MARE: 遺伝子発現量検索エンジンの構築に関する一考察

梅澤香矢乃† 瀬々 潤†

† お茶の水女子大学理学部情報科学科 〒 112–8610 東京都文京区大塚 2–1–1
E-mail: umezawa@sel.is.ocha.ac.jp, sesejun@sel.is.ocha.ac.jp

あらまし 次世代の医療として、個々の特性に応じた診療をするテーラーメード医療が注目される。テーラーメード医療では、患者から採取した遺伝子情報と診断、治療、予後情報をデータベース化し利用することが求められる。その実現に向け、患者から採取した遺伝子情報が過去のどの患者の遺伝子情報に類似するかを検索する遺伝子検索エンジンの実装を目指した。しかし、検索対象とする患者の数は将来膨大になることが見込まれ、さらに患者一人あたり約 3 万もの遺伝子を持っている。このような大規模で高次元のデータから、診断に必要な部分を的確に取り出すことは容易ではない。よって本発表では、ヒトより情報の少ない酵母の遺伝子情報を用いた検索エンジンを作成したので紹介する。さらに、本実装に於いて生まれた問題点及び、ヒト遺伝子への拡張において乗り越えるべき点を提示する。
キーワード 検索、推薦、遺伝子

MARE: A Prototype for Developing a Gene Expression Profile Search Engine

Kayano UMEZAWA† and Jun SESE†

† Dept. of Computer Science, Ochanomizu Univ. 2–1–1 Otsuka, Bunkyo-ku, Tokyo, 112–8610 Japan
E-mail: umezawa@sel.is.ocha.ac.jp, sesejun@sel.is.ocha.ac.jp

Abstract In the post-genomic era, the selection of medical treatments using patients' genomic information, called tailor-made medicine, is one of the important problems. Recent bio-technological advances such as DNA microarrays allow us to generate databases about gene expression profiles observed from patients. Our aim in this study is to develop a search engine for the expressions because we will make an accurate prognosis from patients having similar expressions. However, since the database would contain enormous data, we here develop a search engine for yeast gene expressions using Ruby on Rails, and clarify the difficulties to develop the human patients' search engine.
Key words search, recommendation, gene expression

1. はじめに

ヒトのゲノム配列が利用可能となった現在、ゲノムから得られる情報を利用した医療としてテーラーメード医療が期待されている。テーラーメード医療とは、個々の患者の体質に合った診断や治療をする、従来の医療よりも個人差を考慮した医療のことである。このテーラーメード医療では様々な患者から採取した遺伝子情報、及び診断、治療、予後の情報をデータベース化し、個々の特性に応じた病状診断と療法の選択をすることが目標となる。次世代のテーラーメード医療の流れを図示したのが、図 1 である。

個々の特性に応じた病状診断や治療をする際には、遺伝子情報が似ている患者の診断、治療、予後の情報を活用する。たとえば、ある薬に対する副作用が出やすい患者 A がいたとする。医師が別の患者 B を診療する際に、患者 B が患者 A と似た遺伝子情報を持っているということが分かれば、その薬に対する副作用が出やすいのではないかと予測できる。このように、テーラーメード医療では、遺伝子情報が似た他の患者の情報をデータベースから探し出して利用することが必要になる。

さらに近年の技術の進歩により、DNA マイクロアレイを用いて網羅的に遺伝子の発現量を採取することが容易になった。遺伝子発現量とは各遺伝子の使われ度合いを示している物であり、現在病院で血液から検査する情報をより細かく見られるものだと考えられる。現在では 20 万サンプルを超える遺伝子発現量が蓄積され、公開されている [1]。本研究では医師が患者から採取した遺伝子発現量を入力すると、過去のどの患者の遺伝子情報に近いかを検索する遺伝子検索エンジンの実装を行うことで、医師の診断を手助けするソフトウエアの構築を目標にした。

本研究が想定しているユーザである医師は必ずしも計算機の

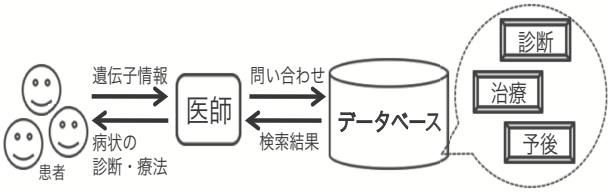

図 1　テーラーメード医療の流れ

扱いに慣れていない可能性があることを考慮し、データの入力や結果の閲覧を容易に行えるように web アプリケーションとして開発する。これによりインターネットさえ確保できれば、誰もがどこからでもアクセスできるため、一人の医者や一病院内だけで保持する患者の情報だけでなく、世界中の患者の情報を集め利用することが可能になる。更に、世界の何処の場所においても均一な診断が出来ること、最新の情報を用いて診断できる事など、多くの利点がある。

また、検索対象とする患者の数は膨大になることが見込まれるため、大規模なデータからの検索が求められている。更に、ヒトの遺伝子は約 3 万あるため、各患者の情報は超高次元データとなる。高次元空間では次元の呪いにより、各データ間の距離がほぼ等しく見える現象が知られており、検索精度が低いことが予想される。診断において検索精度が低いことは致命的な欠陥となるので、この点も考慮する必要がある。

そこで本研究では、まずヒトより遺伝子数が少ない酵母の遺伝子情報を用いて遺伝子発現検索エンジンのプロトタイプの作成を行うことにした。酵母は発酵に用いられる工業的に重要な種であり、現在でも活発に研究対象とされているので、実装意義がある。本実装では、マイクロアレイで得られた酵母の遺伝子発現量情報を入力し、過去のどの実験状況に近いかを検索する。

本研究で構築した web アプリケーション Micro Array Retrieval Environment (MARE) の概略を図 2 に示す。MARE は様々な環境における酵母の遺伝子発現量をデータベース化している。ユーザはブラウザを通じて遺伝子と発現量を MARE に入れることで、類似した過去の実験を検索することができる。

前述したヒト遺伝子検索エンジンの場合は、保持するデータベースは様々な患者の遺伝子発現量、ユーザが入力するのは患者の遺伝子情報、ユーザに提示する情報は類似する患者の情報となる。このように扱うデータが異なっていても検索エンジンの仕組みは変わらない。

本稿では、まず 2 章で関連研究を示し、3 章では実験状況同士の類似度の計算手法について述べる。また 4 章では実際のデータを用いた実行例を示し、5 章でまとめを行う。

2. 関 連 研 究

マイクロアレイ解析用ソフトウエアはデスクトップアプリケーションとして行われている物と、Web ブラウザを通じて利用する物がある。本研究は、Web ブラウザを通じて利用するものを構築する。マイクロアレイの解析をする Web アプリケーションとしては、DAVID [3] が有名である。これは、マイクロアレイから得た高発現、低発現遺伝子リストをサーバ上に蓄積することで、関連する遺伝子機能や遺伝子ネットワークを検索してくれる物である。関連する知識は、広く公開されているものを利用しているが、各研究室内で取得したマイクロアレイ同士しか比較が出来ない。医療への応用を考えた場合、他の医療機関で採取された物も比較することが必要であり、DAVID では応用が難しい。

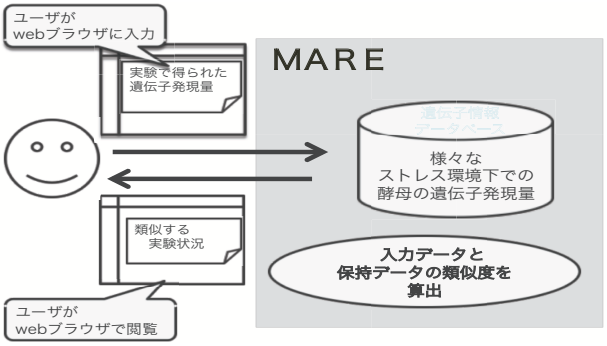

図 2　本研究の流れ

マイクロアレイの検索を行う研究として CellMontage [4] がある。この研究では、対象となるマイクロアレイの数は、最大でも約 15,000 であり、多くの場合は数十の実験に限定した上で検索をする。また、検索に 20 秒程度の検索時間を要しており、今後のマイクロアレイの増加に対応した高速な検索エンジンが必要とされる。さらに、ノンパラメトリックな順位和検定を用いているため、検出感度が低いと考えられる。本研究では、Pearson の相関係数を利用することで、感度高い検索を目指す。

3. 手　　法

MARE の内部では、様々な実験状況下における多数の酵母の遺伝子の発現量をデータベースとして保持している。ユーザは、実験でマイクロアレイから得られた酵母の遺伝子発現量情報を、web ブラウザから入力することができる。そして、MARE は入力データとデータベース内の各実験状況下における遺伝子の発現量と比較し、ユーザの行った実験状況に類似している実験状況を調べてユーザに表示する。

3.1　データベースから類似する実験状況を算出

3.1.1　Pearson の相関係数

ユーザが実験で得て、MARE に入力した酵母の遺伝子名とその発現量を入力データとし、入力データともっとも類似度の高いデータベース内の実験状況をユーザに返す。ユーザが入力した実験状況と MARE の保持するデータベース内の各状況の類似度は、相関を調べるためによく用いられている Pearson の相関係数で計算する。Pearson の相関係数では、2 つのベクトル $\boldsymbol{x}_i$ と $\boldsymbol{x}_j$ が $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in}), \boldsymbol{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jn})$ として与えられた場合に、$\boldsymbol{x}_i$ と $\boldsymbol{x}_j$ の相関係数は、以下の式で定義される。

$$\rho(i,j) = \frac{\sum_{k=1}^{n}(x_{ik}-\bar{x_i})(x_{jk}-\bar{x_j})}{\sqrt{\sum_{k=1}^{n}(x_{ik}-\bar{x_i})^2}\sqrt{\sum_{k=1}^{n}(x_{jk}-\bar{x_j})^2}}$$

表 1　保持しているデータベース

実験状況	遺伝子 a	遺伝子 b	遺伝子 c
1	3.68	1.22	−1.23
2	1.75	3.72	−0.25
3	−3.28	0.41	−0.21

ここで $\bar{x}_i$ は $\boldsymbol{x}_i$ の平均値を表し、$\frac{1}{n}\sum_{k=1}^{n} x_{ik}$ である。

相関係数はつねに $-1 \leqq \rho(i,j) \leqq 1$ をとる。相関係数は 0 に近いと相関が小さいことを表す。1 に近い値をとると正の相関が高く、-1 に近い値をとると負の相関が高い。今回は入力データとのデータベース内の実験状況との相関係数を求め、正の相関が高い実験状況を、入力データとの類似度が高い状況であると考える。

3.1.2　算 出 方 法

入力した実験状況のデータとデータベース内の実験状況の n 個の遺伝子における発現量から Pearson の相関係数が計算できる。

入力データと、データベース内の実験状況 i での遺伝子発現量が、$\boldsymbol{x}_0 = (x_{01}, x_{02}, \ldots, x_{0n}), \boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ として与えられた場合を考える。x_{0k} は入力した実験状況のデータの k 番目の遺伝子の発現量、x_{ik} はデータベース内の実験状況 i の k 番目の遺伝子の発現量を表す。$\boldsymbol{x}_0$ と $\boldsymbol{x}_i$ の Pearson の相関係数は以下のようになる。

$$\rho(0,i) = \frac{\sum_{k=1}^{n}(x_{0k}-\bar{x_0})(x_{ik}-\bar{x_i})}{\sqrt{\sum_{k=1}^{n}(x_{0k}-\bar{x_0})^2}\sqrt{\sum_{k=1}^{n}(x_{ik}-\bar{x_i})^2}}$$

また、本来 Pearson の相関係数の値を求める際には、データベース内の各実験状況に対して上記の式を独立に求める必要がある。しかし、本手法では $\sqrt{\sum_{k=1}^{n}(x_{0k}-\bar{x_0})^2}$ の値がデータベースに寄らず入力データのみに依存するため、一度計算することで、何度も使い回すことが出来、高速化が計れる。

以上の計算を、データベース内の各実験状況について行う。

このように、入力データとデータベース内すべての実験状況との類似度を求め、類似度の高い順にデータベース内の実験状況を並べる。その中で類似度の順位が高い実験状況をユーザに提示する。

具体的に計算してみよう。ユーザが実験で得られた遺伝子発現量が、遺伝子 a、遺伝子 b、遺伝子 c においてそれぞれ $3.01, 0.98, 0.37$ であるとし、表 1 のデータベース内の各実験状況との類似度を計算する。

入力した実験状況のデータとデータベース内の実験状況 $1,2,3$ の 3 個の遺伝子 a,b,c における発現量をベクトル $\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ とすると、$\boldsymbol{x}_0 = (3.01, 0.98, 0.37), \boldsymbol{x}_1 = (3.68, 1.22, -1.23), \boldsymbol{x}_2 = (1.75, 3.72, -0.25), \boldsymbol{x}_3 = (-3.28, 0.41, -0.21)$ である。$\boldsymbol{x}_0$ と $\boldsymbol{x}_1$ との相関係数を計算すると $\rho(0,1) = 0.95$ となる。同様に $\boldsymbol{x}_0$ と、$\boldsymbol{x}_2, \boldsymbol{x}_3$ との相関係数を計算すると $\rho(0,2) = 0.22, \rho(0,3) = -0.93$ となる。相関係数の値は、$\rho(0,1), \rho(0,2), \rho(0,3)$ の順に大きい。よって、この場合は類似度が高い実況状況として $1,2,3$ の順にユーザに提示する。

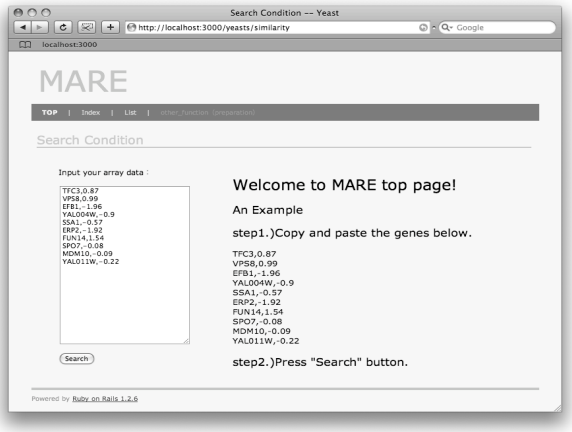

図 3　MARE の Top ページ

4. 実　行　例

この章では実際の MARE の利用を紹介し、DNA マイクロアレイから得られた遺伝子名とその発現量のデータから類似する実験情報を求めるツールとして有用であることを示す。

4.1　MARE のシステム

MARE は現在、酵母の 6152 遺伝子に関し、173 通りの様々なストレス環境下で発現量を観測した値をデータベースに保持している。[2]

システムの構築には Ruby on Rails 1.2.6、MySQL5.0.51 を用いて、MacOS X 上で実行している。

MARE の検索ページに、ユーザが類似する実験状況を知りたい遺伝子発現量のデータを入力すると、MARE は類似する実験状況を 10 位ずつ結果ページに表示する。以下に MARE の具体的な検索手順を説明していく。MARE では、図 3 の Top ページで遺伝子とその発現量の入力をするだけで、類似する実験状況を調べることができる。

4.2　実行例１ー熱ショック応答

MARE の保持するデータベース内の実験の中から、酵母に熱ショックを与えた後 10 分経過後のサンプルである、Heat Shock 10 minutes hs-1 から 50 個の遺伝子をランダムに選ぶ。この遺伝子発現量のデータを、仮に実験で得られたデータとして MARE に入力する。MARE は約１ 秒ほどの検索を行い、出力結果のページは図４のようになる。もっとも似ている実験状況として、熱ショック後 10 分の状況を 1 位に表示する。つまり、仮定した実験状況そのものが、類似度１ 位として表示されている。さらに、この入力結果の場合は 2 位には熱ショック後 15 分、3 位には熱ショック後 30 分、4 位には熱ショック後 20 分そして 9 位には熱ショック後 5 分の状況を表示している。

このように少ない入力データでも、MARE は適切な結果を示すことが分かり、提案手法が有効であると言える。

また、入力データと各実験状況のデータがどの程度の類似しているのかをを示すために、Pearson の相関係数の値と、その値に応じた棒グラフを表示している。このように、結果を表示する際にはユーザビリティにも配慮している。

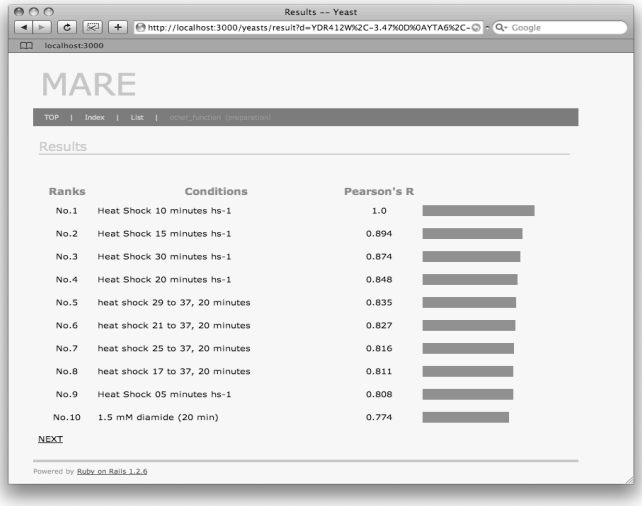

図 4　実行例 1 の結果

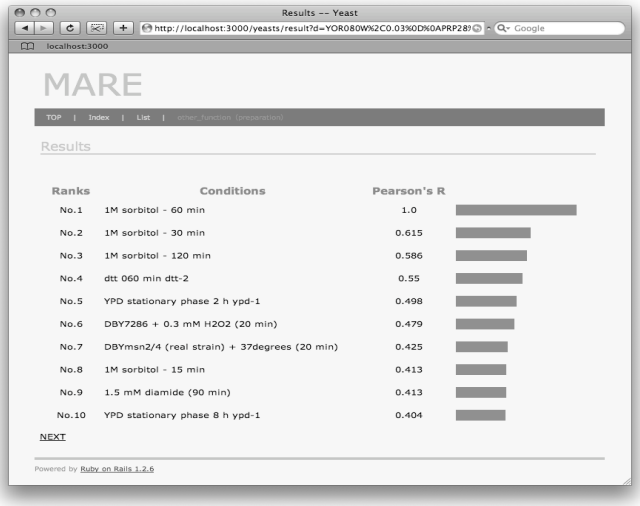

図 5　実行例 2 の結果

4.3　実行例２ー糖代謝

MARE の有用性を示すため、異なったデータを用いて検索しよう。MARE の保持するデータベース内の実験の中から、酵母に糖を与えた後 60 分経過後ののサンプルである、1M sorbitol-60 min から 500 個の遺伝子をランダムに選ぶ。この遺伝子発現量のデータを、仮に実験で得られたデータとして MARE に入力する。MARE は約 10 秒ほどの検索を行い、出力結果のページは図 5 のようになる。もっとも似ている実験状況として、糖投与後 10 分の状況を 1 位に表示する。つまり、仮定した実験状況そのものが、類似度 1 位として表示されている。さらに、この入力結果の場合は 2 位には糖投与後 30 分経過後、3 位には糖投与後 120 分経過後、8 位には糖投与後 15 分経過後の状況を表示している。

本実行例のように、入力データが増えても MARE は適切な結果を示すことが分かり、提案する計算手法が有効であると言える。

4.4　実行例３

最後に、より多くの遺伝子発現量のデータを入力した場合を示す。実行例１、２と同様の方法で、ランダムに 3000 個の遺伝子を選んで入力する。結果は変わらないが、検索時間に数十秒かかってしまう。また 6,152 個のすべての遺伝子発現量を入力した場合、1 分ほど待っても結果は得られなかった。

より膨大なヒト遺伝子をデータベースとして保持するシステムへ拡張していくためには、さらに改良を考えていく必要がある。

5.　まとめと今後の課題

本稿では、ユーザがマイクロアレイで採取した酵母の遺伝子発現量を入力すると、様々な実験状況下における発現量のデータの比較を自動で行い、ユーザの行った実験状況に類似した実験状況を表示するアプリケーション MARE を提案した。

今後の課題としては、はじめにで挙げたように膨大なヒトの遺伝子情報に応用させていきたい。

実行例より、比較的少ない入力データであっても、適切な結果が得られることが分かった。しかし、酵母は単細胞の比較的単純な生物なので、高等生物であるヒト遺伝子を扱う際にも、同様に適切な結果が得られるかは、今後の課題である。少ない入力数で適切な結果が得られない場合、入力数を多くする必要があるが、入力数が増えると計算時間が遅くなる。

さらに、保持しているデータベースも、酵母遺伝子とヒト遺伝子では規模が異なる。より膨大なヒト遺伝子のデータベースを扱う際には、高精度の検索をするための適切な指標の選択や、算出速度をより速めるための計算方法や、データベースからの検索が必要である。

さらに、表示方法を工夫しユーザビリティを高める工夫も行っていきたい。

文　　献

[1] Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R., “NCBI GEO: mining millions of expression profiles–database and tools”, *Nucleic Acids Research*, vol. 33, Database issue,D562–D566, 2005.

[2] Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO.,“Genomic expression programs in the response of yeast cells to environmental changes”, *Molecular Biology of Cell*,vol.11,no.12,pp.4241-4257,2000.

[3] Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA., “DAVID: Database for Annotation, Visualization, and Integrated Discovery”, *Genome Biology*,vol. 4, issue 9, R60, 2003.

[4] Fujibuchi W, Kiseleva L, Taniguchi T, Harada H, Horton P., “CellMontage: similar expression profile search server”, *Bioinformatics*, vol. 23, issue 22, 3103–3104, 2007.