

## 科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 25 年 6 月 5 日現在

機関番号：17104

研究種目：基盤研究（B）

研究期間：2009～2012

課題番号：21310131

研究課題名（和文） 相互作用情報と構造情報を統合した分子ネットワークの構築

研究課題名（英文） Construction of Integrated Molecular Network Based on Interaction and Structural Information

研究代表者

皿井 明倫（SARAI AKINORI）

九州工業大学・大学院情報工学研究院・教授

研究者番号：20221286

研究成果の概要（和文）：

本研究では、生体分子の構造、相互作用、ネットワークの情報を統合し、分子レベルからネットワークレベルまでをシームレスに結びつけるプラットフォーム、PDBnet、を開発した。また、これらの情報の関係を視覚的に理解できるように、検索や可視化インターフェイスを作成し、インターネットに公開した。さらに、PDBnetのデータを用いて、構造、相互作用、ネットワークと機能の関係について系統的な解析を行うことにより、PDBnetの有効性を示すことができた。

研究成果の概要（英文）：

We have developed an integrated database/tool of biomolecular network, PDBnet, based on the structural and interaction information, in order to get useful information from the structurome. We have developed various interfaces for searching data and visualizing the results, which enables users to get insight into the relationship among structure, interaction, network and function of molecules. Using the data in PDBnet, we have made systematic analyses on the relationship among various properties at molecular structure, interaction and network levels to demonstrated the utility of PDBnet.

交付決定額

（金額単位：円）

	直接経費	間接経費	合計
2009年度	3,600,000	1,080,000	4,680,000
2010年度	3,000,000	900,000	3,900,000
2011年度	3,200,000	960,000	4,160,000
2012年度	3,000,000	900,000	3,900,000
総計	12,800,000	3,840,000	16,640,000

研究分野：生物学

科研費の分科・細目：ゲノム科学・ゲノム情報科学

キーワード：分子ネットワーク、相互作用、分子構造、データベース

## 1. 研究開始当初の背景

すでに800以上の生物種で全ゲノム配列が明らかにされ、ゲノムスケールでの生物機能の体系的解析が可能となりつつある。細胞内のシグナル伝達や遺伝子発現制御は最も重要な生物機能のひとつであり、それは膨大な

数の分子の複雑な相互作用のネットワークで実現されている。現在この分子ネットワークを解明するためにさまざまな研究が精力的に行われている。ひとつには、分子間の相互作用が網羅的に解析され、分子の相互作用ネットワークが明らかになりつつある。これにより、さまざまな生物機能をシステムとし

て理解したり、疾病のメカニズムを明らかにすることが可能になりつつある。一方では、シグナル伝達や遺伝子発現制御の本質的な理解やそれに基づく医薬品の開発には、分子レベルでの構造と機能の理解が必須である。現在、世界各国で構造ゲノムプロジェクトに代表される構造解析研究が推進され、生体分子の構造データが急速に増大している。すでに、Protein Data Bank (PDB) [www.rcsb.org/pdb/]には5万個以上のエントリーが登録されている。最近では、機能的な関心から、多分子間の複合体の構造が主体となっており、個々の分子の数にすると12万個以上の分子構造がわかっていることになる。このように生体分子の構造空間(ストラクチュローム)はますます巨大化、複雑化している。これまでに上記のトップダウン的なアプローチとボトムアップ的なアプローチは個別にすすめられてきたが、今後の生物機能の解析や創薬にとっては両者の統合が必要となる。しかし、残念ながらまだ、分子からネットワークの階層にわたって包括的な解析を行うためのインフラが十分に整備されていない。

## 2. 研究の目的

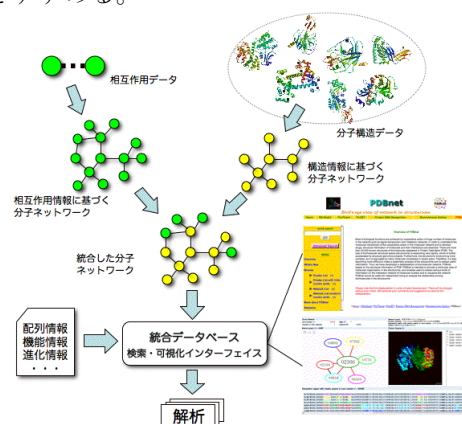
現在、ゲノムスケールでの分子相互作用の解析が精力的に行われ分子ネットワークが明らかになりつつある。一方、分子の構造解析の進展にも目を見張るものがあり膨大な数の分子構造が明らかにされている。しかし、分子とネットワークのレベルでの研究の間には大きなギャップがある。今後、分子ネットワーク機能のメカニズムを分子レベルで明らかにしたり、疾病の原因を分子からネットワークまで統合的に理解し創薬に結びつけたりするには、分子の構造レベルからネットワークレベルまでの多様なデータを統合し、新たな研究を促進する基盤を確立する必要がある。そこで本研究では、生体分子のゲノムスケールでの相互作用情報と分子構造の情報を統合し、分子レベルからネットワークレベルまでをシームレスに結びつけるプラットフォームを開発する。また、このシステムを用いて解析を行うことによりその有効性を検証することを目的とする。

我々はすでに、分子構造データベースの構造情報を用いて、分子の構造、分子間の相互作用、ネットワークを統合し、ストラクチュロームの全体「構造」を俯瞰することのできるツール、PDBnet、を作成し公開してきた[www.abren.net/pdbnet/]。これはいわばストラクチュロームの鳥瞰図であり、利用者は個々の分子の構造や相互作用とそれらのネットワーク上での位置関係を確認することができる。ただ、この構造情報から生成され

たネットワークは細胞の分子ネットワークの部分集合である。生体分子間の相互作用は、構造解析以外にさまざまな生化学実験等により、個別の分子間相互作用あるいはゲノムスケールで網羅的な相互作用による解析がこれまで行われている。そこで、このような分子相互作用の情報を構造情報に基づく相互作用ネットワークとPDBnet上で統合する。これにより、拡大した相互作用ネットワークを構築し、構造情報と生化学実験に基づく相互作用情報を補完し、新たな研究を促進すること目的とする。

## 3. 研究の方法

本研究では、分子とそれらの相互作用の構造からネットワークまでをシームレスに統合し、そこに含まれる多様な情報を包括的に解析できるようなプラットフォームを構築する。そのために、次図のような手順で研究をすすめる。



まず、分子構造データベースの構造情報を用いて、ストラクチュロームを分子構造からネットワークまで俯瞰するためのツールを作成する。構造情報は、PDBのデータを用いた。PDBの構造は一本鎖の蛋白質や核酸などの分子からなる「chain」と呼ばれる構造単位からなる。蛋白質chainの場合、配列が同等あるいは類似の構造がいろいろな複合体構造に含まれている。したがって、chainの情報をそのまま使って相互作用ネットワークを構築すると、冗長で複雑なものになってしまう。そこで、このような配列の冗長性を除去するため、類似のchainをグループ化(クラスタリング)する必要がある。このために、蛋白質配列の総合データベースであるUniprotのIDを用いた。Uniprotでは蛋白質に対して、PDBのアノテーションも行っている。このデータを利用することで、PDBchainのクラスター化を行う。

蛋白質では通常、構造の一部で他の部分とは独立に進化して、機能をもつドメインからなる。それぞれの蛋白質ドメインはコンパクト

トな三次元構造を形成し、ポリペプチド鎖の一部が他から独立した密で安定な構造に折り畳まれている。蛋白質ドメイン情報は、ドメインデータベース Pfam [pfam.sanger.ac.uk/] から取得した。一方、Pfam でドメインと定義されていないリンカーなどの非ドメイン領域も機能にかかわっていることがある。そこで、非ドメイン領域も同時に抽出し、配列上での互いの位置関係の情報もデータベース化した。

複合体に含まれる各 chain はそれぞれが相互作用している。PDB の三次元構造情報をもとに、chain の各原子間の物理的距離を計算し、閾値（通常は最短距離が 5Å）以下となる chain 間の相互作用を抽出した。蛋白質の場合は、ドメインレベルでの相互作用しているペア（蛋白質ドメイン-蛋白質ドメイン、蛋白質ドメイン-非ドメイン、非ドメイン-非ドメインの3種）のリストを作成した。このように作成した相互作用リストからノードをそれぞれの蛋白質あるいは核酸とし、エッジを相互作用として表示した。蛋白質の場合には、さらにドメインまたは非ドメインをノードとし、エッジをドメイン/非ドメイン間の相互作用として表示した。これにより、構造情報に基づく chain レベルの相互作用ネットワークあるいはドメイン/非ドメインレベルの相互作用ネットワークを構築した。

相互作用の情報については、さまざまな生化学実験により、個別の分子間相互作用、あるいはゲノムスケールで網羅的な相互作用がこれまでに解析されている。そこで、このような分子相互作用の情報を構造情報に基づく相互作用ネットワークと統合する。生化学実験などで得られた情報は、IntAct、MINT、DIP、HPRD、BioGrid の5つのデータベースを使用した。生化学実験情報はさまざまな質のデータを含むので、測定の方法や網羅性などの観点からその信頼性評価を行った。さらに、分子の配列や機能などに関する情報も統合した。

この統合データベース、PDBnet、は検索や可視化のインターフェイスをとおしてインターネット上に公開した。Web インターフェイス作成には、主に PHP を使用した。ネットワークの描画には Cytoscape Web を使用した。Cytoscape Web はネットワーク描画のためのライブラリを無料で提供していて、Java と Java Script をベースに作られている。

またこれと並行して、PDBnet のデータを用いた解析を具体的に行うことにより、その有効性を検証した。解析のひとつとして、ハブ蛋白質の同定と分類を行った。ここでは、ハブ蛋白質は、生化学実験による相互作用が 5 以上確認されている蛋白質、または、構造解析による相互作用が 3 以上確認されている蛋白質として定義した。生化学実験による相互

作用と構造解析による相互作用が 1 である蛋白質はノンハブ蛋白質と定義した。これらのハブ蛋白質/ノンハブ蛋白質の分類と、相互作用による構造変化と機能の関係について解析を行った。機能の特徴解析には、BiNGO を使用した。BiNGO により統計学上有意に変動する Gene Ontology (GO) を抽出した。もう一つの解析として、ドメイン/非ドメイン間の相互作用と機能の関係について解析を行った。相互作用に関与する部位は、アミノ酸配列や立体構造上の特徴的なパターン、すなわちモチーフを形成する。このモチーフを集めたデータベースである PROSITE[13] を利用して、モチーフをドメイン/非ドメイン領域にマッピングした。この情報に基づいて、モチーフを持つドメイン/非ドメインの共起関係などの統計解析を行った。

#### 4. 研究成果

本研究では以下のようなテーマについて解析を行い、次のような成果を得た。

##### (1) 構造情報を用いた分子ネットワークの構築

この課題では、ストラクチュロームの全体像を俯瞰するために、PDB に含まれている約 87,279 個の構造データから 214,809 個の chain、すなわち構造単位としての分子、を抽出した。蛋白質ドメインデータベース Pfam に登録されている 143,533 個のドメインについて、chain 毎に蛋白質ドメイン情報と照合した結果、Pfam ドメインをもつ chain 数は 107,365 個であった。また、非ドメイン領域の同定を行った結果、326,786 個の非ドメインが得られた。このようにして、Pfam ドメインと非ドメインのカタログを作成した。

蛋白質 chain 間の原子間相互作用の計算の結果、直接相互作用 (chain 間の原子間距離が 5Å 以下) する chain 間相互作用の数は 152,586 個であった。そして、これらの蛋白質間の原子間相互作用情報に基づいて、それらのドメイン/非ドメインレベルでの相互作用情報のカタログを作成した。

PDB chain は配列の類似度に基づくクラスター化を行い、クラスターのメンバー間の構造上での相互作用を網羅的に調べた。この解析に基づきクラスター間にリンクを作り、リンクを繋げていくことにより独立のネットワークを構築した。これまでに、約 1,000 個のネットワークが得られた。

##### (2) 生化学実験による相互作用情報の分子ネットワークへの統合

これまでに、いくつかの生化学的な実験方法で蛋白質・蛋白質相互作用がゲノムスケールで網羅的に解析されている。それらのデー

タから IntAct、DIP、BOND などのデータベースが作られている。これらの網羅的な相互作用から得られた分子ネットワークは、細胞内の多くの蛋白質間相互作用を反映していると考えられる。そこで、これらの中から公開されているデータを取得し、我々のデータベースと統合した。IntAct、MINT、DIP、HPRD、BioGrid の 5 つのデータベースから得られた生化学実験情報を持つ蛋白質は 82,481 個であった。このうち、構造情報を持つ蛋白質は 7,148 個であった。生化学実験情報で相互作用している蛋白質のペアの数は、は 571,617 個であった。このうち構造情報を持つ蛋白質のペアは 3,830 個であった。これらの情報は、PDBnet 上で構造情報による相互作用ネットワークに統合した

相互作用の信頼性スコアの解析の結果、全体として信頼性スコアの高いものが多かった。しかし、網羅的の実験のみの信頼性スコアの低い相互作用情報もかなり含まれていた。

### (3) 統合データベースと検索・可視化インターフェイスの作成

データベースは、PDBnet を土台としてこれを拡張する形で情報を統合化した。これまでの PDBnet では、分子情報から検索し、その構造、クラスターやそれらのネットワークを可視化できるようになっている。また、分子ネットワークと対応する分子の構造、相互作用する領域、などがマップされた配列が同じ画面に表示され、配列、構造、相互作用とネットワークの関係が視覚的に理解できるようになっている。今回作成したドメイン・非ドメイン情報を PDBnet に統合することによって、PDBnet 上でのドメイン・非ドメイン情報とそれらの相互作用の可視化を行った。具体的には、PDBnet の Web インターフェイス上のクラスター情報のページにて描画したネットワークのエッジから下図のようなドメイン情報のページへのリンクを作成し、ドメイン情報のページにてドメインレベルでの相互作用ネットワークと chain 間の立体構造相互作用部位の描画を行った。

QuickTime's C<sup>o</sup>  
RUIEVECEEA  
C<sup>o</sup>CA<sup>o</sup>CAEENE<sup>o</sup>EE<sup>o</sup>VA<sup>o</sup>EE<sup>o</sup>CE<sup>o</sup>CE<sup>o</sup>CA<sup>o</sup>B

この図で大きな 2 つのノードはそれぞれの chain を表しており、ノードの色は Jmol で表

示した立体構造の色と統一してある。さらにサブグラフ内のドメインノードやエッジをクリックした場合の Jmol 描画を以下に示した。

*IydpA* のノードの  
*PF00129* をクリック  
した場合

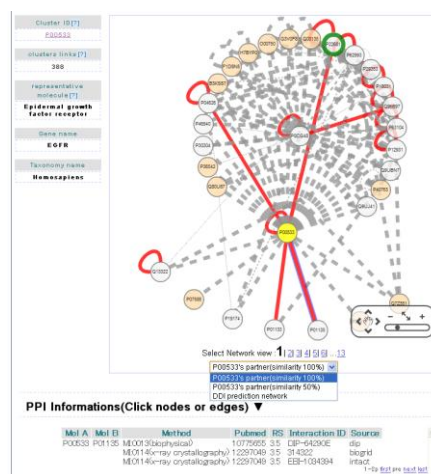
QuickTime's C<sup>o</sup>  
RUIEVECEEA  
C<sup>o</sup>CA<sup>o</sup>CAEENE<sup>o</sup>EE<sup>o</sup>VA<sup>o</sup>EE<sup>o</sup>CE<sup>o</sup>CE<sup>o</sup>CA<sup>o</sup>B

*IydpA* の *PF00129* と  
*IydpB* の *PF07654* を  
結ぶエッジをクリック  
した場合

QuickTime's C<sup>o</sup>  
RUIEVECEEA  
C<sup>o</sup>CA<sup>o</sup>CAEENE<sup>o</sup>EE<sup>o</sup>VA<sup>o</sup>EE<sup>o</sup>CE<sup>o</sup>CE<sup>o</sup>CA<sup>o</sup>B

このようにネットワークと Jmol による対応する構造の描画を同期的に動作することができるようにインターフェイスを作成した。エッジをクリックした場合には、各ドメインの相互作用部位のカラーを magenta にし、各ドメイン間の距離が 5 Å 以内に存在する chainA の原子と chainB の原子を結ぶラインを同じく magenta で表示した。

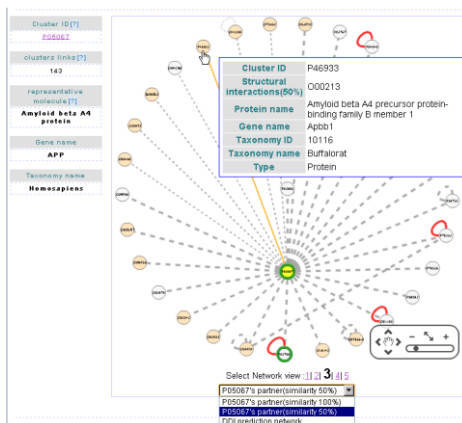
PDBnet で遺伝子名や Uniprot ID での検索後に、構造情報による蛋白質・蛋白質相互作用 (PPI) ネットワークか生化学実験情報による相互作用ネットワークを選べるようにインターフェイスを作成した。



生化学実験 PPI のネットワーク表示では、白のノードが構造解析されている蛋白質であり、薄ピンク色のノードが構造解析されていない蛋白質とわかるようにした。また、緑で円を囲っている蛋白質は遺伝子変異による疾患が確認されている蛋白質である。遺伝子

変異による疾患情報は OMIM [www.ncbi.nlm.nih.gov/omim/] から取得した。エッジは、赤の実線が構造解析による相互作用が確認されているもので、点線が生化学実験情報のみの相互作用情報である。点線に関しては、信頼性スコアが高いほど、エッジの太さが太くなるようにした。巨大なネットワークの表示はブラウザに負荷を与えるので、信頼性スコアの高い順に 30 個ずつ表示させるようにした。エッジをクリックすると PPI 情報に非同期で表示される仕様にした。相互作用している蛋白質の情報としては、蛋白質名、生物種や相互作用の信頼性スコアを表示した。

その他、ネットワークの切り替えが行えるようにした。相互作用している蛋白質の相同性 50% のクラスターへの変換や、構造情報によって計算された蛋白質ドメイン間の相互作用情報を用いた相互作用の予測がある。前者の場合、構造解析の際に、Homo Sapiens の蛋白質で解析ができず、生物種の異なる同じ蛋白質で構造解析を行うことがある。そのような場合に、相同性 50% の変換をすることで、相互作用の有無を推測することができる。変換後に構造的相互作用が確認されている PPI に関しては、エッジがオレンジ色に変わるようにした (下図)。



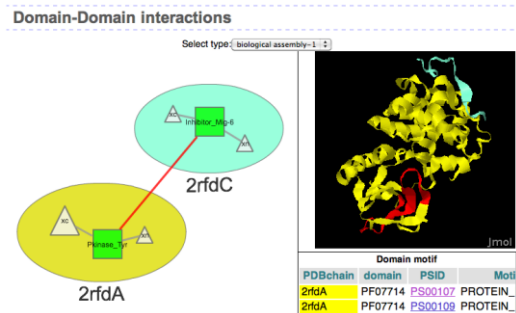
ドメイン間相互作用を用いた相互作用予測については、構造情報から得られたドメイン間相互作用の情報を使い、仮にある蛋白質ペアに対して相互作用するドメインが双方に含まれていれば蛋白質間に相互作用があると予測する。このように予測された相互作用は、エッジの色を変えて表示するようにした。

#### (4) PDBnet のデータを用いたパイロット研究

今回統合されたデータベースの情報を用いて、分子構造、相互作用からネットワークの間の関係について解析を行った。

まず、ドメイン・非ドメインの情報を用い

て、構造と機能の関係を解析した。ドメインの構造と機能についてはよく調べられているので、ここでは主に非ドメインの機能解析を行った。Prosite [prosite.expasy.org/] が提供している ps\_scan と呼ばれるモチーフ配列の検索プログラムを用いて、Prosite から取得したモチーフ配列と非ドメインとの配列相同性検索を行った。その結果、326,786 個の非ドメインにおいて、207,340 個にモチーフが確認された。得られたモチーフは 1,016 種類であった。これらのモチーフは構造上にマップし視覚化できるようにした (下図)。



このようにして得られたモチーフが蛋白質構造中でどのような機能を果たしているかを解析するため、周囲の Pfam ドメインとの位置関係を含む関係性から有意な非ドメインにおけるモチーフと Pfam ドメインとの組み合わせを抽出した。これらのペアについて統計的に有意なものを抽出し、実際のいくつかの例について考察した。その結果、多くの例では、非ドメイン中のモチーフはドメインと協調することにより機能を果たしていることが示唆された。

PDB の蛋白質に対して、ハブ蛋白質の単体と複合体の構造変化による分類を RMSD (Root Mean Square Deviation) を尺度として用いて行ったところ、相互作用の際に構造変化の大きいハブ蛋白質は 55 個、構造変化の小さいハブ蛋白質は 130 個、構造変化の大きいノンハブ蛋白質は 14 個、構造変化の小さいノンハブ蛋白質は 37 個であった。構造変化の大きいハブ蛋白質群の機能で有意な結果が得られたのは、結合に関するものであった。その中には、DNA グリコシラーゼ、SUMO、カルモジュリン、トポロニン C など、カルシウム結合、シグナル配列結合、損傷 DNA 結合、翻訳関連因子活性などが含まれていた。一方、構造変化の小さいハブ蛋白質群の機能で有意な結果が得られたのは触媒活性グループ、加水分解、酸化還元、リアーゼ活性などであった。次に大きなグループが結合に関するグループであり、キナーゼ結合や補酵素結合などの機能があった。その他にもキナーゼ阻害

活性の機能もあった。全体的に酵素が多く、アネキシン、シトクロム C、トリプシン、ユビキチン輸送蛋白質などがあつた。これらの結果から、構造変化の大きいハブ蛋白質は、分子間相互作用により形を変え複合体となり、情報伝達や機能発現する蛋白質が多いと考えられる。一方、構造変化の小さいハブ蛋白質は、化学反応によって情報伝達や機能発現する蛋白質が多いと考えられる。このように、情報的なアプローチによって、ハブ蛋白質の特徴を抽出することができた。今後増加するデータをさらに詳細に解析することにより、構造、相互作用とネットワークの関係が明らかになるであろう。

#### 5. 主な発表論文等

(研究代表者、研究分担者及び連携研究者には下線)

[雑誌論文] (計 16 件)

- ①. S. Yamasaki, T. Terada, H. Kono, K. Shimizu and A. Sarai "A New Method for Evaluating the Specificity of Indirect Readout in Protein-DNA Recognition" *Nucleic Acids Res.* 40, e129 (doi: 10.1093/nar/gks462) (2012). (査読あり)
- ②. M. Fernandez, Y. Kumagai, D. Standley, K. Mizuguchi, A. Sarai and S. Ahmad "Prediction of dinucleotide-specific RNA-binding Sites in Proteins" *BMC Bioinformatics* 12(Suppl 13), S5 (2011). (査読あり)
- ③. H. Okayama, T. Kohno, Y. Ishii, Y. Shimada, S. Ahmad and A. Sarai "Analysis of electric moments of RNA-binding proteins: implications for mechanism and prediction" *BMC Struct. Biol.* 11, 8 (2011). (査読あり)
- ④. S. Ahmad, O. Keskin, K. Mizuguchi, A. Sarai and R. Nussinov "CCR XP: Exploring clusters of conserved residues in protein structures" *Nucleic Acids Res.* 38 Suppl:W398-401 (2010). (査読あり)
- ⑤. M. Fernandez, S. Fujii, H. Kono and A. Sarai "Evaluation of DNA intramolecular interactions for nucleosome positioning in yeast" *Genome Inform.* 23, 13-20 (2009). (査読あり)
- ⑥. M. Andrabi, K. Mizuguchi, A. Sarai and S. Ahmad "Prediction of mono- and dinucleotide-specific DNA-binding sites in proteins using neural networks" *BMC Struct. Biol.* 9, 30 (2009). (査読あり)

[学会発表] (計 36 件)

- ①. Yuji Kawakami, Satoshi Fujii and Akinori Sarai "Integration of Structural and Biochemical Information into Protein Interaction Network" 生命医薬情報学連合大会、東京、2012 年 10 月 14 日-17 日
- ②. Satoshi Komori, Satoshi Fujii and Akinori Sarai "Analysis of Domain and Non-Domain Interactions in Protein-Protein Interaction Network" 生命医薬情報学連合大会、東京、2012 年 10 月 14 日-17 日

[その他]

ホームページ等

<http://www.abren.net/pdbnet/>

#### 6. 研究組織

##### (1) 研究代表者

皿井 明倫 (SARAI AKINORI)

九州工業大学・大学院情報工学研究院・教授

研究者番号：20221286

##### (2) 研究分担者

藤井 聡 (FUJII SATOSHI)

九州工業大学・大学院情報工学研究院・助教

研究者番号：40452825