ヒトゲノム研究の展開

鈴木 裕太 SUZUKI Yuta（東京大学）

１はじめに

観測技術の継続的革新により牽引される近年のゲノム科学は、生命現象の深い理解のための基盤的領域としての地位を確保した。本論では、ゲノムの基礎研究で特筆すべき成果である「ヒトゲノムの完全解読」の技術的側面からの解説に続き、別の話題として医学応用における有力な新潮流である「多遺伝子性リスクスコア」を紹介する。

２. ゲノムの意義

「ゲノム」とは個体における遺伝情報の総体を指していう用語で、その物理的な実体は DNA（デオキシリボ核酸）、すなわちアデニン(A)、シトシン(C)、グアニン(G)、チミン(T) のいずれかの「塩基」を含む４種の単位が配列したポリマーである。生体の構造・機能において多様な役割を果たすタンパク質は 20 種のアミノ酸からなるポリマーで、このアミノ酸の配列は DNA の塩基配列に対応して精密に決定されている。ゲノムが生命の設計図とよばれるゆえんである。

生物を構成する全細胞は原則として同じゲノムを保持し、世代間で（ヒトの場合は両親から半分ずつ）継承される。細胞分裂を繰り返す過程で塩基配列に変化（変異）が蓄積し、個体ごとの多様性の一部を（環境要因とあいまって）形作ると同時に、長い時間をかけて種間の差異をも生み出している。

３. ヒトゲノムの解読

ヒトゲノムでは個人間に違いのある位置は約 1%未満であるので、平均的なゲノム配列を「参照ゲノム」として共有して変異情報などの整理に活用されている。最初の参照ゲノムは 2004 年の「ヒトゲノム計画」完了の成果として得られたが、配列未確定の領域（ギャップ）が多数残っていた。

ゲノム解読では DNA 配列（ヒトゲノムの総長は 30 億塩基対）はリードとよばれる短い断片（例えば 2,000 塩基対）としてバラバラに取得されるため、元の配列を復元する「ゲノムアセンブリ」という情報処理過程が必要となる。しかしヒトゲノムには「繰返し配列」が多く存在し、中には特に類似度が高く、繰返し回数も大きいために正確なゲノムアセンブリの困難な「難読領域」があり、これがギャップとして残されたのである。

４. ヒトゲノムの「完全解読」

2013 年の GRCh38 以来、約 10 年ぶりに更新された参照ゲノムである CHM13-T2T/hs1（2022 年）は、ついに難読領域をほぼ解決した、いわば完全版のゲノムである[1]。この達成には DNA シーケンサと呼ばれる DNA 解読装置の高性能化が決定的な役割を果たした。

まず、現在でもヒト集団の遺伝的多様性の分析や疾患研究で活用される、いわゆる次世代シーケンサ（NGS）は、従来の Sanger 法に基づくシーケンサと比べて圧倒的な大出力を誇りさまざまな生物種のゲノムが NGS により決定された。ただし技術特性としてリード長が短い(数百塩基対)ため繰返し配列の解析には不向きで、解読されたゲノム配列の品質も必ずしも高くはなかった。

この状況は第３世代シーケンサ（ロングリード技術）の登場で一変した。アメリカ西海岸の Pacific Biosciences（PacBio）とイギリスの Oxford Nanopore Technology (ONT)の２社が競争している。その名の通り長い断片長（数万塩基対以上）のリードが得られるので、繰返し配列を含む難読領域の解析に特に有用である。NGS と比べて高コスト・低出力で低精度という問題があったが、次第に改善されている。２社の比較では PacBio は高精度（>99.9%）、ONT は超長鎖（例えば 100 万塩基対）のデータが得られる点が特色である。

さらに DNA の３次元的な折り畳み構造（近接情報）を網羅的に取得できる Hi-C 法（本来はゲノムの複雑な発現制御を解析するための技術）も、再構築したゲノム配列の誤り検出に転用された。

ヒトゲノムの「完全解読」はこうした複数の技術の組合せにより実現された。ヒト以外の生物種についても同様の手法で高品質の参照ゲノムが続々と構築・報告されている。

５. ゲノム医学の展開

ゲノム科学と医学の接点は多岐にわたるが、ここでは大規模集団データを統計的に分析して疾患等の遺伝的基盤の解明を目指す手法を紹介する。

代表的手法の全ゲノム関連解析(Genome-wide Association Study, GWAS)は、疾患群と対照群それぞれの検体から変異情報を取得し疾患等の表現形質と相関する変異を絞り込む手法だが、一方で真の原因変異と疾患の機序を同定し新規治療法の開発を実現することは、当初の想定より困難であった。また典型的な GWAS は対象疾患別に設立された国際コンソーシアム等が実施してきたために取得したデータの統合・再活用のハードルも高い。

そこで近年は、数十万から百万人規模のデータを収集するイギリスの UK biobank やアメリカの All of Us などのバイオバンク（日本にはバイオバンク・ジャパン(BBJ)がある）のデータを活用した新たな展開が注目されていて、そのひとつが多遺伝子性リスクスコア(Polygenic Risk Score, PRS)の開発である[2]。こうしたデータベースに含まれる糖尿病やぜんそくなどの患者数の多い疾患は、GWAS が得意とする比較的重篤・少数の変異が原因なのではなく、むしろ個々の影響が微小な、多数の変異が関与しているという仮説に基づき、こうした多数の変異の寄与を総合して罹患リスクを個人別に定量するのが PRS の思想である。

相対的な罹患リスクの高低を正確に知ることは、適切な介入や行動変容を促すなど疾患予防に役立つと期待される。私見では、今後は疾患リスク予測のために個人の既往歴・生活習慣・健康診断での測定値など、ゲノム以外の医療・健康情報も取り込んだ機械学習(AI)手法の活用が本格化するとともに、個人情報・プライバシー保護の視点の重要性も高まってくると予想している。

６. ゲノムと多様性の問題

PRS は臨床応用や公衆衛生における有望なアプローチなのだが、実は（データが潤沢な）欧米人集団で構築された PRS モデルを別の集団に適用した場合に性能が低下してしまうという問題が知られている[3]。こうした状況を改善するため、現在では欧米人以外の集団を対象とした研究や、複数の集団を同時に扱うことで、不公正を解消し、ゲノムの多様性と普遍性を明示的に理解する試みが精力的に進められている[4]。

また従来の参照ゲノム配列も原則として欧米人のゲノム情報から構築されている。そのため南米・アジア・アフリカなど異なる地域に由来する個人・集団の解析に用いられた場合にミスリーディングな結果を生じる危険がある。ここでも、モダンなゲノム配列決定技術をフル活用して、ヒト集団全体の遺伝的多様性を包括的に表現したパンゲノム（Pan-genome, 汎ゲノム？）の構築が国際協力体制で進められている[5]。

以上の例が示すように、個人や集団にとって繊細な情報であるゲノム情報を公正に活用してゆくためには、ゲノム情報の特性に関して市民の正確な理解を増進することと併せて、研究者の側にもかかる負託に応えるための高い倫理性が求められている。「遺伝学」が「優生学」に転がり落ちていった歴史を繰返してしまう危険は無視できない。

引用文献

[1] Nurk et al. “The complete sequence of a human genome.” *Science*. Apr; 376(6588):44-53 (2022)

[2] Khera et al. “Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations.” *Nat. Genet*. 50: 1219-1224 (2018).

[3] Martin et al. “Clinical use of current polygenic risk scores may exacerbate health disparities.” *Nat. Genet.* 51:584-591. (2019).

[4] Sakaue et al. "A cross-population atlas of genetic associations for 220 human phenotypes." *Nat. Genet.* 53.10: 1415-1424. (2021).

[5] Liao et al. “A draft human pangenome reference.” *Nature* 617, 312-324 (2023).

基礎からの深層学習

丹治　寛樹、TANJI Hiroki（明治大学　兼任講師）

１．はじめに

近年、AI や機械学習の技術が急速に進化し、特に深層ニューラルネットワーク（DNN）を用いた機械学習の手法である深層学習[1]が各分野で革新をもたらしている。深層学習が多くの社会課題を解決し人々を惹きつける一方で、学習用データの収集による著作権侵害[2]や、データセンターの消費電力量[3]が問題視されている。そこで、本稿では、深層学習の技術研究の観点から深層学習の仕組みについて簡単な例とともに概説する。さらに、深層学習の技術的な性質から、なぜ AI 開発による社会問題が発生するか論じる。

２．深層学習

2.1　人工ニューロン

DNNは人間の脳内の神経細胞によって構成されるネットワーク（ニューラルネットワーク）から着想を得ている。人間の脳には 1000 億個程度の神経細胞が存在し、それぞれがシナプスを通じて情報を伝達している。神経細胞Aで発生した電気信号はシナプスを通じて神経細胞Bに伝わり、神経細胞Bは神経細胞Aを含め、他の神経細胞からも十分な電気信号を受け取ると、神経細胞Bの後段に繋がる神経細胞に電気信号を伝える。

神経細胞のはたらきを数式を用いて模擬したものを人工ニューロンと呼ぶ。図 1 に人工ニューロンを示す。x_iは他の人工ニューロンから受け取った信号、w_i は他の人工ニューロンとの結合の「強さ」を表す。ある人工ニューロンは受け取った信号x_iに重みw_iを掛けて溜め込み、溜め込んだ信号と閾値 θ を比較してyを出力する。図 1 において、f は活性化関数と呼ばれており、人工ニューロンに溜め込んだ信号と閾値 θ に基づいて出力yの値を決定する。

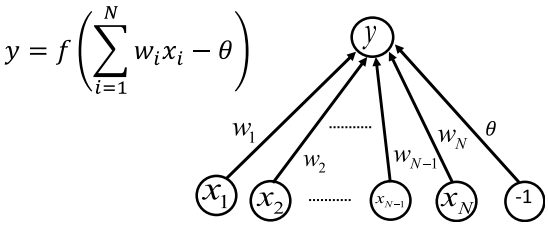

図 1：人工ニューロン

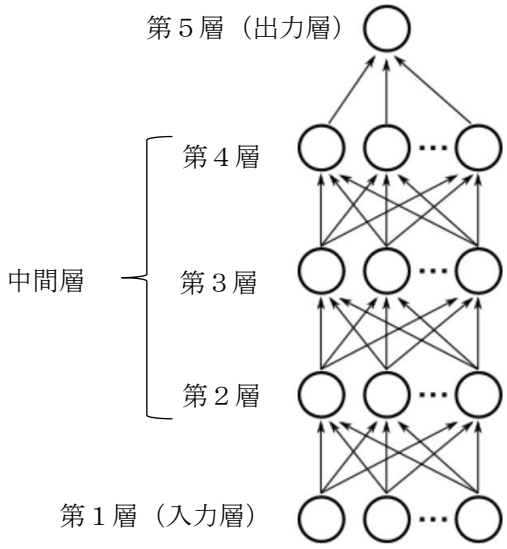

図 2：人工ニューラルネットワーク

2.2　深層ニューラルネットワーク

人工ニューロンを層状に結合することでニューラルネットワークを人工的に構築できる。図 2 に人工ニューラルネットワークの構造の例を示す。図 2 では、ネットワークへの入力を入力層の人工ニューロンが受け取り、第 2 層に伝達する。さらに、第 2 層の人工ニューロンは図 1 に基づいて次の層の人工ニューロンに情報を伝える。この手順を各層で繰り返すことで最終的な出力を決定する。人工ニューラルネットワークの中でも、層の数が多いものを深層ニューラルネットワーク（DNN）と呼ぶ。

例えば、簡単な例として、ある画像を入力すると、その画像に猫が写っているか判別する DNN を考えよう。この場合、DNN への入力は猫が写っているか判別したい画像となる。また、DNN の出力は猫が写っている確率とすることが一般的である。

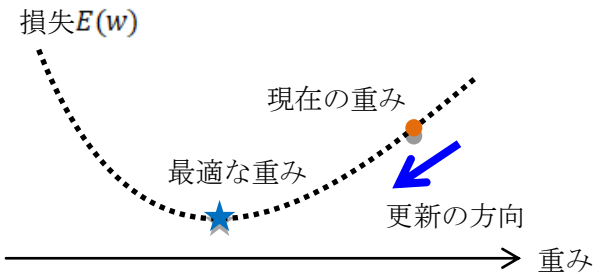

図 3：DNN の学習

2.3　DNN の学習

DNN への入力と DNN を構成する各人工ニューロンの重みおよび閾値が定まれば、DNN の出力が定まる。しかし、重みや閾値の値をいい加減に決めても意味のある出力は得られない。そのため、何らかの方針に基づき DNN の重みと閾値を調整する必要がある。DNN の重みと閾値を調整する一連の手続きを DNN の学習と呼ぶ。

例えば、前述の猫の画像を判別するDNNを学習するには、まず判別したい画像とその画像に対して猫が写っていれば 1、写っていなければ 0 を取るラベルを準備する。画像はDNNに入力され、DNNはその画像に猫が写っている確率を出力する。DNNを学習するには、DNNの出力と画像に付与されたラベルの乖離度を評価する。もし入力画像に猫が写っている場合、DNNの出力がなるべく 1 に近くなるように重みと閾値を調整する。

実際の DNN の学習では、猫が写っていない画像も含めて複数の画像を DNN に入力する。入力された全画像についての乖離度の総和を損失とし、DNN のある重みに着目して損失を求めると図 3 のようなグラフを描ける。DNN から適切な出力を得るには、丸で示した現在の重みの値が、星で示した損失を最小にするという意味で最適な値に近づくように重みを更新すれば良い。このように重み（および閾値）を調整することを最適化と呼び、DNN の学習は最適化アルゴリズムにより達成される。

３．人工ニューロン深層学習には何が必要か

最適化アルゴリズムにより DNN を学習しても、必ずしも DNN は知識を獲得できるとは言えない。2.3 節で説明した例において、ほんの数枚の画像を使って DNN を学習すると、DNN は学習に用いた画像に対しては正しい結果を出力するが、学習に用いなかった画像に対しては判定を誤るだろう。これでは知識を獲得したとは言えない。

一般に、深層学習の目的は DNN の重みと閾値の最適解を見つけることではなく、DNN へのあらゆる入力に対して何らかの意味で「正しい」出力を得ることである。この目的を達成するには、ありとあらゆるデータを DNN に入力し学習する必要がある。そのため、深層学習には、問題の複雑さや DNN の規模に応じた大量のデータが必要になる。また、大量のデータで学習するには、大量の演算をし、大量のデータと演算結果を保持するための大規模な計算資源が必要になる。特に画像生成のような複雑な問題ほど多くのデータと計算資源が要求される。

深層学習を含めた AI の研究で最も重要なことは、データの収集と計算資源の確保であり、大規模な AI ほどこの 2 つが少なくない社会問題を引き起こす。大量のデータを収集しようとすると、しばしば個人情報保護や著作権保護と矛盾する。また、大規模な計算資源を利用しようとすると、それを維持するために多くの電力を消費することになる。

４．まとめ

本稿では、深層学習の仕組みや技術的な性質について概説した。深層学習は研究や利用が進むほど、その技術的な性質から社会との矛盾を発生させる。人類社会と共生する AI を開発できるか、今後も議論を深めたい。

引用文献

[1] Ian Goodfellow 他「Deep learning」2016 年
[2] 文化庁「AI と著作権」
https://www.bunka.go.jp/seisaku/chosakuken/aiandcopyright.html （2024 年 9 月 29 日アクセス）
[3] JST 低炭素社会戦略センター「情報化社会の進展がエネルギー消費に与える影響 (Vol.2) -データセンター消費エネルギーの現状と将来予測および技術的課題-」2021 年 2 月

琉球列島で多発する樹木の新興病虫害

−その自然的・社会的誘因−

亀山　統一　KAMEYAMA Norikazu（沖縄支部・琉球大学）

１．マツ材線虫病（「松くい虫」被害）とは

マツ材線虫病（法定病害名は「松くい虫」被害）は、北米原産のマツノザイセンチュウを病原とする致死的な感染症である。病原は 20 世紀初頭に日本に持ち込まれ、日本在来のカミキリムシに媒介されてマツの幹枝に寄生加害する。

媒介者マツノマダラカミキリは、幼虫と蛹の時期を、枯れたばかりのマツの幹枝の樹皮下と材内で過ごす。羽化脱出した成虫は生きたマツの若枝の樹皮を摂食し、衰退・枯死直後のマツの幹枝に産卵する。マツノザイセンチュウは、枯死木から羽化脱出するカミキリ成虫に乗って移動し、若枝の摂食の傷口から感染する。侵入されたマツでは、センチュウを異物として感知して抵抗反応が起こり、周囲の細胞が死んでいく。センチュウが動くだけで、マツの樹皮・木部の組織が死に、木部の水分通導の機能が破壊されていくのである。水分通導阻害が蓄積したある時点で、地上部の水需要を満たせなくなり、全木が一気に枯死する。感染から数ヶ月で死に至る、激しい病害である。

材線虫病の病原と媒介者は 1960 年代に明らかにされ、急速に防除法が開発されていった。だから、適切な措置を講じれば、ある地域の病気の流行を収束させうる。防除法には次の 2 種類がある。(1)感染枯死木をカミキリの羽化前にことごとく伐倒焼却し、またマツ林の枝葉に殺虫剤を散布して飛来するカミキリを殺すことで、カミキリ密度を大きく下げること。(2)ぜひ守りたいマツの木には、事前に殺線虫剤を注射（樹幹注入）しておき、感染時にセンチュウを殺すこと。

しかし、伐倒駆除と樹幹注入は極めて高コストであるし、農薬散布は環境影響が大きい。そこで、ある地域に病害を侵入させない（カミキリが飛んでこない）ようにして、その対策をくぐり抜けた少数の感染木を確実に処理するのが最もよい。そのためには、守りたいマツの林や巨樹・並木のような文化財を特定して樹幹注入処理しておき、そのまわり数キロのマツを伐るなどして、極力マツがない緩衝地帯をつくり出す。緩衝地帯を越えて飛来したカミキリによる少数の被害木さえ除去すればよくなる。本土の文化財（天橋立や三保の松原など）はこうして守られているのである。

マツ材線虫病の研究は、センチュウのゲノム解読など、林木でありながら農作物の病害に匹敵するほど精緻な研究が進行してもいる。だが、適用可能な防除技術は極めて保守的である。

２．マツ材線虫病の侵入と流行の経緯

材線虫病が全く新しい場所に侵入するのは、センチュウとカミキリ幼虫が入ったマツ枯死木丸太などを人為的に移動した場合だけである。

戦前に日本本土に侵入した際には、例えば、佐世保軍港で初発し、軍港後背のマツ林でまん延して民間地に拡大し、戦争に伴う徴兵で労働力を失った農山村に大流行を引き起こすに至った。

沖縄では 1973 年に、公共事業受注業者が九州から被害材を（不法に）持ち込んで病害が初発した。民間地では初期防除が奏功したが、米軍基地内に拡大して防除が行われずまん延し、沖縄島に定着した。戦前の本土の轍を踏んだのである。

それからは、他島に拡大させない沖縄県の努力もあり、リュウキュウマツ資源は守られてきた。（残念ながら、鹿児島県下の奄美諸島では侵入・流行が繰り返された。）

３．久米島への材線虫病の侵入

だが、2021 年に久米島への材線虫病の侵入が明らかになった。久米島には樹木病害の専門家が在住せず、発見時には既に侵入後数年経過しており、初発の枯死木を全木駆除して撲滅させるとい

う初期対応がもうできない状況だった。発見の遅れは、初発場所が、警戒していた港湾、貯木場、建設現場でなく山中であったことも原因である。しかも、林業技術者の数が足りず、伐倒駆除など人的コストのかかる事業が十分に行えない。

そこで、保護すべき文化財、シンボルツリーへの樹幹注入処理を機敏に行い、それらの保全の見通しが立った。だが、それ以外のマツ林では敢えて激甚な流行を放置して、島内のマツ個体数・密度を急速に低下させて流行を早期収束させる戦術を採らざるを得なかった。また、被害跡地では、マツ稚樹が生育すると将来流行が再燃するので、広葉樹林に誘導する作業も今後必要だ。

久米島の教訓とは、流行病の侵入防止、初期防除、長期の防除戦略の遂行、流行後の森林管理の全局面で、専門家だけでは解決できず、住民の深い理解と協力が不可欠なことである。

４．続発する新興病虫害

(1) デイゴ　2000 年代初頭、沖縄県の県花デイゴ（インド原産のマメ科樹木、「島唄」の冒頭に歌われる）の葉と新梢を加害する害虫デイゴヒメコバチの被害が突如発生した。このハチの飛翔能力は小さいのに中国、台湾、琉球列島に同時に侵入しており、明らかに人為的な移動なのだが、侵入経路はつかめなかった。防除策は農薬か天敵生物の導入しかないから、生物多様性の高いこの地域での防除は困難である。しかも、枝先の被害だけでなく、突如大木の幹が腐って枯れる被害が続出し、後者はカビによる新興の流行病と分かった。2 つの病虫害が同時発生したのである。

(2) アカギ　2010 年代後半から、沖縄島以南に生える有用樹種であるアカギに突如ヨコバイ（昆虫）が大発生し、吸汁された葉が全て落ちて木が衰退する被害が起こっている。このヨコバイは中国原産と思われるが、現地では重大な被害を起こしていない。薬剤防除は可能だが、都市でも森林でも、農薬の使用には副次的な問題が伴う。

(3) ソテツ　2022 年に、奄美大島でソテツの外来カイガラムシ被害が確認された。直後に沖縄島北部でも同じ被害が確認された。ソテツにはすでに外来のシジミチョウによる深刻な食害が発生しているが、カイガラムシはさらに難防除性であり、防除のコストは大きい。ソテツは奄美で花卉として圃場で栽培されるほか、街路樹や庭園樹の植栽も多い。また、琉球列島では、最も土壌の薄い急傾斜の岩山の斜面を被覆する植物であり、世界遺産・国立公園において重要な景観をつくる樹種である。その生育場所へのアクセスは当然、極めて悪い。商業価値は大きくないがこの地域で極めて重要なソテツを誰がどう守るのだろうか。

(4) 南根腐病　熱帯・亜熱帯地域で極めて多様な樹木を侵して致死的な腐朽病（南根腐病）を起こすシマサルノコシカケは、琉球列島にもともといた種と思われるが、近年被害を広げている。防除が困難な上、マンゴーなど果樹にも被害が起こってしまった。ところが、日本には果樹の枝や根の病気の研究者が養成されていない。森林科学者の知見は農業の現場では活かせないのである。

５．新興病虫害防除のために必要なこと

新興病虫害の多くは外来の病原や媒介者による。在来の病害の場合も、何かの環境変化が流行の引き金となる。被害の初発は、1 ヶ所ないし少数の場所で、この異変を素早く察知して早期防除を行えば、仮に原因が不明でも撲滅も可能だ。しかし、そういう態勢をつくれないまま、たくさんの病虫害が次々発生している現状がある。

久米島の材線虫病、デイゴ、ソテツの被害は、侵入直後に発見して撲滅するのが難しかった事例である。南根腐病も在来の病害なので、撲滅は不可能である。すると、どうすれば流行初期のうちに防除策を実施できたかを顧みるべきである。

流行初期には、高度な防除手法が未開発であったり、現地の実施態勢が未確立だったりする。だから、病虫害の専門家、業者、行政の担当部署だけで技術的対応しても成功しない。実際に、行政が強い権限と人的資源を持つ中国も、材線虫病の流行拡大阻止に失敗している。ここでは、高度な技術を駆使した対応以上に、「島の森や緑をどうしていくか」という、持続可能な地域環境を構築する住民の自治的活動づくりが必須である。