



アルゴリズムの頂が持つ力はいかに？ — ゲノコン2021開催記録—

♡ 16



情報処理学会・学会誌「情報処理」
2022年4月15日 08:56



清水 佳奈（早稲田大学）



坂本 一憲（WillBooster（株）／早稲田大学）



笠原 雅弘（東京大学）

筆者らは2021年夏にヒトゲノム配列解析を題材とした競技プログラミングコンテスト，ゲノコン2021を主催した．本稿では開催の背景と結果について報告する．

▼ 目次

競技プログラミング

マラソン方式

競技プログラミングとゲノム配列解析

ゲノコン2021の開催

実問題！ 実データ！ 未公開データ！

目を見張る競技者たちの実力，優勝は非専門家！

コンテストを盛り上げた競技者間の助け合い

実解析を出題する難しさ

アルゴリズムの頂と科学研究の未来

競技プログラミング

競技プログラミング（通称「競プロ」）とは，アルゴリズムの設計力と実装力を競う競技である．競プロのコンテストはさまざまな形態をとり得るが，オンラインジャッジシステムによるコンテストが主流だ．一般的なオンラインジャッジシステムでは，出題者が用意したジャッジサーバーに競技者がプログラムを提出する．ジャッジサーバーは，提出されたプログラムに対してあらかじめ用意した入力を与え，出力結果と正解（期待する出力）を照合して自動で得点を与える．それでは競プロではどのような問題が出題されるのか？ ここでは，ゲノコン2021のB問題を紹介しよう．

s と t は文字A, C, G, Tからなる文字列である． s と t の i 番目の文字はそれぞれ， $s[i]$, $t[i]$ と記述する．ここで， s と t の適当な位置に空白を表す文字 '-' を挿入して長さをそろえ，スコア $S = \sum_{i=1}^N \text{sim}(s[i], t[i])$ を

最大化することを考える. $\text{sim}(x, y)$ は x と y の類似度を示す関数である. $x = y$ の時は 1 , $x \neq y$ の時は -3 , x と y のいずれかが '-' の場合は -5 を出力する. スコア S を最大化するように '-' が挿入された s と t を出力せよ.

(紙面の都合上, 問題文を一部改変している. 詳しくは[AtCoderの問題ページ](#)を参照のこと)

この問題は文字列比較の典型問題であると同時に, ゲノム配列の基礎的な解析法でもある. 解法を知りたい方は, Needleman-Wunsch algorithm というキーワードで調べていただきたい.

マラソン方式

前述の問題には正解が与えられており, 正解を出力するプログラムには同一の得点が与えられる. 便宜上この方式を正解照合方式と呼ぶことにしよう. 競プロでは正解照合方式以外にも, 最適解の導出が困難な問題を出題して競技者の求めた解の良さに応じた得点を与える方式もある. これはマラソン方式と呼ばれる. 正解照合方式は数時間以内に解答の提出が求められるのに対して, マラソン方式では数週間程度の解答提出期間が設けられる. 競技者は期間中に熟考を重ね, 1ポイントでも評価の高い解が得られるようにプログラムを磨き上げていく. このような特徴を持つマラソン方式が実社会で解くべき課題と相性が良いのは想像に難くない. 実際, さまざまなコンテストにおいて社会課題を想定した問題が出題されている. たとえば2017年から毎年開催されている[日立北大コン](#)では, アニーリングマシンの前処理技術や買い物支援サービスに必要な技術等が扱われた.

競技プログラミングとゲノム配列解析

DNAが担う遺伝情報は4種の文字から構成される文字列で記述することができる。生命の設計図ともいえるこの文字列は「ゲノム配列」と呼ばれ、ヒトであれば30億文字、文庫本にして約3万冊の分量となる。近年、DNAからゲノム配列を読み出すコストが劇的に下がり、膨大な量のゲノム配列が生み出されている。これらを詳細に分析することによって、さまざまな生命現象を解き明かす手がかりが得られる。

さて、ゲノム配列というと敷居の高い響きがあるが、文字列であることに変わりはない。この文字列にどのような分析が必要なのかというと、パターンの比較や発見など、自然言語解析でも扱われる文字列処理である。ここでは、実解析において頻出する問題を紹介しよう。

長さ30億のヒトの塩基配列 S が与えられたとする。クエリは S 上の位置 x を始点として S の一部をコピーした配列である。ただし、コピーする際にごくわずかに置換／欠損／挿入が発生する可能性があり、コピー元と完全一致するとは限らない。クエリの長さは150である。 x を求めよ。 S に対して数時間程度の前処理を許すが、1CPUコアあたり100万クエリを3分以内で処理する必要がある。

この問題の解決には高度な索引アルゴリズムに加えて、配列の類似を評価するアルゴリズムが必要となる（興味を持たれた方は、[索引アルゴリズムを](#)

概説した本会誌の記事もご参照いただきたい)。実は、上記にいくつかの要件を追加したものが、著名なプログラミングコンテストTopCoderにおいて出題されたことがある。その結果は大変興味深いもので、ゲノム配列解析の未経験者が最先端の既存ツールを大きく超える性能の解法を提出して優勝したのである。優勝者は後日行われたインタビューにおいて、「アルゴリズムの問題として面白く楽しんで取り組めた」と述べている。現代の生命科学で強く求められているのは、膨大な塩基配列データを超高速、超精密に分析することのできるアルゴリズムである。そのため、目的に応じてさまざまなアルゴリズムを自在に操る力を試す競技プログラミングとゲノム配列解析は非常に相性が良いのである。

ゲノコン2021の開催

前述のように、ゲノム配列解析の諸問題は競技プログラミングとして十分成立する。特に、実解析の課題はマラソン方式と相性が良い。腕の立つ競技者がどのような解法にたどり着くのか大変興味深く、また高度なアルゴリズムが生命科学の問題解決に役立つことも広く伝えたい。このような動機から、マラソン方式のプログラミングコンテスト「DNA配列解析チャレンジ（ゲノコン2021）」を開催する運びとなった。幸いにして、ゲノコンは競プロ人気の火付け役であり、多くの競技者が集うAtCoderにおいて開催することができた。

ゲノコン2021 — DNA配列解析チャレンジ - AtCoder

プログラミング初級者から上級者まで楽しめる、競技プログラミングコンテストサイト「AtCoder」。オンラインで毎週開催プロ
atcoder.jp



AtCoder創設者の高橋直大氏は2021年8月号の国会誌巻頭言において、
(AtCoderはアルゴリズムの) “頂に到達したくなるような山であり続ける”, と述べている。果たして、頂を目指す競技者たちはどのようにしてゲノム配列解析に取り組んだのか。

実問題！ 実データ！ 未公開データ！

コンテスト開催にあたってとりわけ注意を払ったのは、できるだけ実解析に近い出題とすることであった。そのため、研究テーマとしても成立し得る実問題を扱うこととした。以下が問題の概要である。詳細な定義はAtCoderの問題ページを参照されたい。

あなたはOxford Nanopore Technologies社の最新機器であるPromethIONを手に入れた！ PromethIONはDNAを読み取る機械である。この機械を用いてあなたは自分のDNA配列を解析しようと試みた。しかし、PromethIONはあなたのDNA配列を端から端まで100%正確に読み取れるわけではない。PromethIONに投入するためにDNAを精製すると、DNAは千切れてバラバラになってしまう。このため、1回の観測で得られる配列はDNA配列の一部だけだ。また、観測エラーもあるため、読み取ったDNA配列の断片配列の集合からあなたのDNA配列を精密に予測することは簡単ではない。あなたのミッションは、PromethIONの出力データからあなたの真のDNA配列を推測することである。あなたのDNA配列には父から受け継いだ配列と母から受け継いだ配列の2本があり、2本ともなるべく正確に推測したい。

問題文中のOxford Nanopore Technologies社（以下、ONT社）は実在する企業で、PromethIONはONT社による最新鋭のDNAシーケンサー（ゲノム配列の読み取り装置）である。PromethIONは数々の医学、生物学研究で用いられており、上述の問題では精度向上が望まれている。ゲノコンにおいて特筆すべきは、ONT社の協力を得て、実際にPromethIONを使って公開許諾済みのヒトDNAサンプル（つまり実在する人のDNA）から読み取ったゲノム配列データをコンテストに用いたことである。因みに、ONT社ではコンテスト開催時に新型の試薬（シーケンサーの読み取り精度に深く関連する薬品）を開発したばかりというタイミングであったため、市場に出ていない最新の試薬による実データを用いることとなった。つまりゲノコンでは、専門家でも扱ったことのない最新データを、現場の研究者に先立って競技者が分析することになったのである。

コンテスト開催にあたってもう一つ重視したのは非専門家への配慮であった。問題文に専門用語を含めないのはもちろんのこと、本問題に先立って練習問題を出題した。練習問題は本問題を解く上での事前知識、あるいは、要素技術となり得る3つの問題から成り、易しいものから順に時間をおいて公開した。本稿の冒頭に紹介した問題は2番目に公開した練習問題である。また、各問題に対して出題の背景となる生命科学に関する解説文を掲載し、出題された問題が実世界ではどのような意味を持つのか、競技者に理解してもらえるように努めた。

目を見張る競技者たちの実力、優勝は非専門家！

ゲノコンには約900名が参加登録し、練習用のA、B、C問題とメインのD問

題にはそれぞれ、672人、398人、234人、79人が解答した。後日実施したアンケートでは、ゲノム配列解析の経験者は回答者の1割にすぎず、他の生命情報解析の経験者を含めてもわずか2割程度であったことが分かった。さてコンテストの結果だが、練習用で最も難しいC問題とメインのD問題で1位をとったのは別々の競技者であった。そして特筆すべきことに、両者ともに生命情報解析の未経験者であった。

D問題では複数の断片文字列を適切に2つのグループに振り分ける処理がポイントの一つとなるが、優勝者の手法ではこの部分をうまく最大カット問題に帰着させていた。実解析に使われる最先端のツールの多くでも同様の方法が用いられているが、優勝者は事前知識なしにこの解法にたどり着いたようだ。限られた時間内で筋の良い解法を見出し、さらに実際に高い精度を達成するプログラムを仕上げたのは本当に素晴らしかった。

D問題を解く上でのもう一つのポイントはデータの特徴をとらえることであった。前述のように、D問題では実際のシーケンサーにより生成されたデータ（しかも専門家も見つけないデータ！）を入力としたため、データを分析してシーケンサーで生じるエラーの特徴を捉えた上で解決法を見出す必要がある。この点は一般的な競プロの問題とは異なりKaggleのような機械学習のコンテストに近い側面であったろう。コンテストでは分析用にいくつかのサンプルを配布すると同時に、ゲノムデータの可視化ソフトも提供した。この可視化ソフトは専門家が用いる本格的なものであったが、アンケートでは3割弱が利用したとの回答があった。また、成績上位者の中には非常に詳細なデータ分析を行った競技者もいた。問題を解くには多かれ少なかれ実データの性質に即したパラメータを導入することになるのだが、成績

2位の解法ではOptunaにより多数のパラメータを決定していたのが興味深かった。データ公開の制限があったため限られた数のサンプルを配布したのだが、データ分析が重要であったことを思えば、もう少し多くのサンプルを公開したかったところであった。

D問題の解答期間は当初3週間の予定であったが、公開に遅れが生じてしまったことから実際には10日程度に短縮された経緯がある。社会人競技者も多いことを考慮すると、参加者の多くが数日程度しか問題に取り組めなかった可能性も否めない。そのためD問題への取り組みには少し心配があったのだが、実際に提出された成績上位者のプログラムは総じてレベルの高いものであった。また、C問題では問題公開からわずか1日の間に（おそらく）最適解を導出するプログラムが提出されたことも申し添えたい。このように、競技者達の実力には目を見張るものがあり「アルゴリズムの頂」が研究の即戦力になり得ることが十分に示されたと考えている。

コンテストを盛り上げた競技者間の助け合い

ゲノコンでは、練習問題に関してはSNS等を通じて参加者間で解法に関して相談してもよいとして、参加者間で活発に議論が行われることを期待した。その結果、運営の予想を超える協力が生まれたことを報告したい。たとえばC問題では、ゲノム配列の並びが可視化されているとプログラム開発のヒントになるのだが、問題公開の翌日には、有志により開発された可視化ソフトが公開され、数日が過ぎたころには、成績上位者による解法記事がブログに公開された。ちなみに解法記事では、手の内をすべて明かすのではなく成績上位一歩手前までの解法が紹介されており、読者に改良の余地を与えている配慮がなされていた。この記事を読んで得点を伸ばした競技者も多かった

だろう。Twitterでは参加者が気軽にコミュニケーションを取り合う様子も伺え、コンテストは楽しい盛り上がりを見せた。問題の背景を読んだとの声も多くあった。アンケートでは、9割以上が問題の背景について興味深く読んだと回答しており、ゲノム解析研究の魅力を伝えられたように思う。

実解析を出題する難しさ

さて、ゲノム配列解析と競技プログラミングは好相性と述べたが作問に苦労がないわけではない。実解析では巨大なデータを扱うため、多数のCPUコアと数百ギガに及ぶメモリを搭載した計算機を用いる。たとえば、D問題で扱った実解析を前処理も含めてヒトゲノムサイズで行うには、数十ギガ以上のメモリと数百時間以上の計算が必要となる。一方で、一般的なジャッジサーバーであれば、最大でも3~4ギガ程度のメモリを用いて数十秒以内で解ける問題にする必要があるだろう（同様に、解の評価に必要な計算時間にも制限が生じる）。ゲノコンでは、問題の本質が損なわれない程度に問題サイズをスケールダウンしたほか、実解析に必要な処理の一部を事前入力に含めることで、競プロの問題としての要件を整えた。

実データを用いた出題の場合、真の解が未知となる問題もある。D問題では、入力データから実在する人物のゲノム配列を復元する精度を競ったが、実は、ヒトゲノム配列を誤りなく測定する方法は現存しない。そのため、コンテストでは最新の科学的知見から正しいと思われるものを正解として扱う必要があった。当然ながら最新の知見が真の解とは限らないため、競技者にとってはすっきりしない点であったかもしれない。

また運営側では、専門家が知識にものを言わせて半ばチートのような形で

解答を推定してしまわないかといった点についてもよく議論をした。どのようなチートが可能かは個別の問題に依存するだろうが、参加者が白けてしまわないような配慮にはそれなりの検討が必要だろう。

アルゴリズムの頂と科学研究の未来

分野外からは敷居が高く見られがちな問題であっても、アルゴリズムの工夫が必要となる部分をうまく切り出して、非専門家が理解できるように「翻訳」できれば、競プロの問題として成立する。研究では解くべき問題を見出して適切な定義を与える部分と、定義された問題を解く部分があるが、前者が得意な研究者が競プロコミュニティの力をうまく活用していく未来も描けるかもしれない。少なくともゲノコンでは、競技者の力を生命科学の問題解決に役立てる可能性を見出せたように思う。

本稿を通じてゲノコンに興味をお持ちくださった方には、以下より2021年日本バイオインフォマティクス学会年会において行われたセッション動画をぜひご覧いただきたい。入賞者の講演やより詳細な背景説明などを楽しんでいただけたと思う。また、AtCoderではコンテスト終了後も練習問題に取り組めるようになっているので、腕に覚えのある方はお試しいただきたい。

【IIBMP2021 一般公開セッション】DNA配列解析チャレンジ



(2022年3月21日受付)

(2022年4月15日note公開)

■清水佳奈（正会員）

2006年早稲田大学より博士（工学）取得。産業技術総合研究所、メモリアルスローンケタリング癌センターを経て、2016年早稲田大学情報理工学科准教授。2018年より同学科教授。生命情報科学の研究に従事。

■坂本一憲

2013年早稲田大学大学院博士後期課程修了，博士（工学）。国立情報学研究所助教，科学技術振興機構さきがけ研究員（兼任）などを経て，2019年よりWillBooster（株）代表取締役。現在に至る。

■笠原雅弘（正会員）

2004年東京大学情報理工学系研究科修士課程修了，東京大学新領域創成科学研究科情報生命科学専攻特任助教，2009年同専攻にて博士（科学），2009年同専攻にて専任講師，2018年より東京大学大学院新領域創成科学研究科メディカル情報生命専攻准教授。