バイオインフォマティクス領域におけるアニーリングアプリケーションの開発
～アニーリング手法による機械学習のアプリケーションフレームワーク～

1. 背景

生命科学では、ポストシーケンス解析として、ゲノミクス（遺伝子/DNA）、トランスクリプトミクス（転写物/RNA）、プロテオミクス（タンパク質）、メタボロミクス（代謝産物）とそれぞれの研究から、統合的なオーミクス解析（総合的な生物システムの解析）へと横断的で複合的な広い領域を対象としたバイオ解析が必要となってきている。

そこで、バイオインフォマティクスでも機械学習の利用の期待が高まり、代謝パスウェイ、遺伝子ネットワーク推定、タンパク質間相互作用ネットワーク解析や、ケモインフォマティクス分野として薬物・標的タンパク質間相互作用の予測やその特徴抽出に、機械学習が利用され始めており、これから更に「機械学習」の活用が広がる可能性が高い。

一方で、物理学を基礎とした情報処理として注目されている「アニーリング手法」は、機械学習の分野でもその活用が期待されている。しかし、アニーリング手法による機械学習は、バイオインフォマティクスでは未開拓であり、今後、生命科学分野における機械学習の手法の活用とともに、研究に利用されることも考えられる。

バイオインフォマティクスにおいては、妥当性が検証されており、かつ利用しやすいソフトウェアが多数存在するため、多くの場合、生命科学研究者にとって汎用性の高いアニーリング技術を用いたアプリケーションを開発する動機はほとんど無い。そのため、アニーリング手法による機械学習をバイオインフォマティクス領域で活用しようとしても、すでに利用されているソフトウェアと同等ないしはそれ以上に簡単に利用できるものでなければ、その利用は促進されない。

したがって、アニーリング手法を用いた機械学習をバイオインフォマティクス分野に浸透させるためには、一般の生命科学研究者が利用しやすいフレームワークを開発し、提供することが重要となる。

2. 目的

アニーリング手法による機械学習のアプリケーションフレームワークを開発することにより、生命科学分野で機械学習を利用するケースにおいて、アニーリング手法による機械学習の利用を促進することが目的である。その目的達成のために、(1)R 言語によるフレームワーク実装、(2)そのフレームワークの妥当性検証を行う。

(1) R 言語によるフレームワーク実装

バイオインフォマティクスでは、R 言語で多くの研究がされている。（全体の半数以上と言われている。）そして、その R 言語には、生命科学研究者向けのその分野独自の専

用ライブラリ群がまとめられたパッケージプラットフォーム（Bioconductor）もあり、1649 software packages（2019 年 3 月 7 日現在）が登録されている。生命科学研究では、この中のパッケージが使われていることが多く、R 言語で生命科学関連のプラットフォームを開発するのは、バイオインフォマティクスにおけるある種のデファクト・スタンダードになっている。

そこで、生命科学関連のフレームワークとしてアニーリング手法の機械学習を R 言語で開発するのは重要である。R 言語で実装したフレームワークが完成すれば、バイオインフォマティクス分野で用いられている既存のソフトウェアやライブラリとの連携も可能となる。そして、このフレームワークを使うことにより、生命科学分野の研究で、アニーリング手法を用いた機械学習の利用を促進することが可能となる。本プロジェクトでは生命科学研究者向けのR言語による機械学習フレームワークを開発することが目的である。

⑵ フレームワークの妥当性検証

次に、R 言語による機械学習フレームワークを開発したときの適用を考える必要がある。そこで、このフレームワークの妥当性検証として、具体的な生命科学の課題を取り上げ、フレームワークが正しく動作するかを確認する。

その具体的な課題として、「TF-DNA 結合問題」を扱う。

「TF-DNA 結合問題」は、アニーリング手法による機械学習の対象として、R. Li et al., npj Quantum Information 4,14, (2018)の先行研究があり、妥当性検証の結果とも比較しやすい。

セントラルドグマ（「遺伝子情報として DNA の塩基配列があり、その遺伝子が発現して RNA に転写され、さらに RNA からタンパク質が合成される」という生命科学の中心的な事象の命題）は、細分化された領域があり、様々なアプローチで研究されている。近年、分子生物学においては、1分子レベルでの化学反応や拡散などの物理現象がよく研究されており、セントラルドグマにおいても遺伝子発現の動的な事象については分子レベルでの研究が進んでいる。特に「DNA とタンパク質の結合」や「RNA とタンパク質の結合」は、その生化学反応や構造解析が盛んに行われている。この先進的な研究の1つである「TF（転写因子タンパク質）-DNA 結合問題」をケーススタディとして、実行可能性の検証を行うことにより、⑴で開発する R 言語によるフレームワーク実装が動作することを示すことが目的である。

3. ソフトウェア開発内容

「アニーリング手法による機械学習」が、新たに「バイオインフォマティクス」分野において利用されるように「フレームワークとして R 言語の実装」を行った。

開発したソフトウェア:

塩基配列＋実験値データから特徴量を学習し、未知のデータにおいてその塩基配列を推論するアプリケーションフレームワーク

主な機能:

・塩基配列データのエンコーディング・デコーディング
・アニーリングマシンへの API 実行
・訓練データによる学習、特徴抽出
・評価データの推論

（処理概要）

【データ処理】

生命科学データベースのデータフォーマット（gcPBM と HT-SELEX）に即してデータを読み込む。

【DNA エンコーディング・デコーディング処理】

DNA 文字列をデータフォーマットに応じて、エンコーディング、デコーディングする。
エンコーディングされた DNA 文字列をイジングモデルに変換する。

【学習プロセス】

学習用の教師データを使って学習し、その特徴量を抽出する。
アニーリング処理は、あらかじめ準備したソルバーを指定して処理する。

【推論プロセス】

学習プロセスで学習したモデルを指定して、入力データに対して推論を行う。

（処理フローの概要）

各処理は次に示すフロー（図 1）で処理される。

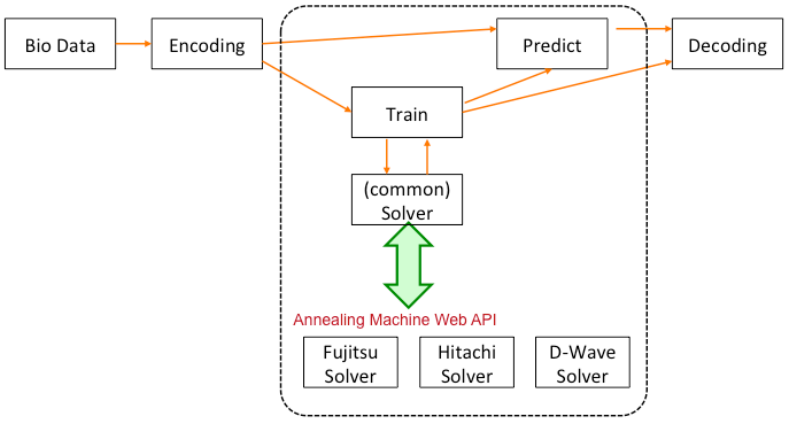

図 1：開発したソフトウェアの処理フロー概念図

（プラグイン機能）

アニーリング処理は、各ベンダーマシンの Web API を利用して実行される。その機能は、フレームワークのブラグインとして提供する。それぞれのベンダーマシンが必要なときに、そのプラグインを個別にインストールすることで、利用可能となる構造とする。

（実行可能性の検証）

富士通デジタルアニーラにおいて、TF-DNA 結合問題の特徴量抽出ができることを確認した。

（実行画面は以下の図 2）

右側が本プロジェクトの開始時に、参考にした R. Li et al., npj Quantum Information 4,14, (2018) から引用した特徴量であり、実行画面は学習イテレーションごとに抽出された特徴量である。黄色枠の部分が、R. Li et al. の結果と対応づけられた塩基配列である。イテレーションを総合すると、R. Li et al.の結果とも対応づけられることがわかる。

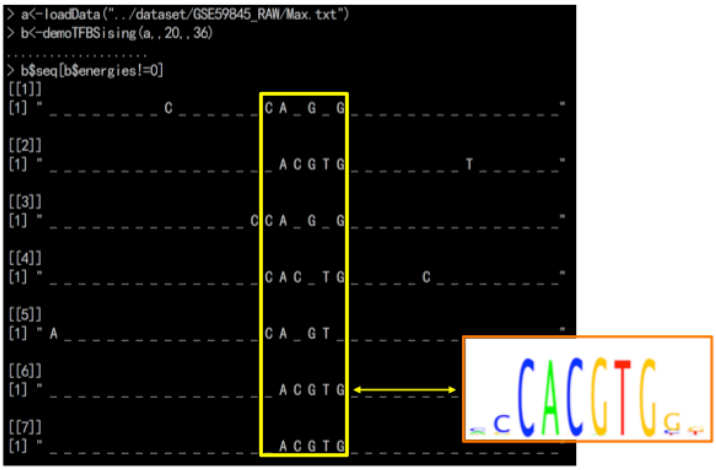

図 2:開発したソフトウェアの実行画面と比較データ

4. 新規性・優位性

アニーリング手法による機械学習は、Python などの R 言語以外では実装の例があるが、R 言語による単純なアニーリング計算の良い実装もなく、アニーリング手法による機械学習の R 言語の実装そのものが新しいため、生命科学分野で利用する用途としては、従来にはないアプリケーションのフレームワークである。

5. 期待されるユーザー価値と社会へのインパクト

生命科学では、生体分子の化学結合に関わる問題や、膨大なデータの中からパターン抽出する問題など至る所に最適化問題が存在する。また、「バイオインフォマティクスは、機械学習をゲノム科学に応用した科学である」と言われるほどに、機械学習の手法が重要であると期待されている。その状況で、生命科学へのアニーリング手法の適用はほとんど行われておらず、バイオインフォマティクスですでに使われているソフトウェアと接続容易なフレームも存在しない。そこで、このフレームワークを開発し提供することは、アニーリング手法を用いたバイオインフォマティクス分野における研究促進にもつながり、重要と考える。

本アプリケーションは、R 言語の標準のインストーラーでインストールすることができ、利用にあたり、一般の生命科学研究者も容易に利用できることを目指して開発した。これにより、バイオインフォマティクスにおける機械学習の利用の際には、アニーリング手法が1つの計算手法の候補に挙げられることが期待される。今後、更に妥当性検証を重ねていき、汎用性を高め、オープンソースとして公開したい。更に実績を積んだフレームワークとして、生命科学分野のライブラリとして認められ、Bioconductor のパッケージ管理システ

ムに取り込まれることを目指している。

6. 氏名（所属）
山崎清仁（有限会社ジェイズコア）