月例卓話

数理と人工知能技術によるゲノム情報と化学情報の解析

阿久津　達　也*

1. はじめに

ヒトの設計図は基本的に ACGT の 4 種類からなる約 30 億文字の DNA 配列に書かれており，ヒトを始めとする各種生物の DNA 配列の解析技術は急速に進歩している．1990 年頃から米国を中心に多くの研究・教育機関の国際共同研究により実施されたヒトゲノム計画では，一人分の DNA 配列決定に十数年の年月と 3,000 億円程度以上の経費がかかった．しかしながら，現在では次世代シーケンサーという解析装置を用いることにより，ヒト一人分の DNA 配列が十日間かつ数十万円程度以下で決定できるようになってきている．これまでに数千種類以上の生物種の DNA 配列が決定し，数万人以上分のヒトゲノムが決定している（ただし，DNA 配列を決定できない箇所がほんの一部残されている）．DNA 配列の変異と疾患との関係が数多く報告されているため，遺伝子検査を行うベンチャー企業が出現し，また，遺伝子検査の結果により予防的な手術が行われることもある．今後，膨大なデータが蓄積し，それと関連データを解析することにより，生命の構築原理・動作原理の解明が進み，新たな治療法の開発にもつながることが期待される．

これらの膨大な DNA 配列データの解析にはコンピュータによる解析が不可欠である．また，化合物に関してもこれまでに数千万種類以上が報告されているため，その解析にもコンピュータの利用が不可欠となってきている．大量の生物情報データ，および，化学情報データを取り扱うためにバイオインフォマティクス，および，ケモインフォマティクスと呼ばれる分野が発展してきた．大量データの解析にはスパコンなどの大規模計算機が必要になるが，単に計算機パワーがあれば十分というわけではない．解析手法（アルゴリズム）も同様に重要である．単純なアルゴリズムを用いたのではスパコンを使って数億年かけても終わらない処理が，良いアルゴリズムを用いるとパソコンでも数分で完了してしまうこともある．良いアルゴリズムの開発のためには，対象となる問題の性質を数理的に解析することが有用な場合が多い．本稿ではそのような例を示す．

一方，近年，車の自動運転，将棋・囲碁のプロ棋士との対戦での勝利，クイズ番組での勝利などで，人工知能（AI）が話題になることが多い．人工知能技術は DNA 配列解析や化学情報解析にも数多く応用されている．現在の人工知能ブームの技術的牽引力の一つは深層学習（deep learning）である．本稿では深層学習の数理モデルであるニューラルネットワークについて，最も基本的な閾値関数に基づくモデルを紹介する．

2. DNA 配列決定のためのアルゴリズム

DNA 配列は次世代シーケンサーなどの解析装置を用いて決定されるが，現在の技術で直接，決定できるのは数十文字から数百文字分の断片

*京都大学化学研究所バイオインフォマティクスセンター教授

第 313 回京都化学者クラブ例会（平成 28 年 7 月 9 日）講演

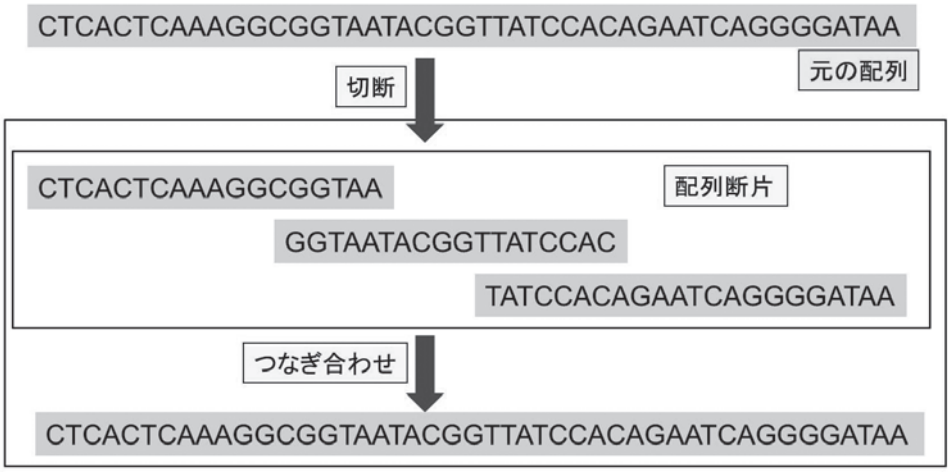

図 1. 断片のつなぎ合わせによる DNA 配列決定．なお，実際の配列のサイズや断片数ははるかに大規模である．

のみである．そこで現在，多くの場合に「もとの DNA 配列から数十文字程度からなる断片を多数作成し，それらの断片の配列を決定した後，コンピュータでつなぎ合わせる」という方法により配列決定を行っている．

ここでは，その最も基本的な定式化とアルゴリズムを紹介する[1)]．この問題は，同じ長さの DNA 配列断片の集合が入力され，「それらの断片のみがちょうど 1 回ずつ含まれる DNA 配列」を出力する問題として定式化される．もちろん，そのような性質を満たす配列が存在しない場合もあり，その際には「解なし」を出力する．例えば，断片の長さを 3 文字とし，ACA, ACT, CAC, CTG という断片集合が入力されたとすると，ACACTG は各断片のみをちょうと 1 回ずつ含むという性質を満たす．一方，ACA, ACT, CAC, CAG という断片集合が入力された場合には，そのような性質を満たす DNA 配列は存在しない．求める配列が存在かどうかは，もとの断片をすべての並び順（順列）を考え，断片をその順番で一文字ずつすらして配置して重なるかどうかを調べることにより判定できる．

例えば，前者の例では ACA, CAC, ACT, CTG という順列の場合を考えると，以下のように同じ列に同じ文字が重なるので，ACACTG という配列（正解）を得ることができる．

```
ACA
 CAC
  ACT
   CTG
```

一方，後者の例では，例えば，ACA, CAC, ACT, CAG という順列については

```
ACA
 CAC
  ACT
   CAG
```

となり，最後から 2 番目の列が T,A となり一致しない．でも，これは順列が悪いせいかもしれないので，すべての順列を試してみないと本当に正解がないかわからない．そこで全ての順列を試してみることにする．

もし，断片が N 個あったとすると，順列の

個数は $N!$（N の階乗）個あることになる．$N!$ は急速に増える恐ろしい数であり，$10! = 362880$, $20! \fallingdotseq 2.43 \times 10^{18}$, $30! \fallingdotseq 2.65 \times 10^{32}$, $40! \fallingdotseq 8.16 \times 10^{47}$, $50! \fallingdotseq 3.04 \times 10^{64}$ である．現在，日本最速のスパコンである「京」は 1 秒間に 1 京回，つまり，1×10^{16} 回の基本演算を行うことができる．よって，この「京」を用いて，1 個の順列のチェックが基本演算と同程度の時間でできたとしても，断片が 30 個の場合ですら，2.65 京秒（>8 億年）の計算時間が必要になってしまう．このようなすべての順列を試すアルゴリズムを用いたのでは，どんなスパコンを持ってきても無理である．

そこで用いるのが数学の力である．ここでは，まず，配列つなぎあわせの問題を，一筆書きの問題に変換する．例えば，CAG という断片があった場合，これを最初の 2 文字からなる CA という点と，最後の 2 文字からなる AG という点を結ぶ矢印に変換する．同様に，ACA は AC と CA を結ぶ矢印，ACT は AC と CT を結ぶ矢印，CTG は CT と TG を結ぶ矢印に変換する．すると，図 2（A）に示す図が得られる．ここで矢印は一方通行であり，楕円の点は大きさがないものとすると，この図は点線の順で一筆書きできる，つまり，すべての矢印を順方向に飛ばすことなしに順番にたどっていくことができる．この場合，ACA, CAC, ACT, CTG とたどることになるが，前に示したように，この順番に一文字ずつずらして重ねることができ，その結果として ACACTG という正解を得ることができる．一方，ACA, ACT, CAC, CAG という断片集合から同様に図を作ると，図 2（B）のようになる．この図は一筆書きができないので，正解がないことがわかる．

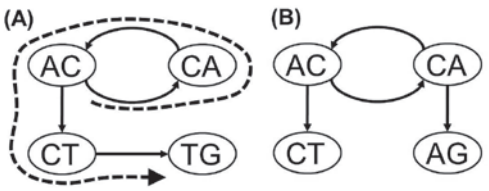

図 2. 配列のつなぎあわせ問題の一筆書きへの変換

もちろん，一筆書きができるかどうか，すべての順番を考えて試すとすると前と同様の問題が生じることになる．ところが，$e^{i\pi} = -1$ などの公式を発見した偉大な数学者であるオイラーは一筆書き可能かの簡単な判定法を発見した．それに基づくと，一筆書きできることは，ある点から他のすべての点に矢印をたどって行くことができ，かつ，次のいずれかの条件を満たすことと等価であることがわかる．

（1）すべての点において「入る矢印の個数＝出る矢印の個数」となっている

（2）1 個の点においては「入る矢印の個数＝出る矢印の個数－ 1」，別の 1 個の点においては「入る矢印の個数－ 1 ＝出る矢印の個数」，残りのすべての点については「入る矢印の個数＝出る矢印の個数」となっている

この条件であれば，それぞれの点を別々に調べていけば良いので，階乗の問題は発生しない．点の個数，すなはち，断片の個数が 1 億個あったとしても通常のパソコンで数分もあれば十分に計算が終了するはずであり，30 個なら一瞬である．また，説明は省くが「他のすべての点に行くことができるか」も簡単かつ高速にチェックすることができ，上の条件を利用して一筆書きの書き方（経路）も高速に計算することができる．

上で述べたアルゴリズムはエレガントで効率的であるが，次世代シーケンサーなどにより得られる配列断片は同じ長さというわけではなく，かつ，読み取り間違いも存在するので，実際の

データ解析に直接適用することはできない．そこで，このアルゴリズムのアイデアに基づき，様々な工学的な工夫がなされている．実際に役立つアルゴリズムの開発のためには数学に加えて工学も必要なのである．

3. 化合物設計支援

断片からの全体構造の推定という考え方は，化学構造の解析や設計にも古くから用いられている．分子式からの化学構造や異性体の数え上げは，原子という断片から化学構造を推定する問題の一種と考えることができ，数学者 Cayley の 1870 年代におけるアルカンの構造異性体の数え上げの研究に見られるように，コンピュータ発明以前より研究されている長い歴史をもつ問題である．スペクトルデータからの化学構造推定にも異性体の数え上げが応用されている．また，化学構造から化合物の活性を予測する構造活性相関（QSAR）は薬剤設計などに応用されているが，近年では指定された活性を持つ化学構造を見つけるという逆構造活性相関も研究されており，そこでも断片からの構造の数え上げが重要な役割りを果たす．

筆者は十年近くにわたり，京都大学数理工学専攻の永持教授らと共同で，分子式からの構造異性体の数え上げ，および，化学構造式からの立体異性体の数え上げという問題に取り組んできた[2]．この場合，一直線上に並んだ配列ではなく，枝分かれやループのある化学構造を扱うため問題が複雑になり，一筆書きを応用することはできない．しかしながら，様々な数理的手法を適用することができる．ただし，すべての化学構造を対象とすると効率化が困難であるので，環構造のない化学構造，もしくは，少数の環構造を持つ化学構造などを対象にして効率的なアルゴリズムを開発してきた．開発したアルゴリズムの一部を組み込んだ EnuMol システムを構築しており，その web server を公開している．このサーバー上で，分子式やその他の制約を入力すると，制約を満たす化学構造を数え上げ，それが画面上に表示される．EnuMol システムの実行例を図 3 に示す．

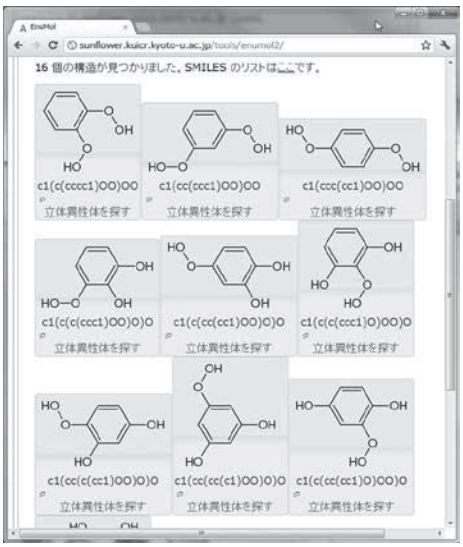

図 3．EnuMol サーバーの実行例

4. ニューラルネットワーク

人工知能は現在，第 3 次のブームを迎えている．以前のブームとは異なり，様々な分野で人間を超える結果をあげつつあるため，今度のブームはバブルではなく永続的なものかもしれない．この第 3 次人工知能ブームを技術面で支えるのが深層学習[3]である．深層学習では，神経細胞のモデルを多数結合したニューラルネットワークという数理モデルを用いて学習を行う．ニューラルネットワークも第 3 次のブームを迎えているが，以前と異なるのは多階層のネットワークを用いるという点である．1980 年代から 1990 年あたりにかけての第 2 時ブームの際に主に用いられたのは，入力層，中間層，

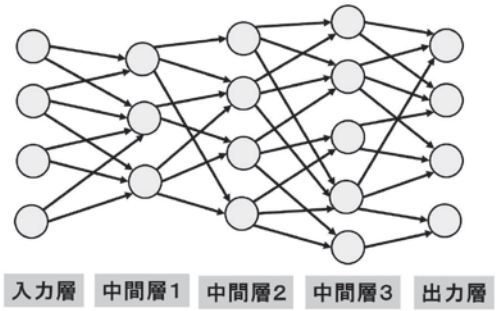

図 4. ニューラルネットワークの模式図（中間層が 3 層の場合の例）

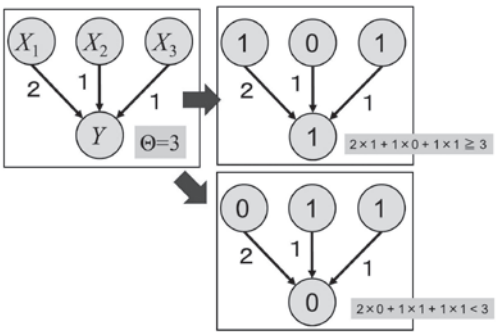

図 5. 神経細胞の数理モデル

出力層の 3 層からなるネットワークである．深層学習では 10 層近くから 20 層くらいからなる中間層を用いるのが特徴であり，深層とは多層であることを意味している（図 4）．もちろん，第 2 次ブーム以前から多層のモデルも研究されていたが，膨大な計算時間や学習がうまく進まないなどの問題があった．しかしながら，計算機の進歩，特に汎用グラフィックスプロセッサの利用による計算時間の大幅な短縮や，学習アルゴリズムの進歩などにより，多階層のネットワークに対しても効率的に学習が進むようになった．その結果，画像の認識や分類などにおいて，既存手法はもとより，人間を上回る認識結果が得られるようになってきた．この技術を DNA 配列データの解析や化合物の活性予測に用いようというのも自然な流れであり，いくつかの問題において既存手法より高い予測精度を得たと報告されている[4, 5]．このように深層学習は様々な問題に対して有効に適用されているがそれは様々な工学的な工夫によるものである．多層にすると何故良いのかは十分に理解できていないのが現状であり，数理的に解明していく必要がある．

神経細胞（ニューロン）の数理モデルは様々なものが提案されているが，最も単純なモデルは各ニューロンが 0（不活性）か 1（活性）のいずれかの状態をもち，そのニューロンへの入力の重み付き和がある閾値以上であれば 1，未満であれば 0 という規則に従って細胞の状態が更新されるというものである[6, 7]．例えば図 5 の例では，Y というニューロンは X_1, X_2, X_3 という 3 個のニューロンから入力を受けており，それぞれの重みは 2, 1, 1 であり，閾値は 3 である．ここで，X_1, X_2, X_3 が 1, 0, 1 という状態にあったとすると，重み付き和は $2 \times 1 + 1 \times 0 + 1 \times 1 \geqq 3$ となるので，ニューロン Y が活性化され $Y = 1$ となる．一方，X_1, X_2, X_3 が 0, 1, 1 という状態にあったとすると，重み付き和は $2 \times 0 + 1 \times 1 + 1 \times 1 < 3$ となるので，$Y = 0$ となる．

このようなニューロンモデルを組み合わせることにより，どのような計算ができるのかを理解すること，特にニューロンの個数，ネットワークの層数と計算可能な問題の関係を理解することは，深層学習の解明，ひいては，人間の知能の解明に役立つ可能性がある．そこで様々な研究が行われてきた．例えば，1 個のニューロンだけでも多様な知識を表現できることが知られている．YES/NO クイズ形式で対象物を探し当てるという知識の表現法は人工知能分野でも幅広く利用されており，その一つに決定木と呼ばれる表現法がある．さらに決定木の簡易

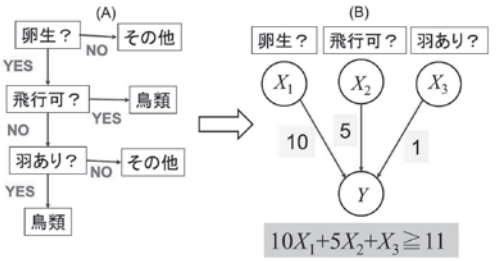

図 6. 決定リストとそれを表現するニューラルネットワーク

版として，直線状に YES/NO で回答していく決定リストというものがある．決定リストの例を図 6(A)に示すが，この決定リストは図 6(B)に示すように(入力部分を除いて)1 個のニューロンで表現することができる[6]．一方，決定木については何個のニューロンが必要十分かはわかっていないと思われる．その他にわかっていることとして，妥当な仮定のものとで，ニューラルネットワークに足し算をさせるには 1 個の中間層で必要十分であるが，掛け算については 2 個の中間層が必要十分であることがある[7]．当たり前のようにも見えるが，これらの結果は数理的にかなり深い解析によるものである．より複雑な計算問題についてはほとんど何もわかっていないのが現状である．筆者はバイオインフォマティクスにおける問題を対象にそのような研究を行っているが，まだ研究途上にある．

5. おわりに

本項では生物情報および化学情報の解析のための数理モデルと数理的手法について紹介してきた．配列データと化学構造データのみを対象としたが，生命の構築原理・動作原理の解明のためには，タンパク質立体構造データ，遺伝子発現データ，代謝反応データ，疾患データなど多種多様かつ大量のデータを組み合わせて解析していく必要がある．そのためには，新規のアルゴリズムや数理モデルを継続して開発する必要がある．また，アルゴリズムをどう工夫しても高速化が難しい問題もあり，そのような場合や膨大なデータの処理が必要になる場合は，やはりスパコンなどが必要になる．

ところで近年，バイオインフォマティクスのアルゴリズムに関して二つの大きな進歩があった．一つは配列比較に関するものである．類似配列のデータベース検索はバイオインフォマティクス研究者のみならず多くの生物学者によっても日常的に利用されている．類似検索の基本となるのは 2 個の配列の相同性比較であり，そのための基本技術が配列アラインメントと呼ばれるものである．配列アラインメント・アルゴリズムの基本版は 1970 年頃に開発されたが，その計算時間は配列の長さの 2 乗に比例するものであった．その理論的な高速化は多くの研究者が取り組んできたことであるが，ある妥当な仮定のもとで「2 乗より本質的に速くすることができない」ことが 2015 年に証明された[8]．一方，RNA 配列から RNA 立体構造における結合塩基対を予測する RNA 二次構造予測という問題について 1980 年頃に配列の長さの 3 乗に比例する計算時間のアルゴリズムが開発されていた．この高速化についても多くの研究が取り組み，筆者自身も取り組んだことがあったが，本質的に改善する結果は得られていなかった．しかしながら，2016 年に 3 乗の壁を破る 2.8244 乗の計算時間のアルゴリズムが開発された[9]．いずれもがバイオインフォマティクスにおける主要な未解決問題であり，前者は約 45 年ぶりに否定的に，後者は約 35 年ぶりに肯定的に解決されたのである．筆者は 30 年くらいにわたりバイオインフォマティクスの研究に携わってきたが，この二つは最大の驚きであった．バイオインフォマティクスは応用分野という考えの

研究者も多く，かつ，応用の重要性も理解できるが，理論的に興味深い問題も数多く存在し，このように数十年かけて解決される問題もある．今後の進展を期待するとともに微力ながらも貢献していきたい．

参考文献

1）Pevzner PA. 1-Tuple DNA sequencing: computer analysis. *J. Biomol. Struct. Dyn.*, **7**: 63–73, 1989.

2）Akutsu T, Nagamochi H. Comparison and enumeration of chemical graphs, *Comput. Struct. Biotechnol. J.*, **5**: e201302004, 2013.

3）岡谷貴之．深層学習．講談社，2015.

4）Alipanahi B, Delong A, Weirauch MT, FreyBJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotech.*, **33**: 831–838, 2015.

5）Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.*, **55**: 263–274, 2015.

6）Anthony M. *Discrete Mathematics of Neural Networks – Selected Topics*. SIAM, 2001.

7）Siu KY, Roychowdhury V, Kailath T. *Discrete Neural Computation - A Theoretical Foundation*. Prentice Hall, 1995.

8）Backurs A, Indyk P. Edit distance cannot be computed in strongly subquadratic time (unless SETH is false). Proc. 47th ACM Symp. Theory of Computing, 51–58, 2015.

9）Bringmanny K, Grandoniz F, Sahax B, Williams VV. Truly sub-cubic algorithms for language edit distance and RNA folding via fast bounded-difference min-plus product, Proc. 57th IEEE Symp. Foundations of Computer Science, in press.