大規模ゲノムデータ処理に対する高速高精度アルゴリズムの開発

●柳浦 睦憲 [1)]　◆宇野 毅明 [2)]　◆小野 廣隆 [3)]
1) 名古屋大学大学院情報科学研究科　2) 国立情報学研究所情報学プリンシプル研究系　3) 九州大学大学院システム情報科学研究院

＜研究の目的と進め方＞

ゲノム研究に関わるデータは巨大なものが多い．ゲノム自身が巨大な文字列データであることをはじめ，遺伝子やたんぱく質，生物種など，多くの項目を持つデータがある．また，マイクロアレイ技術の発達により，多くの実験を短時間で行えるようになったことも，実験結果のデータを巨大化させている．

これら巨大なデータベースを解析し，全体的な特徴の観察や，類似する項目の発見･グループ分け (類似検索･クラスタリング)，確からしいルール・特徴ある部分構造の発見 (ルール／データマイニング) を行うことは，ゲノム研究において非常に重要な計算処理である．しかし，データが巨大であるため，従来の素朴な方法では計算に莫大な時間がかかる．1000 万件の項目を持つデータベースから類似する項目のペアを見つける問題は，単純に全ての組を比較すると１年以上の時間を要する．類似検索ツールを用いれば計算時間を相当短縮できるものの，それでもなお大きな時間がかかる．並列計算機を用いればさらに計算時間は短くなるが，それでもかなりの時間が必要であり，かつ金銭的にも労働力の面でも多大なコストが必要となる．

しかし，この種の問題では，出力する解の数 (解数という) は全ての組合せよりはるかに小さいことが多い．類似する項目を例に挙げれば，１つの項目は他の高々数個の項目とのみ類似する場合が多い (これを解の疎性とよぶ)．そもそも，非常に多くの項目が類似するようなデータベースは，類似する項目を全て列挙すること自体に意味がない．全ての項目を総当り比較するのではなく，探索対象を類似する可能性のあるペアだけに絞り込むことが効率良くできれば，極めて短時間で計算を終了することが可能である．実際にいくつかの問題では，解の疎性が成り立ち，この種の高速化が可能である．

このような性質は組合せ的な構造を検索する問題 (パターンマイニング) においても有効である．全ての組合せの中から必要な組合せのみを調べることで，大幅な計算時間の短縮が期待できる．同種の技術は最適化問題においても有効である．必要と判断される解のみを調べることにより，短時間で精度の高い解を得ることができる．ゲノム列のアセンブリングや複数のゲノム列のアラインメントは，大規模なデータベースを入力とする最適化問題であるため，このようなアルゴリズム技術が大きな役割を果たすであろう．

また，ゲノム情報学の問題における顕著な特徴として，データがあいまいさを含むことが多いということが挙げられる．このあいまいさは，実験のエラーや種の進化によって起こり，計算コストの増大や，解精度の減少を引き起こしている．よって，あいまいさやエラーを許容した高速計算手法の開発が急務となっている．

本研究班では，ゲノム情報学に現れる基礎的なデータベース解析問題や最適化問題に対して，精度が高くかつ高速なアルゴリズムを開発することを目的とする．基礎的な問題に焦点を当てた理由は，情報科学的に先鋭的な結果を追求するのではなく，あくまでもゲノム情報学でのアルゴリズム理論の効率的な利用を目的とするということと，基礎的なアルゴリズムほど，より多くの問題に対して応用が可能だからである．また，アルゴリズム技術的な観点から，あいまいさのモデル化を行い，あいまいさを許容する計算手法の開発を行う．アルゴリズムの開発は，従来の総当り的な手法の改良を行うのではなく，データベースの疎性，解の疎性など，問題の持つ基礎的な構造と，それに効率良く適用するアルゴリズム理論的な技術の組合せによって行う．データベース解析の他，アセンブリングや系統樹作成といった基礎的な離散最適化問題についても，大規模な問題を効率良く解く手法の開発を目指す．

＜ 2007 年度の研究の当初計画＞

本研究班のメンバーは，現在までに，基礎的な類似検索・データマイニング・最適化の問題に対して，精度が高く効率の良いアルゴリズムを開発することに成功している．これらは，学会の文献賞や，プログラムコンテストの最優秀賞を受賞するなど，世界的に見てもレベルが高いと考えられる．しかし，これらの問題は，あくまで情報科学の観点から見て基礎的なものであり，ゲノム情報学の観点から見て基礎的とは限らない．

本年度は，ゲノム情報学で基礎的な問題の中から，実験結果の解析に使われるパターンマイニング，最適分類規則発見，配列の決定やアセンブリなどで用いられる相同性の発見アルゴリズムと並び替えを行うアルゴリズムの開発に関して，最適化・アルゴリズム的な技術を適用して改善できる点を見つけ出し，そこに新たな技法の開発を計画した．

＜ 2007 年度の成果＞

・与えられたグラフから，クリークに近い構造を全て見つける問題，データベースから多くの項目にあいまいさを許容した意味で含まれる集合を全て見つけ出す問題に対するアルゴリズムを開発した．クリークとは，完全グラフとなっている部分グラフのことであり，クリークに近い構造は枝密度の濃い部分グラフのことである．また，多くの項目にあいまいさを許容した意味で含まれる集合は，データから互いに深く関連する要素のグループを発見するには良いモデルである．

・これらの実装はhttp://research.nii.ac.jp/~uno/codes-j.html にて公開されている．データベースから頻出するパターン・確からしいルールを全て発見するもの，グラフ・ネットワークに含まれる，パス，サイクル，木，クリークといった基本的な構造を列挙するものなどである．これら実装は，日本を始めとする世界の多くの研究者によって使用されており，他の手法に比べて非常に高速であるとの評価を受けている．

・ベクトル集合の各要素に真か偽が与えられているデータ集合に対するパターン抽出の基本問題に関する性質を解析した．具体的には，真のベクトルの多くに当てはまるが偽のベクトルにはほとんど当てはまらないようなパターンの列挙を考える．数値データや文字列データから種々の特性を抽出して2値データに変換した上で知識発見を行う手法も数多く提案されており，このようなパターンの生成は，ゲノム情報をはじめとする様々なデータ集合からの知識発見における基本問題のひとつである．この問題において，データ本来の構造とは関係のない偽パターンが生成されないために必要なデータ数について考察した．ランダムデータから生成されるパターン数が十分小さくなる程度のデータ数が必要であるとの仮説を置き，現実のデータに対する計算実験によりこれを検証した．また，この仮説が示唆するデータ数の漸近的な振る舞いを理論的に解析した．

・集合被覆問題に対する高速近似解法を設計する上で有効な手法を検討し，知見を得た．集合被覆問題は，要素集合族とそれらのコストが与えられたとき，全要素を被覆できる集合の組でコストの総和が最小となるものを見つける問題である．代表的なNP困難問題のひとつであり，データからの知識発見をはじめとする多くの応用を持つ．数理計画法の発展は，困難な組合せ最適化問題の最適解を求める分枝限定法や分枝カット法などの厳密解法だけではなく，良い近似解を求める発見的解法にも大きく寄与してきた．とくに，集合被覆問題に対しては，線形計画緩和やラグランジュ緩和を用いた発見的解法が提案されており，大規模な問題例に対する有効性が認知されている．今年度は，集合被覆問題に対して，緩和法を中心とした数理計画の手法に基づく発見的解法を比較検討した．そして，代表的なベンチマーク問題例に対する数値実験を通じて，解法の構成要素の中でとくに大規模な問題例に有効なものをいくつか明らかにすることができた．
・DNA解析等で利用される，所定の熱力学的制約を満たしたDNA配列集合を自動的に生成（設計）するアルゴリズムを提案した．　解析などで用いられる DNA配列は相補配列の熱力学的な振る舞いが強い制約となるため，このような配列集合で大きなサイズのものを設計することは一般に容易ではない．本研究では，強力な探索能力を持つ可変近傍型の局所探索法を用いたアルゴリズムを提案した．通常，局所探索アルゴリズムは非常に多くの解評価計算を要するため単純な適用では良い解が得られたとしても非常に大きな計算時間を要することとなる．提案手法では，近傍構造の類似性を最大限に利用することにより，高速な配列生成を実現している．計算シミュレーションでは，いくつかの既存手法よりも安定した性能を得られることが確認された．

＜国内外での成果の位置づけ＞

パターン発見や最適化の問題は，データマイニングなどデータベース工学の中では中心的な位置にあるが，構造が簡単であり，かつ巨大なデータでも実用的な時間・メモリで動くように設計されたものはそれほど多くない．これは，モデルの開発に研究の中心がおかれ，アルゴリズム的な研究や，実際にそれを用いた研究は後手に回っているためと思われる．このような状況は，あいまいさのモデルに関しても同様である．これに対して本研究のアルゴリズムは，計算量と実装上の効率の両方の面から効率化を行っており，構造が単純であることも手伝って，実装が容易に行える上，非常に高速な実装が得られている．メモリ使用量の点でも他のアルゴリズムより大きな優位性がある．このような実装は，特に萌芽的な研究を行う場合に役立つと考えられる．

＜達成できなかったこと，予想外の困難，その理由＞

・今年度は，高速相同性発見アルゴリズムを用いたアセンブリングアルゴリズムの開発を行う予定であったが，達成できなかった．これは，他研究者とのディスカッションで，アセンブリングよりはあいまいさを許容したデータマイニング問題に対する需要がより高いことがわかったため，研究の順序を入れ替えたという理由による．

＜今後の課題＞

・アセンブリングに関しては，数理的な意味での最適解が実際の生物学的な意味合いでのもっともらしい解になっているとは限らないため，このギャップを埋めるべく，実際に配列決定を行っている研究者とのディスカッションを行ってモデル化を進める必要がある．また，大規模な問題に対応するため，現在のアルゴリズムと実装をより改良する必要もある．
・これまでに開発した集合被覆発見アルゴリズムは，パターンマイニングと組合せることで分類規則発見に応用が可能である．しかし，実用上有効なアルゴリズムを開発するためには，最適化型のアルゴリズムから列挙型のアルゴリズムへの変更，正確な被覆のみを求めるのではなく，ある程度あいまいさを許容した被覆を求められるようにする，といった改良が必要である．これらの改良は今後の課題として残っている．
・今回開発したマイニングアルゴリズムは，アイテムセットと呼ばれる，各項目がアイテムの集合からなるデータベースを対象としたものである．これをマイクロアレイの実験データ解析などに用いることはできるが，塩基配列など文字列型のデータの解析に直接的に用いることはできない．文字列型のデータを許容するためには，アルゴリズムの改良が必要である．また，文字列には，編集距離やハミング距離といったエラーの尺度が数多く存在するため，計算速度を損なわない尺度をどのように設計しモデル化するかという点も今後の課題である。

＜成果公表リスト＞

1. 0801101744
   Uno, T., and Arimura, H., Efficient Polynomial Delay Algorithm for Pseudo Frequent Itemset Mining, Lecture Notes in Artificial Intelligence 4755, (Proceeding of Discovery Science 2007), 219-230 (2007).
2. 0801101751
   Saigo, H., Uno, T., and Tsuda, K., Mining complex genotypic features for predicting HIV-1 drug resistance, Bioinformatics, 23, 2455-2462, (2007).
3. 0801101756
   Uno, T., “An Efficient Algorithm for Finding Similar Short Substrings from Large Scale String Data”, The Pacific-Asia Conference on Knowledge Discovery and Data Mining 2008, to appear (2008).
4. 0801141559
   Haraguchi, K., Yagiura, M., Boros, E., and Ibaraki, T., A Randomness Based Analysis on the Data Size Needed for Removing Deceptive Patterns, IEICE Transactions on Information and Systems, to appear.
5. 0801141734
   Umetani, S., and Yagiura, M., Relaxation Heuristics for the Set Covering Problem, Journal of the Operations Research Society of Japan, 50(4), 350-375 (2007).
6. 0801151829
   Kawashimo, S., Ono, H., Sadakane, K., and Yamashita, M., Neighborhood Searches for Thermodynamically Designing DNA Sequence, Preliminary Proceedings of the 13th International Meeting on DNA Computing, Memphis, Tennessee, 211-220 (2007).