

情 報

日本作物学会紀事を利用した作物学データベースの構築

佐々木治人¹⁾・石丸健²⁾・伊藤聖¹⁾・小曾根睦²⁾・柏木孝幸²⁾・上原直子¹⁾・山岸徹¹⁾

(¹⁾ 東京大学大学院農学生命科学研究科, ²⁾ 農業生物資源研究所生理機能研究グループ)

近年, 独立行政法人農業生物資源研究所が中心となってイネゲノム研究プロジェクトを進めており, 現在世界をリードしている (<http://rgp.dna.affrc.go.jp/>). このイネゲノム研究プロジェクトでは塩基配列データ, プロテオームやマイクロアレイ等の網羅的解析が進められ, 膨大なデータが得られているが, それを効率的に利用するにはバイオインフォマティクス (生命情報科学) 技術を用いてゲノム情報と生命科学並びに情報科学を融合し, 遺伝情報を解析する必要がある. これをふまえて, コンピューター上でイネ等の農作物の品種改良実験を可能とするイネ・ゲノムシミュレーター (仮想実験システム) の開発を目指す「イネ・ゲノムシミュレーターの開発」プロジェクト (独立行政法人農業生物資源研究所) が平成 13 年に開始された. 当プロジェクトは, (1) イネゲノム機能解析研究を促進するためのデータベースの構築とソフトウェアの開発, (2) イネゲノム研究の成果を品種改良に役立てるためのシミュレーター等の新しい情報科学手法の開発からなる.

今までにイネに関する栽培方法や生長特性, 収量性, 品種間差等についてきめ細かな調査が行われ, 日本作物学会もその一翼を担ってきた. そして, その膨大な解析結果は学会誌 (日本作物学会紀事) として蓄積されている. その対象領域は生理学, 形態学, 生態学, 育種学, 遺伝学, 栽培管理, 品種, 品質, モデリング, リモートセンシング等を包含しており, また研究対象も遺伝子, 細胞レベルから個体群やグローバルな食糧問題まで幅広く扱っている.

ゲノム研究と作物学研究として蓄積されてきた情報及び育種現場での特性データ等を関連付けて利用していくためには, 情報を広く提供することが必要となる. 私たちは, 「イネ・ゲノムシミュレーターの開発」の課題の一つとして「作物学データベース」の構築を行った.

作物学データベースに要求されること

作物学関連のデータベースとしては Biological abstracts や AGRICORA 等の論文検索や国際イネ研究所 (IRRI) による論文タイトル集がある. しかしながら, それらはタイトルや英文アブストラクト中にある文言や単語のデータのみが対象であり, 体系化されたキーワード検索もできず, 論文中のデータを精査して抽出した解析データベースといえるものではなかった. また作物学は

- ・専門用語が特殊でわかりづらい
- ・用語間の関係がわかりづらい
- ・多様な要因の複合された結果 (収量等) を研究対象としているため, 整理しにくい

といった問題を抱えており, 作物学研究で培ってきた時系列的な生長解析やシンクソース関係等の収量構成解析, 品種間比較の貴重なデータを異なる学問分野の研究者が利用することは困難であった. これらのことより今までに行われた作物学研究における多量の解析データを整理並びに分類し, ゲノム研究における利用を前提にした作物学データベースの構築を行う場合,

- 1) 検索データとして多くの論文を保有すること
- 2) 和文論文についての検索が可能であること
- 3) 一般の人や他分野の研究者でも容易に検索できること
- 4) 目的とする論文に的確に到達できること
- 5) 検索された論文を検索者へ提供できること
- 6) 今後の新しい論文や新しい知見に対応できることが要求される.

作物学データベースの構築

日本作物学会紀事 50 年分 (22 巻から 71 巻まで) の全文書と, 文献として登録されていないため利用することが困難である農業技術研究所報告書 30 年分の作物学関連論文等についてカバーする作物学データベースを構築した. なお作物学データベースを以下のアドレスで公開している (<http://mitochon.gs.dna.affrc.go.jp:81/csdb/>).

作物学データベースが所有する機能については次の 6 つがあげられる.

1) 自然文による論文検索

キーワードが分からない場合や曖昧な記憶からでも, 思いついたままの言葉や文章を入力するだけで, システムが検索文の内容を自動的に解釈し, 適切な検索結果を提供する.

2) 専門用語を利用せずに論文を検索する機能

作物学の専門用語を体系化した辞書を作成し, 検索文と同義, 類義の内容を含んだ論文を検索する. 作物学用語を知らなくても検索可能である.

3) 類似性に応じた検索結果の提供

論文中の用語の出現頻度や用語の重みを利用して、論文同士の類似の度合いを数値化、検索文と類似度の高い順に検索結果を表示する。

4) 検索文に関連する専門用語を表示

検索文と関連性の高い用語を表示する。関連用語を掛け合わせた絞込み検索を可能にする。

5) 全文書の PDF ファイル化

検索された文献に関しては直接ダウンロードを可能にする (注 1)。

6) 論文追加機能

新規に追加された論文や新しい知見に基づく情報をデータベースに加えることが可能である。

これらの機能を所有し、絞込みが可能な用語による検索 (限定用語検索) と自然文による検索 (自然文検索) を可能にした。

支援システムの構築

支援システムとして、キーワードとなる用語を体系化した辞書を作成した。シソーラスもそのような辞書であるが、もともとインデクシングの効率的運用のために考案されたものである。作物学データベースでは用語の概念をもとに体系化された辞書のみでなく、それぞれの学問分野の知見をもとに体系化された辞書を作成した。

本辞書は作物学データベースの検索システム用に作成された。しかし、本辞書の用語体系は学問的知見を背景としていることから、逆に、学問的知見を体系的に浮き彫りに

することも可能であり、新規の知見や用語間関係の創出の際に効果的に機能する。

作物学データベースが今後の研究に与えるもの

以上のように作物学データベースはゲノム研究や他分野の研究者へ収量や生長解析等の重要な作物学情報を提供することを主眼において作製された。このデータベースにより、生長特性や収量性といった農学的に重要な形質の情報を活用し、ゲノム研究と作物学研究の連携による研究のスピードアップ、遺伝子改変のターゲットの設定、超多収性のスーパーライスの育成等が可能となると考えられる。また網羅的な遺伝子解析により、今まで関係が論じられてこなかった形質間に新規に関係が生じ、その新規の関係について作物学データベースを用いて再検討することにより、新しい研究テーマが創生すると考えられる。ハード面ではこの作物学データベースがテンプレートとなり、他の分野においても過去のデータ、特に和文論文のデータベース化が可能になり、分野間の情報交換を盛んにすることにより、科学が発展することを希望する。

参考文献

- 佐々木治人・石丸健・山岸徹 2001. イネを中心とした作物学データベースの構築. 日作紀 70 (別 2): 267-268.
山岸徹・石丸健・佐々木治人 2002. 作物学データベースのための検索用辞書の作成. 日作紀 71 (別 1): 280-281.

(注 1) 日本作物学会紀事および Plant Production Science の PDF は日本作物学会の提供による。農業技術研究所報告書の PDF は独立行政法人農業環境技術研究所の提供による。