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

2008－MPS－72 (28)
2008－BIO－15 (28)
2008／12／18

相同遺伝子を考慮したGOアノテーションへの多階層分類の適用

木野 嘉祐† 関 和 広†† 上 原 邦 昭†

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．提案手法は，所与の遺伝子とその相同遺伝子との対応関係に基づき，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．

Gene Functional Annotation by Ortholog-based Hierarchical Classification

Yoshihiro Kino,† Kazuhiro Seki†† and Kuniaki Uehara†

This paper proposes a novel method for gene functional annotation in the framework of hierarchical classification that uses as constraints 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 annotated with each gene in model organism databases, such as FlyBase and MGI. These gene functions are described using Gene Ontology (GO), common vocabularies to enable uniform access to different model organisms databases. Our proposed approach exploits gene functions of orthologous gene as constraints, dynamically creating classifiers from training data available under the constraints. The effectiveness of the proposed approach is demonstrated in various experiments.

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．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）．

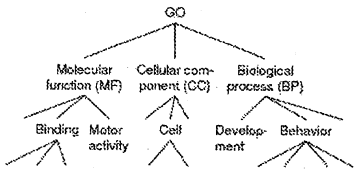

図1 Gene Ontologyの基本構造

図1のように，GOには第一階層に三つの遺伝子機能（GOドメイン）が存在し，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

† 神戸大学大学院工学研究科
Graduate School of Engineering, Kobe University
†† 神戸大学自然科学系先端融合研究環
Organization of Advanced Science and Technology, Kobe University

る．本研究では，遺伝子機能付与の対象となる所与の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を目指す．

2. 関連研究

遺伝子機能アノテーションは，その対象となる遺伝子数，遺伝子機能数の膨大さ，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の自動化を実現できれば，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 ☆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代表的手法，および関連研究に関して簡単に説明する．

- TREC Genomics トラック 2004
  Mouse の遺伝子を対象に，3 種類の GO ドメインを推定するタスクである．
  関とモスタファ[3] はカイ二乗統計値と複数の重み付け法を用いた kNN 分類器による推定手法を提案した．具体的には，論文中で言及されている個々の遺伝子が記述されたテキスト断片の抽出や，教師付き重み付けを行い，kNN 分類器によって GO ドメインの推定を行った．
- BioCreative
  2003 年に開催され，Human の遺伝子を対象に，全ての GO タームを推定するタスクである．
  Stoica と Hearst[2] 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，類似の機能を持つ可能性が高い．ここでは，Human の遺伝子の相同遺伝子として Mouse の遺伝子に付与されている遺伝子機能を制約として設け，その中で文字列のマッチングを適用した．一つ目のマッチング手法は，文献内に制約となる遺伝子機能（制約遺伝子機能）の記述の 75%以上の単語が存在した場合に当該遺伝子機能として付与する方法である（Cross Species Match（CSM））．二つ目は，すべての遺伝子機能の記述を文献内で探索し，完全一致した遺伝子機能と相同分子種に付与された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，閾値を越えた場合に該当遺伝子機能として付与する手法である（Cross Spicies Correlation（CSC））．さらに，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．

☆ http://biocreative.sourceforge.net/

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，Stoica らが利用した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収集し，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 ☆☆や Gene Ontology Annotation（GOA） ☆☆☆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 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，transcription factor activity，nucleus などの遺伝子機能が付与されているため，これらを Mouse の遺伝子 Sox21 に付与する遺伝子機能候補として分類を行う．

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．

3.3.1 データ抽出

生物医学分野では，様々な情報を保有したデータ

☆☆ http://www.informatics.jax.org/
☆☆☆ http://www.ebi.ac.uk/GOA/

```
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:         Feature Selection by Chi-Square.
10:        Make Classifier.
11:        add Classfied_GF to GO_Term.
12:      CommonGF - -;
13:  else if(RestrictGF == 1)
14:    add RestrictGF to GO_Term.
```

図 2 GO ターム推定

ベースが存在する．本研究では MGI，GOA，そして EntrezGene ☆の三つの生物データベースより，各生物種の詳細情報や生物種間のマッピング表などの様々な情報を抽出する．また，Gene Ontology（GO）データベース ☆☆に記載された GO の詳細情報より階層関係の取得を行う．GO タームの推定を行う際には，分類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．分類階層を探索するためには各遺伝子機能の階層・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多くの訓練事例を収集しなければ分類を行うことができない．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 2 に表す．

- 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
  各文献における制約遺伝子機能数は必ずしも同じであるとは限らないため，制約遺伝子機能数に応じた処理を行う必要がある．制約遺伝子機能が 1 つの場合，その制約遺伝子機能を当該遺伝子機能として断定し，付与し，2 個以上の場合は，分類可能と考えられ，分類器を作成するために必要な素性選択を行う．ここで，素性単位を単語として，各分類器特有の素性を訓練事例を用いることで動的に生成する．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その素性を基に分類を行う．

☆ http://www.ncbi.nlm.nih.gov/sites/gquery/
☆☆ http://www.geneontology.org/

表 1 GO ドメイン推定結果

	Precision	Recall	F-Score
TREC（BEST）	0.441	0.769	0.561
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. 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．

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

しかし，このデータには GO タームが付与されていないためデータとして不十分である．そこで，既存データベースより GO タームを抽出することで，GO ターム推定のための評価データ（604 件）として使用する．

```
12411446 MGI:2448704 Afmid MF TAS GO:0004061
```

4.2 多階層分類の評価実験

4.2.1 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 Precision，Recall，F 値を用いた．

4.2.2 GO ドメイン推定

関とモスタファ[3]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 1 に示す．表 1 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．

4.2.3 GO ターム推定

制約遺伝子機能を用いて多階層分類による GO ターム推定を行う．ここで，Rat の制約遺伝子機能を用いて訓練事例数の閾値による分類の比較実験を行う．その結果を図 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，ここで唯一の制約遺伝子機能のみを持つ文献の遺伝子機能の付与を行う．Human，Rat，Human と Rat の三種類の制約遺伝子機能に対して同様に付

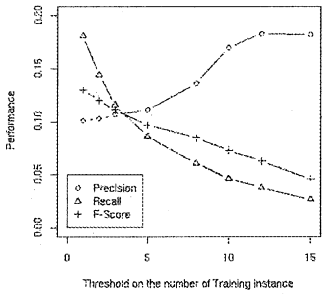

図 3 訓練事例数の閾値による実験

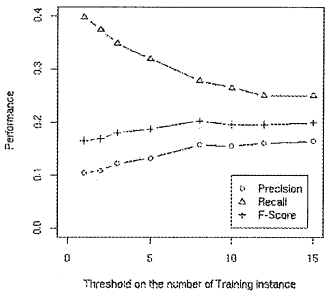

図 4 追加実験

表 2 相同分子種別比較実験

	Precision	Recall	F-Score
提案手法（Human）	0.158	0.184	0.170
提案手法（Rat）	0.115	0.283	0.164
提案手法（Human と Rat）	0.119	0.252	0.162

与し，評価実験を行う．

表 2 より，Human と Rat のどちらか一方の遺伝子機能を制約遺伝子機能とした方が良い結果を示すことがわかる．Human は Precision が高いのに対して，Rat は Recall が高いことが相対的に見てとれる．

4.3 追加実験

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図 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．これは訓練事例を持たない遺伝子機能が付与される可能性が十分に高いことを表している．そのため，訓練事例の閾値によって除去されてしまった遺伝子機能が存在し，評価値が低下したと考えられる．そこで，訓練事例に満たない遺伝子機能に対して新たな処理を加えた追加実験を行う．

追加実験方法は，Rat の制約遺伝子機能に対して，訓練事例が一つでも存在する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簡易的な処理を行う．ここでは，経験的に $n = 1$ とする．反対に一つも存在しない場合には，判断材料がないため，マッチング手法を組み込む．文献内に存在する遺伝子機能の記述を探索し，編集距離を用いることで制約遺伝子機能と完全一致した場合，付与することを決定する．

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+25%増しの結果となった．追加実験として，分類器を作成することができない未知の遺伝子機能への処理を追加することで，本来の分類では，見落とされてしまう遺伝子機能を網羅することができ，本研究の有用性を確かめることができたと考えられる．

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行った．まず，実験結果より，Mouse の相同分子種を制約遺伝子機能として用いることは有用であったと考えられる．また，Human と Rat の 2 種類の相同分子種はそれぞれ別の指標で高い結果を得ることができた．さらに，未知の遺伝子機能に対する処理を付加することでさらに評価値を向上させることができた．今後，提案手法を他のデータで利用することで，本研究の有効性を証明していくことを考えている．

参考文献

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) E. Stoica and M. Hearst, Predicting Gene Functions from Text Using a Cross-Species Approach, Pacific Symposium on Biocomputing 11:88-99, 2006.
3) 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.