トランスクリプトーム解析におけるバイオインフォマティクス要素技術～私の相場観～

東京大学大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究ユニット
門田 幸二（かどた　こうじ）
http://www.iu.a.u-tokyo.ac.jp/~kadota/
kadota@iu.a.u-tokyo.ac.jp

門田の部分は

- バイオインフォマティクス人材育成講座
  - 7月：**マイクロアレイ**受託実験や遺伝子発現データ解析
  - 8月：**次世代シーケンサー**見学
  - 8月：**統計解析言語R**や遺伝子発現データの**クラスタリング**

  } このあたりの補講

- 第一部のねらい
  - クラスタリング（やエントロピー）などのバイオインフォマティクスの**基本的なスキルを身につけるだけで様々な局面に応用可能である**という二つの事例を紹介するとともに、これらの具体的な計算例を示すことで数式アレルギー緩和に貢献

- 第二部のねらい
  - 次世代シーケンサーのデータだって、**統計解析言語Rでお手軽に解析**できる
  - **遺伝子発現行列**にしたら後は同じ
  - **高度なプログラミング能力がなくても**バイオインフォマティクスの世界で生存可能
  - 必要な情報はインターネットのみで十分（＠沖縄）

自己紹介

- 1995年3月
  - 高知工業高等専門学校・工業化学科　卒業
- 1997年3月
  - 東京農工大学・工学部・物質生物工学科　卒業
- 1999年3月
  - 東京農工大学・大学院工学研究科・物質生物工学専攻　修士課程修了
- 2002年3月
  - 東京大学・大学院農学生命科学研究科・応用生命工学専攻　博士課程修了
  - 学位論文:「cDNAマイクロアレイを用いた遺伝子発現解析手法の開発」(指導教官:清水謙多郎教授)
- 2002/4/1~
  - 産総研・生命情報科学研究センター　産総研特別研究員
- 2003/11/1~
  - 放医研・先端遺伝子発現研究センター　研究員
- 2005/2/16~
  - 東京大学・大学院農学生命科学研究科　特任助手
- 2007/4/1~現在
  - 東京大学・大学院農学生命科学研究科　特任助教

アグリバイオインフォマティクスプログラム

高専時代の成績もたいしたことない門田が、かれこれ10年以上バイオインフォマティクスの分野で楽しくやってます。

トランスクリプトームとは

- ある特定の状態の組織や細胞中に存在する全RNA（転写物、transcripts）の総体
- 様々なトランスクリプトーム解析技術
  - マイクロアレイ
    - cDNAマイクロアレイ、Affymetrix GeneChip、タイリングアレイなど
  - 配列決定に基づく方法
    - EST、SAGEなど、次世代シーケンサー
  - 電気泳動に基づく方法
    - Differential Display、AFLPなど

調べたい組織でどの遺伝子がどの程度発現しているのかを一度に観察

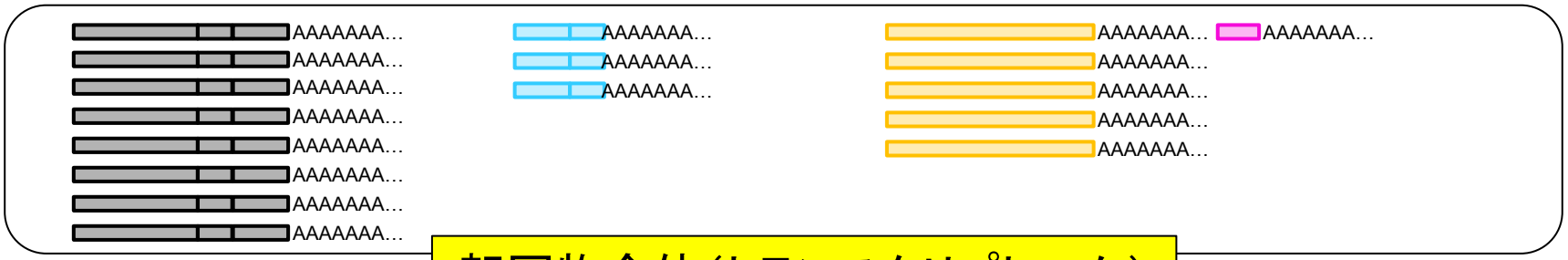
トランスクリプトームとは

- ある状態のあるサンプル（例：目）のあるゲノムの領域

ヒト

遺伝子全体（ゲノム）

・どの染色体上のどの領域にどの遺伝子があるかは調べる個体（例：ヒト）が同じなら不変（目だろうが心臓だろうが…）



転写物全体（トランスクリプトーム）

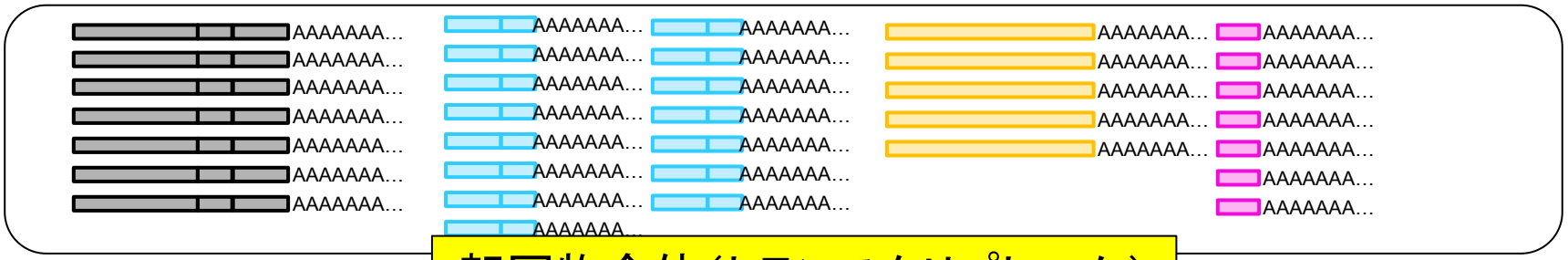
・遺伝子1は沢山転写されている（発現している）
・遺伝子4はごくわずかしか転写されてない
・…

トランスクリプトームとは

- ある状態のあるサンプル（例：目）のあるゲノムの領域

遺伝子全体（ゲノム）

・どの染色体上のどの領域にどの遺伝子があるかは調べる個体（例：ヒト）が同じなら不変（目だろうが心臓だろうが…）

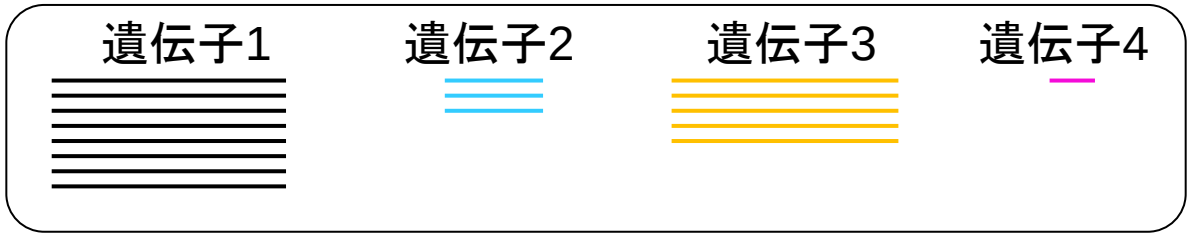


転写物全体（トランスクリプトーム）

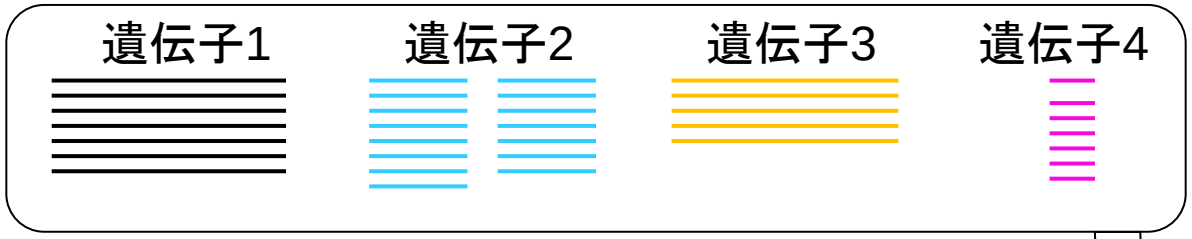
・遺伝子2は光刺激に応答して発現亢進
・遺伝子4も光刺激に応答して発現亢進

トランスクリプトーム情報を得る手段

- 光刺激前（T1）の目のトランスクリプトーム



- 光刺激後（T2）の目のトランスクリプトーム



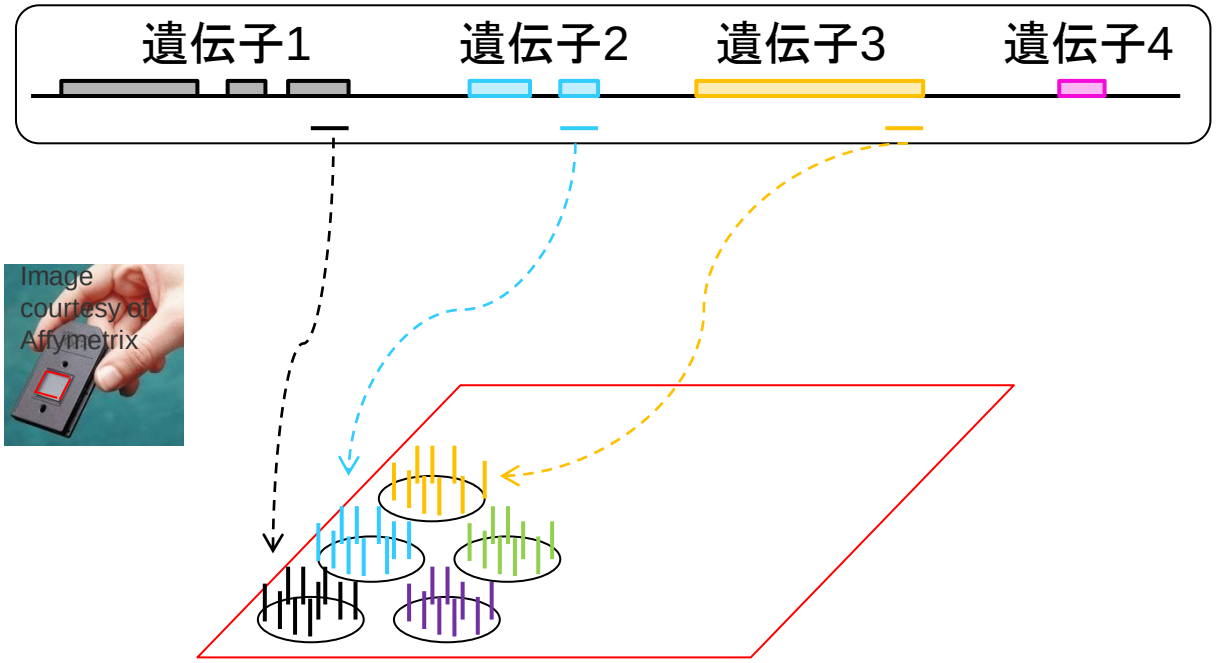
これがいわゆる
「遺伝子発現行列」

	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
…	…	…

- マイクロアレイ
- 電気泳動に基づく方法
- 配列決定に基づく方法

トランスクリプトーム取得（マイクロアレイ）

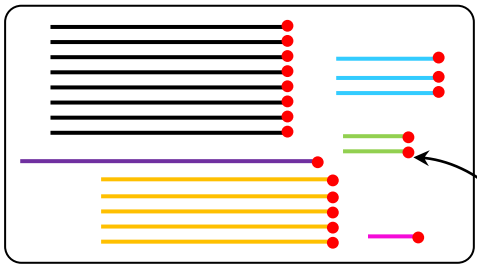
- よく研究されている生き物は多数の遺伝子（の配列情報）がわかっている



わかっている遺伝子（の配列の相補鎖）を搭載した”チップ”

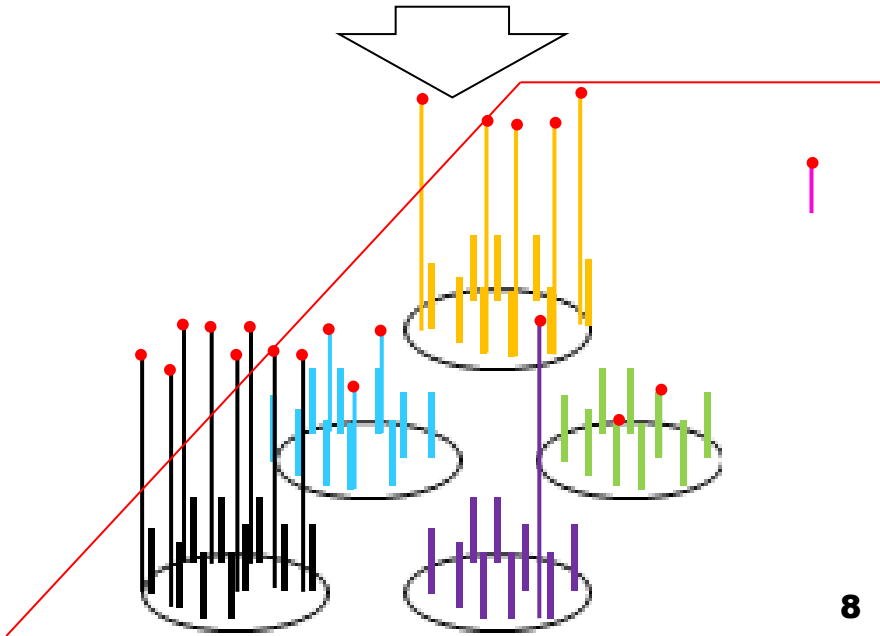
・メーカーによって搭載されている遺伝子の種類が異なる
→搭載されていない遺伝子（未知遺伝子含む、例：遺伝子4）の発現情報は測定不可…

光刺激前（T1）の目のトランスクリプトーム



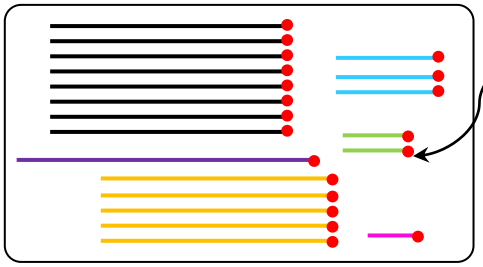
蛍光標識

ハイブリダイゼーション（二本鎖形成）

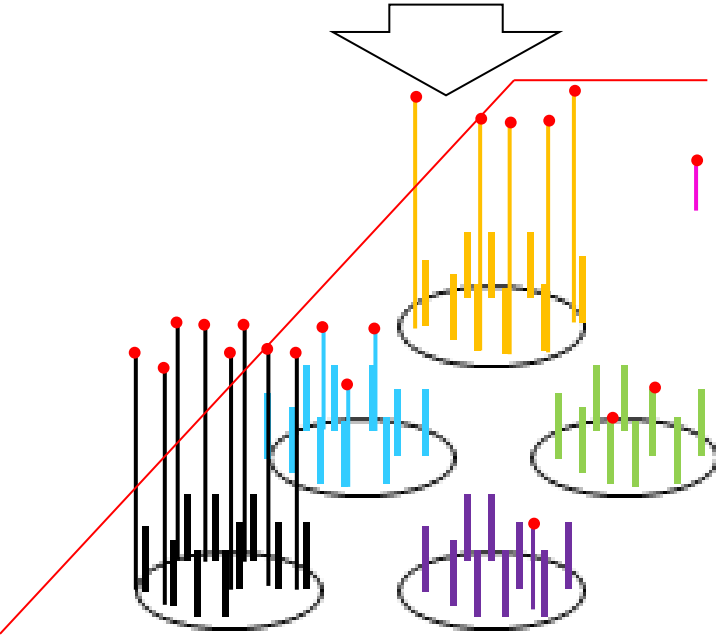


トランスクリプトーム取得（マイクロアレイ）

- 光刺激前（T1）の目のトランスクリプトーム

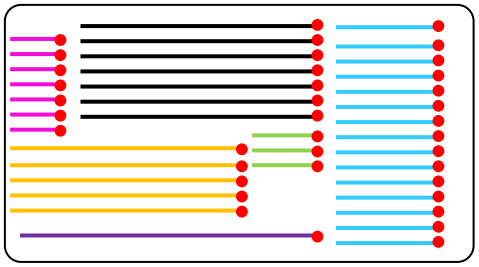


ハイブリダイゼーション
（二本鎖形成）



専用の検出器で各遺伝子に対応する領域の蛍光シグナル強度を測定

光刺激後（T2）の目の
トランスクリプトーム



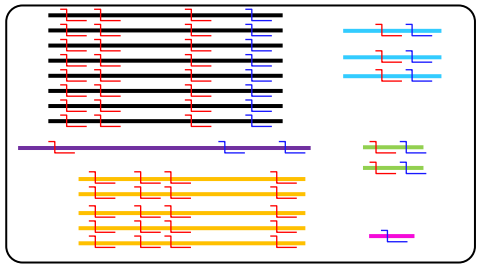
ハイブリダイゼーション
と
シグナル検出

	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	?	?
遺伝子5	…	…
…	…	…

トランスクリプトーム取得（電気泳動）

- cDNA-AFLP（HiCEPの場合）

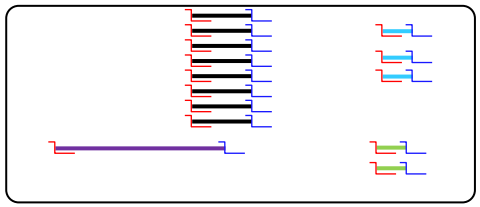
様々な制限酵素認識部位をもつトランスクリプトーム



二種類の制限酵素で断片化

MspI　　MseI

NCCGGTT [■■■■■■■■] GTTTAAN



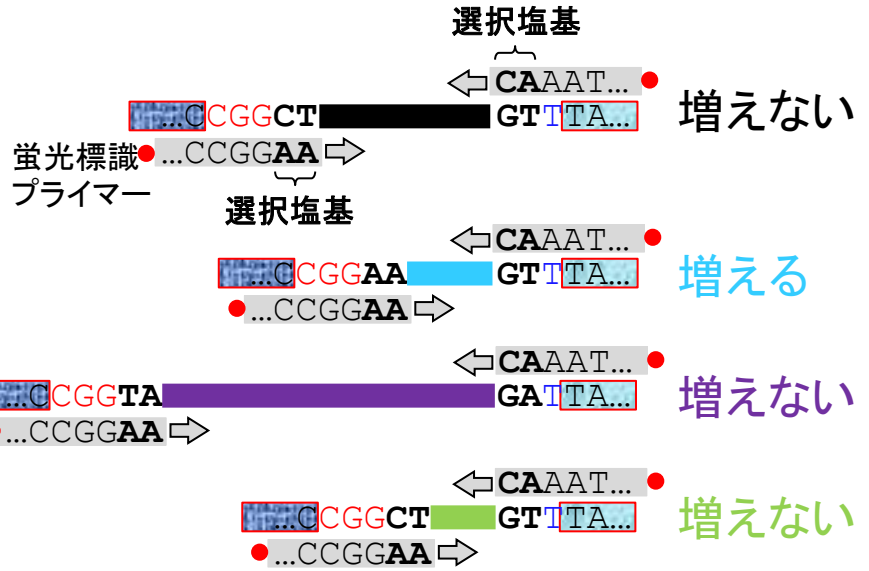
各々のアダプター配列を付加

CGGTT [■■■■■■■■] GTT

...CCGGTT [■■■■■■■■] GTTTA...

選択的PCR増幅

選択塩基の組合せ例：**AA-CA**

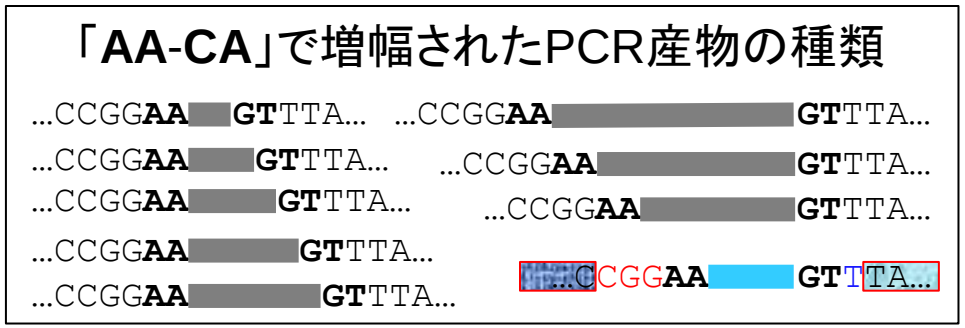


「AA-CA」で増幅されたPCR産物の種類

...CCGGAA [■■■■■■■■] GTTTA...
...CCGGAA [■■■■■■■■] GTTTA...
...CCGGAA [■■■■■■■■] GTTTA...
...CCGGAA [■■■■■■■■] GTTTA...
...CCGGAA [■■■■■■■■] GTTTA...
...CCGGAA [■■■■■■■■] GTTTA...
...CCGGAA [■■■■■■■■] GTTTA...
...CCGGAA [■■■■■■■■] GTTTA...
...CCGGAA [■■■■■■■■] GTTTA...

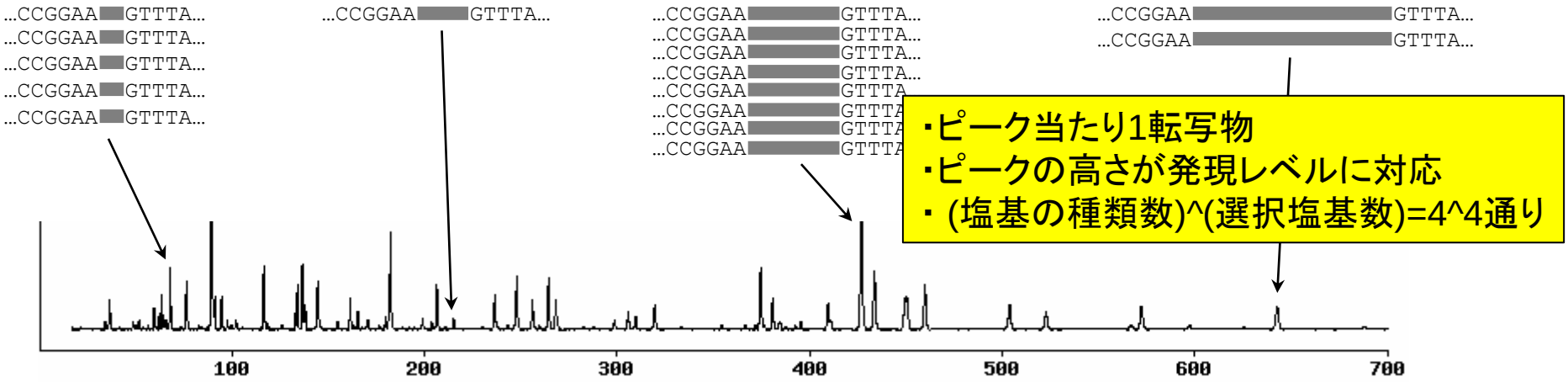
トランスクリプトーム取得（電気泳動）

- cDNA-AFLP（HiCEPの場合）



「転写物由来配列断片（Transcripts Derived Fragments; TDFs）」という表現が業界標準（たぶん）

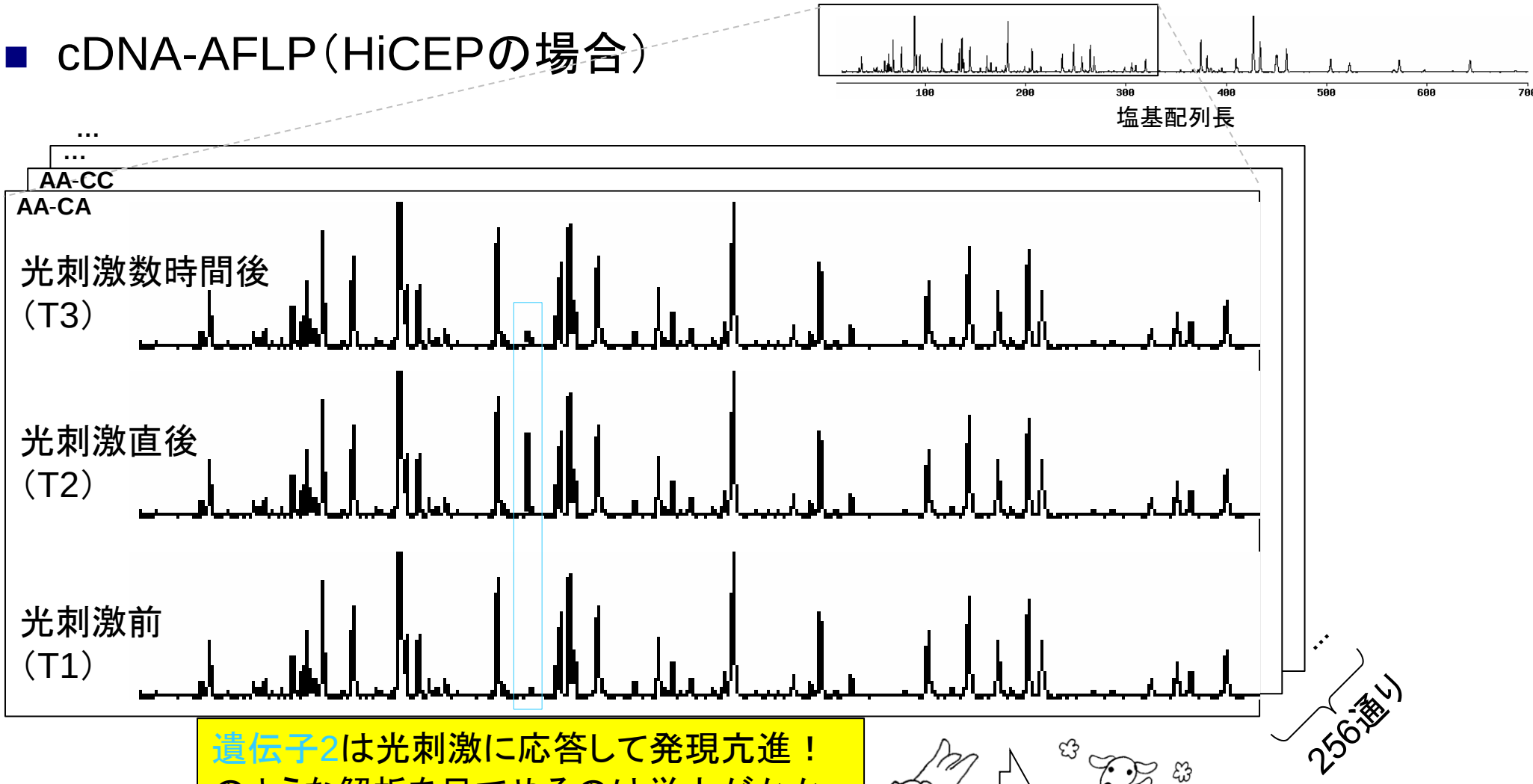
電気泳動
（長さで分離）



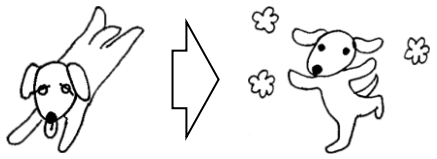
塩基配列長（正確にはアダプター配列を含むTDFの長さ）

トランスクリプトーム取得（電気泳動）

- cDNA-AFLP（HiCEPの場合）



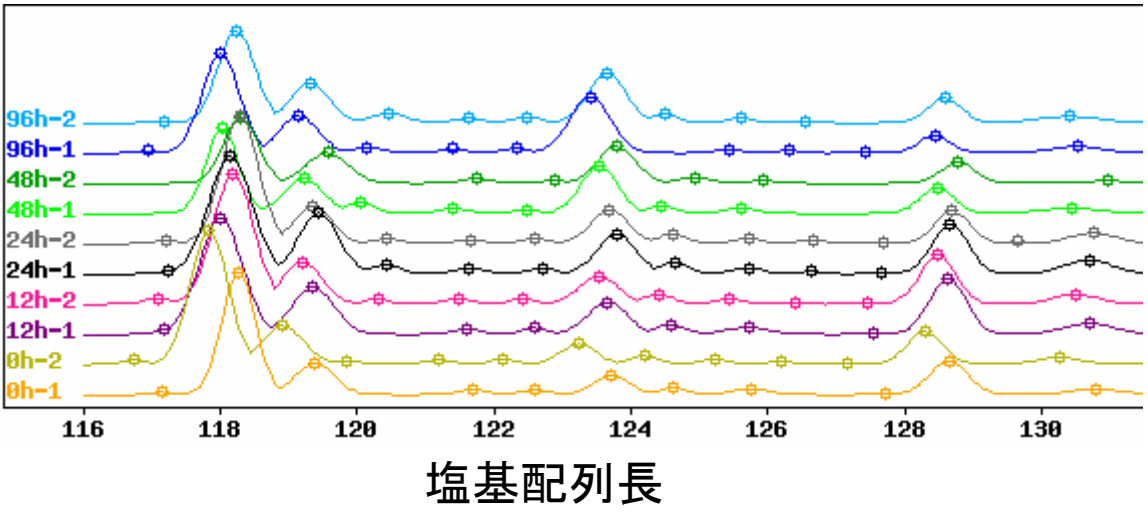
遺伝子2は光刺激に応答して発現亢進！
のような解析を目でやるのは労力がかかり大変です → **バイオインフォマティクス**



電気泳動データ → 遺伝子発現行列

- マイクロアレイ（や塩基配列データ）では**遺伝子発現行列**が出発点
- 電気泳動データは**遺伝子発現行列**の作成が困難

比較する実験数が増えるほど、同一遺伝子の認識（アラインメント）精度が下がるから



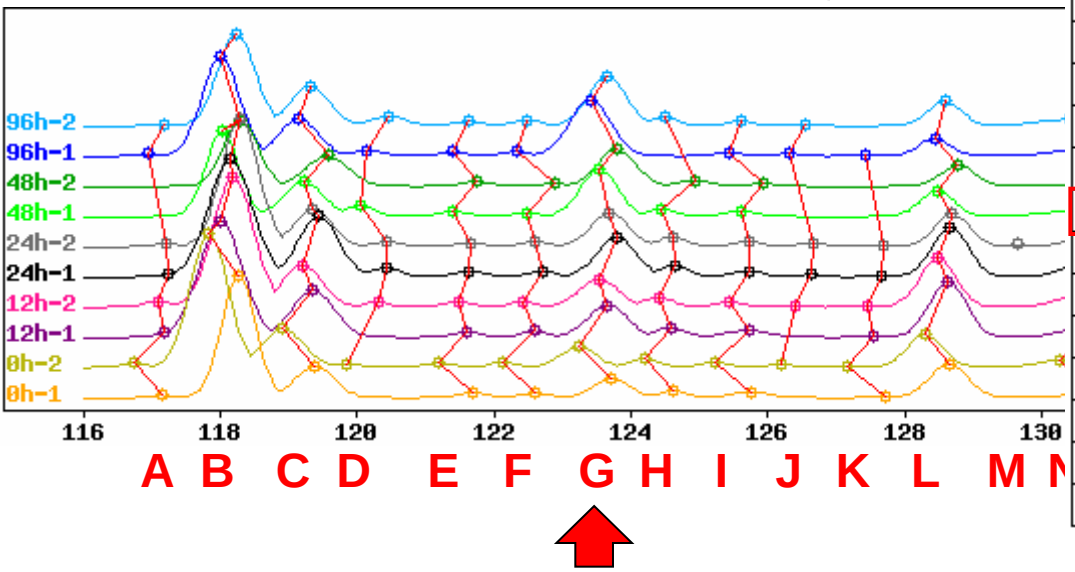
10サンプルの比較

電気泳動データ → 遺伝子発現行列

- マイクロアレイ（や塩基配列データ）では**遺伝子発現行列**が出発点
- 電気泳動データは**遺伝子発現行列**の作成が困難

比較する実験数が増えるほど、同一遺伝子の認識（アラインメント）精度が下がるから

理想的なアラインメント



遺伝子発現行列

	0h-1	0h-2	12h-1	12h-2	24h-1	24h-2	48h-1	48h-2	96h-1	96h-2
A	11	23	21	29	13	15			10	9
B	607	664	576	649	582	634	421	326	491	456
C	156	191	233	209	301	186	172	151	181	195
D		19		24	44	25	53		27	48
E	23	21	28	25	28	19	24	22	20	21
F	21	25	30	26	28	26	19	15	24	29
G	93	100	160	139	196	166	234	184	276	245
H	33	41	47	49	55	48	34	27		43
I	28	24	34	25	30	27	25	14	18	23
J		16		8	13	17			14	7
K	7	9	8	9	8	9			4	
L	168	163	275	242	246	165	126	102	85	123
M						10				
N	30	34	54	49	68	52	25	11	33	31

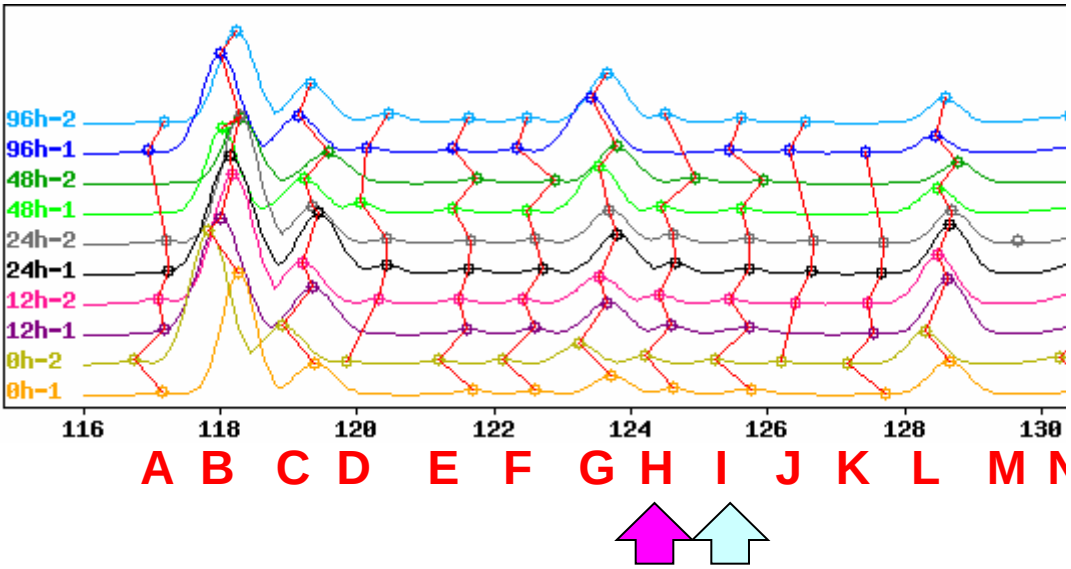
Gの発現が徐々に上昇している

電気泳動データ → 遺伝子発現行列

- マイクロアレイ（や塩基配列データ）では**遺伝子発現行列**が出発点
- 電気泳動データは**遺伝子発現行列**の作成が困難

比較する実験数が増えるほど、同一遺伝子の認識（アラインメント）精度が下がるから

理想的なアラインメント



遺伝子発現行列

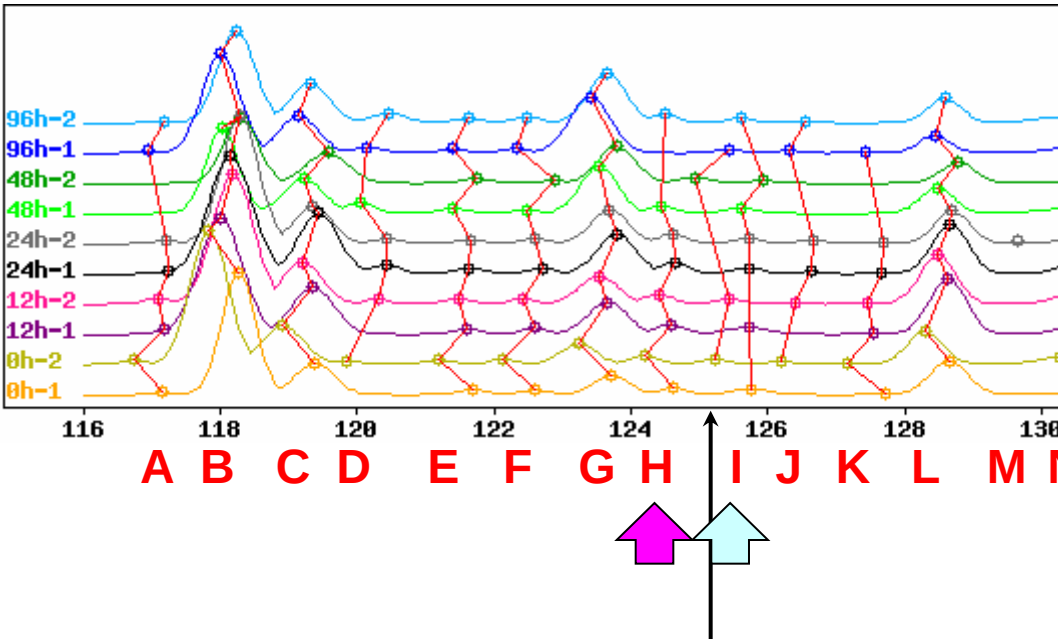
	0h-1	0h-2	12h-1	12h-2	24h-1	24h-2	48h-1	48h-2	96h-1	96h-2
A	11	23	21	29	13	15			10	9
B	607	664	576	649	582	634	421	326	491	456
C	156	191	233	209	301	186	172	151	181	195
D		19		24	44	25	53		27	48
E	23	21	28	25	28	19	24	22	20	21
F	21	25	30	26	28	26	19	15	24	29
G	93	100	160	139	196	166	234	184	276	245
H	**33**	**41**	**47**	**49**	**55**	**48**	**34**	**27**		**43**
I	**28**	**24**	**34**	**25**	**30**	**27**	**25**	**14**	**18**	**23**
J		16		8	13	17			14	7
K	7	9	8	9	8	9			4	
L	168	163	275	242	246	165	126	102	85	123
M						10				
N	30	34	54	49	68	52	25	11	33	31

電気泳動データ → 遺伝子発現行列

- マイクロアレイ（や塩基配列データ）では**遺伝子発現行列**が出発点
- 電気泳動データは**遺伝子発現行列**の作成が困難

比較する実験数が増えるほど、同一遺伝子の認識（アラインメント）精度が下がるから

現実…



遺伝子発現行列

	0h-1	0h-2	12h-1	12h-2	24h-1	24h-2	48h-1	48h-2	96h-1	96h-2
A	11	23	21	29	13	15			10	9
B	607	664	576	649	582	634	421	326	491	456
C	156	191	233	209	301	186	172	151	181	195
D		19		24	44	25	53		27	48
E	23	21	28	25	28	19	24	22	20	21
F	21	25	30	26	28	26	19	15	24	29
G	93	100	160	139	196	166	234	184	276	245
H	**33**	**41**	**47**	**49**	**55**	**48**	**34**			**43**
H'I'		**24**		**25**				**27**	**18**	
I	**28**		**34**		**30**	**27**	**25**	**14**		**23**
J		16		8	13	17			14	7
K	7	9	8	9	8	9			4	
L	168	163	275	242	246	165	126	102	85	123
M						10				
N	30	34	54	49	68	52	25	11	33	31

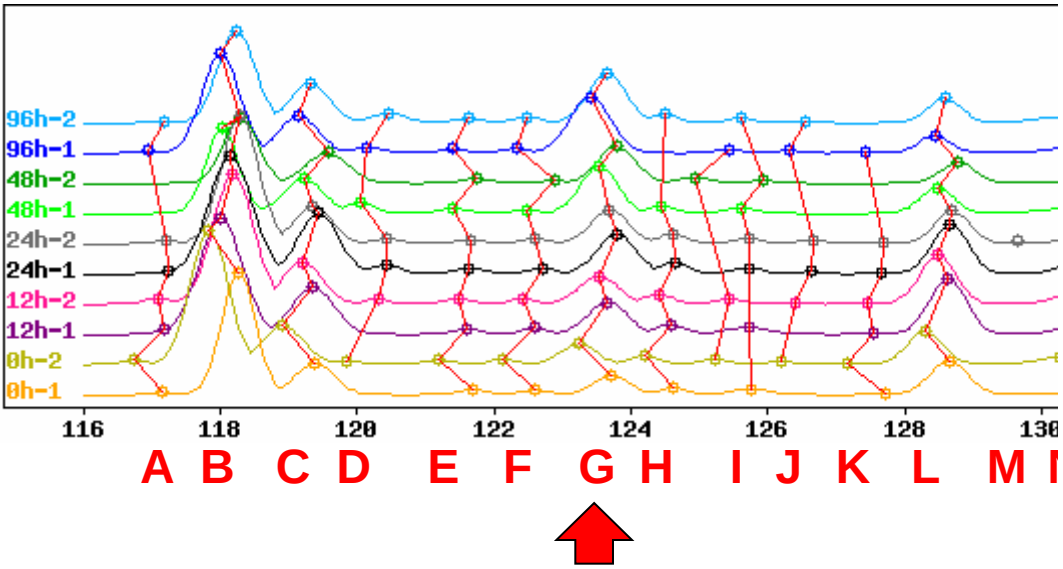
H’I’

電気泳動データ → 遺伝子発現行列

- マイクロアレイ（や塩基配列データ）では**遺伝子発現行列**が出発点
- 電気泳動データは**遺伝子発現行列**の作成が困難

比較する実験数が増えるほど、同一遺伝子の認識（アラインメント）精度が下がるから

現実…

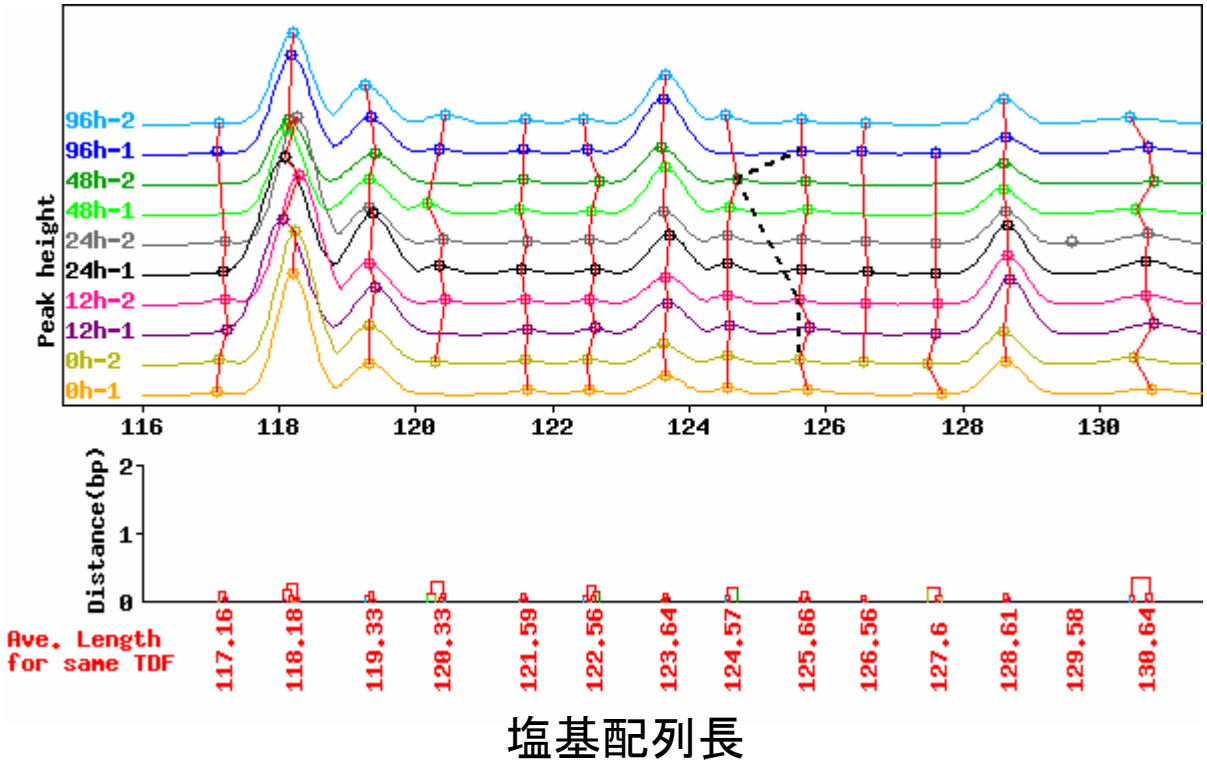


遺伝子発現行列

	0h-1	0h-2	12h-1	12h-2	24h-1	24h-2	48h-1	48h-2	96h-1	96h-2
A	11	23	21	29	13	15			10	9
B	607	664	576	649	582	634	421	326	491	456
C	156	191	233	209	301	186	172	151	181	195
D		19		24	44	25	53		27	48
E	23	21	28	25	28	19	24	22	20	21
F	21	25	30	26	28	26	19	15	24	29
G	93	100	160	139	196	166	234	184	276	245
H	33	41	47	49	55	48	34			43
H'I'		24		25				27	18	
I	28		34		30	27	25	14		23
J		16		8	13	17			14	7
K	7	9	8	9	8	9			4	
L	168	163	275	242	246	165	126	102	85	123
M						10				
N	30	34	54	49	68	52	25	11	33	31

Gの発現パターンは本当に全部G由来？！

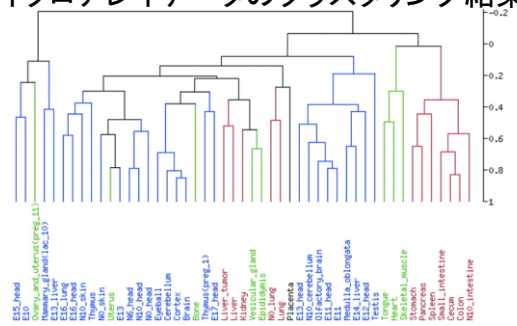
GOGOT法



塩基配列長

「フラグメント長補正」 ＋
「Complete-linkage clustering」

マイクロアレイデータのクラスタリング結果

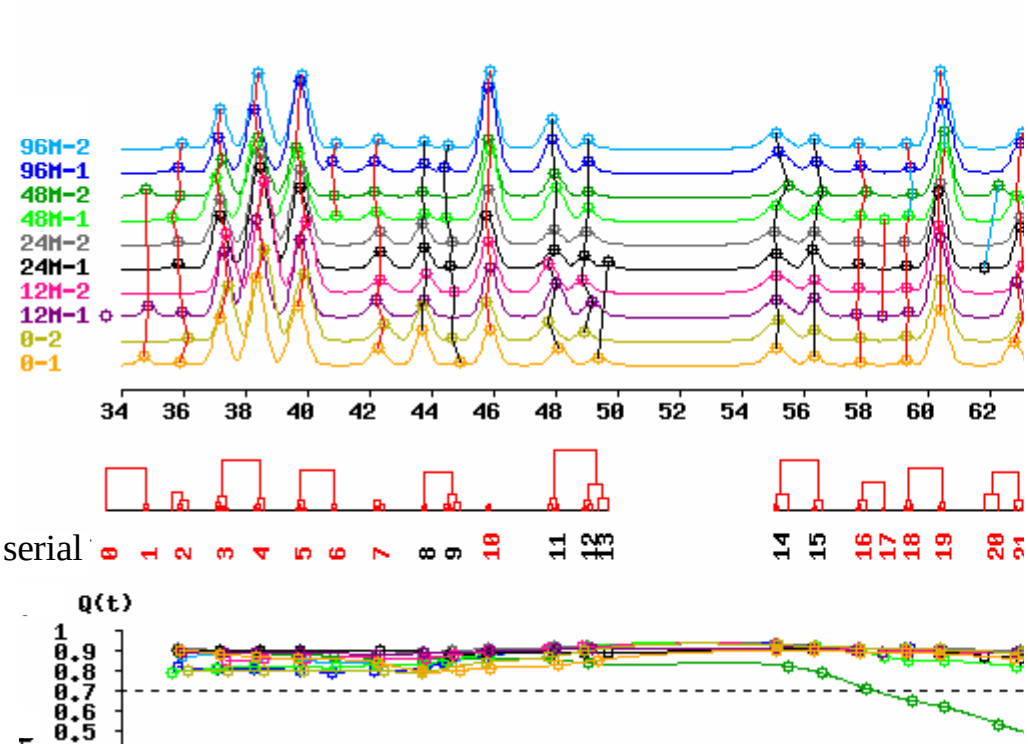


赤線を引く（同一ピークの認識）手段として、マイクロアレイ解析などでよく用いられている「クラスタリング」を利用可能

であることを講義中にふと思いついた（2006年ごろ）

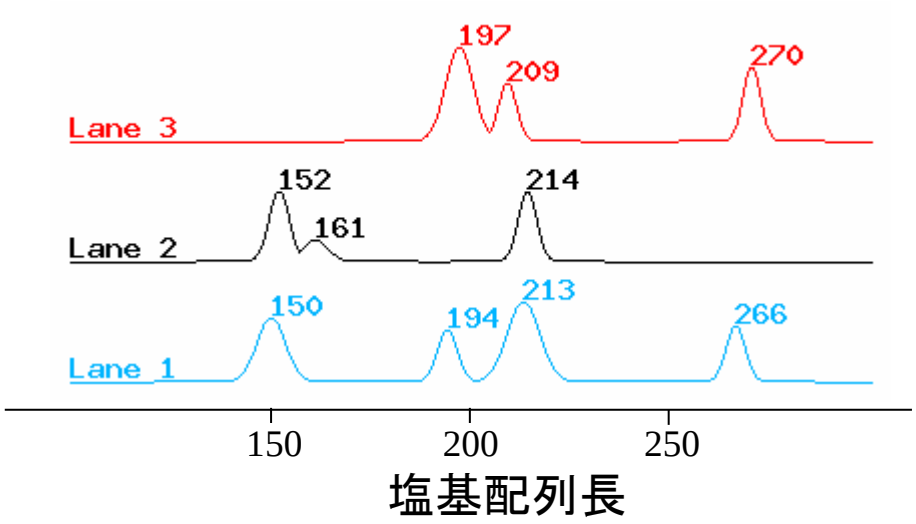
電気泳動データ → 遺伝子発現行列

- ピークのアラインメントがとれている　＝　遺伝子発現行列を作れている



serial	0h-1	0h-2	12h-1	12h-2	24h-1	24h-2	48h-1	48h-2	96h-1	96h-2
0			5							
1	41		58					39		
2	16	17	26		21	16	16	8	24	23
3	235	276	330	297	266	222	213	185	179	199
4	437	456	480	554	507	455	379	300	320	375
5	300	333	381	357	409	371	332	249	452	369
6							27	8	52	25
7	81	82	86	63	83	64	49	27	51	41
8	172	174	84	91	108	105	30	23	41	31
9	11	16		6	11	13	15		22	17
10	173	191	245	255	266	274	359	289	424	384
11	84	91	169	144	96	71	164	114	163	144
12		44	73	64	62	64	60	30	53	43
13	39				37					
14	87	102	85	85	75	64	77	58	101	76
15	48	53	93	60	95	66	55	28	57	45
16	13	15	18	21	22	16	24	27	34	25
17			7				7			
18	21	23	24	28	18	18	23	14	24	27

解析例（Peak alignment via clustering）

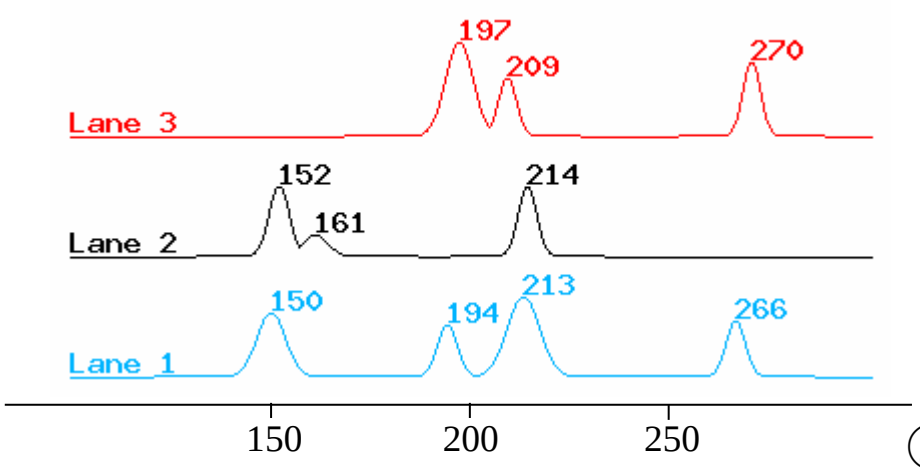


Lane	M.W.
1	150
1	194
1	213
1	266
2	152
2	161
2	214
3	197
3	209
3	270

①分子量でソート

Lane	M.W.
1	150
2	152
2	161
1	194
3	197
3	209
1	213
2	214
1	266
3	270

解析例（Peak alignment via clustering）



Lane	M.W.
1	150
1	194
1	213
1	266
2	152
2	161
2	214
3	197
3	209
3	270

①分子量でソート

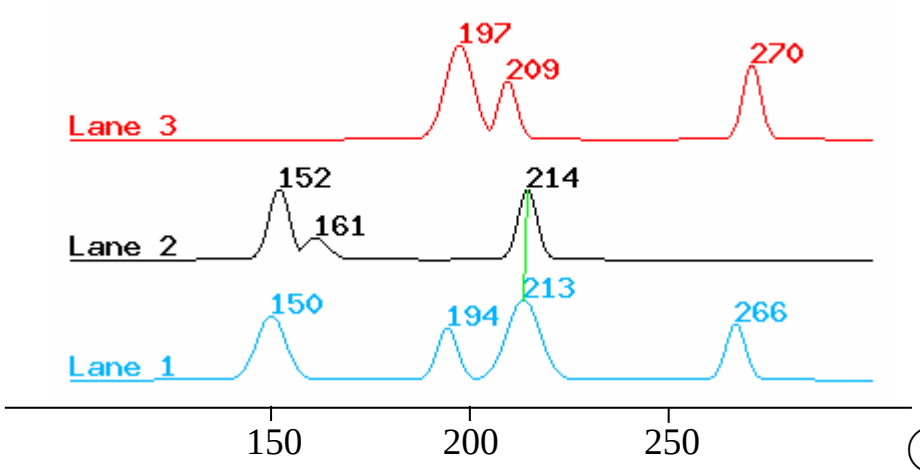
Lane	M.W.	
1	150	
2	152	2
2	161	9
1	194	33
3	197	3
3	209	12
1	213	4
2	214	1
1	266	52
3	270	4

②ピーク間距離を計算

②’クラスター間距離が最短のものをマージ

c.	TDF
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0

解析例（Peak alignment via clustering）



②ピーク間距離を計算

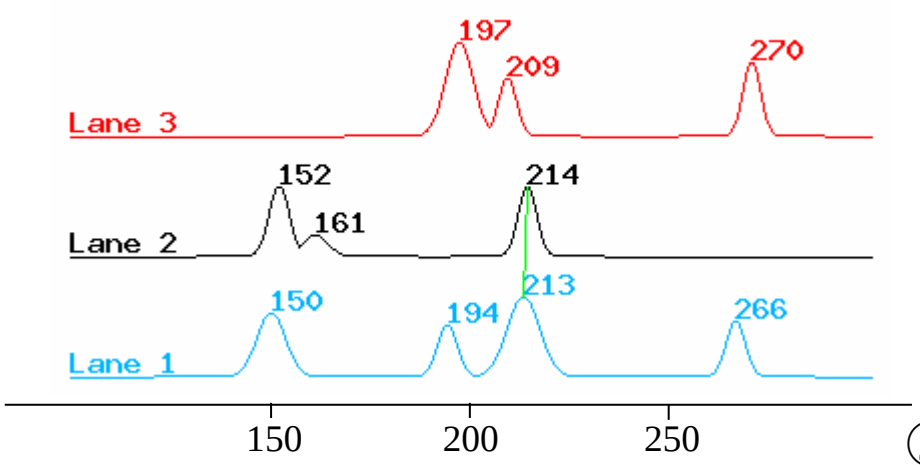
Lane	M.W.	
1	150	2
2	152	9
2	161	33
1	194	3
3	197	12
3	209	4
1	213	1
2	214	52
1	266	4
3	270	

②’クラスター間距離が最短のものをマージ

c.	c.	TDF	TDF
1	1	0	0
2	2	0	0
3	3	0	0
4	4	0	0
5	5	0	0
6	6	0	0
7	7	0	1
8	7	0	1
9	8	0	0
10	9	0	0

マージされたクラスター内のメンバーが同一レーン由来のものでない場合に同じTDF IDを付与

解析例（Peak alignment via clustering）



②ピーク（or クラスター）間距離Dを計算

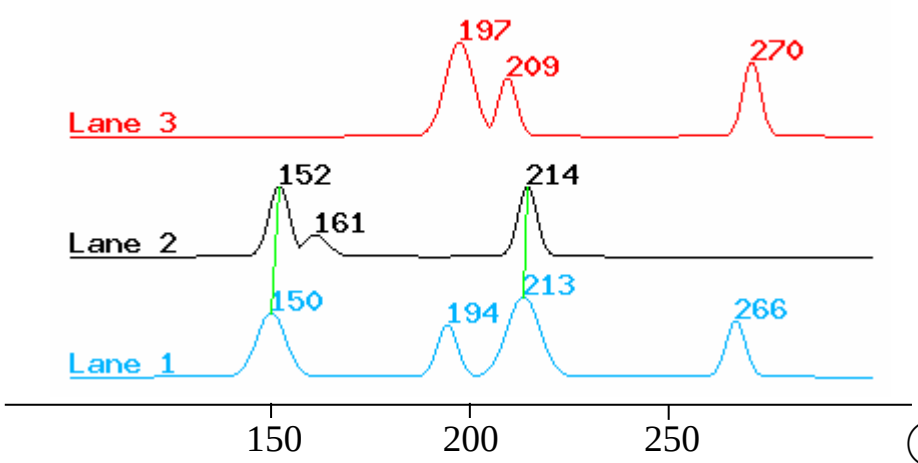
②'クラスター間距離が最短のものをマージ

c.	Lane	M.W.
1	1	150
2	2	152
3	2	161
4	1	194
5	3	197
6	3	209
7	1	213
7	2	214
8	1	266
9	3	270

2
9
33
3
12
D(6, 7)=5
D(7, 8)=53
4

c.	TDF
1	0
2	0
3	0
4	0
5	0
6	0
7	1
7	1
8	0
9	0

解析例（Peak alignment via clustering）



②ピーク（or クラスター）間距離Dを計算

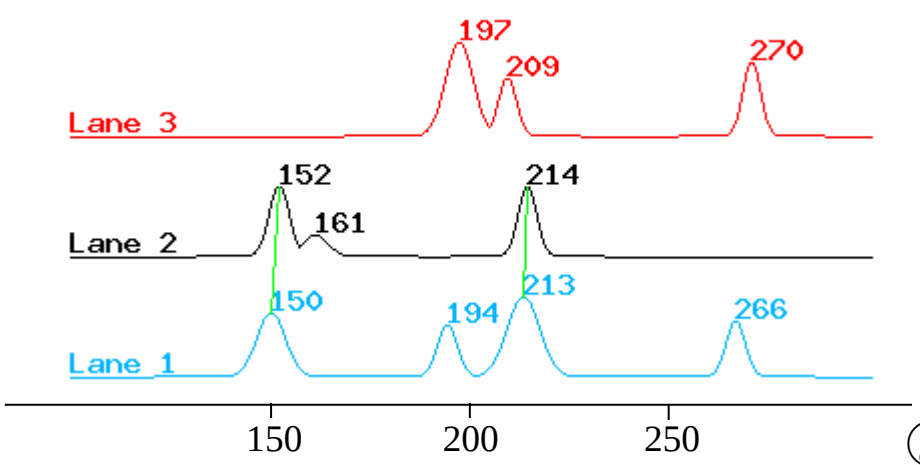
c.	Lane	M.W.	
1	1	150	2
2	2	152	9
3	2	161	33
4	1	194	3
5	3	197	12
6	3	209	D(6, 7)=5
7	1	213	D(7, 8)=53
7	2	214	4
8	1	266	
9	3	270	

②'クラスター間距離が最短のものをマージ

c.	c.	TDF	TDF
1	1	0	2
2	1	0	2
3	2	0	0
4	3	0	0
5	4	0	0
6	5	0	0
7	6	1	1
7	6	1	1
8	7	0	0
9	8	0	0

マージされたクラスター内のメンバーが同一レーン由来のものでない場合に同じTDF IDを付与

解析例（Peak alignment via clustering）



②ピーク（or クラスター）間距離Dを計算

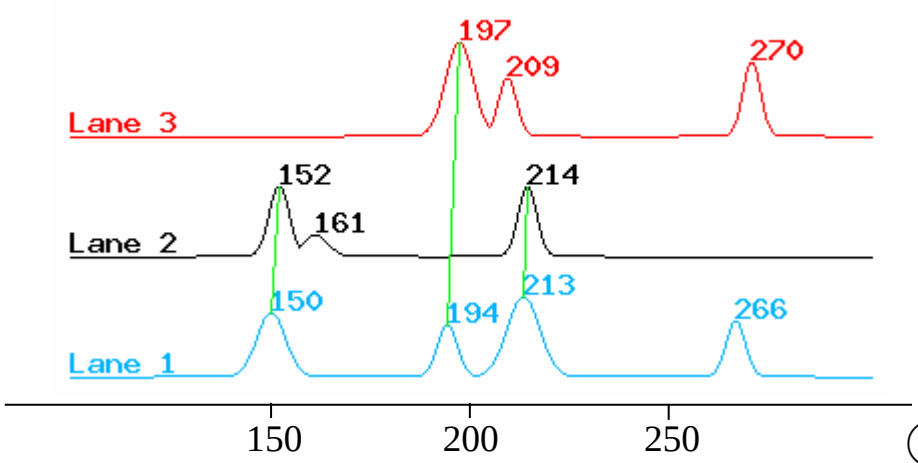
②'クラスター間距離が最短のものをマージ

c.	Lane	M.W.
1	1	150
1	2	152
2	2	161
3	1	194
4	3	197
5	3	209
6	1	213
6	2	214
7	1	266
8	3	270

D(1, 2)=11
33
3
12
D(5, 6)=5
D(6, 7)=53
4

c.	TDF
1	2
1	2
2	0
3	0
4	0
5	0
6	1
6	1
7	0
8	0

解析例（Peak alignment via clustering）



②ピーク（or クラスター）間距離Dを計算

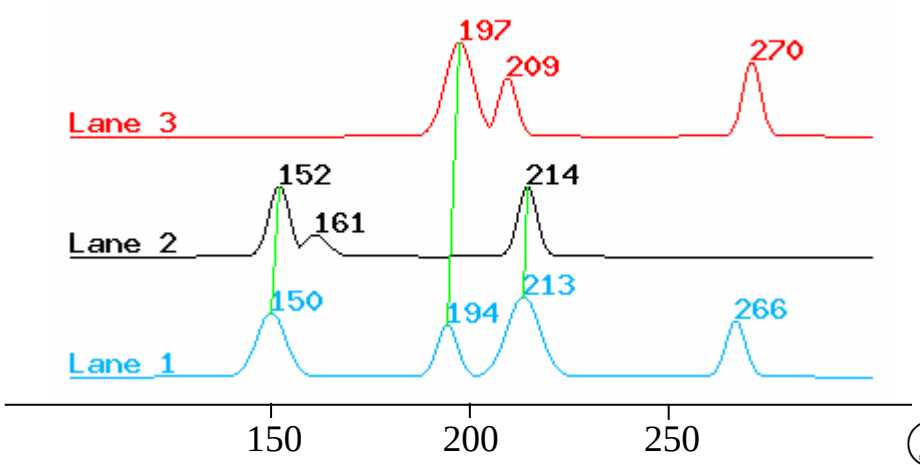
c.	Lane	M.W.	
1	1	150	
1	2	152	D(1, 2)=11
2	2	161	33
3	1	194	3
4	3	197	12
5	3	209	D(5, 6)=5
6	1	213	D(6, 7)=53
6	2	214	4
7	1	266	
8	3	270	

②'クラスター間距離が最短のものをマージ

c.	c.	TDF	TDF
1	1	2	2
1	1	2	2
2	2	0	0
3	3	0	3
4	3	0	3
5	4	0	0
6	5	1	1
6	5	1	1
7	6	0	0
8	7	0	0

マージされたクラスター内のメンバーが同一レーン由来のものでない場合に同じTDF IDを付与

解析例（Peak alignment via clustering）



②ピーク（or クラスター）間距離Dを計算

②'クラスター間距離が最短のものをマージ

c.	Lane	M.W.
1	1	150
1	2	152
2	2	161
3	1	194
3	3	197
4	3	209
5	1	213
5	2	214
6	1	266
7	3	270

D(1, 2)=11

D(2, 3)=36

D(3, 4)=15

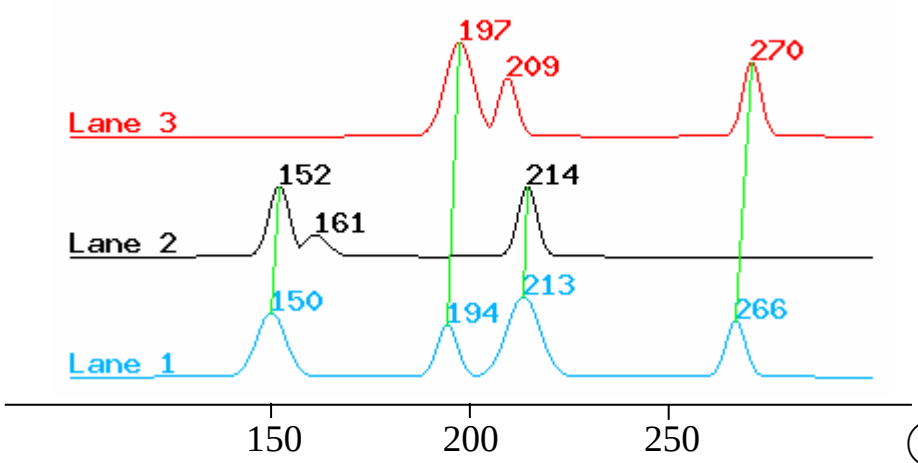
D(4, 5)=5

D(5, 6)=53

D(6, 7)=4

c.	TDF
1	2
1	2
2	0
3	3
3	3
4	0
5	1
5	1
6	0
7	0

解析例（Peak alignment via clustering）



②ピーク（or クラスター）間距離Dを計算

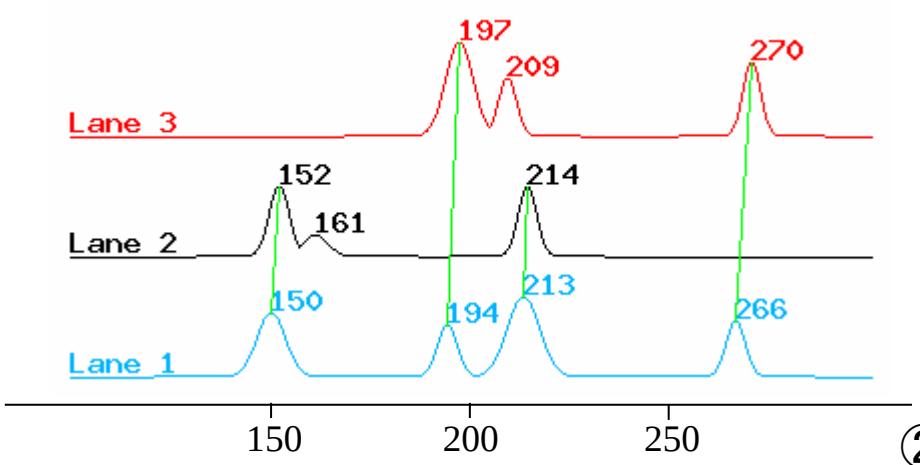
c.	Lane	M.W.	
1	1	150	
1	2	152	D(1, 2)=11
2	2	161	
3	1	194	D(2, 3)=36
3	3	197	D(3, 4)=15
4	3	209	
5	1	213	D(4, 5)=5
5	2	214	D(5, 6)=53
6	1	266	D(6, 7)=4
7	3	270	

②'クラスター間距離が最短のものをマージ

c.	c.	TDF	TDF
1	1	2	2
1	1	2	2
2	2	0	0
3	3	3	3
3	3	3	3
4	4	0	0
5	5	1	1
5	5	1	1
6	6	0	4
7	6	0	4

マージされたクラスター内のメンバーが同一レーン由来のものでない場合に同じTDF IDを付与

解析例（Peak alignment via clustering）

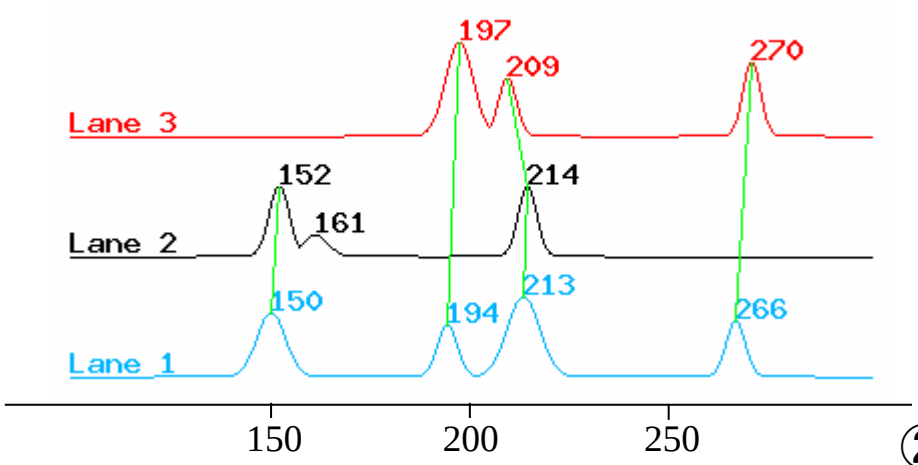


②クラスター間距離Dを計算

②'クラスター間距離が最短のものをマージ

c.	Lane	M.W.		c.	TDF
1	1	150		1	2
1	2	152	D(1, 2)=11	1	2
2	2	161		2	0
3	1	194	D(2, 3)=36	3	3
3	3	197	D(3, 4)=15	3	3
4	3	209		4	0
5	1	213	D(4, 5)=5	5	1
5	2	214		5	1
6	1	266	D(5, 6)=57	6	4
6	3	270		6	4

解析例（Peak alignment via clustering）



②クラスター間距離Dを計算

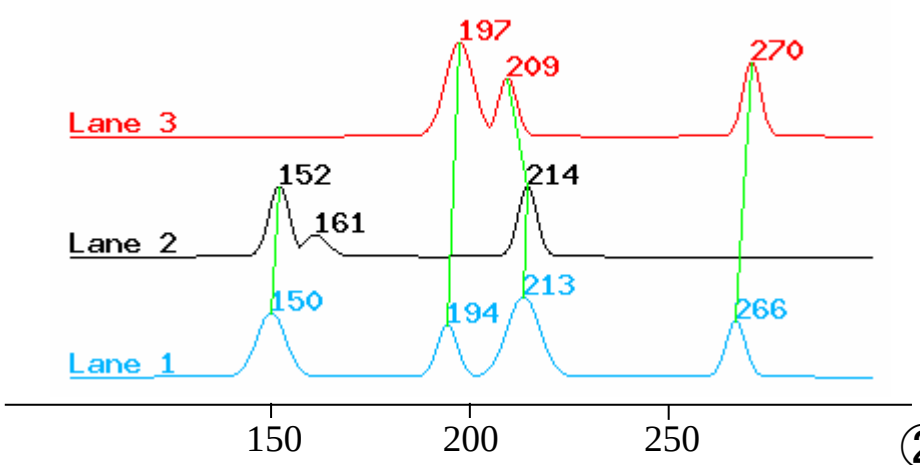
c.	Lane	M.W.	
1	1	150	
1	2	152	D(1, 2)=11
2	2	161	
3	1	194	D(2, 3)=36
3	3	197	D(3, 4)=15
4	3	209	
5	1	213	D(4, 5)=5
5	2	214	
6	1	266	D(5, 6)=57
6	3	270	

②'クラスター間距離が最短のものをマージ

c.	c.	TDF	TDF
1	1	2	2
1	1	2	2
2	2	0	0
3	3	3	3
3	3	3	3
4	4	0	1
5	4	1	1
5	4	1	1
6	5	4	4
6	5	4	4

マージされたクラスター内のメンバーが同一レーン由来のものでない場合に同じTDF IDを付与

解析例（Peak alignment via clustering）



c.	Lane	M.W.
1	1	150
1	2	152
2	2	161
3	1	194
3	3	197
4	3	209
4	1	213
4	2	214
5	1	266
5	3	270

②クラスター間距離Dを計算

D(1, 2)=11

D(2, 3)=36

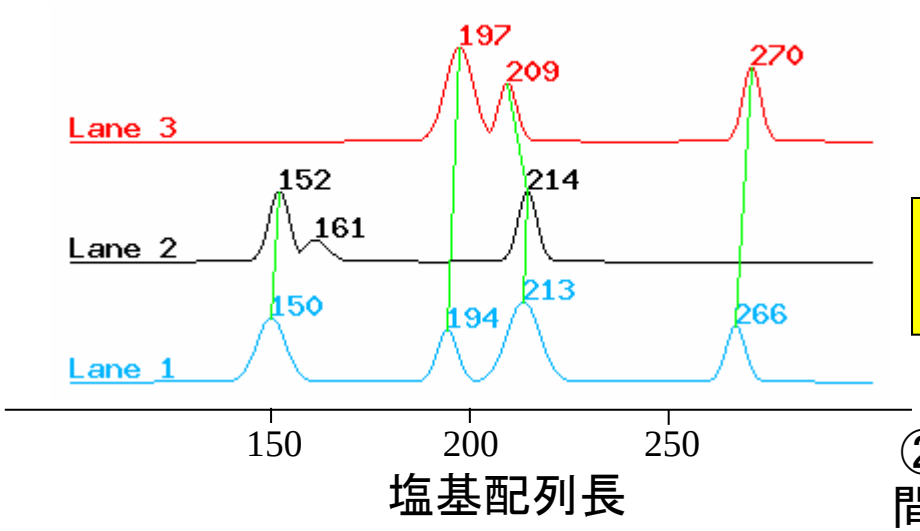
D(3, 4)=20

D(4, 5)=61

②'クラスター間距離が最短のものをマージ

c.	TDF
1	2
1	2
2	0
3	3
3	3
4	1
4	1
4	1
5	4
5	4

解析例（Peak alignment via clustering）



クラスターとしては無条件にマージ

同じレーン由来のピークが含まれる場合には、新たなTDF IDを付与

②クラスター間距離Dを計算

②’クラスター間距離が最短のものをマージ

c.	Lane	M.W.
1	1	150
1	2	152
2	2	161
3	1	194
3	3	197
4	3	209
4	1	213
4	2	214
5	1	266
5	3	270

D(1, 2)=11

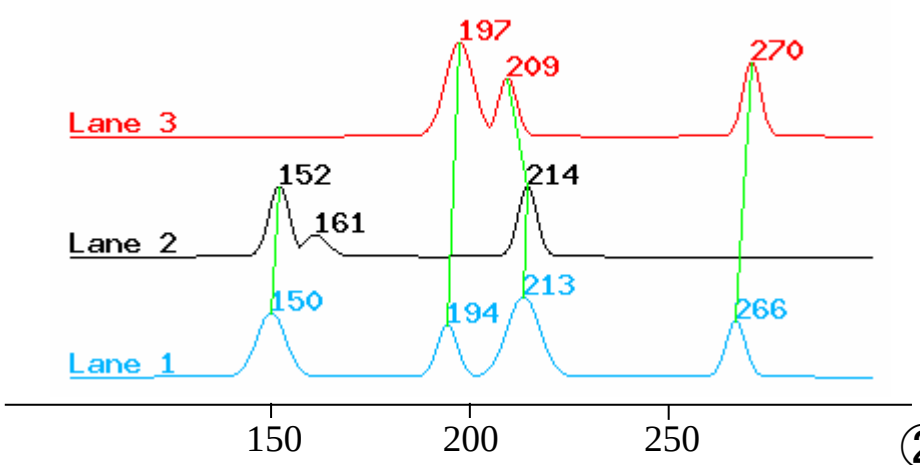
D(2, 3)=36

D(3, 4)=20

D(4, 5)=61

c.	c.	TDF	TDF
1	1	2	2
1	1	2	2
2	1	0	5
3	2	3	3
3	2	3	3
4	3	1	1
4	3	1	1
4	3	1	1
5	4	4	4
5	4	4	4

解析例（Peak alignment via clustering）



②クラスター間距離Dを計算

②'クラスター間距離が最短のものをマージ

c.	Lane	M.W.
1	1	150
1	2	152
1	2	161
2	1	194
2	3	197
3	3	209
3	1	213
3	2	214
4	1	266
4	3	270

D(1, 2)=47

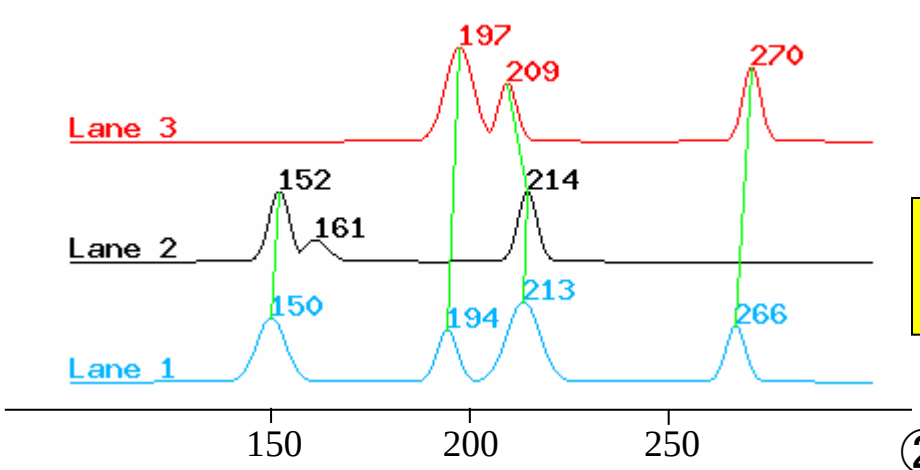
D(2, 3)=20

D(3, 4)=61

c.
1
1
1
2
2
3
3
3
4
4

TDF
2
2
5
3
3
1
1
1
4
4

解析例（Peak alignment via clustering）



クラスターとしては
無条件にマージ

②クラスター間距離Dを計算

②’クラスター間距離が最短のものをマージ

c.	Lane	M.W.
1	1	150
1	2	152
1	2	161
2	1	194
2	3	197
3	3	209
3	1	213
3	2	214
4	1	266
4	3	270

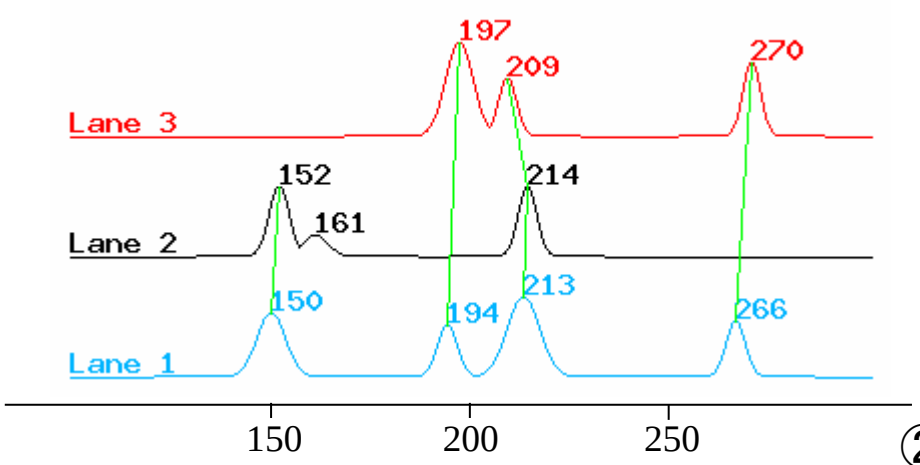
D(1, 2)=47

D(2, 3)=20

D(3, 4)=61

c.	c.	TDF	TDF
1	1	2	2
1	1	2	2
1	1	5	5
2	2	3	3
2	2	3	3
3	2	1	1
3	2	1	1
3	2	1	1
4	3	4	4
4	3	4	4

解析例（Peak alignment via clustering）



②クラスター間距離Dを計算

②'クラスター間距離が最短のものをマージ

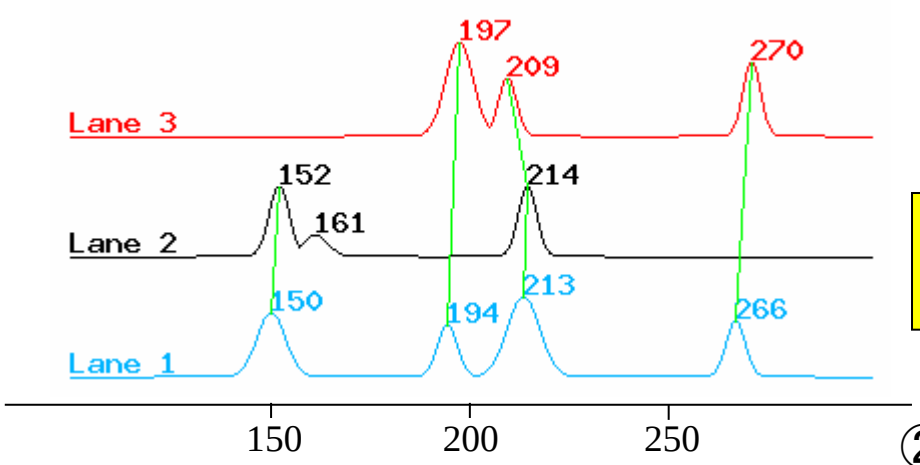
c.	Lane	M.W.
1	1	150
1	2	152
1	2	161
2	1	194
2	3	197
2	3	209
2	1	213
2	2	214
3	1	266
3	3	270

D(1, 2)=64

D(2, 3)=76

c.	TDF
1	2
1	2
1	5
2	3
2	3
2	1
2	1
2	1
3	4
3	4

解析例（Peak alignment via clustering）



クラスターとしては
無条件にマージ

②クラスター間距離Dを計算

②’クラスター間距離が最短のものをマージ

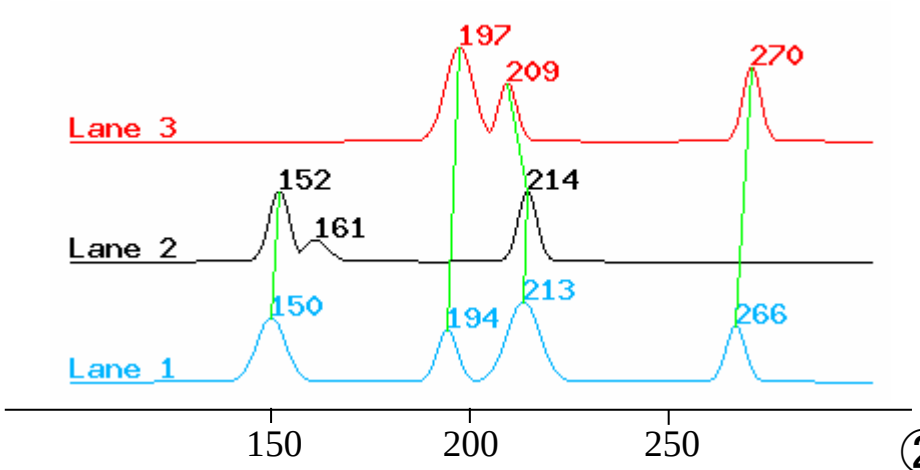
c.	Lane	M.W.
1	1	150
1	2	152
1	2	161
2	1	194
2	3	197
2	3	209
2	1	213
2	2	214
3	1	266
3	3	270

D(1, 2)=64

D(2, 3)=76

c.	c.	TDF	TDF
1	1	2	2
1	1	2	2
1	1	5	5
2	1	3	3
2	1	3	3
2	1	1	1
2	1	1	1
2	1	1	1
3	2	4	4
3	2	4	4

解析例（Peak alignment via clustering）



②クラスター間距離Dを計算

②'クラスター間距離が最短のものをマージ

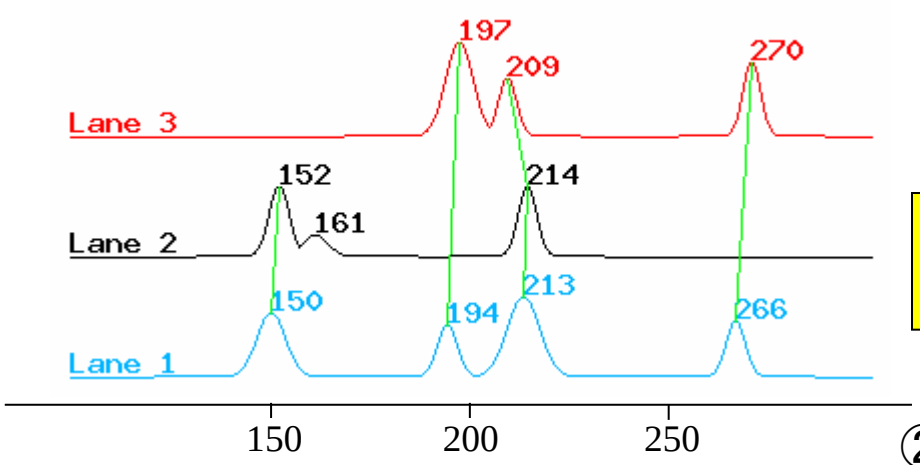
c.	Lane	M.W.
1	1	150
1	2	152
1	2	161
1	1	194
1	3	197
1	3	209
1	1	213
1	2	214
2	1	266
2	3	270

D(1, 2)=120

c.
1
1
1
1
1
1
1
1
2
2

TDF
2
2
5
3
3
1
1
1
4
4

解析例（Peak alignment via clustering）



全ピークが同じクラスターになれば計算終了

②クラスター間距離Dを計算

②'クラスター間距離が最短のものをマージ

c.	Lane	M.W.
1	1	150
1	2	152
1	2	161
1	1	194
1	3	197
1	3	209
1	1	213
1	2	214
2	1	266
2	3	270

D(1, 2)=120

c.	c.	TDF	TDF
1	1	2	2
1	1	2	2
1	1	5	5
1	1	3	3
1	1	3	3
1	1	1	1
1	1	1	1
1	1	1	1
2	1	4	4
2	1	4	4

様々な遺伝子発現行列

心臓 脳 腎臓 肺

1. 二群間比較

	A群 A1	A2	…	B群 B1	B2	…
gene *1*	$x^A_{1,1}$	$x^A_{1,2}$		$x^B_{1,2}$	$x^B_{1,2}$	
gene *2*	$x^A_{2,1}$	$x^A_{2,2}$		$x^B_{2,2}$	$x^B_{2,2}$	
…	…	…	…	…	…	…
gene *i*	$x^A_{i,1}$	$x^A_{i,2}$		$x^B_{i,2}$	$x^B_{i,2}$	
…	…	…	…	…	…	…
gene *n*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	

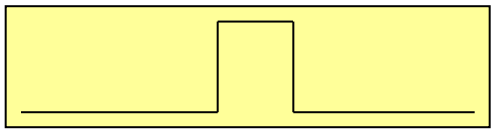
2. 様々な組織（条件）

	S1	S2	S3	S4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

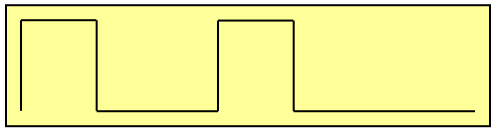
3. 時系列データ

	T1	T2	T3	T4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

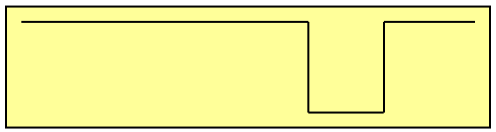
脳特異的高発現



心臓と脳特異的高発現



肺特異的低発現



組織特異的遺伝子検出にエントロピーを利用

- 遺伝子iのエントロピー $H(x_i) = -\sum_{j=1}^{N} p_{ij} \log_2(p_{ij})$, where $p_{ij} = x_{ij} / \sum_{j=1}^{N} x_{ij}$

N:組織数(jの数) = 8

Hの取りうる範囲:$0 \leqq H \leqq \log_2 N \rightarrow 0 \leqq H \leqq 3$

x_{ij} (i: 遺伝子, j: 組織)

j	遺伝子1	遺伝子2	遺伝子3	遺伝子4	遺伝子5	...
組織1	1	5	6	4	10	...
組織2	0	2	6	4	10	...
組織3	0	1	6	4	10	...
組織4	9	2	6	4	10	...
組織5	0	4	6	10	4	
組織6	0	6	6	4	10	
組織7	0	3	6	4	10	
組織8	0	5	6	4	10	
$\Sigma_j x_{ij}$	10	28	48	38	74	

p_{ij} (i: 遺伝子, j: 組織)

j	遺伝子1	遺伝子2	遺伝子3	遺伝子4	遺伝子5	...
1	0.10	0.18	0.13	0.11	0.14	...
2	0.00	0.07	0.13	0.11	0.14	...
3	0.00	0.04	0.13	0.11	0.14	...
4	0.90	0.07	0.13	0.11	0.14	...
5	0.00	0.14	0.13	0.26	0.05	
6	0.00	0.21	0.13	0.11	0.14	
7	0.00	0.11	0.13	0.11	0.14	
8	0.00	0.18	0.13	0.11	0.14	
Σ_j	1.00	1.00	1.00	1.00	1.00	

$-p_{ij} \log_2(p_{ij})$ (i: 遺伝子, j: 組織)

j	遺伝子1	遺伝子2	遺伝子3	遺伝子4	遺伝子5	...
1	0.33	0.44	0.38	0.34	0.39	...
2	0.00	0.27	0.38	0.34	0.39	...
3	0.00	0.17	0.38	0.34	0.39	...
4	0.14	0.27	0.38	0.34	0.39	...
5	0.00	0.40	0.38	0.51	0.23	
6	0.00	0.48	0.38	0.34	0.39	
7	0.00	0.35	0.38	0.34	0.39	
8	0.00	0.44	0.38	0.34	0.39	
Σ_j	0.47	2.83	3.00	2.90	2.96	

組織特異的遺伝子は低いエントロピー

そうでないものは高い値

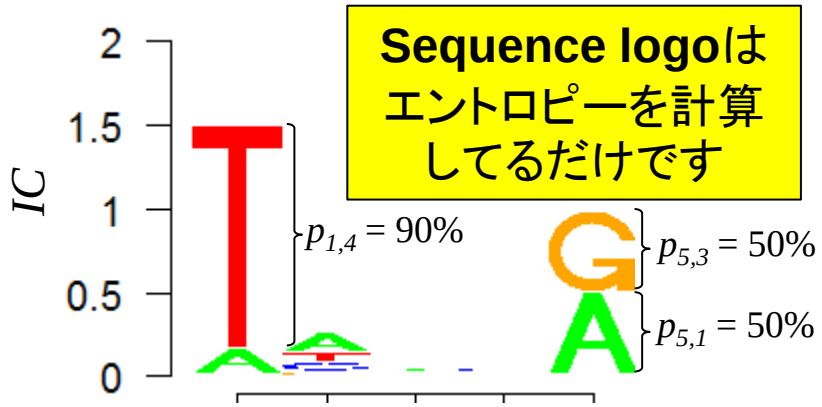
配列モチーフなどの表現にエントロピーを利用

- position iの情報量 $IC_i = \log_2(N) - H(x_i)$
  $= 2$

N: 塩基の種類数= 4
Hの取りうる範囲: $0 \leqq H \leqq \log_2 N$

position i

	1	2	3	4	5	…
配列 1	T	A	C	G	G	…
配列 2	T	A	A	C	G	…
配列 3	T	G	T	A	G	…
配列 4	A	C	T	T	A	…
配列 5	T	T	G	G	A	…
配列 6	T	C	A	A	G	…
配列 7	T	A	C	T	A	…
配列 8	T	T	G	C	A	…
配列 9	T	A	A	C	A	…
配列10	T	A	C	T	G	…



Sequence logoは
エントロピーを計算
してるだけです

IC	1.53	0.24	0.03	0.03	1.00	…

x_{ij}	1	2	3	4	5	…
Aの数 (j=1)	1	5	3	2	5	…
Cの数 (j=2)	0	2	3	3	0	…
Gの数 (j=3)	0	1	2	2	5	…
Tの数 (j=4)	9	2	2	3	0	…
$\Sigma_j\ x_{ij}$	10	10	10	10	10	

j

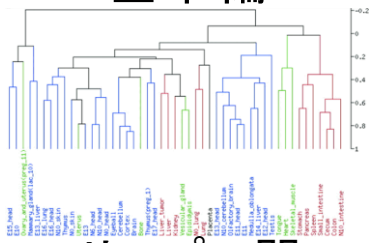
p_{ij}	1	2	3	4	5	…
1	0.1	0.5	0.3	0.2	0.5	…
2	0.0	0.2	0.3	0.3	0.0	…
3	0.0	0.1	0.2	0.2	0.5	…
4	0.9	0.2	0.2	0.3	0.0	…
Σ_j	1.0	1.0	1.0	1.0	1.0	

$-p_{ij}\log_2(p_{ij})$	1	2	3	4	5	…
1	0.33	0.50	0.52	0.46	0.50	…
2	0.00	0.46	0.52	0.52	0.00	…
3	0.00	0.33	0.46	0.46	0.50	…
4	0.14	0.46	0.46	0.52	0.00	…
$H = \Sigma_j$	0.47	1.76	1.97	1.97	1.00	

まとめ（第一部）

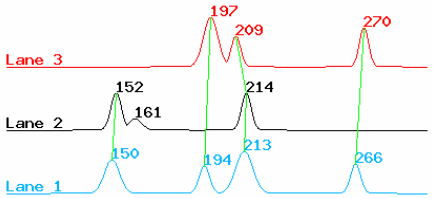
クラスタリング

ー基本編ー



サンプル間

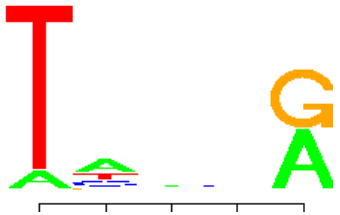
ー応用編ー



ピークマッチング

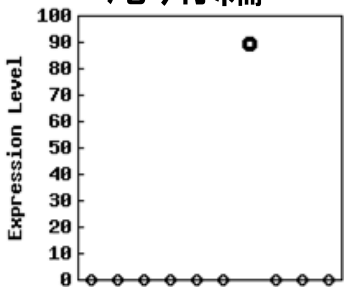
エントロピー

ー基本編ー



sequence logo

ー応用編ー



組織特異的遺伝子

第二部　　第一部

次世代シーケンサデータもRのコピペで解析可能
→ 頭脳労働

バイオインフォ要素技術の習得は大事だが、それだけでも様々な種類の実験データに対応可能

謝辞

東京大学　大学院農学生命科学研究科

清水謙多郎　教授

中井雄治　特任准教授

放射線医学総合研究所　先端遺伝子発現研究G

安倍真澄　グループリーダー

荒木良子　チームリーダー

グラント

若手研究(B)（H19-20年度）:「全生物種のトランスクリプトーム解析を加速するHiCEPデータ高速解析手法の開発」(代表)

挿絵担当:門田雅世