

ライフサイエンスデータベース統合推進事業
統合化推進プログラム（統合データ解析トライアル）
平成 25 年度採択課題における事後評価について

バイオサイエンスデータベースセンター

ライフサイエンスデータベース統合推進事業 統合化推進プログラム(統合データ解析トライアル)における平成25年度研究開発課題の事後評価を「ライフサイエンスデータベース統合推進事業の実施に関する規則」に基づき実施した。

1. 事後評価の目的

「ライフサイエンスデータベース統合推進事業の実施に関する規則」に基づき実施。

(1) 評価の目的

研究開発の実施状況及び研究開発成果等を明らかにし、今後の研究開発成果の展開及び事業運営の改善に資することを目的とする。

(2) 評価項目及び基準

- ・当初計画の達成度
- ・開発したツールを活用してデータベースの新たな活用方や有用な知見の発見の状況
- ・開発したツールの将来性への展望等

2. 対象となる研究開発課題

(1)「大規模なタンパク質データ解析のための高速な局所配列特徴抽出法の開発」

研究代表者:蝦名 鉄平 理化学研究所 脳科学総合研究センター 研究員

(2)「マルチオミクスデータを用いたゲノム規模代謝モデリングのためのネットワーク解析システムの開発」

研究代表者:西田 孝三 理化学研究所 生命システム研究センター テクニカルスタッフ

(3)「KNpSACK を用いた植物の効能メカニズム解明のための基盤構築」

研究代表者:西村 陽介 京都大学 化学研究所 大学院生)

(4)「機械学習を用いたタンパク質—リガンド結合部位予測ツールの自動生成パイプラインの開発」

研究代表者:番野 雅城 東京大学 大学院農学生命科学研究科 大学院生

(5)「植物代謝物プロファイリングデータベース AtMetExpress の開発とオミクスデータ統合化の推進」

研究代表者:福島 敦史 理化学研究所 環境資源科学研究センター 研究員

(6)「共起関係解析によるタンパク質の機能モジュール探索法の開発」

研究代表者:藤井 聡 九州工業大学 大学院情報工学研究院 助教

(7)「タンパク質—糖鎖間の糖鎖結合部位の解明のためのツール改良及び解析」

研究代表者:細田 正恵 創価大学 大学院工学研究科 大学院生

(8)「MicrobeDB.jp データを用いたメタゲノム解析 Web アプリケーションの開発」

研究代表者:森 宙史 東京工業大学 大学院生命理工学研究科 助教

3. 研究開発実施期間

平成 25 年 10 月 1 日から平成 26 年 1 月 31 日まで(4ヶ月)

4. スケジュール

1) 研究代表者が研究開発終了報告書を提出(平成 26 年 2 月)

2) 研究開発終了報告書等を総括、評価委員に送付(平成 26 年 2 月)

3) 研究開発成果報告会の開催(平成 26 年 3 月)

・統合化推進プログラム 統合データ解析トライアル研究総括、評価委員出席の下で一般公開の研究開発成果報告会を開催し、被評価者からの報告、意見交換をもとに評価委員が事後評価記入票を作成

4) 総括による事後評価結果報告書の作成(平成 26 年 4 月～平成 26 年 6 月)

・各評価委員が作成した事後評価記入票をもとに、事後評価結果報告書案を作成

・各研究代表研究者に事後評価結果報告書案を提示し、意見交換

5) 業務運営会議等への報告・HP にて一般に公開(平成 26 年 8 月)

6. 評価者

・研究総括

高木 利久(東京大学 大学院新領域創成科学研究科 教授)

・評価委員

伊藤 武彦(東京工業大学 大学院生命理工学研究科 教授)

大浪 修一(理化学研究所 生命システム研究センター チームリーダー)

川本 祥子(情報・システム研究機構 ライフサイエンス統合データベースセンター 特任准教授)

五斗 進 (京都大学 化学研究所 バイオインフォマティクスセンター 准教授)

藤 博幸 (産業技術総合研究所 生命情報工学研究センター 副研究センター長)

戸田 達史(神戸大学 大学院医学研究科 教授)

松田 秀雄(大阪大学 大学院情報科学研究科 教授)

光山 統泰(産業技術総合研究所 生命情報工学研究センター 研究チーム長)

*五十音順、所属は、平成 26 年 3 月時点のもの

7. 事後評価結果

NBDCでは、データの共有と統合化を推進することにより、我が国のライフサイエンス研究から生み出されるデータの価値を最大化することを目指して統合化推進プログラムを実施している。しかしながら、それらのデータベースを十分に使いこなすためのさまざまな解析ツールや人材の不足などから、それらのデータベースの能力を十分には生かしきるまでにはいたっていないのが実情である。

そこで、そのような問題の解消を図るために、統合データベース解析トライアルを実施した。ここでは、データ解析をするツール等を開発し、それを用いて新たな知識発見を目指す研究開発提案の募集を行い、13件の応募があり、8件を採択した。若手研究者や学生が熱心に取り組んだ本プログラムにおいて、4ヶ月という非常に短い開発期間(研究費:80万円程度)ではあったが、当初予定の研究開発期間を終了したことに伴い、事後評価を行った。

各研究課題の事後評価結果については、以下の通り。

(1)【大規模なタンパク質データ解析のための高速な局所配列特徴抽出法の開発:蛭名 鉄平 研究代表者】

巨大なタンパク質データベースにも適用可能な高速なクラスタリング法を利用し、特定の構造に対応するタンパク質の局所配列の特徴を迅速に抽出するための手法として、RVSB (Representative Vector Selection using BOOL)を開発した。このツールにより、PDBjのデータを利用して、タンパク質のループ構造の予測を、従来法と同程度の予測効率で、大規模なデータに適用することが可能となった。増え続けるPDBデータに対して将来的にも応用可能な方法論の開発である。既存のデータベースの新たな活用法、有用な知識の発見という点では、まだこれのみの成果で判断するには材料が乏しいと考えられるが、今後の研究開発により、ループ構造以外の他の構造の予測や、さらに詳細に解析することによって、機能との関連まで結びつけることができる可能性も示されている。本ツールはタンパク質に限らず利用できる汎用性の高いツールであり、様々な応用の可能性が期待される。

(2)【マルチオミクスデータを用いたゲノム規模代謝モデリングのためのネットワーク解析システムの開発:西田 孝三 研究代表者】

近年いくつかのモデル生物で全ゲノム規模での代謝モデルの再構築結果が公開されているが、これらモデルに「オミクス実験情報」、「KEGGに代表されるパスウェイ、そしてKEGG MEDICUSのようなゆらぎ物質としての薬剤情報を集めたデータベース」を統合したネットワーク解析システムはまだ発展の途上にある。本研究では、CytoscapeプラグインであるKEGGscapeを開発し、KEGGデータベースのみでは扱えないDrugBankのデータやマイクロアレイの発現データ等をCytoscape上で統合可視化し、これまで作られて来たデータベースのみでは見る事のできなかつた情報を俯瞰する事ができるなどKEGGパスウェイを拡張したマルチオミクスモデリングを可能にした。ツールの成熟度について今後の発展が望まれる部分があるものの、大腸菌以外の微生物や動物、植物にも適用できる可能性を秘めており、将来的にはKEGG MEDICUSの応用や、植物メタボロームへの応用などが期待される。

(3)【KNAPSAcKを用いた植物の効能メカニズム解明のための基盤構築:西村 陽介 研究代表者】

分子レベルの機能情報である代謝産物の相互作用タンパク質情報や薬効・適応症・毒性情報をKNAPSAcKに統合することにより、個体レベル、分子レベルの両者の機能情報を関連付け、それらの因果関係を元に効能未知の植物について、メタボローム情報等を利用した効能予測を行うための基盤となるPCIDB (PhytoChemical Interactions DB)を構築した。代謝産物側の様々な指標に基づく分類と、分子機能や疾病情報の関連性について新しい知見を得ようとするものであり、統合データベースの新しい活用法

のひとつといえる。開発したツールを活用して生命科学的に有用な知識を発見するには至っておらず、ヒトに対する植物の効能の予測までは達成できていないが、医薬品に比べて植物の作用はそのメカニズムが解明されていないものが多く、本ツールはその解明の第一歩になるものと考えられる。

(4)【機械学習を用いたタンパク質—リガンド結合部位予測ツールの自動生成パイプラインの開発:

番野 雅城 研究代表者】

PDB 中のリガンド結合情報をまとめた二次データベースを構築し、そこから相互作用情報を抽出して機械学習を行うことで、タンパク質—リガンド結合部位の予測を実現するパイプラインを開発した。予測精度の向上や、もう少しウェット研究者の視点に立ったインターフェース等に工夫が必要であると思われるが、出力として得られるツールは RDF 形式に対応したフローとなっており、他データベース、ツール群への展開等、将来性も期待できる。現時点ではツールの検証段階であり、有用な知識の発見には至っていないが、リガンド結合タンパク質予測ツールやタンパク質間相互作用部位予測ツールに発展することが期待される。

(5)【植物代謝物プロファイリングデータベース AtMetExpress の開発とオミックスデータ統合化の推進:

福島 敦史 研究代表者】

シロイヌナズナの代謝物プロファイリングデータに関して、MassBank、KNAPSAcK、KEGG と海外のデータベースを統合化し、シロイヌナズナで検出された報告のある代謝物と現在の質量分析の技術で実測できる代謝物とを関連づける AtMetExpress データベースを開発した。生命科学的な新たな知識の発見には至っていないが、代謝物プロファイルデータを対話的に可視化し表示できるなど、従来にない解析機能を持っており、このツールの利用を通じて、代謝物プロファイルデータの統合が促進されるなどの展開が望まれる。

(6)【共起関係解析によるタンパク質の機能モジュール探索法の開発:藤井 聡 研究代表者】

PDB に登録されているタンパク質立体構造中の機能モジュールの共起関係を定義することによって、立体構造と機能モジュールを有機的に結びつけており、PDB の活用法を見出したと言える。活性部位付近の共起関係について、PDBnet-Co-occurrence Search Tool -として作成しており、これをより詳細に解析することにより有用な知識の発見に結びつく可能性は高い。現状では、機能モジュールの単位として PROSITE モチーフのみを対象としており汎用性は高くないが、共起関係の解析という一般的な枠組みはできているので、他の機能モジュールや部位に応用することは可能だと考えられる。対象とするデータベースを PROSITE 以外も可能なようにするなど、データベースの適用範囲を広げたり、発見結果を検証する方法を用意したりすることで、もっと有用な手法となっていくことが期待される。

(7)【タンパク質—糖鎖間の糖鎖結合部位の解明のためのツール改良及び解析:細田 正恵 研究代表者】

糖鎖を対象として、マルチプルアライメントを行うツール MCAW(Multiple Carbohydrate Alignment with Weights)をもとに、改良を行った。レクチンフロンティアデータベースの糖鎖構造データに適用し、タンパク質との結合に関与すると考えられる単糖や、単糖と結合の特異性をプロファイルとして表示することが可能

となった。アライメントアルゴリズムの改良、および、アライメントの結果を解釈して、アライメントの改良や意味付けができるような改良がされることにより、糖鎖リガンドとタンパク質との相互作用解析には必要となるツールとなっていくことが期待される。

(8)【MicrobeDB.jp データを用いたメタゲノム解析 Web アプリケーションの開発:森 宙史 研究代表者】

fastq形式のメタゲノム配列データをアップロードして系統組成を返すMeGAP(MetaGenome Annotation Pipeline)の開発を行った。ユーザー側の大量配列データに対応した解析サービスは、これからの統合データベースの活用法としては自然な応用で、今後このようなサービスが充実することが期待される。統合データベースを活用する「新たな解析手法」を開発したわけでは無いが、これまでバイオインフォマティクスが行ってきた統合データベースを活用した解析を、実験生物学者自らが行うという「新たな解析の様式」を可能にするツールであり、メタゲノムデータベースの充実と共に、今回開発されたツールの有用性が増すことが期待でき、メタゲノム解析の研究に貢献すると考えられる。

以上