
# 6 ゲノム情報学と言語処理

辻井 潤一
CREST（科学技術振興事業団）／
東京大学大学院情報理工学系研究科コンピュータ科学専攻
*tsujii@is.s.u-tokyo.ac.jp*

## 言葉の処理とゲノム情報学—交流の活発化

情報技術と生命科学という，急速に進展している2つの分野の境界にゲノム情報学がある．複雑で多様な生命現象を理解する上で，情報技術は欠かせない，その中でも，言葉の処理が，ポスト・ゲノムの生命科学の鍵となる技術として認識されている．

このような事情を，5～6年前から，本特集号編者の1人，高木先生に聞かされていた．そのときは，半信半疑で文部省（当時）のゲノム科学特定研究や，東京・恵比寿で毎年開かれているゲノム情報学の国際会議GIW☆1に参加していた．半身の姿勢だった私が，この分野に強い関心を持ったのは，2000年の年初にハワイで開催された国際会議PSB☆2からである．このゲノム情報学の会議には，言語処理のセッションが設けられていた．

2001年には，同会議の言語処理セッションのオーガナイズとチュートリアルを行い，コペンハーゲンでの同種の国際会議ISMB☆3でも，米国・ヨーロッパの研究者たちと一緒に，「テキスト・マイニングと言語処理」のチュートリアルを行った．いずれのチュートリアルにも，大勢の参加者があり，ゲノム情報学からの言語処理への期待を実感した．ちなみに，2000年には論文集めに苦労していたPSBの言語処理セッションは，2002年には，最も多くの投稿論文が集まったセッションとなっている．

また，この動きと連動して，言語処理研究者の間にも，急速に興味が高まっている．言語処理分野の最大の国際会議であるACL☆4（2001年）では，逆に，ゲノム科学の専門家による言語処理研究者向けのチュートリアルが設けられた．また，NA－ACL（ACLの北米支部会議，2001年）では，「ゲノム科学と言語処理」の特別講演，DARPA☆5主催のHLT☆6（2002年）では，同様なテーマの特別セッションが設けられている．

ゲノム科学と言語処理の両分野は急速に接近し，言語処理はゲノム情報学の中核の1つになりつつある．本稿では，この急速な接近がなぜもたらされたか，また，どのような研究プログラムが構想されているかについて考える．

## 言語と生命

言語と生命は，いずれも，環境との複雑なインタラクションを通じて進化してきた系である．また，生命科学の進展は，生命現象の中に，遺伝，あるいは，細胞間の信号伝達といった，「言語的なもの」による情報伝達の存在を明らかにしてきた．特に，DNA系列によ

☆1 GIW: International Conference on Genome Informatics
☆2 PSB: Pacific Symposium on Biocomputing
☆3 ISMB: Intelligent Systems for Molecular Biology
☆4 ACL: Association of Computational Linguistics
☆5 DARPA: Defense Advanced Research Projects Agency
☆6 HLT: Human Language Technology Conference

る遺伝情報は，4種類の塩基という文字からなる，神が書いたテキストにたとえられ，このテキストの解読が生命の神秘を解明する，とされる．遺伝情報の解読は，ロゼッタ・ストーンのような未知言語の解読ということになる（図-1）．

このような漠然とした共通性の認識に，どれだけの科学的根拠があるかは疑問であるが，ゲノム情報学と言語処理に共通する技術は多い．たとえば，両分野は，機械学習研究の格好の応用分野である．MIT Pressから出版された2つの教科書，「機械学習からのアプローチ」という副題を持つバイオインフォマティクスの教科書と，統計的な言語処理の教科書は，非常によく似た技術内容をカバーしている[1), 2)]．

膨大なデータの解析が必要なゲノム情報学は，データマイニング・機械学習といった情報技術を1つの基盤としてきた．ポスト・ゲノムの時代を迎えた生物学は，この流れをさらに一歩進めて，生命現象のシステムとしての理解へと向かっている．個々のDNA系列，タンパク質の構造の解明から，それらが全体的なシステムの中で果たす役割を同定しようとする機能的な生物学への移行である（図-2）．

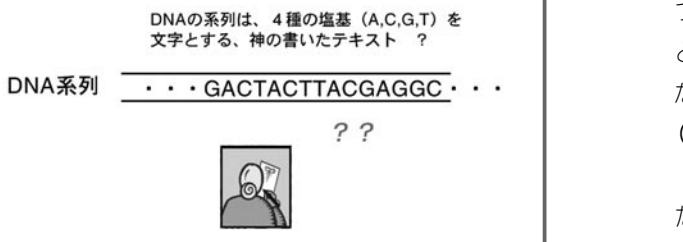



図-1　DNA系列の解読は未知言語の解読

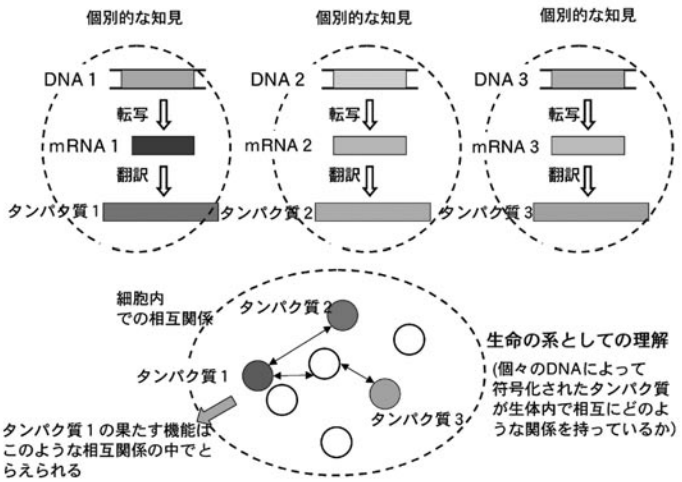



図-2　個別的な知見から総合的な理解へ－機能的な生物学

この視点の変化が，言語処理技術への関心の高まりの基盤にある．具体的には，2つの研究方向が見られる．1つは，言語処理手法をより高次なDNA系列の解析に適用すること，もう1つの方向は，テキストで表現された膨大な知見を有機的に体系化するのに言語処理の成果を使うことである．

## 言語的なもの

「言語的なもの」とは，何かを少し整理しておこう．言語とは，基本の記号（たとえば，単語）を組み合わせることで，複雑な意味を表現できる記号の系である．この系は，

**(1)［記号の組合せ方の規則］**

小さな単位の表現を組み合わせて大きなまとまりの表現を作る構造規則，たとえば，「赤い」と「靴」が結びついて「赤い靴」となり，さらにそれが「赤い靴を買う」という文の構成要素となるといった規則．規則に従わない単語の並びは，文とはならない．

**(2)［表現と意味の対応規則］**

個々の表現からその表現が表すものへの写像規則，たとえば，「小泉純一郎」という語の意味は，人物としての小泉純一郎である，といった対応規則である．

という2つの規則からできている．

日本語，英語など，人間が日常使う言語（自然言語）では，この構造と意味とが密接に関係している．個々の単語は，それ単独で存在するわけではない．周囲の単語と結びつき，より大きなまとまりの単位（句や節，文）を構成し，そのまとまりがさらにその周囲との関連で主語・目的語といった機能を持つ．同じ単語や句も，より広い文脈との相対関係で意味を変える．

前述のバイオインフォマティクスと言語処理の2つの教科書を比較すると，同じ統計処理・機械学習でも，言語処理の教科書では，語・句・節・文といったまとまりを自動的に見つけること，あるいは，大きな構造的なまとまりの機能の同定に力点が置かれている．

ただ，上のような2つの規則性を示す記号の系は，必ずしも，自然言語だけではない．ゲノム科学の対象となるDNA系列は，4種の塩基という基本の記号からできた言語であると考えることもできる．いま，ゲノム科学は，DNA系列とそれが符号化するタンパク質との対応という，いわば，文字の列から単語に相当する単位とその意味の認定から，それらの相互関係というより高次の構造と意味（機能）に関心を移行しつつある．ゲノム情報学が，構造が単純な領域での機械学習から，自然言語のような構造的に複雑な対象を取り扱う技術に興味を持つのは，自然な流れであろう．

この方向からの研究成果は，まだ，散発的ではあるが，機械学習的な手法を言語解析に適用していた研究者が，大量にこの分野に流入する構えを見せており，新しい成果が得られる可能性は高い．

## 生命科学における知識と文献

ゲノム情報学と言語処理のもう1つの接点は，言語処理・テキストマイニングの技術をゲノム情報学における知識の体系化と整理に適用することである．

個々のDNA系列とタンパク質の関係が分かったとしても，個々のタンパク質が他のタンパク質，生体分子とどのような相互関係を持つかを解明しない限り，生命系の全体は理解できない．生命現象の全体像の理解には，これまでの個別的な知見の相互関係を明示的に捉えることが必要となり，ここに，言語処理・テキストマイニングの技術が役に立つ．

実験・観察データという一次のデータを解釈した知見は，言語という手段（すなわち，論文）を使って表現される．この論文という知見の集合を再びデータとみなして，そこから新たな知識を発見する技術がテキストマイニングである．

実際，DNA系列，数値，タンパク質の構造といった一次的なデータを結び付ける我々の知見の大部分は，言語という手段を使って表現されている．すなわち，我々の対象認識を表現するには，自然言語が提供する分類化と構造化が不可欠である．

言語の意味規則の例として，「小泉純一郎」と人物としての小泉純一郎の対応を考えた．しかし，我々が使う語は，このような個体に付けられたラベル（固有名詞）だけではない．「保守政治家」，「改革者」といった分類，「首相」といった役割（機能）を表す単語が数多く現れる．こういった分類や機能にかかわる単語を使って，我々は，一般化された知識や，ある見方を反映した知識の系を構築している．個体をさまざまな基準で分類し，それにある見方から見た機能を付与していくことで，知識の系が作られていく．

特に，生物学，医学，薬学，生化学といった生命に関する科学は，観察対象や事象を分類することを基礎にする科学であり，そこでの知識はテキストによって表現される部分が大きい．物理学のような数式による知識の表現をとる科学とは，対照的である[3]．

語で表現される分類は，さらに，「真核細胞では，転写調整の機構が複雑である」という知見の表現に使われる．ここでは，「原核細胞」と「真核細胞」という分類の一方が取り上げられ，その分類に属する細胞の特徴を述べる命題が作られる．この命題へのまとめあげに，主語，目的語，述語という言語の構造が使われる．

ただ，言語によって表現された知見は，現在のところ，人間が個別の論文を逐一読むことによってしか消費できない．ある論文に現れた知見と別の論文中の知見とが関係するとしても，その相互関係は，2つの論文全体を読んだ読者にしか認定できない．関連ある知見を含む論文の数が，数万，数十万となると，論文中に散在する知見の相互関係を認識し，それを知識の系として再編成する作業には，禁止的な労力が必要となる．

この限界を打破する技術，たとえば，複数の論文の内容を統合し，1つのまとまりのある抄録を作るプログラム（複数文献の自動抄録）とか，膨大な文献集合に散在する特定の話題（たとえば，特定のタンパク質）についての情報を集約するプログラム（情報抽出，Q/A課題）の開発が，現在の言語処理の中心テーマとなっている．

言語的な手段による知識記述の度合いが大きい生命科学が，そのさまざまな分野での個別な知見を統合し，その体系化をめざしたとき，このような言語処理技術に期待するのは当然であろう（図-3）．

## 具体例 —シグナル伝達ネットワーク

話が抽象的になった．文献中に散在する個別知見の統合が必要な例として，生命科学者が興味を持つシグナル伝達系の問題を考えてみよう．

シグナル伝達系の研究とは，細胞外からの刺激が細胞核内に伝達され，ある特定の細胞応答を引き起こす制御の系を明らかにする研究である[4]．この研究は，

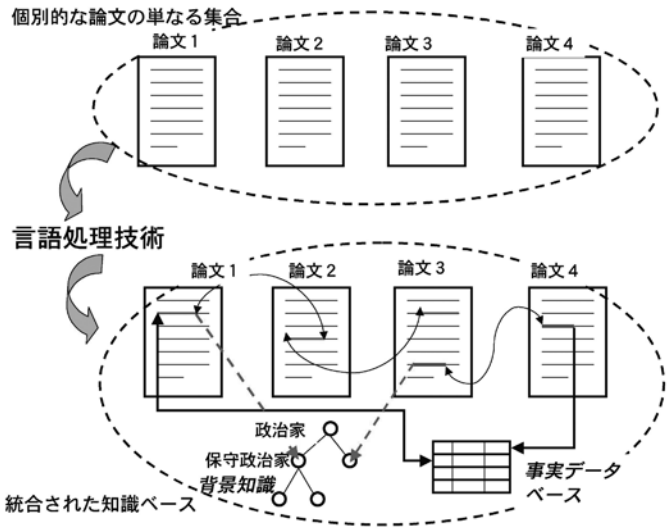
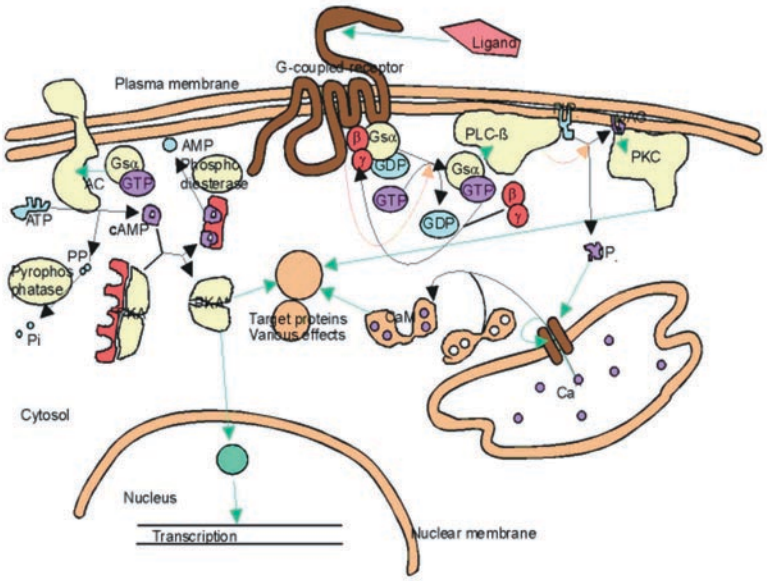



図-3　文献集合と統合された知識ベース




＊この図には，各種のタンパク質，生体分子，あるいは，タンパク質の部分が相互にどのように反応し，その結果細胞内でシグナルがどのようにやりとりされているかを示している．

＊＊図の上部が細胞の外，下部が細胞の内部，上下を分ける2重線が細胞膜である．また，図の最下部が細胞核で，その核内にDNAがある．図-2の下方のネットワークは，このような複雑なタンパク質の系を模式的に示したもの．

注）この図は，ドイツのTranspathプロジェクトのプロジェクト・サイト（http://transpath.gbf.de/maps/GcoupledR.html）からの引用．

図-4　シグナル伝達系の例

個々の細胞内での各種のタンパク質や化学物質，生体分子がどのように相互にシグナルを伝達し，制御の系を構成しているかを明らかにする，いわば，個別の生物種の，個別部位での伝達系を明らかにする個別研究として始まった．図-4に，そのようなシグナル伝達系の一例を示す．

実際，DNA系列の役割の解明には，そのDNAによって符号化されたタンパク質の役割，すなわち，そのタンパク質がシグナル伝達系の中で果たす役割を見なければならない．また，シグナル伝達系の理解は，DNA系列の異常と疾患の因果関係の理解，したがって，治療法や新薬開発など，実用上も，重要なものである．

しかし，実際には，生物種を人間に限定しても，あらゆる部位の中での細胞内での伝達系を明らかにすることは，多大の労力を必要とする．個々の部位での伝達系の全貌を，それ以前の研究成果を参照せずに，逐一解明するのは無駄である．別の生物種の伝達系での知見は，人間にも適用できる可能性が高い．また，タンパク質間やタンパク質と生体分子のインタラクションも，個々のタンパク質についての知見だけではなく，共有する部分構造など生化学的な一般化が可能である．

ただ，これは言うのは簡単であるが，実行は難しい．過去の伝達系研究の論文に埋もれた，現在の伝達系の研究に有効な知見をうまく参照しなければならない．また，数十万種あるタンパク質，その相互のインタラクションは，伝達系の研究とは独立にさまざまな生化学の論文に分散的に報告されている．結局，分野をまたがった膨大な数の論文に散在する知見を，うまく統合することで，個々の個別的な伝達系の解明を加速する必要がある．

このような知見の再編作業は，通常の文献検索システムだけで行えない．文献検索では，結局，数十万という膨大な数の論文を，再編作業をする人間が読む必要がある．むしろ，過去の縦型の研究をまとめた論文を横に切って，すでに報告されている知見をカタログ化することで，個別の伝達系の解明を加速する必要がある．関連文献を検索するだけでなく，そこに埋もれている個別知見の再編成を自動化する必要がある．

情報のカタログ化の一例を図-5に示す[5]．現状は，個々の論文から人手でこのカタログ化が行われているが，必要なカタログのごく一部がカバーされてい

るだけである．出版される論文数は加速度的に増大しており，それに伴って，論文中に埋もれた知見を知識の系に組み込む作業は，ますます難しくなっている．

## 情報抽出とシグナル伝達系

前節のカタログ作りは，テキストからある特定のタイプの情報を引き抜き，それを定型的なテンプレートに構造化する情報抽出（Information Extraction，以下IE）の典型的な課題になっている．

IEは，1987年から98年まで続いたMUC☆7で熱心な研究が行われた技術である．MUCでは，年度ごとに課題を変え，同一のテキスト集合からどれだけの精度で情報が抽出できるかを競争する枠組みがとられた．課題は年ごとに変わったが，視点を固定すると（たとえば，「いつ，どの会社が，どの地域で，合弁会社を作ったか？」），非定形なテキストから特定の情報を抽出し，構造化できることを示した．図-6は，合弁会社のIEの例である．

図-5のシグナル伝達系のカタログと図-6とは，いずれも，非定形なテキスト情報をテンプレートに整理するタスクになっており，その類似性は明らかであろう．新聞記事からの抽出（図-6）が可能であれば，論文からの情報の抽出（図-5）も可能なはずである．

単純化すると，

**事象のテンプレート**:どのような生物種のどの部位の細胞で，また，細胞中のどの場所で，どのようなタンパク質が，どのようなインタラクションをするか

という情報を抽出することが，カタログ作りの第一歩となる．実際，ケンブリッジ大学，シェフィールド大学，あるいは，我々のグループの研究は，MUCで開発したIE技術をタンパク質間のインタラクション情報の抽出に適用している[6]．

しかし，ゲノム情報の抽出にIEシステムを適用した結果は，あまり芳しいものではなかった．これには，いくつかの理由が考えられるが，最大の理由は，2つのタスクが持つ意味的な複雑度が大きく異なることである．合弁会社のIEが，会社名・地名・資本金・製品・地域といった比較的少数の意味クラスだけを対象とすれば良かったのに対して，ゲノム情報のIEでは，タンパク質・タンパク質のクラス・部分構造・さまざまな生体分子など，多様な意味クラスが出現する．

実際，我々はシグナル伝達系のIEに必要な意味クラスと，その階層関係を整理したGENIAオントロジーを構築しているが，このオントロジーには，30種類以上の意味クラスがある[3]．5～6種類の意味クラスを対象としたMUCのIEに対して，この分野のIEは，自然言語処理の観点からも，大きなチャレンジ課題となる．

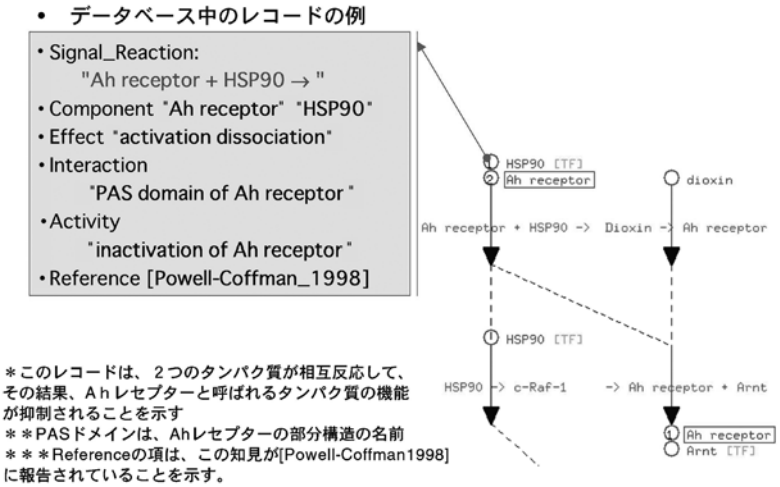



図-5　シグナル伝達のデータベースの例

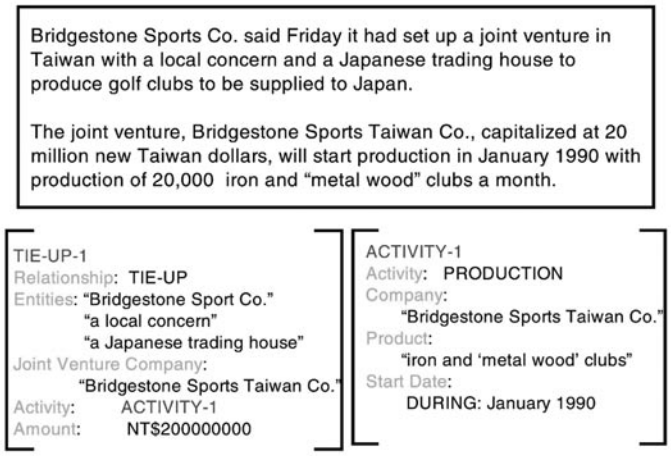



図-6　情報抽出の例

☆7 MUC: Message Understanding Conference

## 知識の枠

さて，現在のゲノム情報学が，これまでの個別的な知見を有機的に総合し，生命現象を系として理解しようとしていると述べた．そして，このトータルな系の理解に，論文からの個別的な知見の抽出と収集が必要である，ということであった．

例としてあげたシグナル伝達系では，前節のテンプレートに合致する個別事象を論文から抽出し，それを部品的な知見として蓄積する．この蓄積された知見を適宜参照することで，別の伝達系の解明を加速する．ここで重要なことは，このような個別事象を蓄えるためのテンプレート，および，それを埋める要素のタイプがあらかじめ定められていること，である．個別の事象に関する情報（個別的な知見）の抽出以前に，どのような形の個別事象が重要かに関する認識がある．インスタンス（具体例）の前に，タイプに関する知識がある，といってもよい．

知識の有機的な統合には，まず，このようなタイプに関する知識，知識の枠がまず設定されていなければならない．ゲノム情報学のIEの難しさは，この知識の枠が，従来のIEタスクよりもはるかに複雑であるということにあったが，それ以前に，このような知識の枠自体に，生命科学者の間に明示的な合意がない．このことは，生命現象に関係する分野が広がり，関与する専門家集団が多様になればなるほど，深刻な問題となる．

実際，共通の知識の枠を設定する必要性は，幅広い知識を共有する必要がある生命科学者の間で強く認識されてきており，分散的に作られたデータベース間の相互アクセスを可能にするための共通のオントロジー（すなわち，共通の知識枠）作りが世界的なコンソーシアムではじめられている．異なった組織や業種間の電子商取引に共通のオントロジーが必要となったのと同じ事情が，急速に分野が拡大する生命科学にも見られる．

言語処理においても，たとえば，テキスト集合から出現語彙の意味クラスを自動的に見つける問題（意味クラスタリング），あるいは，重要な事象クラスの自動認識（IEにおける抽出パターンの自動学習）などの研究が活発化しており，これらの技術をゲノム情報の知識の枠を設定する課題に適用していくことが，興味のある課題となっている．

## テキスト，知識，情報

本稿では述べられなかったが，たとえば，タンパク質の相互関係の予測にデータマイニング的な手法とテキストマイニングの手法を組み合わせることで，その精度が向上したとか，テキスト中での共起単語の統計的な性質だけを使ったクラスタリングで，生物学的に意味のあるタンパク質のクラスが発見できた，という報告もある．

1次の観察・実験データだけでなく，論文の集合を再びデータとしてみて，そこから新たな知識を発見するテキストマイニングの技法は，言語的手段に頼る度合いが高い生命科学には，特に有効であろう．

また，生命科学のように，環境とのインターラクションが大きい対象の研究では，情報の文脈依存性がきわめて高い．たとえば，論文から抽出された，「タンパク質Aはタンパク質Bと相互関係を持つ」といった個別的な知見も，実験環境や生体内・細胞内での部位，生物種といった文脈に依存する．抽出され，論文という文脈から切り離された個別的な知見は，それだけでは役に立たず，個別論文にグラウンディングされている必要がある．ここでは，整理され，一般化された知識（データベース，オントロジー）と状況に依存的な個別知識（テキスト）とが統合的に管理されていなければならない．

この統合的な情報管理システムでは，CBR☆8や類似文脈の判定などの言語処理技術が役に立つ技術となろう．言語処理が，将来のゲノム情報学に貢献する分野は非常に大きい．

**参考文献**

1) Bald, P. and Brunak, S.: Bioinformatics－ The Machine Learning Approach (second edition), MIT Press (2001).
2) Manning, C.D. and Schutze, H.: Foundations of Statistical Natural Language Processing, MIT Press (1999).
3) 辻井潤一，建石由佳: 生命の理解とオントロジー，数理科学，No.458，サイエンス社 (2001).
4) Fukuda, K., et al.: Knowledge Representation of Signal Transduction Pathways, Bioinformatics, Vol.17, Issue 9, OUP, pp.829-837 (2001).
5) Takai, T., et al.: A Database for Cell Signaling Networks, Jour. of Computational Biology, Vol.5, No.4, pp.747-754 (1998).
6) Yakushiji, A., et al.: Event Extraction from Biomedical Papers using a Full Parser, In Proc. of Pacific Symposium on Biocomputing 2001, pp. 408-419 (2001).

(平成13年11月12日受付)

☆8 CBR: Case Based Reasoning