データベース
Databases

オミクスは、十分には予測できなかった知見を得て、医学や生物学における発見を加速させることが目的である。産出される膨大なデータを効率よく解析し、できるだけ早く発見に結びつけるには、研究者自身が得られたデータをこれまでの知見と比較し、また学習していかなければならない。これまでに人類が得た知識を最大限に利用するためには質の高いデータベースが欠かせない。メタボロミクス自体がまだ若い分野であるので、データベースも日々進歩している状態であるが、信頼性が高く、かつ利用しやすいデータベースが世界中の研究者の手でつくり出されている。ここでは、メタボロミクス分野で利用されている主なデータベースについて紹介する。

5.1　データベースの種類

メタボロミクスで利用されるデータベースには、1) 生データやメタデータなどの代謝プロファイルを蓄積したもの、2) 特定の生物種についての情報をまとめたもの、3) 多くの生物種の様々な状態での代謝プロファイルを含むもの、4) 各生物種において既知の代謝物質をリスト化したもの、5) すでに確立された生物学的情報をまとめたものが存在する[118]。これらは実用的には、1) 代謝物質リファレンススペクトルデータベース、2) 代謝プロファイルデータベース、3) 代謝パスウェイデータベース、4) 実験ワークフローを保存したLIMS(Laboratory Information Management System) データベースに分類される[119] (表 5.1)。LIMS データベースについては、直接的な研究用データベースとは異なるため、ここでは割

愛する。

5.2　データベース各論

5.2.1　代謝物質リファレンススペクトルデータベース

代謝物質を同定するために測定機器から得られるスペクトルを参照できるライブラリやデータベースは、特に GC-MS 法による測定分野では初期の頃から構築が進んでいる。これら初期のデータをもとに、マックスプランク植物生理学研究所 (独) では、GMD(the Golm Metabolome Database) というデータベースを公開している [120][121]。また、また、カリフォルニア大学デービス校 (米) のオリバー・フィーンのグループによる GC-MS スペクトルライブラリ [122] をもとにアジレント・テクノロジー社がメタボロミクス用の GC-MS ソフトウェアを開発している。

現在では GC-MS だけでなく、LC-MS や NMR スペクトルも含めたデータベースも構築されている。掲載物質数が膨大であり、最もよく利用されている米国立標準技術研究所 (NIST) の NIST11 には、EI-MS(electron ionization-MS) に 243,893 スペクトル (212,961 化合物)、10,065 イオントラップ MS スペクトル (4,628 化合物由来の 9,194 種のイオン)、85,344 コリジョンセルスペクトル (3,877 化合物由来の 7,172 種のイオン)、346,757 の GC データ (70,835 化合物) が登録されている。

MassBank は慶應義塾大学先端生命科学研究所の西岡孝明のグループによる MS スペクトルの公共データベースで、主要代謝物質だけでなく植物などの二次代謝物質を含む代謝物質を対象としている。検索機能が充実しており、統合スペクトル (merged spectrum) を対象とした検索により、精度と速度を向上している [123]。

METLIN は、スクリプス研究所 (米) のギャリー・シウスダクのグループによって構築された MS/MS スペクトルデータベースで [124]、4 万種以上の化合物情報が含まれている。

5.2.2　代謝プロファイルデータベース

HMDB(Human Metabolome Database) は、アルバータ大学のデイビッド・B・ウィシャートとカルガリー大学、国立ナノテクノロジー研究所 (カナダ) の共同プロジェクトにより構築された [125]。このデータベースは、疾患患者のデータを含め、ヒト体内に存在する代謝物質の百科事典であり、登録代謝物質数は 7,928 に達している (Ver.2.5, 2011 年 8 月現在)。各代謝物質情報は MetaboCard と呼ばれる情報コンテンツにまとめられ、物理化学的パラメータを含む化合物情報、報告されている血液 (血漿か血清かの区別はない)、脳髄液、尿

中濃度、関連する酵素情報が入手できる。

PubChem は、米国立バイオテクノロジー情報研究所 (NCBI) が運営するデータベースで、収載化合物数は 3,000 万、生化学反応は 1 億 3,000 万に達する (2011 年 8 月)。生体物質の基本情報とともに薬理作用や受容体への結合性などの情報が得られる [126 127]。

5.2.3 代謝パスウェイデータベース

代謝パスウェイデータベースは、遺伝子、酵素、代謝物質の情報にもとづき、生化学的パスウェイや反応を生物種ごとに定性的に描写するものである。

京都大学化学研究所の金久實のグループにより構築された KEGG(Kyoto Encyclopedia of Genes and Genomes) は、生物体内で相互作用する分子と環境で相互作用する生物に関する機能情報を理解することを目的としている [128 129]。データベースには代謝物質、生化学反応、転写、遺伝子情報が含まれ、それらが 1)PATHWAY と BRITE、2)GENES と ORTHOLOGY、3)LIGANDS の 3 つのカテゴリーに分類されている。メタボロミクスで最も用いられるのは PATHWAY であり、生物種に特有な代謝マップや遺伝子発現制御、酵素活性制御を可視化できる。LIGANDS はゲノムと代謝物質情報のギャップを橋渡しするようにデザインされており、代謝物質、薬剤、生化学反応情報が含まれている。ここから、各代謝物質や反応を検索することが可能である [130]。

BioCyc は米人工知能センター (Artificial Intelligence Center) のピーター・カープによって構築されたデータベースで、大腸菌 (EcoCyc)、酵母 (YeastCyc)、シロイヌナズナ (AraCyc)、ヒト (HumanCyc) 統合生物 (MetaCyc) の各パスウェイ / ゲノムデータベース (PGDBs) からなる [131]。代謝パスウェイが機能的に分類されており、反応描写も直感的に理解しやすい。

MeTaBoard は、HMT 社が独自に提供する代謝物質データベースであり、同社の解析サービス利用者に公開されている。代謝分子の生物学的知見や実試料での測定値を収載しており、メタボローム解析で得られた結果から標的分子が関わる考察を行う際に役立てることができる。現在登録されている内容は一次代謝に含まれる主要代謝分子のみであるが、今後は定期的なアップデートが予定されている。

表 5.1　メタボロミクスで利用される主なデータベース

データベース	特徴
リファレンススペクトルデータース	
NIST 11	化合物の EI-MS および MS/MS スペクトルを収載。GC での測定パラメータや分子構造も充実。20 年以上の実績あり。
GMD	代謝物質や生理活性物質の GC-MS パラメータを収載。
MassBank	GC-MS, ESI-MS, FAB-MS などにおけるスペクトルを収載。データ比較ツールなども充実。
METLIN	4 万種以上の化合物の LC-MS、-MS/MS, FT-MS によるスペクトルや分析パラメータ、化学構造を収載。
MMCD	NMR や LC-MS パラメータや化学構造を検索可能。関連物質の検索も充実。
代謝プロファイルデータベース	
HMDB	ヒト身体の代謝物質を収載。NMR、GC-MS、MS/MS データを収載。代謝経路情報や濃度データが文献とともに収載。
DrugBank	FDA 承認薬やサプリメントなど 6,700 種の薬理情報や分子構造、物理化学的パラメータを収載。
LMSD	脂質に特化。物理化学的パラメータだけでなく、酵素情報も収載。
PubChem	化合物の物理化学的パラメータの他に、薬理作用や受容体への親和性などのデータも収載。
ChEMBLdb	化合物の物理化学的パラメータや生理的機能も収載。文献も豊富。
代謝パスウェイデータベース	
KEGG	最もよく使われる代謝物質、代謝経路の百科事典。
BioCyc	代謝物質、タンパク質などの総合的 DB。代謝に特化した MetaCyc、ヒトに特化した HumanCyc、大腸菌の EcoCyc などが含まれる。
Reactome	ヒトに関連した代謝反応の DB。
MeTaBoard	主要代謝分子の基礎情報、代謝、検出値などを収載。

EI: electron ionization, ESI: electrospray ionization, FAB: fast-atom bombardment, FT: Fourier transforming, GC: gas chromatography, LC: liquid chromatography, MS: mass spectrometry, MS/MS: tandem MS, NMR: nuclear magnetic resonance, RT: retention time (LC, GC)