分節と記号としての生命－ LLM による科学的理解の可能性

Segmentation and Life as Symbolic Code:
Exploring the Potential of LLMs for Scientific Understanding

イ・ソヒョン *
Seohyun Lee

1. はじめに

ウィキペディアによれば、「人工知能（AI: Artificial Intelligence）」という概念は、古代に遡る多くの神話や物語とともに始まったという [1]。古代科学を研究する Adrienne Mayor は、人工的な知性をもつ存在という発想自体が、紀元前 2700 年頃の神話や伝承の中にすでに描かれていると述べている。たとえば、ヘシオドスの『神統記（Theogony）』に登場する「パンドラの箱」は、鍛冶神ヘーパイストスによって造られた人工的存在として描かれており、これが人類における人工知能のイメージの源泉のひとつと解釈されている [2]。これらの神話的物語において、人工物はしばしば天界に属するものとして描かれ、それが人間と相互作用する時には、パンドラのように混乱や破壊をもたらす存在として現れる。しかし同時に、人工的な知性という発想が、すでに古代においても人間の想像力の射程にあったことを示している。

図 1.　箱を開けるパンドラ
（ChatGPT により生成）

このように、AI という概念は古代からその萌芽が見られるものの、実際に本格的な研究が始まったのは 1950 年代とされ、比較的近年のことである。とはいえ、現代は技術と社会の変

* 東京大学大学院情報学環・学際情報学府
キーワード：分節、AI 技術、大規模言語モデル（LLM）、エビジェネティクス、符号化

化があまりにも速く、「子どもを育てるには、もはや“子”ではなく“孫”を想定しなければならない」と言われるほどであり、1950 年代を「近年」と呼ぶのはためらわれるかもしれない。いずれにせよ、1956 年にダートマス大学で開かれたワークショップを出発点とする AI 研究は、当初アメリカ政府の期待を集めながら急速な発展が期待された。しかし、およそ 20 年後には、AI 研究の難しさを実感した研究者たちの間で失望が広がり、研究開発費の急激な減少とともに、いわゆる「AI の冬（AI Winter）」と呼ばれる停滞期を迎えることになる。

長く続いた「AI の冬」の終わりに、強力なハードウェアの進化を背景として、AI の「春」が 2000 年代中盤に訪れる。半導体の微細化に伴う高効率チップの開発や、汎用 GPU の革命が大きな転機となり、膨大な計算量を必要とする AI の実装が現実的なものとなった。いわゆる「AI ルネサンス」と呼ばれる時代の正確な始点は、学術的にはさらに前に遡るかもしれないが、多くの人々にとって記憶に残る象徴的な瞬間は、2016 年にソウルで行われたアルファ碁とイ・セドル九段による Google DeepMind チャレンジ・マッチだろう。1997 年にチェス世界王者を破った IBM の「ディープ・ブルー」の例があるとはいえ [3]、囲碁はその展開の複雑さから、AI が人間に太刀打ちするのは不可能だと長らく考えられていた。そうした中での「人類 vs 人工知能」という映画のような対決で、アルファ碁はイ・セドルに 4 勝 1 敗で勝利を収めた。この出来事は、まさに「AI 時代の本格的な到来」を世界中に印象づけた瞬間であった [4]。

その後の AI の進展は、学術界以上にむしろ日常生活の中で実感されている。機械学習、

図 2.　イ・セドル九段と AlphaGo による、2016 年の囲碁対局（ChatGPT により生成）

ディープラーニングという段階を経て、近年では「大規模言語モデル（LLM: Large Language Model）」[5] という言葉が注目を集めている。特に、トランスフォーマー（Transformer）構造を基盤とし、生成型モデルへと進化したGPT（Generative Pre-trained Transformer）アーキテクチャ [6] を用いた ChatGPT を皮切りに、さまざまな大手テック企業が生成型言語モデルを市場に投入している。登場初期には、その有用性や正確性に懐疑的だった人々も、いまや人間以上に「人間らしい」応答をする LLM に魅了され、有料モデルを利用するまでに至っている。数年前、東京大学の中央食堂がリニューアルされた直後、まだ Wi-Fi が設置されていなかった頃、ひとことカードに「この食堂には、人間の生存に必要な三大要素の一つが欠けている：水、空気、そして Wi-Fi」と書かれていたのを見たことがある。数年後には、こうした“現代版生存必須要素”に「人工知能」または「LLM」が加わる日が来るのかもしれない。

2. 魔法の仕組み：LLM の構造

ChatGPT をはじめとする大規模言語モデル（LLM）を使って、単なる翻訳から研究の解釈、さらには人生の悩みに至るまで相談したことのある人であれば、「高度に発達した科学技術は魔法と見分けがつかない（Any sufficiently advanced technology is indistinguishable from magic.）」というアーサー・C・クラークの言葉 [7] に、思わずうなずかざるを得ないだろう。筆者自身も、近年 LLM への依存度が急激に高まっていることを日々実感している。日常のちょっとした疑問からメールの言い回し、今後の研究の方向性にいたるまで、現代人が LLM に相談していることは、まるで紀元前のパルテノン神殿でアテナ女神から神託を受けていた古代人のそれと大差ない。LLM の答えが神託のように美しく感じられる瞬間もある。人間の素朴な問いかけに対して、どうしてここまで深く、思慮深い応答が可能なのか。LLM を崇拝する新興宗教が現れても、不思議ではないと思えるほどである。

図 3.　AI という現代の神託
（ChatGPT により生成）

しかし、LLM が魔法のように見えたとしても、それが本当に魔法のように現れたわけではなく、LLM の答えが神託のように響いたとしても、神託として生成されたわけではない。では、この“魔法”のようなふるまいの裏側には、いったいどのような科学的な仕組みがあるのだろうか。ここで、ChatGPT などの LLM の基盤となっている自然言語処理アルゴリズム、トランスフォーマー（Transformer）の構造を一度覗いてみる必要がある。

トランスフォーマーは、Google の深層学習研究者である Vaswani らによって 2017 年に発表された論文「Attention is All You Need（『注意こそがすべて』）」［8］で初めて提案された自然言語処理モデルである。トランスフォーマーの構造は大きくエンコーダ（encoder）とデコーダ（decoder）に分かれており、エンコーダでは入力文を単語やトークン（文を処理しやすい単位に分割した要素）ごとに分解し、それぞれの語が文中の他の語とどのように関連しているかを、「自己注意（self-attention）」という仕組みによって捉える。こうして、文中の各トークンが互いに与え合う意味の「重み」は、自己注意によって計算される。その結果、それぞれのトークンが持つ意味や文脈的な役割は、数値的な特徴のかたまり（＝ベクトル）として表現される。いわば、ひとつひとつの単語に「意味の地図」が与えられるようなもので、コンピュータはそれをもとに文全体の構造や流れを理解していくのである。このトランスフォーマーのエンコーダ部分のみを用いた代表的なモデルが BERT（Bidirectional Encoder Representations from Transformers）［9］であり、文脈の双方向的な理解に長けている。一方で、デコーダ部分のみを用いて構成されたのが GPT（Generative Pre-trained Transformer）［10］であり、前の単語列をもとに次に来る語を逐次的に生成するという特性をもつ。

たとえば、「私は朝コーヒーを飲んだ」という完全な文があったとする。BERT のようなエンコーダ専用モデルでは、このような文を学習する際、一部の単語（たとえば「コーヒー」）を意図的に［MASK］に置き換えて、「私は朝［MASK］を飲んだ」のように変換する。モデルはその文脈（「私は」「朝」「を」「飲んだ」）から、［MASK］の位置に入るべき語として「コーヒー」を推論するように学習されている。このように、文の前後両方の情報を同時に利用して、双方向的な理解を実現するのが BERT の特徴である。一方で、GPT のようなデコーダ専用モデルでは、入力文は「私は」「朝」「コーヒー」「を」の順で処理され、それぞれの単語が過去の単語にのみ注意を向けながら、次に来る語（たとえば「飲んだ」「淹れた」など）を予測する。つまり、GPT は自己回帰的（autoregressive）に動作し、「次の語を順に生成すること」に特化している。エンコーダを持たないとはいえ、GPT 内部でもトランスフォーマー層を通じて、入力されたトークン列は位置情報を含む表現ベクトルとして符号化（＝エンコード）されており、その点では内部的な“意味理解”のプロセスを持つといえる。つまり、私たちが LLM から“神託”のように受け取っているものの正体は、次に来る可能性が最も高い単語の並びである。そして、それこそがこの魔法の本質なのかもしれない。

3. 生命を読む AI：自然科学への応用

筆者はコンピュータサイエンスの専門家ではないが、物理学を専攻し、生命現象を扱う生物物理を研究分野とするなかで、さまざまな AI 技術を積極的に研究に導入しようと試みてきた。なかでも、機械学習を応用して生きている細胞の中で物質や情報を伝える“メッセンジャー”である小胞（vesicle）の動きを解析する研究と、LLM を用いた遺伝子発現の調節機構（エピジェネティクス）の解析に取り組んでいる。

ここでは、これまで取り組んできた研究のいくつかを紹介したい。まず、小胞が細胞内でどのように複雑な細胞骨格ネットワークと相互作用しながら移動するのかというパターンを、単純な機械学習モデルを用いて再現しようとした研究がある。小胞は、細胞外から取り込んだ物質（たとえば栄養素など）を包み込むようにして、細胞膜が内側にくぼむことで形成される。その後、球状の構造をもつ独立した小胞として、細胞骨格と相互作用しながら細胞内を移動する。この移動は、ダイニン（dynein）やキネシン（kinesin）と呼ばれるモータープロテインによって実現される。ここでいう細胞骨格（cytoskeleton）とは、細胞の形を維持するための骨組み構造であり、その代表的なものに微小管（microtubule）やアクチンフィラメント（actin filament）がある［11］。興味のある方は「motor protein walking」などで検索すると、微小管上でモータープロテインが小胞を運ぶ驚くべきプロセスを、3D で可視化した動画を YouTube 上で多数見ることができる［12］。

当時、機械学習の原理を学びはじめたばかりだった筆者が提案したのは、小胞の移動軌跡から得られる物理的な特徴―たとえば、移動速度、直線的な動きを示した総時間、直線的な移動間で測定された角度など―を主要な特徴量として用い、単純なサポートベクターマシン（SVM）を使って、小胞の移動がどの細胞骨格と関係しているか（たとえば、微小管からアクチンフィラメントへ、またはその逆など）を予測するモデルであった［13］。このような単純な機械学習の枠組みであっても、90％以上の予測精度を示したことから、当時「数理モデル派」として機械学習に懐疑的だった多くの工学系の知人とは対照的に、筆者はむしろ「無知は時に大胆である」という言葉どおり、機械学習を研究により積極的に応用するようになった。

正直に言えば、それ以降はある種の個人的な興味本位で、さまざまな AI 技術を研究に取り入れていった。当時は主に細胞の撮影画像データを解析対象としていたこともあり、画像処理に特化した畳み込みニューラルネットワーク（CNN: Convolutional Neural Network）［14］を用いた正常細胞とがん細胞の画像分類や、敵対的生成ネットワーク（GAN: Generative Adversarial Network）による細胞画像の生成といった研究を行った。

4. LLM と配列の物語：ゲノム生命科学の新たな読み方

筆者がLLMの世界に本格的に足を踏み入れたのは、2021年秋に定量科学研究所のComputational Genomics研究室で新たな研究を始めたことがきっかけである。当時の筆者は遺伝学に関してまったくの素人で、「ゲノム（Genome）」と「ジーン（Gene）」の違いすら理解していない状態だった（筆者注：Genomeは生物の全DNA配列、つまり全遺伝情報を指し、Geneはその中でも特定のタンパク質やRNAをコードする領域を意味する。ヒトゲノムにおいて、遺伝子に該当する部分は全体のわずか1～2%程度とされている）。一方で、AIやディープラーニングにはある程度触れてきた経験があったため、「AIを使って遺伝子研究をしてみたい」という気持ちだけが先走っていたとも言える。そうした中で、現在も研究面でお世話になっている中戸先生から提案されたアイデアが、「LLM、とりわけBERTを用いて遺伝子発現の制御機構（エピジェネティクス）を解析する」という研究テーマだった。

話が複雑になる前に、ここで関連する用語を一度整理しておきたい。まず「ゲノム（genome）」とは、人間の体を構成する一つひとつの細胞の核の中に存在する、すべての遺伝情報の総体を指す。この遺伝情報は、生物が生物として成り立つために必要な設計図のようなもので、ふつうは染色体（chromosome）という構造の中に折りたたまれ、非常にコンパクトな形で保存されている。この染色体は、細胞が必要に応じて遺伝子の情報を読み出すときにクロマチン（chromatin）というかたちでほどける。クロマチンを顕微鏡レベルで詳しく見ると、DNAがヒストンと呼ばれるタンパク質構造に巻きついている様子が見えてくる。DNAは、アデニン（A）、チミン（T）、グアニン（G）、シトシン（C）という4種類のヌクレオチド（塩基、リン酸、糖からなる化合物）で構成される二重らせん構造をしており、AはTと、CはGとしか結合しないという「塩基の対合則」に従って、もう片方の鎖の情報も自動的に決まるようになっている。このDNAの並び順（塩基配列）は、そのまま生命にとっての重要な情報となる。DNAの中でも「遺伝子」として機能する特定の配列は、RNAポリメラーゼと呼ばれる酵素によって読み取られ、RNAへと転写される。さらにこのRNAの情報をもとに、細胞内では特定のタンパク質が合成される。この一連の流れが、「遺伝子の発現（gene expression）」と呼ばれている［15］。

このとき、特定の遺伝子の塩基配列が認識され、読み取られることでその遺伝子が発現するということは、言い換えれば、塩基配列には、その遺伝子が発現する可能性や、発現を制御するための手がかりとなる情報が含まれていると言える。すなわち、ATATCGAGCT…といったヌクレオチドの配列パターンから、その遺伝子が発現する可能性や制御に関わる特徴をある程度読み取ることができるというわけだ。実際に、RNAポリメラーゼが特異的に結合する塩基配列モチーフの中には、「TATA」が繰り返されるような配列があり、これはTATAボックスと呼ばれている。TATAボックスは、転

写の開始点付近に見られるプロモーター領域の一部として知られており、遺伝子発現の制御において重要な役割を果たす。　このように、生物学的な機能と関連する特定の DNA 配列のことを、一般に「モチーフ（motif）」[16] と呼ぶ。先ほど述べた RNA ポリメラーゼが結合する TATA ボックスもモチーフの一例である。こうしたモチーフの発見は、遺伝子発現や制御メカニズムの理解において極めて重要な鍵となる。

これまで、モチーフの探索には、統計的な出現頻度の分析や、既知のモチーフとの比較、あるいは位置ごとのスコア行列（Position Weight Matrix: PWM）[17] などの手法が用いられてきた。しかし、A・T・G・C といったアルファベットで構成されていて、一見すると自然言語のようにも見えるのに、実際には「読めない」こうした塩基配列こそ、LLM を用いて“解読”するのに極めて相性が良い題材なのかもしれない。実際、LLM の一種である BERT を応用して DNA モチーフを見つけ出そうとする試みが DNABERT［18］というツールの目的であり、従来のツールと比較して、より高い精度でモチーフを検出できることが示されている。

とはいえ、LLM は本来「自然言語」を処理対象として設計されているため、そのままでは「読めない」DNA 配列を直接入力として扱うことはできない。そこで活用されるのが、「k-mer（ケーマー）」と呼ばれる手法である。これは、DNA 塩基配列を先頭から一定の長さ（k）で区切り、文字列の断片を「単語」のように扱うことで、配列全体を自然言語の文と同様のかたちに変換するものである。たとえば、k が 4（すなわち、4-mer）の場合、ATATCGAGCT… という配列は、ATAT、TATC、ATCG、TCGA、CGAG、GAGC、AGCT といったふうに、1 文字ずつずらしながら切り出された短い配列（トークン）の並びとして表現される。

このようにして、DNABERT では DNA 配列を「単語列」として再構成することで、自然言語における語のパターン認識と同様の処理が可能になるよう設計されている。自然言語の単

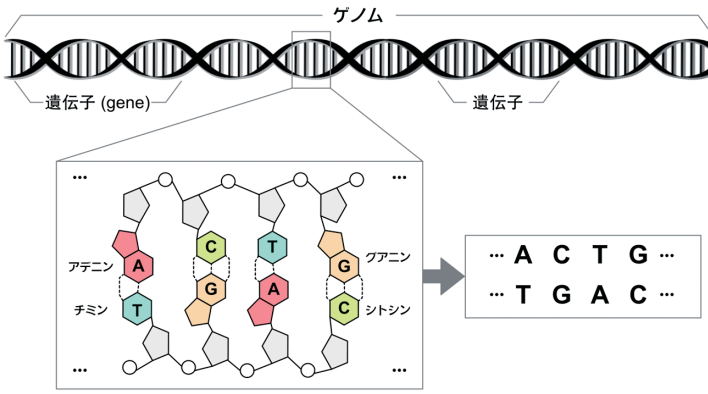

図 4.　DNA は 4 種類のヌクレオチドで構成され、塩基対によって二重らせんを形成する。その塩基配列は、意味のある並びとして読み取ることができるという点で、言語的な特徴を備えている（著者作成）。

語は、語彙としてある程度まとまった意味を持つが、DNA 配列においては明確な「区切り」や「文法」が存在しない。そのため、任意の長さの断片（k-mer）を滑らせながら切り出すことで、未知のモチーフや機能的パターンを網羅的に捉えることが可能になる。とくにモチーフは数文字程度の短い配列で構成されていることが多く、固定長の k-mer 化は、そうした局所的な特徴を捉えるのに適しているとされる［19］。

DNABERT のように、塩基配列をトークン化して学習させるという手法は、自然言語と DNA が「分節可能な記号列」として共通性を持つことを改めて示している。では、こうした発想を、より複雑で多層的な生物情報—たとえばエピジェネティクス（Epigenetics）［20］—にも応用することはできないだろうか。そうした問いから出発して筆者が取り組んでいるのが、クロマチン状態の配列を記号的に読み解く LLM モデル「ChromBERT」［21］の開発である。

ここで再び話が複雑になりすぎないよう、登場する新しい概念を整理しておこう。まず、「エピジェネティクス（epigenetics）」とは、ギリシャ語の epi-（上に、周囲に）と genetics（遺伝学）に由来する言葉で、DNA の塩基配列そのものを変えることなく、遺伝子の発現が行動や環境要因によってどのように変化するかを研究する分野である［22］。具体的には、DNA や、それに巻き付くヒストンというタンパク質に加えられる化学的修飾によって、どの遺伝子がオンまたはオフになるかという仕組みを扱う。たとえば、DNA メチル化やヒストン修飾といったメカニズムは、DNA 配列を変えずに遺伝子の発現状態に影響を与える典型的な例である。

なかでも「ヒストン修飾」とは、DNA に巻き付いているヒストンというタンパク質の一部（主に N 末端）に加えられる化学的な後修飾である。これによりクロマチンの構造が変化し、

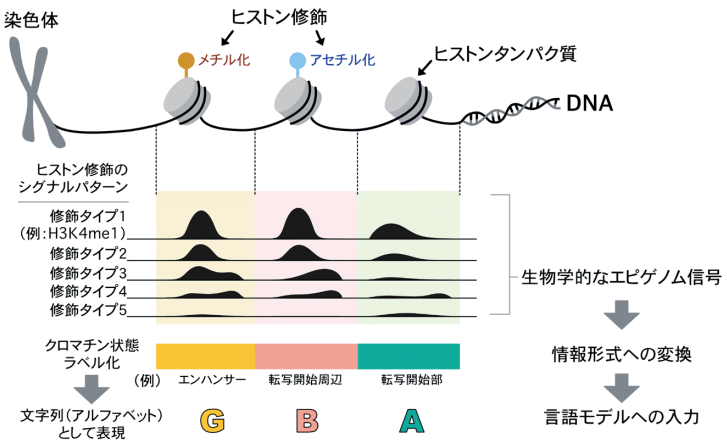

図 5.　ヒストン修飾に基づくクロマチン状態のラベリングと、言語モデルへの入力形式への変換の概略図。エピゲノム上のヒストン修飾パターンをもとに、エンハンサーや転写開始部といったクロマチン状態が分類・ラベリングされ、記号列（アルファベット）として表現される。これにより、生物学的な信号情報を言語モデルの入力データとして活用できるようになる（著者作成）。

遺伝子の発現が促進されたり抑制されたりする[23]。ヒストン修飾には、アセチル化、メチル化、リン酸化などがあり、それぞれ遺伝子のスイッチ（オン／オフ）を調節するエピジェネティックなマークとして機能する。たとえば「H3K4me3」と呼ばれる修飾（ヒストン H3 のリジン 4 番目が三重メチル化された状態）は、活発に転写されている遺伝子の開始点によく見られる。（筆者注：転写とは DNA が RNA ポリメラーゼによって読み取られ、メッセンジャー RNA が合成される過程）つまり、同じ DNA 配列をもつ遺伝子であっても、ヒストン修飾の違いによって発現状態が大きく変わることがある。

ヒストン修飾には非常に多くの種類があり、その効果も多様である。先述の H3K4me3 が遺伝子の活性化に関与する一方で、たとえば H3K27me1（ヒストン H3 のリジン 27 番目が単一メチル化された状態）は、遺伝子の不活性化と関連が深いことが知られている［24］。このように、それぞれのヒストン修飾が特定の生物学的機能と対応する場合もあるが、現実のゲノム上では、ひとつの領域に複数の修飾が重なって存在することが多い。そこで大規模なゲノムデータベースでは、複数種類のヒストン修飾の組み合わせをもとに、その領域の「クロマチン状態（chromatin state）」を定義するというアプローチが用いられている。この「クロマチン状態」は、使用するデータベースによって異なるが、一般的には 15 種類または 18 種類の状態に分類されている［25］。

このようなクロマチン状態が重要な遺伝情報とされる理由は、ゲノムが持つ遺伝子が発現するかどうかに関する情報を担っているためである。つまり、DNA の塩基配列に特定の遺伝子の情報が書き込まれていたとしても、その遺伝子が実際に発現するかどうかは、DNA 配列そのものからは判断できない。その判断には、DNA が巻き付いているヒストンタンパク質にどのような修飾が加えられているかというエピジェネティックなマークの情報を参照する必要がある。

たとえるなら、DNA が本に書かれた文字であるとすれば、クロマチン状態はその文字を音声として出力するスピーカーのオン・オフを決める装置のようなものだ。たとえ本に内容が書かれていても、スピーカーが読み上げなければ、その情報は表に現れない。つまり、遺伝子は存在していても、クロマチン状態によって「発現する／しない」が制御されているということである［26］。興味深いのは、このようなクロマチン状態などのエピジェネティックなマークは、DNA 配列とは異なり、常に固定されているわけではないという点である。環境要因によって変化したり、発生の過程で再プログラムされたりすることがあるのだ。こうした特性こそが、エピジェネティクスの研究をより魅力的なものにしている［27］。

固定された DNA の塩基配列からゲノム全体の情報を読み解くのではなく、ゲノム全域をクロマチン状態の情報として捉え、そこから生物学的な意味を抽出しようとする試みがある。筆者が開発した「ChromBERT」は、まさにその発想から生まれたものである。これは、LLM（大規模言語モデル）の一種である BERT を用いて、先に紹介した DNA モチーフ検出のように、「ク

ロマチン状態のモチーフ」―すなわち、特定の生物学的機能と関連づけられる状態パターン―をゲノム上から見つけ出すことを目的としている。

ChromBERT では、DNA 配列をトークン化する DNABERT と同様に、クロマチン状態の配列を「単語」のように分割して処理する。ただし、DNA が A・T・G・C のわずか 4 種類の塩基からなるのに対し、クロマチン状態は通常 15 種類（あるいは 18 種類）で定義されるため、A から O（18 種類の場合は A から R）までのアルファベットを借用する必要があり、トークンとしての語彙数は格段に多くなる。このようにしてトークン化されたクロマチン状態の系列をもとに、まず BERT アーキテクチャによる事前学習（pretraining）を行う。ここでは、文中の一部を［MASK］に置き換え、その前後の文脈から適切な状態を予測するという、自己教師あり学習の形式が用いられる。クロマチン状態の連なりから、次に来る状態や特徴的な組み合わせ（モチーフ）を予測することで、モデルは状態パターンの意味的な関係性を学習していく。その後のファインチューニング（fine-tuning）では、RNA-seq（筆者注：細胞内でどの遺伝子がどの程度発現しているかを定量的に測定する手法で、RNA を網羅的に読み取って発現量の違いを数値として捉えることができる）の発現データと結びつけて、あるプロモーター領域が「高発現の遺伝子に対応するか、または低発現の遺伝子に対応するか」を分類するタスクにモデルを適用した。これにより、学習済みの ChromBERT が、クロマチン状態の系列から遺伝子発現との関連性を予測する能力を持つことが検証された。

ここで、前述したように、G・B・A はいずれもヒストン修飾の組み合わせに基づいて定義されたエピジェネティック・マークであり、G はエンハンサー関連、B は活性化された転写開始点（TSS: Transcription Start Site）の周辺、A は活性型の転写開始点そのものを表している。クロマチン状態は、DNA とは異なり、ヒストン修飾が比較的広範なゲノム領域に影響を与えるため、同じ状態を示すアルファベットが連続して現れる傾向がある。そのため ChromBERT では、音声認識などでも用いられる「動的時間伸縮法（Dynamic Time Warping: DTW）」［28］を応用し、類似した配列のモチーフが同じクラスターにまとめられるように処理している。たとえば、「GGGBBAAA」「GBBBBA」「GGGGGBA」のように、G → B → A という順序構造を持つ配列は、DTW によって同一クラスターに分類され、「G-B-A」という代表的なモチーフとして抽出される。

話を総合すると、遺伝子が発現するかどうかといった「生物学的な特徴」が、DNA に巻き付くヒストン修飾の組み合わせという「化学的状態」として自然界に現れている情報を、アルファベットという本来は何の生物的意味も持たない記号に置き換えることで、LLM によってその中に潜むパターンを「言語」として読み解こうとする試みが ChromBERT であると言える。つまり、自然界の現象を記号化し、それを再び人工知能によってパターンとして可視化することによって、自然に内在する「意味」や「構造」をあらためて発見し直すプロセスがここにあるのだ。

5. 記号の設計図：分節化がもたらす理解

このように、自然界の情報を文字のかたちで記述し、パターンとして読み解こうとするプロセスにおいて、重要な前提となっているのが、「記号が分節化されている」という構造そのものである。人間の言語もまた、音や意味の連続体ではなく、単語、音素、文といった区切られた単位＝分節（segment）によって構成されている[29]。たとえば「私は朝コーヒーを飲んだ」という文は、「私は」「朝」「コーヒー」「を」「飲んだ」という言語的単位に分けられ、それぞれが意味や文法的機能を担う。こうした分節的構造があるからこそ、私たちは言語を分析・理解し、再構成することができるのである［30］。同様に、LLM もまた、トークンという離散的な単位に基づいて入力を受け取り、意味的な関連性を学習する仕組みで動いている［31］。つまり、記号が連続したままでは意味を捉えられず、あえて切り分けられているからこそ、構造や関係性が現れる。このように、「分けられている」こと自体が理解を可能にしているという視点は、自然科学に応用された LLM、すなわち ChromBERT のようなモデルの本質を捉えるうえで、自然科学と認知科学の交差点に位置する重要な着眼点となる。

人類の歴史を振り返ると、言語や記号を通じた情報伝達において「分節化された構造」が果たしてきた役割は計り知れない。最も原初的な例として、古代メソポタミア文明における楔形文字や、古代中国の甲骨文字のように、自然界の出来事や人間の行動を「区切られた記号」で記録する文化が現れたことは、情報を「単位ごとに切り出す」という思考が早くから存在していたことを示している［32］。これらの文字体系は、単なる描写ではなく、意味を分割・再構成するという知的処理の原点と見ることができる。このような「分けて記す」という発想は、後のアルファベットや音節文字、表意文字などにも受け継がれ、やがては文法というルール体系の発明につながっていく［33］。たとえば、古代ギリシャ語では、語を形態素（morpheme）に分け、それぞれの語尾や語幹に機能を持たせることによって、意味や文法構造を効率的に表現できるようになった［34］。言語は、連続的な音声や思考を、意味単位に分節することによって、共有可能な「情報」として抽象化することに成功した。この「切り分ける」という知的戦略がなければ、言語による知識伝達や記録は不可能であったと言っても過言ではない。

こうした分節的な言語表現が、人間の認知において重要な役割を果たしていることは、現代の認知科学でも指摘されている。人間は、言語を通じて提示された複数の情報を「カテゴリ」や「プロトタイプ」として統合し、そこから抽象的な概念やイメージを構築する能力を持つ［35］。たとえば、「鳥」という語を聞いたとき、私たちが即座に思い浮かべるのはペンギンでもダチョウでもなく、スズメやハトのような「典型的な鳥」であることが多い。これは、言語的な単位がもたらす意味が、脳内で「代表例」として再構成されるプロセスがあることを示している。

人間は、代表的なイメージを思い浮かべるだ

けでなく、さらに複雑で重層的な感情や経験をも、言語によって切り分け、定義しようとしてきた。言葉とは、曖昧で入り組んだ内面の状態を蒸留し、単語というかたちで「分節化」することによって、他者に伝達可能な情報として再構成する装置である。たとえば、ドイツ語の「Schadenfreude（他人の不幸から得る喜び）」[36] は、語りづらいが誰もが少なからず抱く感情を一語で表す。南アフリカの「Ubuntu」[37] は、「私は私たちである」という共鳴的な哲学を体現し、日本語の「侘び寂び」[38] は、不完全さや儚さの中に美を見出す感性を象徴する。こうした言葉は、複雑な情動の「重ね合わせ（エモーショナル・スーパーポジション）」を一つの単語に凝縮し、それを受け取った相手が、文化的・経験的な文脈の中でその意味を「解凍」することによって、理解が成立する [39]。つまり、人間は感情の微細な機微を言語化することで情報として切り出し、それを再び他者が「解読」することによって、相互理解というプロセスが可能になっているのである。

このような言語における「分節化」が感情や概念の理解を可能にしているように、ChromBERT の試みもまた、従来は「名前のない」複雑な生物学的シグナルを、言語的に取り出し、理解可能な単位に変換しようとする営みである。ヒストン修飾の組み合わせは計測可能であっても、その多様なパターンの意味はあまりにも重層的で、私たちはそれに明確な「名前」を与えることができずにいた。ゆえに、それらを一文字のアルファベットで「ラベリング」し、さらに LLM によってその連なりを「言語のように読む」ことは、単なる技術的な工夫ではなく、未定義なシグナルから意味を抽出するための認知的な飛躍でもある。このような「読み方」を可能にしているのが、私たち人間の言語理解を模倣する大規模言語モデルの役割である。名づけることすら困難だったシグナルのパターンに「ことば」としてのかたちを与え、それを意味のある情報として読み取り、優れた通訳者として他者へと届け、再びシグナルとして解読可能にする。

6. おわりに

本稿執筆の契機となったのは、ChromBERT の初稿に寄せられた、あるレビュアーによる、痛烈ながらも示唆に富んだ批判であった。「なぜ、明確に測定可能な生物学的シグナルを、あえて“言語風”に翻訳する必要があるのか。それはあまりに回りくどく、不要ではないか」—その問いは、査読の過程で繰り返し投げかけられた。もちろん、筆者はその問いに対して丁寧に答えたつもりである。査読への応答というかたちでは、定量的な評価や既存手法との比較、精度や解釈性の観点から、その「有用性」についてはすでに十分に説明した。けれども、どうしても心の奥にわだかまりのように残っていたのは、その問いに含まれていた「なぜ、言語にこだわるのか？」という、より根源的なレイヤーの部分だった。

そしてそれこそが、本稿であらためて丁寧に言葉を尽くしてみたかった動機である。生命現

象を意味として捉えるには、それを「切り分け」「記号化」し、「文脈のなかで読み解く」というプロセスが必要であり、まさにそこに、人間と言語、そしてLLMが持つ本質的な共通性がある。測定可能である、ということと、理解可能である、ということのあいだには、越えるべき知的な翻訳作業がある。その橋渡しの形こそ、「分節された記号列」であり、言語モデルはその記号列を扱うことに特化した、新しい“科学の読解者”なのだ。

人間は長い歴史のなかで、複雑な現象を理解するために、それを小さな単位に分け、記号として書きとめ、そして「意味」を読み取るという知的営みに取り組んできた。古代の文字体系も、現代の言語モデルも、その系譜に連なる存在である。本稿で紹介したChromBERTの試みは、その営みを生命科学の世界に持ち込み、記号化されたクロマチン状態の中に潜む意味を、あたかも「読む」かのように可視化しようとする挑戦だった。そこには、自然界の構造を言語的に捉えるという逆説的なアプローチによってこそ、初めて浮かび上がる「理解の形」がある。

参考文献

[1] ‘History of Artificial Intelligence’, in Wikipedia, 10 April 2025 <https://en.wikipedia.org/w/index.php?title=History_of_artificial_intelligence&oldid=1284831409>.

[2] ‘Ancient Myths Reveal Early Fantasies about Artificial Life’ <https://news.stanford.edu/stories/2019/02/ancient-myths-reveal-early-fantasies-artificial-life>［accessed 14 April 2025］.

[3] Wired. “IBM’s Deep Blue Beats Chess Champ Kasparov on May 11, 1997.” Wired, 11 May 2011. https://www.wired.com/2011/05/0511ibm-deep-blue-beats-chess-champ-kasparov/

[4] “AlphaGo,” Wikipedia, last modified June 30, 2024, accessed July 2, 2025, https://en.wikipedia.org/wiki/AlphaGo

[5] Zhao et al., “A Survey of Large Language Models,” arXiv（2023）

[6] IBM. “What Is GPT?” IBM Think Blog. Accessed July 2, 2025. https://www.ibm.com/think/topics/gpt

[7] Arthur C. Clarke, Profiles of the Future: An Inquiry into the Limits of the Possible, Millennium ed（Indigo, 2000）.

[8] Vaswani et al., “Attention Is All You Need,” arXiv（2017）

[9] Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv（2018）

[10] Radford et al., “Improving Language Understanding by Generative Pre-Training,” OpenAI Technical Report（2018）

[11] Alberts et al., Molecular Biology of the Cell, 4th ed., Garland Science, 2002, Chapter 16: “The Cytoskeleton”.

[12] XVIVO. Inner Life of the Cell Animation. YouTube video, Uploaded by XVIVO Scientific Animation, https://www.youtube.com/watch?v=wJyUtbn0O5Y（2002）

[13] Lee et al., “3D Nanoscale Tracking Data Analysis for Intracellular Organelle Movement using Machine Learning Approach,” IEEE ICAIIC（2019）

[14] Krizhevsky et al., “ImageNet Classification with Deep Convolutional Neural Networks,” In Advances in Neural Information Processing Systems 25（2012）: 1097–1105.

[15] Alberts et al., Molecular Biology of the Cell, 4th ed. Garland Science, 2002, Chapter 4: “DNA, Chromosomes, and Genomes”.

[16] Bailey, T. L. et al. “MEME: discovering and analyzing DNA and protein sequence motifs,” Nucleic Acids Research 34:W369–W373（2006）

[17] Xia et al., “Position Weight Matrix, Gibbs Sampler, and the Associated Significance Tests in Motif Characterization and Prediction,” Scientifica（2012）

[18] Ji et al., 'DNABERT: Pre-Trained Bidirectional Encoder Representations from Transformers Model for DNA-Language in Genome', ed. by Janet Kelso, Bioinformatics, 37.15 (2021) , pp. 2112–20, doi:10.1093/bioinformatics/btab083.
[19] Consens et al, "To Transformers and Beyond: Large Language Models for the Genome," arXiv (2023)
[20] Bannister et al., "Regulation of chromatin by histone modifications," Cell Research (2011)
[21] Lee et al., "ChromBERT: Uncovering Chromatin State Motifs in the Human Genome Using a BERT-based Approach," arXiv (2024)
[22] Weinhold, "Epigenetics: the science of change," Environmental Health Perspectives (2006)
[23] Kimura, "Histone modifications for human epigenome analysis," Journal of Human Genetics (2013)
[24] Peterson et al., "Histones and histone modifications," Current Biology (2004)
[25] Roadmap Epigenomics consortium, "Integrative analysis of 111 reference human epigenomes," Nature (2015)
[26] Callinan et al., "The emerging science of epigenomics," Human Molecular Genetics (2006)
[27] Alegría-Torres et al., "Epigenetics and Lifestyle," Epigenomics (2011)
[28] Dynamic Time Warping. In: Information Retrieval for Music and Motion. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74048-3_4 (2007)
[29] Liberman et al., "Perception of the speech code," Psychological Review, 74 (6) , 431–461 (1967)
[30] Saussure, F. de., "Course in General Linguistics," Ed. Charles Bally and Albert Sechehaye. Trans. Wade Baskin. McGraw-Hill, 1966 (originally 1916) .
[31] Brown et al., "Language Models are Few-Shot Learners," arXiv (2020)
[32] Gleb, I. J., "A Study of Writing," University of Chicago Press (1952)
[33] Pae et al., "The effects of writing systems and scripts on cognition and beyond: An introduction," Reading and Writing, Vol. 35, 1315-1321 (2022)
[34] Dionysius Thrax, "The Art of Grammar," c. 100 BCE. In: Kempson, R.M. (Ed.) Semantic Theory. Cambridge University Press (1977)
[35] Rosch, "Natural Categories," Cognitive Psychology 4 (3) , 328-350 (1973)
[36] "Schadenfreude." Encyclopaedia Britannica, https://www.britannica.com/topic/schadenfreude (accessed July 3, 2025) .
[37] Thompson, K., "What's in a Word: The Meaning of Ubuntu." Dandelion Philosophy, https://www.dandelionphilosophy.com/blog/whats-in-a-word-the-meaning-of-ubuntu (accessed July 3, 2025) .
[38] "Wabi-sabi." Wikipedia, https://en.wikipedia.org/wiki/Wabi-sabi (accessed July 3, 2025) .
[39] Barrett, "How Emotions Are Made: The Secret Life of the Brain," Houghton Mifflin Harcourt (2017)

SEOHYUN LEE (イ・ソヒョン)
[専門] バイオインフォマティクス・生物物理
[主たる著書・論文]
• Seohyun Lee and Ryuichiro Nakato, "Advances in Chromatin State Analysis Tools and Their Applications," JSBi Bioinformatics Review (in Japanese, review article) , 2025.
• Seohyun Lee, Che Lin, Chien-Yu Chen, and Ryuichiro Nakato, "ChromBERT: Uncovering Chromatin State Motifs in the Human Genome Using a BERT-based Approach," bioRxiv, 2024.
• Seohyun Lee, Hyuno Kim, Hideo Higuchi, and Masatoshi Ishikawa, "Visualization Method for the Cell-level Vesicle Transport Using Optical Flow and Diverging Colormap," Sensors, 2021.
[所属] 情報学環・学際情報学府
[所属学会] 日本生物物理学会・日本バイオインフォマティクス学会・日本分子生物学会・アメリカ光学会・米国生物物理学会

Segmentation and Life as Symbolic Code: Exploring the Potential of LLMs for Scientific Understanding

Seohyun Lee*

Segmentation, a fundamental structure in human language comprehension, also plays a central role in the architecture of artificial intelligence technologies, particularly large language models (LLMs). The linguistic mechanism by which continuous streams—such as speech or text—are divided into meaningful units closely mirrors the operational principles of LLMs. This structural resonance offers promising avenues for the analysis and reconstruction of scientific information. Drawing from the author's experience in epigenetics research, this article examines how biological signals—such as DNA sequences and chromatin states, represented as symbolic strings—can be segmented, encoded, and interpreted by language models, shedding light on emerging approaches to scientific knowledge discovery.

* Graduate School of Interdisciplinary Information Studies, Interfaculty Initiative in Information Studies, the University of Tokyo

Key Words：Segmentation, Language comprehension, Large Language Models (LLMs), Artificial Intelligence (AI), DNA sequences, Encoding.