連続講演会「東京で学ぶ 京大の知」シリーズ 16　社会に浸透する情報技術　第 2 回

ゲノム情報のコンピュータ解析

― 高校数学＋αによる先端的解析手法 ―

京都大学が東京・品川の「京都大学東京オフィス」で開く連続講演会「東京で学ぶ 京大の知」のシリーズ 16「社会に浸透する情報技術」。9 月 22 日の第 2 回講演では、化学研究所バイオインフォマティクスセンター長の阿久津達也 教授が「ゲノム情報のコンピュータ解析―高校数学＋αによる先端的解析手法―」と題して、ゲノム情報の解読に必要な情報解析手法の一端を紹介した。

●ゲノム情報とバイオインフォマティクス

「大学での研究は難しいものと思われがち。でも、高校数学＋αレベルでも先端的な研究ができるということを知ってほしい」と阿久津教授

近年、急速に進むゲノム解析において、重要な役割を担うバイオインフォマティクスを専門とするのが、化学研究所バイオインフォマティクスセンター長の阿久津達也教授である。

「バイオインフォマティクスとは、コンピュータの情報処理技術を、広く生命現象の解明に応用する学問。研究の鍵を握るのは、コンピュータでの計算の仕方、つまりアルゴリズムの工夫ですが、ここで必要なのは“高校数学＋α”レベルであることをご紹介したいと思います。その前にまず、ゲノム情報とはどういうものか見ていきましょう」

ゲノムとは、各生物の持つすべての遺伝情報のこと。DNA の塩基配列に情報として記録されているのだが、塩基はアデニン（A）、シトシン（C）、グアニン（G）、チミン（T）という 4 種類がある。ヒト一人の場合、文字数は 30 数億に及ぶ。

ヒトのゲノム情報が解明されたのは、1990 年頃から 2004 年にかけて進められたヒトゲノム計画による。これは、アメリカを中心にした国際共同研究で、約 13 年の年月と 3,000 億円程度の費用をかけて解読作業が進められた。

同計画が立ち上がって以降、バイオインフォマティクスという言葉が広く知られるようになる。バイオインフォマティクスとは、バイオ（生物）とインフォマティクス（情報学）が融合した学問であり、DNA 配列やタンパク質構造などをコンピュータで解析する方法の開発、コンピュータによる解析方法を用いた生物学的知識の発見という 2 つの目的がある。

技術の進歩はめざましく、今では、次世代シーケンサーという装置が解析速度を大きく向上させ、5 人分のゲノム解析を 10 日間、70 万円ほどの費用で行えるまでになっている。

「次々と解読された大量のゲノム情報は、医療に活用されています」

例えば、さまざまな疾患は DNA 配列の変異と関連があることが判明しており、BRCA1 という遺伝子に変異があると、将来、乳がんになる確率が高いことから、アメリカの女優が乳房切除手術を受けたことは記憶に新しいだろう。また、遺伝子検査も普及しており、数十から数百種類の遺伝子を調べ、病気リスクを提示する会社も増えている。

一方で、議論の対象となる課題もある。数年前は、ヒト遺伝子の 3 分の 1 程度が特許出願され、多くが認可されていた。しかし、アメリカの会社が乳がんと関連する遺伝子の特許を持ち、診断を独占していたことから訴訟が起こり、2013 年、米国最高裁は「遺伝子は特許対象ではない」との判決を下した。ただし、人工的に合成された遺伝子は特許対象との見解がなされており、今後も議論は続くと考えられる。

●人間の設計図をいかに解くか？

ヒトゲノムは、いわば人間の設計図。ここには、臓器のつくり方、脳のつくり方、顔のつくり方、知能、本能などがすべて書かれている。

「設計図を入手したにもかかわらず、設計図がどのように書かれているか、ほとんど分かっていません。心臓のつくり方はこの塩基配列、といったことは何ひとつ分かっていないのです。設計図は壮大なパズルであり、これを解くことが、21 世紀の重要研究課題です」

DNA 配列の 30 数億文字は、膨大な数に思えるが、実は CD-ROM1 枚少々に収まる量だ。最近のゲームソフトやビジネスソフトなどは、もはや CD-ROM1 枚に収まらない。

「驚くべき事実です。きっとそこには、数理的・情報学的原理があるはず。それを解明したい、というのが研究の原動力となっています」

人間の設計図という壮大なパズルを解くには、「疾患の種類による DNA 配列の違い」などのデータ収集が欠かせないが、一方で、大きなパズルを解くには、いくつもの小さな数

理的パズルを解くことも必要になる。

小さなパズルにこそ、「高校数学＋α」で解ける問題が多くある、と阿久津教授は言う。

●「配列のつなぎ合わせ」というパズル

30 数億文字もある DNA 配列を一度に解析することは、次世代シーケンサーをもってしても不可能である。

そこで行われるのが、長い DNA 配列を短く切ってから、それぞれをつなぎ合わせ、元の配列を推定していくという方法だ。

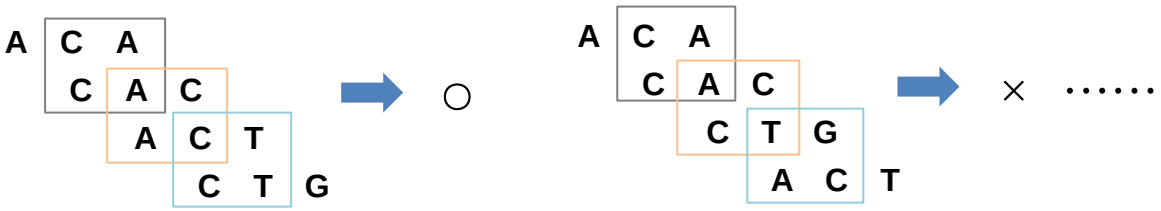
「わずか 3 文字の配列断片を例に、どのようにつなぎ合わせるのか見てみましょう」

＜問題＞ 【ア】「ACA」「CAC」「ACT」「CTG」 【イ】「ACA」「CAC」「ACT」「CAG」という 2 種類の配列断片がある。それぞれの配列断片が、ちょうど 1 回ずつ出てくるような配列はあるか。

この回答は、【ア】には「ACACTG」という配列があるが、【イ】には配列はない。

単純な解き方は、断片のすべての順列に対して、1 文字ずつずらして重なるかをチェックするという方法だ。

例えば、【ア】を調べる場合、次のようになる。



この方法だと、断片の個数の階乗通りを調べなければならない。例題の 4 の階乗（1×2×3×4＝24）なら問題はないが、DNA 配列の断片は数百万にも及ぶ。

スーパーコンピュータ「京」は、1 秒間に 1 京回の演算処理能力があるが、30 の階乗は「≒2.65×10^{32}」、1 京は 10^{16} だから 1 京の 1 京倍以上にもなり、「京」をもってしても計算は不能である。

では、どう解決するのか。「そこで登場するのが、数学の力です」

応用するのは「一筆書き」である。点と線から構成される図形を、一筆書きができるかどうか判定するにはどうすればいいか。1 つずつ確かめるとなると、階乗通り レベルの話と

なるが、数学者のオイラーが 1736 年、この問題を解決した。オイラーの定理は次の通りだ。

基本的に次のどちらかの条件を満たす時、一筆書きができる。

（a）どの点についても

・入ってくる矢印の数＝出て行く矢印の数

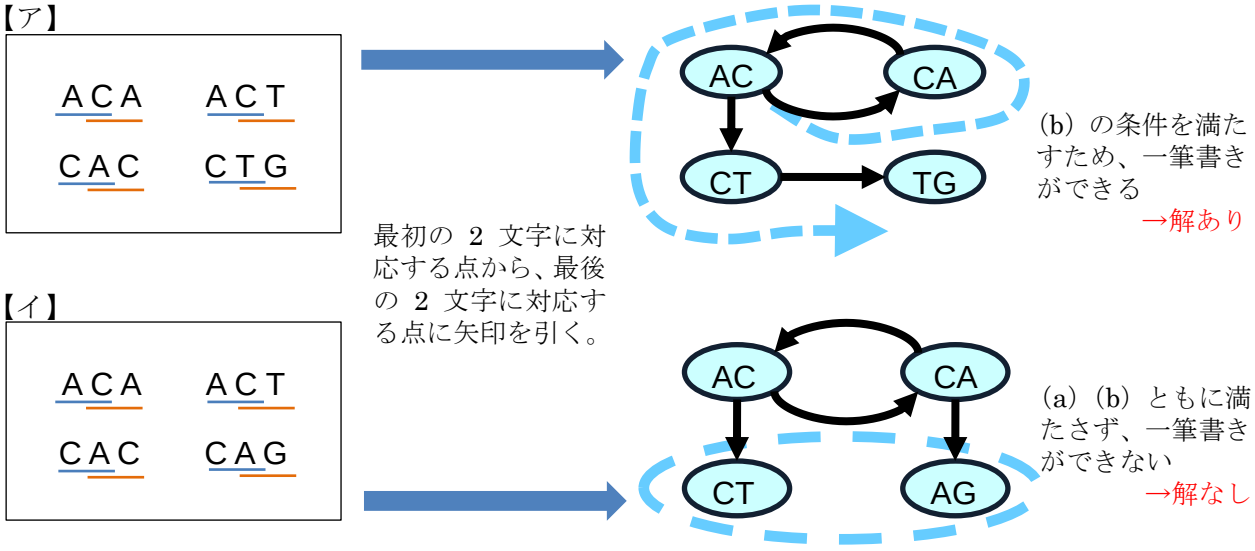
（b）2 点以外は上と同じで、残りの点は、それぞれ以下を満たす

・入ってくる矢印の数＝出て行く矢印の数－1

・入ってくる矢印の数－1＝出て行く矢印の数

この定理を使えば、各点について矢印の数を計算するだけなので、点が 1 億個になったとしてもコンピュータによる計算は可能である。

オイラーの定理を DNA 配列のつなぎ合わせにどう応用するか、前述の断片配列の例で見てみると、次のようになる。



「高校数学＋α レベルの原理で、スーパーコンピュータでも不可能だった計算が可能となるのです」

なお、阿久津教授は、異性体を数え上げる研究も行っている。

異性体とは、分子式は同じだが、原子の結合状態や立体配置が違うため、異なった性質を示す化合物のこと。原子の数が多くなるほど、異性体の数も膨大になるため、計算をいかに高速化するかが研究の目的だ。

「これも順列や組み合わせの応用ですから、基本的には高校数学＋α レベルです」と阿久津教授。

阿久津教授は、京都大学情報学研究科の永持教授との共同研究により、既存手法より高

速なアルゴリズムを開発。化学組成式を入力すれば、異性体が列挙される「EnuMol」システムをウェブ上で公開している。

●アルゴリズムの工夫＋スパコン利用

「前述のように、アルゴリズムの工夫により、処理の高速化が可能な一方で、どんなに工夫しても大幅な高速化が難しい問題が存在しています」

これは NP 困難問題と呼ばれるが､高速化が本当に不可能かどうかは重要な未解決問題となっている。多数の配列の同時比較も、NP 困難問題の 1 つである。

配列のつなぎ合わせにおいては、元のデータが大きく、大量の計算が必要である。しかも、1 回あたりの計算をどうしても効率化できない問題も存在する。「だからこそ、アルゴリズムの工夫とスパコンの利用の両方が大事なのです」

バイオインフォマティクスセンターでも、スーパーコンピュータを運営し、アルゴリズムの工夫と併用しながら研究を進めている。その 1 つの集大成が、「GenomeNet」という生命情報データベースだ。DNA 配列やタンパク質構造が検索できるゲノム統合データベース、化合物や疾患情報などを検索できる KEGG データベースなどを格納しており、世界中から数多くのアクセスがある。

「今回の副題は、高校数学＋α による先端的解析手法ですが、以前、学生から“先生は高校数学＋α しか使っていないと謙遜している”と言われたことがあります」

その時、阿久津教授が返した答えは、「謙遜しているのではなく、逆に、高校数学＋α しか使っていないけれど、新しく重要な分野を切り拓く先端的な研究をしている、ということなんだよ」。阿久津教授はそう講演を結んだ。

「ゲノム解析によるオーダーメイド医療の進展」､「遺伝子特許が医療や製薬にもたらす影響」､「遺伝子や社会的環境と、疾病との関係」など、参加者の質問は多岐にわたった