

発現量に依存した表現形の局所集中に基づいた 表現形への影響タンパク質の推定手法

楽 元斌[†] 井上 悦子[‡] 吉廣 卓哉[‡] 中川 優[‡]

[†]和歌山大学大学院システム工学研究科 〒640-8441 和歌山市栄谷 930 番地

[‡]和歌山大学システム工学部 〒640-8441 和歌山市栄谷 930 番地

E-mail: [†]sl21010@center.wakayama-u.ac.jp, [‡]{etsuko,tac,nakagawa}@sys.wakayama-u.co.jp

あらまし 遺伝子やタンパク質の発現量と生物の表現形の関係を解明することは一つの重要課題である。数値で表される表現形を考えると、複数のタンパク質の発現量が近い生物個体に対して、表現形が近い値をとるならば、その表現形とタンパク質の間には何らかの関係があると考えられる。本研究では、タンパク質の発現量データと表現形データを入力にとり、その表現形との関係が推定される2つのタンパク質の組を抽出するアルゴリズムを提案する。また、提案手法を実データに適用することで、その有用性を評価する。

キーワード プロテオーム、2次元電気泳動、表現形、発現量、データマイニング

1. はじめに

近年、ヒトゲノムプロジェクトに代表されるゲノム解読プロジェクトが完了し、ポストゲノムの研究として遺伝子やタンパク質の機能の特定や、遺伝子やタンパク質間で起こる相互作用によって生じる生命現象の解明を目指した研究が盛んに行われている。その中でも、プロテオーム解析はそのようなポストゲノム研究の一分野として数えられる。プロテオーム解析とは、タンパク質全体を網羅的に解析することである。生体内のタンパク質の相互作用やその機能の全体像を明らかにすることで、生体を構成する有機システムの全体を解明しようとする研究が行われている。

プロテオーム解析の一手法として、サンプル組織中の各タンパク質の含有量（発現量と呼ぶ）を測定・比較する手法が知られている。二次元電気泳動技術の進歩により、多くのタンパク質の発現量を同時に測定できるようになり、またその精度も向上した。この手法の出現により、タンパク質発現量データを用いた様々なデータマイニングが盛んに研究されるようになった。

本研究では、タンパク質発現量を用いたデータマイニングの中でも、タンパク質と生物の表現形（表出する生物個体の特徴）との関係を取り扱う。数値で表される表現形を対象として、2つのタンパク質の発現量が近いサンプルに対して、表現形値も近い値をとっているような場合を発見することで、2タンパク質と表現形の関係を推定するアルゴリズムを提案する。さらに、実データを用いて評価実験を行い、その有用性を評価する。

2. 想定する表現形とタンパク質の関係

2.1. 表現形

表現形とは、生物のもっている性質や特徴のことである。遺伝によって子孫に伝えられる。生物はそれぞれ、様々な表現形を持つ。例えば、牛には角が生えている牛と角がない牛がおり、犬の毛の色、猫の尻尾の長さ、蝶の羽の模様も多様である。また、経済的に重要な表現形として、例えばブランド牛の場合には、肉用牛の枝肉重量（出荷できる枝肉の総重量）や霜降りの度合い（BMS）等が挙げられる。このような表現形をデータ分析に用いる場合には、多くの場合に、表現形を何らかの数値で表現する。本研究でも、入力される表現形データは、各表現形を数値として表したデータを想定している。

2.2. タンパク質の発現量

タンパク質は生物に固有の物質である。生物のあらゆる部位に存在し、生命活動を支える最も重要な物質の一つである。細胞の中で合成され、生物の構造そのものとなり、あるいは酵素などとして生命現象に利用される。生体内のタンパク質の働きはまだ完全には解明されていないが、ヒトには十万種以上のタンパク質があるといわれている[4]。タンパク質発現量とは、各タンパク質のサンプル組織中の含有量のことである。タンパク質の発現量は、二次元電気泳動等の方法によって測定される。この方法は、サンプル中に含まれるタンパク質の発現量を網羅的に測定できることから、広くタンパク質発現量測定に用いられている。

2.3. 想定する表現形とタンパク質の関係

表現形はある程度が遺伝によって子孫に伝えられる。遺伝子が生物の表現形の情報の一部を担っている。遺伝子の情報を元にタンパク質が合成され、生物の生命現象に影響を及ぼす。タンパク質は遺伝子よりも直接的に生物の表現形に影響を与えている。タンパク質は生体内で、相互作用をしながら機能を果たしている。つまり、タンパク質の表現形への影響を考えるうえでは、タンパク質の組み合わせを考えることが重要である[4]。生物は様々な表現形をもつゆえに、生物個体ごとに同じタンパク質が存在していても、発現量は異なる。

ある2つのタンパク質の発現量から構成した二次元平面上で、一つ一つのサンプルが成す点の配置は通常規則的にはならない。さらに、このグラフにある表現形の表現形値を三次元目の軸として加え、2つのタンパク質の発現量を残りの2軸とする3次元空間を考えると、もし2タンパク質と表現形の間に関係がないならば、サンプルは規則的には配置されない。ここで、もし偏った表現形値をとるサンプルのみが集合している領域があれば（例えばある表現形の値が大きな値をとるサンプルが局所集中しているなど）、それは偶然に発生する事象ではないため、その2タンパク質と表現形の間には何らかの関係が隠れていることが推定される。本研究ではそのような、表現形値が近いサンプルが局所集中するような領域を探索することで、表現形に対して影響力を持つ2タンパク質を特定することを目的とする。

3. 発現量平面上で表現形へ影響するタンパク質推定手法

3.1. 入力データ形式

入力となるデータは二種類ある。タンパク質発現量データと表現形データである。発現量データは、2次元電気泳動で得られるデータを想定し、各サンプルに対して含まれる各タンパク質の発現量が数値として表現されたものを想定する。発現量データは、適切な正規化手法を用いて正規化されているものとする。表現形データは、生物学的な実験で得られる各サンプルに対して各表現形の程度を数値として表現したものを想定する。

3.2. 求めたい領域

本研究は2次元平面上で偏った表現形値のもつサンプルのみが集まったような領域を求めることにより、表現形へ影響するタンパク質の組み合わせを推定する。図1に求めたい領域の例を示す。この図では、縦軸と横軸に1つのタンパク質AとBの発現量を取り、各サ

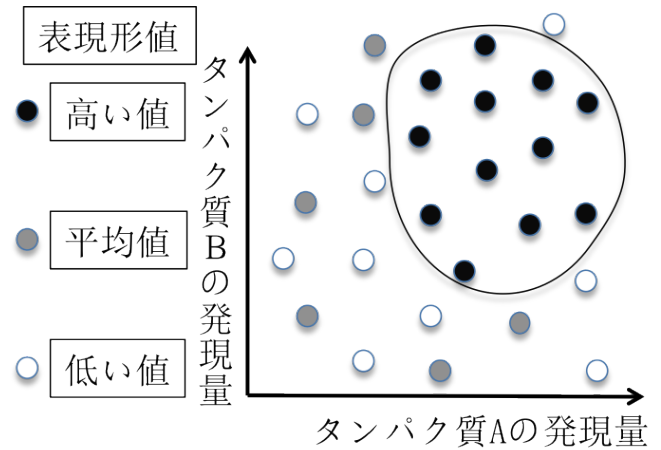


図1：求めたい領域の例

ンプルがこの2次元平面上に配置された散布図である。サンプルを表す点の色はある表現形の値の大きさを表しており、色が濃いほど大きな値、薄いほど小さな値であることを表す。

求めたい領域は、ただ連続しているだけでなく、以下に示すような条件を満たす必要がある。

第一に、領域内の表現形値の分散が小さいことである。本研究で想定する表現形とタンパク質の関係は、表現形の値が近い点のみが集まる領域が存在することである。このため、領域内に含まれるサンプルの表現形値の分散が、他の部分に比べて小さいことが求められる。

第二に、形がいびつではないことである。領域の形として、大きな凸凹もなく細長過ぎることもない、形の良い領域である必要がある。

最後に、領域がそれなりに広いことである。領域内のサンプルの数は全サンプル数のある割合を示さないと、偶然に発生する事象も捉えてしまうため、統計的に信頼できる結果とは言えないからである。

3.3. 提案アルゴリズムの概略

3.2節で述べたような領域を発見するアルゴリズムを設計した。図2（図の中の1,2,3,4は以下の手順に該当するものとする）及び以下に、手順の概略を示す。

1. 全表現形の中から一つを選ぶ。この表現形に対し、二つのタンパク質を成す平面を作成するように全組み合わせを作る。一つの表現形と二つのタンパク質を一つのペアとして、領域を抽出する。
2. 各タンパク質の組み合わせが成す平面上で、サンプルの隣接関係を作成する
3. 求める領域を計算する。手順の概略は以下のようになる。
 - a) 出発点を選ぶ。1点から成る初期領域が決まる。

- b) 領域に隣接する点を候補点とする。
 - c) 各候補点に対してコスト計算する。このコストは拡張コストと名づけ、どの候補点が最もふさわしいかを評価する目的がある。
 - d) コストの最もよい値を持つ候補点を選び、領域に加えて、領域を拡張する。これを繰り返し拡張していく。
 - e) 領域を拡張するたびに領域内の分散を計算する。それをランキングスコアとする。
4. 各タンパク質の組み合わせの中で計算してできたたくさん分散の中で、最も値の小さい分散値をこのタンパク質の組み合わせのランキングスコアとする。そして、表現形ごとにスコアをランキングする。

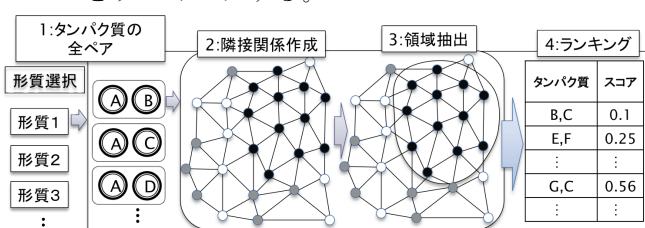


図2. 提案手法の流れ

3.4. 隣接関係の生成

領域を抽出する第一歩として、サンプル間の隣接関係を計算する必要がある。このために、ボロノイ図を作成する。このボロノイ図に対応するドロネーグラフを構築し、このグラフ上で隣接しているサンプル間に隣接関係があるとする。

3.5. 拡張コストの計算

拡張コストは、領域を拡張する際に、どの候補点を取ればいいかを判断する基準として計算する。拡張コストをCとし、Cは二つの部分に分けられる。形状の評価コストTと、表現形値の評価コストZである。形状コストTは値が大きければ大きいほど形が悪くしていき、表現形コストZは値が大きければ大きいほど表現形がばらついていく。コストCは形状の評価コストTと表現形値の評価コストZを掛け合わせたものである。

コストTは候補点がどの程度領域に囲まれたかを評価する。隣接関係を元に候補点の周囲にある候補点との距離が一定になる点の数の内に、すでに領域内に入っている点の数をカウントする。この割合は候補点の包围度を表わしている。式は $T = (1 - (M1/N1 + M2/N2 + M3/N3))$ である。ここで N_n は候補点との最短距離（ホップ数）がnであるすべての点の数である。例えば、 N_1 は最短距離が1のすべての点の数を表す。 M_n は二つの条件を満たす点の数である。条件の一つ

は N_n のうち、領域にある点のことである。もう一つは、 N_n のうち、最短距離一つノード前の点も領域内にある点のことである。nは3までを計算する。各最短距離で求めた割合を足し、1から引くと値の最も大きい点は評価が最も高いものとなる。

コストZは候補点の表現形値がどれだけ現領域内の表現形値に近いかを評価する。現領域内の点の集合をxとする。xは正規分布に従うと仮定する。すると、評価値としてz値の絶対値で表すことができる。Z値とは、平均値からの距離を標準偏差の倍数で表した単位のことである。集合xの平均を μ とし、標準偏差を σ とし、候補点の表現形値をaとすると式は $Z = |(a - \mu) / \sigma|$ である。

3.6. ランキングの作成

領域を拡張する際、点一個加えて拡大するたびに、領域内の分散を計算する。領域を最後の一点まで拡張し、計算した分散値の中で、値の最も小さく、かつ領域内の点の数が全体の20%以上になるような値をそのタンパク質組み合わせのランキングスコアとする。ランキングスコアを元に作成したランキング表の上位にあるタンパク質の組み合わせは表現形への影響が強いと推定する。

4. まとめ

本稿では、発現量に依存した表現形の局所集中に基づいた表現形への影響タンパク質の推定手法を提案した。統計的な視点から、偶然では起こらない、表現形値の近似値ばかりを持つサンプルが集まる領域を抽出することで、タンパク質と表現形間の関係を推定することが可能である。

領域抽出アルゴリズムを提案し、実装・評価を行った。実データ（和歌山県地域結集型共同研究事業[3]により得られたウシのタンパク質発現量データと表現形データである。この発現量データと表現形データは文献[4]に記載されているプロテオーム解析支援システムにより得られたものである。）へ適用することで評価を行った。その結果として、表現形への影響タンパク質の組み合わせを表わすランキング表ができ、その中に適切な結果が見つかった。アルゴリズムは提案手法に準じて作動していることが確認した。これらの結果を、最終稿では記述する予定である。

謝辞

本研究の一部は生研センターイノベーション創出基礎的研究推進事業の支援を受けたものである。ここに記して謝意を示す。

参 考 文 献

- [1] 中野明彦, 遠藤斗志也, タンパク質の一生, 共立出版, 2000.
- [2] 池内俊彦、畠中 寛、タンパク質と遺伝子、オーム社出版、平成 8 年
- [3] 和歌山県地域結集型共同研究事業,
<http://www.wakayama-kessyu.com/>.
- [4] 永井宏平, 吉廣卓哉, 井上悦子, 池上春香, 園陽平, 川路英哉, 小林直彦, 松橋珠子, 大谷健, 森本康一, 中川優, 入谷明, 松本和也, 黒毛和種肥育牛の枝肉表現形バイオマーカーの探索 I : 大規模プロテオーム解析情報と血統・枝肉表現形情報の統合情報管理システムの構築, 日本畜産学会報, Vol.79, No.4, 2008