計量生物学における高次元統計解析の可能性

筑波大学・数理物質系　　青嶋　誠
筑波大学・数理物質系　　矢田　和善
(株)Rhelixa・代表取締役社長　　仲木　竜

ゲノム科学・情報工学・金融工学などの現代科学の一つの特徴は，データがもつ次元数の膨大さにある．特に，マイクロアレイや次世代シークエンサによる遺伝子発現データなどでは，次元数が数万を超える事例も解析の対象となる．こういった高次元データの特徴は，次元数が標本数を遥かに超えることであり，それゆえ高次元空間に豊富な潜在情報を有するものの，それが巨大なノイズに埋もれて見つけ難いことである．

既存の高次元データ解析では，データに内在する潜在情報やノイズがスパースであることを前提とし，スパースモデリングを適用することで何かしら疎な解を得ている．しかし，計量生物学などに見られる高次元データは，潜在情報が次元数に依存して非スパースな構造をもち，そして，ノイズも非スパースである．非スパース性を考慮せずにデータがスパースだと思って解析すると，得られた解はノイズに埋もれたものでしかなく，豊富な潜在情報は全く抽出できていないこともある．計量生物学の最先端のデータを扱うためには，巨大なノイズを処理する方法論が求められる．

講演者の研究グループは，これまで，高次元データの統計解析を行うための新しい理論と方法論を構築してきた．それらは，高次元データの幾何学的表現に基づいており，高次元統計解析と名付けられ，以下の 3 つの柱に纏められる．詳細は，日本統計学会和文誌の寄稿論文 [1] と [2] を参照されたい．

(I) 高次元データの統計的推測
高次元 2 標本問題，判別分析，変数選択，パスウエイ解析など．

(II) 高次元データの新しい主成分分析
ノイズ掃き出し法，クロスデータ行列法，クラスタリング，信号行列の推定など．

(III) 巨大なノイズに対処する非スパースモデリング
強スパイクモデル，データ変換，高次元漸近正規性など．

本講演では，計量生物学という高度な技術を要する研究領域に，高次元統計解析の可能性を見出したい．具体的には，巨大なノイズの処理・潜在情報の抽出・統計的推測の精度保証といったテーマに，高次元統計解析の理論に裏打ちされた方法論を提案できればと考える．一つの解析例として，ゲノム構造活性データに高次元統計解析の立場からアプローチする．細胞ごとのゲノム構造活性は，細胞特異的な遺伝子活性パターンを読み解く上で重要である．加齢黄斑変性の発症群と非発症群でのゲノム構造活性を比較し，発症状態を定義するためのゲノム構造活性を予測する．

参考文献

[1] 青嶋 誠 (2018). 日本統計学会賞受賞者特別寄稿論文: 高次元統計解析: 理論と方法論の新しい展開, 日本統計学会誌, **48**, 印刷中.

[2] 青嶋 誠, 矢田和善 (2013). 日本統計学会研究業績賞受賞者特別寄稿論文: 高次元データの統計的方法論, 日本統計学会誌, **43**, 123–150.