オミックス・プラットフォーム：バイオ・ビッグ・データに挑む

池田　俊・桂樹　哲雄・小野　直亮・中谷　淳至
中村由紀子・森田　晶・金谷　重彦*

ゲノム，トランスクリプトーム，プロテオーム，メタボロームに代表されるオーム情報をいかに効率よく解釈するかということはバイオインフォマティクスの大きな課題である．新型シーケンサ（NGS）の出現により，生物さらにはヒト個人のゲノム，トランスクリプトーム情報が生産される．そこで得られる情報を如何に理解するかという視点で，筆者らが構築するデータベース（KNApSAcK Family）を紹介する．さらに，ヒトの健康維持の恒常，植物代謝物多様性の理解においてバイオ・ビッグ・データをいかに解析するかについて考察する．

オミックス・プラットフォーム

ヒトは，動植物・微生物から栄養を獲得し恒常性を維持する．すなわち，食物の機能性を考慮し，最適化することによりヒトの健康は維持される．このことを，ゲノムサイエンスの知識により図1のように表すことができる[1–6]．このプラットフォームに従ってデータベース（KNApSAcK Family DB）の構築を進めてきた．

まずはじめに，人類の歴史により培われた食/薬用植物における知識を活用し，生薬においては日本における漢方薬（KAMPO DB，278薬用生物，336処方），インドネシアにおけるジャム（JAMU DB，1133薬用植物，5310処方）といったヒトの体調により組わせて摂取すべき食/薬用植物の情報を体系化しDBに整理した．食用に使用されている生物を食用DB（Lunch Box，709生物），さらにはハーブとして利用されている生物をハーブDB（Tea Pot）として整理し公開した．また，国ごとの食/薬用植物の利用を把握する目的で，世界の薬用植物データベースを構築した（WorldMap DB，222地域，50,000地域－薬/薬用植物の関係）．さらに生薬を含む生物において生合成される代謝物を科学文献より網羅し，生物種－代謝物関係DB（KNApSAcK Core，101，500

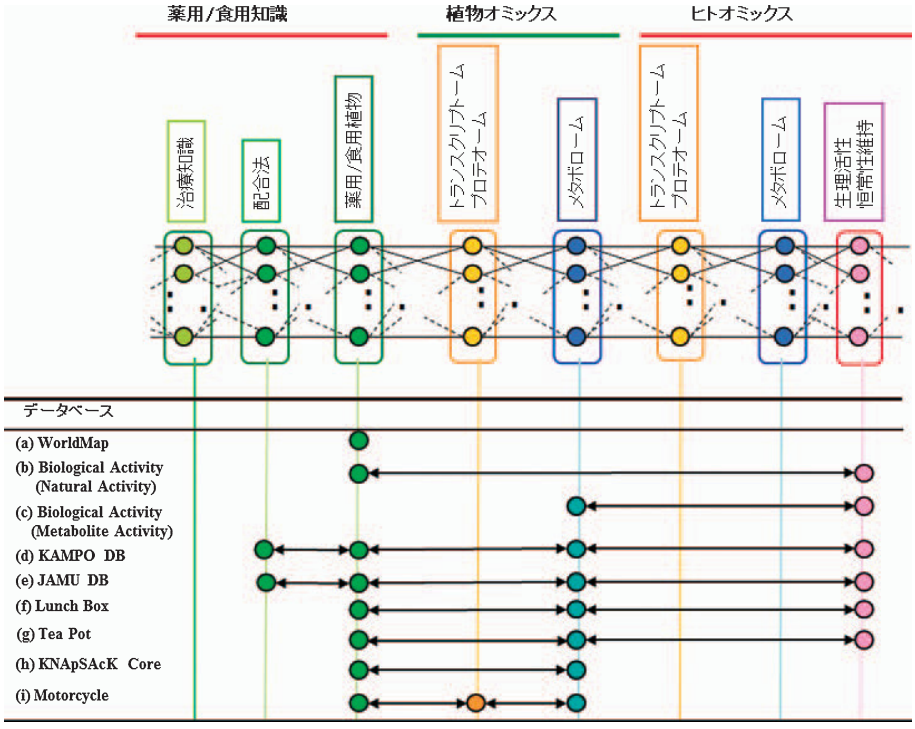


図1．植物-ヒト相互作用における健康維持のためのオミックスプラットホーム

*著者紹介　[1]奈良先端科学技術大学院大学情報科学研究科計算システムズ生物学研究室　E-mail: skanaya@gtc.naist.jp

対の生物種－代謝物の関係，5万代謝物種，2万生物種）を構築した．たとえば，タマネギは食欲増進に効果があるなど，その生物のヒトに対する活性および抗菌作用のようにその生物が他の生物に対してどのような活性があるかが報告されている．このような，生物が他の生物へ与える活性を生物活性DB（Biological Activity DB，2,418種の生物活性における3万対の植物とその活性の関係）として整理・公開した．同様に，個々の代謝物が生物へ与える影響を代謝物活性DB（Metabolite Activity DB）として整理した．また，薬用・植物における有効成分の生合成過程を把握する目的で，植物-酵素-酵素反応からなるデータベース（Motorcycle DB）の構築を進めている．さらに今後，ヒトの代謝経路と恒常性維持の関係を把握するためのデータベース構築を計画している．このようなデータベースを完備すると薬用/食用生物の知識からヒトの恒常性維持に関わるビッグ・データを基盤とした悉皆的な分子メカニズムの解明が可能となる期待される．その中で，特に，新型シーケンサ（NGS）などのDNA配列決定技術の進歩に伴い，遺伝子診断ができるようになった時に，ヒトの健康恒常性を維持するためのパーソナル健康戦略を立て，そのためにどのような栄養をとるべきかという医・薬・栄養学を総合的に融合した情報学の展開が望まれる．

植物代謝物の多様性

地球上で植物が生合成する代謝物種はどのくらいあるのだろうか．この問いに答えられれば，未知の代謝物がどのくらいあり，特に，ヒトの健康と関わる代謝物がどのくらい自然界に存在するのかといった予測をたてることができる．KNApSAcK Coreにおける生物種と代謝物の関係から地球上の植物が生合成することのできる代謝物種の数の推定を試みた．KNApSAcK Coreにおける生物種と代謝物の関係から地球上の植物が生合成することのできる代謝物種の数を推定した．多くの生物にとって共通の成分である一次代謝物ではなく，特定の生物グループに固有である二次代謝物に着目している．図2の横軸は生物種数の対数表示であり，縦軸は代謝物数の対数表示である．すなわち，図の左端は1つの生物のみで報告されている代謝物種が4000種程度あることと対応する．この図から，多くの生物種で報告されている代謝物の数の関係に注目すると，生物種数と代謝物数の間には破線のような明瞭な直線関係が得られる．一方，少ない生物種で報告されている代謝物の数の間には，このような直線性は得られず，むしろ生物種数が1に近づくに従って，代謝物数の増分は小さくなる．このことは，一つの生物において固有の代謝物を発見することが任意の生物で見つかった代謝物の他の生物の有無の確認に比べて，非常に困難であるという天然物構造決定の研究における難しさの特徴が反映されており興味深い．地球上

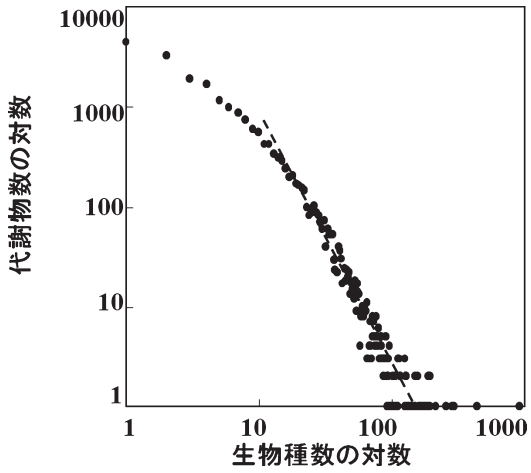


図2．生物種数の対数と代謝物数の対数．たとえば一つの生物種のみで報告されている代謝物種の数は約4000種であるという関係をプロットした．

の生物種数と報告されている代謝物数の関係が線形性（破線）を保たれていることを利用して，KNApSAcK Core DBの中で蓄積されるべき代謝物数を推定すると98,800代謝物となる．顕花植物において現在までに化学構造が決定された代謝物の総数は多く見積もっても10万程度であろうと考えられるので，本データベースに含まれる代謝物種の推定数である98,800は妥当であると考えられる．

いまKNApSAcK Core DBには20,741生物種が格納されている．一方，地球上の顕花植物の数223,300種を考慮すると，地球上で顕花植物が生合成する代謝物種の数は106万種と推定される[1]．いままでに構造決定された代謝物数が10万種であるとすると，メタボローム測定において既知化合物が出現する確率は10%程度である．すなわち，メタボロミクスにおいては，地球上の代謝物のうち約1割の構造既知の代謝物により科学的考察が加えられていることになる．つまり，多くの新発見へとつながる新規代謝物が，まだまだメタボロームには潜んでいることになる．

今後，NGSにおけるゲノムおよびトランスクリプトーム情報の蓄積により，106万種の二次代謝物を生合成するための酵素が解明されていくと期待される．つまり，NGSから得られる大量DNA配列から二次代謝の合成過程を予測することが今後必要になる．そこで筆者らの研究室では，文献情報をもとに酵素のアミノ酸配列と代謝反応を関連づけるためのデータベース（Motorcycle DB）の開発を進めている．以降では，このMotorcycle DBを紹介する．

二次代謝反応データベース（Motorcycle DB）

地球上全体の顕花植物で生合成される二次代謝物（約

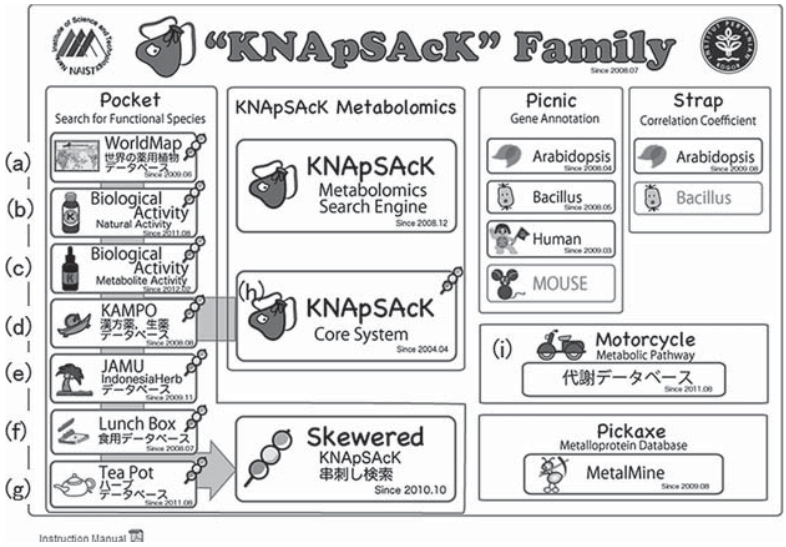


図3．KNApSAcKメインウインドウ，図中の(a)–(i)は図1と対応する．http://kanaya.naist.jp/KNApSAcK_Family/

106万種）の多様性を遺伝子レベルで理解することが，このように多様な二次代謝物を創出する進化のメカニズムを解明することへつながる．そこで，これらの二次代謝に関わる酵素の情報を生物種，酵素反応，酵素のアミノ酸配列と関連づけたデータベース（Motorcycle DB）の構築を進めている．Motorcycle DBの利用法を紹介しつつ，配列情報と酵素反応の多様性の関係を検討する．

まず，Google検索エンジンにおいてキーワードを「KNApSAcK Family」と入力すると，図3のメインウインドウへのリンクがみつかるだろう．このウインドウの左側のPocketは生物および代謝物の機能性を検索するために開発されたDBである．また，ユーザーが測定した質量スペクトルの精密質量から代謝物候補を列挙するためのシステムMetabolomics Search Engineと，生物種－代謝物関係データベースKNApSAcK Core Systemの二つがKNApSAcK Metabolomicsから公開されている．メインウインドウの下方にpdf形式でマニュアルが用意されている．

代謝データベースMotorcycleは，メインウインドウの図3の(i)をクリックすることによりアクセスできる．図3の(i)をクリックすると，図4の画面が得られる．代謝物から酵素とその反応を検索したいときには，代謝物名を図4のように入れて，ラジオボタンEquationを選択の後，リストボタンをクリックすると，図5に示す酵素反応リストが得られる．Motorcycle DBでは，基質と生成物が既知の酵素反応の登録を進めている．検索結果として酵素反応が出力される点が本データベースの特徴である．このリストから反応IDにおいてKR0001659を選択すると，この反応の詳細情報を得ることができる（図6）．さらに図6におけるReaction Mechanismをクリックすると反応メカニズムが得られる（図7）．

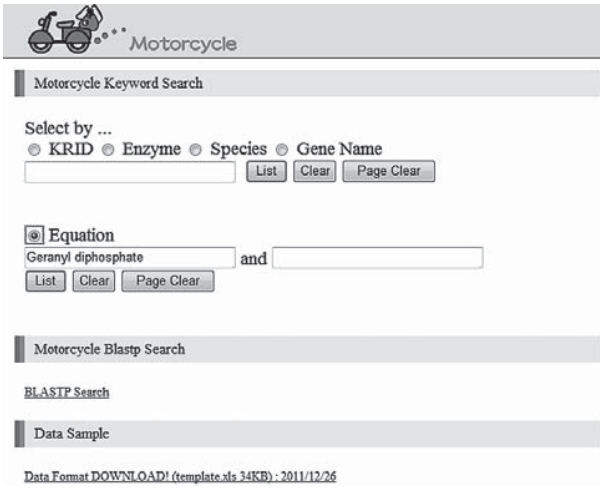
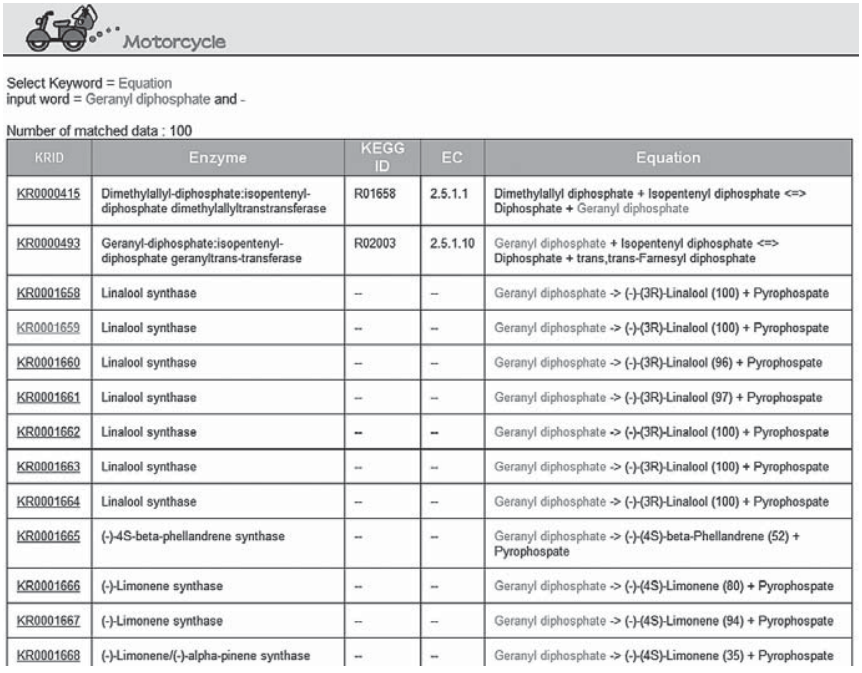


図4．Motorcycleウインドウ，Geranyl diphosphateと入力しEquationを選択しListをクリックする

ペプチド配列から酵素反応を検索することも可能である．例として，図4のBLASTP SearchによりLinalool synthase（QH1遺伝子，KR0001659）のペプチド配列をもとに反応検索を行うと図8が得られる．出力結果において，酵素大分類，代謝物種の大分類と小分類が得られる．QH1の場合，自分自身以外の検索結果においても酵素大分類はMonoterpene synthase，代謝物種の大分類Terpene，小分類はMonoterpeneと出力されるため，この酵素の特徴を代謝反応として得ることができる．一方で，ペプチド配列が非常に高いレベルで類似であってもその代謝反応は非常に多様であることがMotorcycle



Motorcycle

Select Keyword = Equation
input word = Geranyl diphosphate and -

Number of matched data : 100

KRID	Enzyme	KEGG ID	EC	Equation
KR0000415	Dimethylallyl-diphosphate:isopentenyl-diphosphate dimethylallyltranstransferase	R01658	2.5.1.1	Dimethylallyl diphosphate + Isopentenyl diphosphate <=> Diphosphate + Geranyl diphosphate
KR0000493	Geranyl-diphosphate:isopentenyl-diphosphate geranyltrans-transferase	R02003	2.5.1.10	Geranyl diphosphate + Isopentenyl diphosphate <=> Diphosphate + trans,trans-Farnesyl diphosphate
KR0001658	Linalool synthase	--	--	Geranyl diphosphate -> (-)-(3R)-Linalool (100) + Pyrophospate
KR0001659	Linalool synthase	--	--	Geranyl diphosphate -> (-)-(3R)-Linalool (100) + Pyrophospate
KR0001660	Linalool synthase	--	--	Geranyl diphosphate -> (-)-(3R)-Linalool (96) + Pyrophospate
KR0001661	Linalool synthase	--	--	Geranyl diphosphate -> (-)-(3R)-Linalool (97) + Pyrophospate
KR0001662	Linalool synthase	--	--	Geranyl diphosphate -> (-)-(3R)-Linalool (100) + Pyrophospate
KR0001663	Linalool synthase	--	--	Geranyl diphosphate -> (-)-(3R)-Linalool (100) + Pyrophospate
KR0001664	Linalool synthase	--	--	Geranyl diphosphate -> (-)-(3R)-Linalool (100) + Pyrophospate
KR0001665	(-)-4S-beta-phellandrene synthase	--	--	Geranyl diphosphate -> (-)-(4S)-beta-Phellandrene (52) + Pyrophospate
KR0001666	(-)-Limonene synthase	--	--	Geranyl diphosphate -> (-)-(4S)-Limonene (80) + Pyrophospate
KR0001667	(-)-Limonene synthase	--	--	Geranyl diphosphate -> (-)-(4S)-Limonene (94) + Pyrophospate
KR0001668	(-)-Limonene/(-)-alpha-pinene synthase	--	--	Geranyl diphosphate -> (-)-(4S)-Limonene (35) + Pyrophospate

図5.　Geranyl diphosphateによる検索結果

Select Keyword = KRID
input word = KR0001659

KRID	KR0001659
Enzyme	Linalool synthase
KEGG ID	--
EC	--
Equation	Geranyl diphosphate -> (-)-(3R)-Linalool (100) + Pyrophospate
C-class	Terpene
C-subclass	Monoterpene
FinalProduct	(-)-(3R)-Linalool
Eclass	Monoterpene synthase
Reaction Mechanism	ME000001.gif
Pathway	--
Curator	Shigehiko KANAYA

Species Name	Artemisia annua
Gene Name	QH1
AA Sequence	GNAYMRIYSTKTTRITANATVNAADTHVRRSANYKPSSWSFDHIQSLSSKYTGDDYVARANTLKDAVKTMIRKSGNSLRTLELVDELQRLGISYLFEEEISNLLETIYYNYYKFPENWNKINLNLKALGFRLLRQHGYHVPQEIFLNFKDKNQNLNSYLLNDVVEMLNLYEASYHSFEDESILDDARDITTKYLKESLEKIDGSIFSSVTHALEQPLHWRVPRVEAKWFIELYEKKNGMSPTLVELAKLDFDMVQAIHLEDLKHASRWWRDTSWDTKLTFARDLIVENFLWTIGFSYLPNFSRGRRTITKVAVMITTLDDVYDVFGTLGELEQFTDVINRWDIKAIEQLPDYMKICFLGLYKSINDITHETLANKGFLILPYLKKAWADLCKAYLVEAQWYHRGHIPTLNEYLDNACVSISGPVALMHVHFLTSVSSIEEIHQCIQRTENIVHYVSLIFRLADDLGTSLGEMERGDTLKSIQLHMHETGATEPEARSYIKLLINKTWKKLNKERATVNSESSQEFIDYATNLVRMAQFMYGEGDEDFGLDVIKSHVLSLLFTPIQGI	
DDBJ GenBank NCBI	AAF13357
Reference	Jia,Arch.Biochem.Biophysics,372,(1999),143

図6.　KR0001659の詳細情報

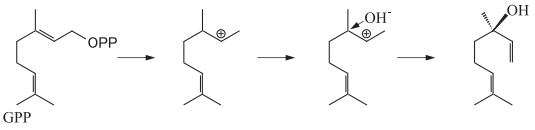


図7. KR0001659の反応メカニズム

図8. QH1遺伝子のペプチド配列による反応検索詳細情報

DBを使うことで把握することが可能である．このことは，NGSによる配列データをもとに，代謝経路を高精度で検討するときに非常に有益な情報となると期待される．現在までに，文献情報をもとに植物を中心に，モノテルペン合成酵素，セスキテルペン合成酵素，ジテルペン合成酵素，トリテルペン合成酵素，P450（CYP）酵素，アルカロイド合成，フラボノイド合成に関わる酵素の反応を整理し，DBへの蓄積がほぼ完了した．

今後の展望

ゲノム，トランスクリプトーム，プロテオーム，メタボロームに代表されるオーム情報をいかに効率よく解釈するかということは，医薬・薬学・栄養学・生物学などの分野の大きな課題である．情報科学では，たとえば世界のウエブ上での情報の集約と情報の流れといった，さらに大きな情報をビッグ・データとよび，そこから，統一解釈を導くためのマイニング研究が盛んに行われている．

バイオサイエンスにより生産されているデータは，このようなビッグ・データに比べれば非常に小さい．しかし，それでもかなり巨大化しており，今後も急速に拡大していくであろう．そして最終的には，生物間相互作用を地球規模で体系化しミクロな事象から地球全体を理解することが可能になるであろう．このことにより，まさに生物を主役とした地球の生態系の悉皆的理解が可能となり，人類が今後どのようにふるまうべきかといった地球規模問題への手がかりを見つけることができるかもしれない．そのキーとなる情報はメタゲノム，ゲノム情報であり，その相互作用を理解するにはメタボローム情報が必須となると考えられる．

また，NGSのさらなる能力向上によりヒト個人のゲノム情報を容易に取得できるようになり，ヒトゲノムの多様性とヒトのさまざまな個性との関係が明らかにされるであろう．その時，健康なヒトそれぞれがその健康をできるだけ長く維持するための手段，例えば，ヒトそれぞれの健康維持の恒常においてどのように食物をとるべきかという情報の提供は，医薬・薬学・栄養学・生物学などの事象をいかに情報科学が統合的に整理し，提供するかにかかっている．

こう考えると，バイオ・ビッグ・データの体系的理解に関する研究分野が構築されてもいい時代かもしれない．「バイオサイエンス研究者と情報科学者のシームレスな関係を構築すべきだ」などという文言をいろいろなところで見かけるが，その意味するところは，バイオサイエンス研究者が手におえないデータ処理を情報科学者に手伝ってもらいたいというのが真意である．しかし，これでは情報科学者の興味を引き続けることはできない．情報科学主導による地球環境とヒトの健康の研究ということをそろそろ真面目に考えるべきであろう．NGSから産出される膨大なデータの登場は，それを如何に効率よく使い新たな学問分野を切り開くかを考えるタイムリーなきっかけとなるのではなかろうか．

本研究はJST-NBDC，JST-CREST，科学研究費・新領域・バイオマシナリープロジェクトなどの支援のもとに進めている．ここに感謝する．また，本原稿をご依頼いただいた生物工学会に深謝する．

文　献

1) Afendi, F. M. *et al*.: *Plant Cell Physiol*., **53**, e1-e12 (2012).
2) Okada, T. *et al*.: *Curr Comput Aided Drug Des*., **6**, 179 (2010).
3) Afendi, F. M. *et al*.: *Curr. Pharm. Pers. Med*., **10**, 111 (2012).
4) 金谷重彦ら：細胞工学, **31**, 101 (2012).
5) 金谷重彦ら：実験医学, **29**, 2460 (2011).
6) 中村由紀子ら：バイオサイエンスとインダストリー, **70**, 267, (2012).