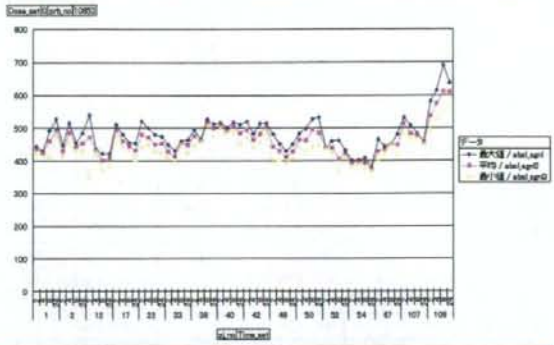
化合物間クラスタリングの課題

- 化合物間(Dose0同士)の距離を決定付けている遺伝子群の中からGc (Prb_No=10853)をグラフ化してみた。発現量が大きく、飽和していると考えられる



このようなプローブセットは排除する必要がある

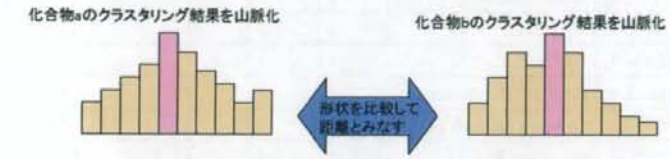
まとめ(1)

- 課題
  - 遺伝子チップのプローブが飽和に近い状況になっていることが考えられる。飽和に近い状況に関して、絶対値変換の際の対策が必要である。
- 本クラスタリングの可能性
  - 今回は、階層型クラスタリングを行った。前段階として、全化合物同士の距離を計算している。これを用いることにより、MADICなどの距離を使用したクラスタリング手法を用いることも可能である。
  - また、距離計算の前段階として各遺伝子間の距離を計算している。これを用いることにより、化合物間の距離に対する各遺伝子の寄与を計算することができ、遺伝子に順位をつけることができる。逆に遺伝子リストを与えることで、飽和等の影響を排除した距離を計算することもできる。

化合物間クラスタリング（山脈構造）

◆同期率計算の山脈構造を用いる。ただし、計算量が多い

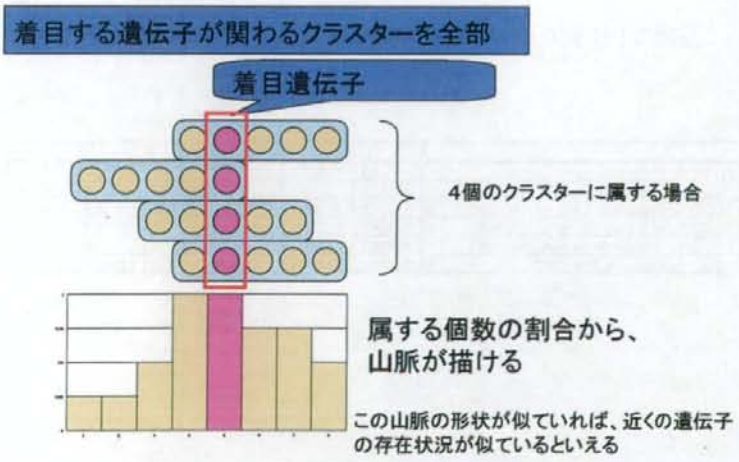
各遺伝子で



課題

1. クラスタリング対象外の取扱
2. ユニークとなった遺伝子の影響度
3. 計算量

同期率計算の山脈作成方法



クラスタリング対象実験

pj_no	prj_name	chemical	sfc_no	sim_no	clst_no	山脈定義用データ数
12	TTG030-L	N-Methylaniline	2	1	1	99,957,189
17	TTG040-L	cisplatin	1	2	1	49,920,803
33	TTG044-L	Clofibrate	1	1	1	6,386,957
42	TTG055-L	N-ethyl-N-nitrosourea	1	1	1	9,802,013
50	TTG052-L	all trans retinoic acid	1	1	1	6,418,489
67	TTG061-L	Paraquat 2nd	1	1	1	54,690,275

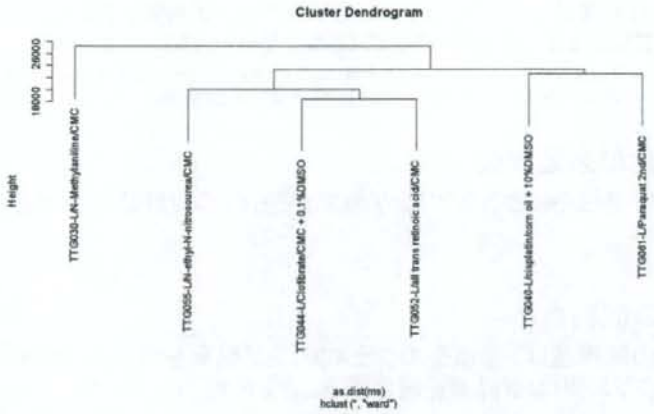
クラスタリング結果

- 距離マトリクス

		1	2	3	4	5	6
TTG030-L/N-Methylaniline/CMC	1	0.0	25037.6	24231.9	24014.0	24928.7	25780.9
TTG040-L/cisplatin/corn oil + 10%DMSO	2	25037.6	0.0	20701.1	22379.2	22281.1	22690.0
TTG044-L/Clofibrate/CMC + 0.1%DMSO	3	24231.9	20701.1	0.0	18535.3	19138.9	20669.2
TTG052-L/all trans retinoic acid/CMC	4	24014.0	22379.2	18535.3	0.0	20175.7	22762.7
TTG055-L/N-ethyl-N-nitrosourea/CMC	5	24928.7	22281.1	19138.9	20175.7	0.0	22461.4
TTG061-L/Paraquat 2nd/CMC	6	25780.9	22690.0	20669.2	22762.7	22461.4	0.0

クラスタリング結果

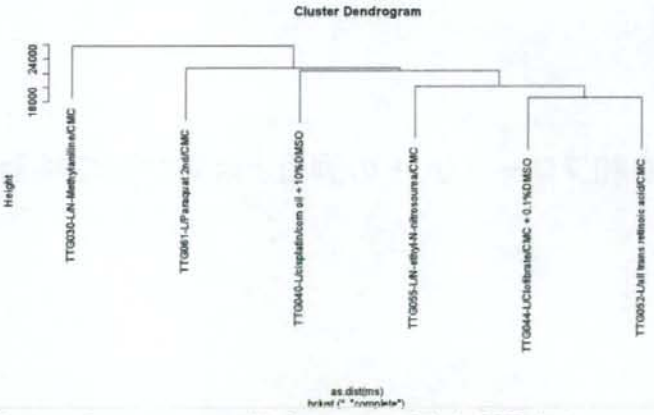
- Ward法



「Clofibrate」と「all trans retinoic acid」が近く、「N-Methylaniline」が違いという結果になった

クラスタリング結果

- Complete法



「Clofibrate」と「all trans retinoic acid」が近く、「N-Methylaniline」が違いという結果になった

まとめ(2)

- 優れた点
  - 山脈構造クラスタリングは、クラスタリング結果を用いるため、プロジェクト(化合物)ごとの偏差を受けにくい
    - 用量を含めたクラスタリング結果を用いるので、単純な化合物の影響を表現してクラスタに分割されるわけではないので結果の判読に留意が必要
- 改良が必要な点
  - 各遺伝子にクラスタリング結果が紐付くので計算量が大きい
- その他利点
  - 山脈構造は、重複型のクラスタリング結果を用いる。複数のクラスタリング結果を統合することも可能

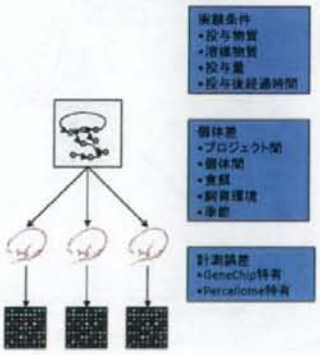
5. 飽和プローブセットの調査と対処方法の検討

180

背景と目的

- 背景
  - 平成18年度(STEP5)のテーマとして、化合物間クラスタリングの検討を行った。この中で、高発現のプローブセットで飽和により、正常に計測されていないと考えられるプローブセットが見つかっている。これらのプローブセットを特定し、適切な対処を行わないと計測結果を用いた研究にて誤った結論を導く可能性がある。
- 目的
  - 飽和していると考えられるプローブセットの特定とその対処方法を検討すること

仮説

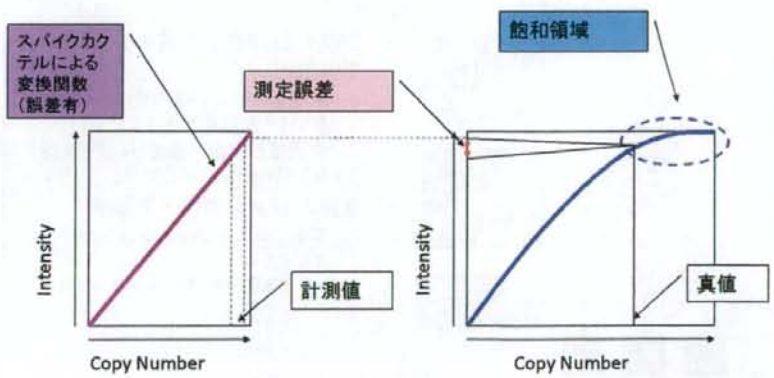


- 仮説1：計測誤差は個体差に比べて小さい
  - 発現量が小さい領域では成り立たない（一般に認知されている）
  - 発現量が大きい領域でも成り立たない(Step5における発見)
- 仮説2：観測誤差は不偏誤差
  - 不偏であるかのチェックがなされていない。
    - 発現量が大きい領域では、不偏ではない

現象

- 高発現域において正常に計測されていないと考えられるプローブセットが見つかった
- 飽和の可能性
  - 完全に飽和しているのではなく、飽和への漸近状態
  - 不偏の条件を満たさない
  - 特に飽和近辺では不偏の条件が明らかに崩れている
- 個体差と計測誤差を別個に取り扱うことが可能か？
  - N=3ならば、正規分布とみなすことが可能
  - 計測誤差が個体差に比べて小さければ、計測誤差が0であるとみなしても影響は少ない
    - 既存の統計手法で分析可能なはず
- 計測誤差が個体差よりも大きく、不偏の条件を満たさないならば、特別の取扱が必要である
  - 低発現域では、計測誤差（チップ偏差）は個体差よりも大きい。
    - 既知の課題として研究対象になっている
  - 中発現域では、計測誤差が不偏の条件を満たすと仮定できる？
    - 既存の分析手法で可能と考えられる
  - 高発現域は、飽和しつつある領域で、計測誤差が大きく、かつ、不偏の条件を満たしていない
    - Percellomeにより量を計測ためには補正が必要と考えられる。

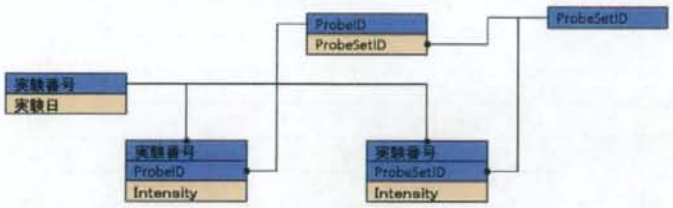
Percellome変換の基礎



- 初期のデータでは、メーカーSOPをそのまま採用していたため、スパイク自身が飽和していた可能性がある
- 飽和領域はプローブごとorプローブセットごとかの検討を行う

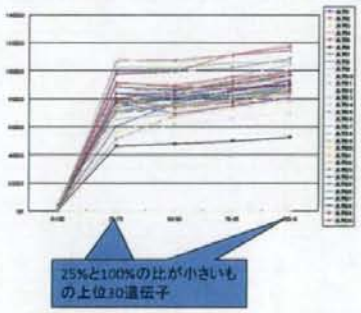
データ構造

- データ分析を行うにあたって、Probe単位のデータを取り扱うためのデータ構造を示す。



データ状況把握（飽和プローブセット）

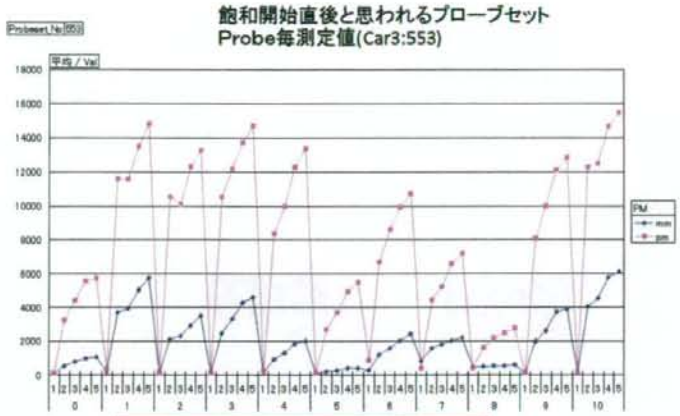
LBM（LiverとBrainの混合サンプル）のLiver100%において飽和していると考えられるProbeSetに着目し、状況を確認する



輸送関連の遺伝子が多い

データ状況把握（飽和プローブセット）

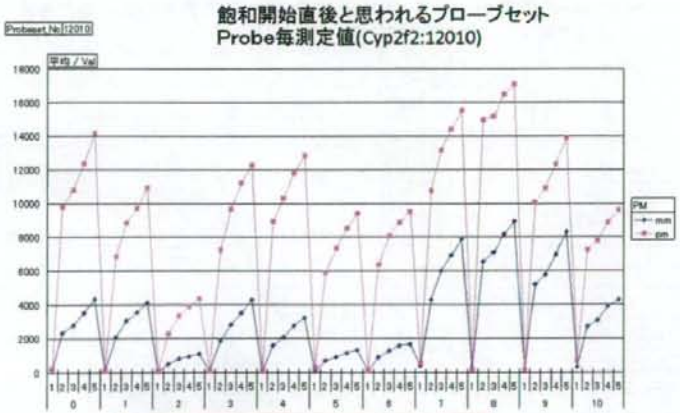
飽和開始直後と思われるプローブセット
Probe毎測定値(Car3:553)



飽和の状況が各プローブで読み取れる

データ状況把握（飽和プローブセット）

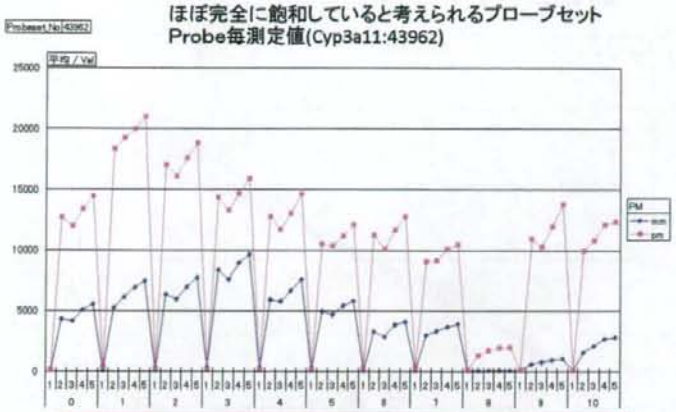
飽和開始直後と思われるプローブセット
Probe毎測定値(Cyp2f2:12010)



飽和の状況が各プローブで読み取れる

データ状況把握(飽和プローブセット)

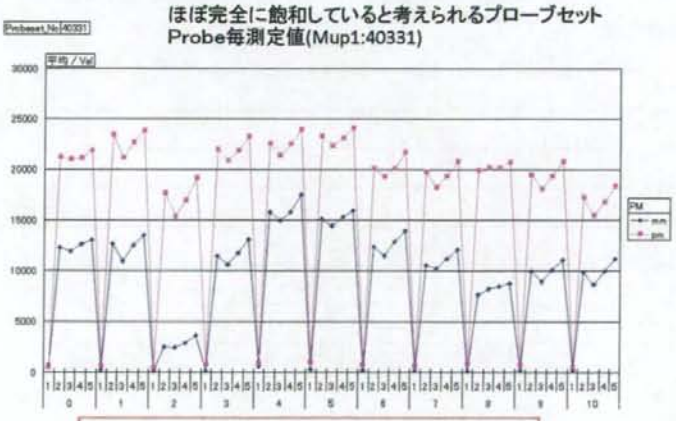
ほぼ完全に飽和していると考えられるプローブセット
Probe毎測定値(Cyp3a11:43962)



プローブ単位では、飽和したものと飽和しつつあるものが混在している。

データ状況把握(飽和プローブセット)

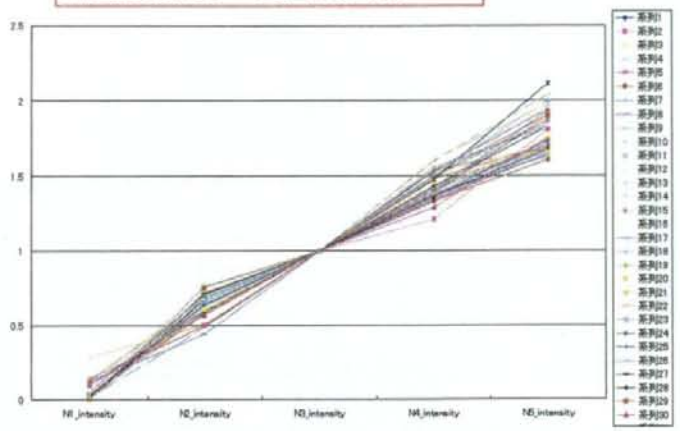
ほぼ完全に飽和していると考えられるプローブセット
Probe毎測定値(Mup1:40331)



プローブ単位でも完全飽和と考えられる状況である

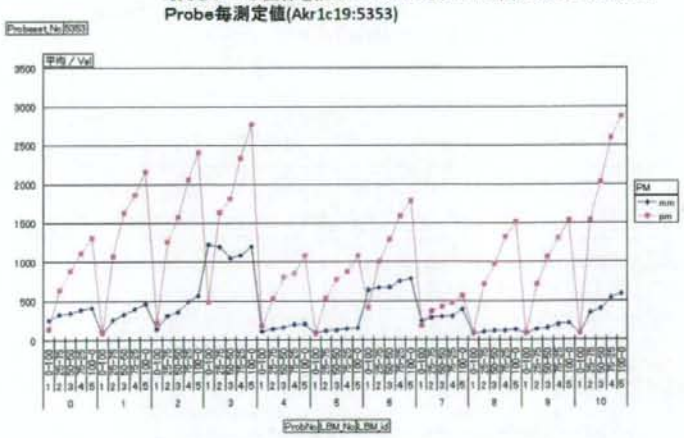
データ状況把握(線形的増加)

線形性が認められるプローブセットを見つけ出す



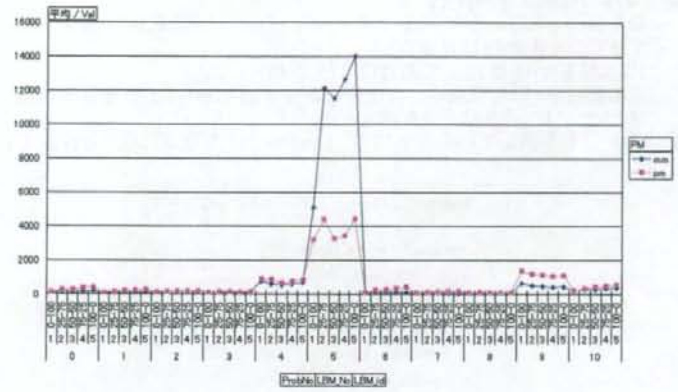
データ状況把握(線形的増加)

最もきれいな直線を描くProbesetでも、Probe単位ではブレがある。
Probe毎測定値(Akr1c19:5353)



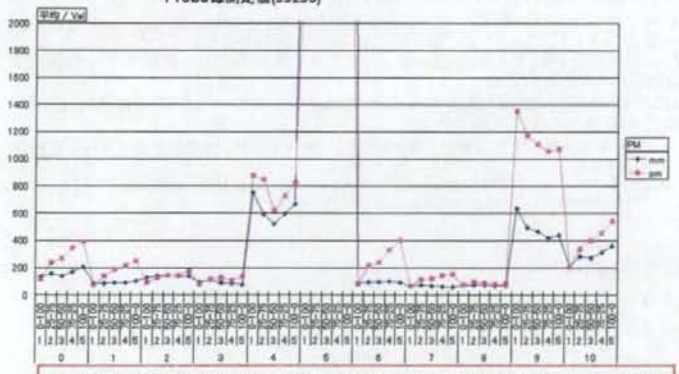
データ状況把握（線形的増加）

MMの方が大きいProbeが存在する。他の遺伝子に合致した可能性がある。
Probe毎測定値(35256)



データ状況把握（線形的増加）

幾つかのProbeで、直線性が見られる
Probe毎測定値(35256)



MMの大きなプローブは無視され、小さい値の線形性がみられるプローブから、線形性の強い遺伝子として検出された。

基本的アイデア①:

Langmuir isotherm equation

- Irving Langmuirによって1918年に導出された理論的な吸着等温式である。以下のような仮定を持っている。
  - 吸着媒には有限な数Nの吸着サイトがあり、そこだけで吸着質分子と結合する。
  - すべての吸着サイトは等価である。
  - 1つの吸着サイトは1つの吸着質分子としか結合しない。
  - 空の吸着サイトM、気相中の吸着質S、吸着サイトに結合した吸着質M-Sの間にM + S ⟷ M-Sの化学平衡が成立する。
- 各プローブが飽和している場合には、Langmuirの方程式に従うと仮定する

$$\log(PM_{pjl}) = \log\left(I_p \frac{k_p c_j}{1+k_p c_j} + bg_p\right) + \varepsilon_{pjl}$$

$p = 1,\ldots,P$: $probe\_index$

j : $concentration\_index$

c : concentration

$l = 1,\ldots,L$: replicate_index

I : Saturation_Intensity

k : equilibriumu_constant

bg : background_component

基本的アイデア②

AIC(Akaike Information Criteria)

- すべてのプローブが飽和しているとは考えられない。また、飽和していたとしても、一定の値を示しているため、飽和しているかを判別できないことが考えられる。これらの状況を判断するため、Langmuirモデル、線形モデル、定数モデルを選択するためにAICを使用する
- AICとは、元統計数理研究所所長の赤池弘次によって1971年に考案された統計モデルの良さを示す指標である
  - モデルは複雑にすれば細かく適合させることができるが、偶然の影響を受ける可能性も高くなる
  - Akaike, Hirotugu (1974). "A new look at the statistical model identification". IEEE Transactions on Automatic Control 19 (6): 716–723.

$AIC = -2\log(L) + 2k$

L : 尤度

k : パラメータ数

確率密度関数 : $P_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

尤度関数 : $L_x(\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$L_{x_1,\ldots,x_n}(\mu,\sigma) = \prod_{i=1}^{n} L_{x_i}(\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\log L_{x_1,\ldots,x_n}(\mu,\sigma) = \log\left(\prod_{i=1}^{n} L_{x_i}(\mu,\sigma)\right) = \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)$$

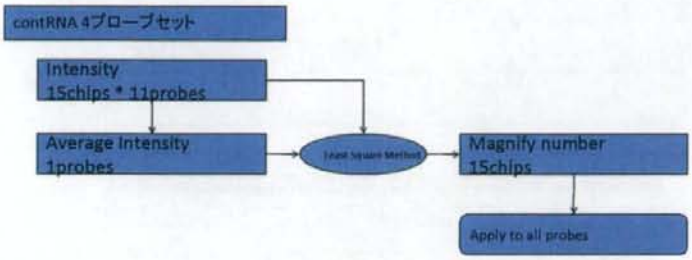
$$= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)$$

$$= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma}}\right) + \sum_{i=1}^{n} \log\left(e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)$$

$$= -n\log\sqrt{2\pi\sigma} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

チップごとの標準化（contRNA）

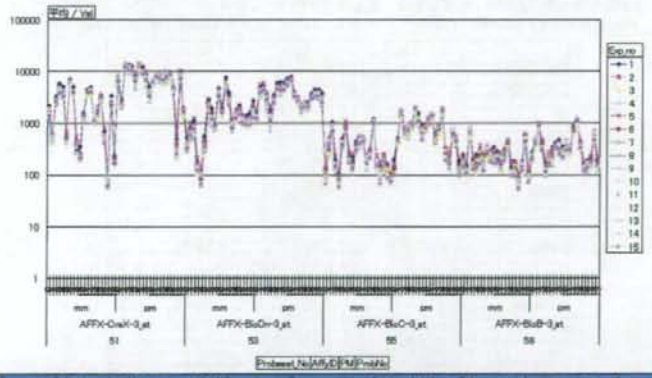
- チップごとの標準化のためにControl用RNAのデータ(contRNA)を使用する
  - 各プローブに対する平均値を求め、それらの平均値に近づくように（最小二乗法）にチップごとの変換倍率を求め、その値をチップの全プローブにかける。
  - 対数領域で計算する
  - ContRNAはハイブリ液に一定量を加えたE.Coli.由来のmRNAの測定データを示す



チップごとの標準化（contRNA）

チップごとの標準化（contRNA）

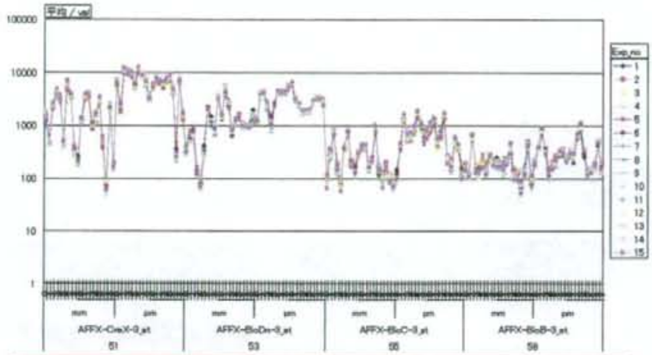
- 露光等の問題により、チップ毎にバイアス値が異なる。このような測定誤差の変動を補正し、標準化を行う。
- contRNAのプローブセットを使用する。



プローブ毎に変動しているが、チップ(Exp_no)毎に似たパターンを示している。
これらの平均値への誤差が最小になるように、定数倍することで標準化を行う。

チップごとの標準化（contRNA）

チップごとの標準化（contRNA）の実施後のグラフ



わずかであるが、線の重なりが増え、チップ間のばらつきが小さくなったことが読み取れる

Langmuir式フィッティング検証

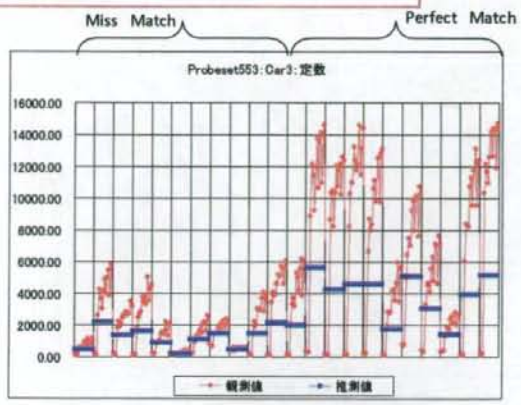
Langmuir式にフィッティング可能かいくつかのプローブセットで試してみる
実験対象プローブセット

- 計算を試みるプローブセット
  - 飽和、線形と思われたプローブセット
  - LBM、50%:50%における、発現値順位、最上位、10%、50%、80%
  - Percellome Spike Cocktail (内2プローブセットは、プローブ数が異なるので今回は省略)

		No	PrbNo	Avg	SD	AffyID	GeneSymbol	CV
飽和	0.1%	46	5509	7700.78	839.10	1460256_at	Cox3	9%
	0.2%	86	12010	6695.40	230.26	1448792_a_at	Cyp2e	3%
	0.1%	80	40867	7099.58	279.33	1415809_at	Cyp3a11	4%
	0.0%	16	40331	6888.57	294.42	1420465_a_at	Mup1	3%
線形	8.0%	3684	5053	777.06	27.06	1455454_at	Akr1c19	3%
	38.1%	17178	35256	60.56	6.63	1426540_at	Otc	11%
最発現値	0.0%	1	42921	10726.10	273.24	1417950_a_at	Apoa2	2%
	0.0%	2	43215	10660.20	406.36	1417556_at	Fabp1	4%
	0.0%	3	26686	10519.77	445.44	1434110_x_at	Mup1	4%
	0.0%	4	5606	10604.17	348.16	1456201_x_at	Apoa1	3%
	0.0%	5	41509	10193.04	234.14	1419233_x_at	Apoa1	2%
中発現値	10.0%	4512	46063	470.25	49.31	1415708_at	Tug1	10%
	10.0%	4514	917	470.18	10.17	1458880_a_at	1110008P14Rik	2%
	10.0%	4515	38843	469.79	15.76	1421863_at	Pbsn1	3%
	10.0%	4516	22322	469.60	69.73	1429845_at	Zcd2	15%
	10.0%	4517	6059	469.59	6.77	1454748_at	Nnat1	1%
低発現値	50.0%	22551	36922	49.64	4.76	1423674_at	Wdr33	10%
	50.0%	22552	8806	49.64	11.65	1454005_at	–	23%
	50.0%	22553	40644	49.62	11.89	1417127_at	Mxd1	24%
	50.0%	22554	18788	49.61	10.41	1444040_at	AU041130	21%
	50.0%	22555	19740	49.61	11.62	1441268_at	Ube2	23%
極低発現値	80.0%	36082	36480	13.48	3.65	1422301_a_at	Tnfrsf18	27%
	80.0%	36083	20485	13.47	2.67	1440311_at	Socs1	20%
	80.0%	36084	4142	13.47	0.71	1456644_at	5730508L13Rik	5%
	80.0%	36085	30240	13.47	9.07	1400558_at	Spag9	67%
	80.0%	36086	14458	13.47	11.50	1446308_at	–	85%
Percellome Spike	18.3%	7838	3	285.81	4.84	AFFX-TrpnX-3_at	–	2%
	8.4%	3945	30	702.92	27.81	AFFX-r2-Bs-dap-3_at	–	4%
	2.4%	1069	37	1832.62	77.83	AFFX-PheX-3_at	–	4%
	0.5%	208	27	4801.25	362.87	AFFX-r2-Bs-lys-3_at	–	7%
	0.0%	15	21	8929.21	358.22	AFFX-r2-Bs-thr-3_s_at	–	4%

Langmuir式フィッティング検証　飽和遺伝子

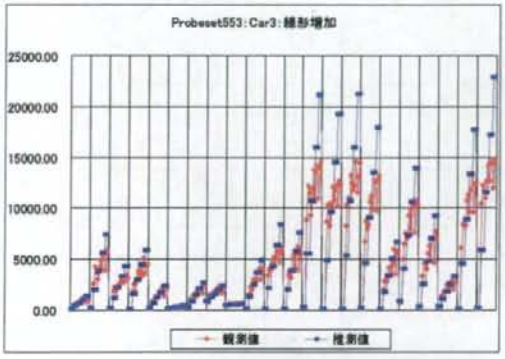
飽和遺伝子:553:Car3:定数



AIC=2568.4

Langmuir式フィッティング検証　飽和遺伝子

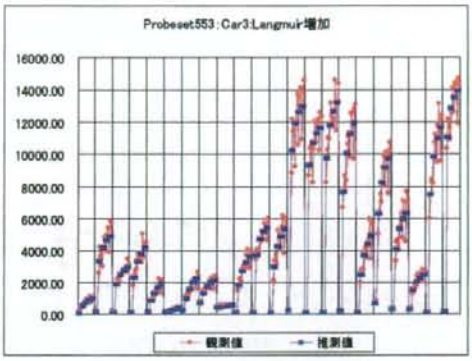
飽和遺伝子:553:Car3:線形(増加)



AIC=927.6

Langmuir式フィッティング検証　飽和遺伝子

飽和遺伝子：553：Car3：Langmuir（増加）

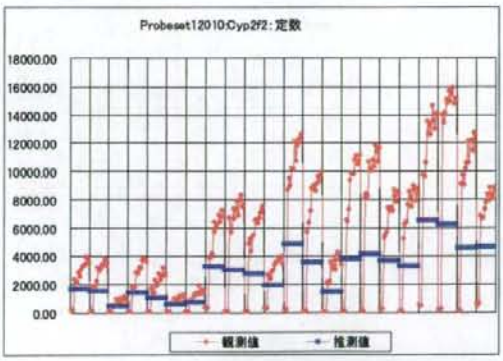


最もAICの小さい最適モデル

AIC=714.3

Langmuir式フィッティング検証　飽和遺伝子

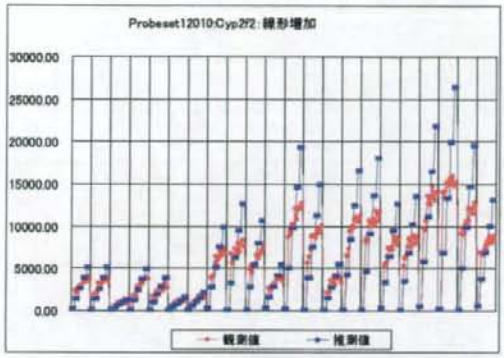
飽和遺伝子：12010：Cyp2f2：定数



AIC=2768.1

Langmuir式フィッティング検証　飽和遺伝子

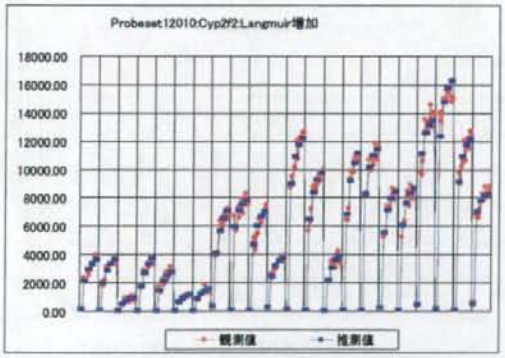
飽和遺伝子：12010：Cyp2f2：線形(増加)



AIC=930.2

Langmuir式フィッティング検証　飽和遺伝子

飽和遺伝子：12010：Cyp2f2：Langmuir(増加)



最もAICの小さい最適モデル

AIC=590.1

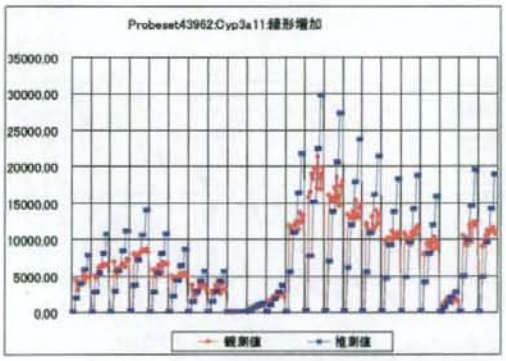
Langmuir式フィッティング検証　飽和遺伝子

飽和遺伝子:43962:Cyp3a11:定数

AIC=2995.2

Langmuir式フィッティング検証　飽和遺伝子

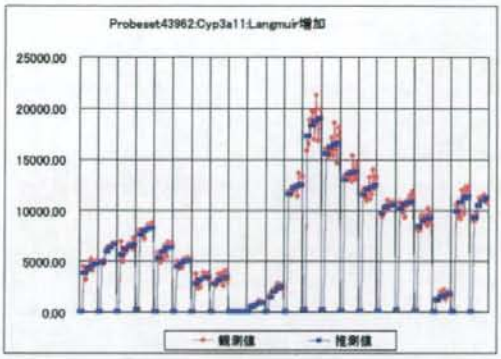
飽和遺伝子:43962:Cyp3a11:線形(増加)



AIC=1076.8

Langmuir式フィッティング検証　飽和遺伝子

飽和遺伝子:43962:Cyp3a11:Langmuir(増加)

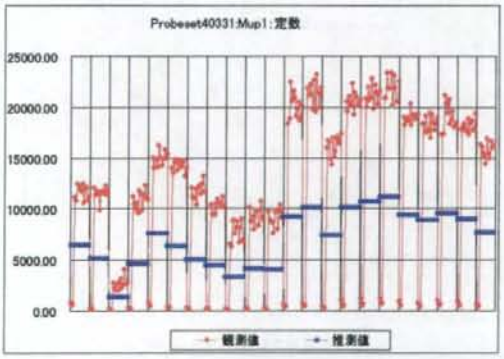


最もAICの小さい最適モデル

AIC=622.9

Langmuir式フィッティング検証　飽和遺伝子

飽和遺伝子:40331:Mup1:定数



AIC=2865.0