INOHプロジェクト：高次機能情報のデータベース化

論文中に埋没している知識をいかにして計算機処理可能にするか？そういった類の仕事にずっと携わっています。生データを解析して知識を得るというよりは、生物学における知識を計算機の上に再構築して知識基盤として提供するための研究です。修士の頃は文献データベース Medline からタンパク質の相互作用情報を自動抽出するための研究に取り組んでいました。当時は、バイオインフォマティクスでなぜ自然言語処理なの？という顔をする人が多かったのを覚えています。しかしその後、多くの方が文献処理に興味を持ってくれるようになりました。そして私自身の研究の興味は、文献から抽出した知識をどのように計算機上に表現して処理するか、というテーマにシフトしていきました。

現在は INOH[(1)]プロジェクトという、論文を介して共有されている高次知識をデータベース化するプロジェクトに取り組んでいます。ここでいう高次知識とは、個々の物質の機能に関する知識と対比して、物質間、物質・現象間の関係で記述される機能の知識を指します。例えば、遺伝子やタンパク質が細胞機能を制御しているネットワーク的な知識構造（パスウェイ）のことです。

INOHプロジェクトには大きく分けて４つの使命があると考えています。１つ目は生物学における高次知識処理技術の開発、２つ目は生物学者によるデータのキュレーション、３つ目は文献に登場する知識を扱うためのオントロジー群の整備、そして４つ目が高次知識データベースシステムの開発です。現在、主にBIRD、JSTの支援の元にこれらの課題に統合的に取り組んでいます。

言葉や図によって表現された知識の特徴はおおざっぱに言うと、不均一な構成要素が入り乱れて登場し、暗黙的な部分構造の知識に言及しがちなことです。例えば、タンパク質と小胞体とカルシウムイオンが対等な役者として登場し、さらにカルシウムイオンはカルシウム依存のパスウェイと直接相互作用するような記述が登場しますが、これらがそれぞれ別々の概念を表していることを理解する必要があります。あるいは、細胞膜のそばに、α、β、γとあれば、あぁ、これはGタンパク質なんだな、ということを理解できないといけません。こういったことを計算機にさせるには、知識を計算機の扱えるうまい形式に変換する必要があります（図１）。また、変換された記述に登場する各オブジェクトの意味を指定するためにはオントロジーが必要となります。最近になってようやく私たちの取り組みの一部を公開できるようになりました（図２）。まだまだ荒削りで現在も精力的に改良を加えているところですが、文献知識に基づく個性的なパスウェイデータベースになっています。

実際に生物学者が読まないとくみ取れないような知識を文献中から抽出し、データベース化するという、ある種、壮大なプロジェクトですので、生物出身の方、情報出身の方を問わず多くの方の協力から成り立っています。そして、逆にプロジェクトの成果も、知識にアクセスする必要のある生物学者、新しい知識処理技術の開発に興味のある情報科学者の双方にとって有用な知識基盤をつくっていけたらよいな、と思っています。

(1) Integrating Network Objects with Hierarchies (INOH).
断片的で曖昧に共有されてきた高次知識を正確に書き下すことで全体を俯瞰した美しい地図に仕上げます。

参考URL　http://www.ontology.jp/FREX/index.html

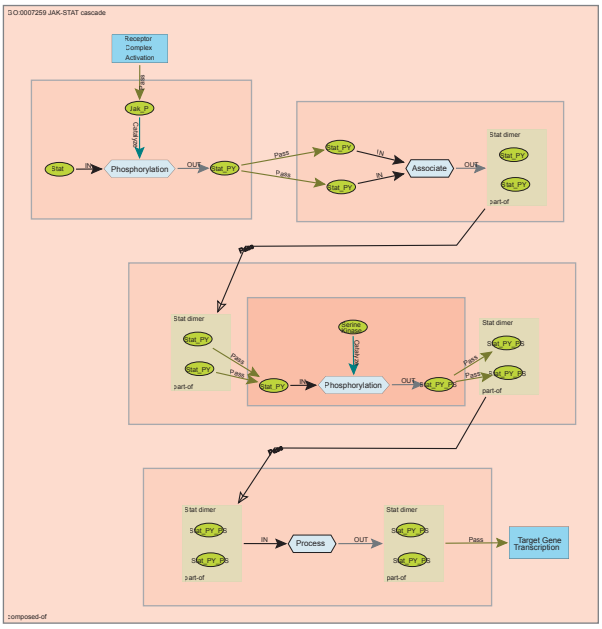

図１：高次知識の階層的な表現

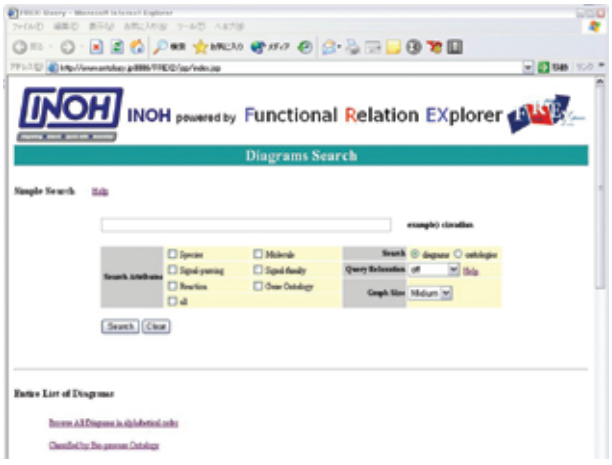

図２：INOHシステムトップページ（予定）

数理モデルチーム
（研究員）
フクダ ケンイチロウ
福田 賢一郎