

理工学部物理学科／非線形物理学研究室  
バイオインフォマティクス、非線形物理学

# 田口 善弘 教授

【プロフィール】 田口 善弘（たぐち よしひろ）▷東京都生まれ。1988年、東京工業大学大学院理工学研究科物理学専攻博士後期課程修了。同年、東京工業大学理学部物理学科助手。1997年より中央大学理工学部物理学科助教、2006年より中央大学理工学部物理学科教授。著書に「砂時計の七不思議」「高校で教わらなかった物理」。



## 物理学の知識を「生命」の分野に応用する バイオインフォマティクスの領域で、 膨大なデータから明白な真実を導き出していく。

田口先生は、「自然現象を数学と結び付けて理論的に解析する学問は全て物理学」と言います。そのなかでも、先生の専門は「バイオインフォマティクス」と呼ばれる分野です。これは、対象を「生命」に求める「生命情報科学」とも称される分野です。「ゲノム（DNAのすべての遺伝子情報）科学」の進展によって、生物学に膨大な量の数値データがもたらされるようになりました。しかし、その貴重なデータを人間にとって有用になるよう解析する技術は実は意外に進んでいません。田口先生は、実験を繰り返すのではなく、既にある膨大なデータを解析することで、新たな薬の開発にも結び付く成果をあげていきます。深遠な生命の世界に、物理学の明晰さで挑む田口先生の研究ワールドに迫ります。

### 膨大なデータを手にしても 解析する技術がない現代

例えば文書作成ソフトのWordで文字を打つ場合を考えましょう。文字はパソコンのメモリーの中では0と1によるデジタル情報の組み合わせで蓄積されます。しかし後で0と1の組み合わせを見ても、実際にどんな文字が打たれたのかは分かりません。これと同じような状況が、生命科学の世界で起きています。

「新薬の開発では、まず何百万個もの化学物質を準備して、最初は（生体外で培養された）培養細胞で試し、次に動物実験で、最後は人間で臨床実験を繰り返し効く薬を見出していました。それでも、効く物質は見つかるのですが、時間も経費も大幅にかかってしまいます。遺伝子治療における研究も、やはり臨床を通してデ



▲田口先生のゼミには生物学に興味があって学ぶ学生も多い。同じ変数選択にチャレンジしているが、「バイオインフォマティクス」の性質上、「粘り強く取り組んでほしい」と願う。

ータから共通項を見つけていく作業が伴います。

ゲノム科学の発達によって、現在は、DNAからRNAに変わる過程の膨大な数値データが得られるようになってきました。したがって、このデータを解析できれば、創薬の効率は大幅に効率化できるのですが、解析する技術はなかなか向上しません」

### サンプル数の少ない アンケートのような現状

「仮に1万項目のアンケートを準備したとして、このアンケートを100人の対象者にしか配布しないなどというようなことは、まずしないでしょ。最低でも1万人以上に配布しないと期待した結果は得られません。

ゲノム科学の領域が発展し、得られる数値情報の種類だけは増加していますが、実際はこうしたアンケートと同じようなサンプル数が不足した状況が生まれているのです」

数値情報つまりアンケート結果は得られましたが、対象者の数が少ないなかで、何が病気の因子なのかを検出することは困難です。それは、ちょうど0と1の配列からは何も分からないパソコン上のWordの状況と同じです。脚光を浴びているゲノム科学ですが、解析技術が進んでいません。そして、まさにその分野をリードするのが田口先生なのです。自らを「自称理論物理学者」と言う田口先生は、生物である人間を研究対象にしていますが、理論的に物理学者の目で解析を進めます。

## 生命の重要なメカニズムを「変数選択」で解明する

物理学が専門の田口先生ですが、「バイオインフォマティクス」を研究される先生の研究を理解してもらうためには、もう少し生体について説明しなければなりません。

この分野で先生が研究対象としているのは、DNA から RNA に変わる際の「メチル化」です。

『メチル化』は、遺伝子を使うか使わないかを制御している印に、メチル基 (CH<sub>3</sub>) が結合するかしないかによって、その後の細胞分化に違いが生じる現象で、人間のような複雑な生物体を形成するためには重要なメカニズムです。具体的に言えば、髪の毛になったり骨になったりするのを決定する働きであり、逆にがんなどの疾患は、このメチル化のパターンに異常が生じて引き起こされることも多いです

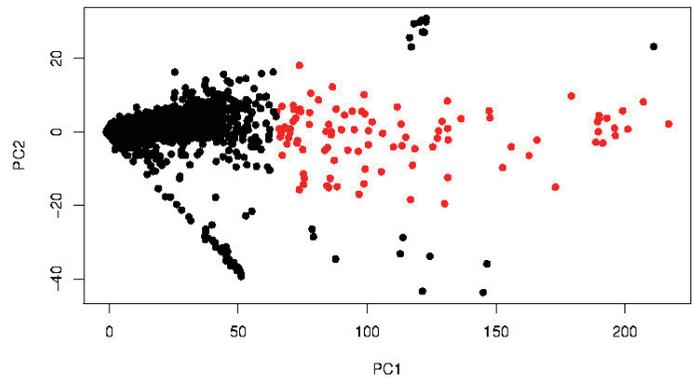
田口先生によれば、このメチル化が生まれる可能性がある場所が 50 万個あるのに、サンプルとなる数は 100 名程度。つまり、100 名の患者を調べるために 50 万個のアンケート項目を用意しているような状況が、ほぼ現実には起きていて言うのです。

「例えば 100 名を健康か病気かで二分するのに 50 万個のアンケート項目は必要ないわけで、理論的には 100 個のアンケート項目があればいいのです。しかし、先ほど述べたように、アンケート項目となる数値情報が膨大に生まれる状況のなかで、調べたい内容のために、どの部分を選択すればよいかを抽出する『変数選択』が重要視されてきています」

現在、サンプル 1 名で約 10 万円の経費がかかるとか。そうした費用的な限界もあるのですが、仮に 100 名を 1 万名に増やしても正確な情報が得られるわけではなく、膨大な数値データを前にした状況は殆ど変わりません。ゲノム科学が直面する、こうした事態を解決するために存在するのが、田口先生の「変数選択」の研究です。これは、人工知能の研究課題でもある「機械学習」の一分野で、データベースなどからある程度の数のサンプルデータを抽出して解析を行い、その結果から有用な法則や判断基準などを導き出す手法です。

## 医学分野での活躍も期待される解析

「例えばメチル化の可能性が 50 万個ある場合、サンプル数が 4 個ならば、1 つのメチル化に対し 4 つ座標があると考えられます。変数選択の過程では、これを 4 次元空間の中の点として一



▲培養細胞のセルラインによる「変数選択」の過程。黒色が病気に関係ない RNA、赤が病気に関係する RNA を示している。この分布で端にある RNA を比較することで膨大なデータから有効な RNA を抽出できる。

つ一つ見ていきます。そして、原点とは外れた端の部分に分布するメチル化サイト (メチル化される可能性があるゲノム上の場所) に注目します。原点近くのサイトは、メチル化の有無を表す組み合わせが平均化されてバラバラになっています。一方、端の方にあるメチル化サイトは、その組み合わせが規則的に並んでいて特長が明確です。また、それがグループとして集まっていれば偶然に生じたものではないことが分かるので、比較の見つけやすく変数として有効になります。例えば人込みの中で集団でダンスを踊っている場合、人数は少なくとも周囲とは全く違う共通のダンスをある程度固まって踊っていれば目立つのと同じことです」

先生は、こうした変数選択における今後の研究課題を次のように語りました。

「物理学は自然現象にある原理を見出す学問だと思うので、解析結果からさらに特長を抽出し、それがなぜ導き出されたかという理由を見つけ出すことが必要なのです。ただ、単に抽象化した理論では現実を語れませんので、解析結果に理由付けができるかどうか、課題と考えています」

また、田口先生は大阪市立大学大学院医学研究科肝胆膵内科学 (村上善基準教授) との共同研究も進めています。医師が臓器に針を刺して組織を採取する「生検」を行わずに、看護師でも可能な血液採取で取得した血中 RNA を用いた解析を行っているのです。そこでは、変数選択を使った解析手法が、医学の現場での実用化に向けて動いています。

さまざまなビッグデータが注目される時代にあって、田口先生の解析手法は、膨大なデータを社会に活かすために欠かせない、重要な研究になることは間違いありません。



◀物理学の視点から身近な現象を解き明かすなど、大人も子供も楽しめる内容の「砂時計の七不思議」など、田口先生独自の眼差しを感じさせる著書。

## Message ~受験生に向けて~

私は物理学を学びましたが、いまはその知識と能力を生物学の分野で使っています。遠回りに見えるかもしれませんが、真実を理解しようとするとは実は遠回りが近道だったりします。「今やりたいこと」もちろん大切ですが、「10年後、20年後にやりたいこと」をよく考えて、大学では、そのために本当に役に立つ学問を選んでほしいと思っています。それが「物理」であれば、一番うれしいです。