Proceedings of the 13th Text REtrieval Conference (TREC) , 2004.
[2] BioCreative, http://biocreative.sourceforge.net/.
[3] B. Settles and M. Craven, Exploiting zone information, syntactic rules, and informative terms in gene ontology annotation of biomedical documents, In Proceedings of the 13th Text REtrieval Conference (TREC) , 2004.
[4] S. Ray and M. Craven, Learning statistical models for annotating proteins with function information using biomedical text, BMC Bioinformatics, (6S1) , 2005.
[5] E. Stoica and M. Hearst, Predicting Gene Functions from Text Using a Cross-Species Approach, Pacific Symposium on Biocomputing 11:88-99, 2006.
[6] Mouse Genome Infomatics, http://www.informatics.jax.org/.
[7] Gene Ontology Annotation, http://www.ebi.ac.uk/GOA/.
[8] The Gene Ontology, http://www.geneontology.org/.
[9] 関　和広，モスタファ　ジャビド, 多様な遺伝子名認識と文書分類を用いた Gene Ontology アノテーション, 電子情報通信学会論文誌 Vol.J91-D, No.04, pp.1033-1041, 2008.
[10] J. Chiang and H. Yu. Extracting functional annotations of proteins based on hybrid text mining approaches. In Proceedings of BioCreative Workshop, 2004.
[11] J. Chiang and H. Yu. Meke:discovering the functions of gene products from biomedical literature via sentence alignment. Bioinformatics, (1911) :1417-1422, 2003.
[12] P. Ruch, Automatic assignment of biomedical categories: toward a generic approach Bioinfomatics 2006 22(6):658-664, 2005.