【課題番号】課題 2

【課題名】RDF データに対する、生物学的な知識発見につながるパラメータ探索

１）課題とその背景

バイオサイエンスデータベースセンター(NBDC)の統合化推進プログラムおよび世界の主要なデータベースセンターから、すでに様々な生命科学のデータがセマンティック・ウェブの標準データモデルである Resource Description Framework (RDF)のデータとして構築・提供されている。RDF の有利な点として、これらの様々なデータを、手を加えることなく、そのまま同じトリプルストアに集積して利用できることがあげられる。しかし、データを一元的に蓄積しただけでは、そこから有用な知識を得ることが簡単になるわけではない。このため、多くの場合すでに知識として知っている関係を検索するのにとどまっているのが現状である。RDF では複数のデータセットが共通の URI を通じて連結されるので、これまで明示的に知られていなかった関係がデータを組み合わせることにより見いだされるはずであり、そのような未知の情報を探索する手法の開発が望まれている。これにより、生命科学的に興味深い統計情報や新しい発見がもたらされる可能性がある。

２）課題の解決方法の概要

膨大なデータに対して、未知の情報に関する何らかの糸口を得るためには、データの特徴をおおまかに把握するためのナビゲーションシステムが役に立つと考えられる。そして、そのようなシステムは、データについて事前に集計された統計情報を必要とする。RDF データは明確にクラスや関係が定義されているため、統計解析をするのに適している。しかし、生命科学のデータには、多種多様な属性が存在する上、その属性間の関係も自明ではないことが多く、生物学的に意味のある属性の組み合わせを見出すのは、簡単なことではない。そこで、TogoGenome 等に蓄積された RDF データを様々な属性でグループに分け（例えば、真核生物と原核生物、独立栄養生物と従属栄養生物等)、さらにグループ内で、グループ分けに使わなかった各種の属性に対して統計情報を集計することで、生物学的に意味のあるグループ分けと属性の組みを提示する手法を開発する。ここで属性の候補としては、オントロジーのクラス等(例えば、遺伝子・環境・表現型など)が利用できる。さらに発展的な課題としては、発見したパラメータの関係を可視化するウェブアプリケーションの開発等が挙げられる。

３）課題の解決方法の概略図