厚生労働科学研究費補助金（化学物質リスク研究事業）
平成 30 年度分担研究報告書

生体影響予測を基盤としたナノマテリアルの統合的健康影響評価方法の提案
(1)ナノマテリアルによる細胞内網羅的遺伝子発現データベースの構築、
(2)機械学習による生体影響予測の試み

研究分担者　花方信孝　物質・材料研究機構 技術開発・共用部門 副部門長

研究要旨：初年度はナノマテリアルによる細胞内網羅的遺伝子発現データベースの構築方法の検討のため利用可能な既存の生命科学系データベースを調査した。その結果、NCBI の GEO データベースを始めいくつかの有用なデータベースを見出した。また、生体影響予測の基盤となる機械学習の実行環境として Tensor Flow を中心に整備するとともに、ナノマテリアルとして ZnO を曝露した細胞系の遺伝子発現マイクロアレイ解析を実施して機械学習に用いる実測データを取得した。さらには、マイクロアレイ解析における一色法と二色法の違いについても検討した。

A. 研究目的

現在一般社会で開発・使用されているナノマテリアルが社会的に受容されるためには、そのリスクについて十分な安全評価手法が必要である。しかしながら、その手法として代表的な動物実験は費用的にも時間的にも高コストであることに加えて、近年の動物愛護原則から敬遠され、代替手法が求められている。そこで *in vitro* 評価法の一環として、初年度はナノマテリアルによる細胞内網羅的遺伝子発現データベースの構築方法および機械学習による生体影響予測モデルを検討することを研究目的とした。

具体的には、データベースを構築する上で元データとして使用する既存のデータベースを選定するために生命科学系データベースの調査と、機械学習を実施するためにコンピュータ環境に関する調査および実環境の整備、機械学習に用いる実測データを得るためにナノマテリアルを曝露した場合の遺伝子発現マイクロアレイ解析を実施することとした。さらに既存データベースにおけるマイクロアレイ解析の一色法と二色法の扱いについても調査することとした。

B. 研究方法

1) 既存データベースの調査

生命科学系データベースのカタログである Integbio データベースカタログ（科学技術振興機構、https://integbio.jp/dbcatalog/）に登録された 1,684 件（調査時）のデータベースを中心にインターネット上で公開されているデータベースの中から遺伝子発現や化学物質の暴露などに関するものなど本研究目的に適したものを検索して調査した。

2) 機械学習の実施環境

使用するコンピュータの OS は MS Windows と Cent OS (Linux)の両方を用意した。ディープラーニング等の機械学習のライブラリとして Google の Keras / Tensor Flow を採用した。プログラムを組むための言語は Python を使用することとした。Python のディストリビューションとして Anaconda をインストールした。

表 1 ZnO 曝露実験におけるマイクロアレイの割り当て

Slide ID	Slide No.	Pos.	Block	Cy3	Cy5
AH72	257236319408	1_1	B1	THP-1_ZnO=0μg/mL_6hr	THP-1_ZnO=300μg/mL_6hr
AH72	257236319408	1_2	B2	THP-1_ZnO=300μg/mL_6hr	THP-1_ZnO=0μg/mL_6hr
AH72	257236319408	1_3	B3	THP-1_ZnO=0μg/mL_24hr	THP-1_ZnO=300μg/mL_24hr
AH72	257236319408	1_4	B4	THP-1_ZnO=300μg/mL_24hr	THP-1_ZnO=0μg/mL_24hr
AH72	257236319408	2_1	B5	A549_ZnO=0μg/mL_6hr	A549_ZnO=60μg/mL_6hr
AH72	257236319408	2_2	B6	A549_ZnO=60μg/mL_6hr	A549_ZnO=0μg/mL_6hr
AH72	257236319408	2_3	B7	A549_ZnO=0μg/mL_24hr	A549_ZnO=60μg/mL_24hr
AH72	257236319408	2_4	B8	A549_ZnO=60μg/mL_24hr	A549_ZnO=0μg/mL_24hr

表 2 本研究目的に適する思われる既存データベース

データベース名	URL	説明
Gene Expression Omnibus (GEO)	**https://www.ncbi.nlm.nih.gov/geo/**	NCBIが運営するマイクロアレイ解析等の遺伝子発現データのレポジトリ。
GEO Data Sets	**https://www.ncbi.nlm.nih.gov/gds/**	GEOのデータセット部分。実験条件等の情報が整備。
ArrayExpress	**https://www.ebi.ac.uk/arrayexpress/**	EBIが運営する遺伝子発現データアーカイブ。
CIBEX	**https://cibex.nig.ac.jp/data/**	DDBJが運営する遺伝子発現データレポジトリ。
CellMontage	**http://cellmontage.cbrc.jp/**	マイクロアレイデータ検索･解析システム。
Open TG-GATEs	**https://toxico.nibiohn.go.jp/**	医薬品等のin vivo/in vitro曝露データ集積データベース。
抗がん剤遺伝子発現データベース	**http://scads.jfcr.or.jp/db/cs/**	化合物の制がん作用と関連する遺伝子発現情報を提供するデータベース。
Comparative Toxicogenomics Database (CTD)	**https://ctdbase.org/**	環境曝露と人間の健康に関するデータベース。
Online Mendelian Inheritance in Man (OMIM)	**https://www.ncbi.nlm.nih.gov/omim/**	ヒトの遺伝子変異と遺伝病のデータベース。
Gene Signature DataBase (GeneSigDB)	**https://genesigdb.org/genesigdb/**	遺伝子発現解析結果からの変動遺伝子群のデータベース。
Integrative Disease Omics Database (iDOx DB)	**https://gemdbj.ncc.go.jp/omics/**	多層的な疾患オミックス解析の統合データベース。
Functional Annotation of the Mammalian Genome (FANTOM)	**http://fantom.gsc.riken.jp/jp/**	ゲノムDNAから転写されているRNAの機能をカタログ化したデータベース。
COXPRESdb	**https://coxpresdb.jp/**	ヒト･マウス･ラット、他4 種の共発現遺伝子データベース。
Human Protein Atlas	**https://www.proteinatlas.org/**	ヒトの器官、組織、細胞におけるタンパク質の発現および局在に関するデータベー
H-ANGEL	**http://www.h-invitational.jp/hinv/h-angel**	H-Invitationalプロジェクトの構築した転写産物データに対する遺伝子発現データ
BioGPS	**http://biogps.org/**	ヒトおよびマウスの遺伝子発現情報データベース。

3) ZnO 曝露細胞のマイクロアレイ解析

細胞：THP-1 細胞および A549 細胞

曝露ナノマテリアル：酸化亜鉛（ZnO）

曝露濃度：300 μg/mL (THP-1 細胞) または 60 μg/mL (A549 細胞)

曝露時間：6 時間または 24 時間

マイクロアレイ：Agilent SurePrint G3 Human GE Microarray 8x60K Ver. 3.0 (G4851C; GEO platform:GPL21185) 1枚

ハイブリダイゼーション：二色法

マイクロアレイの割り当て：表 1

マイクロアレイスキャナー：Agilent SuresScan G2600D

画像数値化処理ソフトウェア：Agilent Feature Extraction v11.5

データ処理ソフトウェア：自作プログラム

Gene Ontology 解析ソフトウェア：DAVID 6.8 (https://david.ncifcrf.gov/)

（マイクロアレイの一色法と二色法）

マイクロアレイ解析には一色法と二色法が存在するが、この2種類では遺伝子発現を測定する根本原理が異なり遺伝子発現データの取り扱い方に影響するため、現在ではどちらが主流か把握するために、GEO データベースから特定のマイクロアレイについて登録されているデータの解析条件情報をダウンロードして、暦年ごとに一色法と二色法のサンプル数を集計した。

C. 研究結果

1) 既存データベースの調査

Integbio データベースカタログを中心に調査を行なったところ、本研究の目的に適するいくつかのデータベースが見出された（表 2）。そのうち最大のものはよく知られている NCBI 運営の Gene Expression Omnibus (GEO) であり、DataSets で 4,348 件、Series で 105,964 件、

表 3　マイクロアレイ解析発現比分布（生データ）

Sample		ZnO/ctrl 6h THP1 [B1]	ZnO/ctrl 6h THP1 [-B2]	ZnO/ctrl 6h THP1 [rep.]	ZnO/ctrl 24h THP1 [B3]	ZnO/ctrl 24h THP1 [-B4]	ZnO/ctrl 24h THP1 [rep.]	ZnO/ctrl 6h A549 [B5]	ZnO/ctrl 6h A549 [-B6]	ZnO/ctrl 6h A549 [rep.]	ZnO/ctrl 24h A549 [B7]	ZnO/ctrl 24h A549 [-B8]	ZnO/ctrl 24h A549 [rep.]
スポット数		60,901	60,901	60,901	60,901	60,901	60,901	60,901	60,901	60,901	60,901	60,901	60,901
Log2値の分布	10以上11未満												
	9以上10未満				1								
	8以上9未満				3	6	1				1		
	7以上8未満				13	19	14				6	1	1
	6以上7未満	6	1	2	61	78	56	9	5	5	17	3	16
	5以上6未満	35	27	30	147	183	142	5	9	7	24	23	8
	4以上5未満	51	69	54	313	370	293	48	26	26	63	36	43
	3以上4未満	108	120	84	825	780	678	102	143	97	187	180	151
	2以上3未満	373	458	279	1,735	1,633	1,406	498	468	372	572	681	464
	1以上2未満	2,047	2,208	1,217	4,128	3,862	2,970	2,301	2,247	1,432	2,185	2,545	1,564
	-1超過1未満	27,377	28,139	23,355	19,358	19,031	15,427	28,293	28,168	24,081	27,467	29,906	24,818
	-2超過-1以下	1,906	2,305	1,084	4,509	4,489	3,466	2,181	2,016	1,174	2,422	2,345	1,391
	-3超過-2以下	169	219	57	1,592	1,518	1,265	170	168	57	265	265	143
	-4超過-3以下	17	12	4	626	635	556	9	13	6	37	25	18
	-5超過-4以下	1			335	263	234		1	1	2	4	
	-6超過-5以下				148	112	103						
	-7超過-6以下				66	51	46						
	-8超過-7以下				42	18	20						
	-9超過-8以下				8	3	2						
	-10超過-9以下				2	1	1						
	-11超過-10以下					1							
total		32,090	33,558	26,166	33,912	33,053	26,680	33,616	33,264	27,258	33,248	36,014	28,617
total(%)		52.69%	55.10%	42.96%	55.68%	54.27%	43.81%	55.20%	54.62%	44.76%	54.59%	59.14%	46.99%
Log2値≧1		2,620	2,883	1,666	7,226	6,931	5,560	2,963	2,898	1,939	3,055	3,469	2,247
-1<Log2値<1		27,377	28,139	23,355	19,358	19,031	15,427	28,293	28,168	24,081	27,467	29,906	24,818
Log2値≦-1		2,093	2,536	1,145	7,328	7,091	5,693	2,360	2,198	1,238	2,726	2,639	1,552
Log2値≧1		8.16%	8.59%	6.37%	21.31%	20.97%	20.84%	8.81%	8.71%	7.11%	9.19%	9.63%	7.85%
-1<Log2値<1		85.31%	83.85%	89.26%	57.08%	57.58%	57.82%	84.17%	84.68%	88.34%	82.61%	83.04%	86.72%
Log2値≦-1		6.52%	7.56%	4.38%	21.61%	21.45%	21.34%	7.02%	6.61%	4.54%	8.20%	7.33%	5.42%
中央値		-0.036	-0.049	-0.050	-0.013	-0.046	-0.036	-0.032	-0.045	-0.039	-0.052	-0.059	-0.055
算術平均		0.013	0.002	0.009	-0.002	0.012	0.007	0.017	0.011	0.018	0.007	0.010	0.012
標準偏差		0.785	0.806	0.715	1.642	1.630	1.649	0.783	0.776	0.708	0.873	0.834	0.782

Samples で 2,783,483 件の遺伝子発現データが登録されている（2018/12/14 調査時点）。今後はこの GEO のデータを主に利用することとした。また、Integbio データベースカタログには収録されていないが、化学物質の曝露と遺伝子や病気との関連をまとめたデータベースとしてノースカロライナ州立大学運営の Comparative Toxicogenomics Database (CTD)が存在し、補足的に利用可能と思われる。なお、公開されている既存データベースの中でも、全データを一括ダウンロードして再利用可能なデータベースと、検索はできるものの一部データしかダウンロードできないデータベースが存在した。後者のタイプのデータベースからウェブスクレイピングにより全データをダウンロードすることは可能かもしれないが、当該データベースのライセンス条件的に問題がないかどうか慎重な検討が必要と思われる。

2) 機械学習の実施環境

機械学習ライブラリ Keras / Tensor Flow をインストールしてプログラムが実行可能であることを確認した。マイクロアレイ発現データを用いた機械学習の前に、GEO データベース上のデータと実測したマイクロアレイ発現データを用いたデータマイニングを試みたが、まだ結果はまとまっていない。

3) ZnO 曝露細胞のマイクロアレイ解析

発現比（生データ）：全 8 アレイおよびカラースワップ間で再現性が取れたデータの発現比の分布を表 3 に示す。THP1 細胞の 24 時間後は ZnO の影響を大きく受けていることが分かる。また、A549 細胞の 6 時間後は、他の条件よりも遺伝子発現の変動が小さいものの ZnO の影響を明らかに受けている。

階層的クラスタリングその1：全 8 アレイのデータのうち 8 アレイとも発現比が求まったプローブ 20,090 個の発現比データで階層的クラスタリングを行なった（図 1）。まず、それぞれのサンプ

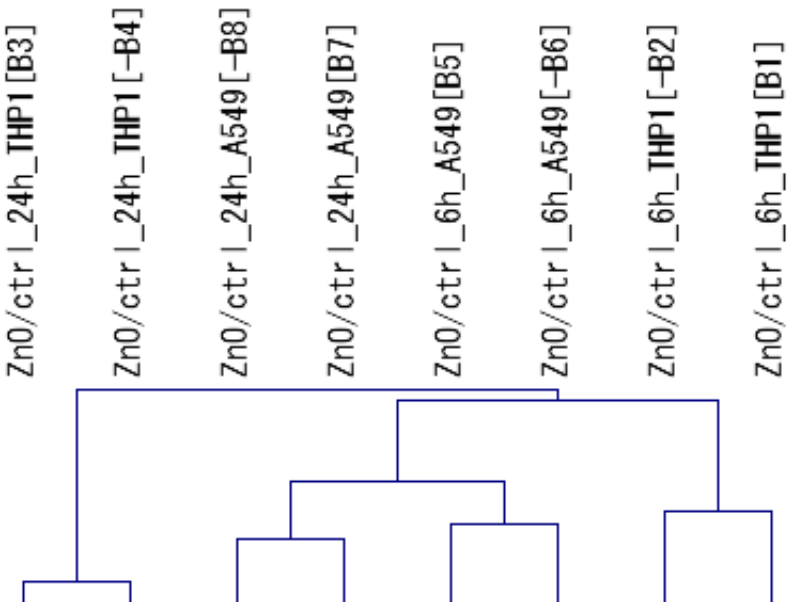

図 1　階層的クラスタリング。アレイ数 8 個、プローブ数 20,090 個。

ル条件のカラースワップ間でクラスタが形成され、マイクロアレイのサンプル条件ごとの再現性が取れていることが確認された。続いて A549 細胞の 6 時間後と 24 時間後がクラスタを形成した。THP1 細胞の 24 時間後は他とは大きく異なる発現パターンであると示されており、やはり ZnO の影響が大きいことの表れだと考えられる。

階層的クラスタリングその 2：続いて、カラースワップ間で再現性を取ったデータのうち 4 つとも発現比が求まったプローブ 17,167 個の発現比データで階層的クラスタリングを行なった場合も同様にクラスタが形成された（図 2）。やはり、THP1 細胞の 24 時間後は他とは大きく異なる。そして、THP1 細胞と A549 細胞では ZnO に対する感受性が異なると考えられる。

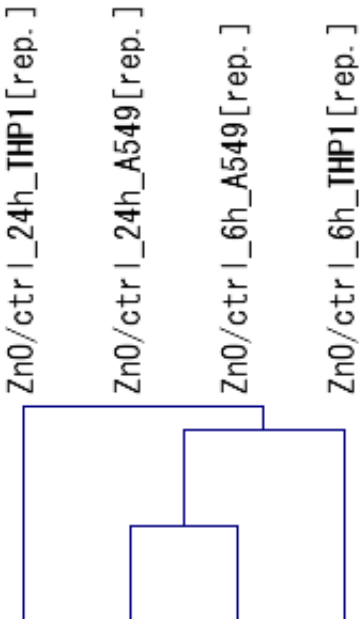

図 2　階層的クラスタリング。カラースワップ処理後、プローブ数 17,167 個

発現比（転写産物）：複数のプローブが 1 つの転写産物に対応する場合があるので、転写産物単位でまとめた発現比データのうち RefSeq (Reference Sequence)の分布を表 4 に示す。なお、1 つの遺伝子から複数の転写産物が転写されるケースがある（transcript variant）。この分布からも THP1 細胞の 24 時間後は ZnO の影響を強く受けていることが分かる。

Gene Ontology 解析：転写産物単位の発現比について Gene Ontology 解析を行なった（表 5）。ZnO の影響を最も受けた THP1 細胞の 24 時間後は発現が上昇する群において GO:0006954 inflammatory response（炎症反応）の転写産物が顕著に多かった。inflammatory response は THP1 細胞の 6 時間後でも有意に多かったことから、ナノマテリアル ZnO は THP1 細胞に対して早期から影響を与えていることが分かる。一方で、A549 細胞では inflammatory response の転写物が有意に増えているとは認められなかった。また、THP1 細胞の 6 時間後では GO:0006364 rRNA processing に分類される転写産物の発現が減少しており、ZnO を曝露した初期段階で rRNA 系が障害を受けていると考えらえる。ただし、THP1 細胞の 24 時間後では rRNA processing の転写物は有意に発現が減少しているとは言えず、時間の経過により rRNA processing は回復しているのかもしれない。一方で、A549 細胞では rRNA processing の転写物は特に発現が減少しておらず、障害を受けていないと見られる。

4)　マイクロアレイの一色法と二色法

GEO データベースから Agilent ヒト遺伝子発現のマイクロアレイを一例として集計した。このマイクロアレイは数年ごとに改定バージョンが登場し、現在は V1 (4x44K フォーマット; GEO プラットフォーム番号 GPL4133, GPL6480), V2 (4x44K; GPL10332, GPL13497), V3 (8x60K; GPL20844, GPL21185)の 3 つのバージョンがある。GEO データベースはマイクロアレイ解析データのレポジトリとして最大級であり、現在ではマイクロアレイ解析の論文発表においてレポジトリへの登録が義務付けられていることが多いことから、世界におけるマイクロアレイ解析の実情を反映していると言える。各バージョンと 3 つの

バージョンの合計について一色法と二色法の
サンプル数の暦年変化を図 3 に示す。

表 4 マイクロアレイ解析発現比分布（転写産物、RefSeq のみ）

Sample		ZnO/ctrl 6h THP1 [rep.]	ZnO/ctrl 24h THP1 [rep.]	ZnO/ctrl 6h A549 [rep.]	ZnO/ctrl 24h A549 [rep.]
転写産物数		21,033	21,033	21,033	21,033
Log2値の分布	10以上11未満				
	9以上10未満				
	8以上9未満				
	7以上8未満		11		
	6以上7未満		31	2	4
	5以上6未満	7	92	2	5
	4以上5未満	19	192	14	21
	3以上4未満	58	396	43	60
	2以上3未満	157	748	199	222
	1以上2未満	691	1,518	759	779
	-1超過1未満	12,062	7,830	12,325	12,754
	-2超過-1以下	544	1,951	615	652
	-3超過-2以下	23	718	31	69
	-4超過-3以下	3	330	5	7
	-5超過-4以下		143	1	
	-6超過-5以下		72		
	-7超過-6以下		34		
	-8超過-7以下		8		
	-9超過-8以下		1		
	-10超過-9以下		1		
	-11超過-10以下				
total		13,564	14,076	13,996	14,573
total(%)		64.49%	66.92%	66.54%	69.29%
Log_2値≧1		932	2,988	1,019	1,091
-1<Log2値<1		12,062	7,830	12,325	12,754
Log_2値≦-1		570	3,258	652	728
Log_2値≧1		6.87%	21.23%	7.28%	7.49%
-1<Log2値<1		88.93%	55.63%	88.06%	87.52%
Log_2値≦-1		4.20%	23.15%	4.66%	5.00%
中央値		-0.055	-0.072	-0.046	-0.065
算術平均		0.014	-0.018	0.012	0.006
標準偏差		0.708	1.736	0.708	0.755

表 5 Gene Ontology 解析結果

(a) THP1, 6h, up-regulated

Term	Count	%	PValue	List Total	Pop Hits	Fold	Bonferroni	FDR
GO:0006954 inflammatory response	56	6.25	1.97E-14	766	379	3.24	6.61E-11	3.58E-11
GO:0006986 response to unfolded protein	18	2.01	1.15E-12	766	42	9.40	3.87E-09	2.10E-09
GO:0043066 negative regulation of apoptotic process	51	5.69	7.34E-09	766	455	2.46	2.47E-05	1.34E-05
GO:0045926 negative regulation of growth	10	1.12	4.94E-08	766	19	11.54	1.66E-04	9.01E-05
GO:0042981 regulation of apoptotic process	29	3.24	4.51E-07	766	213	2.98	1.52E-03	8.22E-04
GO:0043547 positive regulation of GTPase activity	54	6.03	4.90E-07	766	565	2.10	1.65E-03	8.93E-04
GO:0071294 cellular response to zinc ion	9	1.00	8.64E-07	766	19	10.38	2.90E-03	1.58E-03
GO:0042026 protein refolding	8	0.89	1.85E-06	766	15	11.69	6.21E-03	3.37E-03
GO:0032496 response to lipopolysaccharide	23	2.57	5.44E-06	766	164	3.07	1.81E-02	9.91E-03
GO:0043491 protein kinase B signaling	10	1.12	1.16E-05	766	33	6.64	3.83E-02	2.11E-02

(b) THP1, 6h, down-regulated

Term	Count	%	PValue	List Total	Pop Hits	Fold	Bonferroni	FDR
GO:0006364 rRNA processing	41	7.44	8.56E-22	475	214	6.77	1.69E-18	1.47E-18
GO:0000462 maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA,	9	1.63	2.20E-06	475	32	9.94	4.34E-03	3.78E-03
GO:0042273 ribosomal large subunit biogenesis	7	1.27	5.50E-05	475	25	9.90	1.03E-01	9.43E-02

表 5 Gene Ontology 解析結果（続き）

(c) THP1, 24h, up-regulated

Term	Count	%	PValue	List Total	Pop Hits	Fold	Bonferroni	FDR
GO:0006954 inflammatory response	120	4.24	3.21E-17	2462	379	2.16	1.90E-13	6.22E-14
GO:0060333 interferon-gamma-mediated signaling pathway	34	1.20	1.50E-10	2462	71	3.27	8.87E-07	2.90E-07
GO:0006955 immune response	108	3.82	3.28E-09	2462	421	1.75	1.94E-05	6.35E-06
GO:0060337 type I interferon signaling pathway	30	1.06	4.03E-09	2462	64	3.20	2.39E-05	7.80E-06
GO:0032496 response to lipopolysaccharide	53	1.87	2.24E-08	2462	164	2.20	1.33E-04	4.34E-05
GO:0045087 innate immune response	105	3.71	8.53E-08	2462	430	1.67	5.05E-04	1.65E-04
GO:0071222 cellular response to lipopolysaccharide	40	1.41	1.03E-07	2462	113	2.41	6.09E-04	1.99E-04
GO:0045766 positive regulation of angiogenesis	40	1.41	1.75E-07	2462	115	2.37	1.04E-03	3.39E-04
GO:0043547 positive regulation of GTPase activity	127	4.49	5.02E-07	2462	565	1.53	2.97E-03	9.72E-04
GO:0030168 platelet activation	39	1.38	5.30E-07	2462	115	2.31	3.14E-03	1.03E-03

(d) THP1, 24h, down-regulated

Term	Count	%	PValue	List Total	Pop Hits	Fold	Bonferroni	FDR
GO:0070125 mitochondrial translational elongation	43	1.41	5.89E-13	2673	85	3.18	3.45E-09	1.14E-09
GO:0070126 mitochondrial translational termination	42	1.38	4.92E-12	2673	86	3.07	2.88E-08	9.52E-09
GO:0055114 oxidation-reduction process	146	4.78	2.65E-08	2673	592	1.55	1.55E-04	5.12E-05
GO:0006364 rRNA processing	64	2.10	4.11E-07	2673	214	1.88	2.40E-03	7.95E-04
GO:0042384 cilium assembly	42	1.38	1.72E-06	2673	124	2.13	1.00E-02	3.33E-03
GO:0030705 cytoskeleton-dependent intracellular transport	11	0.36	1.30E-04	2673	18	3.84	5.33E-01	2.51E-01
GO:0043434 response to peptide hormone	18	0.59	2.31E-04	2673	44	2.57	7.42E-01	4.47E-01
GO:0060070 canonical Wnt signaling pathway	27	0.88	3.43E-04	2673	83	2.04	8.65E-01	6.61E-01

(e) A549, 6h, up-regulated

Term	Count	%	PValue	List Total	Pop Hits	Fold	Bonferroni	FDR
GO:0006986 response to unfolded protein	16	1.64	7.92E-10	844	42	7.58	2.60E-06	1.44E-06
GO:0000122 negative regulation of transcription from RNA polymerase II promoter	74	7.59	7.16E-09	844	720	2.04	2.35E-05	1.30E-05
GO:0045944 positive regulation of transcription from RNA polymerase II promoter	84	8.62	1.78E-06	844	981	1.70	5.82E-03	3.23E-03
GO:0045892 negative regulation of transcription, DNA-templated	50	5.13	5.55E-06	844	499	1.99	1.81E-02	1.01E-02
GO:0007264 small GTPase mediated signal	30	3.08	1.74E-05	844	246	2.43	5.55E-02	3.16E-02
GO:0035914 skeletal muscle cell differentiation	12	1.23	2.41E-05	844	49	4.87	7.61E-02	4.38E-02
GO:0043065 positive regulation of apoptotic process	33	3.38	4.89E-05	844	300	2.19	1.48E-01	8.89E-02
GO:0045444 fat cell differentiation	14	1.44	6.16E-05	844	73	3.82	1.83E-01	1.12E-01
GO:0006351 transcription, DNA-templated	134	13.74	1.37E-04	844	1955	1.36	3.62E-01	2.48E-01
GO:0045599 negative regulation of fat cell differentiation	10	1.03	1.94E-04	844	42	4.74	4.71E-01	3.52E-01

(f) A549, 6h, down-regulated

Term	Count	%	PValue	List Total	Pop Hits	Fold	Bonferroni	FDR
GO:0051301 cell division	41	6.51	7.65E-12	556	350	3.54	1.74E-08	1.33E-08
GO:0007067 mitotic nuclear division	34	5.40	9.14E-12	556	248	4.14	2.08E-08	1.60E-08
GO:0006334 nucleosome assembly	19	3.02	7.01E-08	556	119	4.82	1.59E-04	1.22E-04
GO:0000070 mitotic sister chromatid segregation	9	1.43	8.91E-07	556	25	10.87	2.02E-03	1.55E-03
GO:0007062 sister chromatid cohesion	16	2.54	1.42E-06	556	103	4.69	3.22E-03	2.47E-03
GO:0000183 chromatin silencing at rDNA	10	1.59	2.40E-06	556	37	8.16	5.43E-03	4.19E-03
GO:0006260 DNA replication	19	3.02	3.76E-06	556	155	3.70	8.51E-03	6.56E-03
GO:0006335 DNA replication-dependent nucleosome assembly	9	1.43	7.06E-06	556	32	8.49	1.59E-02	1.23E-02
GO:0007059 chromosome segregation	12	1.90	1.29E-05	556	68	5.33	2.89E-02	2.26E-02
GO:0032200 telomere organization	8	1.27	2.07E-05	556	27	8.95	4.59E-02	3.61E-02

(g) A549, 24h, up-regulated

Term	Count	%	PValue	List Total	Pop Hits	Fold	Bonferroni	FDR
GO:0043065 positive regulation of apoptotic process	38	3.59	1.05E-06	878	300	2.42	3.37E-03	1.91E-03
GO:0000122 negative regulation of transcription from RNA polymerase II promoter	67	6.33	5.74E-06	878	720	1.78	1.82E-02	1.04E-02
GO:0030308 negative regulation of cell growth	20	1.89	1.60E-05	878	121	3.16	4.99E-02	2.90E-02
GO:0006986 response to unfolded protein	11	1.04	4.58E-05	878	42	5.01	1.37E-01	8.31E-02
GO:0006351 transcription, DNA-templated	139	13.14	1.16E-04	878	1955	1.36	3.09E-01	2.09E-01
GO:0035914 skeletal muscle cell differentiation	11	1.04	1.84E-04	878	49	4.29	4.45E-01	3.33E-01
GO:0006954 inflammatory response	38	3.59	1.92E-04	878	379	1.92	4.60E-01	3.49E-01
GO:0001666 response to hypoxia	22	2.08	2.55E-04	878	172	2.45	5.58E-01	4.61E-01
GO:0045926 negative regulation of growth	7	0.66	3.01E-04	878	19	7.05	6.19E-01	5.44E-01
GO:0071294 cellular response to zinc ion	7	0.66	3.01E-04	878	19	7.05	6.19E-01	5.44E-01

(h) A549, 24h, down-regulated

Term	Count	%	PValue	List Total	Pop Hits	Fold	Bonferroni	FDR
GO:0001525 angiogenesis	23	3.32	1.93E-05	606	223	2.86	5.57E-02	3.47E-02

このグラフから 2009 年までは二色法が主流であり、それからわずか 2 年後の 2011 年には二色法に替わって一色法が主流となっている。これはマイクロアレイ解析のサンプル数が増える中で効率の良い一色法の方が好まれるためだと思われる。また、後継バージョンである V2 や V3 では 80%程度は一色法で利用されており、二色法ではあまり使われていないことが分かる。一方で初期バージョンの V1 は二色法での利用割合がここ 8 年間ほどは 40%程度を維持している。また、一色法について見てみると、V1 から V2 や V3 への乗り換えが進んでいることが分かる。

D. 考察

既存の生命科学系データベースは多種多様であったが、本研究に最も適しているのはマイクロアレイ解析の生データが登録されている GEO データベースであると考える。補足的に CTD データベースが利用可能と思われる。これらのデータを利用して機械学習を行なう場合に一番の問題になると思われることは、サンプル条件などのラベル付けである。人間が手動でラベルを付けるのは困難なため、データのメタ情報からうまくラベルを生成する方法を検討する必要がある。また、Gene Ontology 解析の結果をラベルとして利用することも検討する価値があると思われる。

機械学習を実行するコンピュータ環境として小規模な Windows ワークステーションおよび Linux サーバーを整備したが、演算能力が不足して現実的な実行速度が得られないことが今後発生した場合にはスパコンの利用も検討するべきかもしれない。

遺伝子発現マイクロアレイ解析の実測データとして ZnO を曝露した細胞の解析を実施したが、mRNA 発現だけでなく microRNA 発現も測定したり、化学物質の毒性のモデルとして ZnO 以外の金属酸化物なども測定したりして、機械学習に利用するデータの幅を広げることを検討したい。

マイクロアレイ解析のデータの扱いにおいて一色法と二色法の違いは重要である。測定原理として一色法はマイクロアレイ上のプローブ量が一定であることを前提としており、製造上の誤差が小さい必要がある。一方で二色法は製造上の誤差があっても問題ない方法であり、2 種類のサンプルを異なる蛍光色素で標識し競合ハイブリダイゼーションを行なうことで発現比を測定する。このように原理が異なるため、一色法と二色法のデータは相互に直接比較することはできない。今後も一色法が主流であると見込まれるが、二色法のデータも有効活用するためには一色法のデータと比較できるようにデータ変換を行わなければならない。同一サンプルを一色法と二色法でマイクロアレイ解析したデータを基に適切なデータ変換法を検討していきたい。

以上を踏まえて、次年度は GEO データベースから利用可能なデータを抽出･整理した上で、機械学習を実施する。GEO データベースからの抽出にあたって解決するべき主な研究課題は、①サンプル情報をどのように解釈し分類するか、②一色法と二色法のデータ変換、③異なるマイクロアレイプラットフォーム間のデータ変換、が挙げられる。また、機械学習の実行にあたっては、①入力する特徴量の削減、②機械学習アルゴリズムの選択と最適化、が検討するべき課題となる。なお、マイクロアレイ解析の実測データも増やしていく予定である。

E. 結論

本年度は、ナノマテリアルによる細胞内網羅的遺伝子発現データベースの構築方法の検討のため利用可能な既存の生命科学系データベースを調査し、今後構築するデータベースの元データに利用すべきものとして NCBI 運営の GEO データベースやノースカロライナ州立大学運営の CTD データベースを含む複数のデータベースを選択した。また、生体影響予測の基盤となる機械学習を実施するコンピュータ環境とし

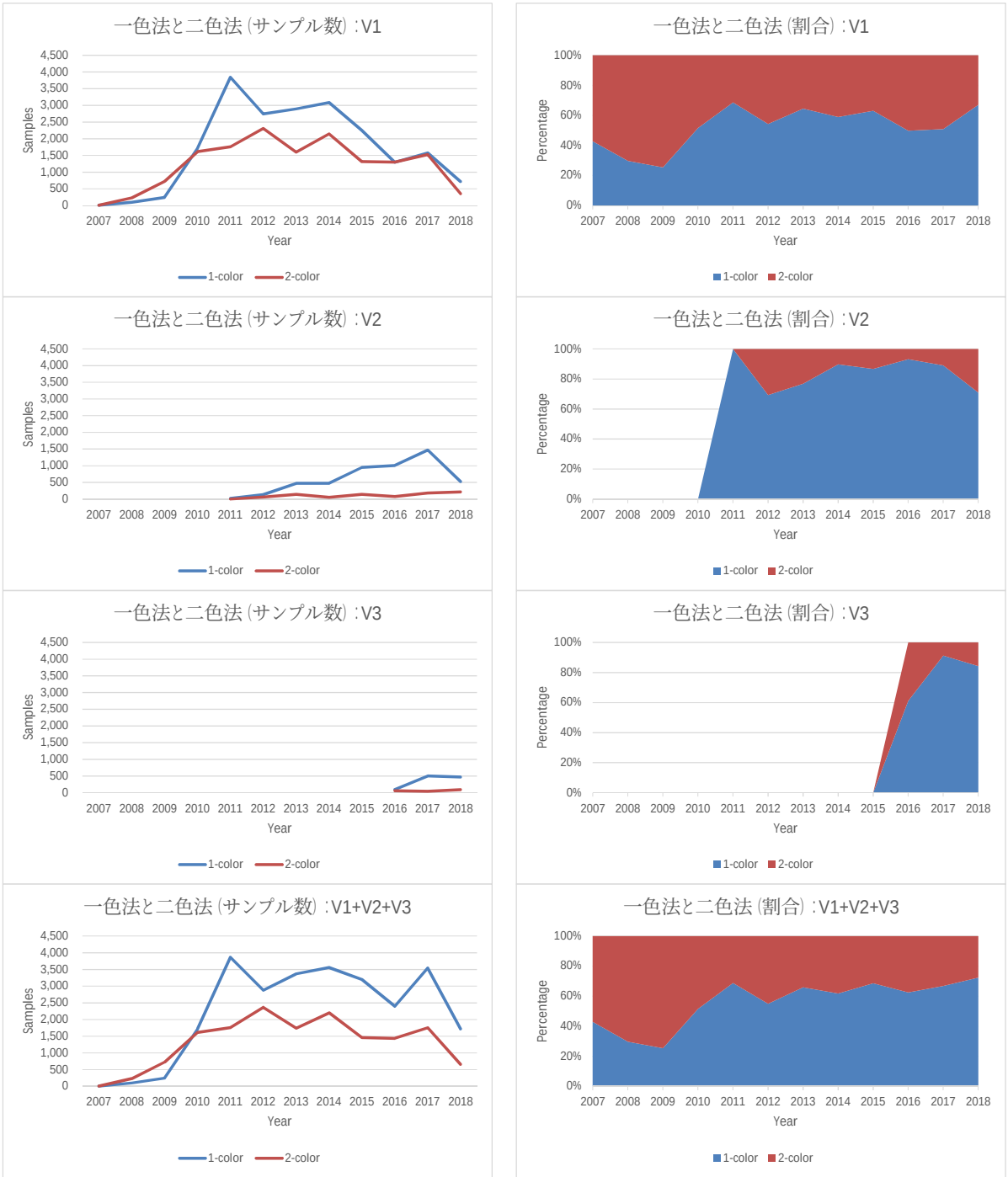

図 3　Agilent ヒト遺伝子発現マイクロアレイの GEO データベースにおける一色法と二色法の遷移

て Google の Keras / Tensor Flow を中心にセットアップを行ない、その動作確認を行なった。ナノマテリアルとして ZnO を曝露した細胞の遺伝子発現マイクロアレイ解析を実施して、機械学習で利用する実測データを取得した。この実測データについて階層的クラスタリングや Gene Ontology 解析などの検討を加えた。さらには、マイクロアレイ解析における一色法と二色法の利用現状について調査し、測定原理の違いによるデータ変換について考察した。

F. 研究発表

本年度はなし

G. 知的所有権の出願･登録状況

本年度はなし