4J4-3

遺伝子オントロジーの階層的クラスタリング

Hierarchical Clustering of Gene Ontology by Non-negative Matrix Factorization

村上勝彦

Katsuhiko Murakami

東京工科大学

Tokyo University of Technology

Gene Ontology (GO) terms are independently annotated in gene database. To elucidate the relationships among the terms, we applied non-negative matrix factorization to gene-term matrix as a clustering method. We extended our approach to find refined smaller clusters. The optimal number of clusters were explored. The evaluation using some coefficients showed good performance even for large number of clusters.

1. はじめに

バイオインフォマティクスでは，遺伝子やタンパク質のメタ情報がよく利用される．論文で明らかになった機能や関連情報は専門のアノテーターによって，データベースに登録され，さまざまな形で利用される [NCBI 15]．例えば機能未知の DNA 配列が実験で容易に得られるため，配列類似検索によって似た配列の機能が未知配列の機能として類推されることに利用される．遺伝子の機能の一部は統制されたターム（Gene Ontology; GO）で表現されている[GO13]．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．IS-A 関係のリンクだけで言えば全用語は3つの木構造でつながっている． 3 種の概念とは，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．

ユーザーがこれらをより良く理解，利用するためには，ターム間に関連がある場合にそれをデータ化することが必要である．ユーザーが1つ1つを見たときに関連語を表示したり，また関連を網羅的に扱うことで高度な解析が可能になる．ここでは遺伝子のアノテーション情報を使ってGOターム間の関連を明らかにすることを目的とする．

細胞を分類するために非負値行列因子分解を適用して階層クラスタリングをする方法がある[Brunet 04]．Brunet らは基底数（クラスター数）k を変化させることによって，データに階層構造が存在する場合には，その部分構造を発見できることを見出した．我々は以前 k を 4 に固定して本手法で生物学的にも意味のあるクラスターを取り出すことを示した[村上15]．しかしこのクラスター数が適当であるかは未検証であった．分解した行列でクラスターの成分をみると小さな係数を持つターム数が多い，いわゆる大きなクラスターがあり，その解釈は明快なものではなかった．そこでより大きなクラスター数 k を設定すれば，より精密なクラスターが多く取れる可能性が残っていた．本研究では，大きな k を設定して精度の高いクラスターを取り出し，互いに関連する GO タームの関係をより多く取り出すことを目的とする．同時にいくつかの指標を用いて最適なクラスター数 k の探索を行う．

2. データと方法

2.1 遺伝子データ

遺伝子リストと機能情報が付与されたデータとして，ヒト遺伝子統合データベース H-InvDB [Takeda 13] (Release 8.3; 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 タームは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．これに付与されていた 533 個の GO タームを使用した．

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．

2.2 非負値行列因子分解 (NMF)

NMF の初期のしごとでは，顔画像のデータに NMF を適用することで顔の特徴を取り出すことで使われた[Lee99]．それ以来，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などで利用されている[Devarajan 08, Kim 03]．

NMF ではデータ行列 V を，以下のように基底行列 W と特徴行列 H の積の形に分解する．

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

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×k，H が k×m である．また，k は基底ベクトルの数（クラスタリングの目的においてはクラスター数）で，解析のときに与える．コスト関数として以下を定義する．

$$F = \|V - WH\|_F^2 \qquad (2)$$

添字の F はフロベニウスのノルムである．初期値としてランダムな非負値を与え，通常の補助関数を用いて更新する．

計算には NMF の Python ライブラリである NIMFA[Zitnik 12] を用いた．収束条件として更新数を最大で 1000 回とした．基底 r の数は 4, 8, 16, 32, 64, 128, 256 について行った．

3. 結果と議論

基底数 k を増加させながら，それぞれの k で行列を分解した．異なる k は異なる階層のクラスターを示すと考えられる．100 回の繰り返しで同じクラスターになったかどうかをみるコンセンサス係数を図1に示す．コンセンサス係数は，GO タームの数を行（列）数とする行列であり，要素の値の大きさは複数回の結果が一致した率であり，0 から 1 の数値をとる．図は青(0)から赤(1)の色になっている．対角領域に赤い四角または非対角領域に青い領域が多ければ安定したクラスターがとれていることを示す．図のように k が 5 程度では不安定であるが，k が大きくなると非対角線上に青い領域が増え，クラスターが安定するようになったことがわかる．

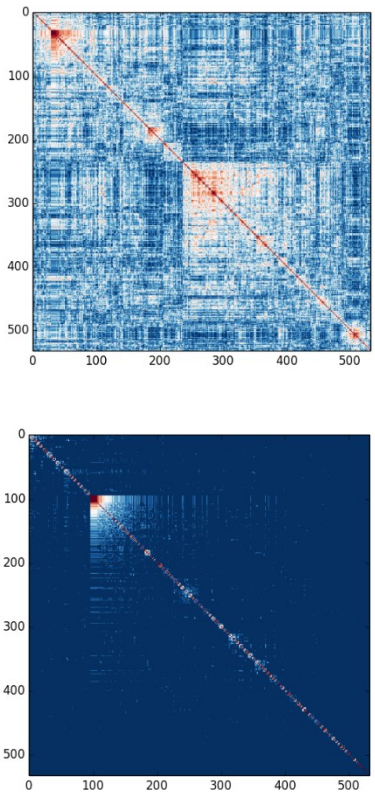

図1　コンセンサス行列．k=4（上）と k=256（下）の場合を示す．

NMF では分解の良さを示す指標の1つとして，コンセンサス行列と階層的クラスタリングとの相関をとった Cophnetic correlation (CC)がある．完全なクラスタリングの場合はコンセンサス行列の要素がすべて 0 か 1 となり，その結果 Cophnetic correlation は 1 となるが，不完全な場合は1以下の値をとる[Brunet 04]．これを k を変えながらピークになったところが最適な k とする方法がある．表1に Cophnetic correlation の変化を示す．K を増加させていくと値は上昇し，k=256 でも良好な値をしめしている．

NMF での分解の良さを示す別の指標として，残差平方和がある（表1）．k の増加と共に残差平方和が減少していることから，k の増加は分解の精度を上げていると考えられる．これより増加させることも可能であるが，GO ターム数の半分に近いため，メンバー数が1つであるようなクラスターが多く作られることが予想される．これは当初の目的からはずれることになる．したがって，これ以上大きいクラスター数(k)の探索はしていない．これを行うには，当初の目的を阻害しないような指標を新たに導入しなくてはならない．これは今後の課題である．

表1　相関と残差平方和のクラスター数による変化

k	4	8	16	32	64	128	256
CC	0.51	0.66	0.70	0.77	0.84	0.89	0.94
RSS	2,441	2,117	1,760	1,353	960	569	244

4. おわりに

本研究ではヒト遺伝子データベースに付与された機能等の情報としての Gene Ontology (GO) タームについて，非負値行列因子分解によるクラスタリングを行い，クラスター数を探索してより精度をあげる方法について検討した．様々なクラスター数 k によって階層的にクラスタリングを行ったとも考えられる．今回の指標では、検討した範囲では最大値 (k=256) が最適であると考えられる．NMF などの行列分解では通常クラスター数は 10 程度で検討されるケースが多いが，本データでは遺伝子数1000個，GO ターム数 533 個の行列で，平均 3 個というスパースな行列であるため，大きなクラスター数 k でも細分化する意味を残しつつ近似を良くできると考えられる．

参考文献

[村上 15] 村上 勝彦, "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," in *第 29 回人工知能学会全 国大会論文集 (JSAI 2015)*, ed. 函館, 2015.

[Brunet 04] J.-P. Brunet, *et al.*, "Metagenes and molecular pattern discovery using matrix factorization.," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 101, pp. 4164-9, 2004.

[Devarajan 08] K. Devarajan, "Nonnegative matrix factorization: an analytical and interpretive tool in computational biology," *PLoS Comput Biol,* vol. 4, p. e1000029, 2008.

[GO 13] The Gene Ontology Consortium, "Gene Ontology annotations and resources," *Nucleic Acids Res,* vol. 41, pp. D530-5, Jan 2013.

[Kim 03] P. M. Kim and B. Tidor, "Subsystem identification through dimensionality reduction of large-scale gene expression data," *Genome Res,* vol. 13, pp. 1706-18, Jul 2003.

[Lee 99] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature,* vol. 401, pp. 788-91, Oct 21 1999.

[NCBI 15] R. C. NCBI, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Res,* vol. 43, pp. D6-17, Jan 2015.

[Takeda 13] J. Takeda, *et al.*, "H-InvDB in 2013: an omics study platform for human functional gene and transcript discovery," *Nucleic Acids Res,* vol. 41, pp. D915-9, Jan 2013.

[Zitnik 12] M. Zitnik, "NIMFA : A Python Library for Nonnegative Matrix Factorization," *Journal of Machine Learning Research,* vol. 13, pp. 849-853, 2012.