GBI(Graph-Based Induction)法の拡張による化学物質からの部分構造抽出方法の検討

田中　栄太朗†　　速水　亜希子†　　稲積　宏誠‡

青山学院大学大学院理工学研究科経営工学専攻†

青山学院大学理工学部情報テクノロジー学科‡

１．はじめに

今日、大量の生物・化学データベースが存在する。それらの基礎データに基づいて新規物質の合成が試みられることになるが、可能な組合せ全てに対して行うことは膨大な時間と費用がかかるといわれている。化学合成や創薬における支援システムは多くの試みが成されているが、我々は、既存の化学構造データベースから、グラフマイニング手法である GBI(Graph-Based Induction)法 [1] に注目した試みを行っている。まず、GBI 法の改良を行いながら共通部分構造を抽出することによって、抽出した部分構造の組合せにより各化学物質を表現し、決定木分析可能とする基礎情報として提供してきた[2][3]。

GBI 法では分類に関わる多くの部分構造が抽出されるが、その部分構造の意味づけや解釈を行うのは容易ではないため、得られた部分構造が有効な知識として活用されないことも多く存在する。そこで、本研究では対象とする問題の領域知識やヒューリスティックスを GBI に組み込むための拡張を行う。そこで、領域知識やヒューリスティックスに基づいて特定の部分構造を指定し、それを起点とした探索領域から得られる部分構造抽出を可能とする方法を検討する。これにより、特徴分類に寄与する部分構造の位置づけが明確となり、その意味づけと解釈が容易になることが期待される。本稿では、特に生理活性物質に対してこの手法を適用し、その有効性について検証する。

２．入力化学物質と抽出部分構造の可視化

抽出した部分構造による解釈を促すためには、専門家にとって、入力に用いた化学物質群や抽出された部分構造やその抽出過程を可視化することが不可欠である。我々は従来、XML による化学物質データベースの仕様である CML(Chemical Markup Language)形式のファイルを入力データとして分析を行い、処理後の分析には化学物質のデータベース・検索ツールである CambridgeSoft 社の ChemFinder を用いてきた。これに対して、データ形式に依存しない分析、Java 3D による化学構造の可視化、GBI 適用条件・抽出過程の表示を実現した。図１に Java Swing による GUI イメージを示す。

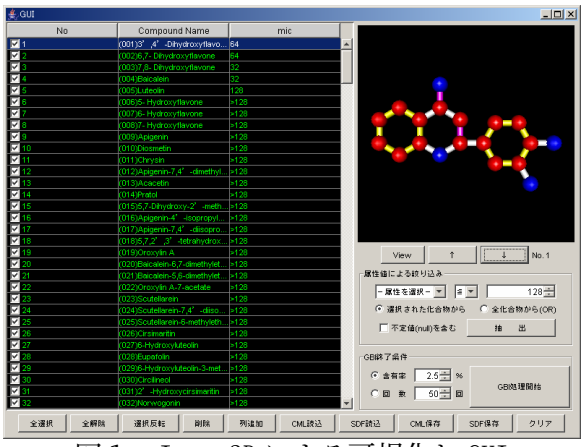

図１　Java 3D による可視化と GUI

３．部分構造抽出処理の一時停止と分岐

有効な部分構造を発見する最も基本的な考え方は、異なるクラスを示す事例集合を対象にして別々に処理を行うこと、クラス分類能力を示す評価基準を部分構造抽出条件に入れることである。しかし、本来期待される特徴上の性質としては、対象とする物質群においてほぼ共通に含まれている共通構造を示した上で、それに付加されるどのような部分構造が性質の違いを決定しているのかという視点であろう。これを実現するためには、特徴の違いによらず共通にもつある一定の部分構造を抽出し、その条件のもとにクラスを特徴づけている付加情報を抽出するというアプローチが必要である。

部分構造抽出処理が進むと共に、抽出部分構造を構成する原子の数が増加するが、それを含む物質数の割合（含有率）は低下する。図２に実験に用いた生理活性を示すフラボノイド類を例にした、抽出部分構造に含まれる原子数(水素原子を除く)と含有率の変動を示す。本研究では、含有率が最大となるペアをチャンクの対象としているので、含有率は単調減少となる。

The substructure extraction from molecules by extension of the GBI method

† Eitaro TANAKA, Akiko HAYAMI　Graduate School of Science and Engineering, Aoyama Gakuin University

‡ Hiroshige INAZUMI　School of Science and Engineering, Aoyama Gakuin University

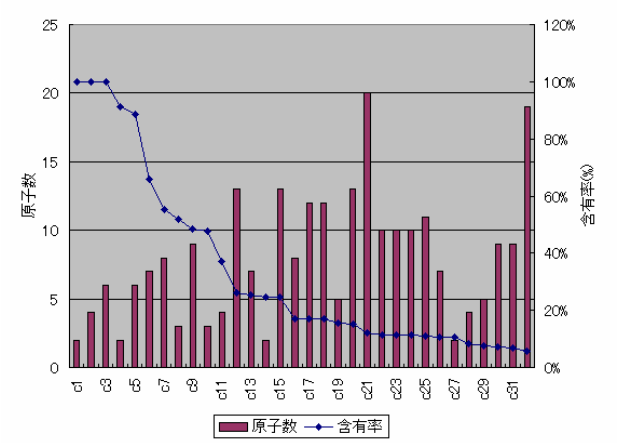

図２　原子数と含有率のグラフ

通常の GBI 法の適用では、小さな部分構造はどの化学物質にも含まれることが多く、クラス分類能力は低く、大きな部分構造は単体の物質そのものを示すことに近くなる。

そこで、まず多くの部分構造を抽出しておき、そこから部分構造の形､大きさ､含有率を考慮し、任意の部分構造 M を選択する。次に、M で一旦処理を停止する。ここで、入力化学物質群を M を含む化合物群と含まない化合物群に分割する。 M を基本骨格に準ずる構造とみなし、 M からどのように置換基が付くかという情報を得ることができる。M を含む化合物群に対する処理においては、与えられたクラスごとの物質群に対して行う。

４．実験例

抗菌活性値により分類されるフラボノイド類325種類に対して実験を行った。終了条件を最小含有率 5%とし GBI 法を適用すると 32 個の部分構造（Chunk1～Chunk32）が抽出された。そこで、その中からフラボノイド類における基本骨格に類似した構造である Chunk15（図３）を選び、活性値(MIC)128を閾値として高活性群と低活性群２つの群に分け、それぞれの群に対して別々の処理を続けた。

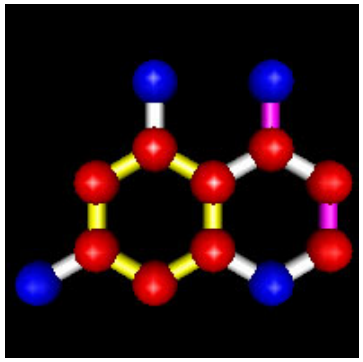

図３　Chunk15

高活性群から抽出された部分構造（Chunk18）を図４に、低活性群から抽出された部分構造（Chunk16）を図５に示す。

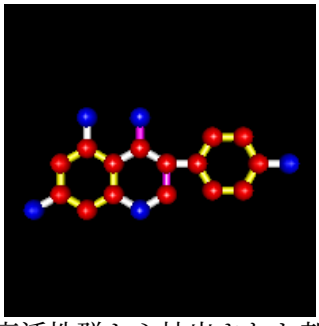

図４　高活性群から抽出された部分構造

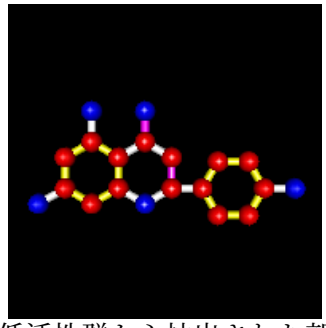

図５　低活性群から抽出された部分構造

このように、同じ置換基でも結合する場所によってクラスが異なることが示された。このような部分構造は、従来の手法では発見できない構造である。

５．結論

本稿で提案した手法により、化学物質や抽出部分構造を可視化、任意の時点で探索戦略を変えることを実装上実現した。これを応用することによって、専門家が事前に指定した構造と一致した部分構造、専門家が抽出結果を見ることによって初めて気付いた部分構造、ある条件を満足した部分構造などを起点とすることによって、探索対象となるデータ集合のグループ分けを行いながら実行することができるようになった。

今後の課題として、分岐後の処理において活性値を考慮したチャンク対象ペアの得点付けを行い、よりクラスの特徴を示す部分構造が抽出されるように改良することが挙げられる。

参考文献

[1] 松田喬、元田浩、鷲尾隆：“一般グラフ構造データに対する Graph-Based Induction とその応用”,人工知能学会論文誌,Vol.16 No.4 A pp.363-374 (2001)

[2] 田中栄太朗、津田哲夫、吉澤有美、稲積宏誠：“化学構造データベースからの有効な部分構造抽出法に関する考察”, 情報処理学会第 65 回全国大会, 3 巻 pp.137-138 (2003)

[3] 速水亜希子､田中栄太朗、吉澤有美、稲積宏誠：“化学構造情報を用いた知識発見と知識表現に関する考察”, 情報処理学会第 65 回全国大会, 3 巻 pp.141-142 (2003)