

化学構造によるグルーピング及びリードアクロスによる神経毒性の *in silico* 予測

研究分担者：吉成浩一 静岡県立大学 教授

研究要旨

化学物質の発達神経毒性（DNT）を動物実験で評価するには多大なコストと時間を要し、また、近年は動物福祉の観点から安全性試験の動物実験代替法の開発が求められている。DNTについても動物実験代替法の開発が期待されているが、公的に利用可能なデータベースがないことなどの理由から研究が遅れている。そこで本研究では、DNT評価におけるインシリコ手法の有用性を提案するために分子記述子を活用したリードアクロス手法の開発並びにクラスタリング解析を行った。文献情報を元に構築した361物質のデータセットを利用して、分子記述子を活用した物質間距離に基づいて毒性予測のための参照物質（ソース物質）を選択し、データセット内の全物質についてリードアクロスを実施したところ、概ね0.6~0.65程度の一致率が認められ、特にDNTと統計学的に関連が認められた分子記述子を選択して使用することで予測精度が向上することが確認された。これは、記述子の選択により陰性物質間の距離が近くなることに起因していると考えられた。また、構造の小さな陽性物質、構造の大きな陰性物質の予測が困難であることが示唆された。クラスタリング解析においては、DNT陽性物質または陰性物質を多く含むクラスターを得ることができ、クラスタリングに基づくDNT予測の精度は、クラスターを形成しなかった物質を除くと0.7程度とかなり高かった。これらのグルーピング手法は、DNT評価において、化学構造に基づく初期スクリーニングに活用できる可能性がある。一方で、予測精度が十分とは言えないこと、また用いたデータセットの不確実性が考えられること、などの課題が見出されたことから、次年度以降はこれらの改善、解消を目指す。

A. 研究目的

化学物質の神経毒性や発達神経毒性（DNT）を動物実験で評価するには多大なコストと時間を要する。また、近年動物福祉の観点から、実験動物を用いた安全性試験の代替法の開発が求められている。しかし、DNTに関しては、公的に利用可能な毒性試験データベースがないことなどの理由から研究が遅れている。そこで我々は、DNT評価におけるインシリコ手法の有用性を提案するために、既存情報からDNT予測のためのデータセットを構築し、DNTを示す物質のグルーピング解析やリードアクロス手法によるDNT評価手法の開発に向けた基礎検討を行ってきた。これまでに、化学構造情報を数値化した分子記述子を利用して化学物質の類似性を評価し、その結果に基づいたグルーピングやリードアクロスを行い、使用する分子記述子の種類、毒性予測に利用する類似物質の数などの検討を進めてきた。今年度は、これまでの成果の再現性の解析を行うとともに、予測結果の具体的な内容を解析し、今後の精度向上のための課題を明らかにすることを目的として研究を行った。

B. 研究方法

1. データセット

文献 (*Neurotoxicol Teratol*, 52: 25-35, 2015) から DNT を示す物質及び DNT 情報がない物質の情報を抽出し、前者を DNT 陽性物質 (164 物質)、後者を DNT 陰性物質 (197 物質) として使用した。データセットの DNT 陽性率は 0.454 である。被験物質の化学構造情報は PubChem から収集した。

2. 分子記述子

分子記述子は alvaDesc ソフトウェア (Alvascience, ver 2.0) で計算し、搭載されている 33 ブロックに分類された 5,666 記述子のうち、化学構造との関連を理

解しやすい 6 ブロック (Constitutional indices, Ring descriptors, Functional group counts, Atom-centred fragments, Molecular properties) に含まれ、被験物質で計算可能であった 376 記述子 (セット A: 6b all) を解析に利用した。

376 種の分子記述子から、次の方法でさらに選択した。セット B~D: セット A の分子記述子のうち、相関係数の閾値を、0.99、0.95 または 0.90 に設定し、相関する分子記述子ペアの片方を削除したセットを、それぞれセット B ($r < 0.99$, 311 種)、セット C ($r < 0.95$, 288 種)、セット D ($r < 0.90$, 261 種) とした。

セット E~G: 376 種の分子記述子について、DNT 陽性物質と陰性物質の間で Welch の t 検定を行い、得られた p 値が 0.2、0.1、0.05 であった分子記述子のセットを、それぞれセット E ($p < 0.2$, 173 種)、セット D ($p < 0.1$, 129 種)、セット G ($p < 0.05$, 81 種) とした。

3. リードアクロス

各分子記述子について、最大値が 1、最小値が 0 となるように標準化したのち、物質間のユークリッド距離を R を利用して計算した。物質間距離は理論上の最大距離 (使用した分子記述子数の平方根) で除して相対距離とした。

本解析では、データセット内の全物質を被験物質として解析に用いた。リードアクロスと予測精度の算出の概略は以下の通りである。361 物質それぞれについて、残りの 360 物質を距離が近い順に並べ、そこから最近傍の 1 物質、または 4、6、8、10、12、14、16 物質を選択し、それらを毒性予測に利用する「ソース物質」とした。ソース物質に占める DNT 陽性物質の割合がデータセットの陽性率 (0.454) よりも大きかった場合、被験物質を陽性と判定し、小さかった場合は陰性と判定した。これらの予測を全 361 物質で実施した。得られた 361 物質の結果を基に、感度 (陽性物質を正しく陽性と判定できた割合)、特異度 (陰性物質を正しく陰

性と判定できた割合)、一致率(全物質の正答率)、ならびに balanced accuracy (BA、感度と特異度の平均値)を算出した。

4. クラスタリング解析

JMP Pro 15 (SAS Institute) を使用し、Ward 法により実施した。

(倫理面の配慮) 本研究では動物実験データを利用した解析を行うため、倫理的配慮を必要とする情報は含まれない。

C. 研究結果

1. リードアクロスを利用した解析

分子記述子を利用したリードアクロスに関する当研究室のこれまでの研究結果から、DNTと統計学に関連する分子記述子を選択して使用することで予測精度が向上することが明らかになっている。これまでの研究では毒性予測に利用する物質(ソース物質)数を奇数個として検討してきたが、再現性解析の過程で、偶数個の選択時には判定結果が異なる事象が認められた。そこで、これまでの解析結果の再現性確認も兼ねて、偶数個のソース物質を利用した解析を行った。

(解析1) 研究方法欄に示したAからGの7つの分子記述子セットを利用して全物質間の相対距離をそれぞれ算出し、各物質について最も近い物質から1、4、6、8、10、12、14または16番目の物質までをソース物質とした。ソース物質のDNT情報から各物質の毒性判定を行う際の閾値は、データセットのDNT陽性率である0.454とした。例えば、近傍物質数が10の条件で陽性物質が5物質であった場合、陽性率は0.5であり、閾値である0.454を超えるため、被験物質を陽性と判定した。本解析では、データセット内の全361物質全てを被験物質として毒性判定を行い、それらの結果に基づいて感度、特異度、一致率およびBAを算出し、その結果をFig. 1に示した。分子記述子を使用せずにランダムに物質を選択した場合の感度と特異度は、データセットの陽性率である0.454と1から陽性率を引いた0.546程度となる。

全分子記述子 (Fig. 1A) を用いた場合、10~14物質をソース物質とした際に、感度と特異度の差が小さく、比較的精度の高い予測結果が得られた。相関係数に基づいて関連性の高い分子記述子を削減した場合は、BAに大きな変化は認められなかったが、削減数が増えると感度と特異度の差が大きくなり、BAが低下するケースも認められた (Fig. 1B~1D)。DNTとの関連性により分子記述子を選択した場合は (Fig. 1E~1G)、8または10物質をソース物質とした際に、0.65を超えるBAが得られ、これはセットAからEで得られた値よりも高かった。ソース物質数を増やすと、特異度は上昇したが、感度が著しく低下した。

以上の結果より、毒性学的に関連性の高い分子記述子を選択して使用することで、リードアクロスの精度が向上することが再現された。また、ソース物質数は10物質程度とすることで、感度と特異度のバランスがよい精度が得られることが示された。

(解析2) ソース物質数を固定した解析では、被験物質と近傍物質の距離を考慮しないため、データセット内に被験物質と類似性が高い物質が存在しない場合は、それほど類似性が高くない物質もソース物質に含まれてしまう。類似物質が存在しないそのような物質に関しては、リードアクロスにより毒性判定を行うことは不適切であるとも考えられる。そこで次に、一定距離内にソース物質が存在しない物質を解析対象外として予測精度を算出した。具体的には、上記解析と同様に近傍10物質をソース物質として選択し、最も遠い10番

目の物質が設定した閾値内に含まれる場合にのみリードアクロスによる評価を行い、閾値内に含まれない場合には解析対象外として精度の算出から除外した。本研究では理論上の最大距離を1として相対距離を算出していることから、物質間の距離は0~1の値をとる。詳細は割愛するが、物質間距離の分布を確認し、本解析での閾値を0.15~0.2の間で変化させて解析した。

全376分子記述子を利用した場合 (Fig. 2A)、閾値を設定することでBAは0.5程度上昇した。閾値を狭くすることで10種のソース物質が閾値内に含まれない解析対象外物質数は増えたが、特異度の上昇は認められたが、感度が低下し、BAや一致率に大きな変化は認められなかった。相関性に基づいて削除した分子記述子セットを用いた場合においても、全分子記述子を用いた場合とほぼ同様の精度の変化が認められた (Fig. 2B~2D)。一方、DNTと関連がある分子記述子セットを利用した場合には、「 $p < 0.2$ 」セットで閾値の設定により特異度が上昇し、その結果BAも0.2程度上昇したが、その他のセットでは顕著な上昇は認められなかった。

以上の結果から、閾値の設定の効果が認められる分子記述子セットは、解析1とは異なることが明らかになった。

(解析3) 機械学習等を利用した毒性予測においては、感度と特異度のバランスのよい予測モデルが良いとされ、本研究でもBAや、感度と特異度のバランスを指標として予測精度を判断してきた。しかし、DNTの毒性学的な重要性を考えると、偽陰性が少ない予測手法が求められる可能性がある。そこで次に、ソース物質数を10物質で固定し、判定基準をこれまでの解析で利用した「データセットの陽性率 (0.454) 以上で陽性」から、「10物質中 n 物質の陽性物質が存在すれば陽性」という基準に変更し、 n を1から10まで1ずつ変化させて予測精度を算出した。結果をFig. 3に示した。

解析1で最も良い精度が認められたセットG ($p < 0.05$) の分子記述子を用いた場合での結果を見ると、ソース物質10物質中1、2または3物質以上が存在した場合を陽性と判断する条件において、感度がそれぞれ0.98、0.96、0.90と高い値を示した。このときの特異度はそれぞれ0.06、0.20、0.32だった。

(解析4) 予測結果の詳細を確認すると共に、分子記述子セットG ($p < 0.05$) の選択による予測精度向上の理由を解析することを目的として、データセットのDNT陽性物質、陰性物質それぞれについて、セットAおよびGの分子記述子を用いた場合のソース物質10物質中の陽性物質数のヒストグラムを作成した (Fig. 4)。その結果、多少のゆがみはあるが、正規分布曲線に近い形状のヒストグラムが得られた。

DNT陽性物質について見ると、セットA (6b all) とセットG ($p < 0.05$) では、感度はともに0.63程度と変わらないことから、ヒストグラムのパターンにも大きな違いは認められなかった。ただし、陽性物質が7物質以上であった物質数は、セットGではセットAの37物質から42物質に増加していた。予測結果が一致しなかった物質については、いずれの記述子セットにおいても陽性物質数が3または4である物質が多かった。

DNT陰性物質について見ると、セットAに比べてセットGの分子記述子を用いた場合の特異度が0.60から0.69に大きく上昇したことと一致して、グラフが全体的に左側にシフトしていた。特に、ソース物質10物質中、陽性物質が1物質以下しか含まれない被験物質数は、15 (セットA) から40 (セットG) に大きく増加した。

予測が難しい物質の特徴を明らかにするために、セットG ($p < 0.05$) の分子記述子を用いた場合に、ソース

物質に陽性物質が0または1物質しか含まれなかった陽性物質（ヒストグラムにa, bで示した各3物質）ならびに陽性物質が8物質含まれた陰性物質（ヒストグラムにcで示した8物質）を抽出し、それらの化学構造を確認した（Fig. 4下）。陽性物質では比較的単純な構造を有する物質がほとんどであったのに対し、陰性物質では構造が複雑で分子量の大きな物質が多く含まれていた。

以上の結果より、本研究で用いた手法において、少なくともソース物質を10物質とした場合には、DNT陽性物質、陰性物質共に大きく予測が外れている物質はそれほど多くないことが明らかになった。さらに、DNTと関連した分子記述子を選択することで、陽性物質同士の距離を縮める効果はあまり認められないが、陰性物質と陽性物質の距離が離れ、陰性物質同士の距離が近くなると考えられた。また、化学構造が比較的小さい、または大きい物質の予測が困難であることが示された。

（解析5）解析4で示唆された「DNTと関連した分子記述子を選択することで陰性物質同士の距離が近くなる」という仮説を検証するために、データセット中の各物質について、各記述子セットを用いた場合のソース物質10物質中における毒性（DNT）が一致するソース物質との距離の平均値を算出した。そして、全DNT陽性物質および陰性物質について、平均距離の分布を比較するために箱ひげ図を作成した（Fig. 5）。

その結果、陽性物質同士の距離ならびに陰性物質同士の距離は、セットAの分子記述子を使用した場合に比べて、セットD~Gの分子記述子を使用した場合でともに減少したが、その程度は陰性物質間の距離でより大きく、解析4の結果を支持していた。

2. クラスタリングを利用した解析

研究方法欄に示したAからGの7つの分子記述子セットを使用して、Ward法による階層的クラスタリングを実施し、全物質を物質数の約十分の一である36クラスターに分類した。セットAの分子記述子を使用して得られたデンドログラムをFig. 6に示した。5物質以上が含まれていたクラスターに着目すると、クラスター12（4/5）、15（12/14）、24（6/7）、30（6/9）に陽性物質が多く集まっていた（カッコ内は陽性物質数/クラスター内物質数）。一方、クラスター6（7/26）、9（6/22）、29（20/61）、32（3/12）、34（10/34）には陰性物質が多く集められていた。

次に、各分子記述子セットを用いたクラスタリング結果について、2物質以上が含まれたクラスターを対象として、クラスターの陽性率がデータセットの陽性率より高い場合にそのクラスター内の物質を陽性と判定、低い場合にはそのクラスター内の物質を陰性と判定し、感度、特異度、一致率、BAを算出した（Fig. 6）。その結果、セットAの分子記述子を使用した場合の一致率とBAはともに0.67程度であり、セットB~Dの分子記述子を使用した場合にも大きな変化は認められなかった。一方、DNTとの関連で選択したセットE~Gの分子記述子を利用した場合には、一致率およびBAは0.7以上に上昇した。これらの結果から、クラスタリング解析においても、DNTと関連する分子記述子を利用することで、DNTの有無がより一致する物質をグループ化できると考えられた。

D. 考察

1. リードアクロスを利用した解析

リードアクロスによる毒性評価においては、毒性試験結果がある物質をソース物質として選択して被験物

質の毒性を予測する。そのため、適切なソース物質の選択が高い予測精度を得るための鍵となる。解析1の結果から、分子記述子を利用したリードアクロスによるDNT評価においては、統計学的にDNTと関連する分子記述子を選択して使用することで、より高い予測精度が得られることが明らかになった。これは、使用する分子記述子計算ソフトウェアのバージョンならびにソース物質の選択方法が異なる当研究室の以前の解析結果と一致しており、得られた結果は妥当かつ普遍的であると考えられた。

一方で、このときのBAは高くても0.65程度であり、予測精度としては十分とは言えない。その理由として、本研究ではデータセットを固定し、その中の全物質を解析対象（被験物質）としていることから、必ずしもデータセット内に構造的に類似する物質が存在しないために予測が外れる物質が存在する可能性が考えられた。そこで解析2では、ソース物質の選択に条件（距離の閾値）を設定して解析した。その結果、全分子記述子を用いた場合や相関係数で選択した記述子セットを用いた場合には、閾値の設定による予測精度の向上が認められた。しかし、解析1で良い精度を示したDNTと関連する分子記述子セットを使用した場合にはそのような効果は認められなかった。これは、事前に適切な分子記述子を選択していない場合には、閾値を設定することで不適切なソース物質が選択される被験物質が解析から除外され、結果的に予測精度の向上が認められたが、DNTと関連する分子記述子を使用した場合には、その選択に伴いソース物質の選択が比較的適切に行われたため、閾値の設定による解析に不適切な被験物質の除外効果が認められなかったのではないかと考えられた。本研究では、近傍第10位の物質に対して閾値を設定して解析対象・対象外を決定したが、この仮説をさらに検証するために、閾値の判定を行う近傍物質を変化させて同様の解析を行う予定である。

インシリコ手法による評価は、コンピュータ上での評価が可能であり、スループットに優れていることから、体系的なDNT評価における初期スクリーニングに有用と考えられる。そのような評価では高い感度が必要であると推察されることから、上記解析で決定した条件（分子記述子セット、ソース物質数）において、ソース物質に基づく判定条件を変化させて、感度を高めた際の全体の予測精度を確認したところ、10物質中3物質が陽性の場合に陽性と判定する条件において、感度が0.90と高い判定精度が得られた。一方、この条件における特異度は0.32、一致率は0.59であり、初期スクリーニングに利用した場合、陰性物質の約7割を偽陽性としてしまうと考えられた。

最後に、予測結果の詳細を確認するために、DNT陽性物質と陰性物質について、それぞれにおけるソース物質中の陽性物質数をヒストグラムとして可視化したところ、一致率が20%未満などの予測がひどく外れている物質数が多いわけではないこと、またDNT関連記述子の選択は、陰性物質同士の距離の短縮に寄与していることが明らかになった。さらに、構造が小さい陽性物質、大きい陰性物質の予測が難しいことが示された。おそらく、データセット内に類似物質が少なく、ソース物質の選択時に適切な参照物質が存在しなかったことが原因と考えられる。今後は予測精度が非常に低かった物質について、どのような物質がソース物質として選択されているのかなど、個別のリードアクロス結果を確認し、得られた情報を元に、予測精度の向上に向けた取り組みを進めていく予定である。また、本研究で用いているデータセットは、2015年の論文を元に構築

していることから、その後の情報更新の確認作業や、論文における陽性、陰性判定根拠の詳細などについて確認し、データセットの信頼性向上も図る予定である。

2. クラスタリングを利用した解析

典型的なグループ化手法である階層的クラスタリングによりデータセット内の物質をグループ化し、化学構造とDNT情報との関連性を解析した。その結果、陽性物質が多いグループや陰性物質が多いグループをいくつか作成することができた。また、クラスタリングに使用する分子記述子の種類の影響を確認するために、種々の方法で選択した分子記述子セットを使用してクラスタリングを実施し、クラスター精度の判定を行ったところ、DNTと関連する分子記述子を使用した場合にクラスタリングの精度が向上すると考えられる結果が得られた。今後、各クラスターに含まれる物質の詳細を確認し、化学構造との関連性を確認するとともに、リードアクロス解析のところでも記載したように、データセットの信頼性確認を行う必要があると考えている。

一方、階層的クラスタリングを実施したところ、1物質や2物質程度しか含まれないクラスターも多く認められた。これは、これらクラスターに含まれる物質に類似する物質がデータセットに含まれていないことを意味しており、今後リードアクロスにおいて予測が適切に行われなかった物質との一致性を確認する予定である。そのような物質は、現データセットを用いる限り、化学構造からのDNT予測が難しい物質と考えられ、さらなるデータ拡充に向けた参考情報として活用したい。

E. 結論

化学構造情報を利用したリードアクロス手法の確立に向けて基礎検討を行った結果、分子記述子を利用した類似性評価、すなわちリードアクロスにおけるソース物質の選択においては、DNTと関連する分子記述子を選択して使用することで、予測精度が向上することが

示唆された。また、階層的クラスタリングを利用した化学物質のグループ化においても、DNTと関連した分子記述子の使用がよりよいグループ化を可能にする可能性が示された。

F. 研究発表

1. 論文発表

なし

2. 学会発表

- 1) 大村奈央、志津怜太、保坂卓臣、菅野裕一朗、吉成浩一：発達神経毒性を予測するためのリードアクロス手法の検討、日本動物実験代替法学会第35回大会、2022年11月18日～20日、静岡市
- 2) 吉成浩一：化学構造情報を用いたリードアクロスによる発達神経毒性の評価、日本動物実験代替法学会第35回大会、2022年11月18日～20日、静岡市
- 3) 大村奈央、志津怜太、保坂卓臣、菅野裕一朗、吉成浩一：リードアクロスによる発達神経毒性評価手法の開発：分子記述子を用いた類似物質選択の有効性、第49回日本毒性学会学術年会、2022年6月29日～7月2日、札幌市
- 4) 吉成浩一：化学構造情報を利用したリードアクロスによる化学物質の毒性予測、第49回日本毒性学会学術年会、2022年6月29日～7月2日、札幌市

G. 知的所有権の取得状況

1. 特許取得

なし

2. 実用新案登録

なし

3. その他

なし

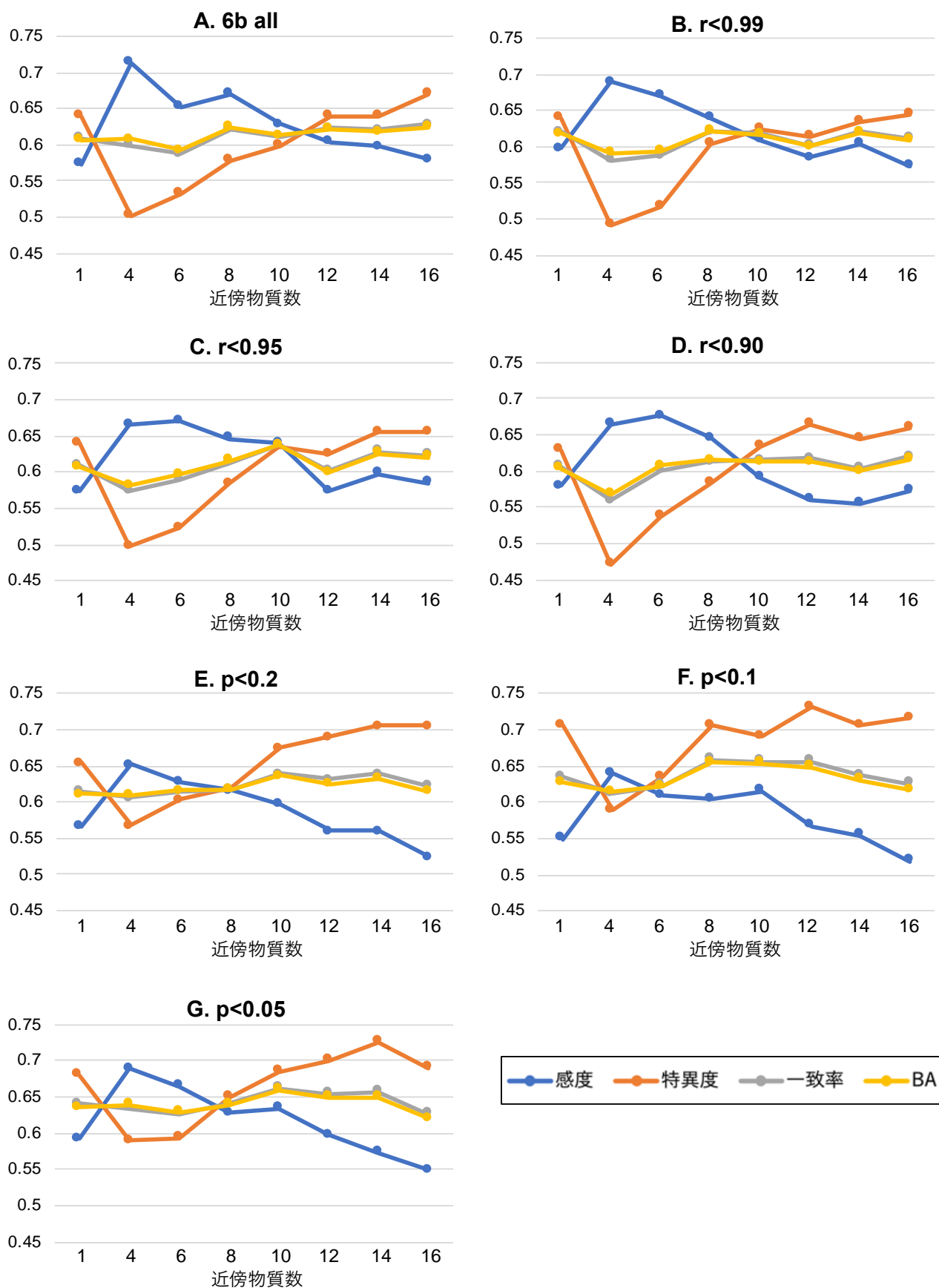


Fig. 1. 分子記述子を利用したリードアクロスにおける分子記述子セットとソース物質数の検討
 各分子記述子セットを利用して物質間距離を計算し、全物質について横軸で示した近傍物質数をソース物質としてリードアクロスを実施した。分子記述子セットごとの予測精度を示した。



Fig. 2. 分子記述子を利用したリードアクロスにおけるソース物質の物質間距離の閾値の検討
 10個のソース物質を利用したリードアクロスを実施する際に、10番目のソース物質が示した閾値内に含まれない場合には解析対象外として、予測精度を算出した。カッコ内の緑の数字は解析対象物質数を示す。

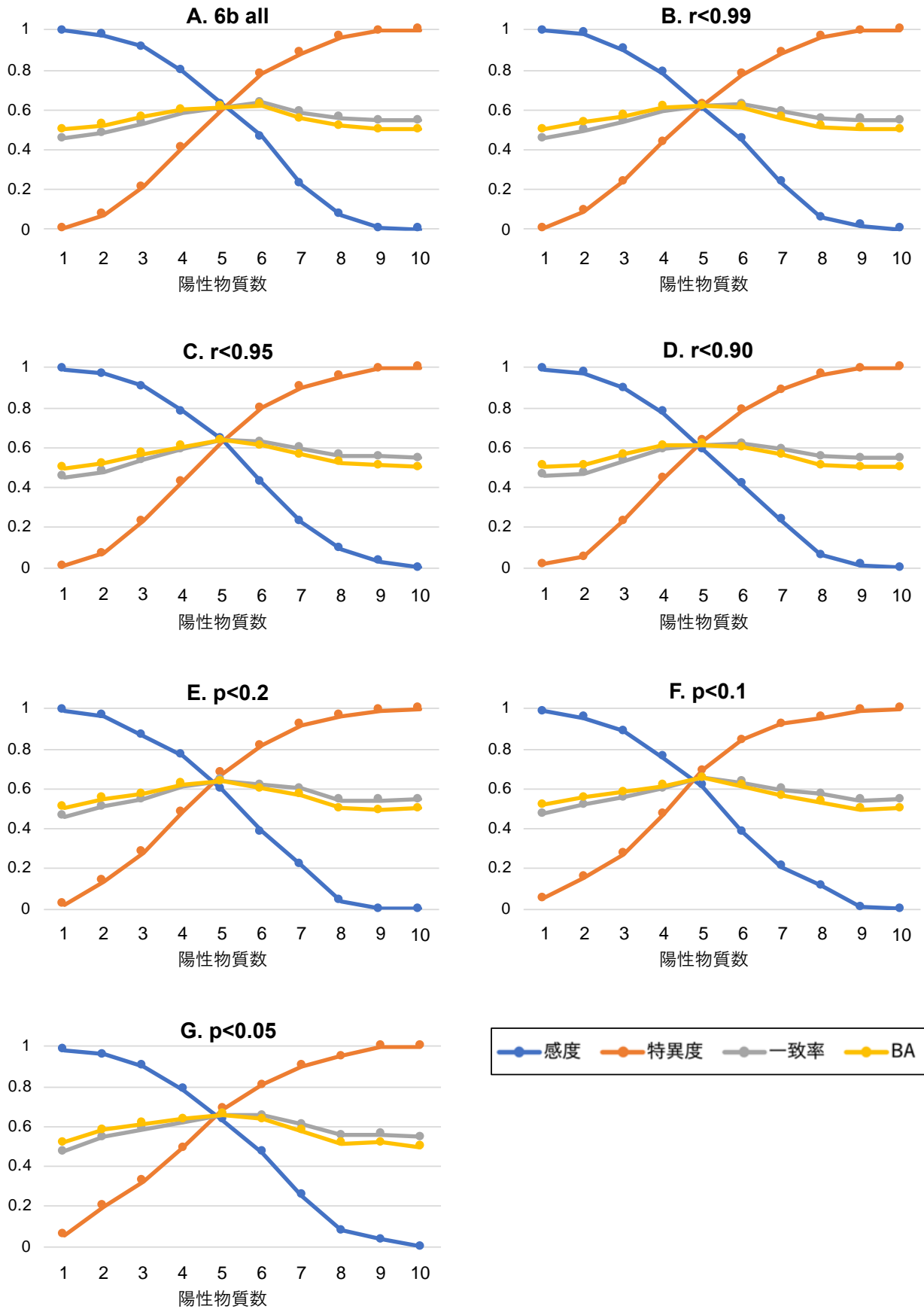


Fig. 3. 10個のソース物質を利用したリードアクロスの性能評価
 10個のソース物質を利用したリードアクロスを実施する際に、陽性判定の際に必要な物質数を変化させて予測精度を算出した。

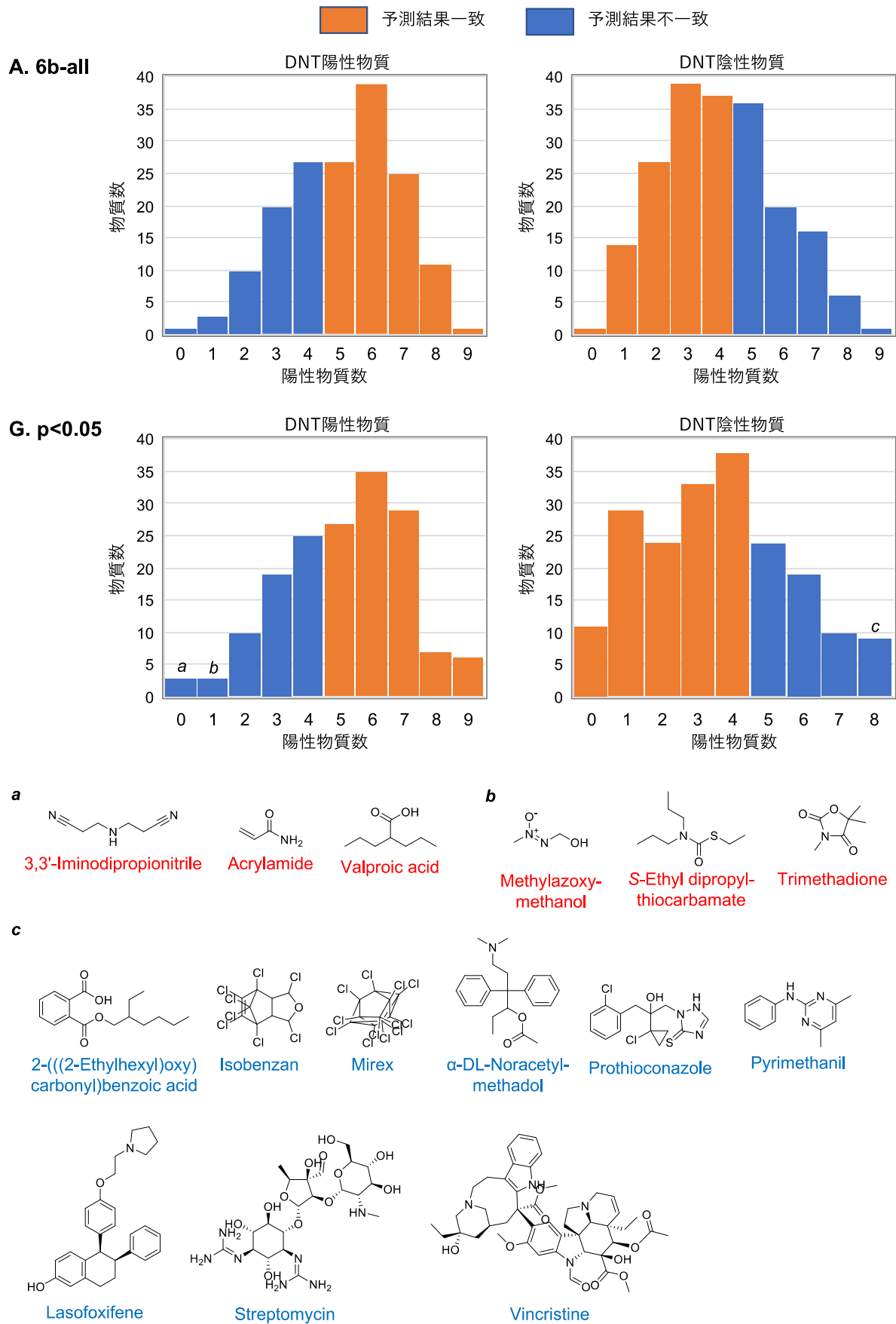
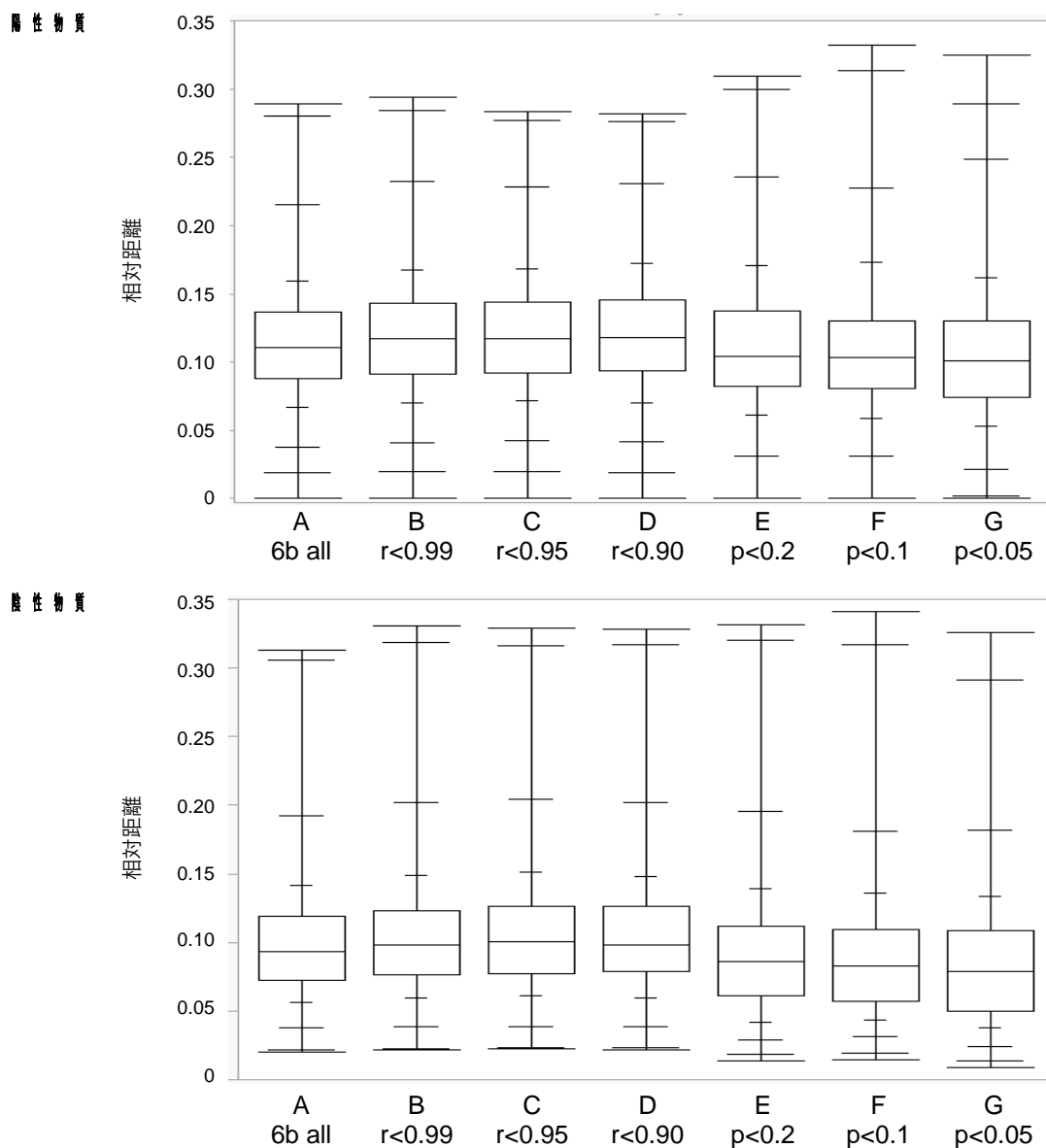


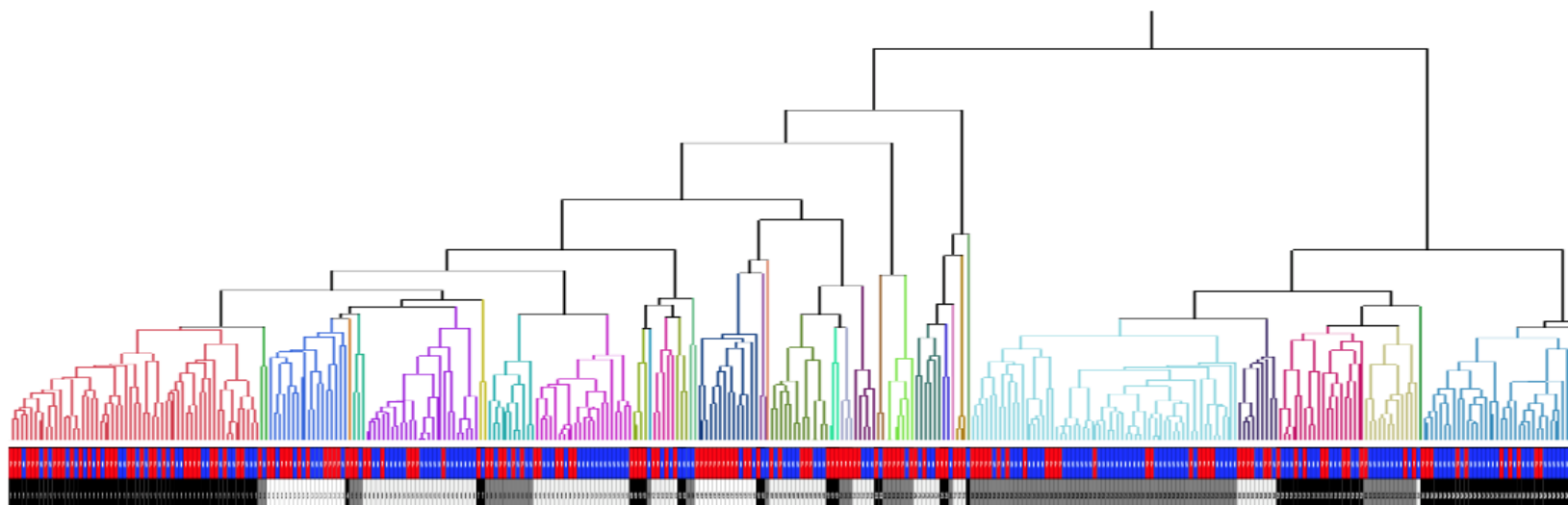
Fig. 4. 10個のソース物質を用いたリードアクロスによる予測結果の特徴解析

分子記述子セットAおよびGを用いたリードアクロスにおいて、各被験物質のソース物質10物質中の陽性物質数を、DNT陽性物質、陰性物質ごとにヒストグラムで示した。予測精度が著しく低かった3区分 (a, b, c) に含まれる化学物質の構造を下に示した。



		A. 6b all	B. r<0.99	C. r<0.95	D. r<0.90	E. p<0.2	F. p<0.1	G. p<0.05
陽性物質	平均	0.1135	0.1198	0.1203	0.1215	0.1117	0.1100	0.1064
	中央値	0.1108	0.1169	0.1174	0.1182	0.1043	0.1033	0.1007
陰性物質	平均	0.0984	0.1031	0.1050	0.1044	0.0907	0.0881	0.0846
	中央値	0.0933	0.0984	0.1005	0.0985	0.0857	0.0830	0.0787

Fig. 5. 10個のソース物質を用いたリードアクロスにおける被験物質とソース物質の物質間距離
 各分子記述子セットを用いたリードアクロスにおいて、DNT陽性物質、陰性物質ごとに、10個のソース物質のうち、それぞれ陽性物質、陰性物質のみの距離の平均を算出し、その分布を箱ひげ図で示した。表には平均値と中央値を示した。



クラスター	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
物質数	57	2	18	1	3	26	2	11	22	4	1	5	3	2	14	2	1	13	3	3	5	2	7	6	2	1	3	1	61	9	20	12	1	34	1	3
陽性物質数	33	1	11	1	2	7	1	6	6	4	0	4	0	0	12	1	0	5	3	3	2	1	6	3	2	0	3	0	20	6	8	3	0	10	0	0
陽性率	0.579	0.500	0.611	1.000	0.667	0.269	0.500	0.545	0.273	1.000	0.000	0.800	0.000	0.000	0.857	0.500	0.000	0.385	1.000	1.000	0.400	0.500	0.857	0.500	1.000	0.000	1.000	0.000	0.328	0.667	0.400	0.250	0.000	0.294	0.000	0.000

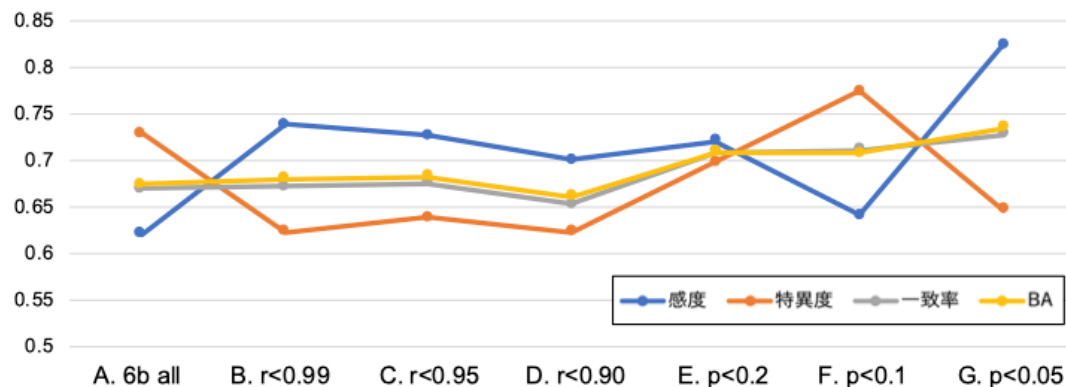


Fig. 6. 階層的クラスタリングによるデータセット内物質のグルーピング

(上) 各分子記述子セットを用いた階層的クラスタリング (Ward法) を実施し、セットAを使用した場合のデンドログラムを示した。デンドログラム下の図では、赤がDNT陽性物質、青が陰性物質を示し、クラスターを3段階のグレースケールで示した。各クラスターに含まれる物質数、陽性物質数を表で示した。

(下) 2物質以上が含まれたクラスターについて、データセットの陽性率を基準としてDNT陽性、陰性の判定を行い、その際の精度を示した。