H_020

モデル植物の購買履歴からの変異体選択支援システム

Mutant selection support system from purchase history of model plant

佐藤 貴命* 賀屋 秀隆§ 松井 藤五郎† 朽津 和幸‡§ 大和田 勇人†

1. はじめに

近年，生物学分野の研究において，モデル植物としてシロイヌナズナ(*Arabidopsis thaliana*) が広く使われている．生物学研究の一例として，プログラム細胞死（細胞が自発的に死ぬ現象）の解明を行なう研究があり，シロイヌナズナ変異体（遺伝子が一部異なるもの）の種子を何種類かストックセンターから取り寄せ，育ててその違いを研究している．代表的なストックセンターの一つである ABRC[1] から変異体を注文する場合，その変異体や注文者に関する情報が購買履歴として残され TAIR[2] にて公開されている．

しかしながら，情報は変異体ごとに個別に公開されており，変異体同士の関連性を調べることはできない．このような関連性は，データマイニングの手法を活用すれば発見できると考えられる．

そこで本研究では，シロイヌナズナの購買履歴に着目し，データマイニングを用いて異なる変異体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提案する．さらに，本研究で構築した変異体選択支援システムについて述べる．本システムは，導出された相関ルールに基づいて関連のある変異体をユーザに推薦する．生物学研究者が本システムを利用することで，変異体を購入する際の変異体選択支援につながることが期待される．

2. 相関ルール

相関ルールと，実際に相関ルールを導出するのに使用するデータマイニングツール MUSASHI について説明する．

2.1 相関ルール

相関ルール(Association Rule) の定義は次のようになる．I = $\{i_1, i_2, \cdots, i_m\}$ をアイテム全体の集合とする（m はアイテム数を表す）．また D をトランザクションデータベースとし，D 中の各トランザクション T は T ⊆ I となるアイテム集合である．

相関ルールとは X ⇒ Y (X ⊆ I, Y ⊆ I, X∩Y = $\emptyset$) という形で記述される関係である．X を含む D 内の c% のトランザクションが Y も含むとき，相関ルール X ⇒ Y は c の確信度(confidence) を持つ．D 内の s% のトランザクションが X ∪ Y を含むとき，相関ルール X ⇒ Y は s の支持度(support) を持つ．

2.2 MUSASHI

MUSASHI (Mining Utilities and System Architecture for Scalable processing of HIstorical data)[3] は，オープンソースとして開発されている知識発見システムの名称である．

MUSASHI は基本データ構造は XMLtable と呼ばれる XML による表形式のデータ構造を採用しており，大規模データを効率的かつ柔軟に処理できる．またデータ処理のためのプログラムとして，単一の機能に特化した小さなコマンド群が提供されており，相関ルールの生成を行なうコマンドもある．

3. 提案手法

本研究で用いるシロイヌナズナの購買履歴情報についての説明や相関ルールの導出方法，そして相関ルールをブラウザで表示するシステムについて説明する．図 1 にシステムの概要図を示す．以下でそれぞれの処理について説明する．

3.1 データの説明と前処理

シロイヌナズナ変異体にはいくつか種類があるが，遺伝子転写を活性化する領域をもつ T-DNA(自身の持つ DNA 内の特定の部分) を，植物ゲノム中に無作為に挿入した T-DNA タグラインと呼ばれる種類の変異体があり，全ゲノムの体系的な機能解析を視野に入れ，数十万の T-DNA タグラインが作成されその解析が進められている．

TAIR で公開されているシロイヌナズナ変異体の種類には SALK, SAIL, GABI, FLAG, EXOTIC, Wisc, RIKEN, CSHL 等の T-DNA タグラインがあり，そのうちの一つである SALK T-DNA の変異体のデータを用いることにする．

SALK T-DNA の公開データからは，変異体に関する生物学的な情報や購買履歴情報を個別に調べることができるが，その中から変異体同士の関連性を調べる際に特に必要となる次のデータを利用する．

[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 がシロイヌナズナ遺伝子のどの位置に挿入されたかがわかる．

[SALK T-DNA 番号] SALK_039515 のような記号で表される．相関ルール導出に用いられる．

[注文者] T-DNA の種子を注文した人名を表す．

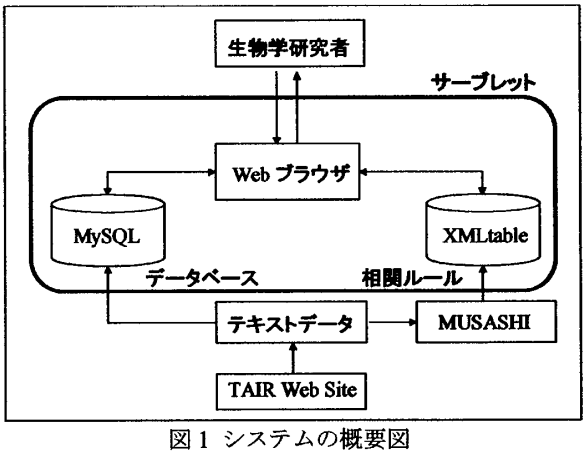

図 1 システムの概要図

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

[注文者 ID] 同姓同名の場合の区別を行なうための番号を表す. 相関ルール導出に用いられる.

[研究室] 注文者の所属している研究室名を表す.

[研究室 ID] 同じ名前の研究室の場合の区別を行なうための番号を表す.

[注文日] T-DNA の種子が注文された日付を表す.

前処理として以上の 7 種類のデータを TAIR に問い合わせ, 抽出し, 各属性をタブ区切りでテキストデータにまとめた.

3.2 相関ルールの導出

MUSASHI を用いて, 前処理されたテキストデータから「この変異体を買った人は他にもこんな変異体を買っている」という相関ルールを導出する. そのため, SALK T-DNA 番号をアイテム(I) とし, 購入された種子の SALK T-DNA 番号を研究者ごとに集めてトランザクション(T) とした. テキストデータを txt2xt コマンドを用いて XML 形式に変換し, xtasrule コマンドを用いて相関ルールを生成する.

3.3 相関ルールのブラウザへの表示

Web ブラウザを利用した検索システムを構築するために, 前処理されたテキストデータをデータベースに格納しておく.

ユーザがある SALK T-DNA についてリクエストを送ると, 変異体の履歴情報と詳細情報, 個人購買履歴, 相関ルールへのリンクを表示する. リンクをクリックするとそれぞれの情報が表示される. システムの構築にあたり, データベースには MySQL[4] を, Web サーバには Tomcat[5] を使用し, Java Servlet を用いて検索システムを実装した.

4. 実験

本研究では, SALK T-DNA のうち 10,000 個の変異体について購買履歴情報を抽出し, 結果 13,451 件の履歴情報を得ることができた. 続いてアイテム集合の数を 2, 最小支持度を 0.0001 として相関ルールを生成した. この結果, 13,451 件の履歴情報から 1,312,849 個の相関ルールが生成された.

これらのデータを使用し, 実験を行なった. 具体的に変異体 SALK_128569 について検索システムを実行した.

図 2 は検索を行った結果である. SALK T-DNA の購買履歴を表示し, 注文者や研究室の詳細情報, 研究者別の購買履歴, そして相関ルールへのリンクが表示される. そのうち詳細情報は直接 TAIR に問い合わせを行なう.

図 3 に相関ルールの情報を表示する. SALK_128569 に関する相関ルールは 14 個あり, その中で最もよい評価を示したルールは支持度が約 0.00138 , 確信度 80%であった. そのルールを次に示す.

SALK_128571 ⇒ SALK_128569

このルールは「SALK_128571 の変異体を買った人の 80% が SALK_128569 の変異体も買っている」ということを示しており, この 2 つの SALK T-DNA には関連性があると考えられる.

5. 考察

生物学研究者に実際に本システムを利用してもらった結果, 生物学の観点からみると, 導出された関連性は興味深く, 例えば, 今後の研究の方向性を決定するのに役立つ可能性があることがわかった. また, 現在は変異体(mutant)同士の関連性を抽出しているが, 遺伝子(gene)同士の関連性を抽出

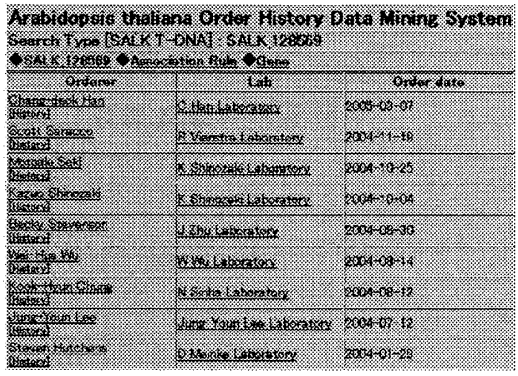

図 2 検索結果

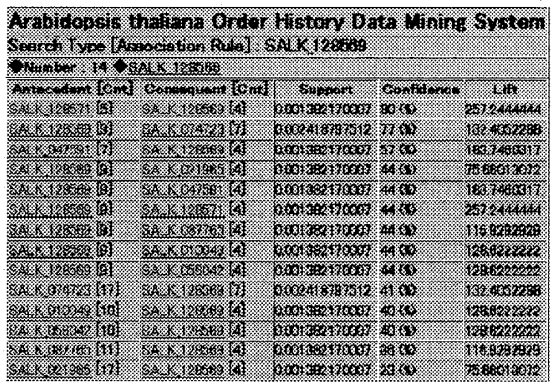

図 3 相関ルール

すれば, さらなる有用な情報を見つけ出すことができるとわかった. 今後は相関ルールが得られた遺伝子の相同性やアミノ酸配列の類似性を調べ, 本システムで生物学的に意味のある関連性が得られるかどうかを検証したい.

6. まとめ

本論文では TAIR で公開されているシロイヌナズナの変異体の購買履歴から変異体同士の関連性を表す相関ルールを導出する方法を提案し, その相関ルールに基づいて関連ある変異体をユーザに推薦する変異体選択支援システムについて述べた.

生物学研究者が本システムを利用することで, 変異体を購入する際の変異体選択支援につながることが期待される.

参考文献

[1] ABRC : The Arabidopsis Biological Resource Center

[2] TAIR : The Arabidopsis Information Resource http://www.arabidopsis.org/

[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 年 1 月)

[4] MySQL, http://www.mysql.com/

[5] Tomcat, http://jakarta.apache.org/