論文審査の結果の要旨

氏名　呉　紅艶

生物学が取り扱うデータは、構造化／階層化して扱うと都合の良いデータが多い。たとえば、組織→細胞→オルガネラ、ＤＮＡ→エキソン・イントロン構造→タンパク質コード領域、Gene Ontology における遺伝子の機能階層などは典型的な例である。このような構造化／階層化をそのまま自然に記述することができる点で、半構造化データ記述言語 XML は相性がよい。そのため多くの生物学的データベースが XML により記述されている。

一方、XML で生物データを記述する際の問題点として、構造化／階層化の記述方式の自由度が大きいため、同じデータを記述するにしても、設計者により大幅に異なるデータ構造が作られる傾向にあることが挙げられる。そのため、データを問い合わせる際には、各データベースの構造化の仕方に忠実に問合せを作成しなければならない。たとえば XML データベース問合せ言語 XPath や XQuery では、必要なデータへアクセスするためのパスを明示的に記述しなければ所望のデータを抽出することができない。プログラミングに近い作業となるため、一般のユーザー、特に生物学者ユーザーにとっては大きな負担となる。

この問題を解決するためにパスを明示的に記述せず、データの属性だけを宣言的に記述するだけで、システム側がパスを推定して計算するようなパスフリー問合せ言語がいくつか提案されている。パスフリー言語は初心者ユーザーにとっては福音である。しかし、計算時間や計算資源を最適化するデータアクセスパスを推定することは自明でなく、難しい問題である。

そこで本研究では XML 用パスフリー言語の要となる演算として近年注目されている Amoeba Join を研究対象として最適化技法を提案している。最適化技法が普及することを意識して、関係データベースシステム RDBMS 上に実装できるように工夫している。最適化の柱は、①木構造を区間木として RDBMS 上に効率よく実装した点、② データ間の関数従属性（Functional Dependency）を考慮した最適化技法である。これらについては第 2 章で詳細に述べられている。

最適化技法の有効性は第 3 章で詳細に検討されている。処理速度の高速化を実証するために、区間木最適化法を利用した場合の高速化率、関数従属性を考慮した最適化の高速化率は、どちらも模擬データを使って実証されている。またデータ数を増やした場合の Scalability も良好であることが示されている。

第 4 章では現実のデータベースを使ってその有効性を議論している。利用したのは出芽酵母のすべての非必須遺伝子を破壊した株の、遺伝子型と表現型を格納したデータベース SCMD である。ゲノム、遺伝子、表現型画像、Gene Ontology という生物学における主要なデータ要素を SCMD は含んでおり、解析対象として適切な選択である。典型的なデ

ータ抽出が現実的な時間で処理できており、当初の研究目標を達成できている。

まとめると、呉紅艶は、構造化した生物学データベースへのデータ問合せをパスフリー言語により現実的な時間で処理することができるシステムを、幅広く普及しているRDBMS 上で実装できる方式を提案することに成功している。情報生命科学の観点から高い貢献である。なお、本論文は、斉藤太郎、森下真一との共同研究であるが、論文提出者が主体となって分析及び検証を行ったもので、論文提出者の寄与が十分であると判断する。

したがって、博士（科学）の学位を授与できると認める。