2ZB-7

モデル植物購買履歴に基づく遺伝子選択支援システム

佐藤 貴命* 賀屋 秀隆§ 松井 藤五郎† 朽津 和幸‡ 大和田 勇人†

東京理科大学理工学研究科経営工学専攻* 同 理工学部経営工学科†
同 応用生物科学科‡ 同 ゲノム創薬研究センター§

1 はじめに

生物学の分野では，遺伝子機能やタンパク質の構造および機能を解析する研究が行われている．遺伝子解析に関する研究をしている生物学者にとって，膨大な量の遺伝子の中からどの遺伝子に注目するかは重要な関心事であり，人手により関連する論文を見つけ，対象となる遺伝子を選び出す方法が一般的である．

だが，遺伝子解析の実験では，人為的操作によるモデル生物の変化を観察・観測するため，実験に長い期間を要する．つまり，発表される論文はかなり昔に研究が始められた場合が多い．このため論文から得られる情報では，今行われている研究で注目されている関連遺伝子を知ることは難しい．

そこで，我々が提案した変異体選択支援システム [1] を改良し，興味がある遺伝子に関連している遺伝子を購買履歴から調べる方法を考える．

本論文では，購買履歴から異なる遺伝子同士の関連性を示す相関ルール [2] を導出することによって，関連遺伝子を発見する方法を提案する．さらに，本研究で構築した類似遺伝子発見システムについて述べる．また，導出された相関ルールに関して，公開されている配列データベースを使用し，配列の類似性を用いることで生物学的な尺度に基づく評価を行う．生物学者が本システムを利用することで，関連遺伝子を探し出し，さらには論文や既存のバイオインフォマティクスの手法では見つけ出すことができなかった関連遺伝子を発見できると期待される．

2 提案手法

本研究で用いるシロイヌナズナ変異体の購買履歴に関して説明し，購買履歴から遺伝子同士の関連性を示す相関ルールを導出する方法について述べる．

Takanori SATOU*, Hidetaka KAYA§, Tohgoroh MATSUI†, Kazuyuki KUCHITSU‡,and Hayato OHWADA†
Department of Industrial Administration, Graduate school of Science and Technology, Tokyo University of Science*, Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science† Department of Applied Biological Science, Faculty of Science and Technology, ‡ Genome and Drug Research Center, ‡

2.1 購買履歴の説明と前処理

本研究では，TAIR* で公開されているシロイヌナズナ変異体の購買履歴から，SALK T-DNA とよばれる変異体の購買履歴を用いる．

公開されている SALK T-DNA の購買履歴から，遺伝子同士の関連性を調べるために必要となる属性を以下に示す．

1. シロイヌナズナ遺伝子番号
   - At1g29050 のような記号で表され, At は *Arabidopsis thaliana* を表す．
   - T-DNA の挿入位置を示す．
2. 遺伝子の説明 (description)
   - 生物学上の遺伝子の性質に関する説明.
3. SALK T-DNA 番号
   - SALK_039515 のような記号で表される．
4. 注文者
   - SALK T-DNA を注文した人名を表す.
5. 注文者 ID
   - 注文者が同姓同名の場合の識別番号を表す.
6. 研究室
   - 注文者が所属している研究室名を表す.
7. 研究室 ID
   - 同じ名前の研究室の場合の識別番号を表す.
8. 注文日
   - SALK T-DNA が注文された日付を表す.

前処理として，自動的に TAIR に問い合わせを行うための Java プログラムを作成し，以上の 8 種類の属性が全て揃っているデータのみ抽出した．各属性をタブ区切りでテキスト形式のデータにまとめた．

2.2 相関ルールの導出

2.1 で作成したデータを，MUSASHI[3] で扱うことができるように変換する．ここで，関連遺伝子を発見するのが目的であるので，アイテムが複数ある場合は考えず，アイテム数を 2 と設定する．遺伝子番号をアイテム (I) とし，購入された変異体の遺伝子番号を注文者ごとに集めてトランザクション (T) とした．ここで，変異体が同一遺伝子上にある場合同じものとして扱い，また

* The Arabidopsis Information Resource. http://www.arabidopsis.org/

バスケット分析とは異なり，注文日は考慮しない．変換したテキスト形式のデータから MUSASHI を用いて相関ルールを抽出するために，次の 2 つのコマンドを使用する．

1. テキスト形式を XMLtable 形式に変換
   `$ txt2xt -F -i (テキスト形式データ) -o (XMLtable 形式データ)`
2. XMLtable 形式のデータから相関ルールの生成
   `$ xtasrule -k (注文者 ID) -f (遺伝子番号)-M (2) -S (最小支持度) -C (最小確信度) -i (XMLtable 形式データ) -o (相関ルール)`

ここで計算時間と相関ルール数のトレードオフを考慮した上で，できる限り多くの相関ルールを抽出するために，データの量に合わせて最小支持度，最小確信度を決定する必要がある．導出された相関ルールは，支持度 (support)，確信度 (confidence)，リフト値 (lift) の 3 つの指標を含んだ形で得られる．

3 提案システム

本研究で構築した，関連する遺伝子を表示するための遺伝子選択支援システムについて説明する．Web ブラウザを利用した検索システムを構築するために，2.1，2.2 でそれぞれ作成したデータをデータベースに格納しておく．ユーザが調べたい遺伝子番号 X のリクエストを送ると，(X,Y) もしくは (Y,X) の組み合わせと，相関ルールの指標として支持度，確信度，リフト値が表示される．この時表示された Y が関連する遺伝子を表している．システムの詳細については，変異体選択支援システムとほぼ同じであるため，本論文では説明を省略させていただく．

4 実験と考察

公開されているデータベースに対して 152,978 個の SALK T-DNA について問い合わせを行い，その結果 131,223 件の購買履歴を得ることができた．続いて，最小支持度を 0.001 として，MUSASHI を用いて購買履歴から相関ルールを抽出し，123,996 個の相関ルールが生成された．

これらのデータをもとに，実験を行った．遺伝子番号 AT1G55020 について問い合わせをした結果，AT1G55020 に関する相関ルールは 172 個抽出された．ここで例として，検索結果の上位 3 件を図 1 に示し，AT1G55020 の関連遺伝子 (AT4G38770,AT1G67560,AT1G17420) について，生物学的な尺度に基づく評価を行う．

Arabidopsis thaliana Order History Data Mining System

Search Type [Association Rule] : AT1G55020

Description
[lipoxygenase (LOX1), identical to SP|Q06327]

◆Number : 172 ◆AT1G55020[G]

Antecedent [Cnt]	Consequent [Cnt]	Support	Confidence	Lift
AT4G38770 R [28]	AT1G55020 R [100]	0.03615328995	357 (%)	123.4821429
AT1G67560 R [23]	AT1G55020 R [64]	0.02313810557	278 (%)	96.20869565
AT1G17420 R [66]	AT1G55020 R [172]	0.06218365871	260 (%)	90.10454545

図 1　検索結果

ClustalW†を用いて各遺伝子のタンパク質の組み合わせに対して配列アラインメントを行った結果，次のようになった．

1. AT1G55020(LOX1), AT4G38770(PRP4)
   ⇒ Alignment Score: 1
2. AT1G55020(LOX1), AT1G67560(AT1G67560.1)
   ⇒ Alignment Score: 2152
3. AT1G55020(LOX1), AT1G17420(AT1G17420.1)
   ⇒ Alignment Score: 2263

配列アラインメントの結果から，1 の組み合わせが 2 や 3 と比較して類似性が低いことがわかる．このことから，2 や 3 の組み合わせ，すなわち AT1G67560 や AT1G17420 は，AT1G55020 と生物学的にもよく類似している遺伝子であることが示された．

5 まとめ

本論文では，購買履歴から異なる遺伝子同士の関連性を示す相関ルールを導出することによって，関連遺伝子を発見する方法を提案し，構築した類似遺伝子発見システムについて述べた．また，導出された相関ルールに関して，公開されている配列データベースを使用し，生物学的な尺度に基づく評価を行った．

生物学者が本システムを利用することで，興味のある遺伝子の関連遺伝子を発見することができる．

参考文献

[1] 佐藤 貴命，賀屋 秀隆，松井 藤五郎，朽津 和幸，大和田 勇人: モデル植物の購買履歴からの変異体選択支援システム．*FIT* 2006, pp433-434, 2006.

[2] R. Agrawal, T. Imielinski, and A.N. Swami.: ”Mining association rules between sets of items in large databases”, pp.207-216, Proc. SIGMOD, ACM (1993).

[3] Yukinobu Hamuro, Naoki Katoh, Katsutoshi Yada, Takashi Washio.: 大規模ビジネスデータからの知識発見システム: MUSASHI, 人工知能学会誌 20 巻 1 号, pp59-66 (2005).

† EMBL. http://www.embl.org/