テンソル分解を用いた二重鎖切断の起こりやすいヒト遺伝子の推定

Conjecturing human genes that are easy to be double strand breaks by tensor decomposition

物理学専攻　保坂　伸生

Physics major Nobuo Hosaka

1. 研究背景

二重鎖切断(DSB)とはDNA損傷の一つで、DNA損傷というのはDNAを構成する塩基と呼ばれる物質が損傷することを指す。中でもDSBは損傷のレベルが高く、二重螺旋構造を持つDNAの二本どちらにも損傷が起きている状態を指す。このDSBを含むDNA損傷は太陽光に含まれる紫外線や呼吸によって生じる活性酸素、食物に含まれる発がん物質などによって生じ、日常的に起こる危険性をはらんでいる。

DSBはDNA損傷の中でも修復が難しく、修復されない場合、細胞が死を選択してしまい、過剰に細胞が減ると病気になってしまうが、修復されたとしても後述する変異によって遺伝子を喪失してしまう可能性がある。DSB以外の修復は比較的簡単で、DNAはアデニン(A)、チミン(T)、グアニン(G)、シトシン(C)の4つの塩基で構成されていて、中でもAとT、GとCが結合することで二重螺旋構造を作る。そのため、一本のみの損傷であれば、反対側の塩基を読み取ることで容易に修復できる。これが二本とも損傷している場合はとても難しく、修復方法には相同組換え(Homologous Recombination(HR))と非相同末端結合(Non-Homologous End Joining(NHEJ))の2つがある。HRでは、損傷したDNA塩基配列に似た配列のものを鋳型として使い、損傷した部位を作り直す。HRは比較的精度が良いがそれでも、似た配列を使っていることから若干の違いが生じる可能性があり、またHRは常にできるわけではなく細胞周期の一部でしか行われていない。NHEJは損傷した先端部分を両者ともに削り、平坦にしたところでそのまま結合するといったものになっている。これは損傷部位を完全に修復することをしていないことから損傷の度合いによっては塩基配列の大きな喪失となる。このような喪失を変異の一つである欠損(Deletion)と呼ぶ。

HR、NHEJどちらも欠損が生じた箇所に遺伝子が含まれていた場合遺伝子の塩基配列が変わることで遺伝子喪失の危険性がある。この遺伝子喪失ががんを抑制するような遺伝子で生じた場合、がんになるリスクが上昇してしまうことになる。このような点からDSBは様々な病気を引き起こす大変危険なDNAの病と言える。しかし、このような危険性に反して、DSBに関する研究はあまり芳しくない。特に機械学習を用いた解析は、多くのデータを必要とする性質上稀である。生物データは倫理、金銭、時間等様々な理由によって多くのデータを集めることが難しいためである。

2. 研究目的

本研究は、DSBとDSBを修復する経路の関係性を調べる先行研究[1]のデータに着目し、ヒトの常染色体1-22を平均化して解析を行っているところを常染色体毎に個別に解析することで、データとして取られているDNAについたタンパク質とDSBの関係性を遺伝子レベルで調べてDSBの起こりやすい遺伝子の位置を推定することを目的としている。

解析手法としてはテンソル分解を用いた教師なし学習による変数選択法[2]をヒトのDNAについたタンパク質のデータとDNAに生じたDSBを測定したデータに使い、DSBの起こりやすい遺伝子の位置を推定する。これにより、データが少ない環境下で行われる生物学に対する機械学習のモデルケースの一つという成果だけでなく、DSBが起こりやすい遺伝子のリストが得られる。DSBの修復で変異が生じやすいことがわかっているので、変異が起こりやすい遺伝子がわかったことになる。この結果から、がんなどの遺伝子変異によって生じる病気の原因である遺伝子を上述のリストから新たに見つけることができるかもしれない。また、未来の技術であるが、遺伝子に対する治療が確立されたときに変異が起こりやすい遺伝子をリストによって絞り込むことで、約2万個の

中から探すよりも効率的に治療が行えると考えられる。

3. 研究方法

本研究では、テンソル分解にHOSVD(Higher Order Singular Value Decomposition)を用いる。これは日本語では高次元特異値分解と訳すことができ、特異値分解が行列つまり二次元を扱っていることから三次元以上の場合の特異値分解と考えられる。行列の表記をN×M行列XをX(N,M)と表記することにすると、特異値分解は以下のように表せる(式(1))。ただし、XはX(N,M)(N>M)、UはU(N,M)、VはV(M,M)、ΣはΣ(M,M)でU、Vは直交行列、Σは対角行列。

$$X = U\Sigma V^T \quad (1)$$

UもしくはVとΣは、両辺に転置したXを掛けることで固有ベクトル固有値として求まる。また、求めていないVもしくはUについては固有ベクトルを使い表すことができる。もし固有ベクトルの数がNもしくはMより少なければ、元の行列Xを低次元で書き表すことができる。

テンソルはunfoldingという手法でもって行列に直すことができる。HOSVDはこれを利用してテンソル分解をする。三次元テンソルX(成分はX_{ijk}(1≦i≦2、1≦j≦2、1≦k≦2))を例にしてunfoldingの具体例(2)、(3)、(4)を示す。

表 1:テンソルX

k=1	j	
i	x_{111}	x_{121}
	x_{211}	x_{221}
k=2	j	
i	x_{112}	x_{122}
	x_{212}	x_{222}

$$X^{i\times(jk)} = \begin{pmatrix} x_{111}x_{121}x_{112}x_{122} \\ x_{211}x_{221}x_{212}x_{222} \end{pmatrix} \quad (2)\ X^{j\times(ik)} = \begin{pmatrix} x_{111}x_{211}x_{112}x_{212} \\ x_{121}x_{221}x_{122}x_{222} \end{pmatrix} \quad (3)\ X^{k\times(ij)} = \begin{pmatrix} x_{111}x_{211}x_{121}x_{221} \\ x_{112}x_{212}x_{122}x_{222} \end{pmatrix} \quad (4)$$

Unfoldingで得られた行列にSVDをして、得られた固有ベクトルを$U^{(i)}$、$U^{(j)}$、$U^{(k)}$とすると特異値分解のように固有値を掛けることで元のテンソルXに戻って欲しくなる。しかし、別々の固有ベクトルを掛けても元には戻らない。そこで、これらの行列が直交行列なので、転置行列が逆行列となり容易に逆行列が求まることから、元のテンソルXに$(U^{(i)} \times U^{(j)} \times U^{(k)})^{-1}$をかけることで、固有ベクトルの積と掛けることで元のテンソルXに戻るようなテンソルGを求める。このテンソルGはコアテンソルと呼ばれていて、SVDによって得られた個々のUが元のテンソルに戻るための重みの役割を持つ。元のテンソルの成分をx_{ijk}とし各SVDの固有ベクトルの数をN、M、K、固有ベクトルの成分を$\boldsymbol{u_{l_1 i}}$、$\boldsymbol{u_{l_2 j}}$、$\boldsymbol{u_{l_3 k}}$とするとHOSVDは以下のように表せる。

$$\boldsymbol{x_{ijk}} = \sum_{l_1}^{N}\sum_{l_2}^{M}\sum_{l_3}^{K} \boldsymbol{G(l_1, l_2, l_3) u_{l_1 i} u_{l_2 j} u_{l_3 k}}\ (\mathbf{5})$$

研究に用いるデータは上述した論文[1]で得られたデータを使っている。このデータはヒトの常染色体1-22の全塩基対について発生したDSBや付着したタンパク質を記録したものとなっている。このデータはDSBを故意に起こさせる薬品を加えたものとそうでないものの二種類ある。これらのデータからタンパク質の数DSBの数を要素に持ち、DSBとタンパク質群、塩基対の番号、薬品の有無の3つの次元を持つテンソルを作った。ただし、塩基対の次元については常染色体1-22全てのままでは非常に大きく解析ソフトで扱いにくいため、この次元を分

けることで 155 個の 3 次元テンソルデータとなっている。これに HOSVD を使い、DSB が起こりやすい遺伝子を見つける。

SVD の話に戻ると、分解後は固有値と固有ベクトルのみで書き表すことができ、固有ベクトルが基底になることから固有値の絶対値が大きいかどうかが元の行列を書き表す上で重要となる。というのも固有値の絶対値が大きいほど元の行列に戻す際、固有ベクトルの寄与が大きくなるためだ。ここで、行列をデータとして考えると、固有ベクトルはデータの特徴と考えられる。これを HOSVD に置き換えて考えると、HOSVD で得られる行列は各次元の特徴を表す固有ベクトルとなっているので、この固有ベクトルから DSB が起こりやすい遺伝子の特徴を見つければよい。

式(5)の$\boldsymbol{u}_{l_1 i}$、$\boldsymbol{u}_{l_2 j}$、$\boldsymbol{u}_{l_3 k}$を DSB とタンパク質群、塩基対の番号、薬品の有無とすると、最終的に遺伝子が求めたいことからパラメータl_2にこの特徴をもたせたい。DSB を表す特徴は$\boldsymbol{u}_{l_1 i}$から i に DSB を代入し絶対値が大きいパラメータl_1を選ぶ。起こりやすいという特徴は、DSB が起こりやすい環境下では普段以上に DSB になるものを DSB が起こりやすいと考えて、薬品有りとなしのときで差が大きいパラメータl_3を選ぶ。選ばれたパラメータl_1、l_3をコアテンソル G に代入し絶対値が大きいl_2を選ぶことでこの特徴を選ぶことができる。

このパラメータを塩基対の行列$\boldsymbol{u}_{l_2 j}$に代入し、仮説検定をして DSB が起こりやすい塩基対を求める。帰無仮説はここで得られた$\boldsymbol{u}_{l_2 j}$が無差別に選ばれておりそれぞれの値が独立な正規分布に従うというもので、対立仮説は DSB が起こりやすいような塩基対を選ぶことができているとなる。この条件で仮説検定をし、5%以下で帰無仮説を棄却したものを DSB が起こりやすい塩基対とする。

現状まだ塩基対を選べただけで遺伝子があるかどうかわかっていない。そこで、遺伝子は塩基対が連続していることから、仮説検定によって選ばれた塩基対の中で塩基対の番号が連続しているものをここでは遺伝子と考える。これによって DSB が起こりやすい遺伝子を解析で見つけることができる。

解析結果が実際に DSB を起こしやすい遺伝子について選べているか調べる。COSMIC[3]というがんに関するデータベースから欠損、挿入という DSB によって生じる変異が発生したことがある mRNA のデータを入手し、UCSC genome browser[4]というゲノムデータベースからヒトの mRNA がどの塩基対にあるかマッピングされたデータを入手する。mRNA を使う理由については、遺伝子に比べて量が多く解析データが多いに越したことはない点、タンパク質を作るのに遺伝子全てを使うわけではなく遺伝子の一部を写した mRNA が使われていることから、mRNA の変異を見るほうがどの塩基配列に変異が生じたか正確にわかる点にある。解析データを含めたこの三種類のデータからフィッシャーの正確確率検定を行い、DSB を起こしやすい遺伝子が選べているかどうか調べる。解析データの塩基対の範囲にある mRNA の数を a、その範囲の mRNA の中で変異を起こしたことがあるものの総数を b、a の mRNA の塩基配列内に解析で得られた$\boldsymbol{u}_{l_2 j}$が含まれるような mRNA の総数を c、c の中で変異を起こしたことがある mRNA に限定して$\boldsymbol{u}_{l_2 j}$が含まれているものを d とする。この 4 つのパラメータからフィッシャーの正確確率検定で p 値を与える。

$$p = \frac{(a-b)!\,(b)!\,(a-c)!\,(c)!}{(a)!\,(a-b-c+d)!\,(c-d)!\,(b-d)!\,(d)!} \quad (6)$$

表 2:フィッシャーの正確確率検定

	解析データと一致しない	解析データと一致する	全データ
変異なし mRNA 数	a-b-c+d	c-d	a-b
変異あり mRNA 数	b-d	d	b
合計	a-c	c	a

4. 研究結果

塩基対について分割したため解析データは155個ある。この155個のうちフィッシャーの正確確率検定で24個のデータが棄却され残った。5%以下で棄却したため、155回の施行で24回も5%以下の施行が起きたこととなる。これは偶然ではありえないため、テンソル分解を用いた教師なし学習法によって、DSBが起こりやすい遺伝子を推測することができたと言える。

5. 結論

我が国において、1981年以来日本寺の死因ランキング一位は悪性腫瘍（がん）であり、国外を見ても特に先進国では悪性腫瘍による死因は多い。そのためがんの治療法を確立することは、とても重要なことである。そんな治療法を確立するためには、その病気について詳しい知識を得ることが必要不可欠である。本研究で注目したDSBは、悪性腫瘍の原因である遺伝子変異を起こす大本の一つと言える現象でそのDSBが起こりやすい遺伝子を推測することが、故意に起こされた薬品上でできたことから、実際のヒトにおいてもできると考えている。しかし、本研究ではmRNAを作る遺伝子に関しての推測という遺伝子の中での制限を設けて推測を行った。そのため、更に拡張し、ノンコーディングRNAについても解析し遺伝子全てに対して推測することが今後の課題だと考えている。

6. 参考文献

[1] Clouaire T, Rocher V, Lashgari A, Arnould C, Aguirrebengoa M, Biernacka A, et al. Comprehensive mapping of histone modifications at dna double-strand breaks deciphers repair pathway chromatin signatures. Molecular cell 72 (2018) 250-262.

[2] Y-hTaguchi. Tensor decomposition-based and principal-component-analysis-based unsupervised feature extraction applied to the gene expression and methylation profiles in the brains of social insects with multiple castes. BMC bioinformatics 19 (2018) 99.

[3] https://cancer.sanger.ac.uk/cosmic/download

[4] https://genome.ucsc.edu/cgi-bin/hgTables