解説

応用

農学系ゲノム科学領域における情報科学・統計科学教育の取り組み

石井一夫
東京農工大学

ゲノム科学の発展と情報科学・統計科学教育

ゲノム科学は，次世代シーケンサや質量分析装置などの高性能機器の実用化に伴いながら，急速に発展している分野である．特に，ゲノム科学の進歩により，医学，農学，環境などの生命科学分野においては，大量のデータ（ビッグデータ）の解析を行う必要に迫られる機会が多い．これらの大規模データ解析においては，プログラミング，データベース，ネットワークなどの情報処理技術やパラメトリック・ノンパラメトリック検定，多変量解析，機械学習などの統計科学が必須である．これらのデータ解析は，昨今のメディアで話題となっているビッグデータ分析で使われている技術と共通するものであり，情報科学や統計科学の知識と技術を併せた境界領域の分野であるため，現在の生命科学系学部において，これらの教育に十分に対応できているとはいいがたい．

すなわちゲノム科学は，ここ数年で急速に発展してきた新しい技術であるため，従来の教育体制でカバーしきれていないのが現状である．

東京農工大学「農学系ゲノム科学領域における実践的先端研究人材育成プログラム（以下，農学系ゲノム科学人材育成プログラム）」では，2011年4月より，ゲノム科学の研究を行いたい大学院生を対象に，ゲノム科学に関する研究テーマを募集し，解析に必要な技術を教授する教育プログラムを実施している．本稿では，これらの教育プログラムにおけるゲノム科学ビッグデータのデータ解析に関する情報科学教育・統計科学教育に関する取り組みに関して報告する．

農学系ゲノム科学でのデータ解析に関する情報科学・統計科学教育の実践

❑ 教育プログラムの創設

近年，生命科学分野において，ゲノムビッグデータ解析の必要性に伴い情報科学・統計科学教育に対するニーズは非常に高くなってきている．

しかし，農学系領域において，これらゲノム科学のデータ解析を十分に行えるような，プログラミング，データベースの構築・取り扱いを含む情報科学教育や，統計学的検定，ベイズ統計，多変量解析に機械学習，自然言語処理なども含めた統計科学教育，およびその基礎となる線形代数や微分積分，常微分方程式，偏微分方程式などの数学的な基礎知識に関する教育が十分に行われている教育機関は多くないと思われる．

東京農工大学では，このような環境のなか，文部科学省の特別経費により，ゲノム科学の研究を実施する必要に迫られた学生，研究者に，ニーズにあった教育を実践し，ゲノム科学研究を実施できる学生，研究者を育成することを目標として，2011年4月より，文部科学省の特別経費により，教育プログラム「農学系ゲノム科学人材育成プログラム」を開始した．

これらの取り組みの中で数学的基礎知識を基盤とした情報科学・統計科学技術に基づくデータ解析を

実施できる学生，研究者を育成する研究教育活動を実施している．

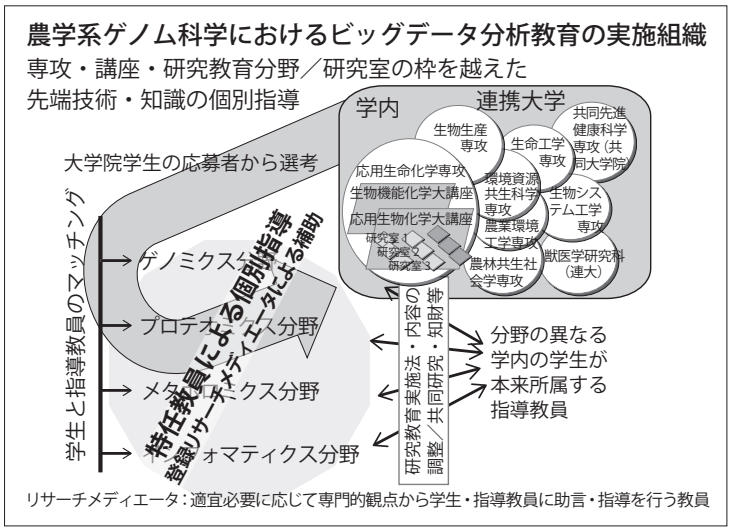

図 -1　農学系ゲノム科学人材育成プログラムの実施体制

❑ 教育プログラムの概要

(1) 教育の対象者

この教育プログラムは，農学系学部（工学部，獣医学部の関連学科を含む）のゲノム科学を専門とする大学院を対象としている．すなわち，学部，研究科，専攻，講座，研究教育分野の枠を越え，東京農工大学だけでなく，東京農工大学と連携する関連の学部（茨城大学，宇都宮大学を含む）も対象としている．

(2) 教育の実施概要

まず，本教育プログラムでは，東京農工大学大学院の学生（修士課程，博士後期課程）からゲノム科学を必要とする研究課題の募集を行う（**図 -1**）[1]．本学の大学院学生であれば，農学府・工学府・Base・連合農学研究科（茨城大学・宇都宮大学を含む）に所属するすべての学生が応募できる．

学内外の識者による審査を経て採択された場合，研究室の個々の研究テーマを実施しながらゲノム科学（ゲノミクス・プロテオミクス・メタボロミクスおよびこれらの応用分野）に関する知識と技術を，主指導教員に加え，ゲノム科学分野を専門とする特任教員およびリサーチメディエータとの連携による個別指導を受け習得することができるしくみになっている．

また，初心者レベルから専門家レベルまでの情報処理技術の習得も含めたゲノム科学全般について，知識・実験技術などに関する講習会・セミナ・シンポジウム等を適宜実施する． セミナや公開講座の実施の際には，状況に応じてゲノム科学のデータ解析を行うことを希望する学内外の教員ならびに一般企業の研究者をも対象に含めた．

(3) 教育の実施過程

以下，本教育プログラムの実施過程をまとめる．

1) 研究テーマの公募と評価，採択

次世代シーケンサ（ゲノム自動解析装置）を用いてゲノム科学研究を行いたい大学院生から研究テーマを公募し，その内容の教育上の妥当性，効果，社会的重要性を評価した上で，有望な研究テーマを採択する．

2) データの取得

採択された研究テーマそれぞれで，その指導教官と学生の打合せを行った後に，次世代シーケンサなどのゲノム解析装置を用いて，ゲノム解析配列データを取得する．

3) データ分析

得られたデータを，UNIX/Linux をプラットフォームとしたデータ解析環境を用いた解析を実施する．その際，プログラミング，データベース，ネットワーキング，統計解析などのデータ分析方法を，マンツーマンでトレーニングする．

4) 講習会，セミナの実施

実施対象の大学院生や，学内外の教員，一般社会人を対象とした講習会，セミナ，シンポジウムを実施する．

5) 研究報告会の実施

各学生の研究成果を発表する報告会を実施する．

❑ 教育プログラムの具体的内容

表 -1 に本教育プログラムで実施したゲノム科学領域における情報科学・統計科学教育の実施内容を示した[2]．

教育プログラムは 3 カ月ごとの区切りになっており，基礎技術レベル，応用技術レベル，アドバンスレベル，専門家レベル，プロレベルと段階を追って

農学系ゲノム科学におけるビッグデータ分析の実施内容

提供する支援レベル（習得技術・内容）

基礎技術レベル（3 カ月）	**E1 : UNIX の操作・データ解析環境の構築・スクリプト作成（Perl/Ruby/Python）** FreeBSD, Linux の操作，インストール，Perl などを用いたテキスト処理
応用技術レベル（3 カ月）	**E2 : DNA 配列アセンブリ・メタゲノム解析・データベース構築（SQL）** DNA 配列アセンブリソフト Velvet, Oases, Trinity などの操作とデータアセンブリ法，原理 データベース管理システム MySQL, PostgreSQL を用いたデータベースの構築と，クエリ，集計
アドバンスレベル（3 カ月）	**E3 : RNA-Seq 解析・ChIP-Seq 解析・統計解析（R/MatLab）** 発現定量データの取得と統計解析，パラメトリック検定，ノンパラメトリック検定， 多変量解析，機械学習，クラスタ解析，グラフィックスによる視覚化
専門家レベル（3 カ月）	**E4 : 上記以外のデータ解析法（QTL・カスタムライブラリの解析）** 遺伝統計解析，統計モデリング（一般化線形モデル，一般化加法モデルなど），モンテカルロシミュレーション，マルコフ連鎖モンテカルロ法，遺伝学的系統樹解析
プロレベル（3 カ月）	**E5 : 新規データ解析法の開発実装（C/C++/Java）** Perl, Python, Ruby, C，C++，Java を用いた新規アルゴリズムの実装

表 -1　教育プログラムの具体的内容

ステップアップしていく．

(1) 最初の「基礎技術レベル」では，① UNIX の簡単な操作の入門にはじまり，②データ解析環境の立ち上げと，③シェルや Perl，Ruby などの簡単なスクリプトの書き方を学ぶ．

(2)「応用技術レベル」は，次世代シーケンサのデータ解析を実際に行うレベルである．

① DNA 配列データの品質管理：品質管理ソフト FastQC を用いた DNA データのクオリティチェックをシェルや Perl などのスクリプトで行い，クオリティの悪いデータを FastX-Toolkit や cutadapt などの簡易ソフトで除く．

② DNA 配列データのアセンブリ（連結），マッピング（DNA 配列の参照配列への整列）：その後，Velvet，Oases，Trinity などの DNA 配列アセンブラ（DNA 配列連結ソフト）で塩基配列のアセンブリを行ったり，DNA 配列マッピングソフト BWA，Bowtie などを用いて参照配列へマッピング（整列）を行ったりする．

③ コマンドによるデータベースの検索，データベースの構築：コマンドによる BLAST を用いた検索や，次世代データを用いた MySQL や PostgreSQL によるデータベースの構築とクエリの方法について学ぶ．統計解析ソフト R を用いた簡単な集計方法についてもここで学ぶ．

(3)「アドバンスレベル」では，次世代シーケンサのデータ解析のうち，より難易度の高いデータ解析を行う．具体的には，RNA-Seq（RNA の網羅的定量データ解析），ChIP-Seq（DNA へのタンパク質の結合様式の網羅的な解析），リシーケンシング（DNA の多型解析）および R を用いた統計解析について学ぶ．

① 発現定量解析（RNA-Seq）ではマッピングソフト Tophat を用いたマッピングと発現解析ソフト Cufflinks によるデータの集計法について学ぶ．

② ChIP-Seq では，BWA により参照配列にマッピングしたあと MACS によるピーク（タンパク質結合部位）検出を行う．その後，MEME や WebLOGO などのソフトによるタンパク質の結合するコンセンサス配列の検出なども行う．

③ リシーケンシング（多型解析）では，BWA により参照配列にマッピングしたあと，SAMtools などによりデータの集計などデータ解析を行う．

④ 発現定量解析については，R による統計検定（パラメトリック検定，ノンパラメトリック検定，分散分析，多重比較の多重補正）などを行う．

(4)「専門家レベル」では，次世代シーケンサのデータ解析のうち，通常のソフトウェアで提供されていない非定型のデータ解析を行う．

①シェルや Perl などのスクリプト言語を用いた自動化パイプラインを構築したり，通常の定型の解析ソフトで行えないようなカスタムメードのデータ解析を行ったりする．遺伝統計解析なども必要に応じて，ここで学ぶ．

② R や Matlab については，統計モデリング（一般化線形モデル，一般化加法モデル），モンテカルロシミュレーションやマルコフ連鎖モンテカルロ法などによる解析法（ブートストラップ法，ジャックナイフ法，並べ替え検定）を学ぶ．③

機械学習，k-means 法，主成分分析，クラスタ分析など．データマイニング手法を学ぶ[4]．

(5)「プロレベル」では，プログラミング言語を用いた新しいデータ解析法の実装について学ぶ．

R による関数の作成とパッケージング．新たな解析方法について，Perl，Python，Ruby などを用いたやや高度なプログラミングを行う．ソフトウェアをインストールする際の，Makefile の読み方やその修正方法，ビルドの際にエラーが出たときの対応方法など C，C++，Java のコンパイル方法について学ぶ．

農学系の学生を対象にしているので，時間的制約もあることから C，C++，Java を用いた新規ソフトウェアの開発まで行うレベルは想定していないが，そのような研究に挑戦する学生が出てくることを期待する．

教育プログラムの実施状況

表 -2 に今回の教育プログラムに参加しデータ解析技術を習得した大学院生の人数をまとめた．

2011 年度に全体で合計 37 名の採択者を受け入れ，そのうち 22 名に対して，ゲノミクス・インフォマティクス分野（表 -2 の GI 分野）の教育指導を行った．残りの 15 名はプロテオミクス分野の教育指導を受けた．

2012 年度には，のべ 85 名の採択者を受け入れ，のべ 63 名に対してゲノミクス・インフォマティクス分野の教育指導を行った．

2013 年度には，のべ 54 名の採択者を受け入れ，のべ 36 名に対してゲノミクス・インフォマティクス分野の教育指導を行った．

年度ごとに採択方法が異なっているので，単純に比較はできないが，順調に教育実践を行った実績を上げたと考える．

まとめ

東京農工大学「農学系ゲノム科学人材育成プログラム」における情報科学，統計科学教育の実施状況を紹介した．まとめに変えて，現在の問題点，今後の課題について述べる．

期間	全採択数（GI 分野）
2011 年度	
第 1 期（7 ～ 9 月）	12 名（うち 7 名）
第 2 期（10 ～ 12 月）	14 名（うち 8 名）
第 3 期（1 ～ 3 月）	11 名（うち 7 名）
2012 年度	
第 1 期（6 ～ 8 月）	27 名（20 名）
第 2 期（9 ～ 11 月）	27 名（20 名）
第 3 期（12 ～ 2 月）	31 名（23 名）
2013 年度	
第 1 期（6 ～ 9 月）	25 名（16 名）
第 2 期（11 ～ 2 月）	29 名（20 名）

表 -2　農学系ゲノム科学における情報科学・統計科学教育の実施実績（2011 ～ 2013 年度）

ほとんどの生物系の学生は，本プログラムに参加するまでに，プログラミングなどの情報科学実習や，統計数理科学の授業などをあまり受けたことがなく，その基礎となっている線形代数や微分積分，偏微分，微分法方程式，確率・統計などの数学的基礎を十分に習得せずに，データ解析を学びにくるのが実状であり，コンピュータリテラシーや，数式の解釈を理解してもらうだけでも相当な苦労がある．

今後，情報科学や統計科学を含むデータサイエンスが工学や理学以外の生物系学部においてきちんとしたカリキュラムとして取り込まれることを期待するが，その実現に関しての見通しは，周囲の理解を得るのはなかなか困難で，決して明るいとはいえない．また，分かりやすい教科書，自習書もあまりないなどの問題もある．しかし，このプログラムを通じて，できることから少しずつ実施して行きたいと考える．

参考文献

1) 文部科学省 連携事業「農学系ゲノム科学領域における実践的先端研究人材育成プログラム」プログラムの概要，http://genome.lab.tuat.ac.jp/~genome/overview.html
2) 文部科学省 連携事業「農学系ゲノム科学領域における実践的先端研究人材育成プログラム」プログラムの内容，http://genome.lab.tuat.ac.jp/~genome/program.html
3) Rizzo, M. : Statistical Computing with R, Chapman & Hall/CRC (2008)（Rizzo M（著），石井一夫，村田真樹（共訳）：R による計算機統計学，オーム社 (2011)）．
4) 石井一夫，佐藤　暁，古崎利紀，有江　力，寺岡　徹：ゲノム科学におけるビッグデータ・データマイニング，日本統計学会誌，Vol.43, No.1, pp.90-111 (2013).

（2014 年 2 月 3 日受付）

石井一夫（正会員） kishii@cc.tuat.ac.jp
東京農工大学農学府農学部「農学系ゲノム科学人材育成プログラム」．