数値属性を対象とした化学構造マイニングツールの開発と変異原性データへの適用

関西学院大学大学院理工学研究科
情報科学専攻岡田研究室　M7429 松本直久

1. はじめに

変異原性を示す物質は，突然変異を引き起こす物質をさし，体内に入ると染色体に作用し，ガンや遺伝子病など染色体の異常による疾患を引き起こす場合がある．変異原性を調べる試験としてネズミチフス菌や大腸菌を用いた Ames 試験や哺乳類培養細胞を用いた染色体異常試験が行われ，特定の化学構造が変異原性の原因ではないかと考えられている．また，どの程度の投与量で影響を与えるかが重要であるため，これらの試験では増殖して形成されるコロニーの数や，被験物質によって細胞の20%に異常を誘発した時の最少用量値(d20)が調査される．形成されるコロニー数が多いことや d20 値が低いことが高い変異原性を意味する．

本研究では染色体異常試験に対して変異原性に影響を与える特徴的な部分構造の発見を目的とする．解析を行うにあたって数値目的変数を使用するため，既存の化学構造精錬システムを数値目的変数が扱えるように発展させた．また導き出されたそれぞれの活性基本部分構造を持った化合物群の重なりを知るため,新たな化学マイニングツール群を作成した．

2. 活性部分構造の発見

当研究室では，生理活性分子における特徴的な部分構造の発見をテーマとして研究が行われている．解析の流れを下図に示す．

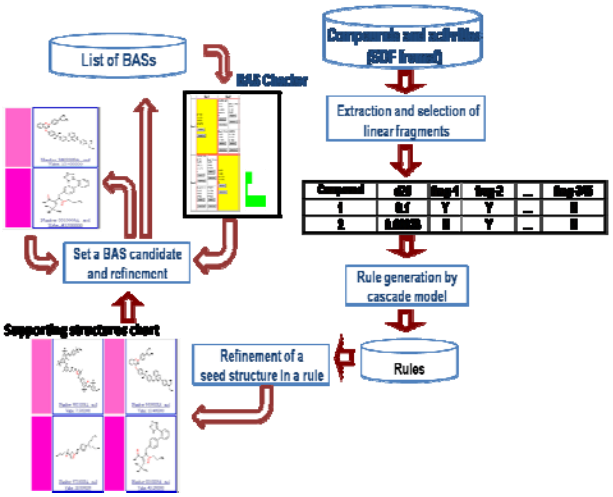

図 1　解析の流れ

まず，与えられた化合物群の構造式データから線形フラグメントを生成する．線形フラグメントとは構造式中の枝別れのない部分構造を文字列で表記したものである．得られたフラグメント群の中で相関の高いフラグメントを除去し，図 1 中央にあるような活性値とフラグメントの有無(y/n)を示した表を作成する．この表からカスケードモデルを用いてルール群を導出する．得られたルールだけでは解釈ができない．そこで構造精錬システムを用いて，より大きな構造に拡大させ，活性基本部分構造(BAS)を同定する[1]．なお，BAS に分類される化合物群間の重なりを確認するために BAS Checker を作成した．

3. 化学構造マイニングシステムの拡張

既存の構造精錬システムを数値目的変数に対応できるように改変すると共に，得られた基本部分構造を確認するために BAS Checker を作成した[2].

3.1 構造精錬システムの拡張

ルール条件文に現れるフラグメントを種として BSS 値が下がらない限り，周辺の原子を付加しより大きな部分構造を生成する．以下に種として与えたカルボン酸が安息香酸に成長した例を示す(丸の部分がふかされている)．

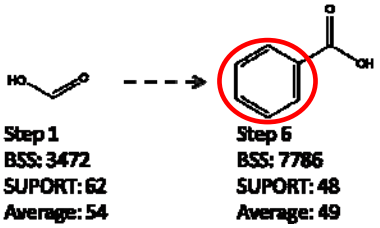

図 2 精錬の例

サポートに示す化合物数が減っているが，d20 値の平均値が減っていることからより特徴的な BAS が抽出されたことがわかる．このように平均値を出力させることで，活性値が全体平均から離れた高活性化合物群や低活性化合物群への精錬が可能となった．

数値目的変数を用いた精錬を行う際に，これまでとは異なり BSS 値の表現に次式を採用して精練を行えるようにシステムを変更した

$$BSS^g = n^g \cdot (\overline{x}^g - \overline{x})^2$$

ここで $\overline{x}^g$は精錬後の活性値の平均値，$\overline{x}$は精錬前の平均値，n^gは化合物数を示している．

支持化合物群の構造式を出力する SSC(Suporting Structures Chart)の画面にて，化合物群を活性値でソートし，図 3 のように活性値の強さによって色を変化させ

た．その結果，数値の高低がわかりやすく表示され，視認性が高まった．

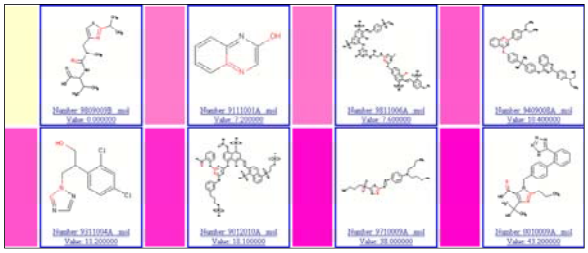

図 3　構造描画画面

3.2 BAS Checker

ここでは精錬システムから BAS1 と BAS2 が同定され，さらに BAS 群の追加を試みていると考えよう．この時，図 4 上部の 2 種の表を出力するようにした．

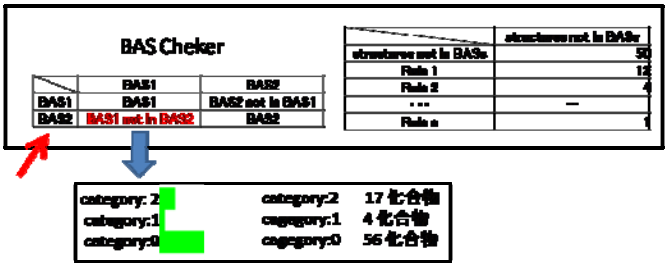

図 4 BAS Checker 出力画面

例えば矢印で指した部分は BAS1 を持ち，BAS2 を持たない化合物数を表す．ここをクリックすると，対応する化合物群の活性度分布を表す棒グラフが図下部のように出力される．また，ユーザは棒グラフの分割値，カテゴリ数を指定できる．さらに右上の表では指定した BAS 群でカバーできていない化合物群(structures not in BASs)とカスケードモデルから得られた各ルールの支持化合物群の重なりを示しており，どのルールを調査すればよいかのヒントを与えている．

4. 結果と議論

4.1 対象データと処理過程の概略

本研究では，有機化合物を対象とした Ames 試験データ 902 件，染色体異常試験データ 882 件のデータをそれぞれ取得した．Ames 試験では変異原性の無い化合物はコロニー数を 0，さらに染色体異常試験において negative, equivocal を示す化合物に関しては d20 値を 100 と設定した．

フラグメントの最大の長さを 10 として，Ames 試験データから 24,250 種，染色体異常試験データでは 24,354 種のフラグメント群を生成した．出現頻度 3%～97%の範囲を選択し，かつ相関係数 0.9 以上のフラグメント対から一方を削除したところ，Ames 試験では 404 種，染色体異常試験では 412 種のフラグメント群を得た．上記のフラグメントを説明変数としてルール導出を行った結果 Ames 試験では 26 種，染色体異常試験では 42 種のルールが導出された．

4.2 解析結果

Ames 試験データから変異原性を示す BAS が 9 種，示さない BAS が 10 種導出された．染色体異常試験では変異原性を示すような BAS が 11 種，示さない BAS が 16 種導出された．これらの BAS により Ames 試験では全体の 27%，染色体異常試験では 34%の化合物群を説明している．図 5 に活性カテゴリごと説明した割合を示す．

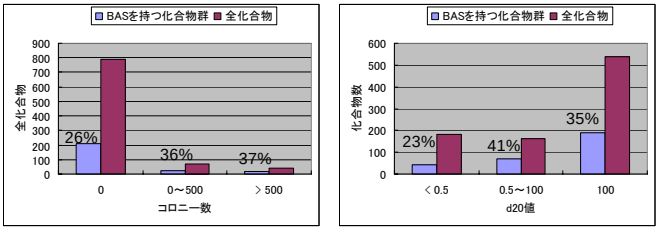

図 5 化合物数の割合

BAS で説明された下の図 6 と図 7 において Ames 試験で得られた変異原性を示す/示さない BAS の一部を紹介する．

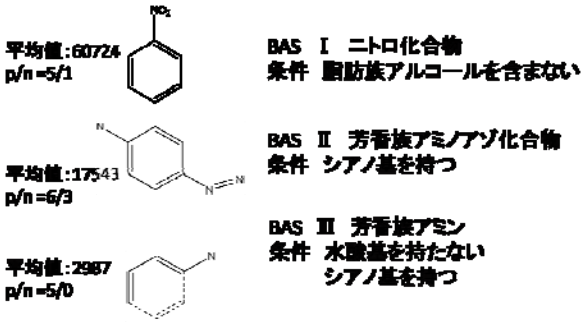

図 6 変異原性を示す BAS

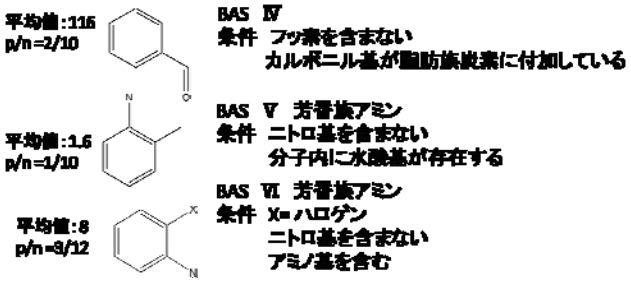

図 7 変異原性を示さない BAS

5. まとめ

対象化合物群のかなりの割合を説明する BAS 群の抽出に成功した．より精練が成功しやすいマイニング法の開発や探索範囲を広げた精細な解析が期待される．

参考文献

[1] S. Fujishima, Y. Takahashi, T. Okada, SCCJ, Vol. 7, No. 2, pp. 63-70, 2008.

[2] 大森紀人，藤島悟志，森幸雄，堀川裕志，山川眞透，岡田孝: "活性構造知識ベース構築とそのソフトウェア基盤", 第 30 回情報化学討論会,pp.13-14, 2007