解 説

ゲノム情報

10. 言語および図表データからの生物知識情報の抽出†

松本裕治†† 宮田高志†† 山下達雄††
藤尾正和†† シバスンタラン スハルナン††

1. はじめに

分子生物学においては，物質の変化や反応など分野固有の膨大な知識が存在し，それらはテキストおよび図表によって表現されることが多い．テキストと図の統合に関する先行研究としては，言語および図表データの統合的解析[1]および統合的生成[2),3)]などの研究があるが，多くは特定の狭い領域を対象としている．本稿では言語および図表によって記述された分子生物学文献からの知識の抽出，体系化および利用を目的としてわれわれの研究の概要を述べる．

2. 言語および図表データからの生物知識情報抽出

ゲノム解析から得られる生物学，医学関連の大量のデータが言語によるテキストおよび関連図などの図模式図的な表現によって蓄積されている．人間にとってはわかりやすい言語および図表に関するデータの相互関連やある特定の遺伝子などに関する知識を整理し知識ベース化する技術を構築することを目的としている．

言語表現を正しく解析することは言語の曖昧性の問題のため現在の技術では困難な問題である．本研究では，分子生物学関係の論文や教科書の中で概念図とテキストによって表現された記述からの知識抽出を行う手法を研究する．当初は，反応や変化をともなう現象を表現した図およびテキストの解析を行い，図表中のコンポーネントと文章中の表現とのあいだの対応関係の自動抽出を行う．知識抽出のための第一段階は，図表現のコンポーネントへの言語表現のリンクにより行われ，これにより，人間の利用者を対象としたハイパテキストの構築が容易になる．さらに，機械処理を考慮した知識の利用のために柔軟な検索や参照に対応する必要があり，その表現法に関する研究も合わせて行う．

一方，分子生物学関連の文献には一般の辞書には登録されていない多くの専門用語が現れる．そのため，辞書への未登録語についての品詞推定および意味クラスへの自動推定を合わせて行う必要がある．未定義語の意味クラス推定については，語構成の情報および共起関係による統計的手法を用いる．

3. 図と言語による生物知識の情報提示

専門分野に関して大量に蓄積されたデータを整理して適切な情報提示することは円滑な研究の進展のために重要である．上述の研究では，テキストと図による記述からの知識抽出を行っているが，こうして得られた知識を，質問者の意図や知識の程度に応じて適切な情報提示を行うことが重要である．

情報提示のための選択肢はいくつかのレベルが存在する．すなわち，どれだけの情報の提示を行うかを決定する内容のレベル，それを図と言語のうちのどの情報媒体をどの内容に対して用いるかを決定する媒体のレベル，および，過不足のない適切な表現を選択する表現のレベルである．

情報提示の方法には一般的な尺度を導入することが好ましいので，提示された情報に対する利用者の処理コストと伝達効果の相関関係によってその適切性をはかる．

図-1 に図表と文の解析による知識獲得および蓄積されたデータからの情報提示の流れを示す．これは長期的なテーマであり，現在は分野依存のテキストから統計的手法により精度の高い言語解析を行うためのパラメータ学習を行っている．

4. 遺伝子情報とテキストデータとの関連づけ

4.1 注釈文を併用したモチーフ抽出

近年遺伝子解析技術の発展にともない，大量の遺伝子情報が蓄積されつつあり，これらに対して計算機を利用したホモロジー探索・モチーフ抽出といった処理が行われている．そこでは手法としては（主に計算機科学の分野で）古くから提案されている文字列検索やパターンマッチが使われているが，一般に効率・精度の点で生物学的な知識をヒューリスティクスとして利用することが不可欠となっている．しかし従来の研究では利用した生物化学的知識はプログラムの中に非明示的に埋め込まれてしまい，どのような知識をどういう形で利用したのかについてはプログラム作成者の説明によるしかなかった．

† Extraction and Formulation of Biological Knowledge from Linguistic and Diagramatic Data by Yuji MATSUMOTO, Takashi MIYATA, Tatsuo YAMASHITA, Masakazu FUJIO and Sivasundaram SUHARNAN (Graduate School of Information Science, Nara Institute of Science and Technology).

†† 奈良先端科学技術大学院大学情報科学研究科

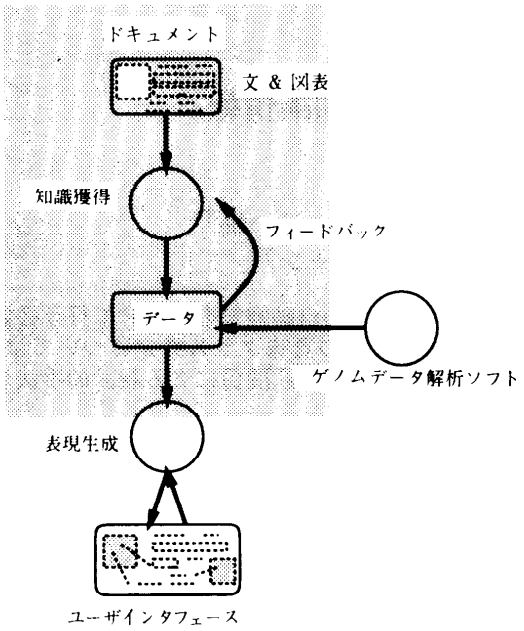

図-1 言語と図表からの知識抽出と情報提示

一方，現在利用可能な遺伝子データベースは単なる塩基／アミノ酸配列の集合ではなく，発見者の名前やその遺伝子／タンパク質の機能などのさまざまな注釈がつけられている．これらの注釈は人間の利便のために用意されたものであるが，使われている単語や構文などに生物化学的知識が反映されていると考えられる．一般に専門用語はその分野に特有な生産的かつ系統的な構造をもっていることが多いので，とくに注釈文中の単語の情報は有効なヒューリスティクスとなる可能性が高い．

そこで本研究では遺伝子データベース中の注釈文を併用したモチーフ抽出を行う．すなわち，塩基／アミノ酸配列とその注釈文の対を1つのデータとみなし，統計的手法に基づいて有用なモチーフを抽出する．

4.2 既知の遺伝子情報と文献データとの有機的結合

大腸菌遺伝子の配列データの解析と文献情報との有機的な結合を実現するデータベースの開発を行っている．

・ゲノム1次配列からコンピュータ解析を行いORF(Open Reading Frame)領域の予測を行う

・大腸菌遺伝子の研究成果である文献情報よりキーワードおよび関連情報を抽出できる電子化データベースを開発する

以上の情報と機能既知の遺伝子との関連づけを行う．機能未知の遺伝子領域に関しては，モチーフ解析，ホモロジ解析，実験から明確にされた結果などからの構造上の類似性より上記既知遺伝子と比較することにより未知遺伝子の関連情報との関連づけを試み，遺伝子群の機能の予測およびその解析を行う．

参 考 文 献

1) 中村，古川：概念図理解を目的としたパターン情報と自然言語情報の統合，情報処理学会論文誌，Vol.36, No.1, pp.196-205 (Jan. 1995).

2) Feiner, S. and McKeown, K. : Coordinating Text and Graphics in Explanation Generarion, AAAI-90, pp.442-449 (1990).

3) André, E. and Rist, T. : Generating Coherent Presentations Employing Textual and Visual Material, Artificial Intelligence Review, Vol.9, No.2-3, pp.147-165 (1995).

（平成8年8月26日受付）

松本 裕治（正会員）

1955年生．1977年京都大学工学部情報工学科卒業．1979年同大学院工学研究科修士課程情報工学専攻修了．同年電子技術総合研究所入所．1984年-85年英国インペリアルカレッジ客員研究員．1985-87年（財）新世代コンピュータ技術開発機構に出向．京都大学助教授を経て，1993年より奈良先端科学技術大学院大学教授，現在に至る．専門は自然言語処理．人工知能学会，日本ソフトウェア科学会，言語処理学会，AAAI，ACL，ACM各会員．

宮田 高志（正会員）

1968年生．1991年東京大学理学部情報科学科卒業．1993年同大学院理学系研究科情報科学専攻修士課程修了．1996年同大学院博士課程退学．現在奈良先端科学技術大学院大学情報科学研究科助手．自然言語処理，特に対話の理解およびその計算機上での実現に興味をもつ．日本ソフトウェア科学会，人工知能学会，言語処理学会，ACM各会員．

山下 達雄

1972年生．1995年広島大学総合科学部総合科学科卒業．同年奈良先端科学技術大学院大学情報科学研究科博士前期課程入学．自然言語処理，特に形態素解析に興味をもつ．人工知能学会学生会員．

藤尾 正和

1972年生．1995年京都大学理学部生物学科卒業．同年奈良先端科学技術大学院大学情報科学研究科博士前期課程入学．HPSGなどの語彙主導型文法における構文解析に興味をもつ．日本ソフトウェア科学会学生会員．

シバスンタラン スハルナン(正会員)

1970年生．1996年山梨大学卒業．現在奈良先端科学技術大学院大学情報科学研究科前期博士課程在学中．