【平成24年4月-12月授与分】博士学位論文内容の要旨及び審査の結果の要旨

氏名・(本籍・国籍)	いわ したゆう じ 岩 下 雄 二 (福岡県)
学 位 の 種 類	博士 (システム生命科学)
学 位 記 番 号	シ生博甲第８９号
学位授与の日付	平成24年4月30日
学位授与の要件	学位規則第４条第１項該当　システム生命科学府　システム生命科学専攻
学 位 論 文 題 目	Genome-wide Repression of NF-κB Target Genes by Transcription Factor MIBP1 and its Modulation by *O*-GlcNAc Transferase (転写因子MIBP１による、NF-κBターゲット遺伝子のゲノムワイドな抑制と、その*O*-GlcNAc転移酵素による調節)
論 文 調 査 委 員	(主　査) 教　授　服 巻 保 幸 (副　査) 教　授　藤 木 幸 夫　　教　授　木 村　　誠 准教授　山 本　　健

論　文　内　容　の　要　旨

転写因子 c-*MYC* intron binding protein 1 (MIBP1) は、c-*MYC* イントロン１をはじめとするゲノム中の様々な制御領域に結合する。この因子は、胎児脳の、分裂を終えた神経細胞で発現が高いことが知られ、また、神経系、免疫系細胞等の分化過程に関与していると考えられている。本研究では、MIBP1 が転写ターゲットとする遺伝子は何か、及び MIBP1 がいかなるタンパク質と相互作用（結合）するか、について網羅的に調べた。すなわち、MIBP1 を恒常的に過剰発現させた HEK293 細胞では、MYC、NF-κB 及び TGF-β パスウェイの下流の遺伝子群の発現が有意に低下することを、マイクロアレイハイブリダイゼーション実験と、その結果の Gene Set Enrichment Analysis を用いたトランスクリプトーム解析によって明らかにした。これら発現の低下した遺伝子群のプロモーター領域における転写因子結合サイトを、oPOSSUM 及び Pscan ソフトウェアを用いた情報学的手法で解析した結果、NF-κB 結合サイトが主要な転写因子結合サイトであることが分かった。また、MIBP1 の発現が高い HT1080 細胞を用いて、内在性 MIBP1 の発現を RNAi 導入によって抑制したところ、NF-κB パスウェイに含まれる遺伝子群の発現上昇が起こることを見出した。さらに、MIBP1 が NF-κB サイトに結合することを、電気泳動移動度シフト (EMSA) 実験によって確認した。これらの結果は、MIBP1 が NF-κB パスウェイの抑制因子であるという考えを支持している。また、免疫沈降産物の質量分析によるプロテオーム解析から、O 結合型 β-*N*-アセチルグルコサミン (*O*-GlcNAc) 転移酵素 (OGT) が、MIBP1 の結合タンパク質であることを決定した。種々の欠失変異体による解析から、MIBP1 中の 154 アミノ酸にわたる領域が、OGT との結合と *O*-GlcNAc 化に必要であることを明らかにした。一方、ルシフェラーゼを用いたレポーター解析から、NF-κB 応答性の発現が MIBP1 によって抑制されること、上記 154 アミノ酸領域を欠く変異体型 MIBP1 の発現によって、さらに強い NF-κB 応答遺伝子の発現抑制が見られることを発見した。これらの結果から、MIBP1 が発現することによる主要な効果が NF-κB パスウェイの抑制であり、この MIBP1 による抑制は *O*-GlcNAc シグナル伝達によって弱められていると結論した。

論　文　審　査　の　結　果　の　要　旨

転写因子 c-*MYC* intron binding protein 1 (MIBP1) は、c-*MYC* イントロン１をはじめとするゲノム中の様々な制御領域に結合する。この因子は、胎児脳における分裂を終えた神経細胞で発現が高いことが知られ、また、神経系、免疫系細胞等の分化過程に関与していると考えられている。本研究では、MIBP1 の転写制御における機能を明らかにするために、そのターゲットとなる遺伝子群、及び MIBP1 の結合タンパク質の同定を目指した網羅的解析を行い、以下の結果を得た。(1) MIBP1 を恒常的に過剰発現させた HEK293 細胞についてマイクロアレイによるトランスクリプトーム解析を行い、Gene Set Enrichment Analysis によって、MYC、NF-κB 及び TGF-β パスウェイの下流の遺伝子群の発現が有意に低下することを明らかにした。(2) 発現が低下した遺伝子群のプロモーター領域における転写因子結合サイトを、oPOSSUM 及び Pscan ソフトウェアを用いた情報学的手法で解析した結果、NF-κB 結合サイトが主要な転写因子結合サイトであることが分かった。(3) MIBP1 の発現

が高いHT1080細胞を用いて、内在性MIBP1の発現をRNAi導入によって抑制したところ、NF-κBパスウェイに含まれる遺伝子群の発現上昇が起こることを見出した。(4) MIBP1がNF-κBサイトに結合することを、電気泳動移動度シフト（EMSA）実験によって確認した。(5) 免疫沈降産物の質量分析によるプロテオーム解析から、O結合型β-*N*-アセチルグルコサミン（*O*-GlcNAc）転移酵素（OGT）が、MIBP1の結合タンパク質であることを決定した。(6) 種々の欠失変異体による解析から、MIBP1中の154アミノ酸にわたる領域が、OGTとの結合と*O*-GlcNAc化に必要であることを明らかにした。(7) ルシフェラーゼを用いたレポーター解析から、NF-κB応答性の発現がMIBP1によって抑制されること、上記154アミノ酸領域を欠く変異体型MIBP1の発現によって、さらに強いNF-κB応答遺伝子の発現抑制が見られることを見出した。以上の結果から、MIBP1の主要な効果がNF-κBパスウェイの抑制であり、このMIBP1による抑制は*O*-GlcNAcシグナル伝達によって弱められていると結論した。

本研究はこの分野において価値ある業績として認められる。よって、本研究者は博士（システム生命科学）の学位を受ける資格があるものと認める。

氏名・(本籍・国籍)	まつ もと しゅんすけ 松 本 俊 介（愛知県）
学位の種類	博士（システム生命科学）
学位記番号	シ生博甲第９３号
学位授与の日付	平成24年9月24日
学位授与の要件	学位規則第４条第１項該当　システム生命科学府　システム生命科学専攻
学位論文題目	A Study on Structure and Function of Oligosaccharyltransferase from a hyperthermophilic archaeon, *Archaeoglobus fulgidus*（超好熱性古細菌*Archaeoglobus fulgidus*由来オリゴ糖転移酵素の構造と機能に関する研究）
論文調査委員	（主　査）教　授　神田大輔 （副　査）教　授　須山幹太　　教　授　石野良純 准教授　稲葉謙次

論　文　内　容　の　要　旨

タンパク質のアスパラギン結合型糖鎖による修飾は、真核生物だけではなく、古細菌や一部の真正細菌にも存在する。オリゴ糖転移酵素（OST）は、アクセプタータンパク質に存在するAsn-X-Ser/Thr（Xはプロリン以外のアミノ酸）のコンセンサス配列を認識し、アスパラギン残基にオリゴ糖鎖を丸ごと転移する。酵母やヒトなどの高等真核生物のOSTは、膜タンパク質複合体をER膜に形成するが、古細菌や真正細菌のOSTは触媒サブユニット単独で細胞膜に存在する。OSTの触媒サブユニットは、真核生物ではSTT3、古細菌ではAglB、そして真正細菌ではPglBとそれぞれ異なる名前で呼ばれている。STT3/AglB/PglBは、N末端側の１３回程度の膜貫通ドメインとC末端側の可溶性ドメインからなる共通のドメイン構造を有している。STT3/AglB/PglBのアミノ酸レベルでの相同性は非常に低いが、可溶性ドメインには高度に保存されたWWDYGモチーフが存在する。先行研究において、古細菌*Pyrococcus furiosus*のAglB（PfAglB）と真正細菌*Campylobacter jejuni*のPglB（CjPglB）の可溶性ドメインの構造が決定されている。その結果、それぞれのWWDYGモチーフの近傍に位置する保存モチーフが同定された。それらはPfAglBではDKモチーフ、CjPglBではMIモチーフと呼ばれている。さらに、これら２つの可溶性ドメインの構造比較をもとに、配列アライメントを解釈するとSTT3/AglB/PglBは、DK、MIそしてDMモチーフを持つ３つのタイプに分類されることが予想された。本研究では、第三のタイプのDMモチーフを持つ古細菌*Archaeoglobus fulgidus*由来のOSTの可溶性ドメインのX線結晶構造解析を行った。

古細菌 *A. fulgidus* のゲノムは 3 つの AglB パラログ遺伝子をコードしている。本研究では、ゲノムデータベースに登録されているすべての STT3/AglB/PglB の中で最小の触媒サブユニットである AfAglB-S1 の可溶性ドメインの構造決定を行った。その結果、AfAglB-S1 の構造を 1.75Å という高分解能で構造決定することができた。AfAglB-S1 の構造は、これまでに構造決定された PfAglB や CjPglB の構造と比べて小さく、コンパクトな構造をとっていた。そして、AfAglB-S1 には、PfAglB や CjPglB に共通して見つかるバレル様のサブドメイン構造が存在しなかった。これら 3 つの OST アミノ酸配列レベルでの相同性が非常に低いにも関らず、WWDYG モチーフ周辺のコア構造は、よく似ていた。しかし、AfAglB-S1 には、PfAglB や CjPglB には存在しないループ構造が DM モチーフの内に挿入されていた。その結果、DK/MI モチーフ周辺の配列アライメントを修正することになり、第三の DM モチーフというよりも、DK モチーフのバリアントという解釈が適切であることが明らかとなった。

次に、AfAglB-S1 の DK モチーフの機能的重要性を調べるために、AfAglB-S1 のアミノ酸置換変異体を用いた構造活性相関解析を行った。*A. fulgidus* の菌体を培養し、その菌体から OST の基質である脂質結合型オリゴ糖（LLO）を脂質粗抽出画分として調製した。次に、大腸菌での全長 AfAglB-S1 の発現系を構築し、AfAglB-S1 組換えタンパク質を含む膜フラクションを調製した。そして蛍光ラベル標識したペプチド（Asn-X-Ser/Thr 配列を含む）の配列最適化を行うことで、AfAglB-S1 の酵素活性測定系を構築した。このアッセイ系を用いて、DK モチーフおよび挿入ループに変異を導入した AfAglB-S1 の酵素活性を測定したところ、DK モチーフの変異によって、その酵素活性が著しく低下することを確認した。

他の研究グループの報告により、真正細菌 *Campylobacter lari* 由来 PglB のペプチド基質との複合体の全長構造が 3.4Å 分解能で明らかとなった。この全長構造から、WWDYG モチーフの Trp-Trp-Asp と MI モチーフの Ile が、コンセンサス配列の Ser もしくは Thr を認識するポケットを作っていることが明らかとなった。

本研究に決定した AfAglB-S1 の結晶構造とこれまでに報告された OST の可溶性ドメインの結晶構造を含めて考察すると、STT3/AglB/PglB の可溶性ドメインに存在する Ser/Thr ポケットは、全ての真核生物と多くの古細菌がもつ DK モチーフタイプと真正細菌とその他の古細菌がもつ MI モチーフタイプの 2 つのタイプに分類することができるという結論に至った。これは今後の STT3/AglB/PglB の構造活性相関の研究を進めていく上で、重要な知見であると考えている。

論 文 審 査 の 結 果 の 要 旨

タンパク質のアスパラギン結合型糖鎖による修飾は、真核生物だけではなく、古細菌や一部の真正細菌にも存在する。オリゴ糖転移酵素（OST）は、アクセプタータンパク質に存在するコンセンサス配列である Asn-X-Ser/Thr（X はプロリン以外のアミノ酸）を認識し、アスパラギン残基にオリゴ糖鎖を丸ごと転移する。先行研究において、古細菌 *Pyrococcus furiosus* の AglB（PfAglB）と真正細菌 *Campylobacter jejuni* の PglB（CjPglB）の可溶性ドメインの構造が決定されている。その結果、それぞれの WWDYG モチーフの近傍に位置する保存モチーフが同定された。それらは PfAglB では DK モチーフ、CjPglB では MI モチーフと呼ばれている。さらに、これら 2 つの可溶性ドメインの構造比較をもとに、配列アライメントを解釈すると STT3/AglB/PglB は、DK、MI そして DM モチーフを持つ 3 つのタイプに分類されることが予想された。本研究では、第三のタイプの DM モチーフを持つ古細菌 *Archaeoglobus fulgidus* 由来の OST の可溶性ドメインの X 線結晶構造解析を行った。

古細菌 *A. fulgidus* のゲノムは 3 つの AglB パラログ遺伝子をコードしている。本研究では、ゲノ

ムデータベースに登録されているすべてのSTT3/AglB/PglBの中で最小の触媒サブユニットであるAfAglB-S1の可溶性ドメインの構造決定を行った。その結果、AfAglB-S1の構造を1.75Åという高分解能で構造決定することができた。AfAglB-S1の構造は、これまでに構造決定されたPfAglBやCjPglBの構造と比べて小さく、コンパクトな構造をとっていた。そして、AfAglB-S1には、PfAglBやCjPglBに共通して見つかるバレル様のサブドメイン構造が存在しなかった。これら3つのOSTアミノ酸配列レベルでの相同性が非常に低いにも関らず、WWDYGモチーフ周辺のコア構造は、よく似ていた。しかし、AfAglB-S1には、PfAglBやCjPglBには存在しないループ構造がDMモチーフの内に挿入されていた。その結果、DK/MIモチーフ周辺の配列アライメントを修正することになり、第三のDMモチーフというよりも、DKモチーフのバリアントという解釈が適切であることが明らかとなった。

次に、AfAglB-S1のDKモチーフの機能的重要性を調べるために、AfAglB-S1のアミノ酸置換変異体を用いた構造活性相関解析を行った。その結果、DKモチーフおよび挿入ループに変異を導入したAfAglB-S1の酵素活性を測定したところ、DKモチーフの変異によって、その酵素活性が著しく低下することを確認した。

今回決定したAfAglB-S1の結晶構造とこれまでに報告されたSTT3/AglB/PglBの可溶性ドメインの結晶構造を含めて考察すると、STT3/AglB/PglBの可溶性ドメインに存在するコンセンサス配列のSerとThrを認識するポケットは、全ての真核生物と多くの古細菌がもつDKモチーフタイプと真正細菌とその他の古細菌がもつMIモチーフタイプの2つのタイプに分類することができるという結論に至った。

以上の結果は、糖質科学および構造生物学の分野で価値ある業績と認められる。よって、本研究者は博士（システム生命科学）の学位を受ける資格があるものと認める。

氏名・(本籍・国籍)	たむらみほ 田村美帆 (山口県)
学位の種類	博士（理学）
学位記番号	シ生博甲第94号
学位授与の日付	平成24年12月31日
学位授与の要件	学位規則第4条第1項該当　システム生命科学府　システム生命科学専攻
学位論文題目	Analyses of BAC sequences from a conifer, *Cryptomeria japonica*（針葉樹スギより得られたBACクローンの塩基配列解析）
論文調査委員	(主査) 教授　舘田英典 (副査) 准教授　Alfred Edward Szmidt　准教授　渡辺敦史 講師　楠見淳子

論文内容の要旨

針葉樹は裸子植物最大のグループに属する樹木であり、多くの被子植物に比べて巨大なゲノムサイズを示し、長い世代時間を持つ。加えて、針葉樹は長い進化の歴史においてほとんどゲノムサイズの変化がないという特徴から、分子進化において興味深い研究対象の一つである。しかしながら、裸子植物のゲノム構造の研究は被子植物に比べて進んでいない。またその中でも針葉樹の研究の多くはマツ科に限られている。スギは針葉樹のもう一つの大きなグループである広義ヒノキ科に属し、日本の重要林業樹種でもあるため、スギのゲノム構造を明らかにすることは重要である。

本研究では針葉樹のモデル植物であるスギのランダムに選んだ8つのBAC配列を使用してゲノ

ム解析を行なった。また、他の既知の被子植物のシロイヌナズナ、イネ、ポプラのゲノム配列に加え、同じ裸子植物であるマツの BAC 配列についても同様の解析を行なった。更にこれらの解析結果とスギの解析結果を比較し、以下に挙げるスギ及び針葉樹の塩基配列の特質を明らかにした。

１つ目の特徴として、スギにおいて様々な繰り返し配列が検出された。また針葉樹の先行研究（マツ、ヌマスギ）と同様に多くの分化した繰り返し配列がみつかった。このことより、針葉樹で多くの Transposable element(TE)は古くに拡大した事が示唆された。しかし、最近まで活性を維持した幾つかの TE もスギで確認された。またこれらの TE のいくつかは、今まで見つかっている TE と相同性がなく、未知のタイプに TE である可能性が考えられた。

次に巨大なイントロンを持つタンパクをコードする遺伝子が、スギの BAC 配列中に存在する事が分かった。これまで被子植物では機能を持った遺伝子においてこのようなサイズのイントロンは報告されておらず、植物は小さなイントロンを持つと考えられてきた。しかし、このスギの遺伝子配列は近縁種ヌマスギの塩基配列との比較より、このような巨大なイントロンを持ちながらも少なくとも最近まで機能を維持している事が確認された。このことより、針葉樹もしくは裸子植物ゲノムは被子植物には無い巨大なイントロンを持つ遺伝子があることわかった。

更に、被子植物の先行研究で報告されている様に、マツ、スギの針葉樹でも CPG 頻度の低下がみられた。しかしこの低下は被子植物（シロイヌナズナ、イネ、ポプラ）でみられるものよりも、強いものであった。この CPG 低下の原因にはクロマチン構造に関与する DNA のメチル化が関係しているのではないかと推測された。

本研究より、針葉樹ゲノムは被子植物のゲノムとは異なる特徴を持つ事が明らかになった。

論 文 審 査 の 結 果 の 要 旨

針葉樹は裸子植物最大のグループに属する樹木であり、多くの被子植物に比べて巨大なゲノムサイズを持ち、また長い世代時間を持つ。加えて、針葉樹は長い進化の歴史においてほとんどゲノムサイズの変化がないという特徴から、分子進化において興味深い研究対象の一つである。しかしながら、針葉樹を含め裸子植物のゲノム構造の研究は被子植物に較べて進んでいない。また数少ない針葉樹の研究も、多くはマツ科に限られている。スギ（*Cryptomeria japonica*）は針葉樹のもう一つの大きなグループである広義ヒノキ科（Cupressaceae）に属し、日本の重要林業樹種でもあるため、スギのゲノム構造を明らかにすることは重要である。

本研究では針葉樹のモデル植物であるスギのランダムに選んだ８つの BAC 配列を使用し、バイオインフォーマティックスの手法を使ってゲノム解析を行なった。また、他の被子植物、シロイヌナズナ、イネ、ポプラの既知のゲノム配列に加え、同じ針葉樹であるマツの BAC 配列についても同様の解析を行なった。これらの解析結果とスギの解析結果を比較し、以下に挙げるスギ及び針葉樹のゲノム配列の特質を明らかにした。

一つ目の特徴として、スギにおいて様々な繰り返し配列が大量に検出された。また針葉樹の先行研究（マツ、ヌマスギ）と同様に、多くの分化した繰り返し配列がみつかった。このことより、針葉樹では比較的古い時代に多くの Transposable element (TE)の増幅が起こった事が示唆された。しかし、少なくとも最近まで転移活性を維持していたと考えられる複数の TE もスギで確認された。これらの TE のいくつかはこれまでに見つかっている TE とは全く相同性がなく、未知のタイプの TE である可能性が高いと考えられた。

二つ目の特徴として、スギの BAC 配列中にサイズが７０ｋｂの巨大なイントロンを持つタンパクをコードする遺伝子が存在することが明らかになった。これまで被子植物の機能を持った遺伝子ではこのような大きなサイズのイントロンは報告されておらず、一般に植物の機能遺伝子は小さな

イントロンを持つと考えられてきた。しかし、このスギの遺伝子配列は近縁種ヌマスギの塩基配列との比較より、このような巨大なイントロンを持ちながらも少なくとも最近まで機能を維持していた事が確認された。このことより、針葉樹もしくは裸子植物ゲノムには、被子植物には無い巨大なイントロンを持つ遺伝子があることが明らかになった。

三つ目の特徴として、被子植物の先行研究で報告されている様に、針葉樹のマツ及びスギでも二塩基組 CpG の頻度低下がみられた。しかもこの低下は、被子植物（シロイヌナズナ、イネ、ポプラ）でみられるものよりも強いものであった。この CpG 頻度低下にはクロマチン構造に関与するDNA のメチル化が関係しているのではないかと推測された。

本研究より、針葉樹ゲノムはこれまでに解析された被子植物のゲノムとは様々な点で異なる特徴を持つ事が明らかになった。

これらの成果は植物のゲノム構造について重要な知見を得たものとして価値ある業績であると認める。よって、本研究者は博士（理学）の学位を受ける資格があるものと認める。

氏名・(本籍・国籍)	ドウ ビン 童 彬（中国）
学位の種類	博士（工学）
学位記番号	シ生博甲第９０号
学位授与の日付	平成24年9月24日
学位授与の要件	学位規則第４条第１項該当　システム生命科学府　システム生命科学専攻
学位論文題目	Dimensionality Reduction with Semi-supervised Learning and Transfer Learning（半教師付き学習と転移学習を用いた次元削減）
論文調査委員	（主　査）教　授　鈴　木　英之進 （副　査）教　授　内　田　誠　一　　教　授　廣　川　左千男

論　文　内　容　の　要　旨

高次元データを低次元空間へ射影する次元削減は、遺伝子発現・テキストマイニング・画像検索のような種々のアプリケーションにおけるデータ分析のもっとも重要な前処理手段のひとつと考えられている。この次元削減は、ラベル情報が利用可能かそうでないかによって、教師ありと教師なしの２つに分類される。一般的に、ラベル情報は次元削減において助けとなる。つまり、異なるラベルをもつデータが次元削減後の空間でも離れて配置されるような射影を求めるときに、ラベル情報は有益である。しかしながら、ラベル情報を得るためにはコストがかかるため、多くの実世界アプリケーションで利用可能なラベル付きデータの数は少ない。したがって、教師情報とともに数多くあるラベルなしデータを利用する半教師付き学習は、実世界での設定によく一致する。一般的に、教師情報には、ラベル情報に加え、2つのデータが同じクラスかどうかを示す対制約がある。実際の問題では、教師情報は関連タスクでは多く存在するが、対象タスクでは少ない場合もよくある。人間は関連タスクにおける教師情報の知識をうまく適用して対象タスクでの問題を解決できるという考えから、転移学習は有効であるといえる。

データマイニングや機械学習などの研究分野では、対象タスクの教師情報が少ない場合に次元削減によって性能が低下しないように、半教師付き学習と転移学習を用いた次元削減が注目を集めている。具体的には、単タスクや複数タスクにおけるラベル情報や対制約を用いて、次元削減後の空間におけるクラスタリングや分類の性能を向上させる手法に関心が寄せられている。しかしながら、対制約を用いた単

タスク問題では、同じクラスのデータが複数のグループに存在する場合、既存の次元削減手法では対制約を適切に利用していない。対制約を用いた複数タスク問題では、複数の元タスクから目的タスクへ対制約の知識を転移させる手法は、重要で難しい問題であると考えられる。ラベル情報を用いた複数タスク問題では、目的タスクでラベル情報が存在しない場合、ラベル情報の知識を転移させることはラベル情報が与えられている場合に比べて困難である。

本論文では、上述の問題に関する３つの系統的な枠組みを提案した。１つ目として、複数のクラスタに同じクラスラベルが存在する場合にも制約を利用することができる半教師付き次元削減のための新しい判別基準を考案した。２つ目として、対制約の知識転移に関し、複数タスクによって共有される共通の低次元空間上でのクラスタリングの枠組みを提案した。この低次元空間では、複数タスクの重心によって形成される構造により対制約の知識が転移される。３つ目として、ラベル情報の知識転移に関し、目的タスクのラベル情報が利用できない場合の転移学習枠組みでの次元削減を提案した。有益な知識は優先的に複数の元タスクにおける元空間と次元削減後の空間との関係に取り込まれ、ガウス過程回帰を用いてうまく目的タスクに転移される。

人工データと遺伝子、病気診断、テキストや画像のような幅広い実データでの実験により、提案枠組みの有効性を評価した。次元削減後の空間でのクラスタリング・分類の実験結果を示し、最新の手法と比べ、半教師付き学習と転移学習を用いたことにより、ほとんどの場合で提案手法が有効であることを確認した。

論文審査の結果の要旨

高次元データを低次元空間へ射影する次元削減は、バイオインフォマティクス、テキストマイニング、画像検索などの種々のアプリケーションにおけるデータ分析において、最も重要な前処理手段のひとつと考えられている。次元削減前の空間におけるデータ配置に関するラベル情報は、得るためのコストが高いため、実問題ではきわめて少ないか存在しないことが多い。このような状況に対し、ラベルありデータとラベルなしデータを併用する半教師付き学習と、関連する源タスクでの情報を対象タスクに転移する転移学習が盛んに研究されているが、頑健性や正確性に問題がある。本研究は、多様体学習などによる幾何情報の利用により、これらの問題点の軽減に取り組んだ。

本論文では、上述の問題に対して３つの系統的な枠組みを提案した。まず、事例のペアに関するラベル情報である対制約を用いる半教師付き次元削減を対象とし、次元削減後の空間において同クラスのラベルを持つ事例が形成するクラスタが複数個ある場合に取り組んだ。対制約の遵守・違背を判断する新しい判別基準を考案することにより、このような現実的な問題設定に頑健な手法を構築した。次に、転移学習でのクラスタリングにおいて、一つあるいは複数の源タスクでは対制約が利用できるが対象タスクではラベル情報がない場合に取り組んだ。源タスクと対象タスクに共通な低次元空間において、源タスクの対制約を対象タスクにクラスタ中心を介して転移する手法を考案することにより、このような現実的かつ困難な問題設定に有効な手法を構築した。最後に、転移学習での次元削減において、一つあるいは複数の源タスクではラベル情報が利用できるが対象タスクではラベル情報がない場合に取り組んだ。転移学習における次元削減問題を回帰問題に転換する方法と、その回帰問題をガウス過程回帰を用いて解く方法を考案し、このような現実的かつ困難な問題設定に有効な手法を構築した。

遺伝子解析と病気診断を含む種々の応用に関する実データ、および系統的に生成した人工データを用いた実験により、提案手法の有効性を評価した。関連する最新手法に比較して、ほとんどの場合で提案手法が優位であることを確認した。

以上、本論文は対象タスクの教師情報が少ないか存在しない場合に有効な判別基準、対制約転移法、回帰学習解法を提案してそれらの有効性を示したものであり、半教師付き学習と転移学習を用いた次元削減に貢献する価値ある業績であると認める。

よって、本申請者は博士（工学）の学位を受ける資格があるものと認められる。

氏名・(本籍・国籍)	グエン　ヒ　テアク Nguyen Huy Thach （ベトナム）
学位の種類	博士（工学）
学位記番号	シ生博甲第９１号
学位授与の日付	平成24年9月24日
学位授与の要件	学位規則第４条第１項該当　システム生命科学府　システム生命科学専攻
学位論文題目	Extended Information Distance and Manifold Learning for Cross-task Learning Problems without Proper Features（適切な属性群がないタスク間学習問題のための拡張情報距離と多様体学習）
論文調査委員	（主　査）教　授　鈴　木　英之進 （副　査）教　授　内　田　誠　一　　教　授　廣　川　左千男

論　文　内　容　の　要　旨

データマイニングにおいて、教師情報は学習タスクに有益である。しかしながら、実際のアプリケーションでは、ラベル情報を得ることは多くの場合高コストで労力が必要となる。そのため、ラベルありデータを集める労力を減らすことは望ましいと考えられる。実際の設定では、関連ドメインにはラベル情報は数多く存在するが、対象ドメインではそうではない場合は頻繁に起こりうる。このような場合には、元ドメインから知識を抽出し、目的ドメインでの学習性能を高める転移学習が有効である。さらに、複数のタスクを同時に学習することによって知識を共有しようとするマルチタスク学習もまた、学習タスクの性能を大きく向上させる。

転移学習の２つの大きな問題は、ノイズの取り扱い、そして元ドメインと目的ドメインにおけるデータ分布の差異の取り扱いである。そして、マルチタスク学習手法の主な問題は、慎重なパラメータ設定と適切な特徴抽出が必要である点である。特徴抽出の失敗や不適切なパラメータ設定により、間違ったあるいは存在しないパターンを見つけてしまう場合もある。近年、特定の特徴ベクトル空間を仮定することなく少数のパラメータのみを必要とするコルモゴロフ定理に基づいたシングルタスク学習アルゴリズムが提案されている。しかしながら、これらの手法は異なるタスク間で知識を共有するメカニズムを欠いており、直接マルチタスク学習問題に適用することは難しい。一方で、転移学習では、ほとんどの研究は適切な特徴空間では元ドメインと目的ドメインのデータ分布は類似すると暗に仮定している。しかしながら、元ドメインと目的ドメインのデータへのラベル割り当てが大きく異なる場合も多く存在する。したがって、元ドメインと目的ドメインのデータ分布の差が小さくとも、ラベル情報を十分に考慮しないかぎり、複数の元ドメインの知識はうまく目的ドメインに転移されない。

本論文では、上述の問題を解決するために特徴空間を要さずパラメータが少ないマルチタスククラスタリングとノイズに強い転移学習の２つの枠組みを提案した。マルチタスク学習の枠組みでは、効果的に異なるタスク間での知識共有を可能にする新しい辞書ベースの圧縮非類似度基準を定義した。各ドメインごとに、クラスター数という１つのパラメータを設定するのみで、特定の特徴集合

を与える必要はない。2 つの提案枠組みでは、転移学習設定でのノイズに頑健なアルゴリズムを作成し、データ分布と、複数の元ドメインと目的ドメインでのラベル割り当ての差を十分に考慮した。知識を転移させるとき局所構造が重要という事実に注目し、局所構造を保持することに細心の注意を払いながら効果的にノイズを除去する密度に基づいたクラスタリングアルゴリズムの変形を考案した。さらに、複数の元ドメインと目的ドメイン間での分布の差を減らす重心を軸としたマッピングを定義し、目的ドメインにおける分類タスクでの性能を向上させるために複数の元ドメインから有用なデータを選んだ。

生物学、言語コーパス、異種混合データ、文書や音楽データセットを用いた実験により、2 つの提案枠組みの有効性を示した。特徴空間を要さずパラメータが少ないマルチタスク学習手法は、適切な特徴抽出の難しいデータセットに対しても比較手法より 10%以上の性能向上を示した。2 つ目の提案手法では、36 の転移学習タスクに対する実験により、提案手法は従来手法に比べ、最大 15%の性能向上を示した。最新の転移学習手法との比較により、提案手法は高ノイズ環境でも最大で 20%の性能向上を示した。

論 文 審 査 の 結 果 の 要 旨

データマイニングにおいて、適切な属性群がない学習問題は頻出する。その対策として、対象タスクに関連する源タスクの情報を用いるタスク間学習問題が盛んに研究されており、中でも各タスクを同等に扱うマルチタスク学習と源タスクを補助的に用いる転移学習が特に重要視されている。適切な属性群がない学習問題は主に、文字列データなどの構造データと数値データなどの高次元データを扱うものに 2 分される。前者においては、理想的だが計算不能である情報距離を模擬する圧縮に基づく非類似度が単一タスク学習問題に数種類提案されているが、それらのタスク間学習への拡張は存在せず、関連タスクの情報が利用できない。後者は、次元削減法による多様体学習で解決するのが一般的であるが、データ分布がきわめて複雑でありノイズレベルが高い場合に有効な手法がなく、現実的な問題設定に対する解法が望まれている。

本論文では、上述の問題を解決するために拡張情報距離と多様体学習に取り組んだ。まず、文字列データのマルチタスククラスタリングのために情報距離を拡張し、その拡張が圧縮に基づく最新の非類似度に比較して情報距離のより厳密な下限であることを証明した。次に、テスト事例が訓練時に得られるトランスダクティブ学習の適切な属性群がない数値データの転移学習版のためのノイズ除去戦略を考案し、その戦略を利用した密度クラスタリングと多様体学習の統合的学習法を提案した。

生物学データ、言語コーパス、異種混合データ、文書や音楽データを用いた実験により、2 つの提案枠組みの有効性を示した。拡張情報距離を用いるマルチタスククラスタリング手法では、適切な属性群がないデータ集合に対しても比較手法より分類精度に関して 10%以上の性能向上を示した。トランスダクティブ転移学習手法では、36 の転移学習タスクに対する実験により、提案手法は従来手法に比べ分類精度に関して最大 15%の性能向上を示した。最新の転移学習手法との比較により、提案手法は高ノイズ環境でも分類精度に関して最大で 20%の性能向上を示した。

以上、本論文はマルチタスククラスタリングとトランスダクティブ転移学習手法を提案してそれらの有効性を示したものであり、適切な属性群がないタスク間学習問題について重要な知見を得たものとして価値ある業績であると認める。

よって、本申請者は博士（工学）の学位を受ける資格があるものと認める。

氏名・(本籍・国籍)	ショウ　コウ 邵　浩（中国）
学位の種類	博士（工学）
学位記番号	シ生博甲第９２号
学位授与の日付	平成24年9月24日
学位授与の要件	学位規則第４条第１項該当　システム生命科学府　システム生命科学専攻
学位論文題目	Adaptive Transfer Learning for Classification through Minimum Encoding（最小エンコードを用いた分類のための適応型転移学習）
論文調査委員	（主　査）教　授　鈴木英之進 （副　査）教　授　内田誠一　　教　授　廣川左千男

論　文　内　容　の　要　旨

今日、データマイニングの分類学習技術は、疾患の診断や生物学的分類など、種々の分野で活用されている。わずかな診断例しかない病気の診断など、医療や生物学のデータ集合においては、新たなタスクのためのラベル付きデータが多くの場合不足しており、分類モデルを正確に構築することは困難である。転移学習と能動学習はラベル付きデータの不足を補うための２つの独立した学習であり、前者は既存のモデルから有用な情報を抽出し、新たなタスクの新規モデルを構築し、後者はラベルの付いていない有用なインスタンスを抽出し、専門家にラベル付けをしてもらう。前者は類似するタスクから知識を借入しようと努めることが自然であるが、数ある転移学習の研究の中でも理論に基づく枠組でパラメータなしの手法は少ない。さらに、タスク間の異なる分布に起因する負の転移問題を考慮していないと、性能が低下する。能動学習において、限られたラベル付きインスタンスから再サンプリングされた最初の仮説はしばしば不正確であり、信頼できない。

本論文では、これらの問題を解決するための効果的な２つの転移学習と１つの能動学習法を設計した。まず、われわれは論理的基礎に基づくパラメータなしの最小記述長原理（MDLP : Minimum Description Length Principle）を基にした転移学習アルゴリズムを提案した。元ドメインと目的ドメインの繋がりを構築するためにコードブックを導入し、MDLP を拡張した。負の転移を避けるため、帰納転移学習設定における超平面識別器のためのコンパクトな符号化法を提供した。次に、この枠組では、2 つのレベルの評価は、最小符号化によるコードの長さで表される異なるタスク間の類似性を測定することを提案している。このような方法で、類似しない元タスクの重みは、目的タスクへの負の転移を避けるために繰り返し減少する。最後に、転移学習知識を用いて能動学習を改良して、われわれのアルゴリズムはラベル付けされていないインスタンスの中で情報価値のあるものだけを選択し、高い分類精度を得るために用いた。同時に、適応手法として不要なインスタンスと下位モデルを排除するように設計した。

提案アルゴリズムの有効性を評価するため、生物学的なデータセットを含む実データと人工データの両方を用いた大規模な実験を行った。分類精度は、最新の既存手法に比べ5%から10%向上した。加えて、われわれのアルゴリズムはノイズに強いことも証明されている。能動学習においては、われわれのアルゴリズムは高い精度で少ない問い合わせを選択することができることが実証されている。

論 文 審 査 の 結 果 の 要 旨

今日、データマイニングの分類学習技術は、疾患の診断や生物学的分類など、種々の分野で活用されている。診断例がわずかしかない病気の診断など、医療や生物学のデータ集合においては、新たなタスクのためのラベル付き事例が不足している場合が多く、特にノイズレベルが高い場合には分類モデルを正確に構築することが困難である。転移学習と能動学習はラベル付き事例の不足を補うための２つの独立した学習であり、前者は源タスクでの学習から有用な情報を抽出して対象タスクでの学習に役立て、後者はラベルなし事例のラベルを問合せて学習に役立てる。転移学習では、理論に基づきパラメータが不要な手法は少なく、その結果ノイズに脆弱である場合が多い。さらに、タスク間の異なる分布に起因する負の転移問題に関する考慮が不十分であるため、源タスクと対象タスクの類似性が低い場合の性能低下が大きい。能動学習においては、転移学習のための問合せ事例選択法がしばしば不正確であり、信頼できない。

本論文では、これらの問題を解決するために、源タスクと対象タスクにラベル情報がある帰納的転移学習に取り組み、２つの転移学習手法と１つの能動学習手法を設計した。まず、理論に基づきパラメータが不要な最小記述長原理を、転移学習用に拡張した。拡張にあたっては、両タスクに共通の符号表を導入して考慮する記述長の種類を2個から5個に増やした。提案した原理はノイズに頑健である最小記述長原理の特長を受け継ぎ、これまで困難であった高ノイズ問題に有効な転移学習手法の提案に成功した。次に、テキスト分類などに必須の超平面識別器の帰納的転移学習のために、負の転移を回避するためのコンパクト符号化法を提案した。提案した符号化法も、データのノイズレベルが高い場合でも正確な識別器を選別する能力に優れており、これまで困難であった重要問題に有効な学習手法の提案に成功した。最後に、クラスノイズがない超高次元データできわめて高い正答率を誇るサポートベクトルマシンについて、転移能動学習のための問合せ事例選択戦略を提案した。この戦略では、ラベルなし事例のうち情報価値があるものを選択するため、より高い分類精度を得ることができる。

提案アルゴリズムの有効性を評価するため、生物学的なデータ集合を含む実データと系統的に生成した人工データの両方を用いた大規模な実験を行った。分類精度は、最新の既存手法に比べ 5%から 10%向上し、提案手法がノイズにきわめて強いことも実証された。なお能動学習において、提案手法が高い精度で少ない問合せを選択することができることが実証された。

以上、本論文は帰納的転移学習における拡張最小記述長原理，超平面識別器用の負の転移を回避するためのコンパクト符号化法、サポートベクトルマシンについての転移能動学習のための問合せ事例選択戦略を提案してそれらの有効性を示したものであり、転移学習と能動学習に貢献する価値ある業績であると認める。

よって、本申請者は博士（工学）の学位を受ける資格があるものと認める。