論文の内容の要旨

論文題目： 遺伝子クラスタ間の関係の可視化に関する研究

氏名： 加納 真

ゲノムサイエンスの分野では，whole genome shotgun 法などによる塩基配列の解読や，マイクロアレイ技術による発現量やゲノムコピー数の測定など，全遺伝子の様々な特徴量が高速測定可能となった．特に，マイクロアレイ技術による発現量測定データは，器官（胃，肺，肝臓等），病気の種類，薬物などの刺激投与からの経過時間，など様々な条件下で測定されたデータが急速に蓄積している．

様々な条件下での遺伝子の特徴量が測定可能となったことで，個々の条件下での遺伝子間の関係に関する知見が取得されてきた．遺伝子間の関係の最も代表的な表現方法は，特定の条件下で類似した特徴を示した遺伝子の集合へのラベル付け，即ち「遺伝子クラスタ」である．“遺伝子配列上流に共通な部分配列を有する遺伝子の集合”，“特定の癌で正常と比べて発現量が増加している遺伝子の集合”，“特定の薬物刺激に対して類似した時系列発現パターンを示した遺伝子の集合”，など様々な関係が，遺伝子クラスタとして表現され，様々な公開データベースで多数の遺伝子クラスタが登録・公開されている．また，各研究機関では，それぞれが独自に行った実験データに基づき，多数の遺伝子クラスタを個別に保有している．このように，測定データの蓄積に伴い，遺伝子クラスタという形態で，フラグメントした知識が蓄積している．このため，遺伝子クラスタ間の関係の分析は，フラグメントした知識を横断的に統合し，新たな生物学的知見の獲得，重要な遺伝子の絞り込み，解析結果の妥当性の検証を可能とすると期待され，非常に有益である．

遺伝子クラスタ間の関係の基本的な分析方法は，二つの遺伝子クラスタ間の重複遺伝子数の統計的評価である．異なる観点で作成された二つの遺伝子クラスタ間の重複遺伝子数が統計的に有意に大きければ，クラスタ間の重複は偶然ではなく，何らかの理由が潜んでいることが示唆される．例えば，ある条件 α 下で類似した発現パターンを示した遺伝子クラスタ A と，転写因子 X の認識配列を配列上流に有する遺伝子クラスタ B との重複遺伝子数が，統計的に有意に大きければ，遺伝子クラスタ A の遺伝子群は，条件 α 下では転写因子 X の制御を受けていることが示唆される．

しかし，通常，ある観点で遺伝子クラスタを作成する場合，同時に多数の遺伝子クラスタが作成されるため，比較するクラスタの組み合わせ数が膨大となり，関係の全体像の把握が容易ではない．例えば，薬物刺激 A に対する遺伝子発現の時系列応答パターンに基づき，遺伝子をクラスタリングした場合，様々な発現パターンの遺伝子クラスタが多数生成される（以下，クラスタセットと呼ぶ）．同様に，別の薬物刺激 B に対する遺伝子発現の時系列応答パターンからも，多数の遺伝子クラスタが生成される．薬物刺激 A と薬物刺激 B に対する遺伝子レベルでの振る舞いの違いを分析する上では，薬物刺激 A で形成された

クラスタが，薬物刺激 B のクラスタセットの中でも保存されるのか，あるいはどのように分割されるのか，その全体像を把握することが重要であるが，組み合わせ数が膨大になると容易ではない（**「クラスタセット間の重複・分割関係の全体像把握」**）．

また，多くの遺伝子クラスタが，その集合を決定するに際し，閾値に依存するため，閾値を変化させた際の影響の把握が必要となる．例えば，薬物刺激 A に対して，類似した時系列遺伝子発現パターンを示した遺伝子クラスタを取り出す場合，発現パターンの類似度に閾値を設定する必要がある．この際，恣意的な閾値では，マイノリティな発現パターンの遺伝子クラスタが，マジョリティの発現パターンの遺伝子クラスタの中に埋もれてしまい，重要な遺伝子クラスタを見落としてしまうリスクがある．また，閾値によりクラスタを構成する遺伝子が変化するため，閾値の変化の影響を把握しないと，遺伝子クラスタ間の重複遺伝子数を適切に評価できない．（**「閾値変化による影響の把握」**）．

さらに，クラスタ間の重複遺伝子数だけではなく，クラスタの背後にある特徴間の関係にも着目する必要がある．例えば，薬物刺激 A の時系列遺伝子発現パターンのクラスタ C_A と，薬物刺激 B の時系列遺伝子発現パターンのクラスタ C_B とを比較する場合，クラスタ間の重複遺伝子数だけではなく，クラスタの平均発現パターンの類似度が重要な評価指標となる．刺激 A と刺激 B がもたらすマクロな現象の違いを分析する場合には，発現パターンが両者の間でまとまって変化している（ex. クラスタ C_A では時間と共に発現が上昇，クラスタ C_B では時間と共に発現が下降）遺伝子群を見つけることが重要となるためである．クラスタ間の重複遺伝子数だけではなく，クラスタの背後にある特徴間の関係を同時評価しないと，重要遺伝子を適切に絞り込むことが難しい場合が多い．（**「クラスタ間の関係の多元的評価」**）

本論文では，遺伝子クラスタ間の関係の分析における上記 3 つの課題－「クラスタセット間の重複・分割関係の全体像把握」，「閾値変化による影響の把握」，「クラスタ間の関係の多元的評価」－を解決するための可視化手法（Cluster Overlap Distribution Map: CODM）を提案した．

CODM は，異なる観点で作成された二種類のクラスタセットを，それぞれ X 軸，Y 軸に配置し，X 軸上のクラスタ X_i と，Y 軸上のクラスタ Y_j の全ての組み合わせに関し，クラスタ間の関係を色付 3D ヒストグラムで可視化する．クラスタ間の重複遺伝子数の統計評価値を高さで，クラスタの背後の特徴間の関係を色で提示することで，「クラスタ間の関係を多元的に評価・可視化」し，ユーザが意味のあるクラスタの組み合わせを，効率的に探索することを支援する．また，X 軸上のクラスタ X_i と Y 軸上の全クラスタとの関係が Y 軸に上に，Y 軸上のクラスタ Y_j と X 軸上の全クラスタとの関係が X 軸に沿って提示されるため，「クラスタセット間の重複・分割関係の全体像把握」が可能となる．さらに，通常は取り扱われない，他に包含されるサブクラスタも可視化対象としており，「閾値変化の影響の把握」を支援する．

本論文では，まず，小細胞肺癌群に共通した染色体上発現異常領域の可視化と，小細胞肺癌検体の発染色体上発現異常領域の可視化に CODM を適用した．その結果，小細胞肺

癌における既知の染色体異常領域を検知することができ，可視化結果の信頼性が裏付けられた．加えて，新規の発現異常領域の候補の全体像も可視化できた．また，単純に各遺伝子の発現量の増減を染色体上の遺伝子の位置にマップ表示する手法では，ノイズに埋もれてしまい情報を読み取ることが難しいが，CODMは，閾値の変化による影響の提示と，各発現異常領域の有意性の統計的の評価により，ノイズに頑強な手法であることを示した．

次に，虚血耐性を持つラットと，虚血耐性を持たない通常ラットとの，虚血処理に対する時系列遺伝子発現応答の比較に対してCODMを適用した．解析の目的は，両群間で，虚血処理に対する時系列遺伝子発現パターンがまとまって変化している遺伝子集合の絞込みであった．クラスタ間の重複遺伝子数の統計的評価と，クラスタ間の発現パターンの相関の二軸から，クラスタ間の関係を評価する必要があり，どちらか一方のみの評価では絞込みが容易ではない．CODMでは，ブロックの高さと色を用いて両評価値を一度に可視化することにより，虚血耐性に関わる遺伝子を見落としリスク少なく効率的に絞り込むことができた．

さらに，癌のサブタイプ分類に固有なパスウェイ変異のトレースバック解析に対してCODMを適用した．発癌や癌の分化においては，染色体異常が起き，その領域の遺伝子の発現が変化し，その変化がタンパク質の制御ネットワークを伝わり，転写因子(Transcription Factor)の活性化／非活性化を引き起こし，転写因子の制御下にある多数の遺伝子の発現プロファイルに影響を与える，パスウェイ変異が発生している．このパスウェイ変異のトレースバック解析を，ゲノムコピー数データ，発現データ，既知のタンパク質制御ネットワークデータなどの統合によって実施，報告したのは筆者が最初である．遺伝子発現クラスタと転写因子の認識配列持つクラスタとの重複遺伝子数の評価で，発現パターンと相関して活性／不活性状態にある転写因子を検出した．さらに，その転写因子の活性・不活性のトリガーに遡るため，遺伝子クラスタの発現パターンと，転写因子の（タンパク質の制御ネットワーク）上流で染色体異常を有する遺伝子の発現パターンとの相関を可視化した．その結果，癌のサブ分類に関係する既知のパスウェイ変異を検出することができ，解析の信頼性が確認された．また，新規のパスウェイ変異の候補も多数検出することができた．

以上のように，本論文の提案可視化手法（CODM）は，様々なタイプの遺伝子クラスタセットの比較に適用可能であり，有用性と解析結果の信頼性が確認された．今後，ゲノムサイエンスのエリアでは，マイクロアレイ技術による発現量やゲノムコピー数の測定など，全遺伝子の様々な特徴量データの蓄積と，そのデータに基づく遺伝子クラスタとしてのフラグメントした知見の蓄積はますます加速することが予測される．本論文の提案可視化手法は，フラグメントした知識を横断的に統合し，新たな生物学的知見の獲得，重要な遺伝子の絞り込むための重要な役割を果たすことができると期待される．