論文内容の要旨

論文題目：

完全長 cDNA 情報を用いた遺伝子アノテーションおよび転写開始点解析

～アピコンプレクサ原虫を例として～

氏名　：　若栗　浩幸

要約

転写開始点(TSS)情報を含む完全長cDNAデータは、プロモーター領域の同定、遺伝子アノテーションをはじめとする様々なトランスクリプトーム解析を行う上で、重要な情報資源である。我々の研究室では5’端の完全なcDNAを抽出可能なオリゴキャッピング法を開発し、完全長cDNAデータの大規模収集を行っている。私は、これまで完全長cDNAデータを有効活用するための情報学的枠組を構築、解析を行ってきた。cDNAのゲノム配列へのマッピング方法、モデル遺伝子への対応付け、プロモーター領域の抽出と転写因子結合領域の推定、近縁種間での遺伝子やプロモーターの対応情報の抽出などといった解析パイプラインを構築し、データベース、DBTSS(http://dbtss.hgc.jp/)としてインターネット上に公開した。

近年、高等動物の他にアピコンプレクサ原虫について解析を進めている。アピコンプレクサ門原虫はマラリアを初めとする、ヒトや家畜動物に感染し世界規模で深刻な被害をもたらしている寄生虫の一群である。本研究では、6つのアピコンプレクサ原虫からオリゴキャッピング法を用いて抽出した61,056の5’端cDNA配列を出発点として用いた。解析の結果、PlasmoDBなど公共データベースに登録されているモデル遺伝子と単離されたcDNAの間には多くの相異が見出され、モデル遺伝子の多くについて修正が必要であることが示唆された。最も多くの差異が認められたトキソプラズマについてはプライマーウォーク法により732の完全長配列を決定し比較した。その結果、41%のモデル遺伝子には、何らかの相異が含まれていた。さらにIllumina GAを使用して他のアピコンプレクサ原虫の完全長cDNA配列決定を行ったところ、現在の遺伝子アノテーションが不十分である多くの例を見出すことができた。データの解析結果は、データベース（Full-Parasites 及び Comparasite; http://fullmal.hgc.jp/）を構築し、一般に公開している。

方法と結果

オリゴキャップcDNAの5’配列のマッピング、クラスタリング及びデータベースの構築

ヒトあるいは多くのモデル生物の様々な組織から構築した完全長cDNAライブラリーに由来するcDNAの5’端配列から、遺伝子構造あるいはその転写制御についての情報抽出を行うための枠組を構築した。これらの情報は、リレーショナルデータベースにテーブルを設計後、データを投入し整備を行った。データベースはFigure1に示すinf，seq，pos，parts，groupの5つのテーブルを中心として構築した。それぞれ遺伝子の名前や転写産物の情報、塩基配列やアミノ酸配列情報、マップ位置情報、エクソン境界情報、マップ情報をクラスターしたデータを保持している。クラスターされたcDNAは同一のgroup_idを割り振ることで同一の遺伝子として対応付けた。これらのデータベーステーブルを用いることでデータを整理し、柔軟なデータの検索や解析が可能となった(Figure 1)。

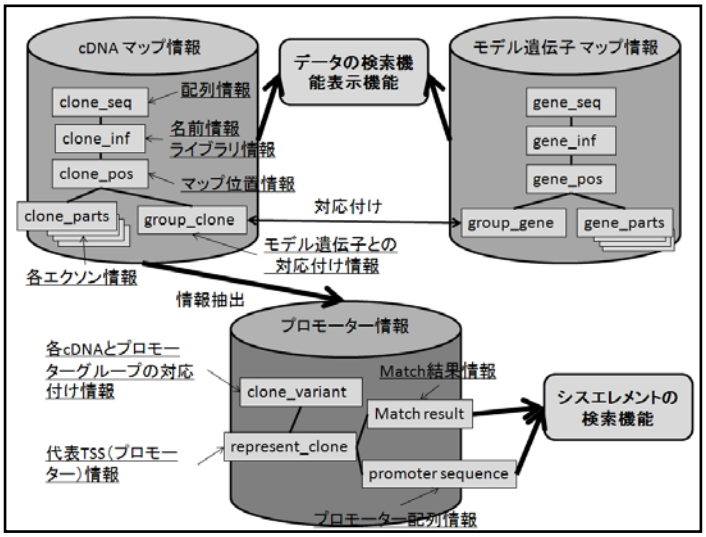

Figure 1: マッピングおよびプロモーター情報のデータ構造

プロモーター領域の抽出と転写因子結合領域の推定

データの構造化に伴い、単一の遺伝子に複数のプロモーターが存在する選択的プロモーターが数多く存在することが明らかとなった。選択的プロモーターの情報を抽出するために、各group_id(遺伝子に対応付けされたcDNA)について500bp以内にあるcDNAを同一のプロモーターと定義し、新規にクラスタリングを行った。クラスター化された全てのプロモーターにIDをアサインしデータベース化した。

各プロモーター領域からは転写因子結合領域の推定と結果のデータベース化も行い、DBTSSから検索できるようにした。

生物種間のプロモーター領域の比較解析

ヒト以外の様々な生物種にも同じ枠組を用いてデータベース構築を行った。マウス、ラット、マカクザル、チンパンジーについては、NCBIのHomoloGeneやUCSCのblastz結果のデータを用いてヒト遺伝子との対応付けを行った。相互に対応する遺伝子については、配列アラインメントを行い、比較解析を可能にした。特に、ヒトとマウスで対応するプロモーター配列については、種間で保存される転写因子結合配列をデータベース化し、シス因子の保存の有無からの検索を可能とした(Figure 2)。

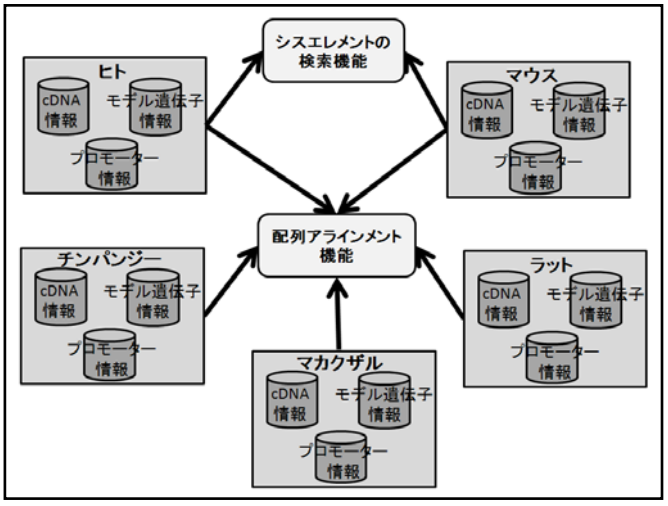

Figure 2: 近縁種での比較解析情報のデータ構造

アピコンプレクサ原虫のオリゴキャップcDNAの5’-ESTのマッピングとクラスタリング

上述のcDNA解析の枠組を、アピコンプレクサ原虫に適用し、完全長cDNAの解析を行った。6つのアピコンプレクサ原虫（*Plasmodium falciparum* (Pf), *Plasmodium vivax* (Pv), *Plasmodium yoelii* (Py), *Plasmodium berghei* (Pb), *Cryptosporidium parvum* (Cp), *Toxoplasma gondii*(Tg)）の5’-EST cDNA配列について公共データベースに登録されているゲノム配列に対してマッピングを行った。マップしたcDNA配列は上記に示した方法でモデル遺伝子に対応付けを行い相互に比較した。比較の結果、平均して16%の遺伝子に相違が認められた（Table 1）。

Table1: オリゴキャップcDNAとモデル遺伝子の差異

生物種	差異のあった遺伝子数 No./Total No. (%)
Pf	133/1,543 (9%)
Pv	258/1,457 (18%)
Py	233/1,340 (17%)
Pb	53/254 (21%)
Cp	32/658 (5%)
Tg	191/780 (24%)
平均	16%

トキソプラズマ原虫の全長cDNAの完全長シーケンシング

上記の結果として、6つの生物種の中でゲノムサイズが最大で、より多くのイントロンを持ち合わせているTgが最も多くの差異を含んでいた。Tgのモデル遺伝子をより精密に評価するために、代表的なcDNAクローンを選出し、プライマーウォーク法で完全長配列の決定を試みた。結果として732cDNAの完全長配列が得られ、そのうち592は1つ以上のモデル遺伝子に対応していた。比較の結果41%ものcDNAに差異が見られ、モデル遺伝子に修正の余地が残っていることが示された。

残りの140のcDNAについては既知のモデル遺伝子とは対応せず、新規の転写産物であることが示唆された。転写を確認するために135個のcDNAをピックアップし、RT-PCRを行った。結果として118についてはポジティブバンドが確認され、実際に転写が行われていることが示された。このように現在のモデル遺伝子にはまだかなりの数の未予測の転写産物が存在すると考えられた。

cDNA-Seqを用いた完全長配列の解読

トキソプラズマ以外の生物種についても、各クラスターから代表的なcDNAを抽出して完全長配列の決定を試みた。配列決定にはプライマーウォーク法に代えて、次世代シーケンサーのIllumina GAを使用した。リード配列を各生物種のゲノムにマップし、マップ位置とリードのカバレージを元にアセンブルを行った。マップされなかったリード配列については2つの配列に分割し、再度マッピングを行うことでエクソン/イントロン境界の判定に利用した。結果として、合計4,319の完全長配列を決定することができた(Table2)。プラーマーウォーク法に比べかなり効率的に全長配列を決定することができたといえる。

Table2: アピコンプレクサ原虫のcDNA-Seqによる全長解読

生物種	read長	read数	ターゲットcDNA数	全長になったcDNA数	成功率(%)※
Pf	36,76	42,495,685	2,898	348	12%
Pv	36,76	77,190,632	3,108	1,538	49%
Py	36	12,551,084	635	311	49%
Pb	36,76	27,399,879	554	329	59%
Cp	36,76	29,122,289	1,200	1,066	89%
Tg	36,40,76	57,175,019	1,877	727	39%

※ 5’-vector, 3’-polyA tailの配列が認められ、全てのイントロンがGT-AGルールをもったcDNAの割合

オルソログ遺伝子を用いた系統樹の生成

オルソログ遺伝子の公共データベースであるOrthoMCL DBから6つのアピコンプレクサ原虫についてのオルソログペアを抽出し、分子進化学的解析を行った。6つの生物種に共通に存在した729遺伝子ペアについて、系統樹を作成したところ、主に3つの樹形に集約された。これらの系統樹はいずれもPy，Pbの共通先祖と、Pf、Pvの分岐の違いであった。これら3つの系統の分岐年代が近い可能性が示唆された。729ペアの遺伝子について、各生物種について直列に結合した配列を生成し、NJ法で系統樹を作成したところ、この3者間は非常に短い枝長となった（Figure 3)。

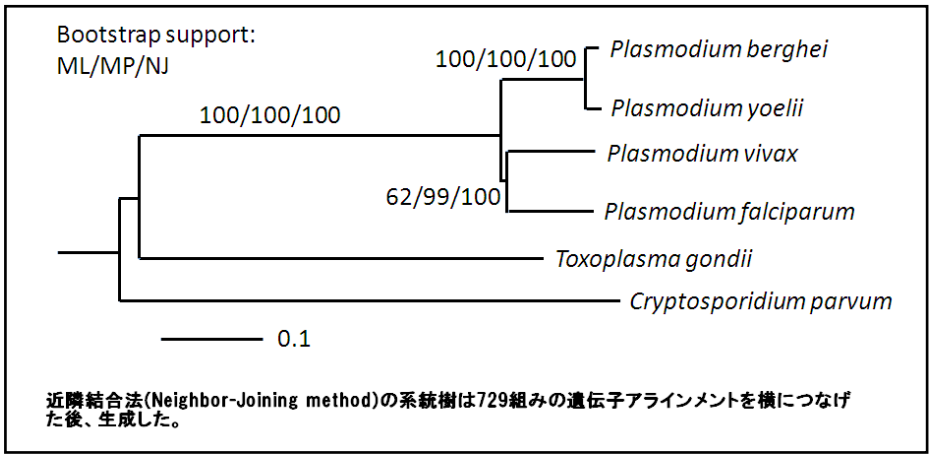

Figure 3: アピコンプレクサ原虫の推定系統樹

Full-Parasites, Comparasiteデータベースの構築

cDNAデータはデータベース化を行い、Full-Parasites及びComparasiteとして一般公開している(http://fullmal.hgc.jp/，http://comparasite.hgc.jp/)。Full-parasitesでは、完全長クローンを検索し、公共の遺伝子との異同の視覚的な解析を可能にし、各モデル遺伝子と全長配列の決定されたcDNAについて、PSORT，Pfam等のアノテーション情報を付加し、それらも閲覧と解析を可能にした。Comparasiteでは、これら6種類の生物種について、アノテーションレベルでの比較がまとめられている。プロモーターに関する配列比較情報も閲覧、検索が可能である。

まとめ

本研究では、大量の完全長cDNA情報を有効活用すべく、データの処理方法や抽出方法について考察した。アピコンプレクサ原虫における応用例では、全長cDNAの5’端配列や完全長配列を用いてモデル遺伝子の修正と分子進化的解析が可能であることを示した。現在、アピコンプレクサ原虫において、次世代シークエンサーの利用により大量に収集された転写開始点の解析についての枠組の構築を検討している。本研究により種間で大きく異なるトランスクリプトーム像の網羅的解析に向けて、組織的かつ横断的な解析が可能になったと考えている。