様式 C-19

科学研究費補助金研究成果報告書

平成２２年　６月１０日現在

研究種目：基盤研究（B）
研究期間：2007～2009
課題番号：19310128
研究課題名（和文）遺伝子発現の周辺確率分布モデル構築

研究課題名（英文）Probabilistic Models of the Marginal Distribution of Gene Expression

研究代表者
Horton　Paul（HORTON PAUL）
独立行政法人産業技術総合研究所・生命情報工学研究センター・研究チーム長
研究者番号：００３７１０７１

研究成果の概要（和文）：
1) マイクロアレイ･データから各遺伝子の発現量周辺分布の確率混合モデルを自動的に構築する手法を確立させた。2) 遺伝子の周辺分布解析を行い、マイクロアレイ・データに伴うノイズが解析の妨げとなっている仮説を立てた。3)mRNA の配列決定を次世代シーケンサーで行った実験データを精密な遺伝子発現解析に使えるための誤読修正法を確立させた。

研究成果の概要（英文）：
1) Established a method for automatically constructing probabilistic mixture models of gene expression from microarray data. 2) Performed analysis on the marginal distribution of genes and hypothesized that the noise in microarray experiments hinders proper analysis. 3) Established a method for correcting the influence of misreads to allow next-generation sequencer mRNA sequence data to be used for precise analysis of gene expression.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	4,700,000	1,410,000	6,110,000
２００８年度	5,000,000	1,500,000	6,500,000
２００９年度	5,200,000	1,560,000	6,760,000
年度			
総　計	14,900,000	4,470,000	19,370,000

研究分野：複合新領域
科研費の分科・細目：ゲノム科学・ゲノム情報科学(2303)
キーワード：遺伝子発現・次世代シーケンサー・TSS-Seq 法・転写頻度・Quality Scores

１．研究開始当初の背景

(1) NCBI の GEO サイトなどから、マイクロアレイで測定された遺伝子発現の公開データが大量に蓄積されていた。

(2) 遺伝子発現制御のモデルとして、各遺伝子の発現状態を{0,1}の二値で近似するブーリアンネットワークや、遺伝子発現を実数値として扱う、ベイ

ジアンネットワークなどは盛んに研究されていた。しかし、遺伝子発現を離散的な値として近似することがどの程度良い近似であるか、ベイジアンネットワークなどの確率モデルに必要となる遺伝子発現分布の適切な事前分布などは知られていなかった。

２．研究の目的
当初の研究目的は、
(1) 多数遺伝子の周辺分布を分析し、遺伝子発現を二値的又は離散値的に扱っても良いか、それとも実数値として扱う必要があるかないかを明らかにすること。
(2) 遺伝子を周辺分布の特徴により分類し、遺伝子の分子機能などと周辺分布との関係を明らかにすること。でした。しかし、計画の途中結果からマイクロアレイデータの質が精密な解析の妨げとなっていると考えるようになった為、
(3) 次世代シーケンサー・データから精密な遺伝子発現情報を得る情報処理技術の確立
を目的に加えた。

３．研究の方法
(1) 周辺分布確率モデルの構築
遺伝子発現の周辺分布として、混合モデルを発現データから学習させた。学習はＥＭアルゴリズムにて行い、混合モデルの成分として、a)正規分布、b)対数正規分布、と c)Γ 分布を試すことにした。適切な成分数の判定は赤池情報基準とベイズ情報基準を試すことにした。遺伝子発現データとして、公開データが豊富なマイクロアレイデータ(NCBI GEO のサイトからダウンロード) を採用した。

(2) 遺伝子発現解析における、次世代シーケンサー誤読による誤差の軽減
研究機関前半の結果を踏まえ、測定精度に問題があるマイクロアレイに代わり、次世代シーケンサーを用いた遺伝子発現データを扱う必要性を認めた。しかし、次世代シーケンサーにも誤読という現象があり、この影響を除く必要がある。そこで、我々は誤読の影響を緩和する為、シーケンサー誤読が、シーケンサー出力全体(各配列とその配列が観察された度数)にもたらす偏りを修正する統計的手法を開発することにした。

４．研究成果
(1) 周辺分布確率モデルの構築
幅広い組織で同一グループが行ったデータセット(ヒト GDS596)と(マウス GDS592) のデータセットを解析した結果、以下の結論に至った。

①成分数を決める判定基準として、ベイズ情報基準は赤池情報基準より正確であった。

②混合モデルの成分として、対数正規分布が最も適切であった。

③２成分や３成分を持つと判定される遺伝子は見つかるものの、ほとんどの遺伝子の周辺分布は「ひとつの山」に近い形を取ることが分かった。
このデータからの結論は、遺伝子発現は連続量として扱うべきものであるが、事前分布として１成分の単純な対数正規分布でも差支えないだろう。

④しかし、「ひとつの山」のような分布は情報を担っているシグナルに見えない。これは恐らく、ノイズを多く含まれるマイクロアレイデータを用いた解析の限界を示している。

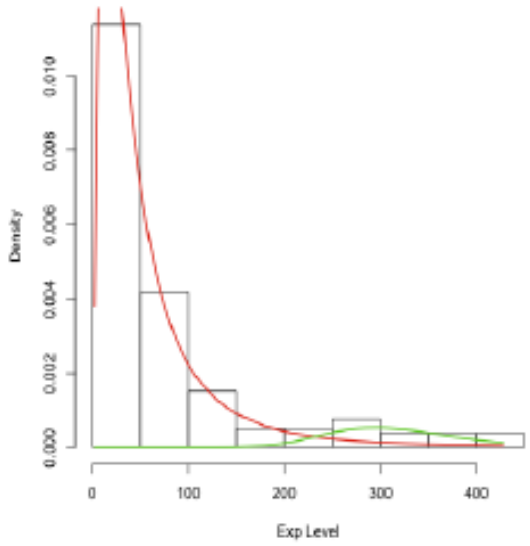
この結果は国際学会 BSBT2008 で査読付き口頭発表を行った。

(2) 遺伝子発現解析における、次世代シーケンサー誤読による誤差の軽減
まず文献調査を行い、シーケンサー誤読の配列度数への偏りを修正する手法を見つけた。Beissbarth et al. 2004 によるＥＭアルゴリズムは提案されていたが、実用的な実装はなかった為、次世代シーケンサーの大量データに耐えられる実装を開発し、RECOUNT と名付けた。

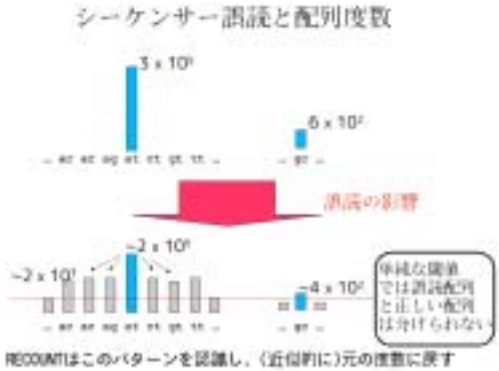
マウスの mRNA 配列をシーケンサーで決定し、遺伝子発現を測定する実験データで RECOUNT の誤読修正性能を検証した結果、信頼できる配列度数を10%増やすことができた。ここで「信頼できる」配列とは、ゲノムにマッピングできる配列である。

近い将来、マイクロアレイ同様、シーケンサー手法により測定された、幅広い組織の遺伝子発現データは入手できるようになる。そのデータに RECOUNT の修正を行えば、より精密な周辺分布解析ができると期待する。なお、RECOUNT の研究成果は査読付き国際学会 GIW2009 で口頭発表を行い、Genome Informatics 誌上にも掲載された。

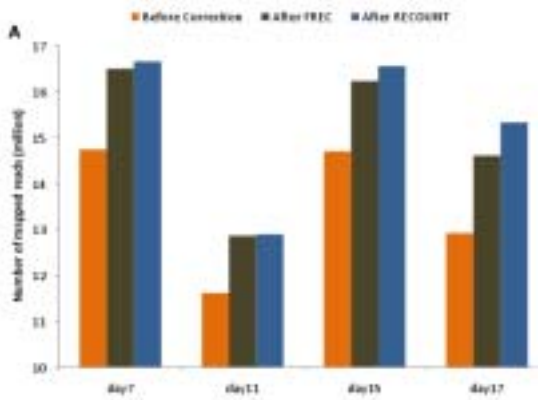
図表



マウス CSN3 遺伝子の発現周辺分布と、本研究で学習した、２成分の対数正規分布を持つ確率モデル。



シーケンサーを用いた遺伝子発現測定における、誤読配列の影響。本研究の成果により、誤読が遺伝子発現解析にもたらす偏りは軽減できる。



RECOUNT の修正を行うと、ゲノムにマッピングできる配列タグは 10%以上増えた。データはマウス胚の転写測定、シーケンサーは Illumina でリード配列長は 36 塩基

５．主な発表論文等

〔雑誌論文〕（計２件）

① Edward Wijaya, Martin Frith, Yutaka Suzuki, Paul Horton, “RECOUNT, Next “Generation Sequencing Error Correction Tool”, *Genome Informatics* **23**(1):189-201, 2009. 査読有り.

② Edward Wijaya and Paul Horton, “Characterizing Genes by Marginal Expression Distribution”, *Communications in Computer and Information Science*, **28**:164-175, 2009. 査読有り

〔学会発表〕（計２件）

①Edward Wijaya, Martin C. Frith and Paul Horton, “RECOUNT, Next Generation Sequencing Error Correction Tool”, 20th international Conference on Genome Informatics (GIW09), 2009 年 12 月 15 日, 横浜

② Edward Wijaya and Paul Horton, “Characterizing Genes by Marginal Expression Distribution”, Bioscience and Biotechnology (BSBT2008), 2008 年 12 月 13 日, 海南、中国

〔その他〕
ホームページ等
http://seq.cbrc.jp/recount/

６．研究組織

(1)研究代表者
Horton　Paul（HORTON PAUL）
独立行政法人産業技術総合研究所・生命情報工学研究センター・研究チーム長
研究者番号：00371071

(2)研究分担者
堀本　勝久（HORIMOTO　KATSUHISA）
独立行政法人産業技術総合研究所・生命情報工学研究センター・研究チーム長
研究者番号：40238803

(3)研究分担者
油谷　幸代（ABURATANI　SACHIYO）
独立行政法人産業技術総合研究所・生命情報工学研究センター・研究員
研究者番号：10361027