タイトルスライド

本日は、DNA 解析における集団層別化のための主成分分析（PCA）の応用についてお話しします。金沢大学電子情報通信学類のナミール・ガリブと申します。よろしくお願いいたします。

はじめに

まず、PCA の基本的な役割についてご紹介します。PCA、つまり主成分分析は、大規模データセットの次元削減に広く使われる手法です。特にゲノム研究では、遺伝的多様性や集団構造を明らかにするために利用されています。この手法により、個体間の遺伝的類似性や差異を視覚化し、祖先集団の推定や遺伝的疾患の原因となる突然変異の特定が可能です。

主要用語

ここでは、今回の研究に関連する主要な用語を簡単に説明します。まず、主成分分析、PCA は、データセットの変数を減らしながら情報を最大限に保持する手法です。一塩基多型、SNP は、参照配列と異なる一塩基の DNA 変異を指します。また、参照ゲノムは、理想的な個体の遺伝子配列を基にしたデジタルデータベースであり、バリアントコーリングは参照ゲノムとの比較を通じて個体間の差異を識別するプロセスです。

研究概要

この研究の主な目的は、C 言語および Rust で効果的なバリアントコーリングと PCA アルゴリズムを実装し、その性能を比較することです。また、既存のツールである GATK や EIGENSTRAT との性能比較も行いました。簡略化のために、入力データはゲノム配列のみに限定し、メタデータや挿入・欠失変異を除外しています。

Variant Calling

次に、バリアントコーリングについて説明します。このプロセスでは、参照配列と各個体の配列を比較し、遺伝子変異の位置を特定します。CpG 部位やクラスター補正係数など、変異の発生率に基づいたスコアリングを導入しました。このスコアはロジスティック変換によって正規化され、各変異に関連する確率を示します。

PCA の数学的定式化

PCA の数学的背景について説明します。データの正規化、中心化、共分散行列の計算、固有値分解を経て主成分への射影を行います。主成分は最大分散を説明する方向を示し、データの次元を効果的に削減します。

データセットの生成

擬似ヒトゲノムデータセットを生成するために、カリフォルニア大学医学部が提供する無料ツールを使用しました。このツールを用いて異なるベース数のデータセットを作成し、それを参照ゲノムとして使用しました。さらに、自作のプログラムを用いて各個体に SNP を導入し、実験用データを準備しました。

実験 1: プログラム実装の検証

最初の実験では、PCA アルゴリズムの実装の妥当性を確認するために、EIGENSTRAT と比較分析を行いました。HERC2 遺伝子領域に限定して解析を実施し、この領域が眼の色に関与する遺伝子である点を考慮しました。1000 Genomes Project から取得したサンプルを用い、自作プログラムと既存ツールの結果を比較しました。

実験 2: 大規模データ処理実験

次に、大規模データを対象とした実験です。30M 塩基対や 30 億塩基対のデータセットを使用し、処理のスケーラビリティを評価しました。初期段階ではスタックオーバーフローが発生したため、メモリ管理をヒープに変更し、スパースモデリングを適用しました。この結果、処理の効率化とメモリ使用量の削減に成功しました。

C と Rust の実装比較

C と Rust の実装を比較した結果、性能の違いは僅かであるものの、Rust のメモリ管理がより安全で効率的であることが分かりました。Rust の &str データ型は文字列の長さを事前に格納するため、O(1) の計算量で操作が可能です。一方、C 言語はヌル終端文字を確認する必要があるため、O(n) の計算量が必要です。

パフォーマンス評価

計算量の評価結果を示します。スパース表現を用いることで、PCA の計算コストを削減しました。特に、行列とベクトルの積や転置行列との積における計算量が O(v) となる点が効率的です。また、各主成分の正規化には Z スコアを適用し、データの次元削減を最適化しました。

結論

PCA はゲノム解析において有用ですが、高次元データにおける非線形な相関関係を完全に捉えることは困難です。C 言語と Rust の比較では、Rust がメモリ管理の安全性と効率性でわずかに優れていることが確認されました。今後の改良点として、非線形次元削減手法や分散コンピューティングの活用が挙げられます。