様式 C-19

科学研究費補助金研究成果報告書

平成２３年６月１日現在

研究種目：基盤研究（Ｂ）
研究期間：２００７～２０１０
課題番号：１９３００１０２
研究課題名（和文）　候補サイトの空間分布に注目した信頼度の高いタンパク質ドッキング判定手法の開発
研究課題名（英文）　An exhaustive all-to-all protein-protein interaction prediction algorithm based on 3-D shape complimentarity and its statistical distribution on protein surface.

研究代表者
秋山　泰（AKIYAMA YUTAKA）
東京工業大学・大学院情報理工学研究科・教授
研究者番号：３０２４３０９１

研究成果の概要（和文）：
複数のタンパク質の立体構造データを入力し、表面形状相補性と静電相互作用に基づいて、各ペアがドッキングするか否かの判定を行うための手法を開発した。既存法 ZDOCK とは異なり、実数のみで表面形状相補性を計算する rPSC モデルの着想を得て、精度を落とさず約４倍の高速化が達成できた。タンパク質の性質に応じて評価関数を動的に調整する手法も開発した。候補サイトの空間分布に注目した後処理により信頼度の向上が達成できた。

研究成果の概要（英文）：
We have developed a novel computing method for rigid-body protein docking and applied it to Protein-Protein Interaction network prediction in systems biology. The original rPSC model uses only real numbers while ZDOCK's PSC uses complex numbers and then our method is four-times faster than ZDOCK without loss of sensitivity. We also proposed a method to dynamically change balance between shape complementarity and electrostatics. Total prediction accuracy is improved with clustering and reranking as post-processing.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	3,800,000	1,140,000	4,940,000
２００８年度	3,500,000	1,050,000	4,550,000
２００９年度	3,500,000	1,050,000	4,550,000
２０１０年度	3,400,000	1,020,000	4,420,000
年度			
総　計	14,200,000	4,260,000	18,460,000

研究分野：総合領域
科研費の分科・細目：情報学・生体生命情報学
キーワード：コンピュータシミュレーション、蛋白質、ドッキング、高性能計算

１．研究開始当初の背景

(1) タンパク質間ドッキングを計算機によって調べようという研究は研究開始当時にも存在していたが、それはおよそ３つほどのアプローチに大別できた。第一は、分子動力学法やその拡張により１対１のタンパク質ドッキングの自由エネルギーを見積もるもの。第二は ZDOCK に代表されるような形状相補性などの簡易計算によりドッキング姿勢を求めるもの。第三は立体構造情報は用いずに、配列モチーフや既知複合体との進化的類縁に基づき、ドッキング可否を判定するもの。

(2) システム生物学の時代が到来し、トランスクリプトーム解析のデータから遺伝子間ネットワークは急速に明らかにされていたが、Y2H や MS/MS によるタンパク質間相互作用の実験データは相互に再現性が悪く、一致

しない等の問題が指摘されていた。計算機による網羅的解析によりこれを支援することが強く望まれていた。

(3) 例えばBlueGeneやPCクラスタを用いて数千CPUコアによる大規模計算がタンパク質構造計算等に応用され始めていた。しかし先述の第一の手法では、計算量があまりに膨大であり１対１の計算でも困難であった。しかも正しいドッキング位置と姿勢が予め判っていなければ計算を開始できない。また第三の手法は、バイオインフォマティクス分野では広く受け入れられていたが、急速に蓄積されつつある立体構造データを活用されないことと、未知の相互作用の発見が期待できないことが問題と思われた。大規模計算を前提とするならば、第二の手法を高速化することが最も有望であると考えた。しかし当時は第二の手法は１対１に適用することが前提であり、当研究のように「1000×1000規模の計算を網羅的に行う」などと標榜することは、ほぼ絵空事のように認識されていた。

２．研究の目的

(1) ドッキング計算の高精度化と高度化

本研究の目的は、高精度なタンパク質ドッキング判定手法を開発することである。このためには、１対１でのドッキング計算そのものを高精度化する必要があり、また目標に掲げた 1000×1000 もの計算を可能にするために高速化も行わねばならない。これらのために必要な技術を開発する。

(2) 候補サイトの分布に注目した後処理

目的を果たす上で本研究の特徴的な点は、計算結果の後処理により精度の向上を図る点であり、この方法論を開発する。ZDOCK などの既存研究が当時は単純にスコア第一位に注目していた点とは異なり、候補サイトの空間分布にも注目し高度な後処理を行う。

(3) システム生物学への応用と評価

少数の例だけでは手法の性能を評価できないため、大きな規模のベンチマーク問題やシステム生物学の実問題への応用を通して、予測精度に関する評価を行う。

本研究の応募時点では、20×20程度の計算を実施済みで、これをせめて実施期間内に数百×数百に拡張したいとの目標を立てた。

(4) 入力立体構造の多様化

システム生物学への応用などに直面すると、ドッキング時の構造(bound 構造)からはかけ離れた構造データ(unbound 構造)しか入手できないケースがある。この場合に入力構造を変化させてアンサンブルを取るなどの手法が考えられるので、これらを検討する。

３．研究の方法

(1) ドッキング計算の高精度化と高度化

以下のいくつかのレベルに分けて研究を行った。

・ドッキング評価関数の最適化
・FFTの底の最適化
・角度刻みの選択と角度毎の候補数の検討
・スレッド並列化による高速化
・MPI並列化による高速化

(2) 候補サイトの分布に注目した後処理

・候補デコイの空間距離に基づくクラスタ化
・クラスタ化手法の検討
・その他の手法によるクラスタ化
・クラスタ化を用いない後処理方法

(3) システム生物学への応用と評価

・ベンチマーク問題による評価
・細菌走化性系に対する応用と評価
・EGFR系に対する応用と評価

(4) 入力立体構造の多様化

・MDによるアンサンブル作成の検討
・アラニン変換による入力構造の検討

４．研究成果

(1) ドッキング計算の高精度化と高度化

・ドッキング評価関数の最適化

本研究の開始当初は、Katchalski-Katzir(1992)の方法に基づいて評価関数を自作していたが、ZDOCK (Chen, et.al, 2002)の性能に比べて予測精度が劣っていたため、ZDOCKのPSC関数についての解析を実施した。その結果、PSC 関数では形状相補性が複素数で表現されているが、必ずしも複素数を用いなくても、実数の範囲でほぼ近似的に同様のことが可能であるとの着想を得た。三次元格子どうしを平行移動しながら畳み込み和を計算する際に、計算量を減じるテクニックとしてフーリエ変換が広く用いられる。このとき、フーリエ変換をするならば元が複素数であっても実数であっても変わらぬ手間がかかるために従来は見過ごされていたのであろうが、静電相互作用等の複雑なエネルギー項を付け加えていくにつれて、FFT の回数が増えて計算時間が延びていた。これに対して、新規提案した rPSC (real Pairwise Shape Complementarity) スコアを用いると、複素数の実部に rPSC を、虚部に別の静電相互作用に関する項を入れて、一度の FFT 計算で２つの相互作用の評価を一度に行うことが可能になる。

当研究の成果は MEGADOCK と称するソフトウェアに反映している。rPSC を導入した結果 MEGADOCK (ver.2.3)が ZDOCK3.0 に比べて高速化されたことを図１に示す。また予測精度は同等であることを図２に示す。

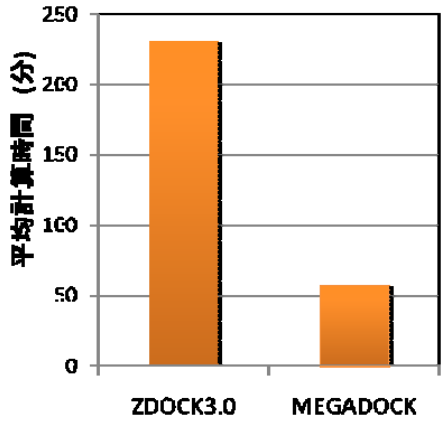

図１　計算時間の比較（右が提案法）

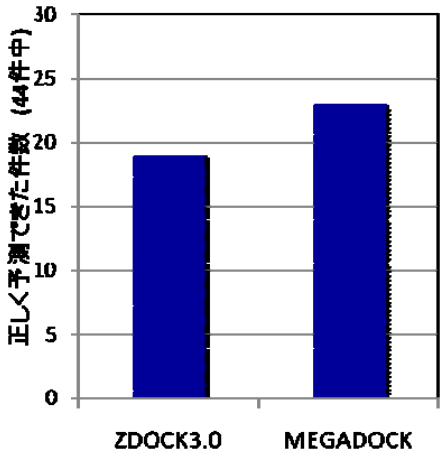

図２　予測精度の比較（右が提案法）

図１は、(3)で後述する 44 件の複合体によるベンチマーク試験を既存法(ZDOCK3.0)と提案法で計算させたときの計算時間の比較。また図２は同じ計算において、ほぼ実験構造に近い構造を得られた件数を比較している。精度には有意な差はないが、速度は４倍程度向上している。なお後述する並列化を行う前の計算時間である。rPSC については、秋山が基本的な着想を行い、秋山研究室の学生の大上が実装および評価等を実施した。

また、rPSC においては、実数部の形状相補性と虚数部の静電相互作用の効果がある重み付け係数によって加算されて評価値となるが、この係数はタンパク質の性質や大きさなどによって動的に決定すべきであるとの着想を得て、係数を決定するための方式を提案した。この研究は秋山研究室の学生の松崎が遂行し、情報処理学会山下記念研究賞を受賞（学会発表⑦）した。

・FFT の底の最適化

研究開始当初は FFT の底は 2 だけであり、タンパク質のモデルの一辺は 2 の冪乗に限定していたが、2, 3, 5, 7 を組み合わせた FFT を可能としたことによりタンパク質の大きさに対して最適なモデルサイズにできた。

・角度刻みの選択と角度毎の候補数の検討

15 度刻みと 6 度刻みでは計算量が 15 倍になるため前者を採用した。ただし角度毎に既存法では１つの平行移動位置のみを報告していたのに対して、任意個数（通常 3 個）を報告することにより、刻みを細かくしたのと似た若干の効果があることを見いだした。

・スレッド並列化による高速化

MEGADOCK ソフトウェアの並列化については、科研費以外の別の研究プロジェクトとして主に実施した。しかしその成果として、１ノードに複数の CPU コアが存在する場合に、回転角度をスレッド並列化して同時計算することにより約８倍までの高速化が達成され、図１に比べてさらに高速化された。

・MPI 並列化による高速化

さらに CPU ノード間については MPI 並列化することにより、複数のリガンドと複数のレセプタのドッキング計算を同時並行的に並列計算機上で実施できるようになった。

(2) 候補サイトの分布に注目した後処理

・候補デコイの空間距離に基づくクラスタ化

15 度刻みでも 3600 通りの回転となり、回転角度毎に 3 個のデコイを報告すると 10800 通りとなる。このうちスコアの上位 6000 個程度を後処理に用いている。

後処理の概念図を図３に示す。空間的に隣接しているデコイをクラスタにまとめて、クラスタ毎に代表デコイを決定して、後のドッキング判定に用いる。

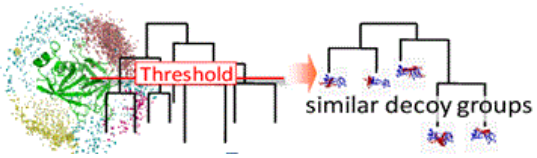

図３　候補デコイ群のクラスタ化

このとき、当初はレセプタタンパク質の周囲のリガンドの重心点により単純にクラスタ化していたが、リガンドの回転角度が反映されないためリガンドから見たレセプタのいちについても考慮する方法を考案した。

・クラスタ化手法の検討

クラスタ化の手法としては、スコア順にグリーディにグループ化を行う単純な方法の他に、ボトムアップクラスタリングの平均距離法、Ward 法など各種を検討して、その性質の差違を明らかにした。しかし現在ではドッキング計算が著しく高速化されているため、後処理手法は簡便な計算法が適している。

・その他の手法によるクラスタ化

秋山研究室の内古閑による、ドッキング時の残基間インターフェース・プロファイル

(IFP)を用いた手法を共同で検討して、良好な結果を得ている。前述の空間的な分布のみによる方法に比べて、より詳細にドッキングの状態を検査し、インタフェースの状態がどれだけ似ているかを候補デコイ間の距離に用いている。

・クラスタ化を用いない後処理方法

計算の目的にもよるが、多数×多数の構造間でドッキングの有無の予測のみを行う場合においては、空間的なクラスタ化が必ずしも重要ではなく、ZRANK 等によるリランキング処理により不適切な解を削除した後に、ドッキングのトップスコアを持つものに注目するだけで予測精度は確保できることなどが判った。これは当研究を開始した当初には予想しなかったことであるが、ドッキング計算がアルゴリズムの改良と各種の並列化により著しく高速になったため、後処理にかけられる時間が相対的に短くなっていることにも関連する。当研究で得たさまざまな後処理は１対１のドッキングには引き続き有効と考えているが、クラスタ化を用いない後処理方法についても、今後は検討を要する。

(3) システム生物学への応用と評価

・ベンチマーク問題による評価

ドッキングの可否を判定する問題に提案手法を用いる際には、(2)で紹介した後処理で得られた評価値が、しきい値よりも高いか否かで判定を行う。そのしきい値を決めるベンチマーク問題が必要であったため、ZDOCK benchmark 2.0 に基づく 44 レセプタ×44 リガンドの集合を設計し、以降の研究に用いた。図４はこのセットに対して提案手法を適用したとき、ドッキングすると判定されたものを赤、中間を黄、ドッキングしないと判定されたものを緑で表している。対角線要素(PDB に登録されていたペア)のみを正解だと厳しく考えると、Recall と Precision の調和平均である F 値は 0.415 である。

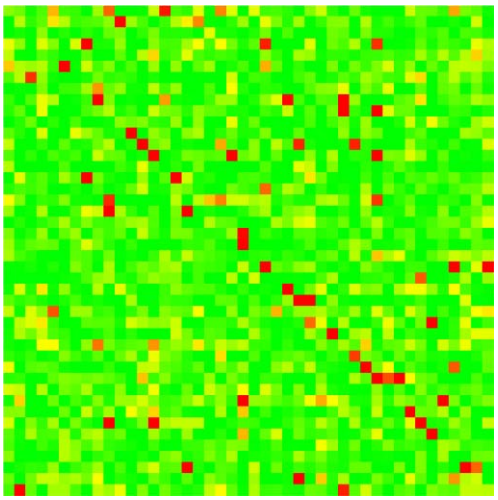

図４　44×44 ベンチマークの予測結果

その後、ZDOCK benchmark が更新されているため、より大型のベンチマークも作成して計測中である。また、秋山研究室の学生の大上が、当手法をタンパク質－RNA 間相互作用にまで拡張し、タンパク質-RNA 複合体を集めた 78×78 のベンチマークも作成した。

・細菌走化性系に対する応用と評価

システム生物学の実際的な例題に本手法を適用するために、細菌走化性系を選択した。これは秋山研究室の松崎由理が以前に走化性系の研究を行っていたことがあり、正解と思われるネットワーク構造について該博な生物学的知識を得られたためである。

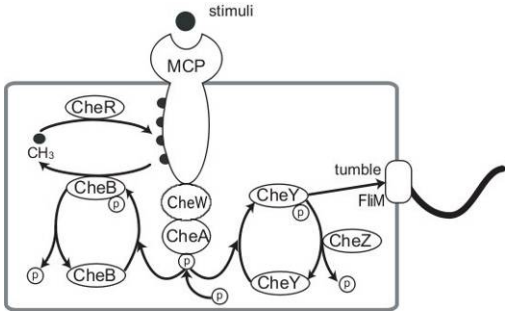

図５　細菌走化性系での既知相互作用

細菌走化性系では図５に示すような回路が知られており、MCP (Tar, Tsr 等)が受容した刺激と内部状態から右端の鞭毛モーターの回転方向が変化する。この系に関係する 13 種のタンパク質について PDB を検索したところ、101 個の立体構造データを取得できた(*E. coli* 42 件、*T. maritime* 26 件、*S. typhimurium* 33 件)ここで同じタンパクの構造が複数登録されていた場合はそのどちらもドッキングに用いて、構造の多様性を増やすことにした。いずれかの組み合わせでドッキングの可能性が指摘されたタンパク質間ではドッキングをすると判定することにした。さらに近縁間でタンパク質は混合して用いることにした。

提案手法によって、相互作用が示唆されたペアを図６に示す。太い実線は既知相互作用と一致したもの。細い実線は False Positive と思われるもの。破線は False Negative である。False Positive のうち、CheD-CheY については生物学的にもあり得る示唆であり、CheD-CheY-CheC の興味深い三体でのドッキング構造形成の可能性を示している。

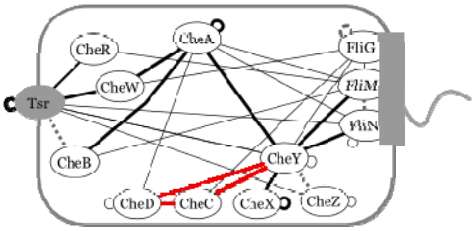

図６　細菌走化性系の相互作用予測結果

・EGFR 系に対する応用と評価
より大型の例題への適用を目指して、ヒトの肺がんなどに関係の深い EGFR シグナル伝達系への適用を行った。
前述の細菌走化性系では、同一タンパク質に対する異なる PDB データを重複して数えたとき、およそ 100×100 級の計算をしていたが、EGFR 系では良くわかっている既知部分だけの計算でも 500×500 規模となった。
すでに 500×500 の計算は終了しており、現在そのネットワークを調べている。
また本科研費課題とは独立に、東大医科研の宮野悟教授との共同研究により、新しい相互作用を実験的に調査中の拡張された EGFR 系について、1500×1500 規模の計算を実施中で、本課題の成果を応用している。
アルゴリズムの高速化と並列化により、1 対 1 の計算は 1 ノード上で 5 分程度まで短縮されており、100×100 級の計算は、数十から数百 CPU ノードが比較的気軽に使える最近では頻繁に実施できるようになった。1000×1000 規模については、大型の並列計算機が必要であるが、我々は本課題の終了以降にも東工大の TSUBAME2.0 や、神戸の次世代スーパーコンピュータ「京」を想定した計算を予定しており、本課題の提案時点では絵空事と考えられた 1000×1000 の計算をほぼ現実化するところまで進めることができた。

(4) 入力立体構造の多様化
・MD によるアンサンブル作成の検討
本研究課題の特徴は、「研究開始当初の背景」の項目でも述べたように、従来は配列モチーフなどを中心に行われていた網羅的なタンパク質間相互作用予測を、立体構造データを活用して行う点にある。しかしそれは同時に弱点でもあり、立体構造データが存在しない場合には当手法はそのままでは適用できない。またさらに本来は柔軟であり、induced fit なども起こしているタンパク質構造を剛体近似していることから、比較的堅い rigid body と見なせる種類のタンパク質以外では予測の精度が上がらないという問題点がある。剛体近似による当手法の守備範囲は、大きな動きを伴わないタンパク質に限られる。しかしそれでも、若干の表面の変化に耐えうる計算とするために、分子動力学法(MD)によりアンサンブルを作成して、検出感度を高める研究を実施した。MD 計算を行うと主鎖がゆるんでタンパクが拡がる傾向があるため、主鎖を拘束しながら側鎖のみ変更するなどの工夫が必要な事がわかった。
・アラニン変換による入力構造の検討
秋山研究室の内古閑との協調により、側鎖を全てカットするアラニン変換により、側鎖の衝突の影響を軽減する手法も提案した。

５．主な発表論文等

〔雑誌論文〕（計１０件）

①Masahito Ohue, Yuri Matsuzaki, Yutaka Akiyama, Docking-calculation-based Method for Predicting Protein-RNA Interactions, Genome Informatics, 25, (in press) 査読有

② 大上雅史, 松崎由理, 松崎祐介, 佐藤智之, 秋山 泰, MEGADOCK: 立体構造情報からの網羅的タンパク質間相互作用予測とそのシステム生物学への応用, 情報処理学会論文誌(TOM), 3 (3), 91-106 (2010) 査読有

③Makiko Kusama, Kouta Toshimoto, Kazuya Maeda, Yuka Hirai, Satoki Imai, Koji Chiba, Yutaka Akiyama, Yuichi Sugiyama, In silico classification of major clearance pathways of drugs with their physico-chemical parameters, Drug Metabolism and Disposition, 28(8), 1362-1370 (2010) 査読有

④Yuri Matsuzaki, Masahito Ohue, Nobuyuki Uchikoga, Takashi Ishida, Yutaka Akiyama, Computer prediction of protein-protein interaction network using MEGADOCK - application to systems biology, TSUBAME e-Science Journal, 2,34-37 (2010) 査読無

⑤ Yutaka Akiyama, Yuri Matsuzaki, Nobuyuki Uchikoga, Masahito Ohue, Exhaustive protein-protein interaction network prediction by using MEGADOCK, BioSupercomputing Newsletter, 3, 8-8 (2010) 査読無

⑥ Yuri Matsuzaki, Yusuke Matsuzaki, Toshiyuki Sato, Yutaka Akiyama, In silico screening of protein-protein interactions with all-to-all rigid docking and clustering: an application to pathway analysis, Journal of Bioinformatics and Computational Biology, 7 (6), 991-1012, (2009) 査読有

⑦ Koki Tsukamoto, Tatsuya Yoshikawa, Yuichiro Hourai, Kazuhiko Fukui, Yutaka Akiyama, Development of an affinity evaluation and prediction system by using the shape complementarity characteristic between proteins, Journal of Bioinformatics and Computational Biology, 6 (6), 1133-1156 (2008) 査読有

〔学会発表〕(計４７件)
① Yutaka Akiyama, Exhaustive Protein-Protein Interaction Network Prediction by MEGADOCK(招待講演), Asia Hub for e-Drug Discovery Symposium AHeDD 2010, 2010 年 12 月 18 日, Yonsei University (韓国)

②Yuri Matsuzaki, Masahito Ohue, Yusuke Matsuzaki, Toshiyuki Sato, Yutaka Akiyama, A computational screening system of protein-protein interactions: connecting protein structural information to biological pathway estimation, 11th International Conference on Systems Biology (ICSB 2010), 2010 年 10 月 12 日, Edinburgh Int'l Conference Centre (英国)

③Masahito Ohue, Yuri Matsuzaki, Yusuke Matsuzaki, Toshiyuki Sato, Yutaka Akiyama, In silico prediction of PPI network with structure-based all-to-all docking, InCoB2010 - the 9th International Conference on Bioinformatics, 2010 年 9 月 26 日, Waseda University (東京)

④Nobuyuki Uchikoga, Takatsugu Hirokawa, Yutaka Akiyama, Searching near-native decoys from various types of protein complexes by cluster analysis with Interaction Finger Print, The 48th Annual Meeting of the Biophysical Soc. of Japan, 2010 年 9 月 22 日, Tohoku University (仙台)

⑤Nobuyuki Uchikoga, Takatsugu Hirokawa, Yutaka Akiyama, Cluster analysis in post-docking process for rigid- body docking problem by using Interaction Fingerprints, CBRC2010, 2010 年 7 月 28 日, Computational Biology Res. Center (東京)

⑥Yuri Matsuzaki, Masahito Ohue, Yusuke Matsuzaki, Toshiyuki Sato, Yutaka Akiyama, A computational screening system of protein-protein interactions: connecting protein structural information to biological pathway estimation, Computing with GPUs, Cells, and Multicores, 2010 年 5 月 10 日, ETH Zurich (スイス)

⑦松崎裕介, 大上雅史, 松崎由理, 佐藤智之, 関嶋政和, 秋山泰, タンパク質の特性に基づく unbound ドッキングのための剛体予測手法の改良, 情報処理学会 第 20 回バイオ情報学研究会, 2010 年 3 月 4 日, 北陸先端大学 (金沢) [情報処理学会平成 22 年度山下記念研究賞　受賞]

⑧Masahito Ohue, Yusuke Matsuzaki, Yuri Matsuzaki, Yutaka Akiyama, Improvement of all-to-all protein-protein interaction prediction system MEGADOCK, Genome Informatics Workshop (GIW2009), 2009 年 12 月 14 日, Pacifico Yokohama (横浜)

⑨ Yuri Matsuzaki, Yusuke Matsuzaki, Masahito Ohue, Toshiyuki Sato, Yutaka Akiyama, In silico screening of protein-protein interactions with all-to-all rigid docking and clustering: an application to pathway analysis, The 10th International Conference on Systems Biology (ICSB 2009), 2009 年 9 月 1 日, Stanford University (米国)

⑩Yutaka Akiyama, Toshiyuki Sato, Yusuke Matsuzaki, Yuri Matsuzaki, Megadock - a rapid screening system for all -to-all protein docking analysis with precalculated fourier library of protein structures, The 2008 Annual Conference of the Japanese Society for Bioinformatics (JSBi2008), 2008 年 12 月 15 日, 千里ライフサイエンスセンター (大阪)

〔その他〕
ホームページ
http://www.bi.cs.titech.ac.jp/

６．研究組織
(1)研究代表者
秋山　泰 (AKIYAMA　YUTAKA)
東京工業大学・大学院情報理工学研究科・教授
研究者番号：３０２４３０９１

(2)研究分担者
該当なし

(3)連携研究者
塚本弘毅 (TSUKAMOTO KOKI)
産業技術総合研究所・生命情報工学研究センター・特別研究員　(２００７年度まで)
研究者番号：９０３９９５０１