４．研究開発課題名　　３Ｄ－１Ｄ法を用いた全遺伝子産物同定システムの研究開発

4.1　代表研究者　　国立遺伝学研究所 生命情報・DDBJ 研究センター
教授　西川　建

4.2 概要

大量に産出されるゲノム配列データをいち早く網羅的に情報解析し、データベース化することが本研究課題の目的である。アミノ酸配列からの立体構造予測を中心とする解析を行い、結果を GTOP データベースとして公開した（終了時点で 41 生物種の収録を達成）。解析結果の吟味により、大腸菌ゲノム中の偽遺伝子の発見等の生物科学的成果も得られた。

4.3 研究開発実施内容

大量に産出されるゲノム配列データをいち早く網羅的に情報解析し、その結果をデータベースとしてまとめ、公開することを本研究課題の目的とした。

ゲノム情報解析は、タンパク質レベルの解析、とりわけアミノ酸配列からの立体構造予測を中心とした。当初、解析ツールとして 3D-1D 法を予定していたが、公表されて間もない PSI-BLAST 法の威力が話題になりつつあったため、両者を比較評価した。その結果、立体構造予測における PSI-BLAST 法の優位（速度、感度、特異度）が明らかになったため、3D-1D 法に代えて PSI-BLAST 法を採用することとした。ゲノム情報解析として立体構造予測の他、配列ホモロジー検索、モチーフ検索、膜貫通ヘリックス／シグナル配列予測、繰り返し配列の同定などのコンピュータ解析も併せて行なった。これらの解析は、新たに配列の決定した生物種を順に解析対象とし、また立体構造予測に用いる立体構造データベース、配列データベースなどを定期的に自動更新して最新のデータを用いることにより解析精度を上げていった。構造予測された遺伝子（ORF）は約 40%に達した（図１）。

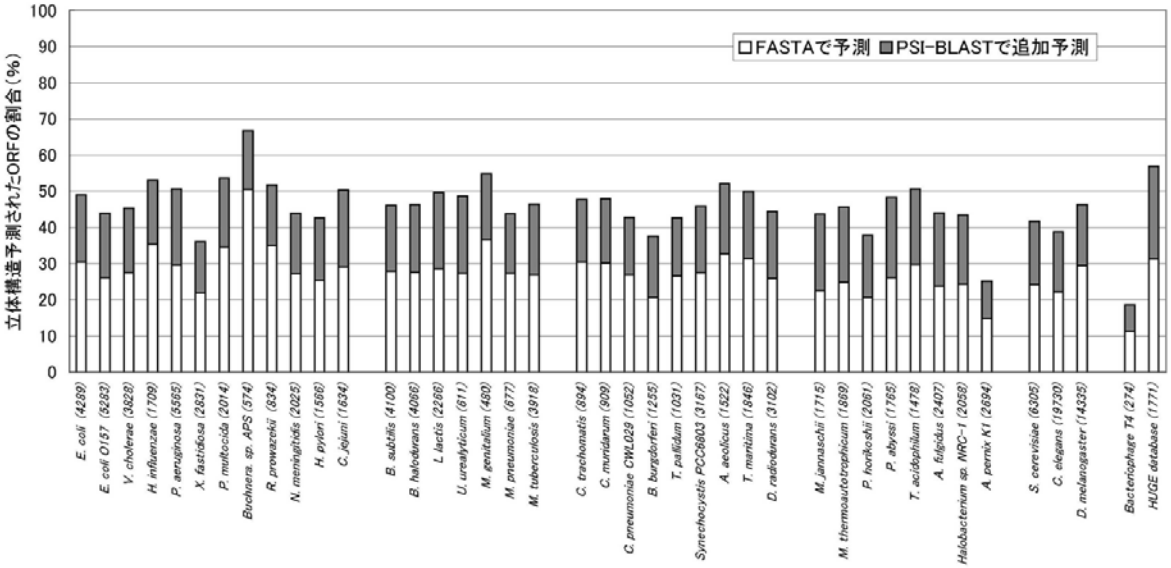

図 1. 立体構造予測結果の概要（立体構造予測されたＯＲＦの割合）

すべての解析結果は「GTOP データベース」[1]としてまとめて公開した。自動解析が完了し、GTOP に収録された生物種の数はプロジェクト終了時点（2001 年 9 月末）で 41 種に達している。GTOP データベースでは、属性値の組合せによる検索（図２）、遺伝子名／タンパク質名／ファミリー名による検索、などを提供している。検索機能により利用者の指定した条件に当てはまるタンパク質（の集合）をリストアップし、その中の各タンパク質に関する様々な解析結果を参照することができる。また、種毎の遺伝子の一覧、着目する遺伝子の近隣遺伝子の表示、配列アライメントの詳細表示・図示、種横断的なホモログ出現パターンなども提供しており、多面的・探索的な分析が行える。さらに、GTOP データベースでは、利用者の手持ちアミノ酸配列に対する BLAST および Reverse PSI-BLAST による立体構造予測サービスも提供している。

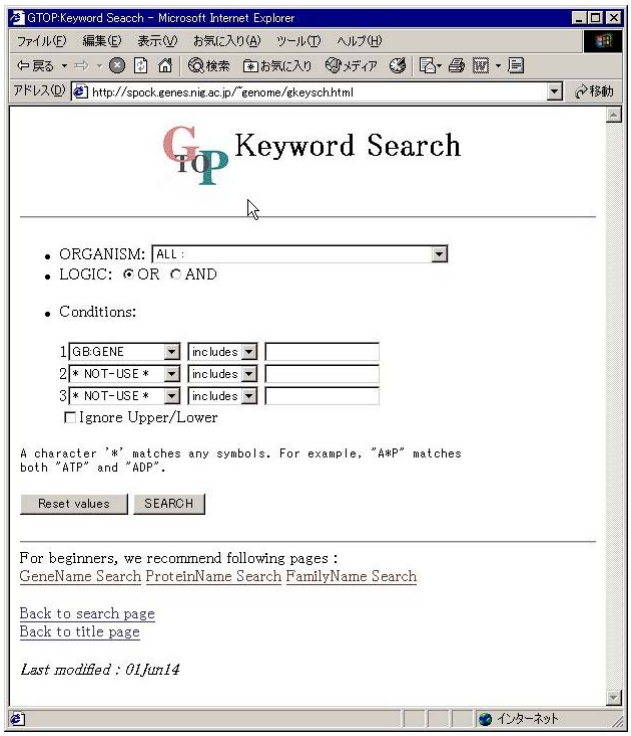

図 3. 隣接した不完全な立体構造

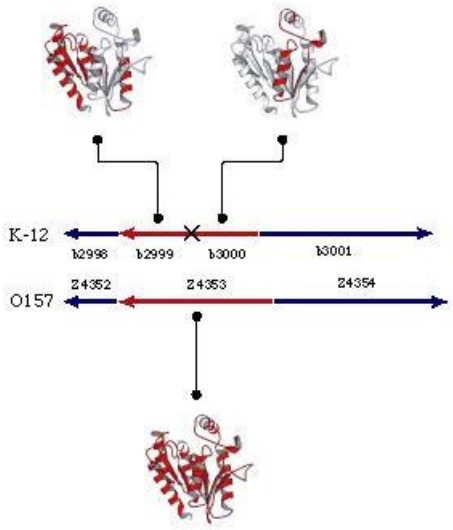

図 2. GTOP 検索画面

このようにして構築されたデータベースは、個別の遺伝子／タンパク質に関する「百科辞典」として意味をもつが、さらに具体的な生物科学的研究と結びついてこそ本当の価値が生まれる。我々は、GTOP の内容を詳細に吟味することにより、ORF の全長が既知の構造ドメインの一部分にしかヒットしない場合があること、あるいはゲノム上で隣接する ORF を２つ合せるとはじめて１つの完全な構造ドメインにヒットする事例が少なからずあることを見出した。大腸菌を対象としてこれらの事例を慎重に解析した結果、これらは配列決定上の単純な実験エラーではなく、本来の遺伝子が壊れた状態、すなわち「偽遺伝子」に違いないという結論に達した（図 3）。大腸菌ゲノム中にこのような偽遺伝子（ORF）が 100 個近く存在することが同定できた。その他、東工大有坂助教授のご協力を得て、T4 ファー

[1] http://spock.genes.nig.ac.jp/~genome/gtop.html

ジの未知タンパクの立体構造（13 個）、機能(3 個)の予測結果を吟味し、発表することができた。

4.4 題目別実施内容

(1) ゲノム上の全遺伝子自動同定システムの開発（総括担当者：西川 建）

初めに、発表されて間もない PSI-BLAST の性能が話題となりつつあったため、立体構造予測性能を 3D-1D 法と比較評価した。その結果、PSI-BLAST の採用を決め、微生物ゲノム、ヒト cDNA 配列などに対して、立体構造予測を中心に情報解析を行なった。構造予測できた ORF は全体の約 40%に達した。さらに、情報解析結果を生物科学的に分析し、T4 ファージの未知タンパク質の構造・機能予測、大腸菌ゲノム中の偽遺伝子の同定などの成果を得た。

(a)全遺伝子自動同定システムの開発（担当者： 太田元規、川端猛、福地佐斗志、伊藤武彦、落合孝正）

◆PSI-BLAST による立体構造データベースに対する検索性能と 3D-1D 法とを比較評価し、前者を立体構造予測プログラムとして採用。解析対象生物種の全遺伝子（ORF）を当プログラムにより立体構造データベース PDB に当て、有意な類似性を抽出

◆主に微生物の遺伝子（ORF）を解析対象とし、微生物以外に HUGO データベース（かずさ DNA 研究所で作成されているヒト cDNA 配列データベース）中の配列なども解析。最終的に 41 生物種を解析。構造予測できた ORF は全体の約 40%に達した。

◆立体構造予測の他、配列ホモロジー検索、モチーフ検索、膜貫通ヘリックス／シグナル配列予測等の解析を実施

(b)生物科学的分析（担当者：本間桂一、西村昭子、鈴木小夜子、中出晋介、松本典子、中島広志、山下紗代、小原収）

◆計算機による情報解析結果を検討し、生物科学的に興味深い事実を発見することを目的とする分析を実施

◆T4 ファージについて、一つずつの解析結果を検討し、機能未知であった ORF の構造・機能について示唆が得られ（東京工業大学の有坂助教授がご協力）、また大腸菌ゲノムの ORF の立体構造予測を詳細に分析し、偽遺伝子の存在を同定することができた。また、大腸菌の機能未知遺伝子の中で GTOP で新たに構造・機能予測されたものの分析とアノテーション付与なども行った。

(2) インターネットから各種データの自動収集モジュールの開発（総括担当者：市吉伸行、担当者： 吉成泰彦、落合孝正）

初めに、データ収集・解析のためのプロジェクトメンバーが共通に使える IMnet に接続された計算機環境を整備した。情報解析に必要な公開データベースは頻繁に更新されて情報量が増加しているため、GenBank、PDB、SwissProt、GenPept、PIR など収集対象となる公開データベースのデータ形式・更新方法に合わせ、定期的にデータを自動収

集し、マージする仕組みを作成した。

(3) 研究成果の公開（総括担当者：西川 建）

遺伝子(ORF)の各種情報解析結果を収めるデータ形式を決め、データを蓄積して行った。それらを属性値（遺伝子名、解析結果の値など様々）で検索できる機能を作成し、GTOP データベースとして Web 上に公開。種ごとの全遺伝子一覧表示、配列アライメントの詳細表示・図示、立体構造の表示などの機能も追加していった。また、利用者の手持ちアミノ酸配列に対する簡易的な立体構造予測サービスを提供した。

(a) 予測結果のデータベース化およびインターネットを通じた公開（担当者： 川端猛、福地佐斗志、伊藤武彦、荒木次郎）

◆各種の情報解析結果を解析対象遺伝子(ORF)毎「属性:値」対として統一して表現し、属性毎に値の条件（およびその論理結合）を指定した検索機能を実現

◆個々の生物種の全遺伝子一覧表示、遺伝子名検索等、様々な視点からの検索・一覧機能、染色体上の遺伝子位置、アライメント状態、類似検索でヒットした既知の立体構造などのグラフィカルな表示を提供

◆「GTOP データベース」としてインターネット上に公開

(b) 遺伝子領域／タンパク質立体構造予測システムのインターネットを通じた公開（担当者：川端猛、福地佐斗志、伊藤武彦、荒木次郎）

◆利用者の入力する配列に対する立体構造予測サービスとして PDB に対する BLAST 検索および SCOP に対する Reverse-PSI-BLAST 検索を提供

4.5 全体の総括と今後の課題

本研究プロジェクトでは、全期間を通して２，３カ月に１度の割合で研究協力者を含めた全プロジェクトメンバーの参加する打合せ会議を開いてきた。毎回用意された報告資料および議事録をもとに全体的な状況を総括する。振り返ってみると、全体としてプロジェクトはそれぞれ１年間ごとの３つの期間に大きく分けて捉えることができる。

最初の１年間（平成１０年１０月～１１年９月）はプロジェクト立上げの時期であり、とくに最初の数カ月は研究員の確保やコンピュータ機器の購入、計算機利用環境の整備など体制づくりに費やされた。具体的な研究内容を検討するための第１回目の打合せ会議を開いたのは平成１１年３月初めであった。ゲノム情報解析と結果のデータベース化という目標に向けてゼロからの出発であったため、やってみなければ判らない問題点が山積していた。最大の問題点は、情報解析の中心となるタンパク質の立体構造予測のための解析ツールとして当初予定していた「３Ｄ-１Ｄ法」が精度面、速度面で実用的に使えるかどうかという点だった。当時、発表されて間もない PSI-BLAST の威力が専門家の間で語られていた。そこで、それぞれの担当者を決め、当面は両ツールを併用することにして両者の予測結果を比較してみることにした。この問題は、PDB に新規に登録されたタンパク構造に対してテストしてみることにより、PSI-BLAST の優位性が明白となり、決着がついた。ゲノム解析の最初の対象として、細菌の代表格である大腸菌と枯草菌、古細菌のメタン菌、お

よびヒト cDNA 配列（HUGE データベース）を選んで解析を行った。その結果、PSI-BLAST によって構造予測される ORF（遺伝子）の割合はゲノム全体の４０％以上に達することが判り、目標達成に向けた見通しを得ることができた。解析結果は、ORF ごとにマスターファイルにまとめ、Web 上で検索／表示させるプロトタイプ版を作成した（Web 版は内部使用のみとした）。これにより、４つの所属機関に亘るプロジェクトメンバーはネットワークを通して同じ Web 画面を見ながら協同作業ができるようになった。

第２の時期（平成１１年１０月～１２年９月）はプロジェクトの成熟期だといえる。情報解析は立体構造予測のみにとどまらず、配列ホモロジー検索（PSI-BLAST、BLAST）、ファミリー分類（Pfam）、モチーフ検索（Prosite）、シグナル配列、膜貫通ヘリックス予測（SOSUI）などの既存ツール、および繰り返し配列の解析ツール（独自開発）を活用し、タンパク質レベルでのオールラウンドの解析ができるようにした。これら各種の解析結果がひと目でわかるように、Web 画面には「カラーバー表示」を取り入れた。この時期に解析の終了した生物種は１２種（真正細菌６、古細菌２、真核生物２、その他２）にまで拡大した。このうち真核生物は酵母（S. cerevisiae）と線虫（C. elegans）であるが、細菌に比べて遺伝子数が１桁上がる（線虫では約２万個）ので、計算時間が膨大になった（開始時に導入した中型サーバ機を使って、大腸菌で約１週間、酵母は１０日以上、線虫では約２カ月を要した）。この原因は主に PSI-BLAST と Pfam サーチの高い計算コストにあるが、この問題は後にコンピュータ機器を増強することによって軽減することができた。また、逆にＴ４ファージはゲノムサイズの小さい例（総遺伝子数２７５個）であるが、我々の解析結果をＴ４ファージの専門家（東工大・有坂文雄博士）に検討して貰ったところ、従来未知タンパク質とされていた ORF のうち立体構造が予測されたもの１３個、さらに機能まで予測されたもの３個という結果を得た。この結果は後に共同研究の成果として論文発表した。以上のように、解析結果は質量ともに充実してきたので、ＧＴＯＰデータベース（Genome TO Protein structure and function）と命名し、平成１２年９月に大々的に公開した。この公開にともない、全体的な解析結果をまとめた各種の統計情報を掲載したサマリーページ、利用者が配列データを入力すると PSI-BLAST による立体構造予測を行い結果を表示する予測サービス、なども用意した。また、関連する他の公開データベースとの間で相互リンクを張って、利用者の便宜を計ることにした。これまでに、遺伝子単位のリンクを張った相手ＤＢは、ＧＩＢ（遺伝研ＤＤＢＪ・ゲノム塩基配列）、ＰＥＣ（遺伝研・大腸菌ＤＢ）、GenoBase（奈良先端大・大腸菌ＤＢ）、ＨＵＧＥ（かずさＤＮＡ研・ヒト cDNA データ）である。

第３期（平成１２年１０月～１３年９月）は、ＧＴＯＰの内容のさらなる拡充とそれを利用した応用研究の時期にあたる。この時期に代表者研究機関（遺伝研）でコンピュータ機器の増強が行われ、自動解析の処理能力は約８倍ほどスピードアップした。これにより、ショウジョウバエ（D. melanogaster、遺伝子数約１万４千）を含めてＧＴＯＰに収録した生物種は４１種に拡大した。また、基本ＤＢである PDB や SwissProt も随時内容が増大しているので、新規の生物種のみにとどまらず全部について自動解析を再度行い、内容を更新した。その結果、立体構造予測される ORF の割合は PDB の増大に伴い１年当り約３%づつ上昇することがわかった。ＧＴＯＰを利用した応用研究として、大腸菌ゲノム中に１００個程度の偽遺伝子を発見することができたのは大きな成果である。それを可能にした要因

として、病原性大腸菌 0157 の全ゲノムが本年初めに発表され、そのデータをいち早く解析しＧＴＯＰに収録したこと、それにより標準株の大腸菌 K-12 と比較できたことが第１に挙げられる。また、偽遺伝子は立体構造で見るとドメイン構造の一部にしか対応しない場合が多いが、ＧＴＯＰでは既知構造にヒットした部分とそれ以外を色分けして表示するように工夫されていたため容易に識別できたこと。さらに、ある遺伝子からみてゲノム上で隣接する位置にある遺伝子を表示する NeighborGenes 機能や、ある遺伝子のホモログが他の生物種に存在するか否かを示す OrgPattern（ホモログ表示)、などもこの研究にとって不可欠であった。このように偽遺伝子探索の研究は文字通りＧＴＯＰに基づいた研究であったといえる。現在論文にまとめ投稿中である。

あらためて全体を顧みたとき、本プロジェクトの所期の目的は十分に達成できたと考えている。プロジェクト開始時点では、大量のデータを取扱うゲノム情報解析は巨大なテーマに思われ、３年間で２，３の細菌ゲノムの解析が完了すれば良い方ではないか、と考えていた。大量データの問題は、解析計算と結果の Web 表示の両工程を自動化することにより解決することができた。しかし何よりも、これだけの計算を処理するためのコンピュータ機器とデータベースとしてまとめ上げる専属の研究員が確保できたことが最大の要因であったと思う。それを可能にしていただいたＪＳＴによる支援に改めて感謝申し上げる次第である。

今後に残された課題は第１に、データベースは更新されない限り価値を失ってしまうので、今後とも更新･拡大を続けながらＧＴＯＰを維持していくことが重要だと考えている。

すでにシステムとしては出来上っているので、ＧＴＯＰの維持に必要な要員は研究員と実務者（ＳＥ）各一名で足りるであろう。ただ、実験的にゲノム配列が新たに決定される生物種は今後ますます増大する傾向にあるので、さらに高性能・大容量のコンピュータ資源が必要となると予想される。そのため将来的には、ＧＴＯＰは遺伝研・ＤＤＢＪに移管し、ＤＤＢＪ活動の一部として維持されるのが望ましいと考えている。第２の、内容的な課題としては、真核高等生物の遺伝子領域が不確定要素を多く含んでいる、という点が残されている。塩基配列データから遺伝子領域を決める問題は、（ＤＮＡレベルの問題なので）ＧＴＯＰの範囲外の問題であるが、不確定なデータに基づく解析結果をどのように取扱えばよいか、という問題点は残る。当面は、遺伝子領域のエキソン、イントロンの区別を情報としてＧＴＯＰに取り込むことが重要だと考えている。この情報と立体構造予測の情報を組み合せることにより、新たな応用研究への展開を期待している。

4.6 研究開発実施体制

代表研究者氏名　　西川　建

所属・役職　　国立遺伝学研究所生命情報・ＤＤＢＪ研究センター教授

(1) ゲノム上の全遺伝子自動同定システムの開発

氏名	所属	役職	研究開発項目
太田　元規	国立遺伝学研究所 生命情報・DDBJ 研究センター	助手	遺伝子産物同定システム研究開発
川端　猛 (H11.4~H12.6)	国立遺伝学研究所 生命情報・DDBJ 研究センター	JST 研究員	遺伝子産物同定システム研究開発
福地　佐斗志 (H11.8~H13.9)	国立遺伝学研究所 生命情報・DDBJ 研究センター	JST 研究員	遺伝子産物同定システム研究開発
本間　桂一 (H12.10~H13.9)	国立遺伝学研究所 生命情報・DDBJ 研究センター	JST 研究員	遺伝子産物同定システム研究開発 偽遺伝子の探索
西村　昭子	国立遺伝学研究所 系統生物研究センター	助教授	大腸菌遺伝子のデータ解析
鈴木　小夜子 (H11.4~H12.3)	国立遺伝学研究所 系統生物研究センター	JST 技術員	大腸菌遺伝子のデータ解析
中出　晋介 (H12.4~H13.5)	国立遺伝学研究所 系統生物研究センター	JST 技術員	大腸菌遺伝子のデータ解析
松本　典子 (H13.6~H13.9)	国立遺伝学研究所 系統生物研究センター	JST 技術員	大腸菌遺伝子のデータ解析
市吉　伸行	三菱総合研究所 フロンティア科学研究部	主席研究員	遺伝子産物同定システム研究開発
伊藤　武彦	三菱総合研究所 フロンティア科学研究部	研究員	遺伝子産物同定システム研究開発
落合　孝正	三菱総合研究所 フロンティア科学研究部	主任研究員	遺伝子産物同定システム研究開発
中島　広志	金沢大学医学部	教授	遺伝子産物同定システム研究開発 塩基組成によるゲノム配列解析
山下　紗代 (H11.4~H13.3)	金沢大学医学部	JST 技術員	塩基組成によるゲノム配列解析
小原　収	かずさ DNA 研究所 ヒト遺伝子研究部	部長	ヒト cDNA 配列データの解析

(2) インターネットから各種データの自動収集モジュールの開発

氏名	所属	役職	研究開発項目
市吉　伸行	三菱総合研究所 フロンティア科学研究部	主席研究員	自動収集システム開発
落合　孝正	三菱総合研究所 フロンティア科学研究部	主任研究員	自動収集システム開発
吉成　泰彦 (H10.10~H11.3)	三菱総合研究所 フロンティア科学研究部	研究員	自動収集システム開発

(3) 研究成果の公開

氏名	所属	役職	研究開発項目
川端　猛 (H11.4~H12.6)	国立遺伝学研究所 生命情報・DDBJ 研究センター	JST 研究員	GTOP データベースの構築
福地　佐斗志 (H11.8~H13.9)	国立遺伝学研究所 生命情報・DDBJ 研究センター	JST 研究員	GTOP データベースの構築
伊藤　武彦	三菱総合研究所 フロンティア科学研究部	研究員	GTOP データベースの構築
荒木　次郎 (H12.7~H13.9)	三菱総合研究所 フロンティア科学研究部	研究員	Web ユーザインタフェースの開発

4.7 本事業により得られた研究成果

(1)外部発表等

(a)原著論文

発表年	論文タイトル	掲載雑誌名 巻・号・ページ	著者名	整理番号
2000	The genomic DNA sequences of various species are distinctively distributed in nucleotide composition space	Res. Comm. Biochem. Cell Mol. Biol., 4, 25-45(2000)	Nakashima, H. and Nishikawa, K	12/10B-4 発 01
2000	ゲノム情報からの立体構造予測	生物物理（231号,307-308頁,2000)	西川建	12/10B-4 発 03
2000	Structural/functional assignment of unknown bacteriophage T4 proteins by iterative database searches	Gene, 259, 223-233(2000)	Kawabata, T., Arisaka, F. and Nishikawa, K.	12/10B-4 発 05
2001	Protein surface amino-acid composition distinctively differ between thermophilic and mesophilic bacteria	J. Mol. Biol., 309, 835-843 (2001)	Fukuchi, S. and Nishikawa, K.	12/10B-4 発 08
2002	GTOP: A database of protein structures predicted from genome sequences	Nucl. Acids Res., in press.	Kawabata, T., Fukuchi, S., Homma, K., Ota, M., Araki, J., Ito, T., Ichiyoshi, N. and Nishikawa, K.	13/10B-4 発 03

2001	蛋白質立体構造予測データベース GTOP	蛋白質核酸酵素,印刷中	川端猛、西川建	13/10B-4 発04
2001	The Escherichia coli genome contains a significant number of pseudogenes	J. Mol. Biol., (投稿中)	Homma, K., Fukuchi, S., Kawabata, T., Ota, M. and Nishikawa, K.	13/10B-4 発05

(b)口頭・ポスター発表

発表年月日 開催場所	発表タイトル	学会等の名称 (予稿集名、掲載ﾍﾟｰｼﾞ)	発表者	整理番号
1999/9/19-24	Oligonucleotide composition analysis of genomic sequence data	第１３回国際生物物理学会（ニューデリ，インド)	中島広志，西川建	11/10B-4 発01
1999/12/7-10	ゲノム塩基配列は生物種特有の塩基組成を持っている	第２２回日本分子生物学会（福岡ドーム）	中島広志，西川建	11/10B-4 発02
1999/12/7-10	マイコプラズマ２種間のオルソロガス遺伝子の解析	第２２回日本分子生物学会（福岡ドーム）	山下紗代，中島広志，西川建	11/10B-4 発03
2000/1/11-15	Protein structure comparison using Markov transition model of evolution	Quantitative challenges in the post-genomic sequence era: a workshop and symposium (サンディエゴ, USA)	川端猛，西川建	11/10B-4 発04
2000/ 2/ 4	ゲノムの機能構造予測 － 情報解析の展望	日本学術会議公開講演会（東京）	西川建	11/10B-4 発05
2000/8/5-8	Distinctive distributions of thermophilic /mesophilic bacterial proteins in the amino acid composition space	14th Symposium of the Protein Society (サンディエゴ,USA)	福地佐斗志，西川建	12/10B-4 発02
2000/9/13	全ゲノム立体構造予測データベース"GTOP"	生物物理学会 第３８回年会(東北大学)	川端猛，福地佐斗志，石井崇洋，太田元規，伊藤武彦，落合孝正，市吉伸行，西川建	12/10B-4 発04

2000/12/15	ゲノム情報からのタンパク質の立体構造／機能予測	分子生物学会，第２３回年会(神戸)	西川建	12/10B-4 発06
2001/2/6-7	全ゲノム立体構造予測データベース“GTOP”	第３回ワークショップ「微生物ゲノム研究のフロンティア(かずさアカデミアホール)	川端猛	12/10B-4 発07
2001/6/2	ゲノム蛋白質構造予測データベース（GTOP）とその応用	第１回蛋白質科学会年会（大阪大学）	西川建	13/10B-4 発01
2001/7/5	ゲノム情報からのタンパク質立体構造予測	第２８回生体分子科学討論会（金沢大学）	西川建	13/10B-4 発02

(2) 成果プログラム等

プログラム名称「繰り返し配列認識プログラム RepAlign」

機能概要：１つのタンパク質において、アミノ酸配列の内部に存在する繰り返し配列（自己相同性配列）を検出するプログラム。長周期の繰り返し配列はタンパク質のドメイン構造を知る上で重要である。また、一般的な相同性検索を行うさいに繰り返し配列が煩雑な結果を与えるので､事前に繰り返し配列の有無を知ることが必要になる。

使用言語：　Ｃ言語

サイズ：　ソースファイル、実行ファイルなどを含めて約 1Mbyte

(3) 特許出願記録

なし。

(4) 新聞記事、雑誌記事、テレビ報道等

なし。

(5) 受賞等

なし。

(6) ワークショップ等(主催分)

なし。