技術論文

「KAIKObase：カイコゲノム統合データベース」の紹介

Introduction of “KAIKObase : An integrated silkworm database”

下村 道彦*
Michihiko Shimomura

カイコは、農業に多大な被害を与える農業害虫が多く含まれる鱗翅（チョウ）目に属し、カイコゲノム解読は、殺虫剤を含む新しい防除に関する方法の開発などに役立つと期待されている。カイコゲノム研究は、日本、中国共同で研究が進められ、その成果が2008年に公開された。これらの公開されたデータを見易く使い易い形で、研究者や一般の方々に提供するために作られた「KAIKObase：カイコゲノム統合データベース」について紹介する。

Silkworm belongs to order Lepidoptera that includes moths and butterflies as well as harmful insects that can cause extensive damages to food crops. As a reference for Lepidoptera, silkworm genome sequencing is expected to facilitate the development of genome-based approaches for sericulture and pest control. The results of extensive silkworm genome analysis being conducted in collaboration with China has been published in 2008. Here I introduce the integrated silkworm database called KAIKObase which was developed to provide researchers and the general public with a repository of silkworm data in a user-friendly interface.

1. はじめに

近年のゲノム解析技術の進歩には目を見張るものがある。1998年にセンチュウに属するCaenorhabditis elegans[1]が解読され、1999年にはHomo sapiens (chr22)[2]、昆虫の分野でも2000年にDrosophila melanogaster[3]、その後Anopheles gambiae[4]、Apis mellifera[5]、およびTribolium castaneum[6]の昆虫ゲノムが公開された。独立行政法人農業生物資源研究所（以降、生物研と呼ぶ）を中心としたグループが、2002年にカイコゲノムのWGSシーケンスのリードを行なったことを契機に、カイコゲノム研究が加速度的に進み、2004年には、日本、中国別々にカイコゲノムの解析論文[7][8]が、ほぼ同時期に発表された。その後、2008年に、日本、中国の共同研究により、双方のデータをあわせたカイコゲノム[9]が公開された。このデータを公開する日本側のサイトとして、「KAIKObase：カイコゲノム統合データベース」[10]が公開（http://sgp.dna.affrc.go.jp/KAIKObase/）され、中国側でもSilkDB[11]が公開された。

カイコは、約5000年前、中国の黄河流域でクワコ（野生種のカイコ）から家畜化された昆虫であり、動物の中で最も家畜化が進んだ昆虫である。発展途上国において、絹糸などの一次産品の生産が活発に行なわれているほか、先進国においては、絹糸やカイコ自身の機能を使い、様々の製品が開発されている。また更なる製品の開発研究も行なわれている。一方、カイコは、農業に多大な被害を与える農業害虫が多く含まれる鱗翅目に属し、この目の中で、唯一ゲノム解析が行なわれた昆虫であり、かつ、これら鱗翅目昆虫の代表（モデル生物）でもある。また、バイオテクノロジの発展に伴い、カイコは組換えタンパクを作るためのバイオリアクタとしても大きな注目を集めている[12][13]。このカイコゲノム情報は、養蚕業に強い影響を与えるだけでなく、殺虫剤を含む新しい防除に関する方法の開発[14]に大きく寄与するものと考えられている。

解析されたゲノム情報を公開する上で、インターネットの様々な道具立てが作り出されてきた。パイオニアは、C. elegans[1]のAceDB[15]が遺伝的地図と物理的地図をもったゲノム情報を表示する道具立てである。この遺伝的地図と物理地図を表示する方法は、INE[16]、NCBI map viewer[17]、Cmap[18][19]などで見ることができる。また、これとは別に、メガベース単位のゲノム情報を表示する道具立てとして、Ensemble[20]、GBrowse[21]、

*つくば事業部 第三技術部

UTGB [22] などが開発された。

本報告は、カイコ地図情報、遺伝子情報などを統合し、新しいビューアなどが組合わされたユーザにフレンドリなカイコゲノム統合データベース”KAIKObase：an integrated silkworm genome database and data mining tool”[10] の紹介である。

2. カイコ研究分野

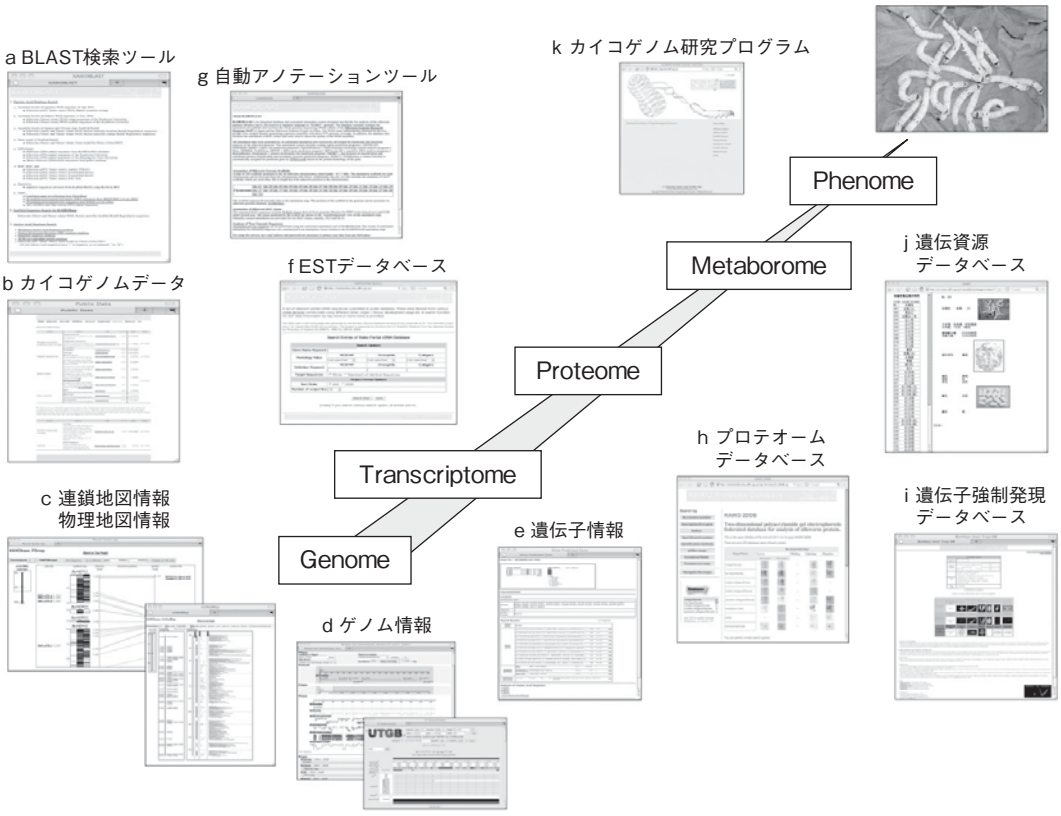
カイコ研究分野において、生物研を中心に開発されてきた様々なデータベースやツールをオーミックス的な分類で仕分けした（図１）。図１に示したオーミックス分類に従い、ゲノム、トランスクリプトーム、プロテオーム、メタボローム、およびフェノーム毎に、開発されてきたデータベース、ツール等を概説する。

2.1 ゲノム（Genome）

ゲノムは、遺伝子と遺伝子以外の配列の全てを指し、染色体の配列とほぼ同じような意味を持つ。 ゲノムの特徴を表すGCコンテンツ、遺伝子分布、リピート配列分布、セントロメア位置等のゲノム構造に関わる情報、遺伝子情報、翻訳開始点などの情報をユーザにとって見易く、かつ関連情報が直ぐに得られるデータベース/情報表示ツールが必要である。カイコゲノムでは、a）連鎖地図および物理地図情報を概観できるブラウザ、b）ゲノム情報（ゲノム上の予測遺伝子、マーカなど）が閲覧できるゲノムブラウザ、c）遺伝子情報（遺伝子モデルなど）が閲覧できるビューアが準備されている。

2.2 トランスクリプトーム（Transcriptome）

トランスクリプトームは、細胞内にある遺伝子転写物の総体で、DNAからコピーされた遺伝情報などを持つmRNA（メッセンジャーRNA）である。mRNA情報は、組織や生育段階で作られる種類や量が異なる。これらをクローン化する技術に、mRNAからcDNAライブラリを



ゲノム、トランスクリプトーム、プロテオーム、メタボローム、およびフェノームというオーミックス分類毎に、データベースおよび解析ツールを分類した。a）BLAST検索ツール：カイコゲノム配列などをデータベースにした検索ツール；b）カイコゲノムデータ：カイコゲノムスキャフォールドデータ、予測遺伝子データなどの公開データ；c）連鎖地図情報-物理地図情報：連鎖地図と物理地図を対比的に表示；d）ゲノム情報：ゲノムブラウザGBrowse、UTGB；e）遺伝子情報：公開されている情報との比較など表示；f）ESTデータベース：カイコ由来のESTをまとめたデータベース；g）自動アノテーションツール：カイコゲノムを中心とした遺伝子予測ツール；h）プロテオームデータベース：カイコ由来のプロテオーム情報をまとめたデータベース；i）遺伝子強制発現データベース：特定の遺伝子をゲノムに導入し、その遺伝子を強発現させたトランスジェニックのデータベース；j）遺伝子資源データベース：遺伝子資源をカタログ化；k）カイコゲノム研究のホームページ。

図１　カイコゲノム関連データベースおよび解析ツール

作成し、そこからcDNAを選ぶ方法がある。カイコゲノムでは、同じ機能を持ったEST（partial cDNAシーケンス）をまとめたESTデータベース、カイコ自動アノテーションツールで生成された遺伝子モデルのデータベースなどが準備されている。

2.3　プロテオーム（Proteome）

プロテオームは、組織にあるタンパク質の総体である。例えば生育段階の違いによるタンパク質の出現頻度や違いなどを比較することにより、生命現象の解明などに用いる。解析でよく使われる方法として、組織別、生育段階別に得られたタンパク質を、先ず等電点、次に分子量よって分類する二次元電気泳動を行なう方法がある。カイコゲノムにおいても、組織別、生育段階別に得られたタンパク質を二次元電気泳動にかけたプロテオームデータベースが準備されている。

2.4　メタボローム（Metabolome）

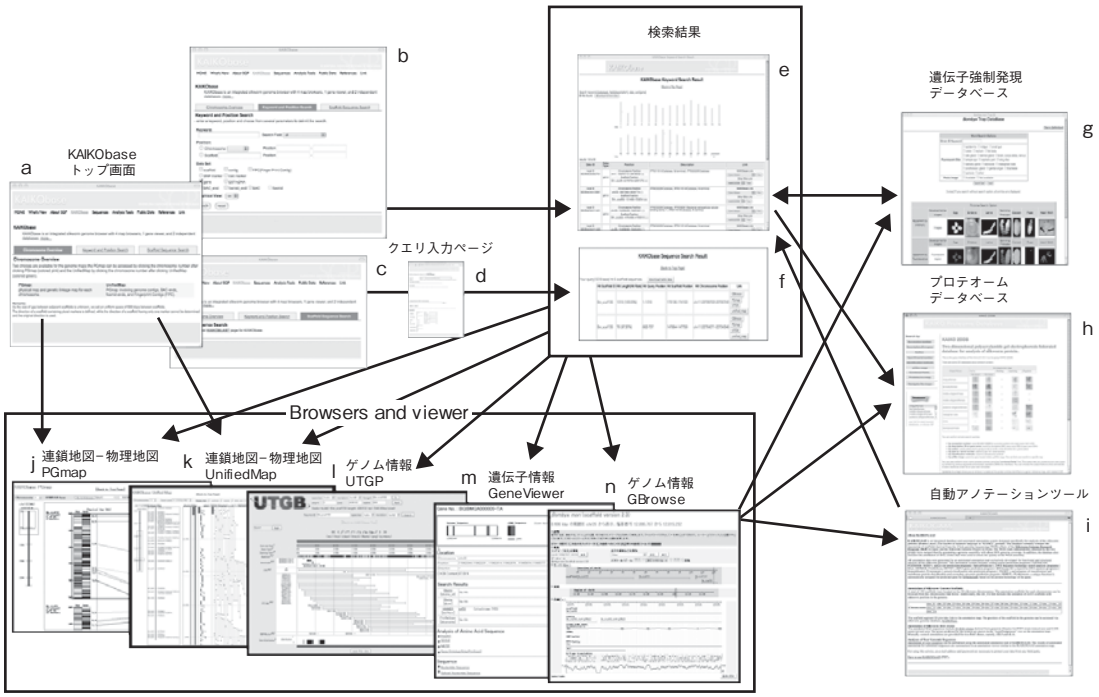
メタボロームは、代謝経路や代謝物質の総体で、酵素（触媒）によって作り出される糖や有機酸などの低分子物質がある。カイコにおいては、代謝系に関するデータを扱うメタボローム情報は、いくつかの報告がなされ始めており、今後の研究が待たれる。

2.5　フェノーム（Phenome）

フェノームは、表現型の総体で、大きさの違い、色の違いなどである。カイコにおいては、カイコ形質ミュータントを集めたデータベースや特定の遺伝子を、トランスポゾンを利用してゲノムに導入して強制的に発現させ、その遺伝子が導入されたゲノム上の位置をまとめたデータベースなどが準備されている。

3. KAIKObase [10]

カイコ研究分野で紹介したデータベースやブラウザなどを統合したデータベース/ツール群KAIKObase：カイコゲノム統合データベースである（図２）。ここで使用されているブラウザ/ビューアは、連鎖地図および物理地図情報を表示するPGmapおよびUnifiedMap、ゲノム情報を表示するGBrowse、およびUTGB、遺伝子情報を表示するGeneViewer、データベースは、プロテオームデータベース、遺伝子強制発現データベース、および自動アノテーションツールである。



a）KAIKObaseトップページで、PGmapとUnifiedMapへのリンクを持つ；b）キーワード検索；c）配列検索（BALST検索）；d）配列検索のパラメータ、配列入力画面；e）キーワード検索結果；f）配列検索結果；g）遺伝子強制発現データベース；h）プロテオームデータベース；i）自動アノテーションツール；j）PGmap の表示イメージ；k）UnifiedMapの表示イメージ；l）UTGB の表示イメージ；m）GBrowseの表示イメージ；n）GeneViewerの表示イメージ。

図２　KAIKObase閲覧のためのフォローチャート

3.1　ユーザインタフェース

カイコゲノムに関する一連の膨大な情報からデータをマイニングする方法として、⑴ブラウザ上の図から情報を絞り込んで行く（ドリルダウン）方法、⑵キーワード検索で、直接閲覧したいものに飛び込んで行く方法、⑶BLAST検索ツールを使用して、塩基配列やアミノ酸配列のクエリから該当するものを探す方法が準備されている。

3.2　ゲノム配列

KAIKObase には、計482Mbの43,462配列 [9] のスキャフォールド、およびコンティグが格納されている。カイコの28本の染色体には、192本のスキャフォールド、およびコンティグがマップされており、これは、カイコゲノムの88%をカバーする長さである。更にKAIKObaseには、81,705 BAC端配列、174,222 fosmid端配列、および166,757 EST配列が格納されている [9]。

3.3　マップ情報

マップ情報として、133,775の遺伝子モデル、1,532のSNPマーカ、770の形質マーカ、および5,419のFPCコンティグが格納されている。遺伝子モデルおよび遺伝子は、中国グループが、GLEAN-basedアルゴリズム [23] によって作り出した14,622の遺伝子モデル、日本グループが自動アノテーションツール（KAIKOGAAS）によって作り出した遺伝子モデル、並びにマニュアルアノテーションにより作り出された1,587のGPCR、OBP、CSP、クチクラタンパク質、およびtRNA [9] の遺伝子である。SNPマーカは、BAC端配列で同定され、形質マーカは、トランスポゾンを利用して導入された遺伝子のゲノム上の位置を表している。FPCコンティグは、BACクローンをフィンガープリント法によりアセンブルしたBACコンティグである。

3.4　プロテオームデータベース [24]

プロテオームデータベースでは、タンパク質の情報が提供され、生育段階、および組織の異なる116の二次元電気泳動イメージが格納されている。各イメージには、分子量、等電子点がスポットされており、これらは、カイコESTや中国グループが予測した遺伝子モデルに対応している。

3.5　遺伝子強制発現データベース

遺伝子強制発現データベースでは、トランスポゾンを利用して導入された遺伝子のゲノム上の位置などの情報を持つ、288個のトランスジェニックカイコの情報が格納されている。

3.6　自動アノテーションツール

自動アノテーションツールは、様々な遺伝子予測ソフトウェアを使ったカイコ用の自動遺伝子予測ツールとデータベースである。このデータベースには、1Kb以上のカイコゲノム（スキャフォールド、コンティグ）配列のアノテーション結果、および55本のBACのアノテーション結果が含まれている。

なお、KAIKObaseの詳細については、参考文献⑽を参照されたい。また、KAIKObaseの使用方法はKAIKObaseのホームページにある使用法を参照されたい。

4. まとめ

ゲノムから形質に至るデータを、データベースやブラウザを連携させ、統一的に閲覧できるKAIKObase：カイコゲノム統合データベースが開発された。今後、不足している情報や欠落している情報の補完、より使い易いユーザインタフェースが構築できれば、研究者のみならず、様々な方々がより効率的に利用できるツールになるものと確信している。

用語説明

ゲノム：遺伝子と遺伝子以外の配列の全て；トランスクリプトーム：細胞内にある遺伝子転写物の総体；プロテオーム：組織にあるタンパク質の総体；メタボローム：代謝経路の総体；フェノーム：表現型（形に現われること）の総体；GCコンテンツ：塩基配列GとCの含有量；BLAST検索：配列同士の相同性を調べるツール；アノテーション：遺伝子などに意味（機能）を付与すること；WGS：Whole genome shotgunの略、染色体の端から端まで読む技術がないため、染色体を微小断片に分割して、配列を読み、そこで得られた配列をコンピュータでつなぎ合わせていく方法；コンティグ：WGSで得られた断片配列が連続的に並んでいる配列；スキャフォールド：コンティグ配列を、ギャップを含んだ形で並べた配列；FPCコンティグ：フィンガープリント法（クローンの重なりを推定する方法）を使ったBACクローンを整列させたコンティグ（固まり）；連鎖地図：組み換えや交差から遺伝子間の距離を基に作成する地図；GPCR：G protein-coupled receptor Gタンパク質共役受容体；OBP：Odorant binding proteins匂い輸送タンパク質；CSP：Chemosensory proteins 化学受容タンパク質および；tRNA：transfer RNA 転移RNA。

謝辞

社内報への発表を許諾して頂いた方々に感謝致します。

参考文献

(1) The C. elegans Sequencing Consortium：Genome sequence of the nematode C. elegans：a platform for investigating biology. Science, 282, 2012-2018（1998）.

(2) Dunham I., et al.：The DNA sequence of human chromosome 22. Nature, 402, 489-495（1999）.

(3) Adams MD, et al.：The genome sequence of Drosophila melanogaster. Science, 287, 2185-2195（2000）.

(4) Holt RA, et al.：The genome sequence of the malaria mosquito Anopheles gambiae. Science, 298, 129-149（2002）.

(5) Honeybee Genome Sequencing Consortium：Insights into social insects from the genome of the honeybee Apis mellifera. Nature, 443, 931-949（2006）.

(6) Tribolium Genome Sequence Consortium：The genome of the model beetles and pest Tribolium castaneum. Nature, 452, 949-955（2008）.

(7) Mita K, Kasahara M, Sasaki S, Nagayasu Y, Yamada T, Kanamori H, Namiki N, Kitagawa M, Yamashita H, Yasukochi Y, Kadono-Okuda K, Yamamoto K, Ajimura M, Ravikumar G, Shimomura M, Nagamura Y, Shin-I T, Abe H, Shimada T, Morishita S, Sasaki T：The genome sequence of silkworm, Bombyx mori. DNA Res., 11, 27-35（2004）.

(8) Biology analysis group, Xia Q, et al.：A draft sequence for the genome of the domesticated silkworm（Bombyx mori）. Science, 306, 1937-1940（2004）.

(9) The International Silkworm Genome Consortium：The genome of a lepidopteran model insect, the silkworm Bombyx mori. Insect Biochem. Mol. Biol., 38, 1036-1045（2008）.

(10) Shimomura M, Minami H, Suetsugu Y, Ohyanagi H, Satoh C, Antonio B, Nagamura Y, Kadono-Okuda K, Kajiwara H, Sezutsu H, Nagaraju J, Goldsmith MR, Xia Q, Yamamoto K, Mita K：KAIKObase：an integrated silkworm genome database and data mining tool. BMC Genomics, 10, 486（2009）.

(11) Duan J, Li R, Cheng D, Fan W, Zha X, Cheng T, Wu Y, Wang J, Mita K, Xiang Z, and Xia Q：SilkDB v2.0：a platform for silkworm（Bombyx mori）genome biology. Nucleic Acids Res., 38, D453-D456（2010）.

(12) Tamura T, Thibert C, Royer C, Kanda T, Abraham E, Kamba M, Komoto N, Thomas JL, Mauchamp B, Chavancy G, Shirk P, Fraser M, Prudhomme JC, Couble P：Germline transformation of the silk- worm Bombyx mori L. using a piggyBac transposon-derived vector. Nat. Biotechnol., 18, 81-84（2000）.

(13) Tomita M, Munetsuna H, Sato T, Adachi T, Hino R, Hayashi M, Shimizu K, Nakamura N, Tamura T, Yoshizato K：Transgenic silk- worms produce recombinant human type III procollagen in cocoons. Nat. Biotechnol., 21, 52-56（2003）.

(14) 山本、三田：カイコゲノム研究の現状とその利用. 植物防疫, 63：12, 735-740（2009）.

(15) Durbin R, Thierry-Mieg J：The AceDB genome database. In Computational Methods In Genome Research Edited by：Suhai S. New York：Plenum Press, 45-55（1994）.

(16) Sakata K, Antonio BA, Mukai Y, Nagasaki H, Sakai Y, Makino K, Sasaki T：INE：a rice genome database with an integrated map view. Nucleic Acids Res., 28, 97-101（2000）.

(17) Dombrowski SM, Magiott D：Using the Map Viewer to Explore Genomes. The NCBI Handbook（http://www.ncbi.nlm.nih.gov/ books/ bv.fcgi?rid=handbook.chapter.ch20）.

(18) Ware DH, Jaiswal P, Ni J, Yap IV, Pan X, Clark KY, Teytelman L, Schmidt SC, Zhao W, Chang K, Cartinhour S, Stein LD, McCouch SR：Gramene, a tool for grass genomics. Plant Physiol., 130, 1606-1613（2002）.

(19) CMap：Visualization of Comparative Maps（http://gmod.org/ wiki/CMap）

(20) Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta- Vidal A, Vastrik I, Clamp M：The Ensembl genome database project. Nucleic Acids Res., 30, 38-41（2002）.

(21) Stein LD, Mungall C, Shu SQ, Caudy M, Mangone M, Day A, Nicker- son E, Stajich JE, Harris TW, Arva A, Lewis S：The generic genome browser：a building block for a model organism system data- base. Genome Res., 12, 1599-1610（2002）.
(22) Ahsan B, et al.：UTGB/ medaka：genomic resource database for medaka biology. Nucleic Acids Res., 36, D747-D752（2008）.
(23) Elsik GM, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM：Creating a honey bee consensus gene set. Genome Biol., 8, R13（2007）.
(24) Kajiwara H, Nakane K, Piyang J, Imamaki A, Ito Y, Togasaki F, Kotake T, Murai H, Nakamura M, Mita K, Nomura R, Shimizu Y, Shimomura M, Ishizaka M：Draft of silkworm proteome database. J Electrophoresis, 50, 39-41（2006）.