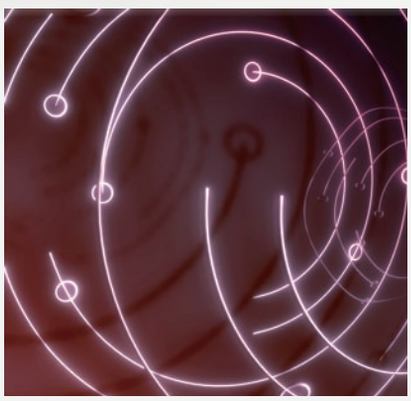
特集

ポストゲノム時代に高まるバイオ自然言語処理への期待：バイオ自然言語処理最新事情

0 編集にあたって



前田 英作
maeda@cslab.kecl.ntt.co.jp / NTTコミュニケーション科学基礎研究所

高木 利久
tt@k.u-tokyo.ac.jp / 東京大学新領域創成科学研究科情報生命科学

背景とねらい

2000年6月26日，米国ホワイトハウスで行われたヒトゲノム解読宣言は，ゲノム（DNAの塩基配列）を読み取る技術を人類が手中にしたことを示す象徴的な出来事であった．そして現在，ゲノム解読を契機とする新しい科学の時代を指して「ポストゲノム時代」という呼び方がされている．このポストゲノム時代の大きな目標の1つは，ゲノム配列や遺伝子発現情報などに対する生物学的医学的解釈を体系的かつ網羅的に行い，生命のより深い理解に迫ることにある．タンパク質のアミノ酸配列，構造，機能などを網羅的に解析するプロテオーム（proteome），細胞内における生理活性物質の代謝系を解析するメタボローム（metabolome）などの研究はそうした試みの例である．ポストゲノム時代には，こうした体系的，網羅的解析から得られる大量の実験データを適切かつ効率的に扱うための情報処理技術（バイオインフォマティクス）がこれまで以上に重要となることは言うまでもない．このあたりの事情は，情報処理2002年1月の特集「ゲノム情報科学」に詳しい．

本特集で扱う「バイオ自然言語処理」（以下，バイオNLP）は，ポストゲノム時代において特に期待の高い情報処理技術の1つである．バイオNLPという言葉を初めて耳にされる方の中には，なぜ「バイオ」と「言語処理」の組合せなのかと訝る向きもあるかもしれない．今，バイオNLPが注目されているのは背景に次のような事情があるからである．第1に，ゲノム解読と並行して進んださまざまな計測技術の進歩が大量の実験データを研究者にもたらすと同時に，それを解釈し報告する文献の量も膨大なものになった．いわば，言語記述の洪水が起きている．第2に，ゲノムという生命を語る共通言語の解析が進んだため，バイオ研究者は自身の一専門分野（たと

えば，ある特定の生物種や臓器など）を超えた広範な領域に対する文献渉猟が必要となった．第3に，実験によって得られた知見は，データベース（DB）というかたちで蓄積，保存されていく．遺伝子DB，タンパク質DBなどがその代表例である．そうしたデータベースの自由記述欄には自然言語を用いてさまざまな注釈が付加される．また，「機能」はこれらのデータベースにとって最も重要な項目であるが，それを表現するほぼ唯一の手段は自然言語である．

こうした状況下において，医学，生物学に携わる研究者自身が，バイオNLP技術そのもの，あるいはそれらを駆使した処理システムを強く必要としている．そこでは，多様な文書中のテキストから遺伝子やタンパク質の機能に関する知識情報を単に検索するだけではなく，関連情報も含めて抽出し，それらを体系化することまで求められている．そのためには，遺伝子名，タンパク質名などの辞書の整備，用語の概念体系を表すシソーラス，オントロジの構築，言語記述を処理するための高速，高性能な言語解析器の開発，さらには，情報検索，知識処理，データベース，機械学習などの情報処理技術も必要となる．一方，医学生物学という学問分野の大きな特徴の1つは，電子化されたテキストが大量に蓄積されているという点にある．先に述べた一連の言語資源や技術を揃えるためには，何よりもまず電子化された大量テキストが必要である．したがって，バイオNLPは自然言語処理の研究者にとっても魅力のある研究領域となっている．

ビジネス分野においても，医学生物学文書からの知識抽出とその体系化を目指した商用システムの開発が進みつつある．しかしながら，実用レベルのものはまだ少なく，この分野への情報処理の専門家の参入が切実に求められている．欧米ではバイオ実験・文献DBの整備と，それを処理するためのバイオNLP技術の研究開発に対して，すでに国家戦略的な取り組みが始まっている．このような現状を踏まえ，情報処理に携わる国内研究者技術者向けにバイオNLPの現状と課題を紹介することによって，より多くの研究者技術者のバイオNLPへの参入を促すとともに，研究開発のさらなる活性化を狙いとして本特集を企画した．

本特集の構成

バイオNLPを取り巻く状況をいろいろな視点から幅広く理解してもらうために，4つの異なる立場の方々にそれぞれの視点から執筆をお願いした．

まず第1に，

1「語る科学へ向けて―データ・知識・生命現象をつなぐ―」

では，生物学と言語処理とのかかわりについてより広い視野からその重要性を解き，長い将来にわたってバイオNLPが目指す方向性とその役割について述べていただいた．

第2は，実際に生き物を使った実験研究（一般に「ウェット」と呼ばれる．それに対し，計算機を使って行う研究を「ドライ」と呼ぶ）に携わっている研究者による2本の解説，

2-a)「遺伝子変異データベース構築のための情報収集と抽出の現状」

2-b)「大規模プロテオミクスからバイオNLPに望むこと」

である．実験研究者が日常使っているバイオNLP技術，データベースの現状と問題点について述べるとともに，生物屋と情報処理屋が協力関係を築こうとするときにしばしばぶつかる問題についても率直に触れられている．

第3は，自然言語処理，知識処理，情報検索の研究者による4本の解説，

3「生命科学文献からの知識抽出と辞書構築―その現状と課題―」

4「バイオNLPのためのコーパスと各種リソースの現状」

5「ゲノムデータの機械解釈」

6「バイオ自然言語処理のための機械学習技術」

である．現在私たちが利用できる言語資源の現状と問題点について俯瞰するとともに，知識処理，知識DB，自然言語処理，バイオDB，機械学習など情報処理のそれぞれの立場から見たバイオNLPの現状と今後について触れる．

第4に，最後の3本の解説，

7-a)「ポストゲノム時代のテキストマイニングミドルウェア」

7-b)「連想統合による医学・生物学知識の活用ソリューション」

7-c)「バイオ医療情報からの疾患関連因子抽出システムについて」

では，バイオNLPを基盤として商用化を視野に入れた企業の取り組みについて紹介していただいた．

本特集が1つのきっかけとなって，研究者，技術者にバイオNLPの重要性を認識していただき，より多くの方々に本分野へ参入していただきたい．また，本特集に関連して，国立情報学研究所と東京大学の共催による国際シンポジウム，“e-Biology Initiative: Towards New Frontiers of Biology”が3月10～11日に東京で開催される予定である．今後，医学生物学，言語処理，知識処理，データベースそれぞれの分野の研究者，技術者がより広く，そしてより深く議論を重ね，バイオ自然言語処理技術を活きた道具として使いこなせるようにしていかなければならない．そしてさらには，より深い生命現象の理解につながる新しい生物学の構築，すなわち，中村桂子氏の言う「語りの科学」の創成に向けた一歩となることを期待したい．

（平成17年1月11日）