H-033

遺伝的アルゴリズムを用いた DNA チップによる配置決定

Reconstruction of DNA Sequencing Information from a DNA Chip Using a Genetic Algorithm

上杉 英司† 外山 史† 東海林 健二† 宮道 壽一†
Eiji Uesugi Fubito Toyama Kenji Shoji Juichi Miyamichi

1．はじめに

DNA 配列の決定はゲノムサイエンスにおいて非常に重要な問題の 1 つである。DNA 配列の決定方法として、SBH(Sequencing By Hybridization)により、塩基配列を決定する方法が注目されている。これは、DNA チップ技術を用いて、塩基配列が未知の DNA から、それに含まれる L 文字長（L=10～12）のあらゆる部分塩基配列の集合を求め、これらの集合から元の塩基配列を再構成する方法である(図 1)。部分配列の集合にエラーが含まれる場合この問題は NP 困難であることが知られている[1]。本研究では、遺伝的アルゴリズム（GA）を用いてエラーが含まれた部分配列から DNA 配列を決定する手法を提案する。GenBank より得た DNA 配列に対してシミュレーション実験を行った結果、エラーを含んだ部分配列において正しい DNA 配列を決定することができた。

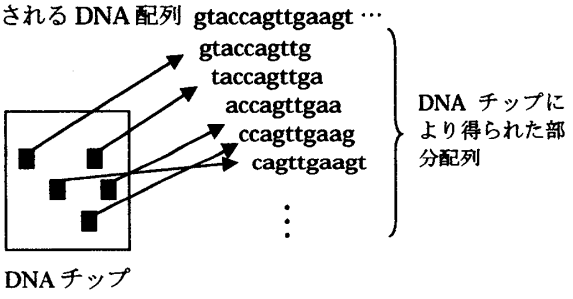

図 1：SBH 法

2．SBH による配置決定問題

SBH 法において DNA チップより得られたデータにエラーが存在しない場合、DNA 配列の決定は逐次的な 1 文字シフトのマッチングにより容易に実現できる。しかし、DNA チップより得られたデータには存在するはずの部分配列が存在しないエラー（false negative）や存在しないはずの部分配列が存在するエラー（false positive）が発生する。このようなエラーがある場合、この問題は、図 2 のように、各部分配列の接続関係を有向グラフで表した場合、この有向グラフにおいて、最適なパスを求める最適パス問題と置くことができる。ここで、図 2 において頂点の番号はラベル付けされた部分配列、辺の番号は重みを表し、何文字シフトで 2 頂点の部分配列がマッチするかを表している。本研究では、この最適パス問題を GA を用いて求める手法を提案する。なお、最適パスにはループが存在しないものとする。また、本研究ではもとの DNA 配列の長さは既知であるとする。

†宇都宮大学

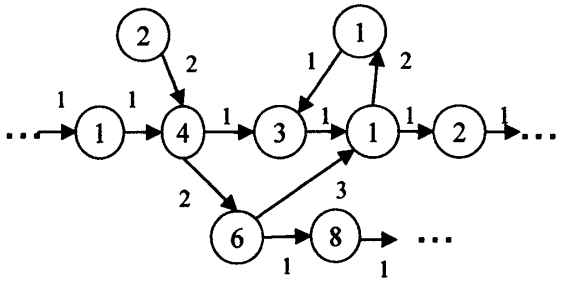

図 2：マッチング関係のグラフ

3. 提案アルゴリズム

3.1 1 文字シフトにおける有向グラフの作成

DNA 配列の長さを N、部分配列の長さを L とする。最初に、ラベル付けされたそれぞれの部分配列に対し、右に 1 文字シフトしてマッチする配列を調べる。次に、このマッチング関係をもとに 1 文字シフトに関する有向グラフを作成する。これは、図 2 の辺番号がすべて 1 となる有向グラフである。

3.2 部分接続関係の決定

1 文字シフトしたときのマッチングにより得られた有向グラフを用いて、部分的な配列を決定する。1 文字シフトにおけるグラフから得られたパスは、部分配列に重複して含まれる塩基数が多いためエラーである可能性が小さい。よって接続関係は決定したとみなし、以降接続関係の変更はしない。ただし、作成した有向グラフにおいて、入次数が 0 又は出次数が 0 となる頂点から長さ T までの頂点は部分配列に重複して含まれる塩基の数が少なく、エラーが含まれている可能性が大きいので、この部分の部分配列の接続関係は決定しないこととする。

3.3 2 文字以上シフトしたときのグラフの作成

1 文字シフトのグラフを作成した後、同様にある配列に対して右に 2 文字シフトしてマッチする配列を調べ、1 文字シフトした時のグラフにつけ加える。このとき、マッチした 2 つの配列に対して、すでにパスが存在しているのならば辺はつけ加えない。3.2 の処理ですでに接続関係が決定している配列はこの処理の対象外である。同様の処理を N_{shift} 文字まで繰り返し、図 2(N_{shift}=3)のような N_{shift} 文字までシフトしたときのグラフを作成する。

3.4 GA を用いた DNA 配置決定

3.3 で作成したグラフから GA を用いて最適なパスを探索することにより DNA 配列を決定する。最適なパスとは、入力部分配列をできるだけ多く含み、パスから得られる配列の長さが、既知である DNA 配列の長さ N と等しくなるパスのことである。

3.4.1 染色体の定義

初めに、パスの先頭となりうる部分配列を得るため、1 文字シフトのグラフと最終的に作成したグラフにおいて入次数が 0 である頂点を探索する。次にグラフにおいて分岐している頂点を探索する。なお、グラフにおいて出ていく辺のシフト数がすべて 3 以上である頂点は、正解配列において、その頂点が表す部分配列とマッチする部分配列がない可能性があるためどの辺も選択しないという選択肢を追加し、分岐点とする。これらを用いて染色体を図 3 のように定義する。ここで、左端の遺伝子はパスの先頭となる頂点のラベル番号であり、2 番目以降の遺伝子の値は複数ある辺の中から、どの辺(何番目の辺。接続する辺のシフト数がすべて 3 以上の場合、辺を選択しないという選択肢も含まれる)を選ぶのかを表している。よって、分岐点の数＋1(左端の遺伝子)が染色体の長さ Nv となる。各遺伝子の遺伝子座に対応する頂点のラベル番号はあらかじめ決っている。例えば、図 3 の左から 2 番目の遺伝子はラベル番号 36 の頂点がどの辺を選択するかを表しており、ここでは 2 なので 2 番目の辺を選択する。

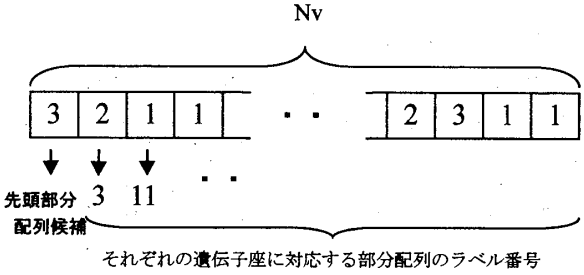

図 3：染色体の定義

3.4.2 適応度の定義

各染色体に対して、どの程度正解に近いかを表す適応度を定義する。DNA チップから得られるデータの大部分は正しいデータであると考えられるため、できるだけ多くの部分配列を含むことが望ましい。さらに DNA チップに入力される DNA 配列の長さは既知であるため、GA より得られたパスから求められる配列の長さは入力される DNA 配列の長さに近い方が良いと考えられる。そのため本論文では両方を考慮した以下の式で適応度 F を定義する。

$$F=\alpha\frac{N-|N-K|}{N}+(1-\alpha)\frac{H-|H-M|}{H} \quad \cdots(1)$$

ここで N は DNA 配列の長さ、K は GA より得られた配列の長さ、H は DNA 配列の部分配列数、M は GA より得られたパスに含まれる部分配列数である。α は配列の長さ、部分配列数のそれぞれの重みを決定するパラメータである。実験では α を 0.5 とした。

3.4.3 適応度の計算方法

作成したグラフにおいてできるだけ多くの部分配列を含み、入力される DNA 配列の長さに近いパスを求める。適応度は、染色体によって表される出次数が 1 以下となるグラフから、最も長いパスを求め、このパスに対して式(1)を用いて適応度を計算する。ただし、最も長いパスを求める際に、ループが発生した場合、染色体によって表される辺番号とは別の、ループの発生しない辺を選択し、その部分の遺伝子を変更することによって、ループの発生を防いでいる。選択する辺が複数ある場合には、ランダムに辺を選ぶようにした。得られたパスとパスの先頭又は最後の頂点を取り除いたパスの適応度を比較し、適応度が高い方をその染色体で表されるパスとする。

3.4.4 遺伝規則

GA において世代交代をする際、上位の数パーセントの個体をそのまま次世代に残す、エリート保存攻略を用いた。交叉の方法は一様交叉を用い、突然変異は、ある確率で遺伝子の値をランダムに変更する方法を用いた。

4 実験結果

GenBank より得た長さが 109、209、309、409、509 の DNA 配列を用い、ランダムに false negative と false positive のエラーを発生させて、これらのエラーの含まれた部分配列から元の DNA 配列を求めた。それぞれの入力部分配列のエラー率は false negative、false positive それぞれ 20%とした。世代交代数、個体数、突然変異率(%)、エリート保存率(%)、部分接続関係の決定に用いるパラメータ T は、それぞれ 5000、500、0.1、3、3 とした。グラフ作成に用いるパラメータ N_{shift} の値は、配列の長さ 109、209 では 6、309、409、509 では 5 とした。実行結果を表 1 に示す。正解率は 40 種類の DNA 配列に対して実験を行い DNA 復元に成功した回数から求めた。ただし、パスの最初と最後の頂点に false negative が存在する場合は、これらの頂点を除いた最適パスを見つけることができれば正解とした。また、Blazewicz らによる、ハイブリッド GA を用いた手法[2]での結果も表 1 に示す。

表1:実験結果

入力部分配列数	100	200	300	400	500	全体の正解率
提案手法正解率	100% (40/40)	75% (30/40)	60% (24/40)	27.5% (11/40)	15% (6/40)	55.5% (111/200)
ハイブリッドGA正解率	100% (40/40)	77.5% (31/40)	50% (20/40)	22.5% (9/40)	12.5% (5/40)	52.5% (105/200)

表 1 より、本手法の方がハイブリッド GA による手法よりも全体的に高い正解率で正しい DNA 配列を決定することができた。入力配列の増加に伴い正解数が減少する原因は、入力配列の増加とともにエラーの数も増加し、作成するグラフにおいて分岐点の数が増加した、N_{shift}、T の値が小さく、正しい接続関係を持ったグラフを作成できなかったことなどが考えられる。

5 おわりに

本手法では、GA を用いて DNA チップからの配置決定を行う方法を提案した。今後の課題としては、生物学的な情報を取り入れることにより、エラー率が高い場合や入力部分配列が多い場合においても、正しい DNA 配列を決定することができるようにすることや、グラフにおけるループ、すなわち同じ配列の繰り返しがある場合について考慮することなどが挙げられる。

参考文献

[1] Jack Blazewicz, Marta Kasprzak, "Complexity of DNA sequencing by hybridization," Theoretical Computer Science, Vol.209, pp.1459-1473, 2003.

[2] Jack Blazewicz, Marta Kasprzak, "Hybrid Genetic Algorithm for DNA Sequencing with Errors," Journal of Heuristics, 8:pp.459-502, 2002.