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　(指導教員：瀬々 潤)

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て個々の患者の体質に合った診断や治療をするテーラーメード医療が期待されている。テーラーメー医療では、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、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示したのが、図 1 である。

近年の技術の進歩により、網羅的な遺伝子発現量の採取が容易になった。遺伝子発現量とは各遺伝子の使われている度合いを示している物で、病院で採取した血液から得られる情報をより細かく見られるものだと考えられる。現在 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。

想定ユーザである医師が計算機の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、世界中の患者の情報を集め利用できること、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ことなども可能になる。

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、大規模な患者数の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、検索精度が低い事が予想される。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。

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量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、現在でも活発に研究対象とされており、実装意義がある。

実装した web アプリケーション Micro Array Retrieval Environment(MARE) は様々なストレス環境下での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、マイクロアレイで得られた酵母の遺伝子と発現量を MARE に入力し、類似した過去の実験を検索することができる。

2　関連研究

マイクロアレイの解析をする Web アプリケーション研究 [2] が行われている。しかし、これらのサーバは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。また、マイクロアレイの検索を行う研究として CellMontage [3] がある。この研究では、対象となるマイクロアレイの数を、数十の実験に限定した上で検索をする（最大では約 15,000 の実験から検索できる）。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。

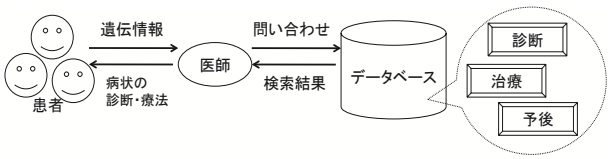

図 1: テーラーメード医療の流れ

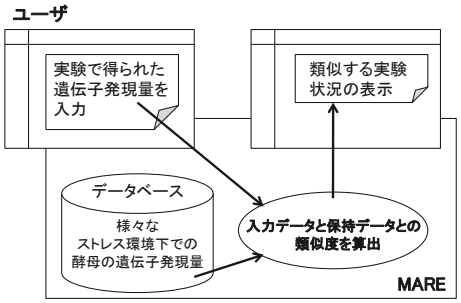

図 2: 本研究の流れ

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。概要は図 2 のようになる。

3.1　利用するデータベース

MARE が利用するデータベースでは、173 通りの様々なストレス環境下における酵母の 6,152 遺伝子それぞれの発現量の値を保持している。[4]

3.2　データベースから類似する実験状況を算出

3.2.1　Pearson の相関係数

ユーザが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、もっとも類似度の高いものをユーザに返す。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 の相関係数とは、2 つのベクトル $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ が $\boldsymbol{x}_1 = (x_{11}, x_{12}, \ldots, x_{1n}), \boldsymbol{x}_2 = (x_{21}, x_{22}, \ldots, x_{2n})$ として与えられた場合、$\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ の相関係数は、以下の式で定義される。

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

$$\rho_{1,2} = \frac{\sum_{k=1}^{n}(x_{1k}-\bar{x_1})(x_{2k}-\bar{x_2})}{\sqrt{\sum_{k=1}^{n}(x_{1k}-\bar{x_1})^2}\sqrt{\sum_{k=1}^{n}(x_{2k}-\bar{x_2})^2}}$$

$$\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ik} (\text{ベクトル}\boldsymbol{x}_i\text{の要素の平均値})$$

相関係数は常に $-1 \leq \rho_{1,2} \leq 1$ の範囲にある。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正の相関が高いデータベース内の実験状況を、入力データとの類似度が高い状況であると考える。

3.2.2 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各実験状況の、n 個の遺伝子における発現量から、Peason の相関係数が計算できる。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。遺伝子 a、遺伝子 b、遺伝子 c の発現量が、それぞれ $3.01, 0.98, 0.37$ である実験状況のデータを入力したとし、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。同様に計算すると $\rho_{0,2} = 0.22, \rho_{0,3} = -0.93$ となる。よって、実験状況 $1, 2, 3$ の順に類似度が高いことをユーザに提示する。

4 実行例

この章では実際の MARE の検索手順を説明し、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。MARE のシステムの構築には Ruby on Rails 1.2.6、MySQL5.0.51 を用いており、MacOS X 上で実行している。

4.1 実行例 1 — 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 2 秒ほどの検索の結果、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、類似度第 1 位として表示されている。また 2 位と 3 位には同じ熱ショック後 5 分と 20 分、すなわち、入力の 5 分前と 5 分後が算出されている。データベースに熱ショック応答実験結果は 5 分おきに保持されているので、入力データに時間が近い 2 個の状況が抽出されていることがわかる。本実行例から提案手法が有効であると言える。

4.2 実行例 2 — 糖代謝

MARE の有用性を示すため、異なったデータを検索しよう。MARE のデータベースの実験の内、酵母に糖を与えた後 60 分経過後のサンプルである、1M sorbitol -60 min から 500 個の遺伝子をランダムに選ぶ。この遺伝子と発現量のデータを、仮に実験で得られたデータとして MARE に入力する。MARE は検索の結果、もっとも似ている実験状況として、糖投与後 60 分の状況を 1 位に表示する。また 2 位と 3 位に同じ糖投与後 45 分と 90 分、すなわち、入力の 15 分前と 15 分後が算出されている。データベースには、1M の糖投与後 15 分ごとのデータが保持されているので、45 分と 90 分は、入力である 60 分後に類似する実験状況である。本実行例からも、提案する計算手法が有効であると言える。この結果の表示には約 10 秒ほどかかる。

5 今後の課題

本研究では、ランダムに抽出した 50 個もしくは 500 個の遺伝子のみを利用して検索を行った結果、高い精度の結果が返ってきた。これは、酵母が単細胞の比較的単純な生物であるために、少ない遺伝子数でも精度が出た物と思われる。しかし、初めに挙げたようにヒトの遺伝子情報に応用することを考えており、高等生物でも同様の結果が得られるかは、今後の課題である。

また、精度が出ない場合入力遺伝子数を多くする必要があるが、現状では入力遺伝子の増加に伴って、急激に実行時間を要している。たとえば、6,152 遺伝子の入力に対しては、10 分以上の計算時間を要する。今後、ヒト遺伝子情報全体を扱えるように、検索精度を保ったまま、類似度算出の速度をより速めることが必要である。

参考文献

[1] Barrett, T. Suzek, TO. *et al.*, “NCBI GEO: mining millions of expression profiles–database and tools”, *Nucleic Acids Res.*, vol. 33, Database issue,D562–D566, 2005.

[2] Dennis, G. Sherman, B.T. *et al.*, “DAVID: Database for Annotation, Visualization, and Integrated Discovery”, *Genome Biology*,vol. 4, issue 9, R60, 2003.

[3] Fujibuchi, W. Kiseleva, L. *et al.*, “CellMontage: similar expression profile search server”, *Bioinformatics*, vol. 23, issue 22, 3103–3104, 2007.

[4] Gasch, AP. *et al.*,“Genomic expression programs in the response of yeast cells to environmental changes”, *Mol.Biol.Cell*,vol.11,no.12,pp.4241-4257,2000. g