1M5-3

行列因子分解による遺伝子データからの潜在的因子の抽出

Extraction of latent factors from genetic data by matrix factorization

村上勝彦
Katsuhiko Murakami

東京工科大学
Tokyo University of Technology

Information or annotation described in genetic database, such as disease and gene functions have relationships each other. However they are described as if they were independent concepts. This causes difficulty to understand in their deep meanings for the user even for the specialists. To manifest such relationships among terms, it is required to extract hidden relationships among biological concepts and add them into databases. For this purpose, we have factorized genetic data set and extracted some hidden relationships suggested by matrix factorization. This re-organization of knowledge is helpful for precise manipulation of information.

1. はじめに

医学生物系のデータを扱う場合，遺伝子（タンパク質）とその役割（機能）の関係は重要で，従来から広く収集活用されている．遺伝子の機能などの情報は日々発見され，データベースが更新される．記述に用いる用語については，オントロジーの研究から次第に整備がすすんでいる．例えば遺伝子機能については Gene Ontology (GO) [GO 00, GO13] が整備されており，著名なデータベースは概ね制御されたボキャブラリーが付与される[NCBI 14, Imanishi 04]. GO は用語の集合であるが，各用語が Directly Acyclic Graph (DAG)の親子関係（例えば IS-A 関係）が付与されている．すなわち全用語は木構造でつながった構造をもっている．木構造のルート（根）は 3 種の概念，すなわち機能(molecular function)，生物学的な反応過程（biological process），および細胞構成要素（cellular component）となっている．制御された用語以外にも，テキストによる自由記述も多い．

生物学でのタンパク質間相互作用を測った大規模な実験結果からも，細胞内の多くの現象が多機能性のタンパク質をとおして直接関連する可能性が指摘されている．したがって，用語として認識され定義された時には独立に見出された用語（概念）であっても，大規模に収集されたデータを解析すると，実はいくつかの事実を経由して関連のある現象かもしれない．このとき具体的な遺伝子をデータベースの知識を用いて候補を列挙することができれば，詳しい解釈を支援したり，正しい推定が可能になる．

ここではヒト遺伝子のデータベースである H-InvDB から遺伝子とその付与された情報，とくに機能情報を記述するための GO をとりあげ，それらの関係性の全体から，隠れた関係性を示す構造をとりだすことを目的とする．

遺伝子名と付与された情報は，関係データベースで格納できるように，テーブル構造のデータに整理することができる．テーブル構造（あるいは行列データ）から，関連する部分をとりだす過去の研究としては，例えば発現データについて，発現が近い遺伝子同士と条件同士（発現組織など）を同時にクラスタリングする「バイクラスタリング」[Madeira 04]が行われてきた．しかしバイクラスタリングは，因子分析でいう因子のような隠れたアトミック的な要素を取り出す解析方法ではない．また，我々はこれまでヒト遺伝子データベースに記述された用語ペア間の相関を解析し，データに記述されていなかった新たな関連を抽出した[Murakami 04]．しかし同時に複数の項目を説明できるような因子や構造については考慮していない．

本研究では，非負値行列因子分解 (NMF)を用いて，機能アノテーションなどの用語を複数同時に取り扱い，非負値行列因子分解によって関連する複数の遺伝子と複数の用語の関係を説明できるような関連性すなわち潜在的因子を抽出することを目的とする．

連絡先：村上勝彦，東京工科大学，応用生物学部，東京都八王子市片倉町 1404-1，murakamikh@stf.teu.ac.jp

2. データと方法

2.1 非負値行列因子分解 (NMF)

NMF は，データ行列 V を，基底行列 W と特徴行列 H の積の形に分解する方法で，

$$V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu} \quad (1)$$

で，表される．ここで，V, W, H についてすべての行列要素が非負値という制約がある．行列の次元は，V が n×m，W が n×r，H が r×m である．また，r は基底ベクトルの数で，解析のときに与えるものである．顔画像の要素の抽出に用いられた例[Lee99]が有名で，行列形式の連続値を要素に持つデータを NMF で解析することにより，顔画像の要素を取り出し，それらの和で元のデータが近似的に表現できることが報告された．これ以降，NMF はさまざまな分野で応用されており，遺伝子データに関しても発現量データを中心に利用されている[Devarajan 08].

2.2 遺伝子データ

遺伝子のリストと機能情報のリソースとしては，ヒト遺伝子統合データベース H-InvDB [Imanishi 04, Takeda 12] (Release 8.3 修正版; http://h-invitational.jp/) を利用した．H-InvDB に登録された遺伝子のうち，GO が付与された 12,261 遺伝子を抜き出した．ユニークな GO の回数（異なり数）は 1,741 個であり，延べ数は 40,871 個であった．1遺伝子あたり平均 3.3 個の GO が付与されていた計算となる．通常の文書解析の場合と異なり，1遺伝子に何度も同じ GO が出現することはない．すなわち，同じ GO 用語は0回か1回の出現しかない．計算時間の都合上，このデータから一部の遺伝子 1,000 個をランダムに取り出した．

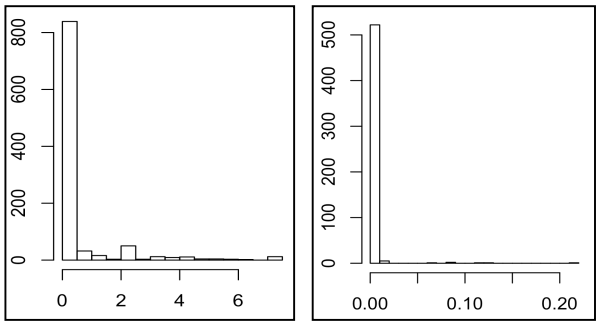
このようにして遺伝子 N 個と，それらに付与されていた用語 M 個から，N 行 M 列の行列 T を作成した．この行列要素は，遺伝子 i に用語 j が付与されていれば，T(i, j)が 1，付与されていないならば 0 となる行列である．この行列に対して，非負値行列因子分解を行った．基底 r の数は4として行った．収束の目安としては，W*H の全要素の分散が減衰していくが，減衰変化分の全分散に対する割合が，1E-5 より下回った場合に，収束したとみなして停止した．これを 30 回実行し，分散が最小のケースを採用し，以下に詳細に分析した．

3. 結果と議論

行列を分解したあと，4つの基底のそれぞれで，要素の数値が高い値をもっているものに対応する行（遺伝子）と列（機能などのアノテーション）を調べた．図1に基底1に対する要素の値のヒストグラムを示す．

W と H で数字の分布が異なるため，比較は W 内，H 内の相対的比較のみが可能である．図は基底1についてであるが，他の基底についても，一部の数値が高く，それ以外のほとんどの数値は0かそれに近い値になっているのは共通であった．

図1　基底1に対する要素の値のヒストグラム



(a)左は W の要素値，(b)右は H の要素値である．

3.1 各基底についての検討

1つ目の基底について大きな値を持つ H（GO ターム側）の要素として最大の3つは，0.21 (GO:0005634, nucleus;核) ，0.13 (GO:0003677, DNA binding; DNA への結合)，および 0.12 (GO:0006355, regulation of transcription, DNA-dependent; 遺伝子の転写制御)に対応していた．行列 W（遺伝子側）で要素値が大きな(1.0 以上)，これに関連する遺伝子は 62 遺伝子もあり，上位10個を示すと ALX4, DLX5, ISL2, ESRRG, HES1, POGK, TOX4, TOX, TSHZ3, EBF4 などであった．これらの遺伝子は細胞核内で転写制御に関連しており，妥当な結果である．

また，上位20遺伝子のうち16個しか (GO:0005634, nucleus; 核)を付与されていなかったので，”nucleus”を付与されているためにスコアが上がったわけではないことがわかる．言い換えれば，”nucleus”を付与されていないものが関連する遺伝子として浮かび上がってきたということである．しかし，それらの遺伝子は他の関連する GO である，regulation of transcription, DNA-dependent)などを付与されていた．

以下，同様の方法で2つ目の基底についても調べた．基底2については，GO の最大であったものは，(GO:0005524, ATP binding)が 0.17， (GO:0003824, catalytic activity)が 0.12，(GO:0008152, metabolic process)が 0.10 であり，これらは酵素反応に関するものに対応することが示された．この基底2に対応するスコア 1.0 以上の遺伝子は 134 個で，遺伝子は ATP1A2, ATP13A3, ATP10B, ABCC10, PYCRL, GFOD1 などであり，ATP のエネルギーを用いて物質の輸送を行う膜輸送体の一群と考えられた．

3つめの基底では，0.30 (GO:0005622, intracellular), 0.24 (GO:0003676, nucleic acid binding), 0.24 (GO:0008270, zinc ion binding) のスコアが高かった．これらはそれまでと違って，いずれも 0.23 以上と高いスコアであるということは，これらのタームは初期行列の説明に比較的大きく貢献しているということである．基底3についてスコアの高い(>1.0)遺伝子は 139 個で，ZNF254, ZNF492, ZMAT4, BCL11A などの転写因子群であった．

第4の基底については，スコアの高い GO は，0.32 (GO:0005515, protein binding)である．以下は低いスコアであるが，0.07 (GO:0008270, zinc ion binding), 0.03 (GO:0006886, intracellular protein transport)が続く．これに対応する遺伝子群は RAB8A, RAB31, KCNS2, RPH3AL など 93 個であった．ここで，GO のスコアが通常ないくらい高いものを示したが，これはこのテーム(GO:0005515, protein binding)の1つで93個の遺伝子が説明できていることを示唆している．

以上のように，遺伝子データの GO ターム（用語）のデータにおいて，複数の GO をまとめた概念ともいえる，隠れていた基底が存在することが明らかになった．言い換えれば GO 情報の冗長性が示唆された．さらにそれらの遺伝子や GO の情報のクラスターは，ことなる基底で扱える可能性が示された．

4. おわりに

本研究ではヒト遺伝子の統合データベース H-InvDB に付与された別のものとされている個々の GO タームについて，複数の用語を同時に取り扱い，非負値行列因子分解によって関連する複数の遺伝子と複数の用語の関係をセットとして抽出した．

これら集合の存在が確認されたことは，GO の冗長性やその数理的な程度があり，それらを定量的に扱える可能性を示唆している．

今後は大規模に行うことと，基底の数を増やして，精度をあげていくことが期待される．

参考文献

[Devarajan 08] Devarajan K: Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. PLoS computational biology 2008, 4:e1000029

[GO 00] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat Genet. May 2000;25(1):25-9.

[GO 13] The Gene Ontology Consortium, Gene Ontology annotations and resources. Nucleic Acids Res, 2013. 41(Database issue): p. D530-5.

[Imanishi 04] Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K.O., Barrero, R.A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M. et al.: Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS biology, Vol. 2, No. 6, pp.e162. 2004

[Lee 99] D.D. Lee and H.S. Seung, “Learning the parts of objects with nonnegative matrix factorization,” Nature, vol. 401, pp. 788-791, 1999.

[Madeira 04] Madeira SC1, Oliveira AL.: Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinform. 2004 Jan-Mar;1(1):24-45.

[Murakami 14] Murakami, K., Imanishi, T.: 遺伝子データからの相関する概念抽出と関係づけオントロジーの作成. 人工知能学会, 1G3-2, 2014

[NCBI 14] NCBI Resource Coordinators.: Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2014 Jan;42(Database issue):D7-17.

[Takeda 12] Takeda, J., Yamasaki, C., Murakami, K., Nagai, Y., Sera, M., Hara, Y., Obi, N., Habara, T., Gojobori, T. and Imanishi, T.: H-InvDB in 2013: an omics study platform for human functional gene and transcript discovery. Nucleic acids research, Vol. 41, pp.D915-919. 2013