

## 〈特集〉

変異原性試験に関する新しい動き  
—— 機械学習による予測と NGS による直接測定 ——松田 知成<sup>1)</sup><sup>1)</sup> 京都大学大学院 工学研究科附属 流域圏総合環境質研究センター  
(〒520-0811 滋賀県大津市由美浜 1-2 E-mail: matsuda.tomonari.8z@kyoto-u.ac.jp)

## 概要

新規化学物質の審査において、変異原性試験は事実上必須項目である。ICH M7 ガイドラインが施行され、医薬品に含まれる不純物に関してはコンピューターによる変異原性の予測を実際の Ames 試験の代わりに使用することが解禁された。これにより、変異原性に関する構造活性相関が改めて注目されている。様々な機械学習アルゴリズムが利用され、それなりに使えるものになってきている。一方、次世代 DNA シーケンサーの登場により、DNA 中の突然変異を直接測定できるようになってきた。極低頻度の突然変異を検出するためには、超正確なシーケンス技術が必要となり、duplex sequencing 法や SMRT 法が開発された。今後、これら正確な変異検出技術と機械学習の進歩のスパイラルを作ることにより、変異原性予測がどんどん正確になっていくことが期待される。

キーワード：変異原性試験，Ames 試験，次世代 DNA シーケンサー，機械学習，QSAR

原稿受付 2019.2.1

EICA: 23(4) 32-35

## 1. はじめに

化学物質の安全性評価は、有害な化学物質から人間の健康や生態系を保全するうえで重要である。主要国では 1970 年代から相次いで化学物質の規制に関する法律が制定された。日本では化審法が 1973 年に、米国では TSCA (Toxic Substances Control Act) が 1976 年に制定されている。これら法律が施行される前に流通していた化学物質を「既存化学物質」、施行後に流通している、または流通しようとしている化学物質を「新規化学物質」と呼んでいる。欧州では、1981 年以前に流通していた化学物質を「既存化学物質」としている。既存物質の総数は、日本の化審法では 19,246 物質<sup>1)</sup>、米国 TSCA では約 8 万物質あり、EU も同程度あると言われている。

新規化学物質の安全性評価は、その開発企業が担うことになる。これは、規制当局から認可を得るのに必要である。一方、日本や米国においては、既存化学物質の安全性評価は規制当局が主として担っている。しかし、予算の都合もあり、既存化学物質の安全性評価は遅々として進んでいないのが現状である。EU では 2007 年に新しい化学物質管理に関する法律 REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) が制定され、既存物質についても、EU 域内で一定量以上流通する場合には、業者に安全性評価が義務付けられ、今まで野放しであった既存化学物質の安全性評価が進むことが期待されている。

変異原性は発がん性と密接な関係があり、化学物質の安全性評価においては事実上の必須項目である。従って、これをより簡便に、より信頼性高く評価する手法が求められている。変異原試験のうち最も広く用いられているのは Ames 試験である。これは、サルモネラ菌を用いた復帰突然変異試験であり、使用する菌株はヒスチジン合成酵素の遺伝子に点突然変異を持っていて、ヒスチジンを含まない培地ではコロニーを作らない。変異原処理により、ヒスチジン合成酵素の遺伝子に復帰突然変異が生じると、ヒスチジンを合成できるようになり、ヒスチジンを含まない培地でもコロニーを作ることが可能になる (復帰コロニー)。この復帰コロニーをカウントすることで、突然変異を検出する方法である。

## 2. 機械学習を使った変異原性予測の動向

Ames 試験自体はそれほど難しいものではないが、構造活性相関を用いて、化学構造から変異原性を直接予測することに対するニーズは多い。例えば製薬において、ほとんどの企業は候補物質の絞り込みにコンピューターソフトウェアを使った *in silico* 予測を取り入れている。現時点では、最終的な製品については実験で確かめなくてはならないが、不純物の安全性評価には、計算による変異原性予測を認める動きがある。特に最近、日米 EU 医薬品規制調和国際会議 (ICH) の M7 「潜在的発がんリスクを低減するための医薬品

中 DNA 反応性 (変異原性) 不純物の評価及び管理」のガイドライン<sup>2)</sup>が施行されたが、この中では安全性評価にコンピューターを使った変異原性予測を取り入れる際の指針が明記されている。

## 2.1 ICH M7 ガイドラインの概要

M7 ガイドラインは、医薬品の原薬及び製剤中の DNA 反応性不純物のリスクマネジメントに関わる指針である。この不純物とは医薬品の製造過程に用いられる出発物質、出発物質から原薬に至る合成ルート上の試薬、及び中間体の他、分解生成物なども含まれる。この時予測される不純物の変異原性が明らかでない場合は、細菌を用いる変異原性試験を実施するか、構造活性相関による評価を実施するべきであるとしており、コンピューター予測で実際の試験を代替できることを示している。これは、一つの製剤に対して、潜在的な不純物は多数あり、これら化学物質すべてを実際に準備して、変異原性試験を実施するのは困難であるためである。不純物が既知の変異原性物質である場合や、または、変異原性試験や、構造活性相関の実施により変異原性ありと判定された場合は、許容摂取量の考え方で製品の安全性を管理する。許容摂取量は臨床試験時には発がんリスクが 100 万人に 1 人のレベルで管理し、製品になった際は、10 万人に 1 人のレベルで管理するとしている。臨床試験時の方が基準が厳しくなっている理由は臨床試験時には薬の明らかなベネフィットが不明なためである。

## 2.2 Ames 試験の結果を予測する構造活性相関法

M7 ガイドラインでは、Ames 試験の結果をコンピューターで予測する際には、互いに相補的な 2 種類の構造活性相関法を適用すべきであるとしている。一つは、専門的経験に基づくルールベースの方法、二つ目は統計ベースの方法とする。これら相補的な二つの構造活性相関法において両方陰性であれば、その不純物には変異原性に関する懸念がないと結論できるとしている。

経験に基づくルールベースシステムとしては、英国ラーサ社が開発・販売する DEREK、オープンソースの ToxTree や OECD (Q)SAR Toolbox などがある。一方、統計ベースのシステム、すなわち機械学習を用いたシステムとしては、MultiCASE 社が販売する MCASE などがある。

しかし、最近は様々な機械学習のアプリケーションが簡単に利用できるようになってきたので、データセットさえあれば大学の研究室レベルで Ames 試験の予測器を構築することが可能である。Hansen らは、Ames 試験に関する公開されているデータベースや文献を整理し、6512 物質からなるデータセットを構築

し、DragonX というソフトウェアで生成した分子記述子を説明変数とした機械学習を行った。サポートベクターマシン、ガウシアンプロセス、ランダムフォレスト、k 近傍法、の 4 種類の機械学習を試したが、いずれも DEREK や MCASE の予測性能を凌駕していた<sup>3)</sup>。また、Xu らも同様に 7617 物質からなるデータセットを構築し、PaDEL descriptor というソフトウェアで分子記述子を生成し、人工ニューラルネットワーク、k 近傍法、サポートベクターマシン、ナイーブベイズ法、決定木、という 5 種類の機械学習アルゴリズムを行ったところ、サポートベクターマシンと k 近傍法で良い成績を収めている<sup>4)</sup>。

## 3. NGS による突然変異の直接測定

機械学習で良い成果を収めるためには、良いデータをそろえる必要があるが、Ames 試験の研究室間の再現性は 85% しかないという報告もあり、より再現性の高い方法が求められている。また、Ames 法は復帰変異を検出する方法なので限られた種類の配列に生じる変異しか検出することはできない。そのため、GC 塩基対の変異を検出する株と AT 塩基対の変異を検出する株の両方を用いる必要があり煩雑である。

それら問題を解決するため、次世代 DNA シーケンサー (NGS) で直接突然変異を検出する方法が提案されている。以下にそれらを紹介する。

### 3.1 コロニー NGS 法<sup>5)</sup>

変異原で処理したバクテリアをしばらく培養し、変異固定後に寒天培地にまき、コロニーを形成させ、各コロニーから DNA を精製して NGS で全ゲノム解析を行う。突然変異はバイオインフォマティクスを用いて *in silico* で検出する。我々は市販のソフトウェア (CLC genomics work bench) を用いて解析しているが、フリーのツールもいろいろ利用できる。

コロニー NGS 法は非常にシンプルな方法であるが、多数のコロニーについて解析しなければ、低頻度の変異上昇を検出しにくいという問題点がある。また、コロニー単離ができなければ使えないという欠点がある。

### 3.2 Duplex-sequencing 法<sup>6)</sup>

Duplex-sequencing 法は 2012 年にワシントン大学の L. A. Loeb の研究室で開発された。仮にエラーを全く起こさない理想的なシーケンサーがあればどんな低頻度の突然変異も検出できるであろう。この方法は工夫によって超正確なシーケンスを実現し、極低頻度の突然変異を検出するものである。

シーケンスエラーは様々な原因で生じる。例えば機械の読み取りミスもあれば、シーケンスライブラリー

作成時の PCR 増幅過程で生じる塩基置換もある。特にサンプル DNA に損傷がある場合は高頻度でエラーが生じる。これらによって生じたシーケンスエラーは、どんなにクオリティーの高いデータを使っても、1000~10000 塩基に 1 個程度生じてしまう。一方、自然突然変異は 1 億~10 億塩基に数個程度なので、変異原試験として使う場合は、エラー率をこれ以下にする必要がある。

そこでいかにして超正確なシーケンスを実現するかであるが、まず、サンプル DNA を超音波等で断片化し、それらに分子固有のタグ配列をライゲーションさせる。こうすることで、その後 PCR 増幅してシーケンスライブラリーを作成しても、タグ配列を読むことにより、共通の祖先配列によって生じたリードを仕分けすることができる。さらに、タグ配列を読むと、そのリードが DNA 2 本鎖のどちらの鎖由来かも解析できる。このようにして、一つの DNA 断片から派生したリードが、DNA 2 本鎖のそれぞれの鎖由来のものが数リードずつ得られるようにライブラリーを調整する。それらをまず、鎖ごとにコンセンサス配列を取ることによって、単純な機械の読み取りミスをはばゼロにすることができる。続けて、DNA 2 本鎖の各鎖のコンセンサス配列が一致したものだけが、正しく読めた配列であると判定する。こうすることにより、自然突然変異レベルを十分に検出できるレベルまで、シーケンスエラーを低減することができる。

Duplex-sequencing 法の良い点は、DNA を抽出してきて Duplex-sequencing するだけで突然変異の定量が可能であり、どのようなサンプルにも応用できる点である。Ames 試験だけでなく、哺乳類培養細胞にも使えるし、実験動物の臓器、人間の臨床検体にも使用可能である。

### 3.3 SMRT 法<sup>7)</sup>

筆者もまた、超正確なシーケンス技術により、任意の DNA に生じている極低頻度の突然変異を検出する技術を構想し、2015 年に SMRT 法を発表した。この方法は、PacBio 社の一分子リアルタイムシーケンサー (Single Molecule Real Time; SMRT) を用いて行う。

SMRT シーケンサーでは、ナノメートルサイズに仕切られた小区画に DNA ポリメラーゼが 1 分子入るようになっており、その中で実際に DNA を合成し、このとき取り込まれるヌクレオチドをリアルタイムで検出する。従来の NGS よりも正確性が劣るという欠点があるが、ロングリードが得られるという利点もある。

この SMRT シーケンサーを用いて極低頻度の突然変異を検出する際には、SMRT Bell ライブラリーというものを調整して行う。これは、サンプル DNA を

超音波で断片化したのち、両端にダンベル型のアダプターをアニーリングする。このダンベル型アダプターは 1 本鎖 DNA であり、その末端の十数塩基が自己相補配列になっており、末端同士がアニーリングして 2 本鎖 DNA になっているが、その他の部分は 1 本鎖となっている。両端にダンベル型アダプターを付けた DNA ライブラリーは、幾何学的には閉じた一本鎖 DNA のループとなっている。SMRT シーケンサーはこのループを繰り返しシーケンスすることができる。これにより、サンプル DNA の 2 本鎖のうち、片側の鎖とその相補鎖を交互に何回も読んだリードが得られる。このリードを Duplex-sequencing 法と同様に、鎖ごとにコンセンサス配列を作り、そのコンセンサス配列が一致したもののみを正しく読めた配列と判定する。

SMRT 法も Duplex-sequencing 法同様、Ames 試験だけでなく様々な生体試料の突然変異検出に適用できる。

## 4. おわりに

以上、変異原性試験の現状について紹介したが、最後に、今後変異原性の評価や管理がどのようになってゆくのか考察したい。まず、コンピューターによる変異原性の予測についてであるが、今後ますます様々な場面で規制に利用されるようになるだろう。

医薬品をはじめ、新規化学物質の開発現場では、コンピューター予測をいかにうまく使うかで開発の効率が大きく変わってくる。さらに、新たに開発した化学物質の認可における安全性評価にも、M7 ガイドラインを皮切りに適用が拡大する可能性が有る。M7 ガイドラインは医薬品の不純物の変異原性物質の評価に限ったものであるが、農薬やその他工業製品に含まれる不純物の評価にも、コンピューター予測が解禁される可能性は高い。また、実際に Ames 試験を行うことが困難な物質、例えば強烈なにおいのする香料などへの評価にも適用されるだろう。

さらに、規制当局側では、膨大な既存化学物質について、評価の優先順位を決定するのにコンピューター予測が利用できる。

コンピューター予測の方式に関して言えば、様々な機械学習アルゴリズムを用いた統計ベース予測器が、DEREK など経験ベースの予測器よりも良い性能を示している。しかし、その予測性能はまだ百発百中とはいえない。論文で最も性能が良かった報告は DragonX で生成した分子記述子をサポートベクターマシンで学習した例では、感度 0.88、特異性 0.64 程度である<sup>3)</sup>。最近はやりの deep learning を用いた例はまだ報告されていない。今後、分子記述子や機械学

習方式の改良により、性能向上が期待される。

機械学習による予測精度をさらに向上させるためには、さらに多くのデータを学習させることが重要であると同時に、データの質の向上も重要である。NGSによる突然変異の直接測定は、バイアスが極めて少なく、疑陽性、偽陰性の起こりにくい系である。Ames試験の結果が怪しいものについて、NGSでデータを取り直すなどして、データの質を向上させる必要がある。

よく言われることであるが、機械学習の結果を実験に反映させ、実験結果をまた機械学習に反映させるループを回すことにより、予測器の性能はどんどん良くなる。決定木など、人間が理解できる機械学習アルゴリズムを用いれば、変異原性の新しい法則の発見も期待でき、学問的にも意義深い。

#### 参考文献

- 1) 独製品評価技術基盤機構ホームページ：  
URL : [http://www.safe.nite.go.jp/jcheck/list3.action?&request\\_locale=ja](http://www.safe.nite.go.jp/jcheck/list3.action?&request_locale=ja)
- 2) 独医薬品医療機器総合機構ホームページ：  
URL : <https://www.pmda.go.jp/int-activities/int-harmony/ich/0036.html>
- 3) Hasen, K. et al., Benchmark data set for in silico prediction of ames mutagenicity, *J. Chem. Inf. Model.*, 49, 2077-2081, 2009
- 4) Xu, C. et al., In silico prediction of chemical ames mutagenicity, *J. Chem. Inf. Model.*, 52, 2840-2847, 2012
- 5) Matsuda, T. et al., A pilot study for the mutation assay using a high-throughput DNA sequencer, *Genes and Environment*, 35, 53-56, 2013
- 6) Schmitt, M. et al., Detection of ultra-rare mutations by next-generation sequencing, *PNAS*, 109, 14508-14513, 2012
- 7) Matsuda, T. et al., Mutation assay using single-molecule real-time (SMRT<sup>TM</sup>) sequencing technology, *Genes and Environment*, 37, 15, 2015