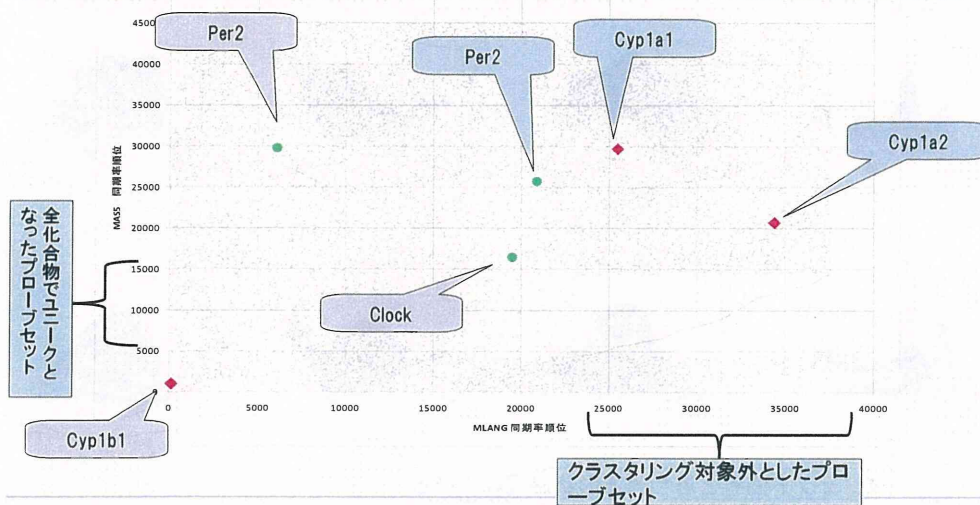


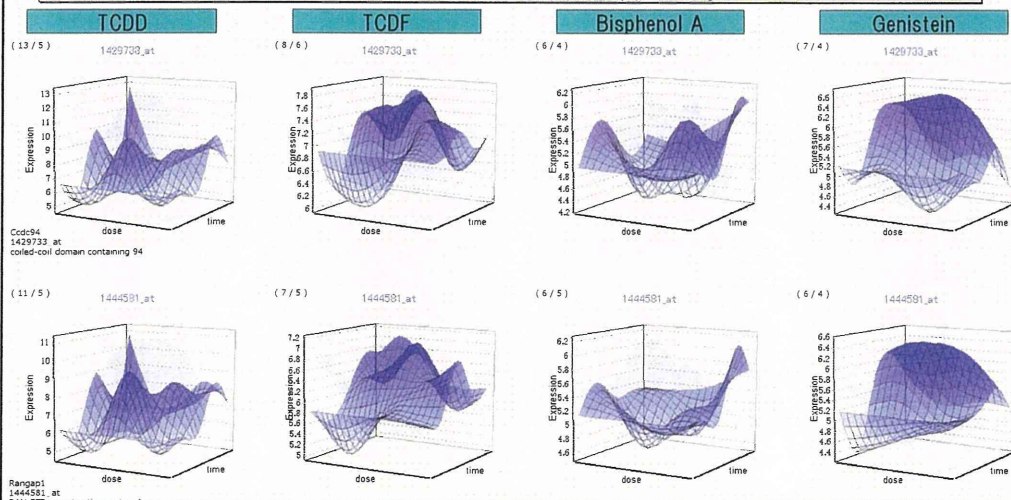
5. 3. 同期率計算結果比較（結果順位）

同期率の順位に対して、幾つかの代表的な遺伝子をプロットした。



5. 4. 同期率計算結果 MLANG同期率上位形状確認

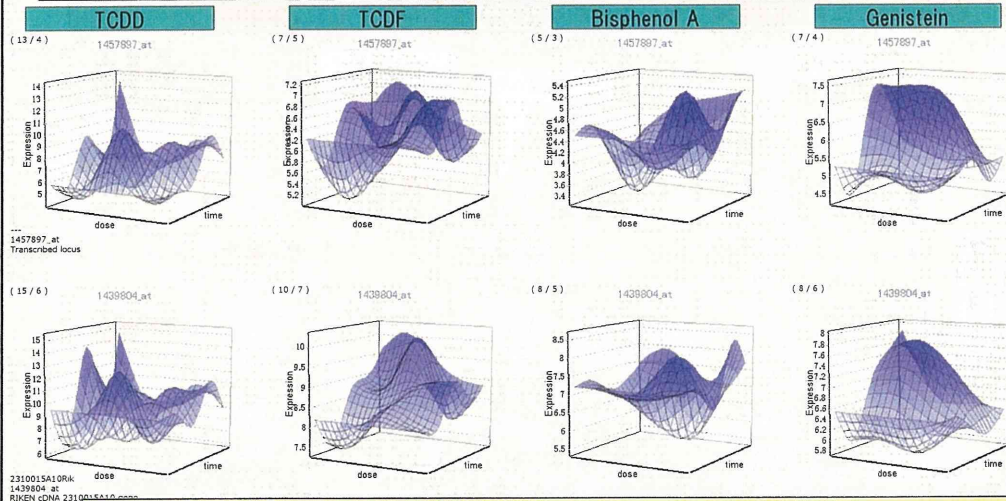
MLANG同期率上位となったプローブセットの形状を確認する。



5. 4. 同期率計算結果 MLANG同期率上位形状確認



MLANG同期率上位となったプローブセットの形状を確認する。



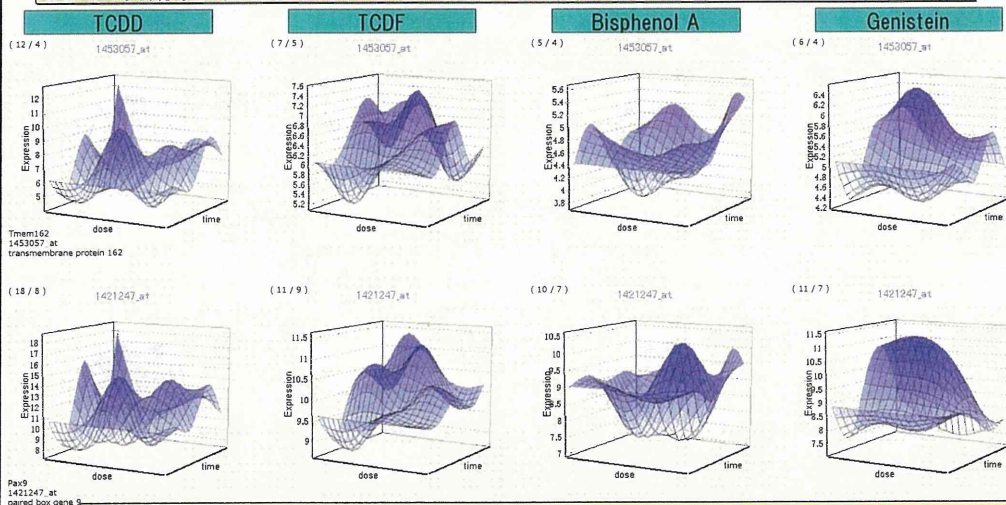
各プロジェクト固有と思われる形状が現れている。

Copyright(C)2009-2010 NTT DATA Corporation

5. 4. 同期率計算結果 MLANG同期率上位形状確認



MLANG同期率上位となったプローブセットの形状を確認する。

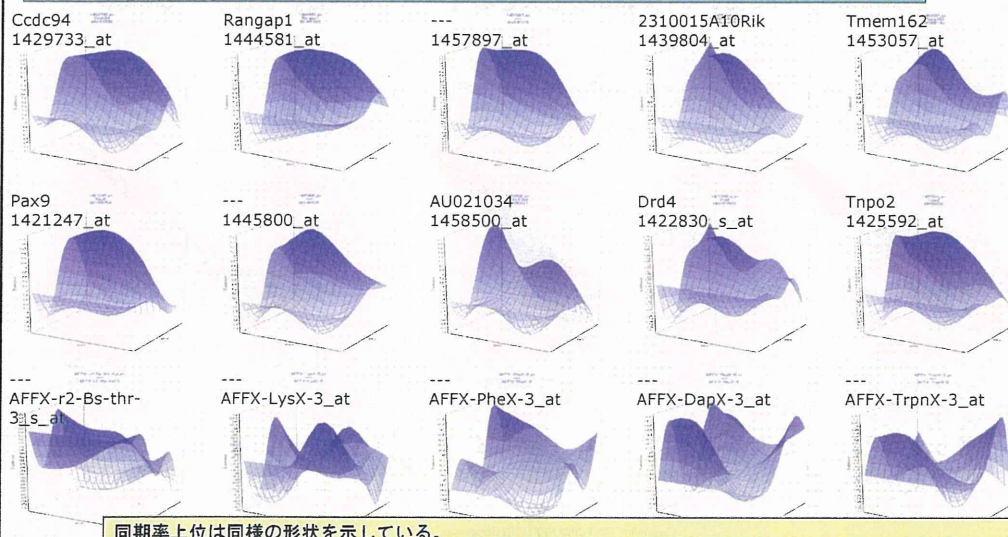


各プロジェクト固有と思われる形状が現れている。

Copyright(C)2009-2010 NTT DATA Corporation

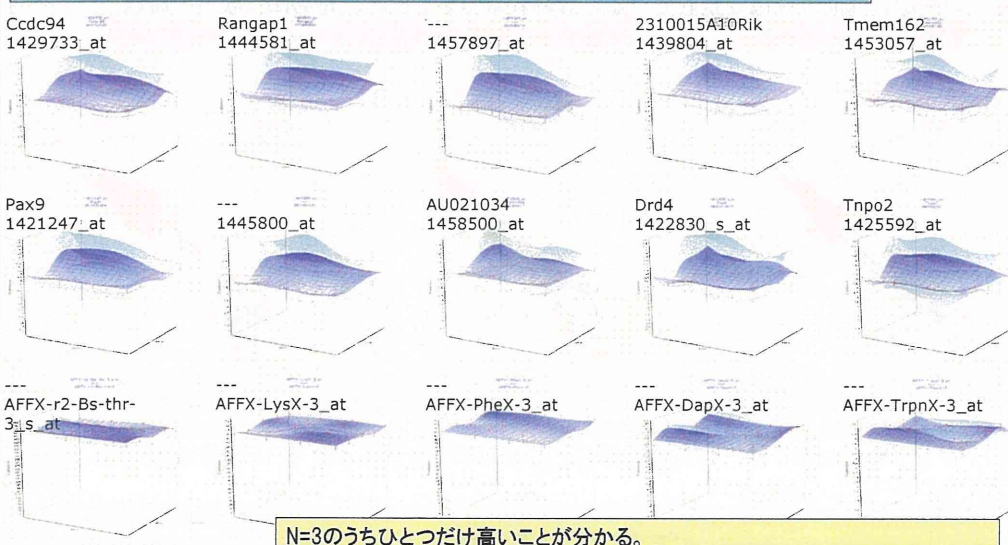
5. 5. MLANG同期率上位およびGSC形状確認 TTG048-L

TTG048-Lは滑らかな形状を示しており、何らかの生物学的な意味が存在するようにもとれる。同期率上位とスパイクの形状を確認する。



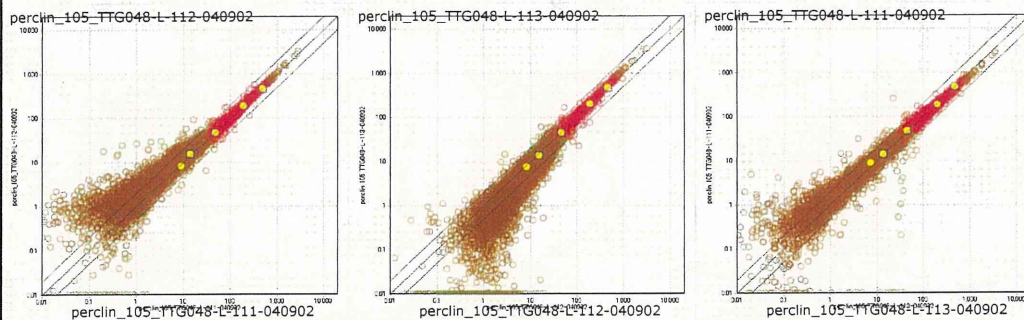
5. 5. MLANG同期率上位およびGSC形状確認 TTG048-L (最小値0固定)

最小値を0と固定し、素点を表示した。N=3のうちひとつだけ高いことが分かる。



5. 6. TTG048-L(8H-MidDose)

TTG048-Lの8H-MidDoseにおいて1実験が少し大きな値を示していた。状況を確認する。

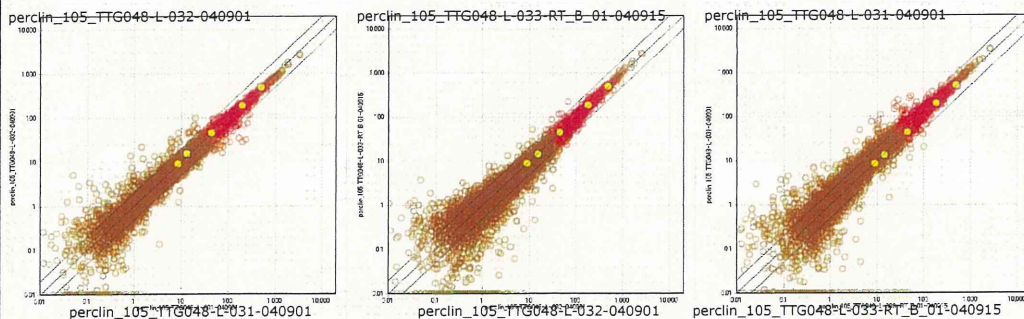


チップ番号112は、低発現域において、他の2チップよりも大きめの値を示しており、対角線からずれている。

Copyright(C)2009-2010 NTT DATA Corporation

5. 6. TTG048-L(2H-MidDose)

TTG048-Lの2H-MidDoseは安定していると思われるので、比較のために、状況を確認する。



2チップ間の比較では、対角線の両側にほぼ均等に広がっており、偶然の結果と考えられる。

Copyright(C)2009-2010 NTT DATA Corporation

6. まとめ

MLANG補正の結果を用いて、同期率計算を実施したところ、次の現象が発生した。

同期率上位に低発現のプロジェクト固有と考えられる形状が現れていた。プロジェクト固有の形状となったプローブセットは、単一のクラスタを形成し、同期しているとみなされたと考えられる。

同期率中程度の値を示すプローブセットとして、clockやPer2などがあり、適切な計算がなされていると考えられる。

同期率計算そのものは当初の目論見どおりに機能していることが確かめられたと考えられる。

プロジェクト固有のパターンが発生しにくいアルゴリズムを組み立てるとともに、発生した場合に、キャンセルするような仕組みが必要であり、今後の課題である。

委託研究報告書 (STEP9)

低発現領域の精度向上のためのアルゴリズム生成研究

Copyright(C)2010-2011 NTT DATA Corporation

1. テーマ

平成21年度の研究では、化合物クラスタリングや同期率計算の精度向上のための検討を行った。この中で飽和現象による影響(H21年度に対処済み)とは独立で、低発現遺伝子の精度もこれらの解析精度に影響を及ぼしていた。平成22年度では、低発現領域における精度向上のための補正方法を検討した。

Copyright(C)2010-2011 NTT DATA Corporation

2.存在量と蛍光値の比例関係を逸脱させる要因

存在量と蛍光値の比例関係(直線性)を逸脱させる要因として、次のことが想定できる

	支配的物理化学的現象	原因・解決策
高レベル	飽和現象 saturation	Langmuir等温式
中レベル	半特異的結合 semi-specific hybridization	MLANG基本概念 competitive hybridization
低レベル	非特異的結合	MLANGで計算不可能な大量な半特異的結合の存在
	光学的背景レベル	偶発的事象が存在する可能性
その他	2次構造	MLANG係数学習による解決
	配列依存結合力	MLANG係数学習による解決
	RNA同士の結合	未知(小さいと想定)
	断片化・増幅	MLANG係数学習による解決
	RNA枯渇	未知(小さい)

高レベル、中レベルに対する処置を講じてきたが、低レベルでの特性を考慮した、補正手法を構築する

Copyright(C)2010-2011 NTT DATA Corporation

3.MLANGのmRNA/プローブの結合名称

MLANGにおける結合を次のように分類し、名称を与えている。

特異的結合 Specific hybridization	•25-merの配列一致
半特異的結合 semi-specific hybridization	•端12-mer以上、中央13-mer以上～24-mer以下の配列一致
非特異的結合 Non-specific hybridization	•端12-mer未満、中央13-mer未満の配列一致

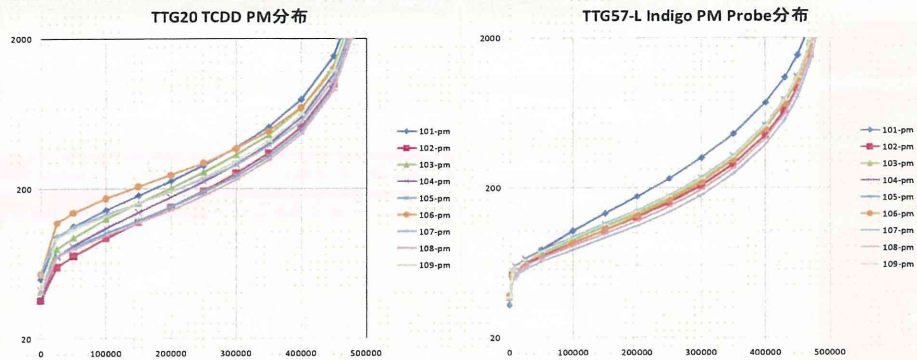
プローブの設計位置以外を含む、全RNAの全ての位置と全プローブの間での関係をチェックして、半特異的結合が発生し得る組み合わせを抽出した

マイクロアレイの設計情報、コンピュータの処理能力を考慮し、このように分類している。

Copyright(C)2010-2011 NTT DATA Corporation

4. マイクロアレイProbe-levelの蛍光値分布

マイクロアレイ実験におけるPerfectMatchプローブの蛍光値の分布状況を、横軸に蛍光値の順位、縦軸に蛍光値を示すグラフで確認した。

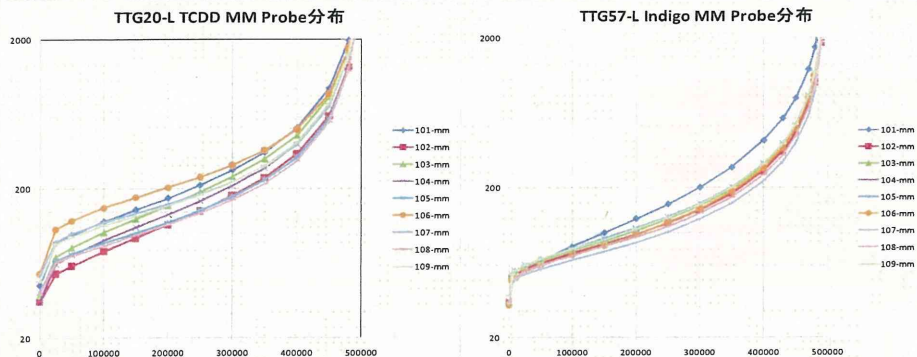


TTG20は、ベースラインが異なるが、似たような分布をしている
TTG57は、1チップを除き、ほぼ同じ値となっている。

Copyright(C)2010-2011 NTT DATA Corporation

4. マイクロアレイProbe-levelの蛍光値分布

マイクロアレイ実験におけるMisMatchプローブの蛍光値の分布状況を、横軸に蛍光値の順位、縦軸に蛍光値を示すグラフで確認した。



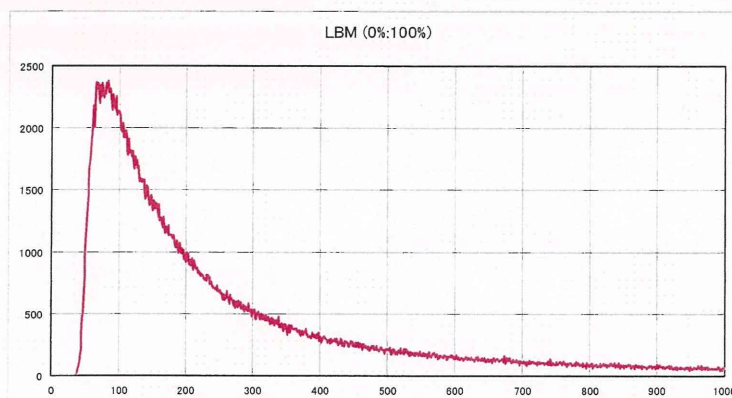
PMとほぼ同じような形状を示す

Copyright(C)2010-2011 NTT DATA Corporation

4. マイクロアレイ Probe-level の蛍光値分布



低発現域の蛍光強度ヒストグラムで確認する



多少の変動を含むが、何らかの規則的な分布をしていると思われる

Copyright(C) 2010-2011 NTT DATA Corporation

5. べき乗分布によるアプローチ



偶然に発達したものはべき乗分布になるものが多い

山火事の規模

地震の規模

都市のサイズ

通販売上

生体内のRNAの存在量は冪乗分布になるのではないかな？

マイクロアレイの蛍光値もべき乗分布すると仮定し、背景補正を提唱している研究者もいる

Wu 2004 : A Model-Based Background Adjustment for Oligonucleotide Expression Arrays

Copyright(C) 2010-2011 NTT DATA Corporation

5.べき乗分布によるアプローチ

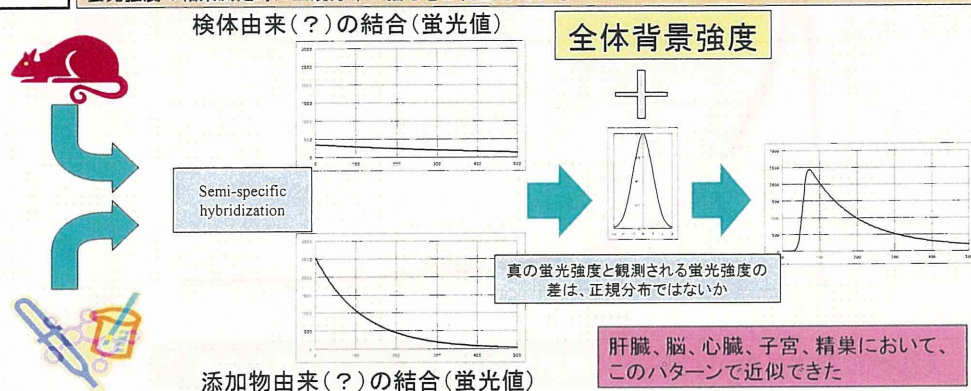
次の仮説を用いて、蛍光値の分布を近似できるか確認する

仮説

発現している特異的結合による蛍光強度はべき乗分布に従う

半特異的結合による蛍光強度は指数分布に従う

蛍光強度の結果測定時に正規分布の揺らぎが発生している



Copyright(C)2010-2011 NTT DATA Corporation

5.べき乗分布によるアプローチ

更に、飽和による影響を考慮した蛍光強度の分布を探る

1. 生体由来の分布(べき乗分布と思われる)
2. 添加物のクロスハイブリダイゼーションによる分布(指数分布)
3. スキャナー誤差(正規分布)
4. プローブ飽和によるLangmuir曲線

仮説

観測蛍光強度 = 真の蛍光強度 + 背景強度 + 観測誤差(正規分布)

真の蛍光強度 = 目標RNA蛍光強度 + 非目標RNA蛍光強度

べき乗分布

指数分布

飽和による影響を受ける

Copyright(C)2010-2011 NTT DATA Corporation

5.べき乗分布によるアプローチ

べき乗分布がクロスハイブリダイゼーションによる変形を受けている分布の数式モデル

べき乗分布

$$f_1(x) = \begin{cases} a_1 \cdot x^{k_1} & x > 0 \\ a_1 \cdot x_0^{k_1} & 0 < x \leq x_0 \\ a_0 & x = 0 \\ 0 & x \leq 0 \end{cases}$$

通常のべき乗分布

低レベルの一定値

結合しないプローブ

定義外

指数分布

$$f_2(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

飽和

レベルの低い部分では、線形であるとみなす

合成

$$h(s) = \int_{-\infty}^{+\infty} (f_1(s-t) + f_2(t)) dt$$

$$d(x) = \int_{-\infty}^{+\infty} h(x+s-bg) \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(s)^2}{2\sigma^2}\right) \right) ds$$

この方程式の解を求めるのは、非常に困難であった。数値計算として求めようとしても、収束せず発散してしまい求められなかった。

Copyright(C)2010-2011 NTT DATA Corporation

5.べき乗分布によるアプローチ

MLANGにおいて、添加物と各プローブの間の係数は用意してある。蛍光値が低くて、学習が進まないのでは、

特別な学習方法が必要

検体無しのハイブリダイゼーション実験（？）

複数実験を用いて、活動していないRNAの場合の輝度を求め、係数を求める

冪乗分布の広がりを表す係数で正規化し、複数の臓器の実験間で比較した。発現していないような状況では、同じ値を示しているようである（発現していない状況を定義するのが困難）

Copyright(C)2010-2011 NTT DATA Corporation

5. べき乗分布によるアプローチ

- ・ 生命内のRNAは、遺伝子カスケードというネットワークを形成している。自然発生的なネットワーク構造である。自然発生的なネットワーク構造はべき乗分布になることが知られており、その存在量はべき乗分布となっていると思われる。
- ・ Semi-specific hybridizationによる蛍光強度、つまり、洗浄工程での残る可能性はゼロではない。洗浄工程の残りではないか？
 - 合致長が1増えると、結合力が1増加し、存在比は4分の1になる。
 - 特に、グリッド識別用の添加物の残りではないか？

べき乗則では、低レベルで膨大な数が存在するが分布だが、プローブの数に限りがある

汎用的な分布法則が見つからない

方針転換

実験結果をもとに係数を見つけ出す

Copyright(C)2010-2011 NTT DATA Corporation

6. 一定比で添加している系列によるアプローチ

- ・ 既定の比率で実験の際に添加しているRNAが2系列存在する。
 - GSC
 - BioB,C,D,Cre
- ・ これらの系列を用いて、背景補正が可能か検討する。

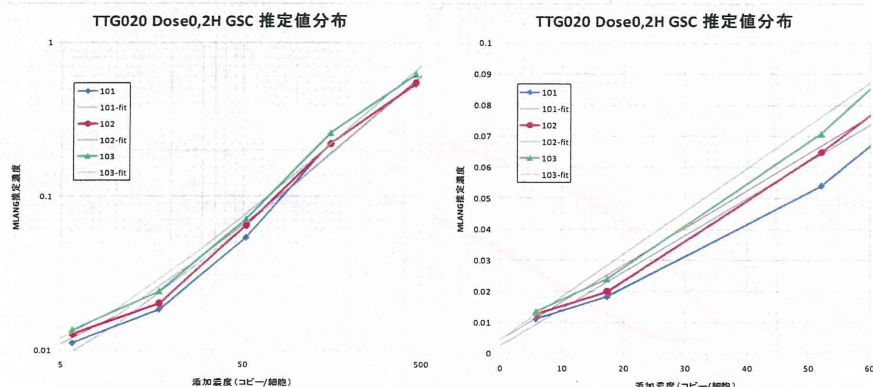
GSC(コピー/細胞)			BioBCD (pM)	
Liver用				
	新	旧		
Trpn	1.450	5.800	BioB	1.5
Dap	4.325	17.300	BioC	5.0
Phe	13.000	52.000	BioD	25.0
Lys	39.000	156.000	BioCre	100.0
Thr	117.025	468.100		

Copyright(C)2010-2011 NTT DATA Corporation

6.一定比で添加している系列によるアプローチ TTG20 GSCの添加量と推定値



原点を通る直線に乗るように、推定結果に何らかの数値を加える



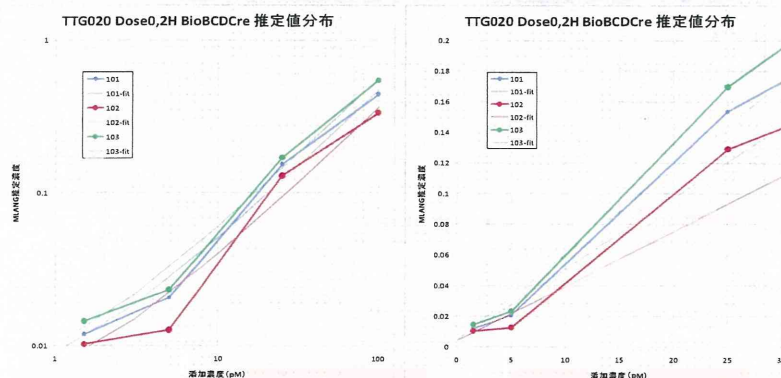
細い線は補正用直線
GSCは、最低濃度のTrpnは、高い値を示す傾向がある。

Copyright(C)2010-2011 NTT DATA Corporation

6.一定比で添加している系列によるアプローチ TTG20 BioBCDの添加量と推定値



原点を通る直線に乗るように、推定結果に何らかの数値を加える

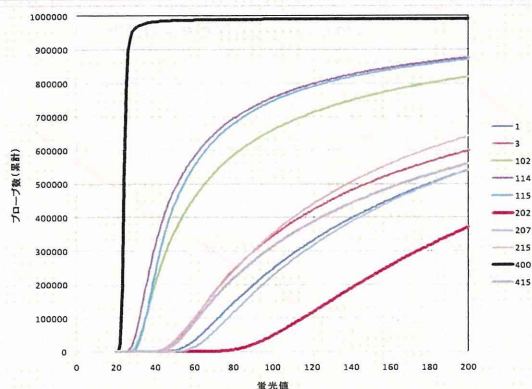


細い線は補正用直線
BioBCDも最低濃度が高めにでる傾向がある。

Copyright(C)2010-2011 NTT DATA Corporation

7.背景ターゲットによるアプローチ

非常に暗いプローブが値を有しているのは、光学的な状況から差が生じるものと考えた。しかし、暗いプローブの蛍光値は、光学的な背景の差よりも、クロスハイブリダイゼーションから発生する蛍光値の方が強い可能性も考えられる。



400(黒線): 検体なしの実験
BIOBCDなどが付着して、蛍光している可能性が存在する

202(赤線): EuLBM-25-75-050518
全スポットで明るくなっている。全体的にクロスハイブリダイゼーションの量が多くなった可能性がある

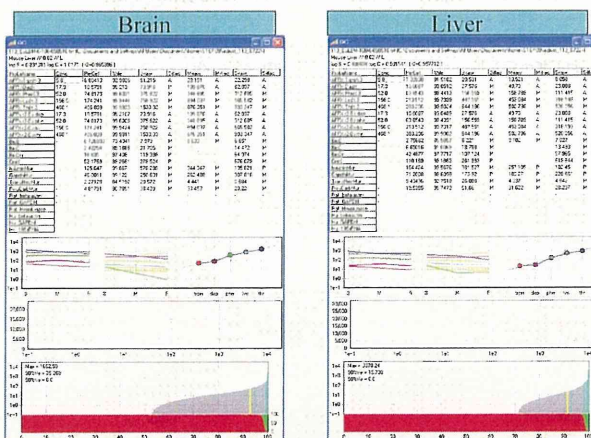
アイデア

下から5%の明るさのプローブに比例した量をMLANG上の背景物質(全プローブと結合し得る仮想ターゲット)として与える

Copyright(C)2010-2011 NTT DATA Corporation

7.背景ターゲットによるアプローチ

背景ターゲットを与え、計算を実施した。



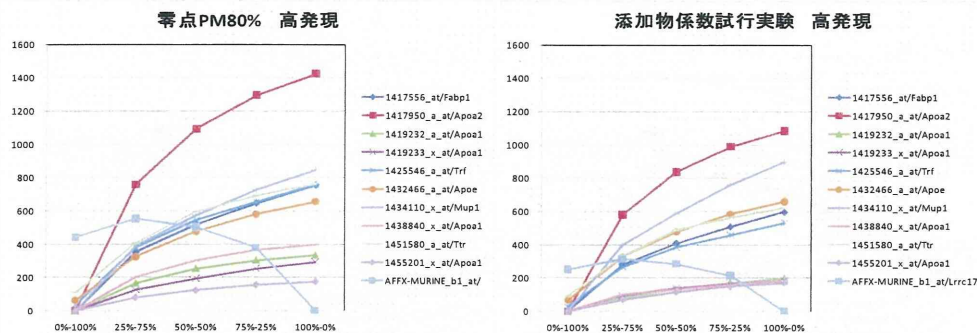
多くのプローブセットが濃度0と推定された。
通常、単一の臓器で使用する遺伝子は限定されており、半分程度が0となっているのは正しい結果と考えられる

Copyright(C)2010-2011 NTT DATA Corporation

7.背景ターゲットによるアプローチ 零点を与えた場合vs.背景ターゲットを与えた場合



高発現域



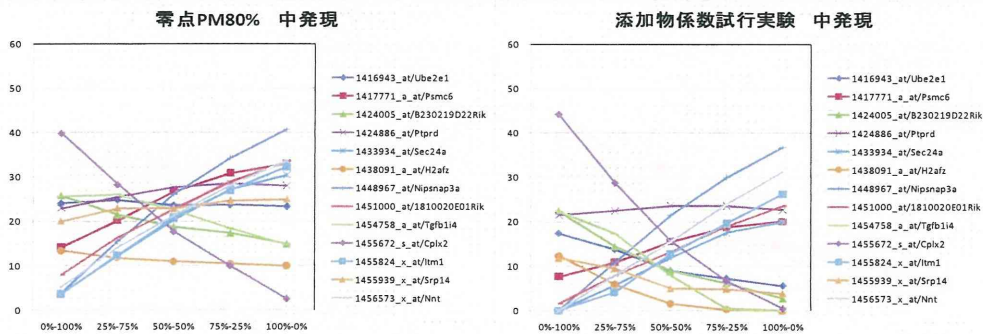
背景ターゲットを与えたほうが、全体的に低い。Apoa1関連プローブセットがほぼ同じ値となった。

Copyright(C)2010-2011 NTT DATA Corporation

7.背景ターゲットによるアプローチ 零点を与えた場合vs.背景ターゲットを与えた場合



中発現域

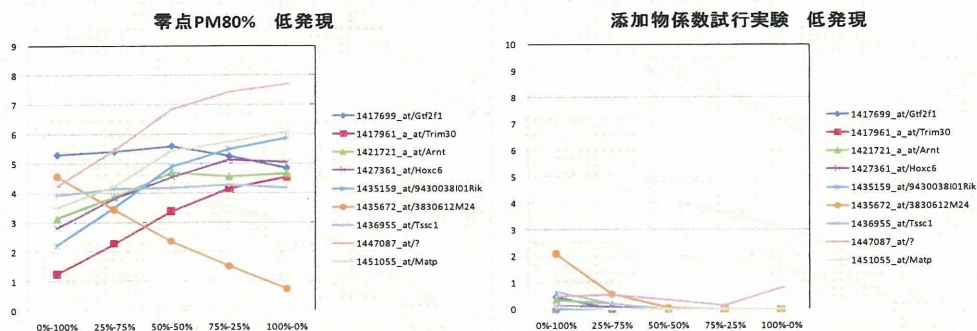


背景ターゲットを与えたほうが、全体的に低い。臓器ごとの発現遺伝子が明確になったものと思われる

Copyright(C)2010-2011 NTT DATA Corporation

7.背景ターゲットによるアプローチ 零点を与えた場合vs.背景ターゲットを与えた場合

低発現域

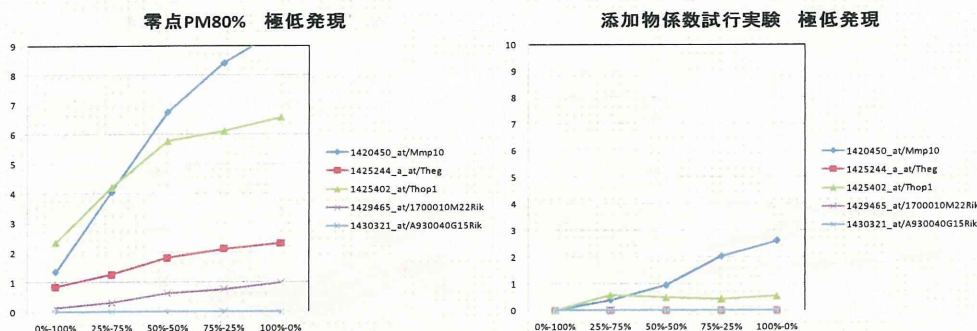


背景ターゲットを与えたほうが、全体的に低い。

Copyright(C)2010-2011 NTT DATA Corporation

7.背景ターゲットによるアプローチ 零点を与えた場合vs.背景ターゲットを与えた場合

極低発現域



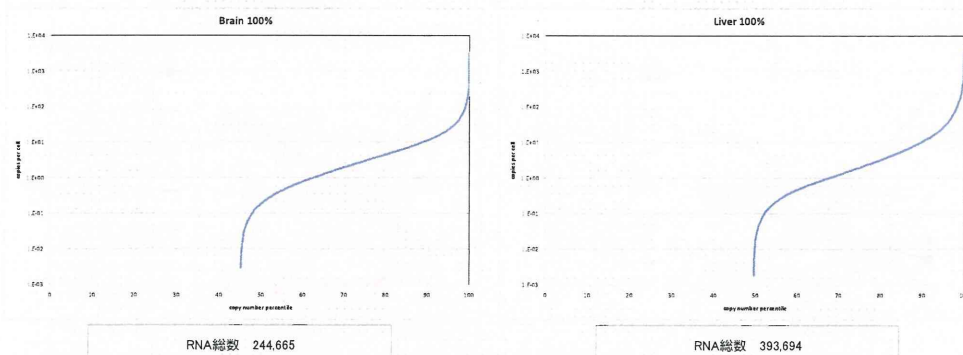
背景ターゲットを与えたほうが、全体的に低い。THEGは精巣特異的な遺伝子なので、脳・肝臓で発現しているとは考えにくく、0であるほうが良好な推定結果と考えられる

Copyright(C)2010-2011 NTT DATA Corporation

7.背景ターゲットによるアプローチ 係数学習後の補正結果



背景ターゲットを設定し、係数学習を進めた係数を用いて補正を行った。



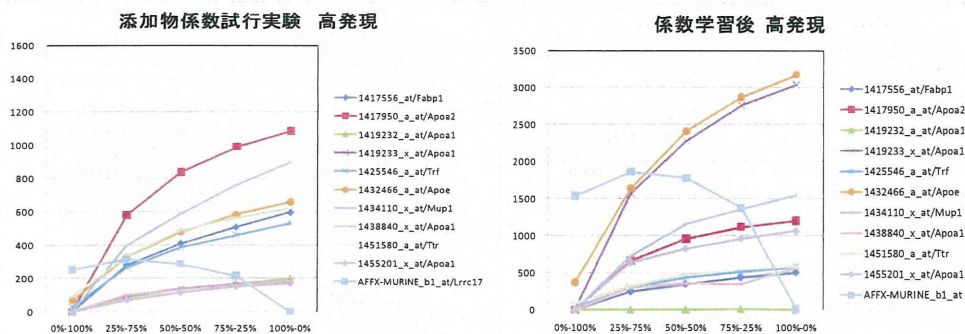
全体の半分近くの遺伝子が0となっている。RNAコピー総数は、脳において約24万、肝臓において約29万であり、肝臓の活動を支えるために多くのタンパク質が必要である状態が窺えて、適切な推計結果と思われる。

Copyright(C)2010-2011 NTT DATA Corporation

7.背景ターゲットによるアプローチ 背景ターゲットを与えた場合 係数学習前vs.係数学習後



高発現域



添加物係数を与えたほうが、全体的に低い。Apoa1関連プロンプセットがほぼ同じ値となった。

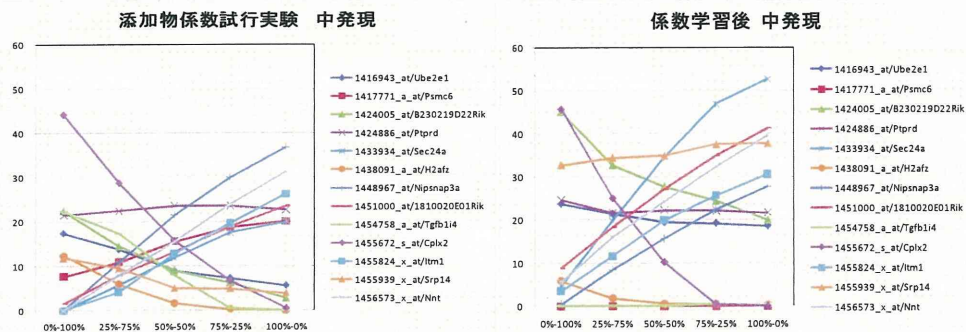
Copyright(C)2010-2011 NTT DATA Corporation

7.背景ターゲットによるアプローチ

背景ターゲットを与えた場合 係数学習前vs.係数学習後



中発現域



添加物係数を与えたほうが、全体的に低い。臓器ごとの発現遺伝子が明確になったものと思われる

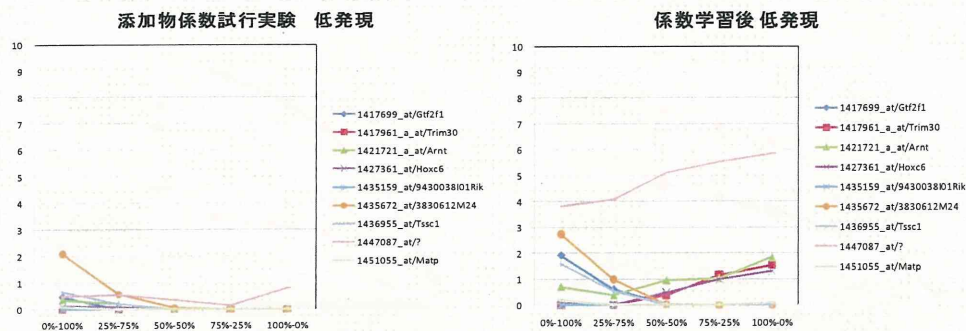
Copyright(C)2010-2011 NTT DATA Corporation

7.背景ターゲットによるアプローチ

背景ターゲットを与えた場合 係数学習前vs.係数学習後



低発現域



添加物係数を与えたほうが、全体的に低い。

Copyright(C)2010-2011 NTT DATA Corporation

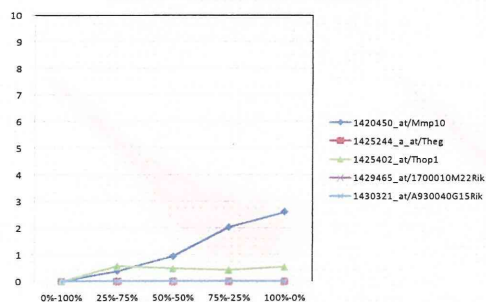
7.背景ターゲットによるアプローチ

背景ターゲットを与えた場合 係数学習前vs.係数学習後

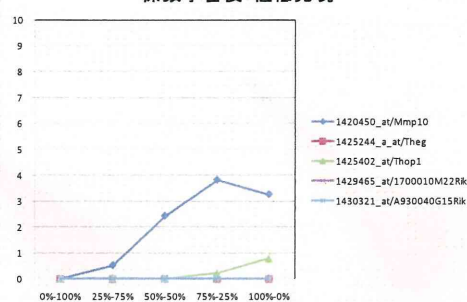


極低発現域

添加物係数試行実験 極低発現



係数学習後 極低発現



添加物係数を与えたほうが、全体的に低い。たとえば、THEGは精巣特異的な遺伝子なので、脳・肝臓で発現しているとは考えにくい

Copyright(C)2010-2011 NTT DATA Corporation

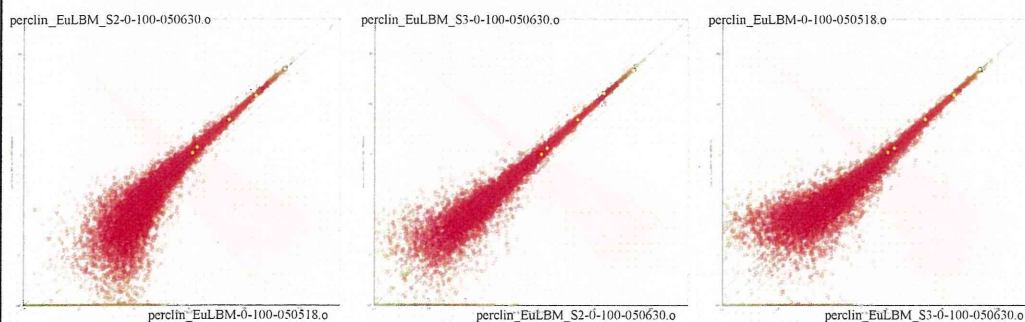
7.背景ターゲットによるアプローチ

散布図による確認



Liver-Brain-Mixtureの三重化実験データを用いて、低発現域における偏差を確認した。

Liver0%-Brain100% 三重化データをサイクリックに散布図を作成した



EuLBM-0-100-050518が他の2チップよりも、少し大きな値となっているが、1コピー以下である

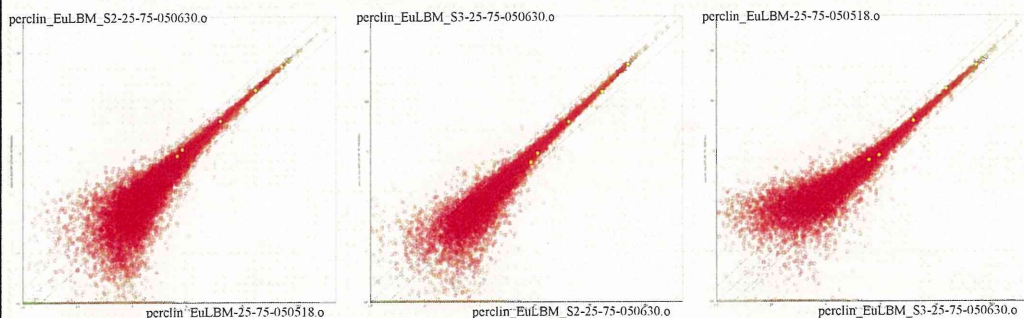
Copyright(C)2010-2011 NTT DATA Corporation

7.背景ターゲットによるアプローチ 散布図による確認



Liver-Brain-Mixtureの三重化実験データを用いて、低発現域における偏差を確認した。

Liver25%-Brain75% 三重化データをサイクリックに散布図を作成した



EuLBM-25-75-050518が他の2チップよりも、少し大きな値となっているが、1コピー以下である

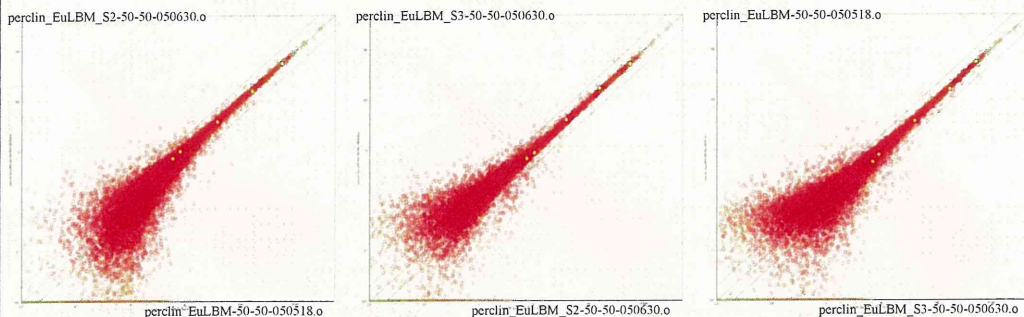
Copyright(C)2010-2011 NTT DATA Corporation

7.背景ターゲットによるアプローチ 散布図による確認



Liver-Brain-Mixtureの三重化実験データを用いて、低発現域における偏差を確認した。

Liver50%-Brain50% 三重化データをサイクリックに散布図を作成した



EuLBM-50-50-050518が最も高く、EuLBM_S3-50-50-050630が最も低く出ている。しかし、その差も1コピー以下である

Copyright(C)2010-2011 NTT DATA Corporation