情報論的学習理論テクニカルレポート 2009
Technical Report on Information-Based Induction Sciences 2009 (IBIS2009)

生物学情報への機械学習解析の応用（Toxicogenomics への展開）
Application of machine learning to Biological data (Toxicogenomics)

武藤裕紀*, 松下智哉, 芦原基起
Hironori Mutoh, Tomochika Matsushita and Motooki Ashihara

Abstract: Microarray technology has been widely utilized in the biological fields. To analyze the data derived from microarrays, bioinformatics technology is necessary because the amount of data is huge and biologically complicated. Toxicogenomics is an area of study where clarification and prediction of the mechanisms of toxicity are found by applying microarray technology. Here we applied Support Vector Machine (SVM) to predict a toxicity of the liver, the proliferation of bile ducts, using toxicogenomics database constructed by the Toxicogenomics Project in Japan (TGP).
Keywords: Microarray, Bioinformatics, Toxicogenomics, Support Vector Machine

1 緒言

マイクロアレイ技術の発展により、網羅的な遺伝子発現データを比較的容易に取得することが可能となり、生物学の多くの分野で活用されている。マイクロアレイデータ解析では、1サンプルにつき数万の遺伝子発現値が得られる。これらのデータは単に膨大なだけでなく、生物学的な複雑さを含むため、統計解析やバイオインフォマティクス技術の使用が不可欠である。

トキシコゲノミクスは、マイクロアレイ技術を毒性学の領域に応用し、薬物を動物や細胞に暴露して網羅的に遺伝子発現解析を行うことにより、毒性発現メカニズムの解明や毒性予測を行う学問領域である。従来の毒性評価法に比べ、創薬研究の初期段階で医薬品候補化合物の毒性を効率的に評価・予測する手法として期待され、発がん、肝毒性等のリスク評価への活用が既に試みられている[1-6]。

毒性予測には、あらかじめ毒性の発現が確認されている（毒性有り）化合物と毒性の発現が認められていない（毒性無し）化合物の情報を与えて予測モデルを作成する教師付き機械学習の手法が有用である。この予測モデルの構築には多数のデータがリファレンスとして必要となるため、毒性予測モデルの構築を目的としたデータベースが構築されている[7-8]。日本においては、独立行政法人医薬基盤研究所、国立医薬品食品衛生研究所および製薬企業の共同研究である「トキシコゲノミクスプロジェクト」において、150を超える化合物を用いた暴露実験が行われ、マイクロアレイのプラットフォームのひとつである Affymetrix 社の GeneChip により取得された遺伝子発現データおよび付帯する毒性関連情報が格納されたデータベースが構築された(TGP2; http://wwwtgp.nibio.go.jp/index.html)。

GeneChip 解析を実施するうえでの必要な前処理として、遺伝子発現値の数値化手法と、解析に有用な遺伝子の絞り込み手法の二つの検討を行う必要がある。

数値化手法に関しては、Affymetrix 社が提唱している MAS5[9]をはじめとして、いくつかの数学的処理が知られているが、処理法によって算出法が変わるため、解析結果に与える影響が異なることが報告されている[10]。

遺伝子の絞込み手法に関しては、学習データに基づいて機械的に絞り込む方法に加え、生物学情報などにより事前に絞り込む方法などがある。GeneChip にはプローブ設計の問題上、実際のターゲットとなる遺伝子発現を捕らえられていないデータも存在するため、これらを事前に取り除き、毒性予測に有用な遺伝子を絞り込むことが重要となる。

我々は、毒性予測モデルの構築を目的として、前述した項目が、機械学習による毒性予測モデルの精度にどのような影響を与えるかについて検討を行った。リファレンスデータベースとして「トキシコゲノミクスプロジェクト」で構築されたデータベースを用い、機械学習アルゴリズムには、近年、未学習データへの汎用性が高いことで注目されている SVM(Support Vector Machine)を用いた。また毒性予測は、肝臓における胆管増生を対象として実施した。

*中外製薬株式会社 研究本部 創薬資源研究部, 〒247-8530 神奈川県鎌倉市梶原 200, tel. 0467-47-6312,
e-mail: mutohrn@chugai-pharm.co.jp, Discovery Science & Technology Dept. Research division, Chugai Pharmaceutical CO., LTD. 200 Kajiwara, Kamakura, Kanagawa, 247-8530 JAPAN

2 データおよび解析手法

2.1. リファレンスデータベース

リファレンスデータベースとして独立行政法人医薬基盤研究所、国立医薬品食品衛生研究所および製薬企業の共同研究である「トキシコゲノミクスプロジェクト」において構築されたデータベースを使用した(TGP2; http://wwwtgp.nibio.go.jp/index.html)。データは Affymetrix 社の GeneChip Rat Genome 230 2.0 Array (14,562 遺伝子、31,099 プローブ)により測定されたものを使用した(Affymetrix 社; http://www.affymetrix.com/index.affx)。

2.2. 学習用データ

本研究では、評価する毒性として、肝臓における胆管増生（病理所見名「Proliferation, Bile duct」）をターゲットとした。毒性有り化合物は、「Proliferation, Bile duct」の病理所見がいずれかの条件において確認された化合物を選択した。毒性無し化合物は「Proliferation, Bile duct」以外の所見の確認された化合物の一部を選択した。この選択指標に基づいて、リファレンスデータベースから毒性有り化合物 6 化合物および毒性無し化合物 10 化合物を選択した (Table 1)。これら計 16 化合物を高用量で 28 日間反復投与し、投与後 29 日に取得したデータ(n=3)および、各化合物の溶媒のみを投与したコントロールデータ(n=3)を学習用データセット(計 96 サンプル)として用いた。

Table 1 List of Training data

Compound	Toxicity	Vehicle	Dose [mg/kg]
acetamidofluorene	Positive	0.5% MC	300
allyl alcohol	Positive	corn oil	30
lomustine	Positive	0.5% MC	6
methapyrilene	Positive	0.5% MC	100
naphthyl isothiocyanate	Positive	corn oil	15
thioacetamide	Positive	0.5% MC	45
amiodarone	Negative	0.5% MC	200
amitriptyline	Negative	0.5% MC	150
clofibrate	Negative	corn oil	300
flutamide	Negative	corn oil	150
furosemide	Negative	0.5% MC	300
hydroxyzine	Negative	0.5% MC	100
imipramine	Negative	0.5% MC	100
metformin	Negative	0.5% MC	1000
omeprazole	Negative	0.5% MC	1000
ticlopidine	Negative	0.5% MC	300

2.3. 評価用データ

リファレンスデータベースに含まれている、薬物投与を行った全実験データを評価用データセットに用いた。

2.4. データ処理および解析手法

GeneChip データは MAS5[9]および GCRMA[11]を用いて数値化し、底を 2 として対数化を行った。溶媒のみを投与したコントロールデータとの Ratio データを生成するため、各化合物に対し溶媒のみを投与したコントロール実験データ（n=3）の平均値を計算し、投与実験データとの差を取った(数式 1)。

$$\log_2(Exp_{treated,i}) - \frac{\sum_{j=1}^{n_{control}} \log_2(Exp_{Control,j})}{n_{Control}} \quad (1)$$

予測モデルの構築には SVM(Support vector machine)を用い、カーネル関数は Linear kernel を用いた。変数の機械的な絞込みには SVM-RFE(Recursive feature elimination)[12]を用い、weight パラメータの小さい変数 5%を繰り返しごとに消去、評価関数の値が下がったところで繰り返しを止めた。データセットは同一化合物の反復実験を含むため、クロスバリデーションは化合物単位の LOOCV(Leave one out cross validation)を実行した(Fig.1)。評価関数には MCC(Matthews Correlation Coefficient)を用いた(数式 2) [13]。また Sensitivity, Specificity は下記に示す数式 3、4 を用いて算出した。

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)\cdot(TP+FN)\cdot(TN+FP)\cdot(TN+FN)}} \quad (2)$$

$TP : TruePositive \quad FN : FalseNegaitive$
$FP : FalsePositive \quad TN : TrueNegative$

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

すべての解析は R version 2.8.1 を用いて行った。また解析パッケージとして e1071, MASS および Bioconductor の affy, gcrma を利用した (CRAN; http://cran.r-project.org/), (Bioconductor; http://www.bioconductor.org/)。

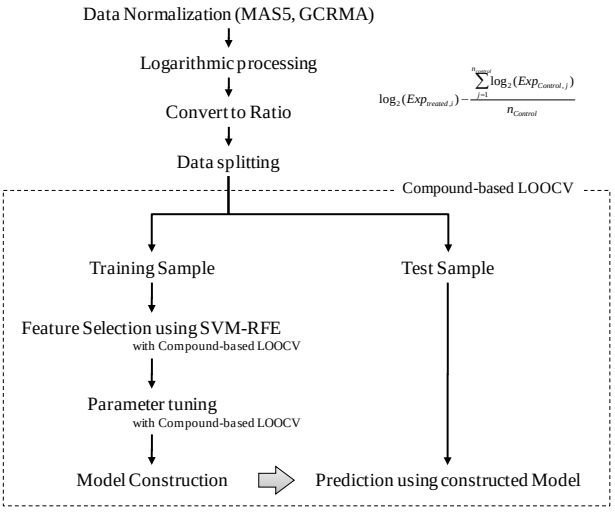

Fig. 1 Analysis Flow

3 結果

3.1. データ前処理検討

GeneChip の数値化手法による影響を調べるため、MAS5[9]および GCRMA[11]により数値化を行い、両者の精度の比較を以下の要領で行った。まず、リファレンスデータベースに含まれる 2 種類の異なる実験プロトコルを用いて Duplicate で取得したラット初代培養肝細胞の遺伝子発現データを MAS5, GCRMA で数値化し散布図を作成することで、数値の再現性への影響を確認した。また、学習用データを用いて実際に予測モデルを構築し、精度の評価を行った。

数値の再現性への影響を比較した結果では、プロトコル間およびプロトコル内のいずれにおいても MAS5 と比較して GCRMA のほうが実験間の再現性が高い傾向が確認された(Fig. 2)。

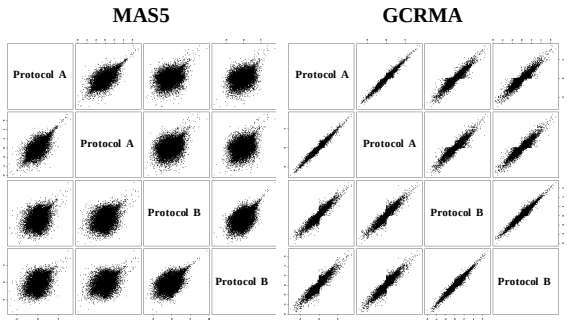

Fig. 2 Scattered plot of Log2ratio
遺伝子発現プロファイルはラット初代培養肝細胞から取得。それぞれisoniazidで処理を行った。24時間後の初代培養肝細胞から、2種類の実験プロトコル(A,B)によりデータを取得。

また毒性予測モデルの精度を比較した結果では、GCRMA により構築されたモデルにおいて Sensitivity, Specificity がそれぞれ 66.67%、100.00%と、MAS5 の 16.67%、100.00%と比べてより高い予測精度が得られた(Fig. 3)。

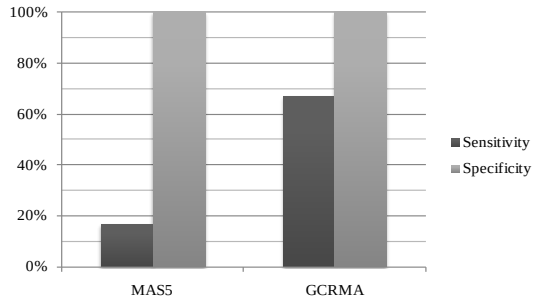

Fig. 3 Accuracy of MAS5 and GCRMA based models
MAS5およびGCRMAにより数値化されたデータから構築したモデルの精度。

次に GeneChip のプローブの絞り込み手法による影響を調べるため、Affymetrix 社の提供するプローブに関する情報と Ensembl を用いた独自の評価法を用いてプローブを分類し比較を行った。Affymetrix 社ではプローブのターゲットへのマッチングにより Annotation Grade を 5 段階に分類している(Affymetrix 社; http://www.affymetrix.com/index.affx)。本研究で使用した Rat Genome 230 2.0 Array について調べたところ、ターゲットへのマッチングが確認されている Grade A のプローブは 16,327 あり、その他の 14,772 プローブは部分的にマッチングしているものや EST クラスターにのみマッチングしているプローブであった。次に、我々は GeneChip のオリゴプローブ配列を Ensembl Transcript (Release54) 配列上へマッピングした(Ensembl; http://www.ensembl.org/index.html)。その結果、ターゲットにマップされたプローブが 8,259、クロスハイブリダイズしていたプローブが 3120、ターゲット以外にマップされたプローブが 1,907、逆向きにマップされたプローブが 1,172、全くマップされなかったプローブが 16,579 確認された(Table 2)。Annotation Grade で Grade A とされているプローブのうち、単一のターゲットにマップされたプローブは 46.6%(7,609)で、その他 53.4%(8,718)はクロスハイブリダイゼーションや逆向きに設計、遺伝子以外の配列を認識しているプローブを含む可能性が示唆された(Table 2)。一方 Ensembl 上で単一ターゲットにマップされたプローブのうち 92.1%は Grade A のプローブで、その他はわずか 7.9%であった。

Table 2 Probe annotation and mapping results

Ensembl Mapping // Annotation Grade	GradeA	Grade B,C,E,R	Total
Mapped (single target)	7,609	650	8,259
Mapped (multiple targets)	52	10	62
Mapped (cross hybridization)*	2,861	259	3,120
Mapped (target not covered)*+	1,534	373	1,907
Mapped Reverse	163	1,009	1,172
Unmapped	4,108	12,471	16,579
Total	16,327	14,772	31,099
*include mapped on Unknown			
+include partially coverd(mutiple targets)			

これらの結果に基づき、あらかじめ解析に用いるプローブセットの設定を変えて(GradeA, Mapped (単一ターゲット)およびそれらの組み合わせ)、毒性予測モデルの構築を行い、精度の比較を行った(数値化処理はすべて GCRMA で実行)。構築されたモデルの予測精度は Mapped および Grade A Mapped で最も高く、いずれも Sensitivity 83.33%, Specificity 100.00%と高い値を示した(Fig. 4)。

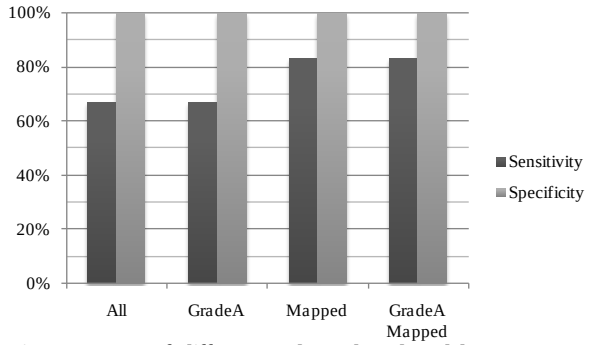

Fig. 4 Accuracy of different Probe set based models
プローブセットを変更して構築したモデルの精度。Mappedは Ensembl上の単一ターゲットにマップされた8,259プローブ、Grade A Mapped はMappedのうち Grade Aであった7,609プローブを利用した。

3.2. 毒性予測

データ前処理検討の結果、最も予測精度の高かった GCRMA で数値化を行い、Mapped プローブを用いて構築されたモデルを利用して、評価用データの毒性予測を行った。その結果、全 151 化合物中、15 化合物について、少なくともいずれかの条件で毒性有りと予測された(Table 3)。

Table 3 Predicted toxic compounds and experimental conditions

Compound // Time point & Dose	4 day			8 day			15 day			29 day		
	L	M	H	L	M	H	L	M	H	L	M	H
acetamidofluorene					T		T	T	T	T	T	T*
allyl alcohol												T*
cisplatin												T
colchicine			T									
ethambutol						T			T			
lomustine									T			T*
lornoxicam			T			T						NA
meloxicam						T						NA
methapyrilene						T			T			T*
monocrotaline						T		T	T		T	NA
naphthyl isothiocyanate						T			T			T*
naproxen			T						T			NA
nitrosodiethylamine						T		T	T		T	NA
phalloidin	NA	NA	T	NA	NA	T	NA	NA	NA	NA	NA	NA
thioacetamide			T						T		T	T*

L:Low dose, M:Middle dose, H:High dose, T:Toxic(predicted), NA:Not Avalable

Shaded:Proliferation, bile duct was observed

*Training sample

評価用データで「Proliferation, Bile duct」の病理所見が確認された 8 化合物中、学習データに用いた 6 化合物では allyl alcohol を除き、いずれも学習に用いた条件以外のサンプルでも毒性有りと予測され、学習データに用いていない nitrosodiethylamine, phalloidin も毒性有りと予測された。また、thioacetamide については所見が確認される以前のサンプルでも毒性有りと予測された。一方、所見が確認されていない 143 化合物では，7 化合物を除いて大半がいずれの条件でも毒性無しと予測された。

4 考察

本研究ではまず GeneChip データの数値化手法による毒性予測モデルの精度への影響を検討した。その結果、MAS5 と比較して GCRMA を用いたときに高い精度の予測モデルが構築できた。GCRMA による数値化では、MAS5 と比較して、低発現でのばらつきが抑えられることが確認されていることから(Data not shown)、コントロールとの Ratio データに変換すると、発現の小さなプローブのばらつきの影響が少なく MAS5 と比較して Fig. 2 のように実験間での再現性が高くなり、その結果、毒性予測精度に差が出たと考えられる。またトキシコゲノミクスによる毒性予測では、化合物により溶媒（化合物を溶解する溶液）が異なるなど(Table 1)、データ採取の条件を統一することが難しく、Ratio のようなコントロールとの比較値を用いることが重要となるため、数値化手法として MAS5 よりも GCRMA を用いることが適切と考えられた。

次に、プローブをあらかじめ選別することによる毒性予測モデルの精度への影響を検討した結果、全プローブを用いるよりも、Ensembl 上で単一ターゲットにマッピングされたプローブを用いたときに高い精度の予測モデルが構築できた。一方、Affymetrix 社の Annotation Grade で Grade A とされているプローブを用いたときは、全プローブを用いたときと予測精度に差がなかった。Table 2 で示したように、Grade A のプローブであっても、クロスハイブリダイゼーションや逆向き、遺伝子以外の配列にマップされたプローブを含んでおり、毒性予測に関連のないノイズを含んでいた可能性が考えられた。

最後に、構築したモデルを用いて評価用データの毒性予測を行った結果、所見が確認された化合物については allyl alcohol を除いて、学習用に用いた条件以外でも毒性有りと予測された。Allyl alcohol については、学習用に用いた条件(29 日、 高用量)では所見が確認されていなかったため、正しく学習できなかった可能性が考えられた。実験条件ごとに見ると、29 日、高用量のデータを用いてモデルを構築したにもかかわらず、それよりも早いタイムポイントや、低用量の条件でも毒性有りと予測されたものも多く見られた。また、thioacetamide については所見が現れる以前のサンプルでも毒性有りと予測されるなど、病理所見が確認されるより早い段階で毒性を予測できる可能性が示唆された。一方で、所見が確認されていない化合物で毒性有りと予測されているサンプルや、所見が確認されている条件で毒性無しと予測されているサンプルもあった。前者については、「Proliferation, Bile duct」の病理所見に先立つ遺伝子発現の変動を捕らえている可能性もあるが、前述したような学習データの特性により、正しく学習できていない可能性もある。したがって、構築された予測モデルの生物学的意義については精査していく必要があると考えられた。

本研究では、生物学情報への機械学習解析の応用として、トキシコゲノミクスデータの解析検討を行った。肝臓における胆管増生を毒性のターゲットとして SVM により予測モデルを構築した結果、精度の高い予測モデルの構築に成功した(Sensitivity 83.3%, Specificity 100%)。本研究の結果から、トキシコゲノミクスデータに機械学習を応用することで化合物の毒性を予測できる可能性が示唆された。また複数のデータ処理の方法を比較し、適当な手法の選択および生物学的背景を考慮した検討を行うことによって、精度の高い予測モデルが構築できた。このことから、最適な手法を選択・使用することが精度の向上に重要な因子であることが確認された。手法の最適化を行うことで、より高精度な予測モデルが構築できると思われることから、今後も検討を進めていく。

参考文献

[1] Nie AY, McMillian M, Parker JB, Leone A, Bryant S, Yieh L, Bittner A, Nelson J, Carmen A, Wan J, Lord PG. “Predictive toxicogenomics approaches reveal

underlying molecular mechanisms of nongenotoxic carcinogenicity.", *Mol Carcinog.* 2006 Dec;45(12):914-33.

[2] Fielden MR, Brennan R, Gollub J. "A gene expression biomarker provides early prediction and mechanistic assessment of hepatic tumor induction by nongenotoxic chemicals.", *Toxicol Sci.* 2007 Sep;99(1):90-100.

[3] Ellinger-Ziegelbauer H, Gmuender H, Bandenburg A, Ahr HJ. "Prediction of a carcinogenic potential of rat hepatocarcinogens using toxicogenomics analysis of short-term in vivo studies.", *Mutat Res.* 2008 Jan 1;637(1-2):23-39.

[4] Steiner G, Suter L, Boess F, Gasser R, de Vera MC, Albertini S, Ruepp S. "Discriminating different classes of toxicants by transcript profiling.", *Environ Health Perspect.* 2004 Aug;112(12):1236-48.

[5] Zidek N, Hellmann J, Kramer PJ, Hewitt PG. "Acute hepatotoxicity: a predictive model based on focused illumina microarrays.", *Toxicol Sci.* 2007 Sep;99(1):289-302.

[6] Hirode M, Ono A, Miyagishima T, Nagao T, Ohno Y, Urushidani T. "Gene expression profiling in rat liver treated with compounds inducing phospholipidosis.", *Toxicol Appl Pharmacol.* 2008 Jun 15;229(3):290-9.

[7] Ganter B, Tugendreich S, Pearson CI, Ayanoglu E, Baumhueter S, Bostian KA, Brady L, Browne LJ, Calvin JT, Day GJ, Breckenridge N, Dunlea S, Eynon BP, Furness LM, Ferng J, Fielden MR, Fujimoto SY, Gong L, Hu C, Idury R, Judo MS, Kolaja KL, Lee MD, McSorley C, Minor JM, Nair RV, Natsoulis G, Nguyen P, Nicholson SM, Pham H, Roter AH, Sun D, Tan S, Thode S, Tolley AM, Vladimirova A, Yang J, Zhou Z, Jarnagin K., "Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action.", *J Biotechnol.* 2005 Sep 29;119(3):219-44.

[8] Ganter B, Snyder RD, Halbert DN, Lee MD., "Toxicogenomics in drug discovery and development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix database.", *Pharmacogenomics.* 2006 Oct;7(7):1025-44.

[9] Hubbell E, Liu WM, Mei R. "Robust estimators for expression analysis.", *Bioinformatics.* 2002 Dec;18(12):1585-92.

[10] Lim WK, Wang K, Lefebvre C, Califano A. "Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks.", *Bioinformatics*. 2007 Jul 1;23(13):i282-8.

[11] Zhijin Wu, Rafael A. Irizarry, Robert Gentleman, Francisco Martinez Murillo, Forrest Spencer, "A Model Based Background Adjustment for Oligonucleotide Expression Arrays.", *Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 1.*, 2004 May.

[12] Guyon IM, Weston J, Barnhill S, Vapnik VN. "Gene selection for cancer classification using support vector machines.", *Mach Learn.* 2002 46:389–442.

[13] Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H., "Assessing the accuracy of prediction algorithms for classification: an overview.", *Bioinformatics.* 2000 May;16(5):412-24.

謝辞

本研究は厚生労働科学研究費補助金 H14-ﾄｷｼｺ-001 および H19-ﾄｷｼｺ-001 による。