Vol.2013-CE-118 No.12
2013/2/9

農学領域におけるゲノム科学ビッグデータのデータ解析

石井一夫[†1]　古崎利紀[†1]

東京農工大学では、ゲノム科学に関する研究を行いたい大学院生を対象に、研究テーマを公募しその内容を吟味して一定の評価を得て採択された学生に研究の実施に必要な技術などを教授する教育プログラムを 2011 年 4 月より実施している。ゲノム科学においては、次世代シーケンサーより産生される数千万～数億断片の塩基配列データを取り扱い、データ解析を行うことが日常的に生じている。このようなビッグデータのデータ解析は、プログラミング、データベース、ネットワークを基盤とする情報処理技術が必須である。本稿では、この農学領域のゲノム科学ビッグデータのデータ解析に関する教育に関する取り組みに関して報告する。

Data Analysis of Genomic Big Data in Agricultural Sciences

KAZUO ISHII[†1]　TOSHINORI KOZAKI[†1]

We have performed a human resource development program in agricultural genome sciences in Tokyo University of Agriculture and Technology since 2011. This program is intended to educate the genomic technology and information technology to graduate students requiring genomic research. In recent genomics research, daily data analysis with large scale of sequence data produced by next generation sequencer is performed. Information technology, such as programming, database and network is essential for big genomic data analysis. Educational activity of Human Resource Development Program in Agricultural Genome Sciences, Tokyo University of Agriculture and Technology is described in this manuscript.

1. はじめに

ゲノム科学は次世代シーケンサーや質量分析装置の実用化にともない、急速に発展している学問分野である。ゲノム科学の進歩により、医学や農学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）のデータ解析を行う必要に迫られる機会が多くなってきた。

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。これらの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が必須であるが、現在の生命科学系学部においては、このような教育が十分になされているとはいいがたい。

このような状況のなか、東京農工大学では、2011 年 4 月より、ゲノム科学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（農学系ゲノム科学人材育成プログラム）を開始した。教育プログラムは、大学院生に個別指導により実施した。

ゲノム科学に関するビッグデータ解析においては、(1) UNIX/Linux をプラットフォームとした、コマンドベースによる操作、(2) シェルや Perl、Ruby、Python などのスクリプト言語によるテキスト処理や、プロセスの連結（パイプラインの構築）、(3) MySQL や PostgreSQL によるデータベースの構築やデータベースに基づいたデータ処理、(4) R や Octave などの統計解析ソフトを用いた解析などを指導した。

また、ゲノム科学の知識や情報リテラシー不足を補うために、年に 10 回以上のセミナー、講習会、公開講座などを実施した。これらのセミナーなどは、学内だけでなく内外の教員や、一般社会人にも公開で行い、ゲノム科学の知識とデータ解析技術の普及に務めた。

ゲノム科学教育プログラムに参加した学生には、その研究成果を卒業研究や、学会発表で報告したほか、研究成果報告会を実施し、その教育成果を披露した。本稿では、これらの教育実践の内容について以下に紹介する。

2. 次世代シーケンサーの普及とビッグデータ解析の必要性

2.1 次世代シーケンサーとは

ゲノム科学の進歩により、医学や農学などの生命科学分野においては、数百ギガバイトの塩基配列データを取り扱うことは日常的になっている。数テラバイトを超えるデータの解析を行うことも珍しくない。これは、2005 年以降に、急速に普及してきている次世代シーケンサーと呼ばれる超並列型自動塩基配列決定装置の普及によるところが大きい。

(1) 次世代シーケンサーの種類

次世代シーケンサーは、固相上に無数の DNA 断片を結合させ、酵素を用いて伸張反応や連結反応を起こさせ、その配列を蛍光や電圧によって並列的に検出する装置である。ロシュ社の 454、イルミナ社の GAIIx、HiSeq、MiSeq、ライフテクノロジーズ社の SOLiD などが主流である。

また、半導体シーケンサーと呼ばれるライフテクノロジーズ社の IonPGM や IonProton、第 3 世代シーケンサーと呼ばれる Pacific Biosciences 社の PacBio RS など、新たな原理に基づいたものや、安価なものも登場しており、その普及

†1 東京農工大学農学系ゲノム科学人材育成プログラム
Human Resource Development Program in Agricultural Genome Sciences, Tokyo University of Agriculture and Technology

はさらに加速化している。

(2) 次世代シーケンサーの特徴

従来のキャピラリー型シーケンサーと異なり、応用範囲が非常に広い。例えば(1) DNA の塩基配列解析だけでなく、(2) RNA の発現定量解析や、(3) DNA のメチル化やタンパク結合領域などのエピゲノムの解析も行える。

発現定量解析においては、従来のマイクロアレイのようにゲノム配列が既知の生物の解析だけでなく、ゲノム配列が未知の生物の解析の発現定量解析も可能であるなど、従来の機器に比べて画期的なほど情報量が多いという特徴を持つ。また、材料の標識方法により、多種類の検体を同時に分析することができるなどいろいろな機能が追加されている。

(3) 次世代シーケンサーのデータ解析法

次世代シーケンサーのデータ解析は、おおまかに以下の3段階に分けられる。

- 一次解析：次世代シーケンサーより産生された画像データを統合して塩基配列データを得る。
- 二次解析：得られた配列データは、配列が既知の参照配列に整列させたり（マッピング）、新たに連結する配列を重ね合わせたり（アセンブリ）して、塩基配列データを得ることができる。また、重なった塩基配列断片の数を計数する。これにより、数千万から数億個のテキストデータ（塩基配列）と数値データ（解読された断片（リード）の数）を得る。
- 二次解析：得られたデータは、既知のデータベースとの相同性を調べたり、検体ごとの数値データを用いて統計解析を行ったりして有用な情報を引き出し、実験結果がまとめられる。

2.2 次世代シーケンサーのデータ解析に用いられる情報技術

次世代シーケンサーのデータ解析に用いているソフトウェアの一覧を右に示す。ここに掲載しきれないがおよそ100種類のデータ解析ソフトを日常的に使用している。

(1) プラットフォーム

データ解析に用いるプラットフォームはもっぱらUNIX/Linuxである。MS Windows やMacOSX などを用いている人もいるが、これらはセキュリティが脆弱であったり、安定性･堅牢性の面でUNIX/Linuxに劣っていたりする。UNIX/Linuxの中でも特に、FreeBSDを中心に解析を行っている。これは、ZFSという大容量で高速のデータ解析を可能とする先進的なファイルシステムを有するのが特徴である。さらに、サーバやデータ解析で業界標準となっているRHELのクローンであるCentOSやScientific Linuxも併用している。合計40コアCPU、850ギガバイト近くのメモリを搭載したパワフルな解析マシンを使用している。

図 1　東京農工大学に導入されている次世代シーケンサーの例:イルミナ社 GAIIx(上)、MiSeq(下)

Figure 1　Example of next generation sequencers in Tokyo University of Agriculture and Technology: Illumina GAIIx (upper), MiSeq(lower)

表 1　東京農工大学でデータ解析に用いるソフトウェア群

Table 1　Software for Genomic Data Analysis used in Tokyo University of Agriculture and Technology

分類	ソフトウェア名
OS	Linux/UNIX (CentOS 6.3, Scientific Linux 6.3 and FreeBSD 9.1)
プログラミング言語	Perl, Python,Ruby, Java, C, C++
データベース	MySQL, PostgreSQL
ゲノム配列データのアセンブリ	Velvet, ABySS, SOAPdenovo, WGS Assembler, MIRA3, Phrap
ゲノム配列データのマッピング	Bowtie, Bowtie2, BWA, Maq, SOAP
RNA 発現解析用ソフト	Tophat, Cufflinks, Trinity, Oases, SOAPdenovo-Trans
ChIP-Seq 解析用ソフト	MACS, Quest, SISSRs, SPP
統計解析ソフト	R/Bioconductor, Octave, Mailab
相同性解析、注釈付けソフト	BLAST, BLAT
生物学データ解析用ライブラリ	BioPerl、BioRuby、BioPython、BioJava

(2) 汎用のフリーソフト群

使用するデータ解析ソフトは、フリーソフトが多い。シェルスクリプトや Perl、Python、Ruby など汎用スクリプト言語を日常的に活用して、テキスト処理や各プロセスの連接（グルー言語）を行う。必要に応じて、C、C＋＋、Java などのプログラミングも行う。また、ビッグデータ処理には、データベースは欠かせない。MySQL や PostgreSQL などのデータベース管理システム（DBMS）も日常的に使用している。

(3) データ解析専用のフリーソフト群

ゲノム科学専用の解析ソフトウェアは、例えばゲノム配列データのアセンブリやマッピングなど多数のオープンソースのソフトウェアがリリースされている。現在は、Velvet、Oases、Trinity、Bowtie、BWA、Tophat、Cufflinks、MACS などを用いることが多いが、ソフトウウェアの移り変わりは非常に激しく、毎月のようにバージョンアップや、新規ソフトの導入を図っている。

統計解析は R と Octave を用いる。特に R は Bioconductor というゲノム解析に特化した専用パッケージを含むために非常に便利である。

BioPerl や BioPython、BioRuby、BioJava など各種のプログラミング言語で使用できる生物学データ解析用もプログラミングの効率化に有用である。

3. 農学系ゲノム科学でのデータ解析に関する情報科学教育の実践

3.1 農学系ゲノム科学人材育成プログラムの目的と実施開始時期

このように、生命科学分野においては、ビッグデータの解析のニーズは非常に高くなってきているが、農学系領域においては、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､データベースの取り扱い、統計解析などの情報科学教育が行われている教育機関は多くない。

東京農工大学では、このような環境で、ゲノム科学のビッグデータのデータ解析を行う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、これらの情報科学技術とデータ解析を実施できる学生、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。

3.2 農学系ゲノム科学人材育成プログラムの実施体制

図 2 に本教育プログラムの実施体制を示す。

(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生を対象としている。すなわち、学部、研究科、専攻、講座、研究教育分野の枠を越え、東京農工大学および、東京農工大学と連携する関連の学部を対象としている。

(2) 教育の実施概要 [5)]

本教育プログラムは、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。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）に所属する全ての学生が応募できる。

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及びリサーチメディエーター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・セミナー・シンポジウム等を適宜実施することとした。

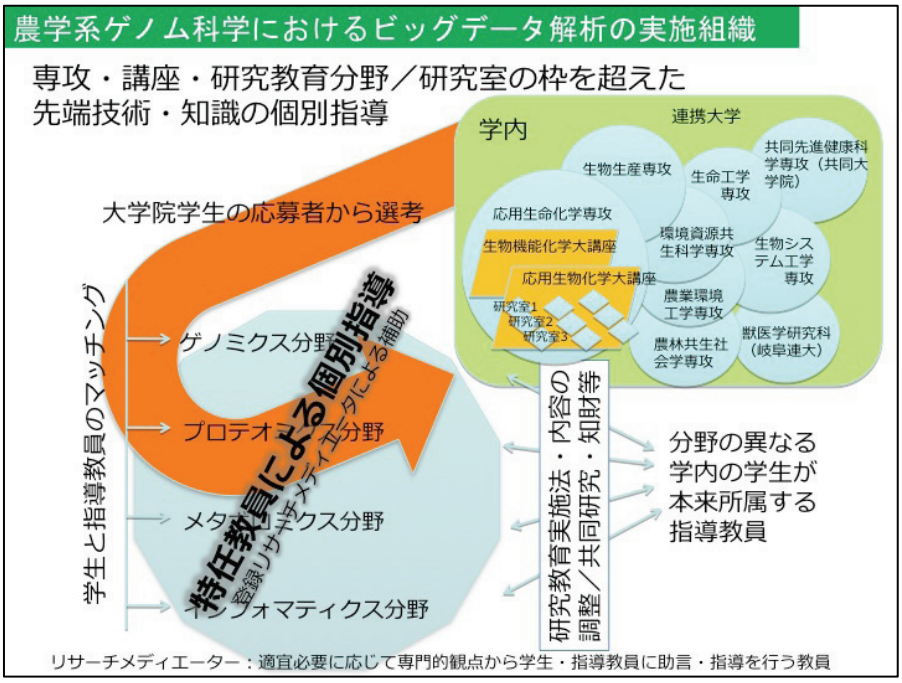


図 2 東京農工大学農学系ゲノム科学人材育成プログラムの実施体制

Figure 2 Organization of Education Program for Genomic Big Data Analysis in Tokyo University of Agriculture and Technology

また、セミナーや公開講座の実施の際には、適宜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。

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
教育内容は、次世代シーケンサー（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) 採択された研究テーマにつき、その指導教官と学生の打ち合わせを行った後に、次世代シーケンサーなどのゲノム解析装置をもちいて、ゲノム解析配列データを取得する。

3) 得られた配列解析データを、UNIX/Linux をプラットフォームとしたデータ解析環境を用いて、解析を実施し、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解析方法を、マンツーマンでトレーニングを実施する。

4) 同様のゲノム科学研究を行いたい学生や、学内外の教員、一般社会人を対象とした講習会、セミナー、シンポジウムを実施する。

5) 研究を実施した学生の研究成果を発表する研究成果報告会を実施する。

3.3 農学領域におけるゲノム科学のビッグデータのデータ解析に関する情報科学教育の実施内容

表 2 に本教育プログラムで実施したビッグデータのデータ解析に関する情報科学教育の実施内容を示した [6]。

教育プログラムは 3 ヶ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ステップアップしていく。

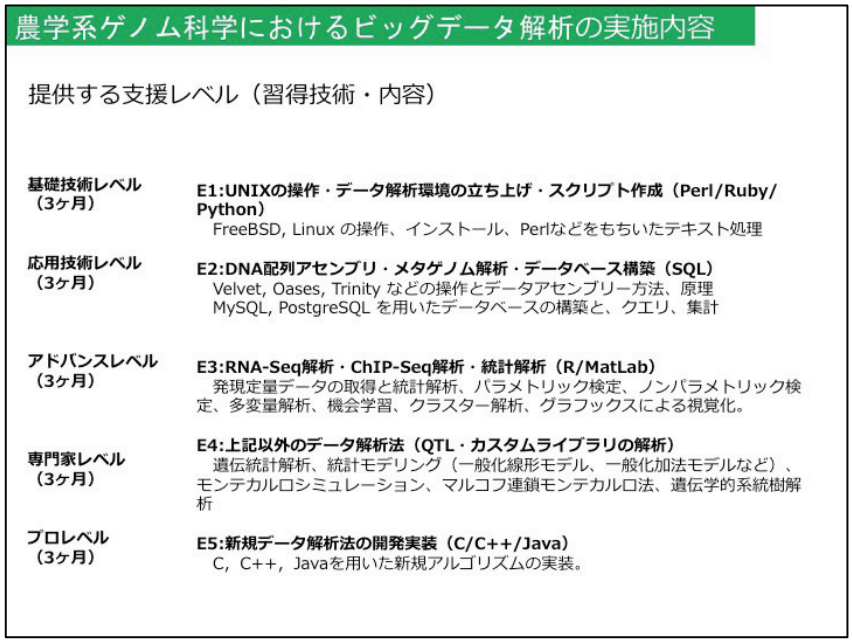
1) 最初の「基礎レベル」では、UNIX の簡単な操作にはじまり、データ解析環境の立ち上げと、シェルや Perl などの簡単なスクリプトの書き方を学ぶ。Perl や Ruby などの入門的な内容を学ぶ。

2) 「応用技術レベル」では、実際の次世代シーケンサーのデータ解析を行うレベルである。

FastQC を用いたデータのクウォリティチェックをシェルや Perl などのスクリプトや、FastX-Toolki や Cutadapt などの簡易ソフトで除く。

その後、Velvet、Oases、Trinity などの配列解析アセンブラで塩基配列のアセンブリを行ったり、BWA、Bowtie などを用いて参照配列へマッピングを行ったりする。

表 3 農学系ゲノム科学における育成プログラムの実施内容
Table 3 Contents of Education Program for Genomic Big Data Analysis in Tokyo University of Agriculture and Technology



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。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、次世代シーケンサー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。すなわち、RNA-Seq、ChIP-Seq、リシーケンシング（多型解析）および R を用いた統計解析について学ぶ。いわゆる 2 次解析に相当する解析技術全般を学ぶ。

発現定量解析（RNA-Seq）では Tophat を用いたマッピングと Cufflinks によるデータの集計法について学ぶ。

ChIP-Seq では、BWA により参照配列にマッピングしたあと MACS によるピーク検出を行う。その後、MEME や WebLOBO によるコンセンサス配列の検出なども行う。

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 などによるデータ解析を行う。

発現定量解析について 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 [2]。

4) 「専門家レベル」では、次世代シーケンサーのデータ解析のうち、非定型のデータ解析を行う。基本的には、3 次解析に相当する部分の解析である。

主に、シェルや Perl などのスクリプト言語を用いて、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。データのフォーマット変換や、Samtools でのデータの集計などは自在にできるようになることを期待している。

遺伝統計解析なども必要に応じて、ここで行う。

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）の使い方。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、クラスター解析など。データマイニング手法を学ぶ[4]。

4) 「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 の読みかたやその修正方法、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。

農学系なので、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。

図 3 にこれらの 5 つの習得レベルにおける実施例を示した。

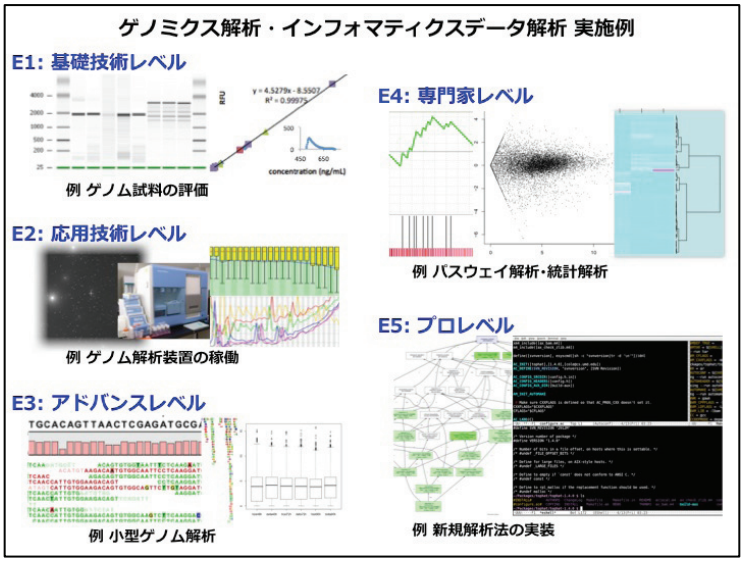


図 3　農学系におけるゲノム科学のビッグデータ解析の実施例

Figure 3　Example of Genomic Big Data Analysis of Genomics Education Program in Tokyo University of Agriculture and Technoloy

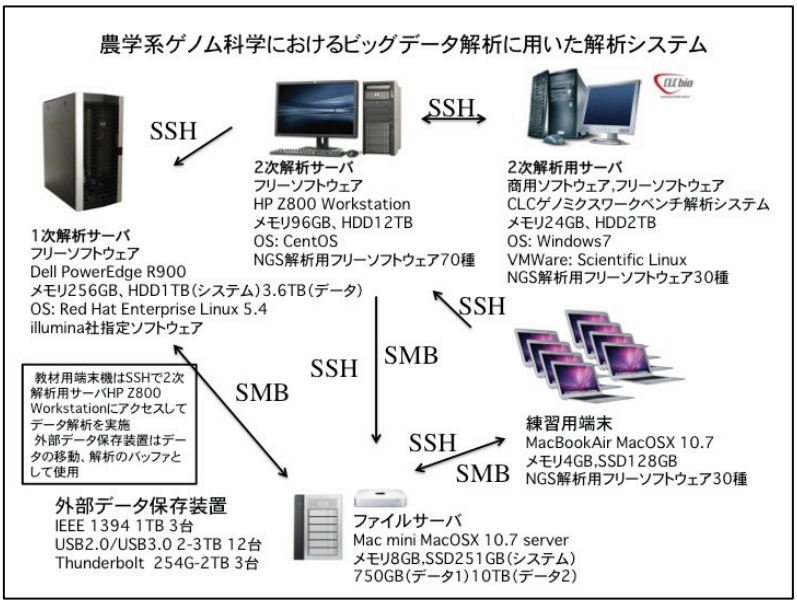


Figure 4　System of Education Program for Genomic Big Data Analysis

図 4　農学系ゲノムにおけるデータ解析に用いたシステム

3.4　農学領域におけるゲノム科学のビッグデータのデータ解析に関する情報科学教育の実施するために整備したシステム

図 4 に、データ解析に用いたシステムの例を示した。現在はこの図に示した内容よりさらに強化されている。ここで示すように、何台かの解析サーバをファイルサーバや SSH などのネットワークで接続し、解析を行う。教育プログラムに参加した学生には、次世代シーケンサー解析用ソフトをプレインストールしたノートパソコンを貸与し、解析演習を実施する。さらに強力な解析が必要な場合には、ネットワークを介して、リモートログインにより接続して、解析を行うこととした。

3.5　農学領域におけるゲノム科学のビッグデータのデータ解析に関する情報科学教育の実施実績

表 3 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。

平成 23 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･インフィマティクス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。

また、平成 24 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･インフィマティクス分野の教育指導を行った。

平成 23 年度と平成 24 年度で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。

表 3　ゲノム科学人材育成プログラムの実施実績（平成 23 年度および平成 24 年度）

Table 3　Activity of Education Program for Genomic Sciences in Tokyo University of Agriculture and Technology(FY2011-2012)

期間	全応募数（担当人数）
平成 23 年度	
第 1 期(7～9 月)	12 名 (内 7 名)
第 2 期(10～12 月)	14 名(内 8 名)
第 3 期(1～3 月)	11 名(内 7 名)
平成 24 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）

3.6　講習会･セミナーの実施

個別指導によるトレーニングで、データ解析の実績を上げていくことは可能であるが、それだけだとどうしても、リテラシーが不足しがちである。このため、適宜セミナーを実施した。

1) 平成 23 年度のセミナー･講習会の実施

平成 23 年度には､次世代シーケンサーや質量分析機や関連試薬の販売、データ解析の受託を行っている企業の技術者や営業担当者にセミナーを行っていただいた。表 4 のように､計 11 回のセミナーを行いのべ 250 名以上の参加があった。

さらに、次世代シーケンサーのデータ解析ソフトをプレインストールしたノートパソコンを用いて Ruby と R の 2 回の講習会を行った。

表 4　ゲノム科学人材育成プログラムのセミナーの実施実績(平成 23 年度)

Table 3　Seminars of Education Program for Genomic Sciences in Tokyo University of Agriculture and Technology(FY2011)

期間	セミナーの内容(受講人数)
9/2	次世代シーケンサー(39)
9/22	次世代シーケンサーデータ解析(36)
10/6	次世代シーケンサー(43)
10/25	質量分析装置データ解析(30)
11/7	次世代シーケンサー(22)
11/14	qPCR (28)
11/22	次世代シーケンサーデータ解析(13)
12/7	次世代シーケンサーデータ解析(15)
12/20	質量分析装置(19)
1/18	次世代シーケンサー(9)
11/22	次世代シーケンサー(7)

2) 平成 24 年度のセミナー･講習会の実施

平成 24 年度には、セミナーを 4 回、講習会を 3 回実施した。さらに公開講座や、データ解析相談会を行いデータ解析の普及に務めた。

表 4　ゲノム科学人材育成プログラムのセミナーの実施実績(平成 24 年度)

Table 3　Seminars of Education Program for Genomic Sciences in Tokyo University of Agriculture and Technology(FY2012)

期間	セミナー･講習会の内容(受講人数)
7/13	セミナー･共焦点顕微鏡(19)
7/18	セミナー･次世代シーケンサー(11)
7/31	講習会･データ解析(36)
8/16	セミナー･次世代シーケンサー(20)
9/3	講習会･データ解析(37)
9/11	セミナー･次世代シーケンサー(9)
10/4	公開講座･データ解析(5)
11/5	講習会･データ解析(26)
11/9-11	データ解析相談会(14)

2) 国際セミナーの実施

ゲノム解析は、中国や、欧米の海外の大規模 DNA 解析センターで盛んに実施されており、ゲノム解析の状況を把握するにはどうしても海外のデータ解析事情を知る必要がある。

このため、中国の北京ゲノム研究所（BGI）から若手の研究者を招聘し、英語を主言語とする国際セミナーを 2012 年 9 月に実施した。講演内容は、2012 年 10 月に Nature 誌に発表される直前の (1) 牡蠣（カキ）のゲノム解析や、(2) 発現データを用いた系統樹解析、(3) 癌のエピゲノム解析および発現解析など、最先端の内容であった。世界最大規模の研究所での最先端の内容の研究発表を聞く機会は、大学院生二は非常に刺激になったものを思われる。以下に、実施した講演者と講演内容をまとめたプログラムを示した。

講演内容:

講演 1 Xiaodong FANG, Plant and Animal Genome Group, BGI
Genome assembly and functional analysis of oyster

講演 2 Likai MAO, Plants and Animals Trans-omics, BGI
The power of phylogeny analysis based on transcriptome data

講演 3 Guangliang YIN, Scientist of Cancer Research of Human Transomics, cience&Techonology Division, BGI
Epigenomics and Transcriptomics Studies in Cancer Researches

以上のようにセミナー、講習会、公開講座、データ解析相談会などいろいろな形で、ゲノム解析やそのデータ解析の普及に務め、データ解析の必要性を訴えた。

3.7　研究成果報告会の実施

2012 年 11 月には、ゲノム解析によるデータ解析の成果を、本教育プログラムを実施した学生により学内で発表を行う成果発表会を実施した。学生からは 10 名、教員から 2 名が発表を行った。

研究発表会のタイトルを以下に示す。発表内容は、発現

解析、ゲノム配列解析、プロテオーム解析などであり、解析した生物も植物、細菌、マウス、昆虫など多彩な内容であった。

発表プログラム
1) トランスクリプトーム解析による植物過敏感細胞死と抵抗性誘導機構の研究
2) トマト萎凋病菌の宿主特異性に関わる因子の網羅的解析
3) 土壌伝染性 *Fusarium oxysporum* の病原力・非病原力因子に関する解析
4) イネいもち病菌の付着器形成時特異的発現遺伝子 *CBP1* における転写制御機構
の解明
5) 温湯消毒時における種子成熟期の水稲種子が示す高温耐性の解析
6) Genome-wide mass spectrometry-based RNA analysis reveals a novel snRNA
metabolic pathway in human cell
7) 麹菌 *Aspergillus oryzae* プロテアーゼ遺伝子の転写解析
8) 花の寿命を支配する分子機構の解明
9) mRNA-Seq によるハダニ科の系統解析
10) トランスクリプトームを用いた低酸素条件下におけるマウスメラノーマ細胞悪化の機構解析
11) トランスクリプトームを用いた、発現変動解析にかける実験区の選定方法
12) Genome Sequencing of the Liverwort Marchantia paleacea var. diptera

これらの発表の成果から、大学院生が、次世代シーケンサーから産生されたゲノム配列データを用いて、自らの手で UNIX/Linux をベースとしたフリーソフトによる情報処理技術によりデータ解析を行いことができるようになった。そして、その成果をまとめて発表することができるようになったことがわかる。

プログラミング、データベース、ネットワークを駆使したデータ解析は、農学系の大学院生には未知の世界であった。しかし、多方面からの個別指導、セミナー·講習会などの実践により、効果的な情報科学教育が達成できたことが示された。

4. 今後の課題

ゲノム科学における情報科学教育プログラムを開始して 2 年が経過しようとしている。これらの、教育実践により、大学院生が、情報処理技術を駆使してビッグデータ処理によるデータ解析を実施し、その研究成果を発表するまでに至った。今後のいくつかの課題として以下のような事があげられる。

1）情報科学·統計科学教育の認知度アップと定着

農学系学部において情報科学·統計科学教育を実施して、それを芽吹かせることができるようになった。これらにより、その重要性や威力が少しずつ認知されるようになったと思われる。しかし、未だ正規の教育過程までには至っておらず、全学的な認知もまだまだ遠い先のことである。今後のよりいっそうの教育成果を積み重ねて、農学系学部における情報科学·統計科学教育の定着を図る必要がある。

2）教材や教科書の整備

ゲノム科学における情報処理に関してまとまった教科書が存在しない。農学系学部の学生が学ぶために必要な教科書が存在していない。このための教科書や教材を整備する必要がある。

謝辞 農学系ゲノム科学人材育成プログラムの実践的活動に賛同し、ご協力いただいた教員の先生方、ならびに参加された大学院生の皆様に，謹んで感謝の意を表する．

参考文献

1) 石井一夫: 日本発のプログラミング言語 Ruby：Web アプリケーションからバイオインフォマティクスまで, 技術士, pp.8-11 (2008)
2) 樋口千洋、石井一夫: 統計解析環境 R によるバイオインフォマティクスデータ解析, 共立出版社 (2007)
3) 石井一夫, 村田真樹訳: R による計算機統計学, オーム社 (2011)
4) 石井一夫: 図解よくわかるデータマイニング, 日刊工業新聞社 (2004)
5) 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
6) 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