創薬・疾患研究のためのデータ統合の実際

水口 賢司
医薬基盤・健康・栄養研究所バイオインフォマティクスプロジェクト プロジェクトリーダー

創薬研究の各段階に関連する実験データは、すでに公共データベースに多数格納されている。しかし、それらをビッグデータとして解析するためには多くの課題を克服する必要がある。例えば、実験条件についての情報が十分に構造化されておらず、必要なデータの取捨選択が難しい、用語や単位が統一されていないなどは、分子レベルのデータから非臨床、臨床データに至るまで共通して見られる問題と言える。

本講演では、我々が遂行している統合データベースや予測モデルシステム開発において、データ統合に関わる課題にいかに取り組んでいるかの具体例を紹介する。薬物動態の基礎データベースとそれに基づく物性・薬物動態関連パラメータの予測モデル構築プロジェクトでは、薬理活性データベースChEMBLや他の幾つかのデータソースから抽出したデータについて、実験条件の精査や単位の正確な変換、その他のマニュアルキュレーションを効率化するためのワークフローを構築した(**図1**；Esaki et al., Mol. Inform. 印刷中)。キュレーションによって訓練データの質を上げることが、予測モデルの精度向上につながることを確認している。創薬初期の探索研究を支援するTargetMineデータウェアハウスでは、複数のデータベースから遺伝子と疾患・表現型、遺伝子と発現組織などの関係性に関わるデータを取得しているが(**図2**)、これらを統合して有効な解析ツールにするためには、用語や概念の統一が大きな課題になっている。これらは、個別の研究領域における例ではあるが、データ整備、オントロジーなど、人工知能技術の応用における共通の基本課題に示唆を与えるものと考えている。

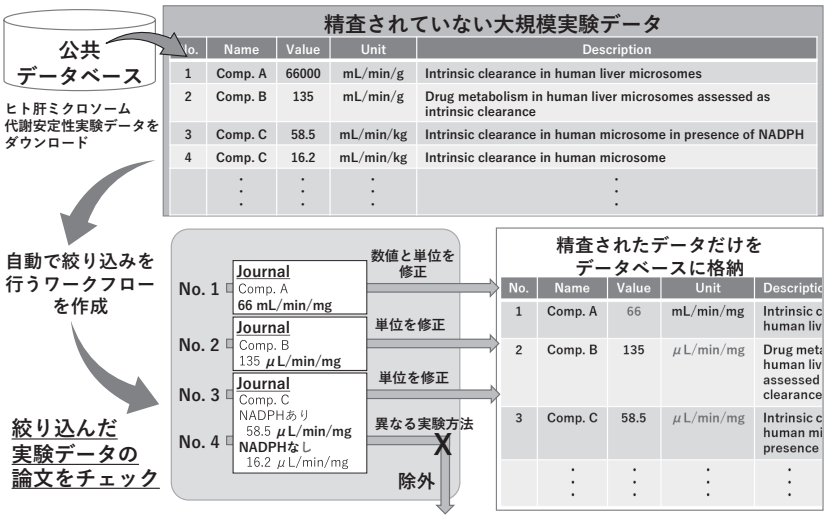

図1　代謝安定性実験データを例にしたキュレーションの実際

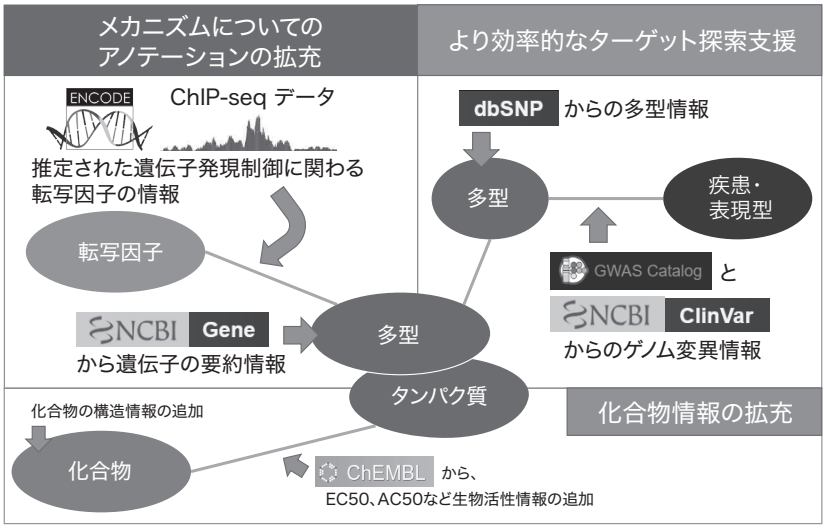

図2　TargetMineデータウェアハウスにおけるデータ統合