ポストゲノム時代のデータマイニング

静岡大学　理学部　生物科学科
石原顕紀

ポストゲノム時代とは

ゲノム：

Gene + ome = Genome

遺伝子 + 集合 = ゲノム

一般的に、ある生物における
全DNA配列をさすことが多い

ポストゲノム時代とは

21 November 2002
Asteroid impacts
Reassessing the hazard
Superconductivity
Plutonium springs a surprise
Diabetes and obesity
JNK cascade as drug target
naturejobs
THE TOMA
the mo
15 February 2001
nature
www.nature.com
the human genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
ヒトゲノム大特集
nature
he chicken genome
THE POTATO GENOME

遺伝子とは

遺伝子は生物の遺伝情報を担う主要因子であると考えられている。全ての生物でDNAを媒体として、その塩基配列にコードされている。

遺伝子とは

Statistics

ヒトゲノムの情報

Summary

Assembly:	GRCh37.p8, Feb 2009	
Database version:	69.37	
Base Pairs:	3,300,551,249	←全DNAの長さ（塩基対）
Golden Path Length:	3,101,804,739	
Genebuild by	Ensembl	
Genebuild method	Full genebuild	
Genebuild started	Jul 2010	
Genebuild released	Apr 2011	
Genebuild last updated/patched	Oct 2012	

DNAの配列はA, G, C, Tのアルファベットで表記される。
つまり文字配列のテキストデータとして扱える。
１文字＝１バイトとすると約3.3GB。
DVD一枚に収まる。

Gene counts

Coding genes:	20,476	←タンパク質をコードしている遺伝子の数
Non coding genes:	22,170	
Pseudogenes:	13,322	
Gene exons:	700,944	
Gene transcripts:	201,816	←mRNAの数

データマイニングとは

データマイニング：

Data + mining = Data mining

データ+ 採掘 = データを取り出すこと

人間の目、脳で解釈することが難しい膨大なデータを、コンピュータ（バイオインフォマティクス）を用いることによって、人間にとって有用な情報を取り出す。

バイオインフォマティクスとは

バイオインフォマティクス：

Bio　+ informatics = Bioinformatics

生物　+ 情報科学 = 生物情報科学

バイオインフォマティクスの２つの側面

＊学問としてのバイオインフォマティクス

データ解析を行うアルゴリズム（解法）を提案する分野。数学的要素が強い。

＊ツールとしてのバイオインフォマティクス

上記アルゴリズムを用いてデータ解析を行う実験情報学的分野。

ツールとしてのバイオインフォマティクス

＊発現解析方法（mRNA定量法）

・ノザンブロット

・RT-PCR

・RNase プロテクションアッセイ

・リアルタイムPCR

→いずれの方法も1遺伝子ずつしか検討できない。

→ゲノムワイドな解析には不向き。

ツールとしてのバイオインフォマティクス

＊発現解析方法（mRNA定量法）

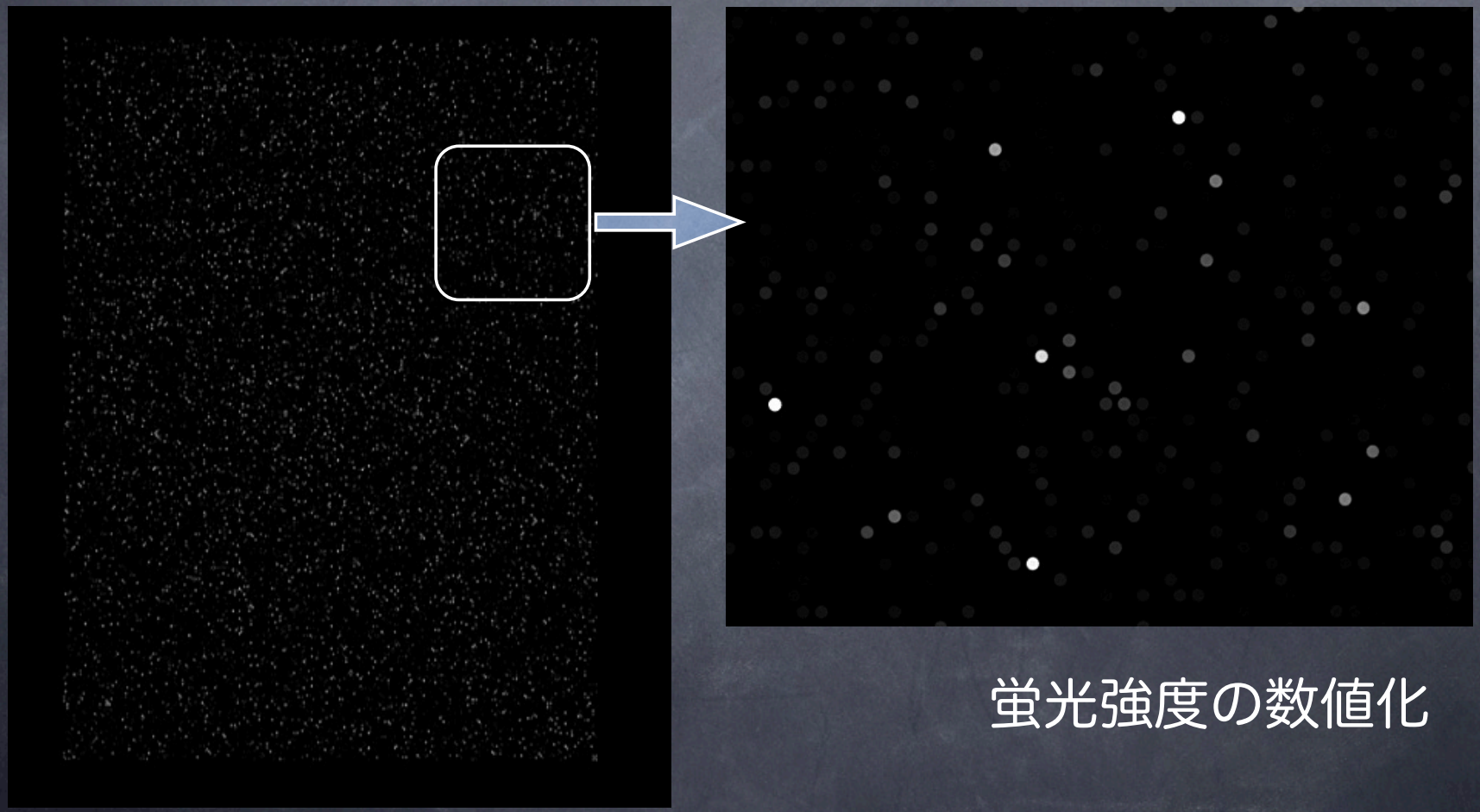
・マイクロアレイ

DNAマイクロアレイはDNAチップとも呼ばれ、細胞内の遺伝子発現量を測定するために、多数のDNA断片をプラスチックやガラス等の基板上に高密度に配置した分析器具のこと。

ツールとしてのバイオインフォマティクス

＊発現解析方法（マイクロアレイ）

結果画像



蛍光強度の数値化

ツールとしてのバイオインフォマティクス

＊発現解析方法（マイクロアレイ）

データマトリックス

サンプル数 →

アノテーション（注釈）→

遺伝子数 ↓

◇	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	AC	AD	AE	AF	AG
1	ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	DMSO	T3	OH-CB159	OH-CB106	GeneSymbol	GeneName	EnsemblID	TIGRID	GO
2	A_10_P000018	0.55	0.84	0.21	0.25	0.13	0.97	0.94	1.09	0.14	0.25	0.28	6.36	6.12	6.3	5.44	5.67	5.84	0.3944956	0.6093174	6.2576415	5.648762	LOC397792	alpha-1-antiproteinase		TC383838	
3	A_10_P000075	3.71	3.36	4.49	3.38	3.68	1.07	1.31	1.03	0.63	0.32	0.37	1.21	1.7	1.2	0.43	0.39	0.69	3.7257368	0.7886832	1.3681403	0.5027059	Xesp-2	embryonic serine protease-2		TC358745	
4	A_10_P000084	2.01	2.09	2.5	1.85	2	1.05	1.19	1.01	0.5	0.22	0.3	1.79	2.05	2.08	2.19	2.17	2.43	2.091036	0.7134385	1.9736458	2.2637699	lk-a	larval keratin		TC397137	
5	A_10_P000170	1.93	1.14	1.74	2.64	2.02	1.14	0.96	2.37	0.95	0.98	1.19	0.83	0.73	0.57	0.53	0.48	0.55	1.8944374	1.2635843	0.7102711	0.5214314	AHR	aryl hydrocarbon receptor		TC357927	
6	A_10_P000252	1.22	1.22	2.51	2.66	2.54	0.44	0.48	0.53	0.35	0.28	0.4	1.52	1.32	1.33	1.21	1.13	1.4	2.0309915	0.4130425	1.3905919	1.2421975	BGP	bone Gla protein			
7	A_10_P000253	1.06	1.1	5.02	4.37	4.63	1.07	0.99	1.06	1.17	0.88	1.05	0.4	0.33	0.55	0.79	0.94	0.84	3.2339485	1.037382	0.4255625	0.8564081	mgp-A	matrix Gla protein		TC371317	
8	A_10_P000267	0.86	0.88	0.4	0.46	0.46	1.67	1.82	1.67	1.47	1.39	1.45	0.87	0.95	0.98	0.52	0.53	0.58	0.612373	1.5769785	0.9329097	0.5441302	LOC398090	TAP2 protein		TC357973	
9	A_10_P000308	1.42	1.54	1.42	1.18	1.17	0.74	0.8	0.75	0.82	0.74	0.72	1.22	1.21	1.22	1.13	1.11	1.14	1.3462446	0.7622302	1.2159824	1.125507	SLBP2	oocyte-specific histone mRNA s		TC365743	
10	A_10_P000358	0.54	0.58	0.58	0.55	0.52	2.79	2.92	3.49	0.57	0.47	0.51	5.85	6.92	5.71	0.41	0.53	0.48	0.5552788	1.7914822	6.1600498	0.475414	im28-a	clone IM28/E3-111 upregulated during intestinal metamorph			
11	A_10_P000373	1.56	1.6	1.72	1.66	1.74	1.05	1.05	1.08	1.07	0.97	0.96	1.11	1.01	1.05	0.49	0.49	0.51	1.6544611	1.0296408	1.0573993	0.4957388	flj20436	hypothetical protein MGC52525		TC374727	
12	A_10_P000440	0.44	0.19	0.33	0.5	0.5	2.27	0.81	0.87	3.04	3.04	3.27	0.79	0.85	0.82	1.24	1.34	1.35	0.3908296	2.2153556	0.8215837	1.3081381	dbx-A	homeodomain protein dbx		TC365931	
13	A_10_P000473	2.23	2.51	7.44	7.06	6.77	0.49	0.57	0.67	1.2	1.17	1.06	0.37	0.39	0.51	0.39	0.51	0.39	5.2008163	0.8588968	0.4215465	0.428668	anf-A	atrial natriuretic factor		TC366806	
14	A_10_P000475	0.75	0.62	0.74	0.86	0.68	1.55	1.25	1.46	1.1	1.19	1.06	1.25	1.17	1.14	0.8	0.8	0.72	0.731925	1.2688618	1.1884609	0.7753875	tcf4	T-cell factor XTCF-4A		TC406625	
15	A_10_P000507	1.29	2.22	1.42	1.44	1.23	0.98	0.76	0.82	0.86	0.78	0.76	1.34	1.35	1.61	0.64	0.75	0.74	1.5212993	0.8258733	1.4324454	0.710258	gcgrp-A	glucagon receptor precursor		TC362012	
16	A_10_P000573	1.29	1.32	2.84	2.39	1.96	0.48	0.45	0.44	0.82	0.92	0.52	0.66	0.73	0.73	1.76	1.53	1.4	1.959324	0.6060286	0.7065479	1.5637813	LOC398311	T-cell receptor gamma		NP460569	
17	A_10_P000657	0.57	0.3	0.43	0.33	0.27	1.57	1.3	1.59	0.3	0.32	0.29	0.68	(										hannel Ka		TC359451	
18	A_10_P000700	0.43	0.29	0.9	1.07	0.81	0.76	1.02	0.65	0.41	0.29	0.52	1.72	1										roxylate aminotransfera		TC373866	
19	A_10_P000705	1.63	1.86	1.51	1.41	1.39	0.95	0.93	0.96	0.62	0.48	0.51	0.71											Na+/K+ transporting, be		TC365137	
20	A_10_P000745	1.8	1.7	2.09	1.77	1.81	1.03	1.01	1.01	1.07	0.83	0.93	0.78	(										rotonin receptor		TC358291	
21	A_10_P001373	1.6	1.91	1.1	1.51	1.58	0.92	1.31	1.37	0.81	0.41	0.87	0.57	(												TC408003	null
22	A_10_P001408	0.56	0.62	0.62	0.44	0.51	2.28	2.47	2.8	2.1	1.98	1.91	0.97	(												TC384897	
23	A_10_P001417	0.77	0.95	1.19	1.03	1.26	1.06	0.98	0.88	0.87	0.87	0.79	0.5	(												TC396959	
24	A_10_P001421	1.06	1.09	0.8	1.03	0.96	0.69	0.69	0.71	0.8	0.89	0.87	1.3	1													
25	A_10_P001529	0.69	0.35	0.78	0.3	0.65	1.39	0.9	1.46	1.47	1.36	1.39	1.32	1													
26	A_10_P001663	0.93	0.98	1.08	1.01	1.02	0.85	0.98	0.79	0.32	0.34	0.37	0.45	(												TC412713	
27	A_10_P001772	3.22	2.52	1.87	1.9	1.79	0.73	0.85	0.73	0.88	0.86	0.87	0.44	(													
28	A_10_P001908	1.36	1.29	1.11	1.49	1.32	1.4	1.31	1.13	1.61	1.48	1.44	0.54	(											null	null	null
29	A_10_P001965	2.23	1.41	1.9	3.75	1.24	0.76	0.94	1.48	0.92	0.69	1.58	0.68	(										like factor 15		TC365184	
30	A_10_P001999	1.82	1.69	2.49	2.81	2.67	0.16	0.22	0.16	0.81	0.82	0.97	0.96	1												NP869135	
31	A_10_P002000	1.58	1.74	7.14	6.79	6.89	0.1	0.12	0.11	0.95	0.98	0.79	0.48	(										r associated with antig		TC357511	
32	A_10_P002104	0.76	0.56	0.61	0.46	0.47	1.41	1.32	1.4	1.01	1.47	1.44	1.36	1										endent kinase inhibitor		TC365585	
33	A_10_P002167	0.66	0.74	0.37	0.39	0.39	1.05	0.91	0.95	1.36	1.16	1.32	1.2	1												TC366704	
34	A_10_P002182	0.87	0.58	0.65	0.85	0.67	1.32	0.78	1.01	0.58	0.45	0.5	2.06	2.36	1.89	2.39	1.76	1.97	0.7251126	0.7727703	2.1053339	2.0407114	serpinc1	serine (or cysteine) proteinase in		TC360723	
35	A_10_P002251	1.1	1.1	1.85	1.82	1.67	0.64	0.68	0.65	0.62	0.63	0.65	1.29	1.27	1.44	1.26	1.24	1.27	1.5093187	0.6467922	1.3339656	1.2567803	MGC52934	similar to apolipoprotein A-I		TC365470	
36	A_10_P002284	2.19	2.08	3.07	2.9	2.98	0.47	0.55	0.51	1.56	1.51	1.55	1.27	1.38	1.37	0.56	0.56	0.62	2.643972	1.0260786	1.3403173	0.5819113	hoxa3a	homeo box A3		TC360883	
37	A_10_P002303	0.4	0.46	0.42	0.41	1.14	0.82	0.36	0.31	1.24	0.37	0.37	2.19	1.44	1.84	11	12.1	10.8	0.5664787	0.5788532	1.8226201	11.304793	fga	Fibrinogen, A alpha polypeptide		TC360342	
38	A_10_P002367	1.62	1.38	1.88	1.85	1.93	0.9	0.72	1.16	1.04	1.17	0.63	0.49	0.66	0.51	0.89	1.15	1.12	1.7344715	0.9360542	0.5529548	1.0512139	foxi1	forkhead box I1		TC362704	
39	A_10_P002398	1.13	1.25	2.3	2.11	2.2	0.81	0.94	0.84	1.41	1.33	1.37	0.92	0.87	0.91	0.53	0.53	0.56	1.7973627	1.1183452	0.8997099	0.5427944	MGC82035	MGC82035 protein			
40	A_10_P002473	0.99	1.16	1.88	1.8	1.87	0.69	0.73	0.75	0.97	0.86	0.94	0.69	0.56	0.67	1.25	1.41	1.35	1.5390202	0.8229748	0.6425872	1.3370467	MGC53012	similar to Fyn proto-oncogene		TC363643	

遺伝子数m

×

サンプル数n

のデータマトリックス

↓

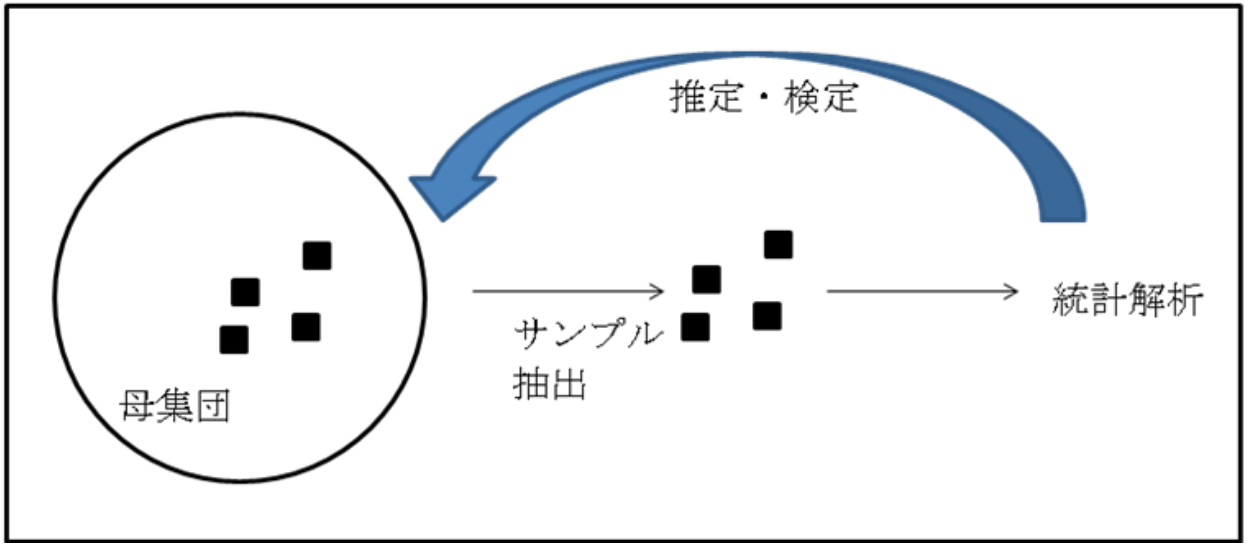
統計解析とプログラムの必要性

統計解析（統計学）とは

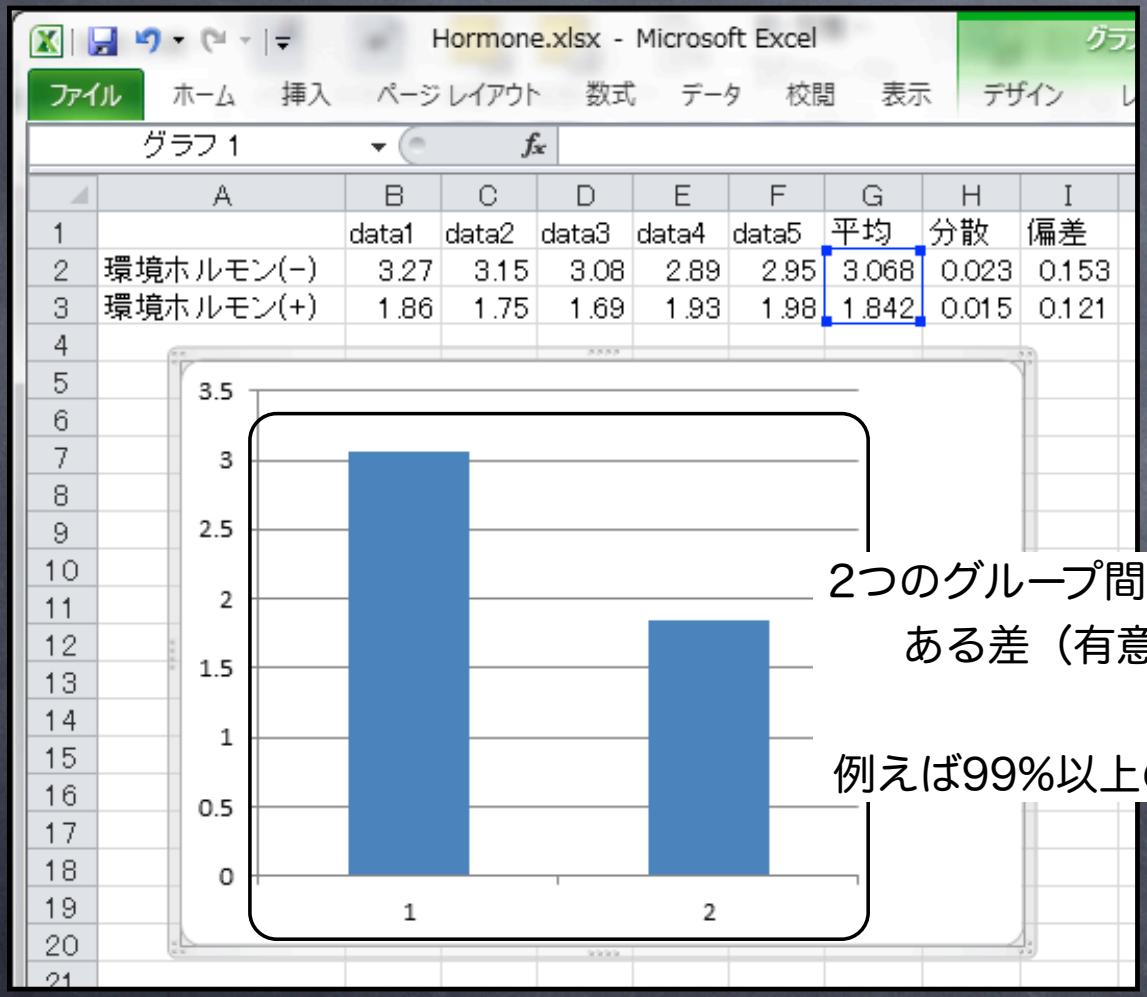
統計学は、経験的に得られたバラツキのあるデータから、応用数学の手法を用いて数値上の性質や規則性あるいは不規則性を見いだす。

統計的手法は、実験計画、データの要約や解釈を行う上での根拠を提供する学問であり、幅広い分野で応用されている。

統計解析（統計学）とは



統計解析（統計学）とは



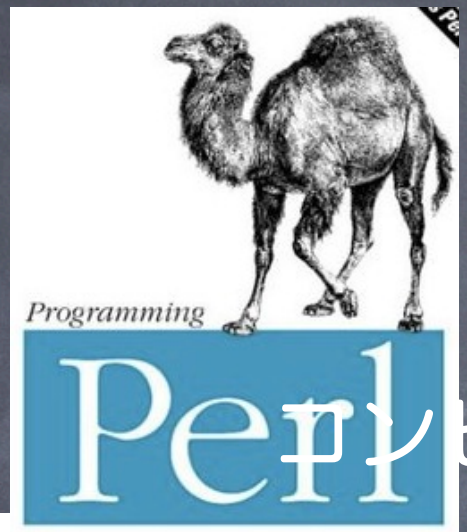
	A	B	C	D	E	F	G	H	I
1		data1	data2	data3	data4	data5	平均	分散	偏差
2	環境ホルモン(−)	3.27	3.15	3.08	2.89	2.95	3.068	0.023	0.153
3	環境ホルモン(+)	1.86	1.75	1.69	1.93	1.98	1.842	0.015	0.121

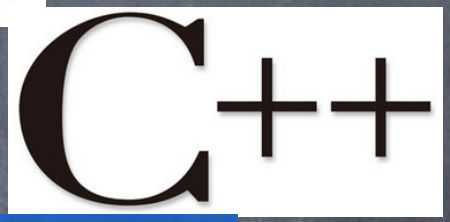
2つのグループ間に、「統計的」に意味のある差（有意差）が存在するか？

例えば99%以上の確率で差がある、等。

プログラム（言語）とは

プログラム（言語）とは





コンピュータ社会における言語

Perl言語とは

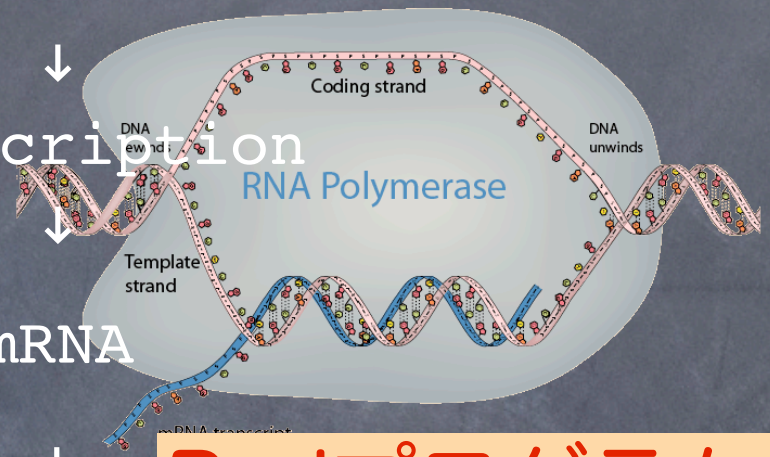
Perl（パール）とは、ラリー・ウォールによって開発されたプログラミング言語である。実用性と多様性を重視しており、Cやsed、awk、シェルスクリプトなど他のプログラミング言語の優れた機能を取り入れている。Webアプリケーション、システム管理、テキスト処理などのプログラムを書くのに広く用いられている。

テキストデータ

DNA: ATGCATCAGTCAGTCAGTACGTAA

↓

Transcription



↓

mRNA

↓

Translat

↓

Perlプログラムの有用性

Amino Acid: MetHisGlnSerValSerThrStop

テキストデータ

Perl言語の実際

```
34 >contig00019 gene=isogroup00001 length=579
35 CTGTACTTGTTACTGTGAGTcATGTATAGGCTGCCTTTGGATGATGTGggTCAAATCTCA
36 GCTGGATACTAGTGGTAAAAGAAATGGACCGTTTCGGAACAAAACAAGTATGTGTAATTT
37 GGCTGGAGCTATTTTTTtCATTAAACTATTTTCTTTTTGTAAGAATTTAATTTTtAAAGA
38 aCAATGGAACTTAGCCATGGTGCATGAAAGGAAAGTTACAGTGTTGTCCATAATAcACAC
39 TTTTtCCTAGCACAGATTATAGTATAAAAaGAATCAAAGGAACAATCTCCCATTTTTTtC
40 TGTTCTGTATTTTTACACATTAGGGTCAGTGATGTAAAGGGCGATTTTAAAATTGTAAAA
41 CTATGCACAAGACTTTGCTGGTTTCTCAGTCTGTTTGCTTTAGTACTTCAATCTGCCAGT
42 TTACTTTAGAAAATTAAATCTGATTTTAGATATTTTAGACACTATTACCCATCATGATGC
43 CCATTTAAGTCTTTTCCAAATTACAAACTTTTAAACTTTACTTGTAGATGTGATATTGCT
44 TGGAGCACTGCTTTACATATGTAGCTAAAGTAATATGTG
45 >contig00020 gene=isogroup00001 length=691
46 AAAAaTGTTCTAGGCATTCATTGATTATACTTTAGTCATGTCAAATTACAGCTTGATGCC
47 TTTAGGCACTTTTGTAACCCTACTATAGTAAATGCTAATATATATATAAAAGCCATTTTT
48 ATGATTAAGCACCTTATTTTTTCAGTTTGCTGTTAGCTAATGTACTGTTTGCTTAGCATA
49 TTACTTTTTTtGTAAAAGTGGATACTGTTGTTTATCTCAAACTAAGCCAGTACAAATTGA
50 GAATACAAACACATCTACACCACTGAGGCAAATGGTAGGTAGAAGAACAGTTGGACAAGT
51 TATGGTAAAATACTCTCTGTTGTTAACAATTTACAAGAAGCTAACAGCTATCAGGGGCTC
52 TATGTTTATTGATGCTAAGCAACGGGATGAATGCTGCTTTAGGTGATTTTCAATTAAAAT
53 TTTCTGAATTTACTTTTATCCTTCAAGCTCTATAGCTGTACCACATTGGTTTATGTAATG
54 AGCAAACATTAGGGGGCAGGTGAAATGAAAAATGAATACTAAATTTTAGAATAAATACAA
55 TACAATACAGTgCACTTACACCCCCTTCCTGCCCAAGCCATTTTTCAGCTTTCAGCTATC
56 ACACTTTGAATGACAATTGcgCAGTCATGCTACACTGTAACCATGTGAAATTTTTaTCaT
57 TTTCTTCATACAAATTGAGCTTTCTTTTGGT
58 >contig00022 gene=isogroup00001 length=1004
59 gTtATtaTATATATATATATATATATATATAtATATATAGAGAGAGAgGAaGAGAGAGAG
60 AGAGAGAGAGAGAGACGATCACTTTAAGAATTAGTAATGCACTTTTTTGCTTCTATTTGG
61 ACACAGAGTTTTCAGTCCATGATGACCTTCACCACCAGATAAGAAGCAATAGTTTACTCC
62 CTTTTTGCTCCAAGAAATGGACTGTCTTATCTAGCTGTTCTGATCTAGTACCGTTATGGC
63 GGCGCTCATGTTTTCTTGCACTTTCTATCAGTCACATGATGTAACCACCTACTTAACTGT
64 TCCCTTGCTACTAAGAAGAAATGAATGCTGATTGGACTAAACTATATGATCAcaCCTATA
65 ACACGCCTGGACCTTATGCATATTCATACTTATAAAATTGTATAGCCTGTAGATGTAACC
66 TCAGAAACACTTTCGCATTACTGCTTTATTGAGAACTGTGTATTGTGTTTCTCCACTGCA
67 GTGTATGGAATAAAAGCACCTAATTTTCTTAAGATTTGGACTCACTGTCTCGTCAGAAAA
68 TGTCTTGTTTGGAGACTCTCTAACCTGATTTAACAATATTTAACTATGTACAGGCAGTTC
```

DNA配列ファイル

＞配列名
配列
＞配列名
配列

while.pl

```
1  #! /usr/bin/perl
2  use strict;
3
4  my ($file) = @ARGV;
5  open (F, $file) or die $!;
6  my $count = 0;
7  while (<F>){
8    chomp;
9    if ($_ =~ /^>/){
10     print "$_\n";
11     $count++;
12   }else{
13     next;
14   }
15 }
16 close F;
17
18 print "Entry count: $count\n";
```

配列名を抽出する
Perlプログラム

バイオインフォマティクス的解析例

１．　祝・山中教授ノーベル賞受賞

～iPS細胞樹立の陰にバイオインフォマティクスあり～

祝・ノーベル賞受賞

平成24年10月8日（月）　産経新聞　号外

産業経済新聞（サンケイ）
THE SANKEI SHIMBUN
発行所　㈱産業経済新聞社東京本社
〒100-8077東京都千代田区大手町1-7-2
☎ 東京(03)3231-7111（大代表）

山中氏にノーベル賞

医学・生理学

iPS細胞を開発

スウェーデンのカロリンスカ研究所は8日、2012年のノーベル医学・生理学賞を、あらゆる細胞に分化する能力があるiPS細胞（人工多能性幹細胞）を開発した京都大の山中伸弥教授（50）と英ケンブリッジ大のジョン・ガードン教授の2氏に授与すると発表した。患者自身の細胞を移植して病気やけがを治療する「再生医療」への道を開いた功績が評価された。

日本人のノーベル賞受賞は2年ぶりで、米国籍の南部陽一郎氏を含め計19人。医学・生理学賞は1987年の利根川進氏以来、2人目の快挙となった。

iPS細胞は、あらゆる細胞に分化する能力を持つ万能細胞の一種。これを基に神経や肝臓、心臓などの細胞を作製し、病気や事故で機能を失った患者の臓器などに移植することで、難治疾患を治療する再生医療の実現が期待されている。

研究はまだ安全性評価の段階で、実際に患者に使われたケースはない。医療分野のノーベル賞は広く実用化した段階で授与されるのが通例で、極めて異例のスピード受賞となった。

山中氏は平成18年、マウスの皮膚細胞に4つの遺伝子を導入することで世界で初めてiPS細胞の作製に成功。19年には同じ手法でヒトの皮膚から作製することにも成功し、世界の第一人者として研究開発をリードしている。

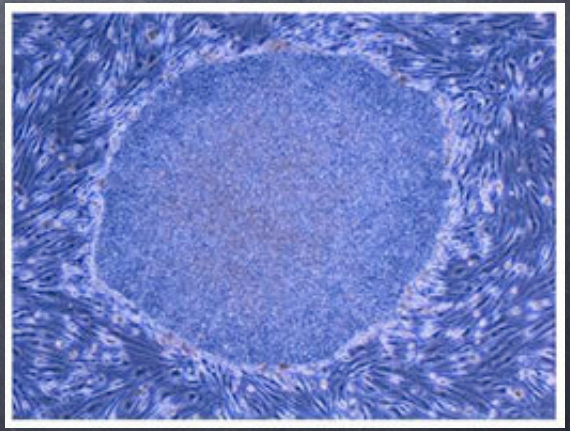
授賞式は12月10日にストックホルムで行われ、賞金計800万スウェーデンクローナ（約9500万円）が贈られる。

産経新聞 購読のお申し込みは 0120-34-4646

MSN産経ニュース　http://sankei.jp.msn.com

分化多能性を有する、人工的な細胞を開発した。

iPS細胞（induced pluripotent stem cell）



多能性幹細胞とは

様々な臓器に分化する能力を持った細胞をさす。
胚性幹細胞（Embryonic Stem Cell, ES cell）が研究の主役だった。動物の発生初期段階である胚盤胞期の胚の一部に属する内部細胞塊より作られる幹細胞細胞株である。

多能性幹細胞とは

多能性細胞の医療への応用：
ES細胞では、倫理的な問題がつきまとうとともに、移植時の拒絶反応が避けられない。

研究の目的：
受精卵を用いず、患者自身の体細胞から人工的にES細胞に匹敵する多能性幹細胞を作り出すこと。

研究上の仮説

ES細胞と体細胞を融合させると、体細胞が多能性を有するようになることが明らかになっていた（ES細胞が多能性誘導因子を有する）。

↓

仮説１：体細胞に多能性を誘導する因子は、ES細胞が多能性を維持するために必要な因子と同じである。
仮説２：ES細胞の多能性を維持している因子は、ES細胞だけで働いており、体細胞では働いていない。

ES細胞の多能性を維持する因子を探すことからスタートした。

因子探索のブレイクスルー

因子探索は暗礁に乗り上げていたらしい。。。

in silico（コンピュータ内）解析、すなわちバイオインフォマティクスの技術を利用した。

ゲノム配列の解読と平行して、cDNAプロジェクトが進行していた。

因子探索のブレイクスルー

ESTとは：Expression Sequence Tagのこと。
cDNAがどんな遺伝子由来か明らかにするために、塩基配列を決定した際の配列。
ライブラリを作成した組織においてmRNAが存在することの証拠となる。

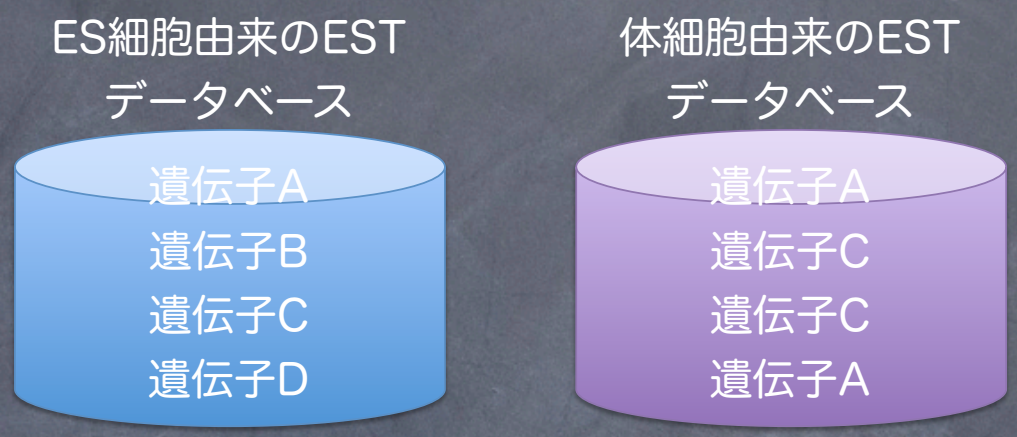
FANTOMとは

Functional Annotatioin of Mammalian Genome

完全長cDNAのアノテーション（機能注釈）を行うことを目的として、2000年に結成された国際研究コンソーシアム

ES細胞由来のESTデータベースと通常の体細胞由来のデータベースを比較し、ES細胞特異的に発現しているcDNA（mRNA）をin silicoで同定した。

FANTOMとは



ES細胞特異的に発現している遺伝子BとD

ES細胞由来のESTデータベースと通常の体細胞由来のデータベースを比較し、ES細胞特異的に発現しているcDNA（mRNA）をin silicoで同定した。

マイクロアレイの利用

EST解析の結果、ある遺伝子がES細胞の多能性維持に重要な役割を果たしていることが分かった。
この遺伝子の制御下にある遺伝子をマイクロアレイを用いて探索し、EST解析の結果得られた重要遺伝子リストに加えられた。

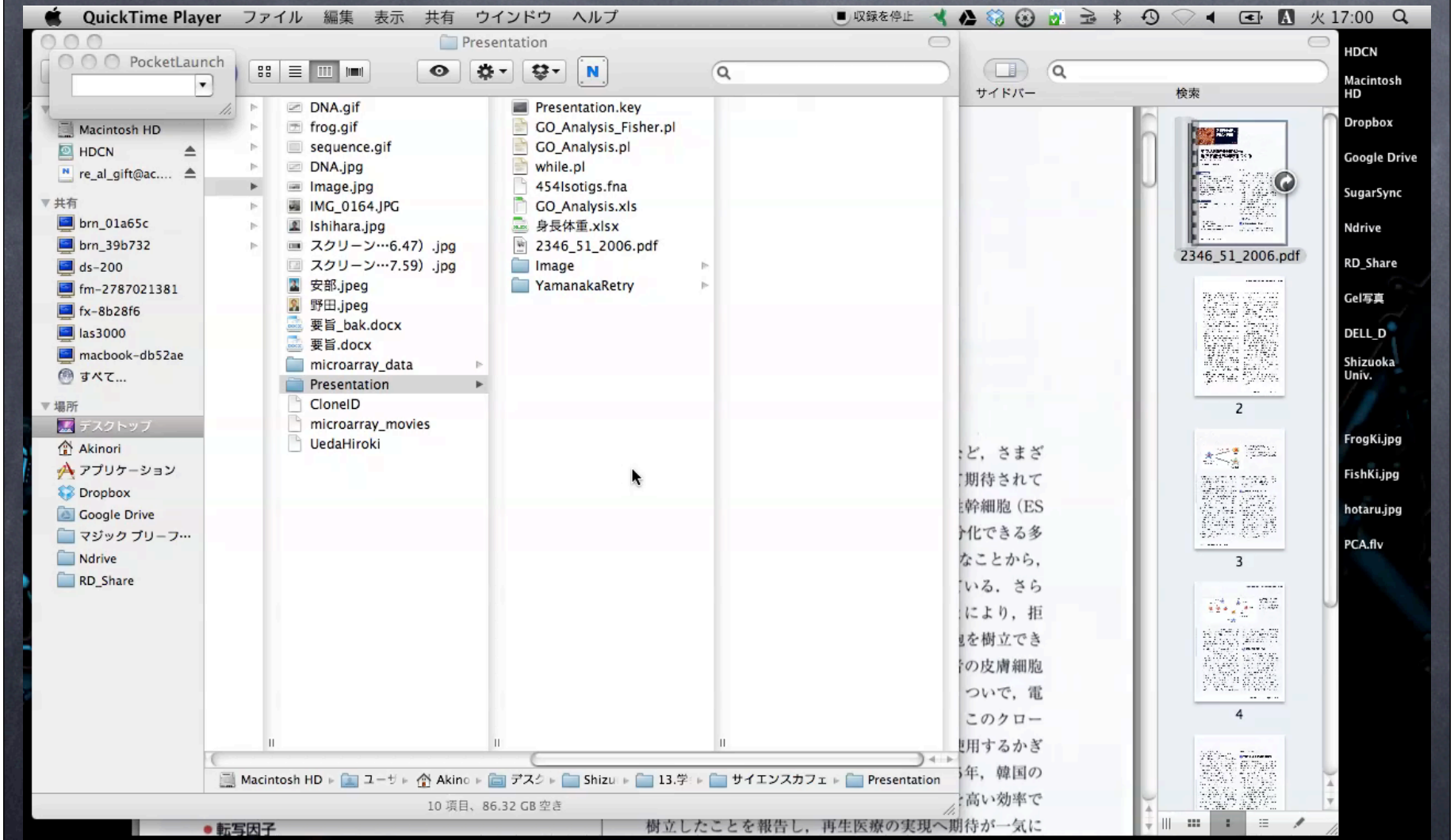
ES細胞の多能性を維持する因子を２４遺伝子に絞り込んだ。

↓

この２４遺伝子の中に、体細胞を分化多能性を有する幹細胞にリセットする因子が存在すると考えられ、最終的に４遺伝子を用いてiPS細胞の樹立に成功した。

因子探索のブレイクスルー

山中先生が行ったin silico解析によるブレイクスルーを再現してみました。



バイオインフォマティクス的解析例

２．　甲状腺系に及ぼす環境ホルモンの影響

The effects of EDCs -- overview

Endocrine disruptors are chemicals that interfere with endocrine (or hormone system) in animals, including humans.
These disruptions can cause cancerous tumors, birth defects, and other developmental disorders.
Specifically, they are known to cause
*learning disabilities 学習障害
*severe attention deficit disorder 注意多動性障害
*cognitive and brain development problems 認識障害、脳の発達の障害
*deformations of the body (including limbs) 身体の奇形
*sexual development problems
*feminizing of males or masculine effects on females, etc.
性ホルモン系への悪影響

Any system in the body controlled by hormones, can be derailed by hormone disruptors. The critical period of development for most organisms is between the transition from a fertilized egg, into a fully formed infant. As the cells begin to grow and differentiate, there are critical balances of hormones and protein changes that must occur. Therefore, a dose of disrupting chemicals can do substantial damage to a developing fetus (baby). The same dose may not significantly affect adult mothers—for a more scientific explanation, see below.

The effects of EDCs -- overview

分類	生物	場所	内分泌撹乱作用	推定原因物質
巻貝類	エゾバイ科 （複数種）	シンガポール、 インドネシア、 マレーシア	雌のインポセックス	船底塗料由来のトリブチルスズの可能性が高い
	イボニシ、 レイシガイ （アクキガイ科）	日本	雌のインポセックス	船底塗料由来のトリブチルスズ及びトリフェニルスズが原因物質
魚類	サケ属	米・五大湖	甲状腺の過形成 （疾病率100%）、 雄の二次性徴欠如、早熟	甲状腺腫誘発物質、未特定なるも多数
	ホワイトサッカー （サッカー科）	米・五大湖 （スペリオル湖）	成熟遅延、生殖巣縮小、 加齢による生殖能の低下、 雄の二次性徴欠如	製紙工場の排水、化合物は未特定なるも多数
	カダヤシ （メダカ目）	米・フロリダ州	雌の雄化（シリビレ伸長）	製紙工場の排水、化合物は未特定なるも多数
	ローチ （コイ科）、 ニジマス	イングランドの河川	精巣発育遅延	ノニルフェノールなどの可能性あるも、排水処理場放水中の女性ホルモンの関与否定できず
爬虫類	ミシシッピーワニ	アポプカ湖 （米・フロリダ州）	雄：ペニス矮小化、精巣機能不全 雌：超雌個体の出現、卵胞での異常卵・多卵、卵巣の退行 幼体：低孵化率、死亡率上昇	湖内に流入したDDT、ディコホルとされるが、データが少なく断定はできない
鳥類	ヤマシギ、ハイタカ、ミサゴ、ミヤマガラス、ヨーロッパヒメウなど	英国 北アメリカ	産卵数減少、繁殖期遅延、産卵失敗、卵の小型化、卵殻薄化、破損卵増加、卵内での死にごもり増加など	野外での使用開始時期と影響の発現との整合性および毒性試験結果より、DDT、DDEと断定
	アメリカオオセグロカモメ	サンタバーバラ島 （米・カリフォルニア州）	雄：雌化、個体数減少 雌：レスビアン個体出現、 生殖器の退行	データの欠如により現段階では特定できない
	セグロカモメ	米・五大湖	甲状腺異常 （肥大、上皮組織過形成など）	含ハロゲン有機化合物（DDT）の可能性もあるが、データ不足のため特定できない
	ハクトウワシ	米・五大湖	低孵化率	PCBまたはDDE（DDT）の可能性あり
	アジサシ、 カモメ	米・五大湖	雄の両性生殖器官具有化	PCB、DDTとの説、知見少なく特定困難
哺乳類	カワウソ、 ミンク	米・五大湖	繁殖激減	餌魚中のPCBの可能性が高い
	フロリダパンサー	米・フロリダ州	雄：精子数減少、潜在精巣症、 雌：不妊	女性ホルモン作用を有する農薬との説があるが、データが少なく特定困難

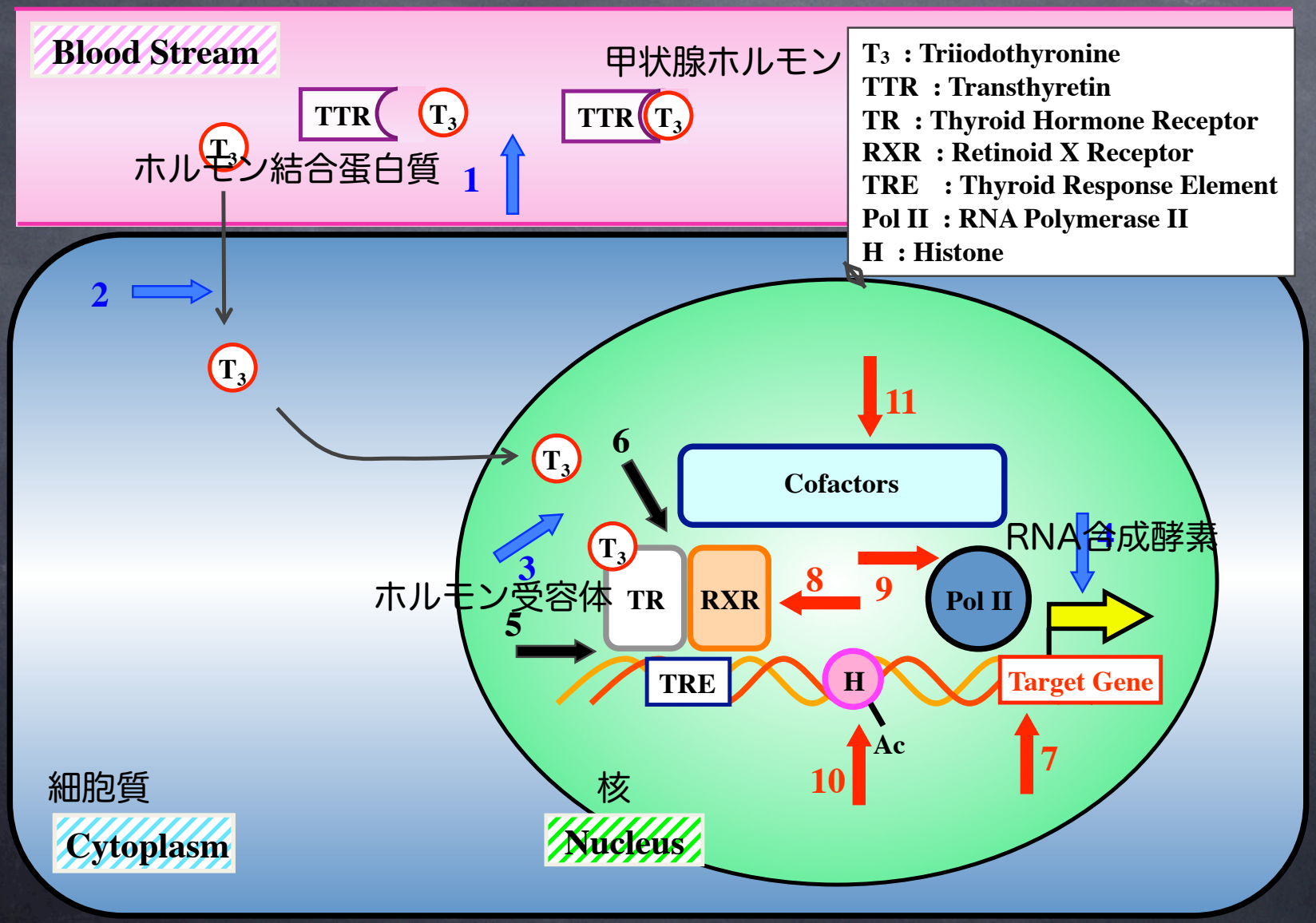
「内分泌（エンドクリン）系に作用する化学物質に関する調査研究」報告書（平成9年6月）

（社）日本化学工業協会、（社）日本化学物質安全・情報センターより

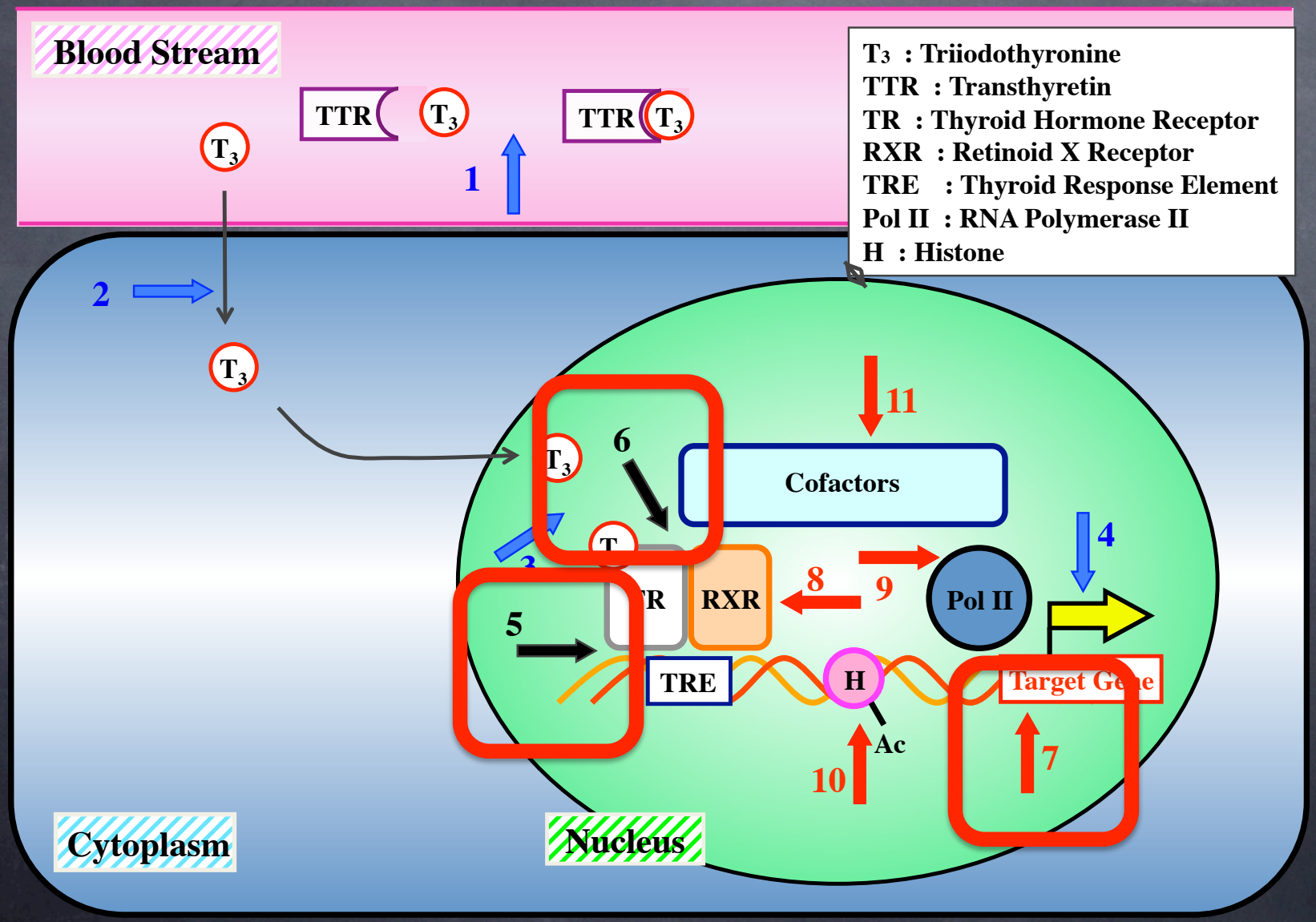
甲状腺ホルモンとは

甲状腺ホルモン (thyroid hormone) とは、甲状腺から分泌され、一般に全身の細胞に作用して細胞の代謝率を上昇させる働きをもつ、アミノ酸誘導体のホルモンのこと。
甲状腺ホルモンは脳や神経の発達に重要な役割を持っている。
甲状腺ホルモンの受容体（受け手）を持たない動物は、多動、学習障害、こだわりなど自閉症に似た症状を示す。

The mechanism of thyroid hormone action and endpoints affected by endocrine-disrupting chemicals.



The mechanism of thyroid hormone action and endpoints affected by endocrine-disrupting chemicals.



T_3 : Triiodothyronine
TTR : Transthyretin
TR : Thyroid Hormone Receptor
RXR : Retinoid X Receptor
TRE : Thyroid Response Element
Pol II : RNA Polymerase II
H : Histone

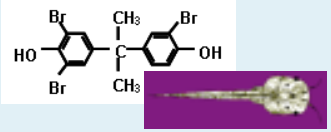
Background and the purpose of experiment.

*Polychlorinated biphenyls (PCBs) have high stability to heat, so they had been widely used for many applications such as dielectric oil, plasticizer, coating material.
*Production of PCBs was banned in the 1970s. However, large amounts of these compounds remain ubiquitous in the environment.
*Hydroxylated PCBs are the metabolites produced from parent compounds by the drug-metabolizing enzyme cytochrome P450. These compounds are suspected to disrupt post-embryonic neural development in the brains of mammals including humans.
*Because thyroid hormones play a role in the perinatal development of the central nervous system, the thyroid system is thought to be one of the important endpoints affected by several endocrine-disrupting chemicals including PCBs. Although it is important to study the effects of hydroxylated PCBs on the brain and thyroid system *in vivo*, studies with mammalian species are fraught with difficulty because each embryo is enclosed by the uterus 周産期ほ乳類のモデルとして両生類を用いた
*Amphibian metamorphosis is one of the most dramatic TH-induced changes, and it shares many similarities with mammalian perinatal development. It could also be a

両生類の変態は、甲状腺ホルモンによって引き起こされる現象であり、甲状腺ホルモン作用を撹乱するかどうかの良いモデル系である

*To clarify the effects of these compounds on thyroid hormone function in the brain, we used metamorphosing tadpoles of the African clawed toad as a model for mammalian post-embryonic development.

Experimental Scheme

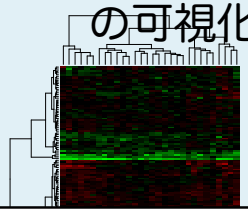


Chemical Exposure 化学物質の曝露

↓ マイクロアレイ

Genome-wide Analysis of Chemical Effects on Gene Expression

マイクロアレイデータの可視化、解釈

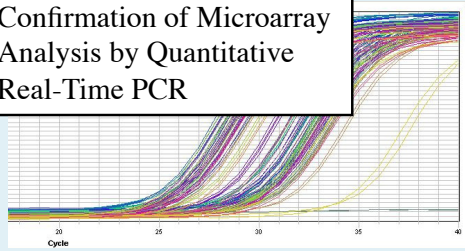


Computational Analysis
*Clustering
*PCA
*GO Enrichment Analysis

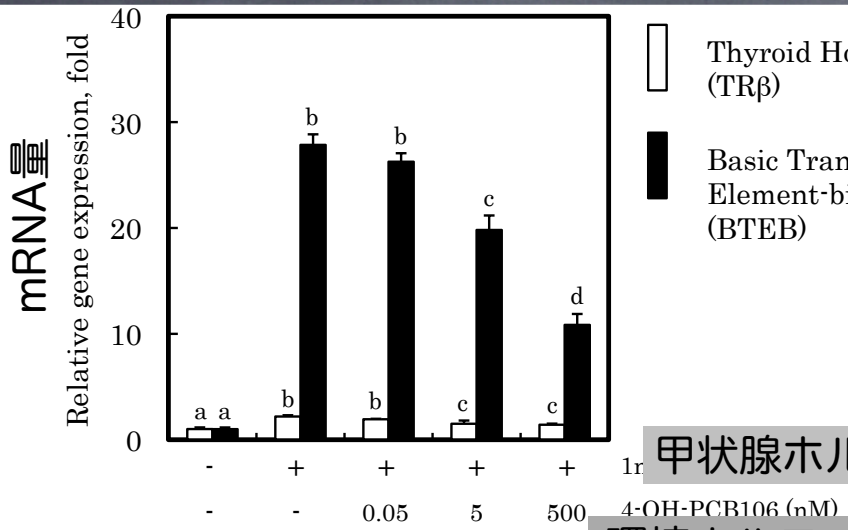
Selection of the Genes Whose Expression is Significantly Altered

マイクロアレイ結果から、
有用な遺伝子群を抽出

Confirmation of Microarray Analysis by Quantitative Real-Time PCR

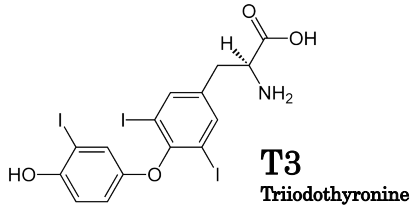


Effect of PCBs on T_3-induced TRβ and BTEB gene expression in the tadpole brain undergoing metamorphosis.

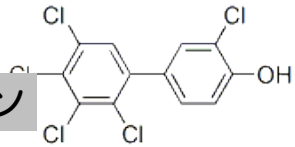


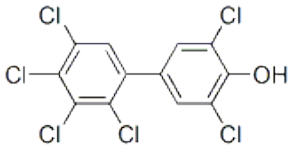
Thyroid Hormone Receptor β (TRβ)

Basic Transcription Element-binding Protein (BTEB)



甲状腺ホルモン

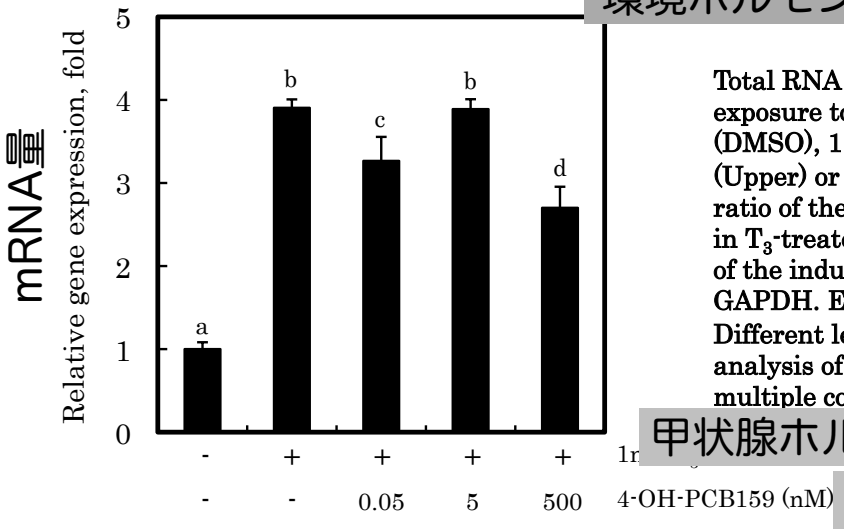




環境ホルモン1

4-OH-PCB106
4-HYDROXY-2',3,3',4',5'-PENTACHLOROBIPHENYL

4-OH-PCB159
4-HYDROXY-2',3,3',4',5,5'-HEXACHLOROBIPHENYL

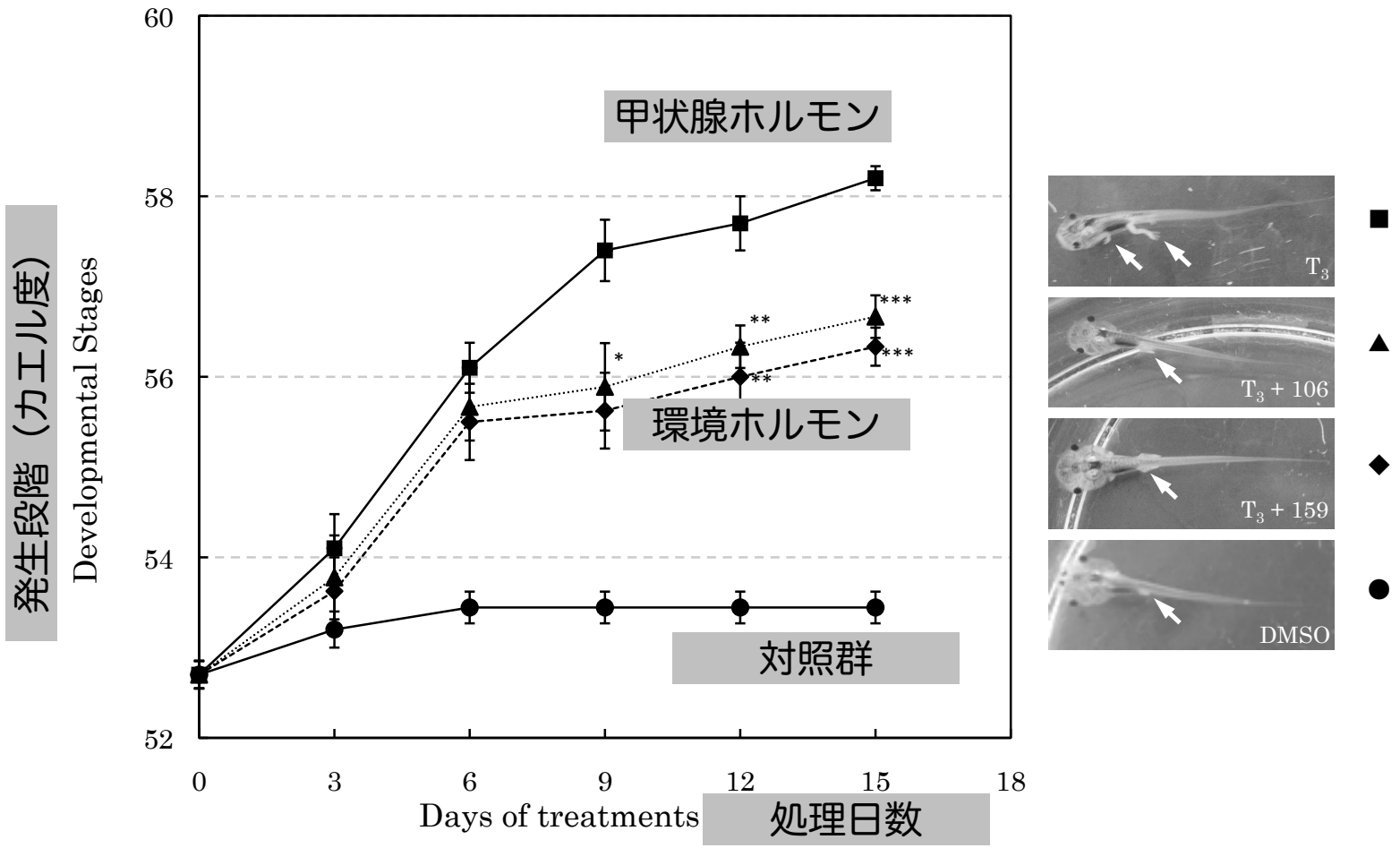


Total RNA was extracted from the tadpoles after short-term (96-h) exposure to hydroxylated PCBs. They were treated with vehicle control (DMSO), 1 nM T_3, 1 nM T_3 and 0.05, 5, or 500 nM of 4-OH-PCB106 (Upper) or 4-OH-PCB159 (Lower). The vertical axis represents the ratio of the amount of TRβ (open bar) and BTEB (solid bar) transcripts in T_3-treated samples to those in T_3-untreated samples as a magnitude of the induction (fold) after normalization with the housekeeping gene GAPDH. Each value is the mean ± S.E. of triplicate determinations. Different letters denote significantly different means ($p < 0.05$; 1-way analysis of variance by Fisher's least significant difference test for multiple comparisons).

甲状腺ホルモン

環境ホルモン2

Hydroxylated PCBs inhibit T_3-induced metamorphosis.



Time-dependent inhibition of T_3-induced metamorphosis by hydroxylated PCBs. The average developmental stages of the tadpoles treated with vehicle control (DMSO; solid circle), 1 nM T_3 (solid square), 1 nM T_3 and 500 nM 4-OH-PCB106 (solid triangle), 1 nM T_3 and 500 nM 4-OH-PCB159 (solid diamond) were plotted every third day. Each value is the mean ± S.E. of surviving animals. An asterisk indicates significant differences in the stages between T_3 and T_3 + hydroxylated PCBs groups (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). Typical images of tadpoles from each group after the experiment are represented in the figure. White arrows indicate limbs.

遺伝子の抽出

甲状腺ホルモンの作用によって、mRNA量が変動し、かつ環境ホルモンによって、そのホルモン作用が撹乱される遺伝子群を抽出したい。
統計解析によって、下記の組み合わせで「有意に」差がある遺伝子群を抽出した。

・対照群vsホルモン処理群
・ホルモン処理群vsホルモン＋環境ホルモン１処理群
・ホルモン処理群vsホルモン＋環境ホルモン２処理群

３つの比較で得られた遺伝子群で、重複するものを抽出した。

The results of genome-wide expression analysis. PartI

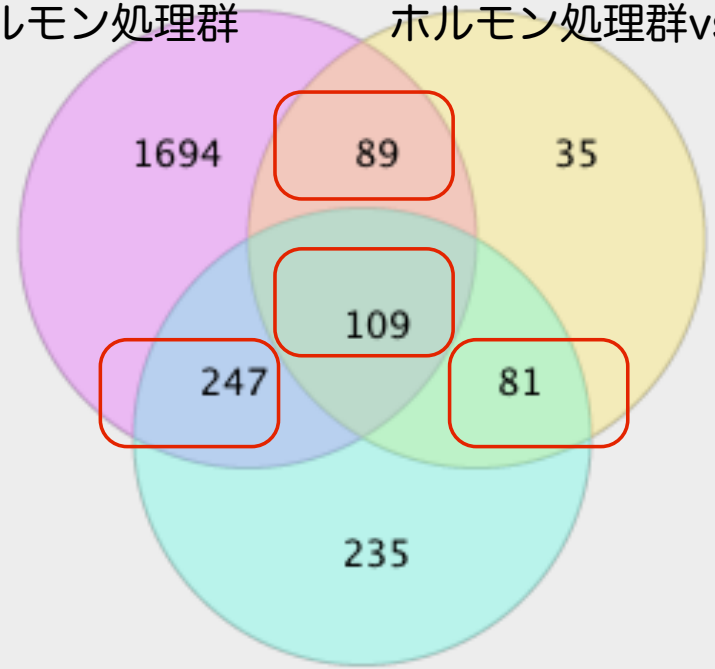
Venn diagrams showing the number of genes altered by T3 or a combination of T3 and hydroxylated PCBs.

DMSO vs T_3
1.5 fold
2139 genes

対照群vsホルモン処理群

T_3 vs T_3 + 4-OH-PCB159
1.5 fold
314 genes

ホルモン処理群vsホルモン＋環境ホルモン１処理群



ANOVA: $p < 0.001$
Total 2490 genes

T_3 vs T_3 + 4-OH-PCB106
1.5 fold
672 genes

ホルモン処理群vsホルモン＋環境ホルモン２処理群

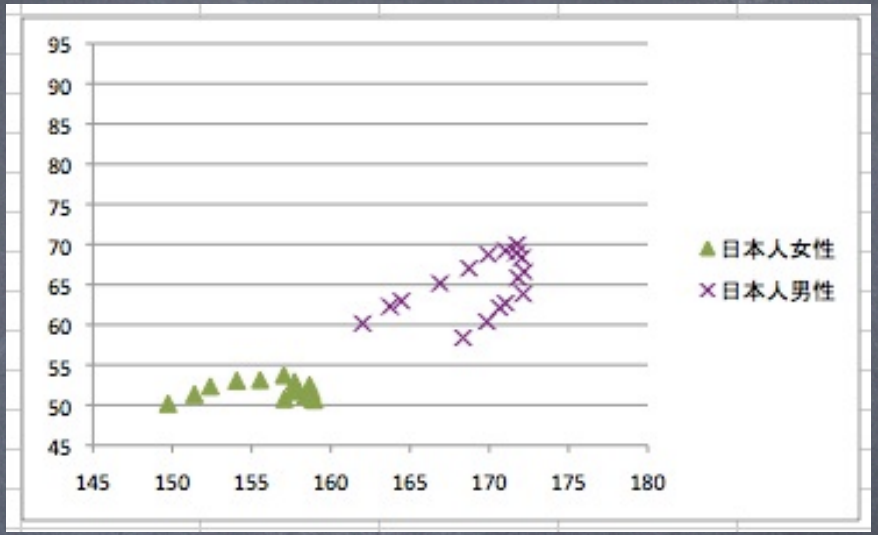
A total of 2139 genes were altered in response to T_3. Among these genes, 356 (247 + 109) or 198 (89 + 109) genes were altered in the presence of 4-OH-PCB106 or 4-OH-PCB159, respectively. About 445 (21%) of the 2139 genes were altered in the presence of either of the hydroxylated PCB. A total of 796 (89 + 35 + 109 + 81 + 247 + 235) genes were altered in the presence of hydroxylated PCBs, while 24% of the 796 genes were commonly altered, and 526 (89 + 109 + 247 + 81) genes were selected as good parameters to distinguish each treatment group.

抽出した遺伝子の妥当性評価

データを可視化することによって妥当性を評価したい。
抽出した遺伝子は526、サンプル数は17。
→526次元空間に17個のプロット。？？？？？？？

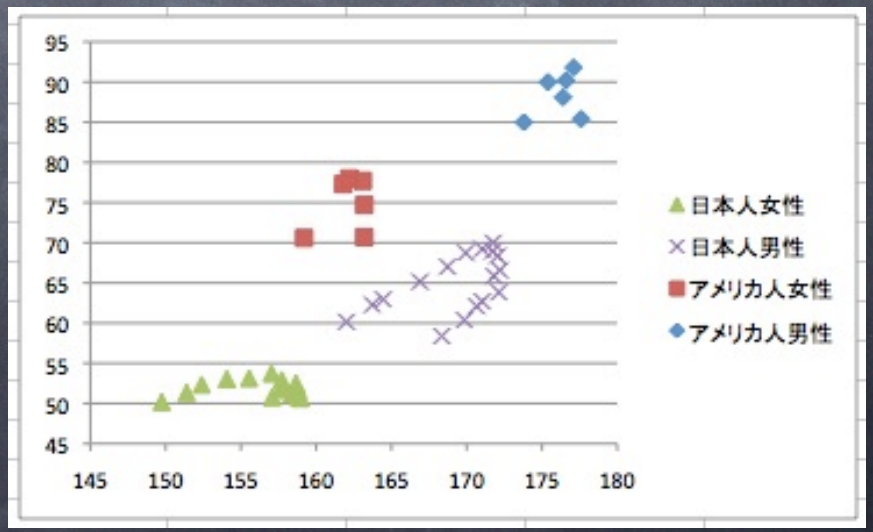
パラメータとサンプルの関係

	A	B	C	D	E
1	日本人	男性		女性	
2		身長	体重	身長	体重
3	15歳	168.35	58.38	157.08	50.77
4	16歳	169.85	60.38	157.62	51.85
5	17歳	170.66	62.1	157.79	52.1
6	18歳	171.03	62.69	157.94	51.7
7	19歳	172.14	63.83	158.41	51.19
8	20～24歳	171.82	65.84	158.84	50.81
9	25～29歳	172.2	66.58	158.96	50.64
10	30～34歳	172.06	68.35	158.6	51.2
11	35～39歳	171.73	68.93	158.81	51.64
12	40～44歳	171.76	70	158.66	52.5
13	45～49歳	171.03	69.21	157.71	52.93
14	50～54歳	169.94	68.73	157.04	53.73
15	55～59歳	168.73	67.01	155.55	53.14
16	60～64歳	166.9	65.15	154.07	53.04
17	65～69歳	164.46	62.95	152.4	52.34
18	70～74歳	163.74	62.27	151.4	51.34
19	75～79歳	162.01	60.16	149.74	50.18
20					



似た性質のデータ集団は空間上の近い位置にプロットされ、クラスタを形成する。

21	アメリカ人	男性		女性	
22		身長	体重	身長	体重
23	20-29 years	177.6	85.4	163.2	70.7
24	30-39 years	176.4	88.1	163.2	74.7
25	40-49 years	177.1	91.8	163.1	77.7
26	50-59 years	176.6	90.2	162.2	78
27	60-69 years	175.4	90	161.8	77.3
28	70-79 years	173.8	85	159.2	70.6



抽出した遺伝子の妥当性評価

データを可視化することによって妥当性を評価したい。
抽出した遺伝子は526、サンプル数は17。
→526次元空間に17個のプロット。？？？？？？？
→身長と体重の例では、２次元空間。

次元を減らすしか無い。
526次元を3次元に。

主成分分析

主成分分析（しゅせいぶんぶんせき、principal component analysis）は、直行回転を用いて変数間に相関がある元の観測値を、相関の無い主成分とよばれる値に変換するための数学的な手続きのことである。

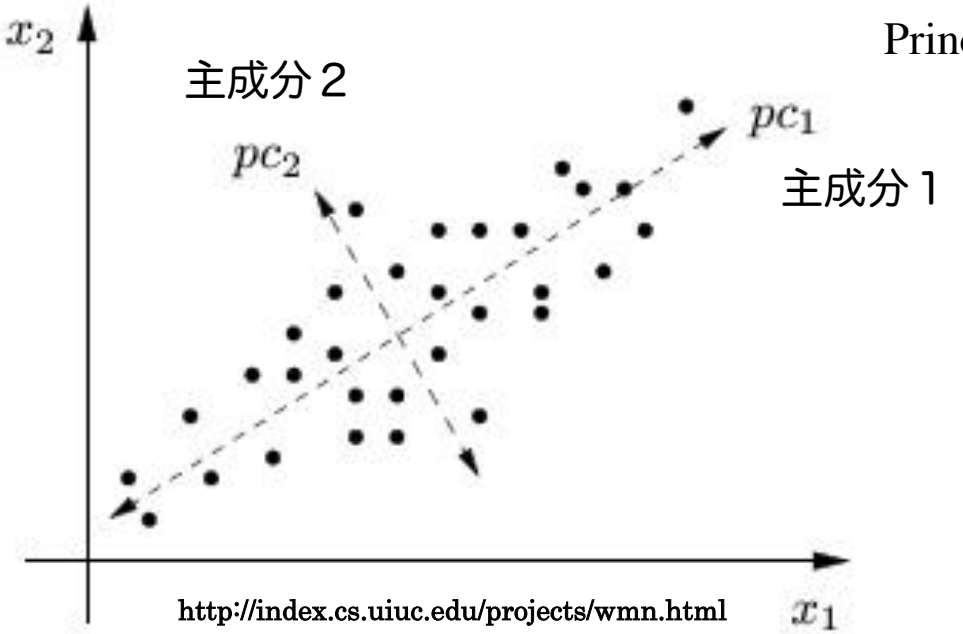
The results of genome-wide expression analysis. PartII

Principal component analysis of gene expression in all the samples by using the significantly selected gene set as analyzing parameter.

Microarray data (m transcripts and n samples)
---> m dimensional n plots.

It's difficult to visualize such high dimensional data…

Principal Component Analysis!

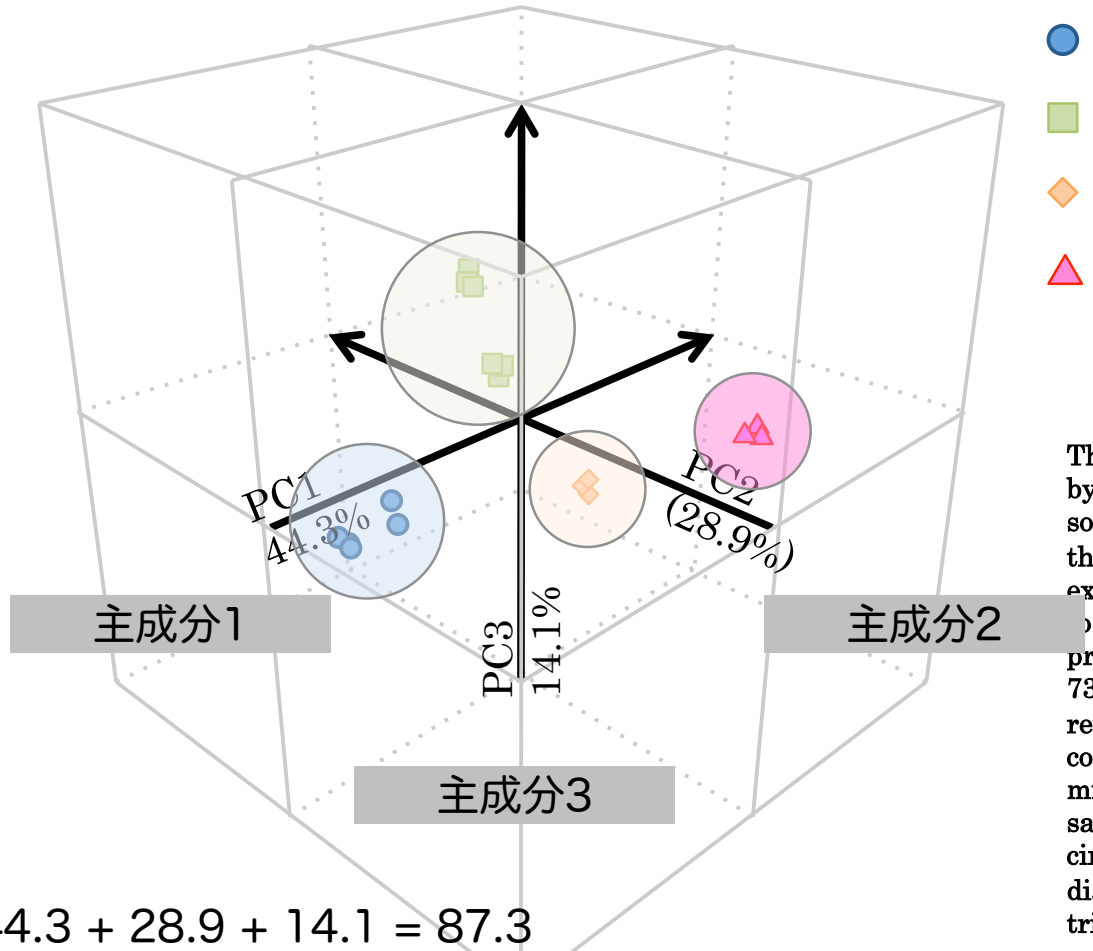


http://index.cs.uiuc.edu/projects/wmn.html

主成分に置き換えることによって、データ情報量の節約ができる。

The results of genome-wide expression analysis. PartII

Principal component analysis of gene expression in all the samples by using the significantly selected gene set as analyzing parameter.



- DMSO
- T_3
- T_3 + 4-OH-PCB159
- T_3 + 4-OH-PCB106

The microarray data were analyzed by PCA by using Subio platform software with the basic plug-in. The three-dimensional plot view of gene expression data is shown with respect to their correlation to the first 3 principal components. Approximately 73.2% variation in the samples was revealed in the first 2 principal components. Each sample used for microarray was plotted with the same symbol for each group (solid circle, DMSO; solid square, T_3; solid diamond, T_3 + 4-OH-PCB159; solid triangle, T_3 + 4-OH-PCB106).

44.3 + 28.9 + 14.1 = 87.3

------> 526次元の情報の87.3%を3次元で表現できたことを意味している。

Gene Ontology (GO) enrichment analysis.

- GO:0008150 biological_process [367271 gene products]
  - **GO:0008283 cell proliferation [5198 gene products]** 細胞増殖
    - GO:0003263 cardioblast proliferation [5 gene products] 心臓芽細胞増殖
    - GO:0071838 cell proliferation in bone marrow [5 gene products]
    - GO:0003295 cell proliferation involved in atrial ventricular junction remodeling [0 gene produc
    - GO:0035736 cell proliferation involved in compound eye morphogenesis [1 gene products]
    - GO:0060722 cell proliferation involved in embryonic placenta development [0 gene products]
    - GO:0061323 cell proliferation involved in heart morphogenesis [9 gene products]
    - GO:0090255 cell proliferation involved in imaginal disc-derived wing morphogenesis [0 gene
    - GO:0072111 cell proliferation involved in kidney development [45 gene products]
    - GO:0035726 common myeloid progenitor cell proliferation [4 gene products]
    - GO:0050673 epithelial cell proliferation [794 gene products]
    - GO:0070341 fat cell proliferation [13 gene products] 脂肪細胞増殖
    - GO:0048144 fibroblast proliferation [277 gene products]
    - GO:0048134 germ-line cyst formation [32 gene products]
    - GO:0014009 glial cell proliferation [48 gene products]
    - GO:0003419 growth plate cartilage chondrocyte proliferation [9 gene products]

For example…

GO000000　あるGO000000が、、、
Total 22000 transcript:　10 c 調べた全遺伝子のうち１０個
Altered 500 transcript:　8 co 絞り込んだ500遺伝子のうち８個
This “GO” is significantly enriched in altered transcripts.
(Fisher’s exact test)
　絞り込んだ500遺伝子に、有意に濃縮されている。

What is “Gene Ontology” ?

オントロジー　＝　語彙

The Gene Ontology project is a major bioinformatics initiative

GeneOntologyは、遺伝子、遺伝子産物であるタンパク質の名前と、それぞれの機能に関する語を関連づけているデータベースである。各機能に特有のIDを割り振っている。階層構造を持っている。

characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process this data.

Table X. The list of Gene Ontology IDs that were associated with selected 526 genes and were significantly enriched in Biological Process.

GO ID	Parent GO ID	Definition	
GO:0040007	GO:0008150	growth	
GO:0008285	GO:0008283	negative regulation of cell proliferation	0.011397899
GO:0042127	GO:0050794	regulation of cell proliferation	9.03E-05
GO:0016477	GO:0048870	cell migration	0.013632849
GO:0030902	GO:0048856	hindbrain development	
GO:0030901	GO:0048856	midbrain development	
GO:0031641	GO:0032844	regulation of myelination	0.001761645
GO:0030155	GO:0050794	regulation of cell adhesion	0.02085799
GO:0050673	GO:0008283	epithelial cell proliferation	0.029762641
GO:0030856	GO:0051239	[illegible]	0.037778386
GO:0045011	GO:0022607	[illegible]	0.046513827
GO:0009825	GO:0016049	multidimensional cell growth	0.027673394
GO:0009792	GO:0009790	embryonic development ending in birth or egg hatching	0.011691047
GO:0030182	GO:0030154	neuron differentiation	0.044955394
GO:0048333	GO:0030154	mesodermal cell differentiation	[illegible]
GO:0002053	GO:0010463	positive regulation of mesenchymal cell [illegible]	
GO:0045766	GO:0045765	positive regulation of angiogenesis	[illegible]
GO:0016339	GO:0016337	calcium-dependent cell-cell adhesion	0.031695492
GO:0042478	GO:0022603	regulation of eye photoreceptor cell development	0.027673394
GO:0043065	GO:0006915	positive regulation of apoptosis	0.011268777
GO:0045579	[illegible]	[illegible]tiation	0.005448753
GO:0006919	[illegible]	[illegible]	0.005620347
GO:0042664	GO:0001714	negative regulation of endodermal cell fate specification	1.08E-03
GO:0045843	GO:0048635	negative regulation of striated muscle development	0.013323335
GO:0048562	GO:0048598	embryonic organ morphogenesis	0.005368416
GO:0051489	GO:0060491	regulation of fi[illegible]	68416
GO:0048335	GO:0045596	negative regula[illegible]	10723
GO:0048925	GO:0048880	lateral line system development	0.01853492
GO:0048514	GO:0009887	blood vessel morphogenesis	0.000772626
GO:0030326	GO:0035108	embryonic limb morphogenesis	0.0388314
GO:0007386	GO:0009952	compartment specification	0.002334391
GO:0050770	GO:0010769	regulation of axonogenesis	0.008384524
GO:0035284	GO:0035282	brain segmentation	8.38E-03
GO:0021545	GO:0021675	cranial nerve development	0.01853492
GO:0010090	GO:0000904	trichome morphogenesis	0.027673394
GO:0030282	GO:0031214	bone mineralization	0.001237237
GO:0051932	GO:0007270	synaptic transmission, GABAergic	0.02085799
GO:0007435	GO:0007431	salivary gland morphogenesis	0.039897319
GO:0007387	GO:0007386	anterior compartment specification	0.036726938
GO:0007388	GO:0007386	posterior compartment specification	0.036726938
GO:0001754	GO:0048592	eye photoreceptor cell differentiation	0.006305561

成長

細胞増殖

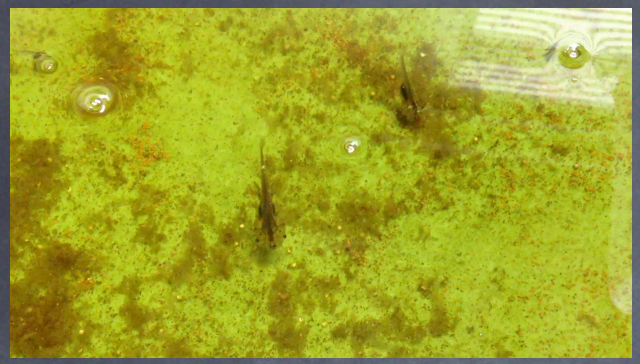
細胞移動

脳の発達

細胞接着

細胞死

結果から得られた現象の概要は？



甲状腺ホルモン処理によって変態が進行する。

・幼生型の細胞の細胞死

・成体型の細胞の増殖、移動、接着

・脳の発達

このような現象が環境ホルモンで阻害されることによって、変態が抑制されている。

まとめ

膨大な情報になりつつある実験データを、人間の目、脳で解釈することは不可能になりつつある。
情報学、数学など異分野と思われる技術、知識を駆使することで、データマイニングすることが重要である。

実験的に行き詰まっても、コンピュータが助けてくれるブレイクスルーが期待できるケースがある。

今後、ますますバイオインフォマティクスの技術は普及していくでしょう。