公募研究：2000～2002、2004年度

線形配置アルゴリズムによる遺伝子データの統合と表示手法の開発

●中谷 明弘

東京大学大学院新領域創成科学研究科

〈研究の目的と進め方〉

複数の遺伝子データの統合を行うことを目的とし、そのために必要となる情報処理技術の開発を進めた。下に示すような遺伝子情報はWWW上から簡単に取得できるようになっているが、いざ自らのデータの解析に用いようとしても難しいことが多い。

・遺伝子発現データ（トランスクリプトーム）
・相互作用データ（インタラクトーム）
・配列相同性データ（パラログ・オーソログ）
・遺伝子注釈データ（オントロジー）
・代謝/制御に関するデータ（パスウェイ）
・遺伝子破壊による形態変異データ

これら遺伝子情報それぞれが異なった構造（ネットワーク構造をはじめ、木構造、時系列的構造、階層構造など）を成しており、複数の遺伝子データの統合（比較・重ね合わせ・相互補完など）が生物学的にも情報科学的にも本質的に難しいことが理由のひとつである。そのため、未知データを既知の情報に結び付けて機能解明の手がかりとすることは、基本的な手法であるにもかかわらず必ずしも十分に行えていないのが現状である。とりわけ、遺伝子間の多対多の関係に注目する場合に、困難さが増大することは周知の通りであり、手動での解析が現実的ではないことも多い。

この問題を解決するために、本研究では遺伝子データの統合（integration）を数理的な手法を用いて行うことを目的とし、そのために必要なアルゴリズムやプレゼンテーション方法を開発し、実用的な応用を目指した。以下のような項目に関するアルゴリズムやプログラムの開発を行った。

・組合せ最適化問題
・線形配置問題
・アソシエーション・スタディー
・階層的クラスタリング
・ネットワーク探索
・ネットワーク比較
・ネットワーク可視化

プログラムはスクラッチから独自開発した。

〈研究開始時の研究計画〉

互いに関連した内容であったが、通算4年度分の個別の公募研究であり、各年度ごとに研究計画を立てた。

2000年度「並列データマイニングによる遺伝子ネットワークからの情報抽出」：当時所属していた京都大学化学研究所ではKEGGデータベースが内部使用可能であったので、その中に蓄積されたネットワークのデータからの情報抽出を行うこととした。

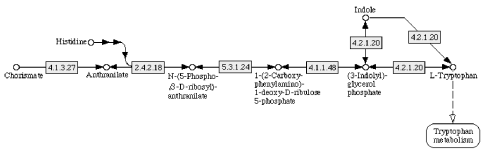
2001年度「並列データマイニングによる遺伝子ネットワークからの相関遺伝子クラスタの抽出」：相関遺伝子クラスタ（複数のネットワークで共通してクラスタ化している遺伝子群）を抽出することとした。2000～2001年度当時は、所属機関等に新規に導入された並列計算機群の使用アカウントをもっていたので、これを活用するために、並列計算機上でのデータ処理技術も開発することとした。

2002年度「相関遺伝子クラスタの抽出に向けた複数グラフの比較手法の開発」：前年度までに開発したプログラムを用いて、複数の酵母相互作用ネットワークの比較や重ね合わせを行うこととした。

2004年度「線形配置アルゴリズムによる遺伝子データの統合と表示手法の開発」：複数の遺伝子データで共通しているネットワーク構造を抽出・表示する手法を開発することとした。階層的クラスタリングの結果得られるデンドログラムの枝の向きを入れ替えて、クラスタ間の関係行列の配置パタンの最適化を行うことを目指した。その際に必要となる線形配置問題を解くアルゴリズムを開発・実装する。

〈研究期間の成果〉

2000年度：　ゲノム上での隣接関係、代謝パスウェイ上での隣接関係、配列の類似関係、立体構造の類似関係などの遺伝子間の相互関係の集合はネットワーク（遺伝子ネットワーク）を構成する。複数の遺伝子ネットワーク上で共通して近接している遺伝子群は機能的により強く相関していると想定して、そのような条件を満たしている遺伝子クラスタの抽出を行った。例えば、下の図は、ゲノム中での位置、代謝パスウェイでの位置（KEGGデータベース）、立体構造の類似性（SCOPデータベース）を示すネットワークから抽出されるネットワークコンポーネントを示している。これは、大腸菌のトリプトファンオペロンに関する例で、これらのゲノム上で隣接する遺伝子は、代謝パスウェイ上でも隣接し、また、立体構造が互いに類似している。このように指定した条件を満たす遺伝子群を自動的にリストすることが可能になった。

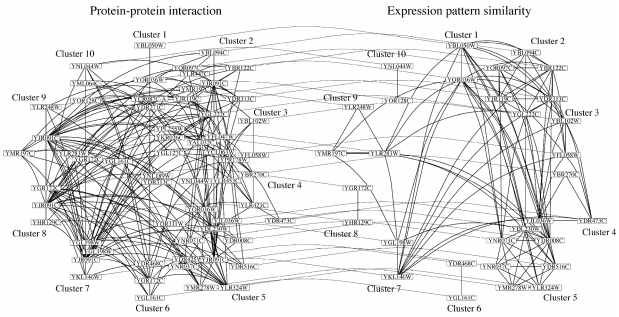


この処理では、複数遺伝子ネットワークからの相同部分（類似トポロジー部分）の抽出を行っているが、この処理は、データサイズの増加に対して、効率よく行うことができないことが知られている。これに対処するために、ヒューリスティックスを導入して探索空間を縮小するグラフの比較手法を開発し並列計算機上で実装した。一連の研究の最初であるので、ソフトウェアの開発を行った。主として、ネットワーク探索と階層的クラスタリングを行うプログラムを作成した。前者は複数のプロセッサ上で並列実行できるように実装し、後者は並列計算機の大量のメモリを活用するように作成した。これらのプログラムは次年度以降も改良を続けて、一部は5年以上経過した現在も使用している。

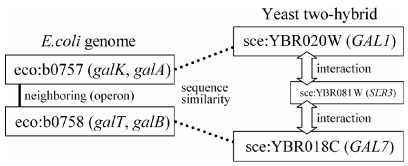
2001年度：　前年度までに開発した手法・実装の改善

を行うとともに、データへの適用を試みた。解析データとして、京都大学化学研究所で開発されているKEGGデータベース内に蓄積されているデータセットを使用した。例として、Y2Hによるタンパク・タンパク相互作用（PPI）のデータのスクリーニングを開発した手法を用いて行うことを試みた。例えば、下の例は、酵母PPIデータから得られた10個のクラスタ（左側）と、それに対応するマイクロアレイでの遺伝子発現パタン類似ネットワーク（右側）の示したものである。このように互いに相関し合っているクラスタを「相関遺伝子クラスタ」と呼ぶこととして、これを自動的かつ網羅的に抽出することを行った。また、得られたクラスタメンバー遺伝子の関係を可視化するプログラムを作成した。

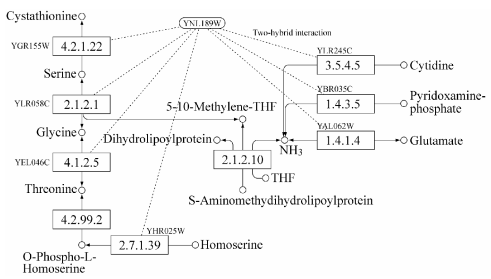
出力はコンピュータ画面上で表示できるほか、この図のようなポストスクリプト形式で行うことができる。



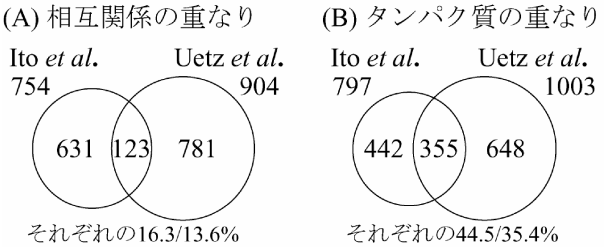
同様にして、京都大学化学研究所で開発されている配列相同のデータベースであるKEGG/SSDBのデータに従って、大腸菌の遺伝子に関するデータと酵母PPIデータとを関連づけることによって、異なる生物種のデータを用いて相関遺伝子クラスタを発見する手法の有用性を確認した。以下は得られた結果の一例である。所謂ロゼッタストーン解析に類似して、特定の生物種でのオペロン様構造の情報を用いて、別の生物種での物理的な相互作用を予測しようとした例でもある。



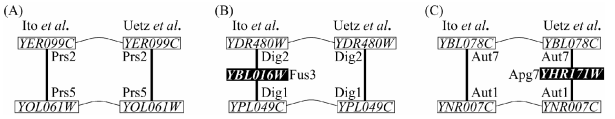
また、下は代謝パスウェイのデータに酵母PPIデータをマッピングした例である。これらの例はいずれも一見すると単純なものではあるが、データの中からこのような情報を人手で探し出すことは不可能である。



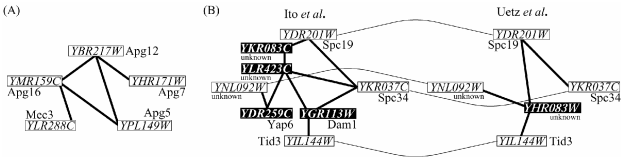
2002年度：　前年度までに開発したツールを用いて、2つの酵母Y2H実験データ（Ito et al., 2001及びUetz et al., 2000）をグラフ構造として表現し比較を行った。両データは、下図(A)に示す通り、相互関係の重なりが小さいことが指摘されている。しかしながら、(B)に示すように両データに含まれるタンパク質の重なりは前者のそれよりも大きい。



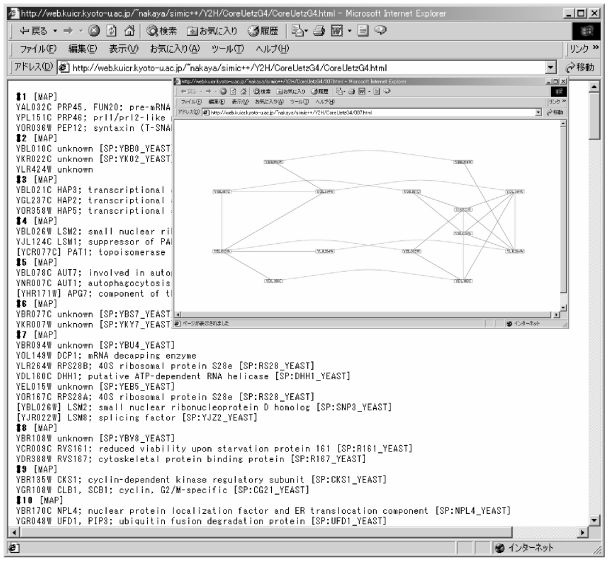
これは、同一のタンパク質ペア間にそれぞれのデータで異なった相互作用関係が存在している可能性を示している。下図に示すように、両データで共通して相互関係しているタンパク質ペア(A)に加えて、(B)及び(C)のような片方のデータでは直接相互関係しているが、他方のデータでは、第3のタンパク質（黒色箱）を介在して間接的に相互作用している例が見つかっている。



この評価方法をタンパク質間の多対多の関係の評価に発展させた。例えば、下図(A)の5つのタンパク間の相互関係は両実験で完全に再現していた。また、同図(B)に示す4つのタンパク（白色箱）は介在タンパク（黒色箱）を介して間接的に相互関係を保存している。



このような部分ネットワークを約150例抽出した。この結果に基づいて両実験の再現性の評価を行った。また、得られた部分ネットワーク群のデータベース化を行って、前述の可視化法を用いて表示できるようにした（下図）。



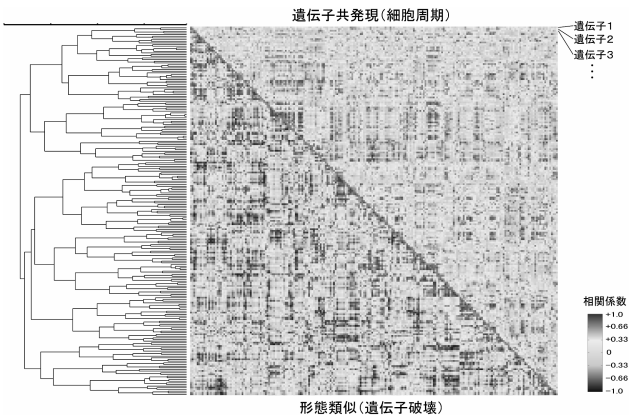
2004年度：　複数の遺伝子データの統合を行うためのソフトウェアを作成し、実データの解析に使用できるようにした。既に述べた通り、細胞形態のデータや遺伝子発現データなどの遺伝子情報はWWW上から取得できるようになっている。しかしながら、いざ自らのデータの解析に用いようとしても難しいことが多い。遺伝子情報それぞれが異なった構造（ネットワーク構造をはじめ、木構造、時系列的構造、階層構造など）を成しており、複数の遺伝子データの統合（比較・重ね合わせ・相互補完など）が生物学的にも情報科学的にも本質的に難しいことが理由のひとつである（そしてそのためのツールが整っていないのが現状である）。

ここでは、従来から使われているクラスタリングやネットワーク構造の特徴抽出などの手法と併せて、遺伝子間の関係の解析に線形配置アルゴリズムを適用することを試みた。

この目的のために、まず階層型のクラスタリングツールを作成した。従来のツールとの違いは、クラスタリングの結果得られる樹形図の形状を最適化する機能を有する点である。この最適化は、クラスタ間の相関を示す行列に対して定義されたエネルギー関数が最小化するように行われる。未知データを既知の情報に結び付けて機能解明の手がかりとすることは基本的な手法であるが、この最適化されたクラスタリングによって、より多くの情報が得られるようになった。

一般的に、より強く関係し合った物（例えば遺伝子）同士がなるべく近くなるように物（遺伝子）の順番を決定する問題は、線形配置（linear arrangement）と呼ばれ、問題サイズの増加時に効率のよい解法が存在しない難しい問題（NP完全）であることが知られている。一方、関係し合ったもの同士を同一のカテゴリに分類する手法は上述の通りクラスタリングと呼ばれる。クラスタリングの結果を一列に並べることによって、線形配置の近似的な解を得ることができる。しかしながら、クラスタ間の全ての順序関係が規定されていないために、近似解の改善を行う余地がある。階層的な手法の場合、樹形図の枝の位置関係は再帰的に反転しても良いことを利用して、近似解の改善を行うことができる。また、ある深さ以上の枝の反転を行っても解が改善しないことが判明した場合には、反転処理の枝刈を行うことができるので計算の高速化が行える。本研究では、この問題を分枝再配置（tree rearrangement）問題と呼ぶこととした。分枝再配置の結果が改善するように複数のデータの重ね合わせを行うことにより、遺伝子データの統合を可能とし、また、分枝再配置処理の結果に基いて、遺伝子間の関係を可視化する手法の開発も行った。

下図は約200の酵母の遺伝子の破壊実験による細胞形態の実験データを画像処理によって数値化したデータと、WWW上から取得した細胞周期に関する遺伝子の発現プロファイルの類似情報（Spellman et al., Mol. Biol. Cell 1998 & Cho et al., Mol. Cell 1998）の統合を試みたものである。



この図では、細胞形態への影響が類似し、かつ、細胞周期中で発現パタンが類似している遺伝子群を階層型クラスタリング（左側の樹形図）を用いて求めている（樹形図の葉が遺伝子に相当）。各クラスタに含まれる遺伝子間の関係は図右側の隣接行列で表示されている（遺伝子は縦軸と横軸に沿って同じ順序に整列）。1つの行列要素が1つの遺伝子ペアに相当（色の濃淡が相関の高低を表す）。左下三角領域は形態類似データ、右上三角領域が遺伝子の発現類似データを表示している。対角線付近の色の濃い領域が互いに相関し合った遺伝子クラスタを表している。対角線から離れた領域はクラスタ間の関係を示している。対角線を挟んで2つのデータでのクラスタ形成のパタンを比較することにより、細胞機能に関する知見を得ることを行う。

この例ではネットワーク構造と木構造をもつ2つの遺伝子データを、隣接行列の配色パタンのブロック化した領域を分離させて遺伝子間の関係をより良く表すように統合している。上述の通り、この問題は線形配置（linear arrangement）に帰着され、効率よく解けないことが知られている。生物学的な背景知識を用いたヒューリスティクスによる近似的な解法など、さまざまな方向への展開を行うことが可能である。

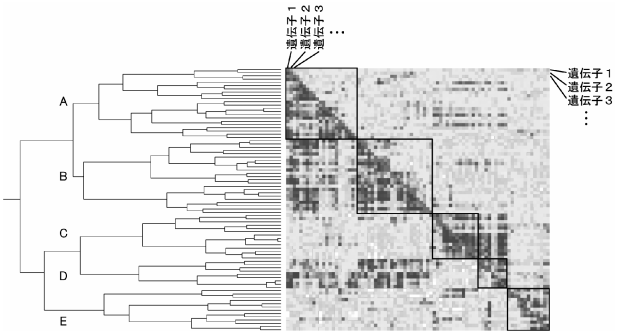
上記では、分枝再配置法の遺伝子データの統合への応用について説明したが、それと並行して結果のプレゼンテーション方法についても開発を進めた。前図では、階層型クラスタリングの樹形図と隣接行列表示されたネットワークの関係を示したが、この表示方法は以下の情報を表示しており、遺伝子データの統合結果の表示に有効である。

重属性クラスタ：　2つのネットワーク（属性）で共通してクラスタ化している遺伝子の集合の情報を表示。

クリーク情報：　ネットワーク中でクリーク様構造を成している遺伝子の情報（対角線上下の直角三角形の濃色領域）を表示。画像パタンの認識によるクリークの抽出および表示。

クラスタ間の関係：　隣接行列の対角線から離れた領域はクラスタ間の情報を表示している（通常のクラスタリングのみではこの情報は欠落）。

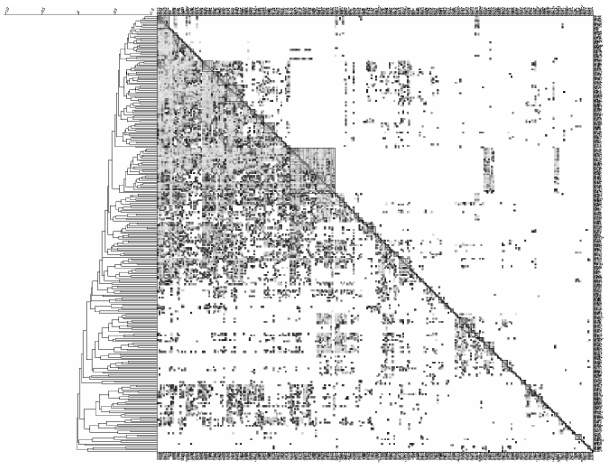
例を用いながら説明する。下図は既に示した酵母の遺伝子破壊による細胞形態への影響を調べた実験の結果と（対角線左下）、細胞周期に関する遺伝子の発現パタンを調べた結果（対角線右上）を統合した図の左上部を拡大したものである。



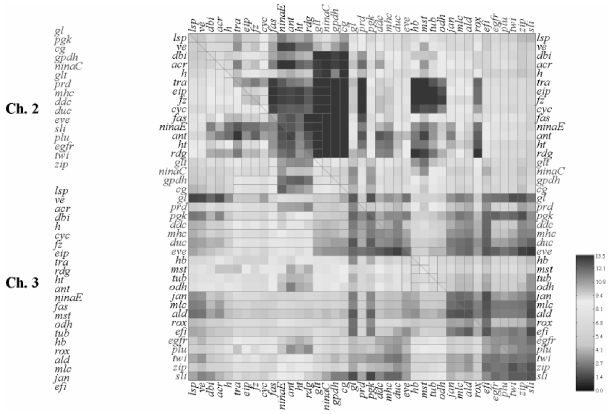
前述の通り、両データはネットワーク構造をしており、その情報はこのような隣接行列として表すことができる。隣接行列の縦横両軸に沿って遺伝子は同じ順番で整列されている。各行列要素は遺伝子のペアに相当して色で相関の度合いを表している（濃いほど相関が高い）。軸に沿った遺伝子の順番を「適切」に決めると、ここで示しているような特徴的な行列の配色パタンが対角線付近に現れる。これは、2つの属性に関して相関しているクラスタを示している。例えば、A〜Eの5つの枝に注目すると、これらの枝以下に該当するクラスタ内の遺伝子の関係は行列側の黒枠で示した対角線付近のブロックに示されている。黒枠内の領域の相関係数が高いということは、該当する遺伝子群が2つのネットワークで共通してクリーク状の構造となっていることを示している。また、黒枠の外側の領域は、クラスタ間の関係を示している。

正方形の対角線の上下それぞれに2つのネットワークの隣接行列を表示した単純なものであるが、線形配置法や分枝再配置法と併用することにより、遺伝子間に潜在する構造の直観的なプレゼンテーションが可能である。

以下では、手法の適用例を挙げる。最初の例は、2つの酵母トランスクリプトームの統合に関するものである。次図は、158個の遺伝子に関して、細胞周期に関する2つの発現プロファイルデータ（Cho他1998とSpellman他1998）の比較を行っている。行列要素は遺伝子ペアの発現プロファイルの相関係数を示している（黒いほど正の相関が高い）。対角線に関して、左下領域にChoのデータ、右上領域にSpellmanのデータを表示。遺伝子の並び順は、階層型クラスタリングの結果に基づいて決定し（左端の樹形図参照）、線形配置問題を解くことによって、樹形図の枝の上下を入れ替えて、行列の着色パタンを最適化する。両データで共通して発現プロファイルが類似している遺伝子群は対角線付近にブロック構造を生成する。この結果を見ると、両データで共通して相関係数が高い遺伝子ペアもあるが、全ての遺伝子間で必ずしも共通しているわけではないことが見てとれる。



2つめの例は、ショウジョウバエの突起サイズのQTL解析に関するものである。次図は、ショウジョウバエの「突起」の大きさ（量的形質）に関する分離データ（Zeng et al., 2000）に示された各個体の突起サイズとマーカー座の遺伝型のデータの関係から、マーカーのペアの突起サイズに対する有意さを算出して、マーカー間のネットワークを構成したものである。このデータは複数の戻し交配を行ったデータを含んでいて、図の対角線左下領域は、D.mauritiana（突起小）へのF2バッククロス集団、右上領域は D.simulans（突起大）へのF2バッククロス集団でのマーカーペアの突起サイズへの有意さを示している。QTL解析を行う際には、染色体に沿って一次元的に、量的形質への有意さの評価を行うことが多いが、このように染色体位置（ここではマーカー座）のネットワークを構成することによって、組合せの効果を評価することができる。これによって、単独では有意ではないが、複数組み合わさって初めて有意となるようなマーカー間の効果（エピスタシス）を抽出することができる。また、複数のネットワークを比較・統合することによってより多くの情報を得ることができる。

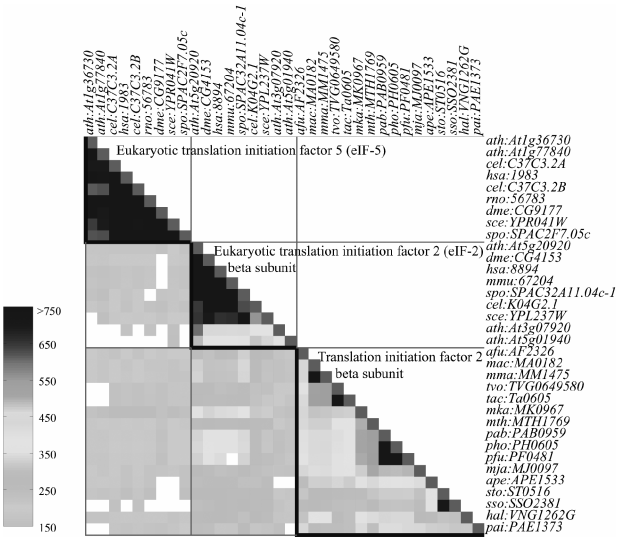


以上の例も示すように、複数の「物」の関係（ここでは遺伝子やタンパク質間の関係であるが）には多くの情報が含まれていることが分かる。

以上の例も示すように、複数の「物」の関係（ここでは遺伝子やタンパク質間の関係であるが）には多くの情報が含まれていることが分かる。

〈国内外での成果の位置づけ〉

線形配置問題を解くアルゴリズムは、階層的クラスタリングの結果として得られたデンドログラムの枝の入れ替えを行って、マトリックスパタンの最適化を行う。これは、入れ替えを行う枝をどのようなパタンで選んだら最適化できるか、という問題（一種のアソシエーション・スタディー）を解いていると言い換えることができる。この枠組みは、その後、植物の網羅的なQTL解析に用いている。また、作成した階層的クラスタリングは、アレイデータ等の解析に用いている他、データベース内に蓄積されている遺伝子のアミノ酸配列間の網羅的なホモロジーに基づくネットワークに適用されて、自動的なパラログ・オーソログ情報の作成に用いられている。次図はごく小さな例であるが、開発した手法を用いて抽出したオーソログを表示したものである。対角線部分のブロック部分がオーソログに該当して、互いに配列の類似性が高くなっている。



〈達成できなかったこと、予想外の困難、その理由〉

期間を通じて、アルゴリズムやプログラムの開発は比較的順調に進んだが、実データへの応用という点から言うと、もう少し時間が必要であった。研究期間の終了後には、ここで作成した手法をベースとして共同研究を開始しているので、タイミングの問題であったと思われる。また、当初に入手を予定していたデータが使えなかったなど、解析の下流段階であるが故の悩みもあったかも知れない。

〈今後の課題〉

ある程度の分量のプログラムは作成できたので、実際のデータへの適用を行うべきであると思われる。また、自前のデータを持たない頼りなさを改めて痛感したのであるが、情報系の立場では、自ら実験を行ってデータを生成することは現実的ではないので、「信頼できるデータ」から二次的データを生成して蓄積していく必要がある。いずれにしても、道具は作ったので如何に使っていくかという部分が課題である。

〈研究期間の全成果公表リスト〉

1）論文／プロシーディング

1. 0202281508
Nakaya, A., Goto, S., and Kanehisa, M., Extraction of correlated gene clusters by multiple graph comparison, Genome Informatics, 12, 44-53 (2001).