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　４日現在

研究種目：特定領域研究
研究期間：２００５～２００９
課題番号：１７０１８０２１
研究課題名（和文）
大規模ゲノム情報の比較技術と知識発見
研究課題名（英文）
Comparative analysis of large scale genome data and knowledge discovery
研究代表者
矢田　哲士　（YADA TETSUSHI）
京都大学・大学院情報学研究科・准教授
研究者番号：１０３２２７２８

研究成果の概要（和文）：
DNA や RNA やアミノ酸の配列を比較することで、分子生物学には数多くの発見がもたらされてきた。一方で、これらの配列を決定する技術は、加速度的に進展し、大量の配列データが蓄積されるようになった。その量は、これまでの配列比較技術が適用できる上限を超えている。ここでは、この大量に蓄積された長大な配列データを精度良く比較する技術を確立するとともに、その技術を活用した生物学的な新しい知識の発見を試みた。その成果のひとつとして、マイクロ RNA の機能が転写のレベルで制御されていることを見いだした。

研究成果の概要（英文）：
We have accumulated a large amount of knowledge concerning with molecular biology by comparing biological sequences, such as DNA, RNA and amino acid sequences. On the other hand, rapid progress of sequencing technology have brought accumulation of a large amount of sequence data. Then, its quantity is now exceeding the upper limit of application of existing sequence comparison methods. Here, we have developed methods which are capable of comparing such a large amount of sequence data with high accuracy, and have tackled the challenge to biological knowledge discovery by applying the methods. As a result, we have revealed molecular mechanism of transcriptional regulation of microRNA function.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００５年度	15,000,000	0	15,000,000
２００６年度	13,800,000	0	13,800,000
２００７年度	13,800,000	0	13,800,000
２００８年度	13,500,000	0	13,500,000
２００９年度	13,500,000	0	13,500,000
総　計	69,600,000	0	69,600,000

研究分野：生物情報科学，計算機生物学
科研費の分科・細目：情報学・生体生命情報学
キーワード：バイオインフォマティックス，ゲノム情報処理

１．研究開始当初の背景

DNAやRNAやアミノ酸の配列を比較することで、分子生物学には数多くの発見がもたらされてきた。一方で、これらの配列を決定

する技術は、加速度的に進展し、大量の配列データが蓄積されるようになった。その量は、これまでの配列比較技術が適用できる上限を超えている。ここでは、この大量に蓄積された長大な配列データを精度良く比較する技術を確立するとともに、その技術を活用した生物学的な新しい知識の発見を試みた。

２．研究の目的

本研究課題では、ポストシークエンス時代に切望されるさまざまな配列比較技術を確立するとともに、その適用による配列情報の解読に取り組む。具体的な研究項目として、（１）miRNA のターゲット推定、（２）ゲノム比較によるアノテーション、（３）ゲノム配列間距離の計測、（４）プロモーター配列の大域的なモデル化に挑んだ。

３．研究の方法

（１）miRNA のターゲット推定では、miRNA とターゲット遺伝子の間では転写情報の一部が共有されているとの仮説を立て、この仮説に立脚したターゲット遺伝子推定法の確立を試みる。miRNA が適切な遺伝子をターゲットするためには、両者が同じ時期に存在しなければならず、その制御の一部は、両者の転写によってなされている可能性がある。すなわち、miRNA とそのターゲット遺伝子の転写は、共通の転写因子によって制御されている可能性がある。そこで、実験的に検証されたヒト miRNA のターゲットデータを網羅的に収集し、miRNA(あるいはそのホスト遺伝子)のプロモーターとターゲット遺伝子のプロモーターに共通のシス因子が存在するかどうかを調べ、その統計的な有意性を評価する。さらに、共通のシス因子に着目して miRNA のターゲット遺伝子を推定することができるかどうかを考察する。

（２）ゲノム比較によるアノテーションでは、ヒトをはじめとする哺乳類ゲノム間で比較解析を行い、哺乳類のゲノムに共通に存在する転写制御領域の解析を行う。また、得られたプロモーター配列のアライメントに基づいて、高精度に転写開始点を予測可能なコアプロモーター予測モデルの構築を目指す。遺伝子の発現制御は多くの因子が絡む複雑な系である上、転写因子と相互作用するゲノム上の制御領域（シスエレメント）は極端に短いことが多い。このため、単一種のゲノム解析や二種程度の比較ゲノム解析では制御領域をうまく捉えることは難しい。本研究では、近縁種間や系統間の複数のゲノムで比較解析を行うことで、プロモーター領域に潜む情報を効果的に抽出する手法を考案する。また、解析においては DNA マイクロアレイを用いた発現情報など配列以外の実験データを取り入れることで、より多角的な視点からプロモーターの情報構造の解明を試みる。

（３）ゲノム配列間距離の計測では、数百万塩基以上のゲノム配列に対して、近年提唱された情報理論（情報距離）に基づいたゲノム配列のマクロ的特徴の抽出、及びゲノム配列間距離において計算コストを抑えて計測する新たな算出法の確立を目指す。

その試みとして、全長数百万塩基以上で構成されているゲノム配列に対し、コルモゴロフ複雑性理論から提唱された正規圧縮距離（ＮＣＤ）の手法を応用し、ゲノム配列間距離の概算及びシンテニー領域の抽出を試みる。mtDNAなど配列長が比較的短い配列の場合、相対エントロピーやK-merの出現頻度比較に比べ、高精度に配列間距離の測定、複数の配列に対する分類分けが行えることが知られているが、染色体全体のような配列長が長い場合、特に遠縁種の比較においては配列間距離の精度の妥当性、及び計算速度の遅さについての問題が残っている。

近年ではＮＣＤによる比較に対して、他分野では特定の文字列データ、画像データに特化して局所的なノイズを排除し、大域的な特徴、情報の抽出及び比較するため非可逆圧縮アルゴリズムを用いるなどの改良手法が試みられている。

我々はこの手法をゲノム配列に応用し、染色体レベルの配列長、及び遠縁種におけるゲノム配列間距離の精度向上を目指す。また、これまではＮＣＤによる算出においては、比較的データサイズの小さい対象のみであるために計算速度は考慮されていなかったが、巨大なデータ比較する場合を配慮し、計算速度の向上を視野に入れた開発を行う。計算速度の遅さの一因として圧縮アルゴリズムの遅さがその原因となっているが、計算コードの最適化、並列計算を採用することにより、計算速度の向上を目指す。

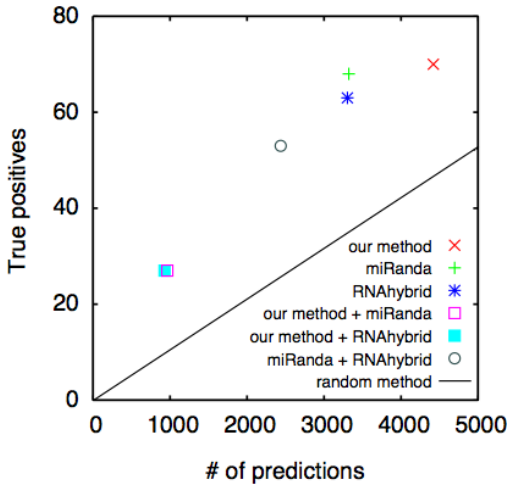
（４）プロモーター配列のモデル化研究では、シス因子が同定されたプロモーター配列のセットから、シス因子が形成する規則性をモデル化する技法を考察する。シス因子が形成する規則性とは、シス因子の並び、組み合わせ、シス因子間の距離などで特徴付けられるものである。ここでは、プロモーター配列の多様性に対応するために、これらの規則性に基づいてプロモーター配列を幾つかのグループに分類し、各グループに関する隠れマルコフモデル(Hidden Markov Model: HMM)を構築する。各分類グループのHMMは、従来から用いられてきたプロファイル型の構造を持つものではなく、汎用性の高い一般的な構造を持つものを採用することとし、その設計アルゴリズムを併わせて考察する。

４．研究成果

（１）miRNA のターゲット推定では、以下

に示す解析により、miRNA とターゲット遺伝子の間では転写情報の一部が有意に共有されていることが示された。ここでは、実験的に検証されたヒト miRNA のターゲットデータ 155 例について、miRNA に関するプロモーターとターゲット遺伝子のプロモーターに共通のシス因子が存在するかどうかを調べた。まず、各々のプロモーターについて、ヒト、マウス、ラット、イヌで保存されている 6 塩基以上のオリゴマーを抽出し、次に、相互作用する miRNA とターゲット遺伝子のプロモーターの間で抽出されたオリゴマーを比較し、両者に共通するオリゴマーを同定した。すると、38%のデータについて、統計的に有意なオリゴマーの共通性が検出された。

次に、シス因子の共通性に立脚した miRNA のターゲット遺伝子の推定法を開発し、その有効性を評価した。この推定法では、まず、与えられたヒト miRNA について、そのプロモーターを同定し、その多重アラインメントから候補シス因子を抽出する。ここでは、ヒト、マウス、ラット、イヌで完全に保存されている 6 塩基以上のオリゴマーを候補シス因子とした。次に、あるヒト遺伝子について、miRNA の場合と同じ手順で候補シス因子を抽出し、miRNA の候補シス因子との間の共通性を検出する。そして、5%の有意水準を満足するシス因子の共通性が観察されると、その遺伝子は与えられた miRNA のターゲットであると判定する。この推定法の精度は、ランダム法の精度を統計的に有意に上回り、miRNA-mRNA 結合部位の種間保存に頼らない従来法の精度とほぼ同等であった。このことは、ヒト miRNA のターゲッティングが転写のレベルで調節される場合があることを示している。



図： miRNA ターゲット推定法の予測精度

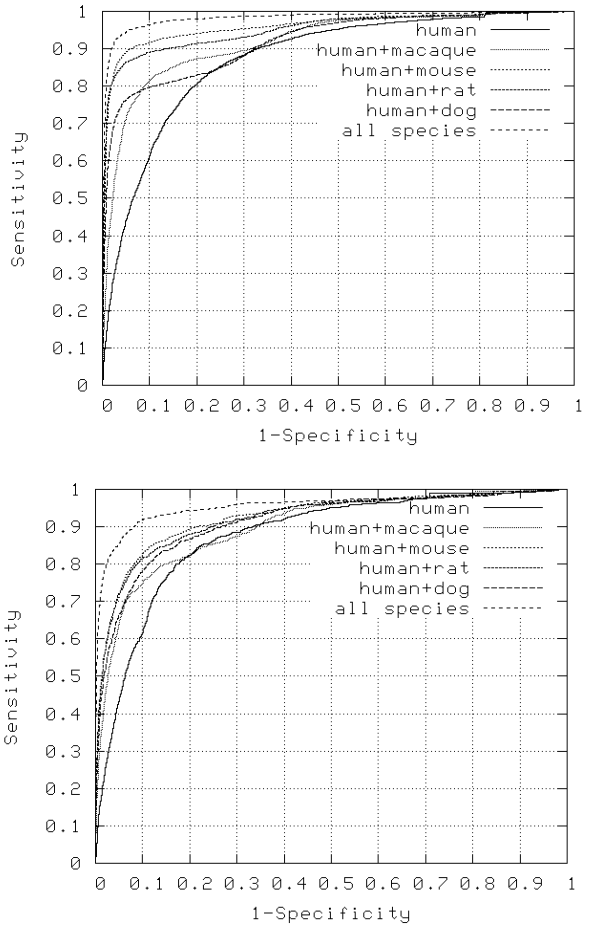
（2）ゲノム比較によるアノテーションでは、ヒト遺伝子の転写開始点（TSS）周辺配列を基準に、他の哺乳類との比較解析を行った。各遺伝子の転写開始点のデータには、ヒト完全長 cDNA 配列を用いて最も転写活性が強いと思われる（より多くの配列でサポートされる）ものを一遺伝子につき一箇所決定し用いた。また、哺乳類遺伝子の多くはプロモーター領域に CpG アイランドを持っており（CpG プロモーター）遺伝子発現に強く影響していると考えられているが、一方で CpG アイランドがないもの（Non-CpG プロモーター）も存在する。これらのプロモーターでは発現制御の様式や配列上の特徴も異なると考えられたため、本解析ではこれらのプロモーターを分けて解析を行った。結果として、CpG プロモーター6,410 個、Non-CpG プロモーター1,292 個の non-redundant な配列が得られた。まずはこれらの配列を用いて、プロモーター領域における CpG アイランドの有無と遺伝子発現の組織特異性との関係を調べた。DNA マイクロアレイ（GNF GeneAtlas2）のデータをもとに発現の組織特異性をシャノンの情報エントロピーとして数値化したところ、エントロピーの低い（組織特異性の高い）遺伝子のほとんどが Non-CpG プロモーターを持つことが示された。

次にヒトプロモーター配列を、マウス、ラット、イヌの相同遺伝子の上流配列とペアワイズで比較したところ、90%以上のプロモーター配列が種間で高度に保存されていることが分かった。このことから、少なくとも転写活性の強いプロモーターは種を超えて保存されており、比較ゲノムによるコアプロモーター同定が十分可能であることが示された。Non-CpG プロモーターの TSS 周辺配列は CpG プロモーターに比べてより強く保存されており、特に-28 前後の領域（TSS を+1 とする）が Non-CpG プロモーターでは重要であることも分かった。CpG プロモーターでは、領域中の CpG アイランドのほぼすべてが種間で保存されていた。他の領域にある CpG アイランドは必ずしも種間で保存されていないことから、プロモーター領域中の CpG アイランドが転写制御に強く関わっている可能性が改めて確認できたといえる。

ペアワイズの比較解析では、TSS や TATA ボックスの出現位置周辺が強く保存されており、そこから上流・下流に離れるほど保存度が低くなることが示された。そこで、コアプロモーターの予測に重要な領域をさらに絞り込むため、ヒトの配列のみを用いて様々な解析範囲で予測モデルを構築し、その精度を調べた。モデルには（TSS を基準とした）位置特異的な 2 次のマルコフモデルを用い、評価は上記データを用いてクロスバリデーション（5-fold）を行った。結果として、[-50,+50]の範囲でモデルを構築したとき AUC 値(ROC カーブの下側面積で最大値は 1)が最大(CpG:0.87, Non-CpG:0.88)となった。この領域には、上流部の TATA ボックスやイニシエーターのほかにも、下流に DPE（Downstream

Promoter Element)などの重要なコアプロモーターエレメントが存在することが知られており、リーズナブルな結果といえる。さらに、アライメント情報をもとにマウス相同領域でもヒトと同じマルコフモデルを構築し、ヒトモデルとマウスモデルのスコアの和をとることで、さらに高い精度（CpG:0.96, Non-CpG:0.93）を実現できた。本予測モデルは確率モデルでありスコアは対数オッズ比であるため、単純なスコアの足し合わせで予測精度の改善が可能であり、より多くの生物種の情報を簡単に取り込むことが出来ることが示された。

最後に、ヒトとマウスに加えて、アカゲザル、ラット、イヌの配列を用いてマルチプルアライメントを行い、それぞれの生物種のマルコフモデルを構築してコアプロモーターの予測を行った。2 種（ヒト＋他）のモデル統合では、CpG/Non-CpG プロモーターともマウスとの組み合わせのとき（上述）に AUC 値が最も大きく、逆にアカゲザルとの組み合わせのとき最小（CpG:0.92, Non-CpG:0.90）であった。より近縁の種であるほど共有する遺伝子も多く比較解析を用いることによる感度の低下も軽減できる可能性は高いが、オルソログが存在するのであれば出来るだけ離れた生物種と比較した方が情報量も多く、予測精度の向上にも寄与することが確認された。5 種すべてのモデルを統合することで、最終的に CpG プロモーターで 0.98、Non-CpG プロモーターで 0.96 というきわめて高い精度での予測が行えた。本予測モデルを用いることで、プロモーターの種類（CpG アイランドの有無）を問わず、哺乳類に共通する転写活性の強いコアプロモーター領域を高精度に予測可能であり、今後の転写制御の研究に役立つものと期待される。



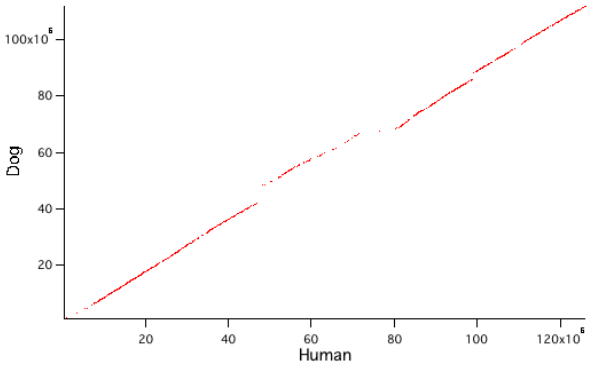
図：CpG(上)、non-CpG(下)プロモーターの判別精度

（３）ゲノム配列間距離の計測では、全長数百万塩基以上で構成されているゲノム配列に対し、計算コストの高いアラインメントを行うことなく、配列に内在する巨視的な特徴に基づいて配列を分類する新しい手法の確立を目指した。

まず、ＮＣＤを巨大塩基配列にも適応させるための複数の改良案を検証するため、Evolsimulator、INDELibleなど人工データを生成する複数のソフトウェアを用いて百万塩基前後の配列を生成、その改良案の対する有効性、適用限界の検証を行った。そして、次の段階として改善が見られた三種類の改良案に対し、実在するゲノム配列における配列間距離の計測、検証を行った。

(a) まず一般のＮＣＤが適応可能と考えられる一定の固定長に配列を「ブロック」に分割、ブロック単位での配列のペアに対して、全てのブロックの組み合わせのＮＣＤの算出化結果から配列全体の配列間距離を近似する手法である。

これは全組み合わせのペアのＮＣＤの結果の中から、全てのブロックを使用し且つ重複が無い条件の下、ブロック単位でのＮＣＤの合計が最小になる組み合わせとなるブロックペアを求め、そのペアの組み合わせを基に、配列全体におけるＮＣＤを再計算することによって配列全体の情報距離を算出する手法である。この方法は、以下の改良案に比べ、計算時間は掛かるが、高い安定性で古細菌、

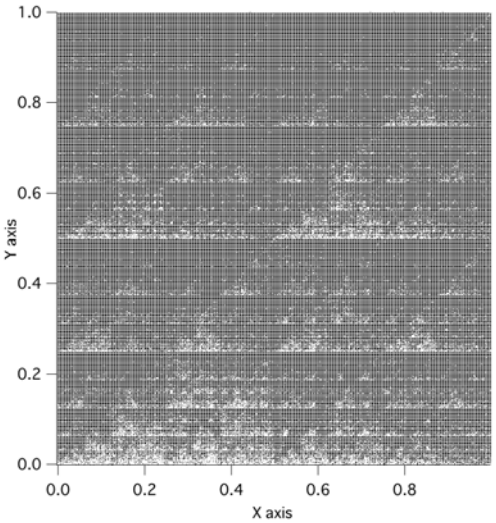


図：ヒトとイヌのＸ染色体におけるシンテニー領域

真正細菌の完全ゲノム、哺乳類の染色体に対するアライメントなど他の手法で得られるシンテニー領域、配列間距離について同等の結果を算出することが確かめられた。

(b) ゲノム配列に対し、１９９０年にH.Jeffreyが提案したChaos Game Representation（ＣＧＲ）によってゲノム配列を二次元カオスマップに写像する方法を応用する手法である。この手法でゲノム配列が持つカオス的な周期的特徴を抽出し、その後にＮＣＤで配列間距離を算出する。これまでゲノム情報をカオスマップに写像する手法において、ゲノム配列をその特徴別に分類する際、あいまいな視覚的分別もしくは各研究者が独自に定義した距離法によって分類されていたが、我々はこの手法にＮＣＤを組み合わせることにより、アルゴリズム情報理論に基づいた分類法の構築を試みた。この手法を遠縁種が多く含んだ古細菌、真正細菌ゲノム配列群を用いた場合、高速に近縁種同士のグループに分別することが確認された。

(b) べき相関が見られるマクロな現象（フラクタル・パターンなど）から特徴を抽出する際に用いられる実空間繰り込み群の手法をゲノム配列に適用し、配列を短い符号列に変換することによって配列情報の縮約を行った。そして縮約後の情報を用いてＮＣＤによる配列間距離の算出を行った。古細菌、真正細菌のゲノム配列を用いた計算機実験ではＣＧＲによる手法に比べて近縁種における配列間距離の精度が高いことが示された。



図：M.tuberculosis のカオスマップ

（４）プロモーター配列のモデル化研究では、シス因子が同定されたプロモーター配列のセットから、シス因子の形成する規則性(シス因子の並び、組み合わせ、シス因子間の距離など)をモデル化する技法を考察してきた。発現プロファイル法などで機能的に分類されたプロモーター配列のセットであっても、そこには異なる制御情報を持つグループが幾つも混在していることが多い。そこでここでは、シス因子の規則性に基づいてプロモーター配列を幾つかのグループに分類し、各グループに関するHMMを構築するアルゴリズムを考案した。各グループのHMMは、従来から用いられてきたプロファイル型の構造を持つものではなく、汎用性の高い一般的な構造を持つものを採用している。アルゴリズムの概略は以下の通りである。(a)与えられたプロモーター配列のセットをシス因子に基づいてアラインメントし、そのアラインメントをleft-to-right型のHMMに変換する。このHMMは、状態でシス因子を出力し、状態遷移でシス因子間の配列を出力する。(b)シス因子間の配列長の分布をFrechet分布を使ってモデル化する。このモデル化に先立ち、Dixon検定を用いて配列長の分布データから外れ値を取り除く。また、モデル化は、Sturges則で作成された配列長分布のヒストグラムに最小二乗法を適用することで行なわれる。(c)各配列に関するHMMの状態遷移列のデータに数量化III類を適用し、状態と配列の各々をWard法によって分類する。(d)状態の分類について、次の条件を満たす状態をマージする。同じグループに分類された状態で、同じシス因子を出力するが、それらのシス因子が別々の配列に由来するもの。この操作により、left-to-right型のHMMは一般的な構造を持つことになる。マージする状態がなくなるまで(c)と(d)の操作を繰り返す。(e)配列の分類について、同じグループに分類された配列の状態遷移列でHMMを分解し、各グループのHMMとする。このアルゴリズムを哺乳動物のヒストン遺伝子のコアプロモーター配列のセット([-250, -1]領域の127配列)に適用したところ、アルゴリズムは、配列セットに含まれる各ファミリーに当たるプロモーターモデルを生成することに成功した。分類精度は64.5%だった。また、H1-H4を特徴付ける主要なシス因子はleft-to-right型の構造でモデル化され、状態のマージ処理は副次的なシス因子間で行なわれていた。

15
$\mathrm{AC}_{.50} \rightarrow \mathrm{GC}_{.61} \rightarrow \mathit{H4TF2/HinF\text{-}P}_{.89} \rightarrow \mathrm{TATA}_{.39}$

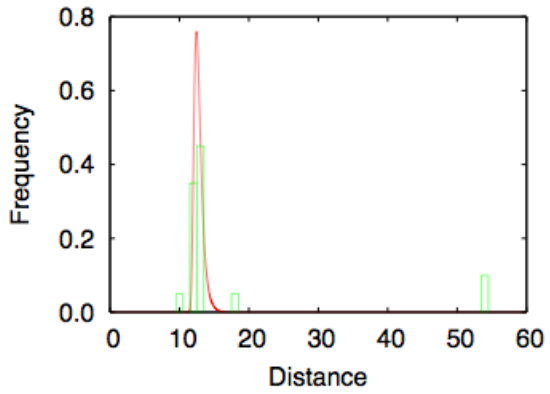
18
$\mathrm{AC}_{.68} \rightarrow \mathrm{GC}_{.55} \rightarrow \mathrm{CCAAT}_{.73} \rightarrow \mathrm{TATA}_{.73}$

17
$\mathit{Oct\text{-}1}_{.76} \rightarrow \mathrm{CCAAT}_{.90} \rightarrow \mathrm{CCAAT}_{1.00} \rightarrow \mathit{E2F}_{.62} \rightarrow \mathrm{TATA}_{.52}$

25
$\mathrm{CCAAT}_{.68} \rightarrow \mathrm{CCAAT}_{.97} \rightarrow \mathrm{TATA}_{.45}$

26
$\mathrm{TATA}_{.34} \rightarrow \mathit{E2F}_{.47} \rightarrow \mathrm{CCAAT}_{.81} \rightarrow \mathrm{CCAAT}_{1.00} \rightarrow \mathit{Oct\text{-}1}_{.59}$

図：生成されたヒストン遺伝子のプロモーターモデル



図：モデリングされたシス因子間距離の例（CCAAT→TATA ）

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計４件）

1. T.L. Cui, H. Nakaoka, K. Akiyama, H. Kamura, K. Hosomichi, J. Bae, H. Cheong, H.Shin, T. Yada, I. Inoue: Positional effects of polymorphisms in probe-target sequences on genoplot images of oligonucleotide microarrays, Genet. Mol. Res., 23, 524-31 (2010).
2. S.J. Park, N. Ichinose, T. Yada: Inferring Probabilistic Conditional Independency from Large-scale Combinatorial Regulation of Transcription Factors, J. Software, (in press).
3. S.J. Park, N. Ichinose, T. Yada: Probabilistic Graphical Modeling for Large-scale Combinatorial Regulation of Transcription Factors, In the Proc.of Workshop on Knowledge, Language, and Learning in Bioinformatics (KLLBI-08),72-86 (2008).
4. N. Ichinose, T. Yada, O. Gotoh, K. Aihara: Reconstruction of transcription-translation dynamics with a model of gene networks, J. Theor. Biol., 255, 378-86 (2008).

〔学会発表〕（計５件）

1. Yada, T.: A novel method for miRNA target prediction: miRNAs and their targets are frequently coregulated by common transcription factors, The 4th Global COR International Symposium 2009 joint with the 19th Hot Spring Harbor Symposium, Fukuoka, Novermber S3-2 (2009).
2. Yada, T., Fujiwara, T., Terai, G., Gotoh, O.: A Novel Point of View on MicroRNA Target Prediction, Proc. of the 2008 Annual Conference of Japanese Society for Bioinformatics, Osaka, P018/T04 2008.
3. S.J. Park, N. Ichinose, T. Yada.: Comprehensive Modeling for the Combinatorial Regulation of Transcription Factors, Proc. of the 2008 Annual Conference of Japanese Society for Bioinformatics, Osaka, P006/T08 2008.
4. N.Ichinose, T.Yada, O.Gotoh: Fast Motif Extraction Tool from a Large Number of DNA Sequences, Proc. of the 2007 Annual Conference of Japanese Society for Bioinformatics, Tokyo, P07 2007.
5. Yada, T.: A motif-based framework for constructing promoter models, The Fifth East Asian Biophysics Symposium & Forty-Fourth Annual Meeting ofthe Biophysical Society of Japan, Okinawa, Novermber 12-16 (2006).

〔図書〕（計４件）

1. 矢田哲士：遺伝子発見，森下真一、阿久津達也編集，生命研究への応用と開発が進むバイオデータベースとソフトウェア最前線，実験医学増刊，26(7)，1050-1055 (2008)
2. 矢田哲士，後藤修：ゲノム比較時代の配列比較技術，藤山秋佐夫監修，比較ゲノム学から読み解く生命システム，細胞工学増刊，21-27 (2007)
3. O. Gotoh, S. Yamada, T. Yada: Multiple Sequence Alignment, Handbook of Computational Molecular Biology, (Aluru, S. ed.), Chapman & Hall/CRC Computer and Information Science Series, 9, 3.1-3.36 (2006).
4. 矢田哲士：ヒトゲノムのアノテーションデータベース，小原・菅野・小笠原・高木・藤山・辻編集，ゲノムから生命システムへ，蛋白質核酸酵素増刊，50(16)，2053-2058 (2005).

〔産業財産権〕
○出願状況（計０件）

○取得状況（計０件）

〔その他〕
なし

６．研究組織
(1)研究代表者

矢田　哲士（YADA TETSUSHI）
京都大学・大学院情報学研究科・准教授
研究者番号：１０３２２７２８