2S05

化学物質の発ガン性の予測

○田辺 和俊[1]、鈴木 孝弘[2]

[1]ティーエヌケー（〒300-0848 土浦市西根西 1-7-12）
[2]東洋大学自然科学研究室（〒112-8606 東京都文京区白山 5-28-20）

1. 緒言

ガンは1980年以降、日本人の死因の第1位であり、近年では死亡者の30％を占めている。ガンの発生には、物理的（放射線、紫外線等）、化学的（発ガン性化学物質）、生物的（ウイルス、細菌、遺伝等）等、様々な原因が上げられる。その内、飲食物や喫煙等により体内に取り込まれる発ガン性化学物質が最大原因であることが明らかになっている。そのため、化学物質の発ガン性の情報を把握することが不可欠であるが、我々の周りには発ガン性不明の化学物質が氾濫しているのが現実である。

発ガン性等の化学物質の毒性の評価には通常、動物を用いた試験が行われ、きわめて長い期間、莫大な費用と多数の動物が必要である。また、近年では動物愛護の観点から毒性試験が問題になっており、特に欧州では動物実験に対して厳しい法規制がとられている。したがって，発ガン性未知の膨大な数の化学物質の全てについて動物実験により発ガン性を評価することは現実には不可能である。

そこで、化学物質の発ガン性評価の動物実験に代わる手段として、コンピュータを用いる発ガン性の予測技術の確立が渇望されている。そのため、化学物質の発ガン性を構造から予測する手法の研究開発が欧米では活発に行われている。芳香族アミン等の同族体については、発ガン性を比較的高い精度で予測できる場合もある。しかし、我々の周囲に存在する発ガン性未知の化学物質の構造は多種多様であり、任意の構造の化学物質の発ガン性を十分な精度で予測できる手法は未だに存在しない。

非同族体の予測が困難な原因の 1 つは、発ガン性予測モデルを開発するために必要な信頼性の高いデータの不足である。発ガン性データは NTP、IARC、EPA）等の幾つかの機関で収集・公開しているが、データの信頼性を表す発ガン性の格付けに統一性がなく、混乱している。

非同族体の予測が困難なもう 1 つの原因は、不特定の化学物質の発ガン性を構造情報のみから予測できるかという点である。QSAR は元来、毒性発現機構が類似する同族体を対象としたものであるが、現在までに発ガン機構が判明している化学物質は数十種程度にすぎない。したがって、発ガン機構に基づいて広範囲の化学物質の発ガン性を予測できる手法を開発することは現状では不可能である。

このようなきわめて複雑な問題に対して、現在、有効と考えられるのが人工ニューラルネットワーク（ANN）である。発ガン性予測に ANN を適用した研究もあるが、その対象は同族体に限られている。しかも、ANN には局所解、過学習、計算時間等、多くの問題があることが指摘されている。発表者らは ANN を用いて PTC の発ガン性データを解析したが、多数の局所解の存在のために最適解が得られず、予測精度を確定できないという問題があった [1]。

そこで発表者らは、近年、非線形解析法として注目されているサポートベクターマシン（SVM）を適用して非同族体の発ガン性予測の可能性を検討してきた。SVM は、ANN において深刻な局所解の問題がないことや、処理がきわめて高速なため大規模な問題にも簡単に実行できること等の利点がある。発表者らは SVM を用いて PTC のデータを解析し、非同族体の発ガン性予測に SVM が有効であることを実証した [2]。しかし、PTC のデータは物質数が少ないため、広範囲の化学物質の発ガン性予測の有効性については明らかにできなかった。そのため、大規模な非同族体の発ガン性データを用いて、広範囲の化学物質の発ガン性を高精度で予測する実用的なモデルの開発が求められている。

発表者らは多種多様な化学物質の発ガン性について信頼性の高いデータを集積した発ガン性 DB の構築、および発ガン性未知の化学物質についてその構造から発ガン性を高精度で予測するシステムの構築の 2 点を目的として研究してきた。本講演ではこれまでの研究成果を総合して報告する。

2. 発ガン性データベースの開発

発ガン性 DB は、IARC、EU、EPA、NTP、ACGIH、JSOH の 6 種の DB からデータを収集した。それぞれの DB では、化学物質の発ガン性は動物実験の信頼度により幾つかのランクに分類されているが、DB によりランクの数が異なっている。また、その信頼度に関する表現が異なり、表現の違いが分かりにくい。さらに、同一の化学物質でも DB によって異なるランク付けがされている物質が多数存在する。

このような各種の発ガン性 DB における信頼性ランクの不統一を解決するために、種々の DB における信頼性を総合的に評価し、ランク付けの統一を図り、発ガン性のランクを 6 段階に統一的に格付けした。以上の手順により発ガン性 DB に収録できた総物質数は 1,512 種である。

3. 発ガン性予測システムの開発

3.1 方　法

3.1.1 発ガン性データ

上記の発ガン性 DB には純有機物以外の様々な物質（金属、金属塩、混合物等）も収録されているので、H, C, N, O, F, Si, P, S, Cl, Br および I 以外の原子を含むもの、キシレン異性体等の混合物、ポリビニルアルコール等の高分子、アスベスト等の化学構造が特定できないもの、等は除いて、構造が明確に規定される有機化学物質 911 種を抽出し、予測モデルの構築に用いた。

3.1.2 記述子データ

以上の化学物質について、Corina プログラムを用いて平面構造から立体構造を生成し、構造を最適化した。次に、記述子作成プログラム Dragon 5.4 を用いて 1,504 種の記述子を作成した。ただし、これらの記述子は物質数に比べて明らかに過多である。そのため、予測に有効かつ不可欠な説明変数をスクリーニングする必要がある。本研究では、変数選択法としては若干厳密性に欠けるが、迅速性を優先して、発ガン性データと各記述子との単相関係数を計算し、相関の高い記述子から採用する個数を変えながら予測精度を調べて最適数を探索する方法を用いた。

3.1.3 SVM のモデル

SVM のソフトウェアは LIBSVM ver.2.89 を用い、発ガン性データに信頼性のランクに応じた重みを設定し、重み付きデータに対する 2 群分類機能を用いた。SVM は ANN と同様、過学習の問題があり、モデルの最適化が必要である。LIBSVM では最適化すべきパラメータが多数あり、中でも g(gamma)と c(cost)の設定が重要であり、これらと記述子数の計 3 個のパラメータについては最適設定が不可欠である。

そこで、以下の Dual Cross-Validation Test により、SVM の学習・テストと最適化を同時に行った。

①解析対象の化学物質を 10 群に分割する。

②その内の 9 群を学習用とし、この群について Leave-One-Out を用いてパラメータ g(gamma)と c(cost)および記述子数を最適化する。

③その最適モデルを用いて、テスト用物質の発ガン性を予測する。

④以上の手順を学習用とテスト用物質を入れ替えながら 10 回繰り返し、全ての物質について発ガン性を予測する。

モデルの性能評価には総合正解率 (OA) を用いた。

$$OA = \frac{TP + TN}{TP + TN + FP + FN}$$

ここで TP は true positive、TN は true negative、FP は false positive、FN は false negative の物質数である。

3.2 各種モデルの予測結果

3.2.1 単一の SVM による一括予測

まず、単一の SVM を用いて全物質 911 種を一括して解析するモデルを検討した。記述子の数を 50 個から増していくと正解率は向上するが、250 個以上に記述子が増えると正解率は減少し、過学習状態に陥る。250 個の記述子を用いた時に正解率の最高値 68.8%が得られた。この正解率は動物実験代替の予測手法としては満足できる性能ではない。しかも、発ガン陽性の物質を陰性と誤判定する FN の比率が

37%と高いことは致命的である。

3.2.2 SVMとアンサンブル学習（ブースティング）の組み合わせによる予測

次に、多数のSVMを組み合わせたアンサンブル学習の中で最もよく知られているAdaBoostを検討した。この方法は、誤分類率に応じて（adaptive）重みを変えるブースティングであり、以下のアルゴリズムで学習を行う。

① N個の全データに最初は均等な重み1/Nを割り当てる。

② 全データを用いて1台目の学習器h_1を学習し、不正解率ε_1から次式により信頼度β_1を求める。

$$\beta_i = \frac{1}{2}\ln\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right)$$

③ 1台目の学習器が正解したデータは重みを$\exp(-\beta_1)$倍し、不正解のデータは重みを$\exp(\beta_1)$倍する。

④ 2台目以降の学習器h_iについて同様の重み付きの学習を繰り返す。

⑤ 以上の方法でM台の学習器を作り、信頼度付き多数決で判別器としての正解率を計算する。

$$y = \frac{1}{M}\sum_{i=1}^{M}\beta_i h_i(x)$$

以上のアルゴリズムに従ってSVMを用いて10台の学習器を作成し、発ガン性データの学習とテストを行った。しかし、学習器が増えるほどその正解率が徐々に低下すると共に、全体の総合正解率も徐々に低下した。この方法は全データを単一のSVMで学習するため、発ガン性データの場合にはそれらを多数組み合わせても全体の正解率は向上しない。

3.2.3 SVMとアンサンブル学習（バギング）の組み合わせによる予測

アンサンブル学習の中のバギングを検討した。この方法では、N個のデータから重複を許してデータK個をサンプリングして数台の学習器を作り、AdaBoostと同様、多数決で正解率を計算する。発ガン性予測では、データ全体でなく部分集合を学習するという利点を生かせば、AdaBoostより高成績が期待できる。そこで、このバギングの修正法として、全データを幾つかのグループに分け、グループごとに学習する方法を検討した。

まず全データをSVMで学習・予測しながら、発ガン性の実測値を用いて陽性と陰性の誤判定（FT、FP）の物質を枝刈りする。この操作を繰り返すと、3段目で陽性211物質、陰性321物質、計532物質の第1群が生成され、このグループでは94.9 %という高い正解率が得られた。次に、ここまでに枝刈りされた残りの379物質について同様の操作を行うと、2段目で陽性187物質、陰性7物質、計194物質の第2群が生成され、このグループでは98.5 %という高い正解率が得られた。残りの185物質は第3群を形成し、このグループでは97.8 %の正解率となった。

このようにして、911種の全物質が532、194、185物質の3群に分けられ、それぞれ90%以上の高率で予測できることは分かったが、全物質をこれら3群に振り分ける方法がない。そこで、全物質についてこれら3個のSVMで予測した結果を親SVMに入力して発ガン性を予測することにした。その結果、TPが268、FPが141、TNが361、FNが141で、全体の正解率は69.0 %となったが、この成績は以上の方法と同程度であり、満足できる正解率ではない。

3.2.4 SVMと決定木の組み合わせによる予測

一括モデルによる予測結果について、陽性・陰性に判定されたグループをさらに個別にSVMで学習し、その操作を続けて多数のSVMを決定木状に組み合わせるモデルを検討した。その結果、最下段だけでなく途中で止まった枝も含めて、TPが297、FPが174、TNが328、FNが112で、総合正解率は68.6 %となった．これまでの方法と比較してFNの比率は低下したが、正解率は同程度であり、やはり満足できる結果ではない。

3.2.5 敗者復活を取り入れた決定木とSVMの組み合わせによる予測

敗者復活を取り入れたSVMを決定木状に組み合わせたモデルを検討した。すなわち、決定木の2段目のSVMで陽性･陰性と判定されたグループ同士をまとめて再解析する方法である。その結果、TPが311、FPが165、TNが337、FNが98となり、FNの比率がかなり低下し、総合正解率も72.0 %となった。これまでの方法の中では最高の予測成績が得られたも

のの、動物実験代替の発ガン性予測手法としてはやはりまだ満足できる性能ではない。

3.2.6 同族体に対する SVM の直列組み合わせによる予測

これまでの 5 種類のモデルでは満足できる結果が得られなかった原因は、機械的なデータ分割法に改良の余地があると考えられ、解析対象の化学物質を発ガン機構に関連させて分割し、個別に解析することにより予測精度の向上が期待される。

そこで、同じ部分構造を含む同族体では発ガン機構が同じであると仮定し、発ガン性を高精度で予測できるような同族体を探索し、それぞれ個別の SVM で予測した結果を組み合わせて全物質の発ガン性を予測するモデルを検討した。そのためにまず、Dragon 記述子を用いて、種々の同族体の物質数を集計した。

次に、これらの部分構造の中から発ガン性を高精度で予測できるような同族体を選定した。その際、同族体の選定条件として、正解率 80%以上と物質数 50～150 程度を目標に同族体の候補を探索した。このようにして 23 種の同族体を選定した。

さらに、これらの同族体を直列に組み合わせるモデルを検討した。すなわち、まず第 1 段において、選定した 23 種の同族体ごとに個別に SVM 解析を行い、その中から正解率最高の ArHC を選出し、これを含む同族体 54 物質について 87.0%の正解率を得た。

次に第 2 段において、残りの 857 物質について同じ操作を繰り返し、Ketone を含む同族体 58 物質について正解率 86.2%を得た。以上の操作を全物質がなくなるまで繰り返した結果、17 段の直列分岐を繰り返すことで、解析対象の全化学物質 911 種がどれかの SVM で予測できるモデルを構築できた。最終段までの予測結果は、TP＝328、FP＝95、TN＝407、FN＝81 となり、80.7%の総合正解率が得られた。

この 80.7%という正解率はこれまでの 5 種のモデルの成績より大幅に向上しており、また、既存の発ガン性予測モデルの性能（約 70%）をはるかに凌駕するものであり、さらに誤判定 FN の比率もきわめて低いので、動物実験代替の発ガン性予測手法としては満足できる性能である。

しかし、このモデルには頑健性の点で問題がある。すなわち、18 種の同族体の内で、第 8 段の Amide の 28 物質、第 10 段の AliAmine の 26 物質等のように物質数の少ないものが幾つか存在する。このような物質数が少ない同族体は若干のデータの追加削除により正解率が大きく変動し、モデル全体の頑健性を低下させる可能性がある。

3.2.7 同族体に対する SVM の並列組み合わせによる予測

同族体の並列組み合わせ法を検討した。すなわち、予測したい化学物質が含む同族体に対応する SVM の予測結果の多数決で陽性・陰性を判定した。そこでまず、23 種の部分構造について分割や融合を多数回試行した結果、20 種の同族体を選定すれば、ここで解析対象とした全物質が含まれ、全同族体での平均正解率は 79.8%となることが分かった。

次に、この結果を用い、各化学物質ごとに該当する同族体の予測値の多数決により陽性・陰性を判定し、全物質に対する正解率を算出した。その際、該当同族体の数が偶数で、それらの予測結果が陽性・陰性同数の場合は正解数を 0.5 とカウントした。その結果、全物質 911 種に対する予測結果は TP＝313.5、FP＝90.5、TN＝411.5、FN＝95.5 となり、正解率 79.6%という結果が得られた。

この正解率は、既存の予測モデルの最高精度よりはるかに高く、また、誤判定 FN の比率もかなり低く、動物実験代替の発ガン性予測手法としては十分な性能である。直列組み合わせ法の正解率 80.7%よりは若干劣るものの、頑健性の点ではこの並列モデルがはるかに優れている。今後、発ガン性のデータが既知の化学物質の数が増えて直列組み合わせ法の頑健性が向上する可能性はあるが、正解率と頑健性の両者を考慮すると、現状ではこの並列組み合わせ法が最適のモデルであると結論される。

謝　辞

本研究で協力いただいたた Bono Lučić、Dragan Amić、栗田多喜夫、西田健次、貝原巳樹雄、小野寺夏生の各氏に感謝します。

引用文献

[1] K.Tanabe, etal, *J. Comput. Chem. Jpn.*, **4**, 89 (2005).

[2] K.Tanabe, etal, *J. Comput. Chem. Jpn.*, **7**, 93 (2008).