Vol.37 No.10 情 報 処 理 Oct. 1996

解 説

ゲノム情報

2. タンパク質配列解析を例題とした並列最適化処理†

石 川 幹 人††

1. マルチプルアライメントの問題

本稿では，代表的な配列解析法であるマルチプルアライメント(Multiple Alignment)の問題をとりあげて，実用的な規模の最適化問題を解決する1つのアプローチを解説する．紹介するアプローチは，筆者らが(財)新世代コンピュータ技術開発機構において取り組んだ研究である．

マルチプルアライメントは，遺伝子やタンパク質の機能・構造予測，生物種の進化系統樹の作成の際に欠かせない解析法である．マルチプルアライメントとは，短的にいえば，タンパク質のアミノ酸配列を複数並べ，配列のところどころにギャップ“－”を入れることで，共通文字(あるいは性質の似たアミノ酸を表す文字)が同じ縦のカラムに並ぶ(図-1(c)参照)ようにする問題である．

マルチプルアライメントの問題は，次の評価スコアを最適化することで，ある程度解決できることが知られている．

$$AlignmentScore = \sum_{i<j}^{seq.pair} \sum_{k}^{column} MatchScore(A_{ik}, A_{jk}).$$

$$MatchScore(A_{ik}, A_{jk}) = \begin{cases} Dayhoff(A_{ik}, A_{jk}). \\ \quad : A\text{がともにアミノ酸} \\ 0 : A\text{がともにギャップ} \\ p : A\text{がアミノ酸と先頭のギャップ} \\ q : A\text{がアミノ酸と2番目以降のギャップ} \end{cases}$$

式のなかで，アミノ酸同士の類似性尺度を与えているのは，Dayhoffの提案したマトリックス[1]で，該当アミノ酸対の遺伝的変異が偶然に対していかに大きいかを数値化したものである．また，ギャップを挿入するときのコストは，ギャップの長さに対する1次式で与える．p, qの値は問題に応じた調整が必要であるが，$p = -8$, $q = -1$程度とするのが一般的である．

この最適化問題としてのマルチプルアライメントは，ダイナミックプログラミング(DP: Dynamic Programming)により，原理的には解決できる[2]．2つのアミノ酸配列をアライメントする場合，この2つの配列を2次元のネットワークの辺に対応させ，斜め方向のアークに，アークの位置に対応するアミノ酸の類似度を，縦および横方向のアークに，ギャップを挿入するときのコストを割り振る．このように問題を表現すると，最適なアライメントを求めることは，このネットワーク上の最良の経路を求めることに対応し(図-2の右端図を参照)，各ノードに至る最良経路を段階的に決定していくことで解決される．

理論的には，DPは一度に何本の配列でも同時に比較でき，与えられた評価値における最適なマルチプルアライメントが得られる．ところが，N本の配列を同時にアライメントするN次元のDPは，概して，配列のN乗の計算量とN乗のメモリ量が必要となり，現実的に可能なのは3次元程度までである．一方で，実用的なマルチプルアライメントでは，通常，長さ100以上の配列を20本以上も扱うので，多次元DPによる厳密解法では，一般のマルチプルアライメントの問題には，とても対処できない．

そこで，準最適解を求める近似解法が必要となるわけであるが，DPで2本ずつアライメントした結果を組み合わせるのが，従来の典型的な近似解法であった．しかし，それでは精度が十分でなく，専門家が人手で行う品質に達していなかった(研究の歴史は文献3)を参照されたい)．

† Parallel Optimization Processing Applied to Protein Sequence Analysis by Masato ISHIKAWA (Matsushita Electric Industrial Co., Ltd.).

†† 松下電器産業（株）マルチメディアシステム研究所

```
SIV  : ---------------GGNQEIDHLVSQGIRQVLFLEKIEPAQEEHSKYHSNIKELVFKFGLPRLVAKQIVDTCDKCHQKGEAIHGQVNSDLGTWQ---------------
HIV1 : ---------------AHKGIGGNEQVDKLVSAGIRKILFLDGIDKAQDEHEKYHSNWRAMASDFNLPPVVAKEIVASCDKCQLKGEAMHGQVDCS---------------
EIAV : ---------------ISQRGDKGFGSTGVFWVENIQEAQDEHENWHTSPKILARNYKIPLTVAKQITQECPHCTKQGSGPAGCVMRSPNHWQADC---------------
SRV1 : ---------------GPIAHGNQKADLATKTVASNINTNLESAQNAHTLHHLNAQTLKLMFNIPREQARQIVRQCPICATYLPVPHLGVNPRGLL---------------
MMTV : ---------------LPGPLAQGNAYADSLTRILTALESAQESHALHHQNAAALRFQFHITREQAREIVKLCPNCPDWGHAPQLGVNPRGLKPRV---------------
HTLV1: ---------------TNLPDPISRLNALTDALLITPVLQLSPAELHSFTHCGQTALTLQGATTTEASNILRSCHACRGGNPQHQMPRGHIRRGLL---------------
BLV  : ---------------SHPIASLNNYVDQLLPLETPEQWHKLTHCNSRALSRWPNPRISAWDPRSPATLCETCQKLNPTGGGKMRTIQRGWAPNHI---------------
```

(a) 初期状態

```
SIV  : ---------------GGNQEIDHLVSQGIRQVLFLEKIEPAQEEHSKYHSNIKELVFKFGLPRLVAKQI-------VDTCDKCHQKGEAIHGQVNSDLGTWQ--------
HIV1 : ---------------AHKGIGGNEQVDKLVSAGIRKILFLDGIDKAQDEHEKYHSNWRAMASDFNLPPV-------VAKEIVASCDKCQLKGEAMHGQVDCS--------
EIAV : ------------ISQRGDKGFGSTGVFWVENIQEA---QDEHENWHTSPKILARNYKIPLTVAKQITQECPHCTKQGSGPAGCVMRSPNHWQADC---------------
SRV1 : ---------------GPIAHGNQKADLATKTVASNINTNLESAQNAHTLHHLNAQTLKL-MFNIPREQARQIVRQCPICATYLPVPHLGVNPRGLL---------------
MMTV : ---------------LPGPLAQGNAYADSLTRILTALESAQESHALHHQNAAALRFQFHITREQAREIVKLCPNCPDWGHAPQLGVNPRGLKPRV---------------
HTLV1: ------------TNLPDPISRLNALTDALLITPVL---QLSPAELHSFTHCGQTALTLQGATTTEASNILRSCHACRGGNPQHQMPRGHIRRGLL---------------
BLV  : ------------SHPIASLNNYVDQLLPLETPEQW---HKLTHCNSRALSRWPNPRISAWDPRSPATLCETCQKLNPTGGGKMRTIQRGWAPNHI---------------
```

(b) 3回変形後

```
SIV  : -----------GGNQEIDHLVSQGIRQVLFLEKIEPAQEEHSKYHSNIKELVF--KFGLPRLVAKQIVDTCDKCHQKGEAIH---GQVNSDLGTWQ---------------
HIV  : -----AHKG-IGGNEQVDKLVSAGIRKILFLDGIDKAQDEHEKYHSNWRAMAS--DFNLPPVVAKEIVASCDKCQLKGEAMH---GQVDCS---------------
EIAV : ------------ISQRGDK--GFGSTGVFWVENIQEAQDEHENWHTSPKILAR--NYKIPLTVAKQITQECPHCTKQGSGPA---GCVMRSPNHWQADC-----------
SRV1 : --------GPIAHGNQKADLATKTVASNI-NTNLESAQNAHTLHHLNAQTLKL--MFNIPREQARQIVRQCPICATYLPVPHL--GVNPRGLL-----------------
MMTV : -----------LPGPLAQGNAYADSLTRILTALESAQESHALHHQNAAALRF--QFHITREQAREIVKLCPNCPDWGHAPQL--GVNPRGLKPRV---------------
HTLV1: ----------TNLPDPISRLNALTDALLITPVLQLSPAELHSFTHCGQTALTL--Q-GATTTEASNILRSCHACRGGNPQHQMPRGHIRRGLL-----------------
BLV  : ------------SHPIASLNNYVDQLL----PLETPEQWHKLTHCNSRALSRWPNPRISAWDPRSPATLCETCQKLNPTGG---GKMRTIQRGWAPNHI----------
                                      #  #                    #  #          #
```

(c) 最終結果

図-1 シミュレーテッドアニーリングによるマルチプルアライメント

2. 3次元DPの並列高速化

従来の近似解法では，2本ずつアライメントした時点に発生するエラーが，結果を組み合わせていくうちに次々と累積してしまうところに問題点があった．そこで，同時に比較する配列の本数を増やすことが近似解法の精度をあげるのに重要と思われた．

我々は，並列計算機(並列推論マシン)を用いて，まず3次元DPの並列化による高速化を行った．具体的には，3次元DPのネットワークを実装し，そこに配列3本のアライメントの問題を流し，次々とパイプライン処理をさせた[4]．64台の要素プロセッサを使用して，37倍の並列効果が得られた．さらに，それらの配列3本のアライメントを組み合わせることで，マルチプルアライメントを形成した[5]．

この方法は，配列2本のアライメントを組み合わせる従来の近似解法に比べて，エラーが少なく精度が高い．しかしながら，実用規模の問題では，3次元DPの処理にともなう実行時間の増大に見合う，十分な品質の解は得られなかった．

3. シミュレーテッドアニーリング

アライメントの精度をあげるには，全配列を同時に評価することが肝要と考え，我々は，次にシミュレーテッドアニーリング(SA: Simulated Annealing)の適用を試みた[6]．SAは，組合せ最適化問題の解空間を，微小変形を繰り返して得られる解を温度パラメータに依存した確率で選択しながら，準最適解を求めて探索するアルゴリズムである．

まず，微小変形を次のように定義した．図-1(a)にあるように，アライメントすべき配列の頭部や尾部に，あらかじめ十分な数のギャップを付加しておく(図の左端のSIVなどは配列の名前を表す)．そして，配列群から1本の配列をランダムに選び，その配列に対して，任意のギャップと任意のカラム位置をそれぞれランダムに選択し，選択されたギャップを選択されたカラム位置へ移動させる(単独変形)．さらに，マルチプルアライメントには，ギャップが固まって入りやすいことから，ギャップを長方形のブロック状にまとめて動かす操作(ブロック変形)も導入した．図-1(b)は，初期状態から単独変形を1回，ブロック変形を2回行った後の状態を示している．

SAの探索には，並列計算機を用いて並列多点探索を行った．解の並列評価を利用して温度パラメータを自動設定する，2つの方式を実装した[7]．図-1(c)は，63台の要素プロセッサでSAの並列探索を行い，1時間ほど(各プロセッサでの微小変形3万回に相当)で得られた解である．共通配列がよく揃っていることがわかる．

しかし，近似解法といえども，実用規模の問題を解くには，依然として実行時間の課題が残る．実用規模の問題は，アミノ酸の数にしてこの4倍以上であり，ギャップの数もまた4倍以上となれば，微小変形の種類は16倍以上となる．すると，

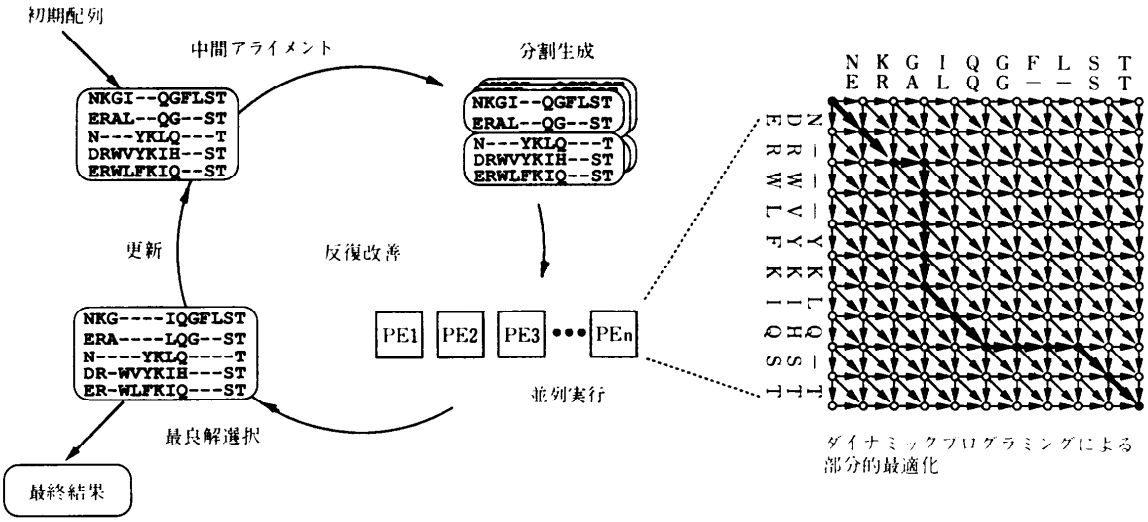

図-2 並列反復改善法（最良優先探索）によるマルチプルアライメント

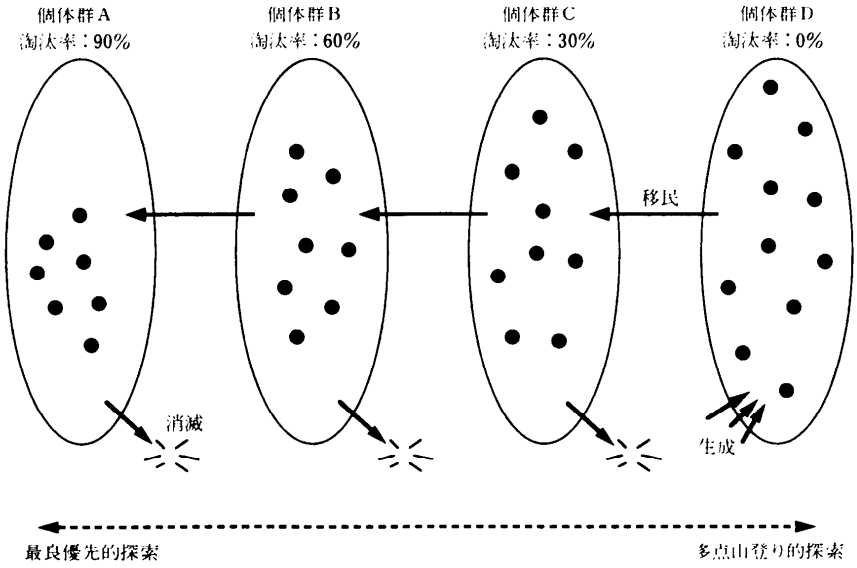

図-3 遺伝的アルゴリズムのマルチ個体群方式

収束までに要する時間は，数十時間から数百時間にのぼることも予想される．

4. 並列反復改善法

前章で述べたSAの解法では，どんな微小変形を導入するかが鍵であった．単独変形に加えてブロック変形を導入したところ，単独変形のみに比べ，解の改善効果に著しい向上があった．この事実から，最適解に近づけるマクロな変形操作を，問題に応じてよく吟味して導入することが，実用的な解決法への近道と思われた．

マルチプルアライメントにおいて，最適解に近づけるマクロな変形操作には，やはりDPが第一候補にあげられた．折しも，マルチプルアライメントに2次元DPを繰り返し適用して，解を改善する方法が提案された[8]．我々は，そのアルゴリズムに限定分割と並列化を導入して発展させ，新たに並列反復改善法を開発した[9]．

並列反復改善法は図-2に示すような構成である．まず，ギャップのない状態を初期アライメン

トとして改善サイクルへ投入する．各改善サイクルでは，配列群を経験的な方略で2つのグループに分ける分割を複数生成し，各要素プロセッサ(PE:Processing Element)に割り当てる．そしてPEごとに，各グループ間のDPによる最適化を並列に行う．その部分的な最適化の結果得られた複数の解のうち，最も評価スコアの高い解を，次の改善サイクルの初期状態として採用する．もはや，改善が見られなくなったならば，最終の解とする．

並列反復改善法を並列計算機上に実装したところ，実用規模の問題でも十数分で，高精度な準最適解が得られることが判明した．我々は，この並列反復改善法を中核の機能とするアライメント編集ツール[10]を作成した．このツールは，生物学研究者に利用され，生物学上の研究に役立っている[11]．

5. 遺伝的アルゴリズム

前章で述べた並列反復改善法は最良優先探索であり，時にはあまりよくない局所解に陥る問題点がある．それを避ける1つの手段は，改善サイクルの内部で並列化せずに，各プロセッサごとに単純な改善サイクルを回し，並列に多点山登り探索することである[9]．だが，そうすると今度は，収束までに時間がかかるという問題点が出てくる．

我々は，並列反復改善法の改善サイクルを変形操作(突然変異)とみなすことで，並列反復改善法を遺伝的アルゴリズム(GA: Genetic Algorithm)の枠組みに定式化できることに気づいた[12]．このGAの定式化では，淘汰率を変えることで，解空間の探索戦略を調整できる．淘汰率を高くすると探索戦略が最良優先的に，低くすると多点山登り的になる．すると，**図-3**に示すようなマルチ個体群方式のGAを適用すると，有効な最適化処理が可能である[13]．

マルチ個体群方式のGAでは，淘汰率を変えた個体群を複数用意し，個体群中の各個体では，1つの解の反復改善を行う．そして，淘汰率が低い個体群(多点山登り的な探索をしている)でスコアのよい解をもつ個体がみつかったならば，淘汰率が高い個体群(最良優先的な探索をしている)へ移民させ，その解の近傍の集中的な探索をする．その一方で，全個体の解が一様になってきた個体群は，局所探索に陥ってきたと判断し，個体数を減らす．空いた計算資源は，解の多様性が残っている淘汰率の低い個体群に割り当て，そこの個体数を増やして多点探索をするのである．

我々は，マルチ個体群方式のGAを並列計算機上に実装した．その結果，最良優先探索の並列反復改善法のような局所最適解に陥ることなく，かつ，単なる多点山登り探索よりも，きわめて早期に解の改善効果が得られることが，実験から明らかとなった．また，部分アライメントを交換するクロスオーバ操作の導入により，初期段階のスコアの向上が，さらに早まる効果が得られた．

6. おわりに

マルチプルアライメントのような実用規模の問題は，最適化問題として定式化すると，どうしても探索空間が大きくなりがちである．そうした問題の解決には，たとえ並列処理による高速化を行ったとしても，無作為な微小変形だけでは不十分である．実用的な時間内の解決には，解の改善に有効な，マクロな変形を積極的に導入することが不可欠である．マルチプルアライメントの問題の場合は，微小変形の単位として，ダイナミックプログラミングによる部分的な最適化を導入することが，実用的な問題解決に有効であった．また，マルチ個体群方式の遺伝的アルゴリズムの枠組みで，戦略的な探索が実現できることが示された．

なお，以上に述べた諸々のソフトウェアは，http://www.icot.or.jp/ より，無償公開されている．

謝辞 共同で研究にあたった当時の(財)新世代コンピュータ技術開発機構の方々，および，ご指導いただいた文部省科学研究費「ゲノム情報」の諸先生方に感謝いたします．

参考文献

1) Dayhoff,M.O., Schwartz,R.M. and Orcutt, B. C. : A Model of Evolutionary Change in Proteins, Atlas of Protein Sequence and Structure, Vol.5, No.3, Nat. Biomed. Res. Found., Washington DC, pp.345-352 (1978).

2) Needleman,S.B. and Wunsch,C.D. : A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins, J. Mol. Biol., Vol.48, pp.443-453 (1970).

3) 石川幹人，金久 實：文字を比較し並べる，ヒト

ゲノム計画と知識情報処理，培風館，pp.65-97 (1995).
4) 戸谷智之，星田昌紀，石川幹人，新田克己：並列3次元ダイナミックプログラミング法による蛋白質の配列解析，情報処理学会プログラミング研究会 5-14, pp.127-134 (1991).
5) 石川幹人，星田昌紀，広沢 誠，戸谷智之，鬼塚健太郎，新田克己，金久 實：並列推論マシンを用いたタンパク質の配列解析，情報処理学会情報学基礎研究会 23-2, pp.1-14 (1991).
6) Ishikawa, M., Toya, T., Hoshida, M., Nitta, K., Ogiwara, A. and Kanehisa,M.:Multiple Sequence Alignment by Parallel Simulated Annealing, Computer Applications in the Biosciences, Vol.9, No.3, pp.267-273(1993).
7) 戸谷智之，石川幹人，星田昌紀，荒木 均：並列シミュレーテッドアニーリングによるアミノ酸配列解析，情報処理学会記号処理研究会 69-1, pp.1-8 (1993).
8) Gotoh, O.: Optimal Alignment between Groups of Sequences and Its Application to Multiple Sequence Alignment, Computer Applications in the Biosciences, Vol.9, No.3, pp.361-370 (1993).
9) 石川幹人，十時 泰，戸谷智之，星田昌紀，広沢誠：並列反復改善法によるタンパク質の配列解析，情報処理学会論文誌, Vol.35, No.12, pp.2816-2830 (Dec. 1994).
10) Ishikawa, M., Totoki, Y., Tanaka, R. and Hirosawa, M.: Multiple Sequence Alignment Editor Featured by Constraint-based Parallel Iterative Aligner, Proc. 3rd Int. Conf. Bioinformatics and Genome Research, World Scientific Publishing, pp.385-396 (1995).
11) Nakahigashi, K., Yanagi, H. and Yura, T. : Isolation and Sequence Analysis of RpoH Genes Encoding Sigma-32 Homologs from Gram Negative Bacteria, Nucleic Acids Research, Vol.23, No.21, pp.4383-4390 (1995).
12) Ishikawa, M., Toya, T., Totoki, Y. and Konagaya, A.: Parallel Iterative Aligner with Genetic Algorithm, Proc. Genome Informatics Workshop IV, Universal Academy Press, pp.84-93 (1993).
13) 戸谷智之，石川幹人：マルチ個体群の並列遺伝的アルゴリズムを用いたタンパク質の配列解析，情報処理学会論文誌, Vol.36, No.11, pp.2549-2558 (1995).

(平成8年6月26日受付)

石川 幹人 (正会員)

1959年生．1982年東京工業大学理学部応用物理学科卒業．同大学院物理情報工学専攻を経て，松下電器産業(株)に入社．1989-95年(財)新世代コンピュータ技術開発機構に出向．東京工業大学大学院非常勤講師，明治大学法学部兼任講師を歴任．博士(工学)．知識情報処理，特に生物学・法律への応用に興味を持つ．第8回元岡賞受賞．人工知能学会，認知科学会，生物物理学会，法とコンピュータ学会各会員．