データベース構築促進及び技術開発に関する報告書

変異タンパク質配列データベースの構築

平成5年3月

財団法人　データベース振興センター
委託先　日本電子計算株式会社

本事業は、日本自転車振興会から競輪収益の一部である機械工業振興資金の補助を受けて作成したものである。

序

データベースは、わが国の情報化の進展上、重要な役割を果たすものと期待されている。今後、データベースの普及により、わが国において健全な高度情報化社会の形成が期待される。さらに海外に対して提供可能なデータベースの整備は、国際的な情報化への貢献および自由な情報流通の確保の観点からも必要である。しかしながら、現在わが国で流通しているデータベースの中でわが国独自のものは１／３にすぎないのが現状であり、わが国データベースサービスひいてはバランスある情報産業の健全な発展を図るためには、わが国独自のデータベースの構築およびデータベース関連技術の研究開発を強力に促進し、データベースの拡充を図る必要がある。

このような要請に応えるため、㈶データベース振興センターでは日本自転車振興会から機械工業振興資金の交付を受けて、データベースの構築および技術開発について民間企業、団体等に対して委託事業を実施している。委託事業の内容は、社会的、経済的、国際的に重要で、また地域および産業の発展の促進に寄与すると考えられているデータベースの構築とデータベース作成の効率化、流通の促進、利用の円滑化・容易化などに関係したソフトウェア技術・ハードウェア技術である。

本事業の推進に当って、当財団に学識経験者の方々で構成されるデータベース構築・技術開発促進委員会（委員長　山梨学院大学教授　蓼沼良一氏）を設置している。

この「変異タンパク質配列データベースの構築」は平成４年度のデータベースの構築促進および技術開発促進事業として、当財団が日本電子計算株式会社に対して委託実施した課題の一つである。この成果が、データベースに興味をお持ちの方々や諸分野の皆様方のお役に立てば幸いである。

なお、平成４年度データベースの構築促進および技術開発促進事業で実施した課題は次表のとおりである。

平成　５　年　３　月

財団法人　データベース振興センター

平成4年度　データベース構築・技術開発促進委託課題一覧

分野	課題名	委託先
社会	1 変異タンパク質配列データベースの構築	日本電子計算㈱
	2 新聞縮刷版見出しデータベースの構築	㈱朝日新聞社
	3 ファジィに関する文献データベースの構築	㈶日本情報処理開発協会
	4 医療用医薬品坑生物質データベースの構築	㈱小田島
	5 交通事故調査データベースの構築	㈱日本自動車研究所
	6 楽器データベースの構築	㈱ダイソメディアサービス
	7 人体計測データベースの構築	㈳人間生活工学研究センター
	8 大学におけるデータベース利用教育システムのプロトタイプ作成	日外アソシエーツ㈱
	9 先進複合材料データベースの構築	㈶次世代金属・複合材料研究開発協会
	10 博物館所蔵地図資料所在情報データベースの構築調査	㈶地図情報センター
中小企業振興地域活性化	11 地域流通最適化データベースのプロトタイプ作成	㈳日本ボランタリー・チェーン協会
	12 異分野研究のための知的オリエンテーション・データベースシステムのプロトタイプ作成	㈱けいはんな
	13 在宅勤務者サポート・データベースの構築調査	㈱志木サテライトオフィス・ビジネスセンター
海外	14 銅基複合材料日本特許英文データベースの構築	神鋼リサーチ㈱
	15 技術協力供与機材データベースのプロトタイプ作成	㈶日本国際協力システム
	16 先端産業分野における専門用語の電子辞書データベース化の調査研究	科学技術情報研究所㈱
	17 マーケティングコードの英文データベースの構築	㈱帝国データバンク
技術	18 安全研究における多重シソーラス・システム構築のための基本安全用語データベースの開発	㈱紀伊國屋書店
	19 3次元マッピングデータベースの技術開発	㈱日本総合技術研究所
	20 データベース検索サポートシステムの調査研究	セントラル開発㈱情報図書館 RUKIT
	21 グループウェアにおけるデータベースシステムに関する調査研究	㈱イフ・アドバタイジング
	22 パーソナルコンピュータとLANの利用による非定形データベースのプロトタイプ作成	㈱メイティック
	23 知的資源型データベースの調査研究	㈱ジャパンコミュニケーションズインスティテュート

目次

１．概要

１．１　目的

タンパク質は、生命活動の基盤となる作用物質である。この生体高分子はわずか２０種類のアミノ酸を素材とし、ペプチド結合という単一の共有結合によるアミノ酸相互の結合により出来ている。タンパク質のアミノ酸配列の決定は遺伝子工学的手法で容易になり、天然のタンパク質のアミノ酸配列データは、約１年半で２倍のペースで増加しており、既に約４０、０００件がデータベース化されている。近年、遺伝子工学の進歩により天然のタンパク質のアミノ酸配列の一部を変え、産業に有用な人工タンパク質の研究開発が進んでおり、２１世紀の主要技術の一つと目されるバイオテクノロジーの着実な果実となることが予測される。

研究者がアミノ酸配列を改変して物性や機能の改変を行おうとする時、現在、試行錯誤的に実験を行うか、過去の文献を調査し推定する等の方法で行われている。これには研究者の膨大な時間と手間が必要とされ、有用なデータベースの出現が求められている。

そこで今回、タンパク質改変実験の過去の報告を集積し知識体系化することを目的に、変異タンパク質配列データベースの構築を実施する。当データベースは、バイオテクノロジー分野の発展に大きく貢献できるとともに、世界的学術的貢献の一助として位置づけられることが期待される。

1．2　実施内容

(1) 学術文献の内容調査

変異タンパク質に関する学術文献について記述されている内容を調査した。

(2) データベースの内容の検討

学術文献の内容調査を踏まえ、データベースとして必要な情報（項目）の検討を行った。

(3) データ作成方法の検討

データをデータベースに効率的に取り込むために、データのチェック方法などを検討した。

(4) データの作成

データの作成方法に基づきデータを約10,000件作成した。

(5) データベースシステムの作成

検索・出力機能のほか、ユーザが作成したデータの取り込み、キーワードの作成などデータベースを作る一連のシステムを作成した。

(6) データベースの検索

データベースシステムが仕様どおり動くかどうかを調べるために、実際の検索で行われる例を想定し、検索・評価した。

2．　学術文献の内容調査

学術文献に記載されている変異タンパク質に関する記述内容がどのようなものであり、それをどのように整備するか検討した。

2．1　文献内容の概略

(1) 変異タンパク質

変異タンパク質は大別して２つのグループに分かれる。１つは蛋白工学的手法で作成された人工変異タンパク質であり、もう１つは遺伝的疾患や自然変異細菌などから古くから見い出されている自然変異タンパク質である。変異タンパク質を研究するうえで、タンパク質のアミノ酸配列の変化と、その結果得られるタンパク質の生物的活性、物理化学的性質の変化の２つが重要な要素と考えられている。この関係を研究するために、目標となる部位に変異を起こしている天然変異タンパク質を探すよりも、目標となる部位を最近の蛋白工学的手法により積極的に変換した人工変異タンパク質を研究している論文数の方が大勢を占めている。

(2) 変異タンパク質名

変異タンパク質が由来した天然タンパク質に対して、文献によっては異なる名称を用いている場合がある。個々の変異タンパク質を識別する名称は、置換前の天然タンパク質のアミノ酸残基名と置換後のアミノ酸残基名を使用して表わしている場合が多く見られるが、必ずしも統一されていない。

(3) アミノ酸残基番号

アミノ酸残基番号は、リボソームにより翻訳された段階（未成熟）のアミノ末端から振られている場合ばかりではなく、成熟した天然タンパク質のアミノ末端から振られている場合もあるし、類似の他の生物由来のタンパク質のアミノ酸残基番号を参照している場合もある。その上、アミノ酸配列の挿入や欠損のために、変異タンパク質と由来した天然タンパク質とで長さが変化した場合、アミノ末端から通し番号を振る場合以外に、変異を起こさせた場所だけ特別な番号を振り、他の部位は天然タンパク質の該当する番号を振っている場合なども見られ、アミノ酸残基の番号づけが統一されていない。

(4) 変異タンパク質作成方法

蛋白工学的手法の中には遺伝子工学的手法（部位特異的変異、発現ベクタなど）、物理的手法（紫外線、放射能など）、化学的手法（変異試薬など）など、様々な手法が用いられており、全ての変異タンパク質作成に共通する様な作成方法は見られない。また結果として同じアミノ酸配列を持つ変異タンパク質であっても、異なる方法で作成されている場合も見られる。

(5) 活性、性質測定条件

変異タンパク質を特に遺伝子工学的に変異させた場合には、大腸菌や酵母など増殖速度が速い微生物に変異タンパク質の遺伝子を組み込み、変異タンパク質を量産させ、最終的に必要な変異タンパク質のみを単離している場合が多い。このとき取り出した変異タンパク質の純度により、生物活性、物理化学的性質の測定値が大きく左右される。またこのときの不純物は、変異方法、発現系などにより変わってくる。また生物的活性や物理化学的性質を測定する際の溶媒の温度、塩濃度、pHなど実験条件が完全に一致している場合はほとんどない。

(6) 変異タンパク質作成数

変異タンパク質において、活性、性質がどのように変化したかは同じ条件下で測定される天然タンパク質の数値との比較で論じられる。そのため測定には天然タンパク質と変異タンパク質が用意される。変異タンパク質数は１つの場合も見られるが、通常同じ部位のアミノ酸残基を数種類のアミノ酸残基に置換したり、別々の部位を置換したりして、数種類の変異タンパク質を用意する。文献によっては１０種類以上の場合もある。

(7) 測定結果の表示方法

天然タンパク質や他の変異タンパク質との特性の違いを比較するために、また、基質、塩濃度、経過時間などに対する変化を見るために表形式で表示される場合が多い。このとき使用される数値は、測定値そのものである場合以外に、天然タンパク質の値を１とした相対値の場合も見られる。また時間経過による変化を記述する場合の様にグラフで示された結果も見られる。

２．２　整備の方針

(1) 記述方法の統一

文献の内容を調査した結果、変異タンパク質名、アミノ酸残基番号の振り方などについて統一されていないことが判明した。そこで整備の第一歩として、データベース上での記述方法の統一を図る。この際使用する基準は、文献の中で比較的多く利用されているものに準じる。また天然タンパク質のアミノ酸配列データベースとして、 PIR-International (Protein Information Resource-National Biomedical Research Foundation(PIR-NBRF), Japan International Protein Information Database(JIPID), Martinsried Institute for Protein Sequence (MIPS)) が作成し、世界的によく利用されているデータベースが存在するので、このデータベースでの記述方法も参考にする。

(2) 測定結果

文献では多くの場合、測定結果が表形式で記載されている。そこで本データベースでも表形式で測定データを表記する。文献にはグラフや図も掲載されているが、入力の負荷がかかること、デー

タ容量が大きいこと、システム開発に時間がかかることから、今回は、グラフ、図の収録は見合わせる。

（3）他のデータベースとのクロスリンク

変異タンパク質に関する論文の多くは、天然タンパク質と変異タンパク質との諸特性の違いについて論じている。したがって、本データベースでもこの点に重きを置くことにし、天然タンパク質に関する内容については、クロスリファレンスなどを記述し、他のデータベースを参照できるようにする。またクロスリンクする機能を持たせることにより、既存の天然タンパク質に関するデータベースと一緒に利用した場合には、直接該当する内容を引用して表示できるようにする。

3．　データベースの内容

前章の文献調査を踏まえて、エントリーの単位、作成するデータベースの項目、及び検索するためのキーワードを検討する。

3．1　エントリー単位の決定

エントリーの単位としては、

(1) 1天然タンパク質を1エントリー（同じ天然タンパク質から派生した変異タンパク質を全て1つのエントリーにまとめる）

(2) 1文献を1エントリー

(3) 1変異タンパク質を1エントリー

という3種類が考えられる。

(1)の場合、ある天然タンパク質についての全ての変異タンパク質の種類を知るのにはとても都合が良いが、測定された諸特性は論文毎に異なり、同じ特性を測定している場合でもその測定条件が一致していることはほとんどない。そのため測定結果を1つにまとめるのはあまり意味をなさず、逆に内容が多くなりすぎ見づらくなると予想される。

(2)の場合は、文献に記載されている範囲で変異タンパク質の諸特性を比較するのには都合がよい。また入力する場合でも、常に新規に登録することになるので、データチェックを行いやすい。しかし1つの変異タンパク質の性質のみを表示させたり、ある特定のアミノ酸の置換を起こしたタンパク質をリストアップする等、個別に変異タンパク質を検索したり表示するには適していない。

(3)は個々の変異タンパク質の諸性質を表示したり、いろいろな検索方法を行う場合には最も適している。しかし、データをこの単位で入力するのでは重複する内容がかなり多くなる。またこの単位での表示は他の変異タンパク質の諸特性との比較には不適当である。

以上の3種類のエントリー方法の長所、短所を検討した結果、検索方法の最も簡単な(3)を採用することに決定した。しかしながら、データ入力の段階では(2)を採用して入力にかかる手間を減らし、最終的なデータエントリーにする際にプログラムで(3)に変換することにした。またエントリーの内容を表示する方法として、各エントリー毎に表示する方法以外に、複数のエントリーの内容をまとめて、変異タンパク質同士の特性を比較できるような表示方法も採用した。

3．2　項目の決定

各変異タンパク質に対して、エントリーの内容を見ただけで変異タンパク質の諸性質の測定を追試できることが望ましい。そのために、各種実験条件について最低限必要であると思われる項目を設定する。さらに詳細な内容を知ることができるように、参照している文献についての情報も含める。データは国際的な学術論文から主に収集しており、各項目の内容として用いられている用語も英語のままであることが多い。そのため項目、内容とも英語で記述する。

(1) ACCESSION NUMBER

個々のエントリーを識別するためのユニークな番号。最初にデータを登録した際に決定される不変番号。アルファベットの大文字のMで始まり、その後に５桁の英数字が続く。

（例）M01021

(2) TITLE

変異タンパク質名と元になった天然タンパク質が由来する生物種名とを ハイフン " － " で結合してそのエントリーの TITLE とする。

変異タンパク質名　－　生物種名

（a）変異タンパク質名

由来した天然タンパク質名の後に、変異を付け足すことによって変異タンパク質名とする。

変異タンパク質名　＝　天然タンパク質名　変異

天然タンパク質名としては天然タンパク質のアミノ酸配列データベースを作成しているPIR-International が採用しているタンパク質名に準じる。

変異の例は以下の通りである。

(7) アミノ酸の部位特異的変異（あるアミノ酸から別のアミノ酸への置き換え）

天然タンパク質の変異を起こしたアミノ酸残基番号の前に天然タンパク質のアミノ酸を、残基番号の後に変異タンパク質のアミノ酸残基をそれぞれアミノ酸残基に対する１文字記号（表３－１）を用いて表示する。

（例）A91G

天然タンパク質の９１番目のアラニンがグリシンに置換している。

表　3－1　アミノ酸に対する1文字及び3文字記号

アミノ酸	1文字記号	3文字記号	アミノ酸	1文字記号	3文字記号
アラニン	A	Ala	メチオニン	M	Met
システイン	C	Cys	アスパラギン酸	N	Asn
アスパラギン	D	Asp	プロリン	P	Pro
グルタミン	E	Glu	グルタミン酸	Q	Gln
フェニルアラニン	F	Phe	アルギニン	R	Arg
グリシン	G	Gly	セリン	S	Ser
ヒスチジン	H	His	トレオニン	T	Thr
イソロイシン	I	Ile	バリン	V	Val
リジン	K	Lys	トリプトファン	W	Trp
ロイシン	L	Leu	チロシン	Y	Tyr

(イ) アミノ酸の挿入

小文字で add と記したあと挿入したアミノ酸配列を "[" "]" で括り、最後に挿入した直前の成熟した天然タンパク質のアミノ酸残基番号を記す。天然タンパク質のアミノ末端にアミノ酸を付け加えた場合には、アミノ酸残基番号は0になる。

(例) add[AGH]15

天然タンパク質の15番目と16番目のアミノ酸残基の間にアラニン、グリシン、ヒスチジンという3残基を挿入している。

(ウ) アミノ酸の欠損

小文字で del と記したあとに欠損した天然タンパク質のアミノ酸残基番号を "[" "]" で括る。アミノ酸が連続して欠損している場合には、欠損している最初のアミノ酸残基番号と最後のアミノ酸残基番号を "－" で結ぶ。

(例) del[21-30]

天然タンパク質の21番目のアミノ酸から30番目のアミノ酸に相当する配列が欠損している。

(エ) 融合（あるタンパク質と別のタンパク質を人工的に結合）

個々のタンパク質の融合した範囲の最初のアミノ酸残基番号と最後のアミノ酸番号を " － " で結び、それを "[" "]" で括り、これらを "=" で繋ぐ。この際元となる天然タンパ

ク質由来ではないアミノ酸配列の "[" の直前にその配列が由来したタンパク質名を書き加える。最後に全体を"(" ")" で括り、先頭に小文字で fusion と付け加える。

(例) fusion([1-152]=IgA1[102-108]=[1-152])

天然タンパク質の１番目から１５２番目の領域２つをIgA1というタンパク質の１０２番目から１０８番目の領域で結合した。

(オ) 上記 (ア)、(イ)、(ウ)、(エ) の組み合わせ

(例) add[AGH]15,del[21-30],A91G

(b) 生物種名

生物種名は、PIR-International が採用している生物種名に準じる。そのタンパク質をコードしている遺伝子が核以外の場合（ミトコンドリア、クロロプラストなど）や組織特異性がある場合にはその組織名も記載する。

(3) ALTERNATE NAME

変異タンパク質の元となったタンパク質に対する別名。別名が存在する場合にのみ記載する。

(4) REFERENCE

エントリーの内容が記載されている学術文献に関する情報を記載する。記載する内容は以下のものである。

(a) 著者名

名字 (family name) をフルスペリングし、その他の名前 (first name, middle name など) は最初の１文字のみを大文字で表記する。著者が複数の場合には、前の著者名の後に "," を打って続ける。この場合、最後の著者名の直前には " and " も付け加える。

(例) Hallewell, R. A., Laria, I., Tabrizi, A. Sousens, L. S., and Mullenbach, G. T.

(b) 雑誌名

学術雑誌名、巻数、掲載ページおよび発行年を記述する。学術文献名はPIR-Internationalが採用している学術文献名に準じる。

(例) J. Biol. Chem. 264, 5260-5268, 1989

(c) 表題

文献の表題を記載する。

(例) Title: Genetically engineered polymers of human CuZn superoxide dismutase.

(5) ORGANIZATION

機能を有する最小単位の成熟したタンパク質の構成を記述する。記述する内容は以下のものである。

（a）サブユニット数及びサブユニット名

変異タンパク質がサブユニットからなり、サブユニット名が付いている場合にはサブユニット名を記載する。また各サブユニット数をサブユニット名の直前に記載する。

（b）サブユニットに該当するアミノ酸配列

成熟したタンパク質に対するアミノ酸配列を、PIR-International のアミノ酸配列データベースの該当するエントリー番号及び該当する領域の最初のアミノ酸と最後のアミノ酸番号で記述する。

（例）2 alpha; 1-141<HAHU>, 2 beta; 1-146<HBHU>

上の例では、alpha サブユニット２つと beta サブユニット２つの計４つのサブユニットから機能を有するタンパク質が構成されている。alpha サブユニットは PIR-International の HAHU というエントリーの１番目から 141 番目までのアミノ酸配列に該当し、beta サブユニットは HBHU というエントリーの１番目から 146 番目までのアミノ酸配列に該当している。

(6) VARIATION

アミノ酸配列がどのように天然タンパク質のものと異なっているかを変異したサブユニットについて記述する。記述する内容は次のものである。

（a）変異タンパク質のアクセッション番号

（b）変異の記述

変異している部位の直前までは ORGANIZATION のように元のアミノ酸配列の領域を示し、"," に続けて変異タンパク質特有のアミノ酸配列を "'" で括って記述する。その後に元のアミノ酸配列と同じ領域を、元のアミノ酸残基番号を用いて記述する。アミノ酸配列が欠損している場合には、欠損直前の領域までと、欠損直後の領域を元のアミノ酸残基番号で記述する。変異を起こした部位が複数の場合には、最初の変異した部位までは先に述べた方法で記述し、続けて変異した部位の間に該当する元のアミノ酸配列の残基番号を記述する。これを最後まで繰り返す。

（例） 11-20,'ANR',24-30,'A',31-40,46-210<DEHUA>

上の変異タンパク質は DEHUA というエントリーの １１ 番目から２１０番目までの領域を基にして作成されている。変異タンパク質の11 番目から 13 番目までのアミノ酸配列はアラニン、アスパラギン酸、アルギニンに置き換わり、変異タンパク質の 21 番目にアラニンが挿

入され、元のアミノ酸配列の 41 番目から 45 番目のアミノ酸配列が欠損している。

（c）変異タンパク質名

TITLE に表示する変異のパターンを行末に "/" に続けて記載する。

（例） 2-6,'A',8-111,'S',113-118<DSHUCZ>/C7A,C112S

(7) SOURCE

変異タンパク質が由来した変異株やその変異タンパク質を産する細胞・組織を記載する。天然変異タンパク質の場合にのみ記載する。

（例） Escherichia coli K12 Ymel trpA34

Red cells from a patient displaying a genetically transmitted deficient enzyme

(8) METHOD

その変異タンパク質を作成した実験方法の一般名称を記述する。人工変異タンパク質の場合のみ記載する。

（例） Substitution : chemical synthesis of gene and mutagenesis

Fusion : two genes linked by human IgA1 hinge sequence

(9) EXPRESSION

変異タンパク質を大量発現させるために使用したプラスミド、ベクター、大腸菌など微生物の種類及び変異タンパク質の生産量を記載する。

（例） Host : Escherichia coli strain sodAsodB

Vector : pCl/1

Yield : About 40 percent of total soluble cell protein

(10) GROWTH RATE

変異タンパク質を大量発現させる際に利用した微生物が示す増殖量を、天然型のタンパク質を含む微生物の場合と比較して表形式に記載する。また表の下に表中で用いた記号等に対する注釈を記載する。

（例）

	Cell growth rate (mu/h)	
	(A)	(B)
No plasmid:	<0.11	0.15
Wild(clone):	0.11	0.17
M10091:	<0.11	0.10

---

mu = (logX2-logX1)t2-tlog2 (X; optical density at time t)

(A) in Minimal media 132

(B) in Minimal media 132 plus arginine and uracil

(11) PURIFICATION

変異タンパク質を単離するのに使用した実験器具、及び単離された変異タンパク質の純度を記載する。

(例) Chromatography: Ultrogel AcA-34, Pharmacia FPLC Mono-Q 10/10

Purity: >95 percent homology

(12) FUNCTION

天然タンパク質と変異タンパク質とで生物活性、及び物理化学的性質がどのように変化したかを表形式で記載する。またこの表の下に実験条件等をコメント形式で記載する。

(例) ATP synthesis

	Km(mM)		Vmax(micromol/min mg)	
Effector:	ornithine	no-ornithine	ornithine	no-ornithine
WILD (clone):	0.005	0.090	0.39	0.39
M10091:	0.008	0.080	0.27	0.33

Condition: pH 7.5, 25 C, variable ADP, carbamoyl phosphate (10 mM), magnesium ion (20 mM)

(13) COMMENT

変異タンパク質が由来した天然タンパク質についての一般的な注釈を記述する。

(例) The binding of GDP and phosphoenolpyruvate does not influence the activity.

(14) KEYWORD

タンパク質名以外で変異タンパク質を特徴付ける言葉がある場合にキーワードに記載する。

(例) Keywords: heterotropic interaction

(15) FEATURE TABLE

変異タンパク質のアミノ酸配列のうち、ある特徴を示す部位に対してそのアミノ酸残基番号（或いは領域）及びその部位が示す特徴を記載する。複数のサブユニットから成る変異タンパク質の場

合には、サブユニットごとに記載する。

（例）

Residues	Feature
19,59,145,169, 461	Binding site: carbohydrate (Asn) (potential)

(16) AMINO ACID SEQUENCE

変異タンパク質間で異なっているアミノ酸の部位はわずかなので、変異タンパク質ごとにアミノ酸配列を記述すると、ほとんど同じ配列を重複して持つこととなる。そこで、この変異タンパク質データベース自身にはアミノ酸配列を記述せず、天然のアミノ酸配列をPIR-Internationalのアミノ酸配列データベースから引用し、ORGANIZATIONとVARIATIONの記述に従って、変異タンパク質のアミノ酸配列をアミノ酸に対する1文字記号を用いて表示する。複数のサブユニットからなる変異タンパク質の場合には、サブユニットごとに表示する。

（例）

```
                5         10        15        20        25        30
  1 A T K A V A V L K G D G P V Q G I I N F E Q K E S N G P V K
 31 V W G S I K G L T E G L H G F H V H E F G D N T A G C T S A
 61 G P H F N P L S R K H G G P K D E E R H V G D L G N V T A D
 91 K D G V A D V S I E D S V I S L S G D H S I I G R T L V V H
121 E K A D D L G K G G N E E S T T K T G N A G S R L A C G V I
```

3．3　検索キーワード

（1）キーワードの取り出し方

キーワードの辞書ともいうべきシソーラスを作成し、それに基づいてキーワードを決める方法が望ましい。しかしその方法ではシソーラスを常に拡充しなければならないので、時間と手間がかかる。そこで今回は、入力内容の最終チェックを行う人間を1名に設定し、その人間の基準で用語の統一を図り、取り出しは機械処理だけで行うことにする。キーワードの取り出しは項目別とし、項目別のインデックスファイルを作成することにする。

（2）キーワードの決定

検索に慣れていない人でも簡単に検索できるように、その項目に現われる単語を、3文字単位で1文字分ずらしながら切り出したものを検索キーワードとして登録する。なおアミノ酸置換の検索用キーワードは、置換前のアミノ酸残基と置換後のアミノ酸残基の種類のみから作成し、何残基目のアミノ酸であるかには依存しないことにする。

4．データの作成

学術文献から人力で１つ１つデータを作成する。１つの文献で複数の変異タンパク質に対して測定が行われている場合、文献、実験方法等に関する情報は共通なので、入力の手間を省きタイプミスの可能性を減らすために、１文献を１エントリーとして通常のエディタを使用して入力し、データのチェック後、プログラムを用いて変異タンパク質ごとに分割する。このようにして作成されたデータの件数は、約10,000件である。

4．1　データの入力作業

(a) 学術文献のコピーの取り寄せ

予め各学術雑誌ごとに担当者を定めておき、変異タンパク質に関する論文を収集する。

(b) アクセッション番号の決定

エントリーごとにユニークなアクセッション番号を決定する。

(c) TITLE, REFERENCE の入力

文献の内容をあまり詳しく読まないでも入力できる TITLE 及び REFERENCE についてすぐに入力を行う。

(d) TITLE, REFERENCE 以外の入力

専門の知識を有する人が論文を読みその内容にしたがって、TITLE、REFERENCE 以外の内容を入力する。

(e) 入力内容のチェック

データを入力していない専門知識を有する人が、プリントアウトされた入力内容と文献の内容との比較、チェックを行う。もし入力ミス等がある場合には (d) に戻す。

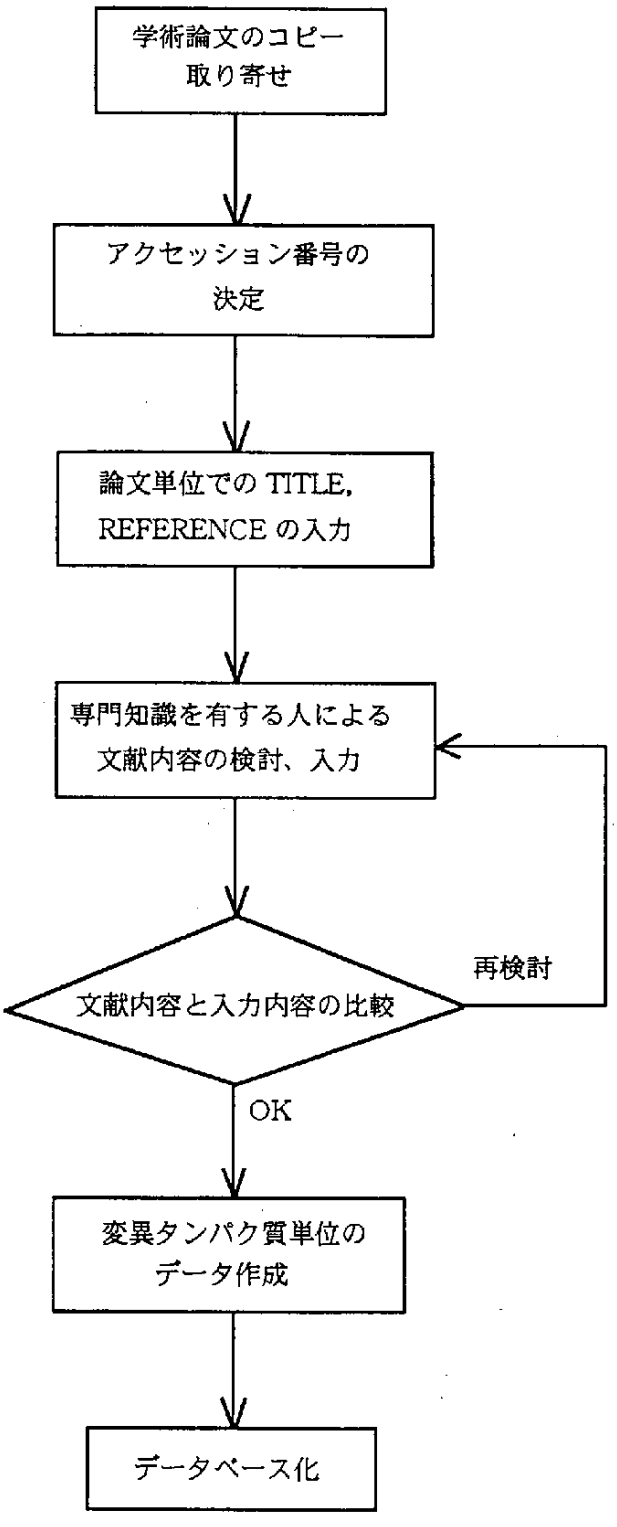


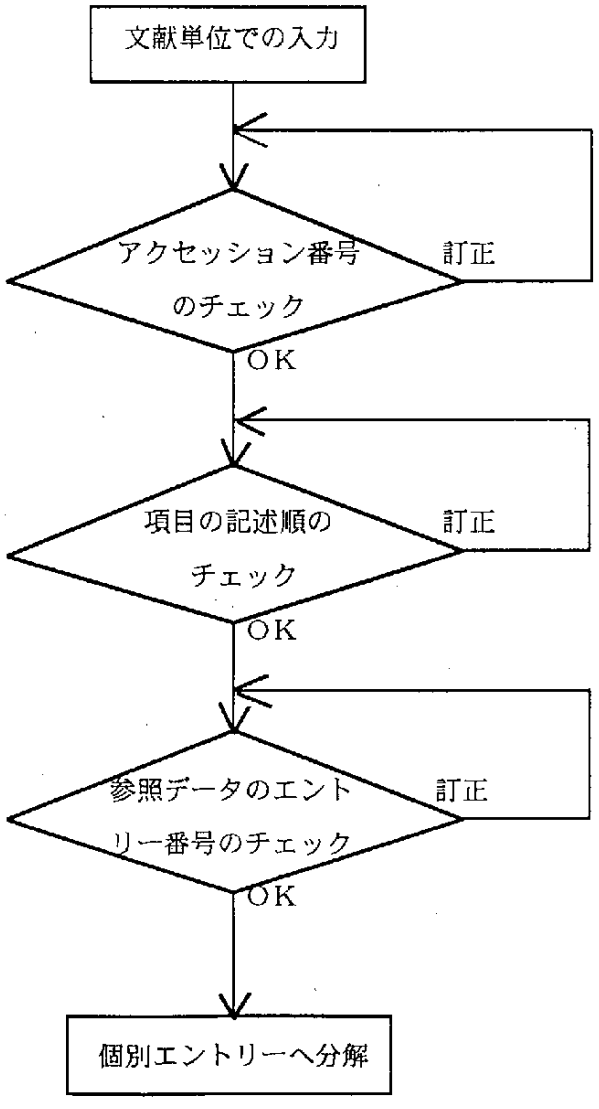
図４－１　データ入力の流れ

4．2　データ作成支援システム

人間が通常のエディタを使用して入力する方法では、タイプミスが避けられない。そこで、タイプされたデータを文献別に入力されている段階でワークステーションでチェックする。

チェックはまず変異タンパク質のアクセッション番号に考えられない番号が付いていないか、アクセッション番号の桁数はあっているかをチェックする。つぎに各項目ごとに、その項目の記述順序に間違いがないかをチェックする。例えば、REFERENCEの先頭の行には著者名が、2行目には雑誌名、3行目には論文名が入るがこの順番は合っているかをチェックする。またPIR-Internationalのアミノ酸配列データベースの該当するエントリー番号の検索を行い、該当エントリー番号にミスがないかのチェックも行う。この時点ではデータは通常のテキストファイルなので、エラーを発見したらその場で修正を行う。これをエラーがなくなるまで行う。

文献単位のエントリーのチェックが終了したら、それを分割プログラムを利用して1変異タンパク質ごとのエントリーに分割する。



5．　データベースシステムの作成

第３章の「データベースの内容」を踏まえて、それを実現するためのデータベースシステムの作成を行う。図５－１にデータベースシステムの構成を示す。このうち、データベースシステムの特徴となる検索及び出力機能、検索インデックス作成機能について詳細に検討する。

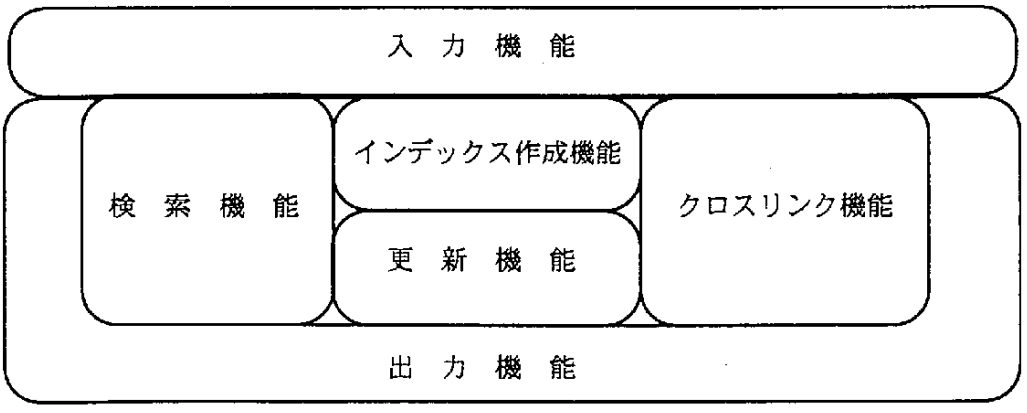


図５－１　データベースシステムの構成

５．１　検索・出力機能

TITLE による検索の場合、変異タンパク質の表わし方が著者毎に異なっている場合が多く、原論文の変異タンパク質名と TITLE 中の変異タンパク質名が異なり、いきなり特定の変異タンパク質を指定するのは困難である。そこで、その変異タンパク質の元になったタンパク質名やタンパク質を産する生物種名での検索ができるようにする。

さらに、似たような働きをするタンパク質について同時に検索する場合もあるので、タンパク質名の一部（３文字以上）の一致でも検索できるように中間一致検索を採用する。

変異タンパク質の検索としては、アミノ酸の置換パターンによる検索も考えられる。これはアミノ酸個々の性質に重きを置いた検索方法であり、タンパク質の種類によらないアミノ酸の置換とその影響の一般則を探る場合である。そのため、あるアミノ酸から別の特定のアミノ酸への置換パターンでの検索を可能にしている。

TITLE部分以外の項目のうち、著者名、雑誌名、METHOD、EXPRESSION、PURIFICATION、FUNCTION、KEYWORD、FEATURE の各項目内での検索もできる機能を組み入れる。これらの項目では統制のない自由な語が使われているので、最低３文字一致した項目を含むエントリーを検索できるようにする。

検索の結果が妥当なものであるかがすぐに確認できるように、該当するエントリーのアクセッション番号と共に TITLE を表示する機能を組み込む。また検索結果はカレントリストに登録させる。

出力は通常データベース内の全項目を出力する場合が多いと思われる。しかし、変異タンパク質

について記載されている原論文に関する情報のみが必要であるなど、検索した内容の一部にしか興味がない場合もあるはずである。そのため、出力は基本的には全項目が出力されるようにしておき、オプション機能として指定した項目のみを表示できる機能を組み入れる。

変異タンパク質の生物活性・物理化学的性質は、天然タンパク質ばかりではなく他の変異タンパク質と比較して表示したほうが便利な場合もある。また多くの文献では、１つのタンパク質に対して複数の変異タンパク質を作成してその諸性質を比較している。そのため、個々のエントリーの内容を順番に表示する以外に、各項目毎にまとめて表示できる機能も組み込む。

以下にデータベースシステムのコマンドを示す。各コマンド名は省略可能であり、"*" の後の部分が省略できる。

（例）q*uit は uit の部分が省略でき、q, qu, qui, quit の４つのいづれでも構わないことを示している。

（１）検索開始コマンド

コマンド　topiqus

【機能】　データベースの利用を開始する。

【書式】　topiqus

（２）検索終了コマンド

コマンド　q*uit

【機能】　データベースの利用を終了する。

【書式】　quit

（３）TITLE 検索コマンド

コマンド　fin*d

【機能】　与えられた検索式でTITLE部分を検索し、ヒット数、ヒットしたエントリーのアクセッション番号及びTITLEを表示する。また検索結果をカレントリストに登録する。

【書式】

find[|オプション]　検索式

（a）[ ] 内は省略できる。

（b）|オプション

以下の6種類が用意されている。(ｱ)、(ｲ)、(ｳ)、(ｴ)同士の組み合わせは不可能であるが、これらのうちの1つと(ｵ)および(ｶ)、または(ｵ)あるいは(ｶ)との組み合わせは可能である。

(ｱ) |current

検索直前のカレントリスト内に検索範囲を限定して検索を行い、検索結果を新しいカレントリストとする。

(ｲ) |add

検索実行直前のカレントリストに検索の結果得られたリストを追加し、新しいカレントリストとする。

(ｳ) |subtract

検索実行直前のカレントリストから検索の結果得られたリストを削除し、新しいカレントリストとする。

(ｴ) |keep

検索の結果得られたリストを表示するが、カレントリストは検索以前のままに保つ。

(ｵ) |brief

ヒット数及びヒットしたエントリーのアクセッション番号のみを表示し、TITLEの表示は行わない。

(ｶ) |out=ユーザファイル名

ディスプレイへの表示を行う代わりに、検索結果を指定したユーザファイルに書き込む。

（c）検索式

3文字以上からなる語を1つ以上並べたものである。複数並べる場合には語と語の間に1つ以上の空白を入れる。複数の語を並べると、並べられた語全てを含むTITLEを持つエントリーのみがヒットされる。

（4）アミノ酸置換検索コマンド

コマンド　　var*iation

【機能】　与えられた検索式に合致するアミノ酸置換をした変異タンパク質を検索し、ヒット数、ヒットしたエントリーのアクセッション番号及びTITLEを表示する。また検索結果をカレントリストに登録する。

【書式】

variation[|オプション]　検索式

（a）［ ］内は省略できる。

（b）|オプション

以下の６種類が用意されている。(ｱ)、(ｲ)、(ｳ)、(ｴ)同士の組み合わせは不可能であるが、これらのうちの１つと(ｵ)および(ｶ)、または(ｵ)あるいは(ｶ)との組み合わせは可能である。

(ｱ) |current

検索直前のカレントリスト内に検索範囲を限定して検索を行い、検索結果を新しいカレントリストとする。

(ｲ) |add

検索実行直前のカレントリストに検索の結果得られたリストを追加し、新しいカレントリストとする。

(ｳ) |subtract

検索実行直前のカレントリストから検索の結果得られたリストを削除し、新しいカレントリストとする。

(ｴ) |keep

検索の結果得られたリストを表示するが、カレントリストは検索以前のままに保つ。

(ｵ) |brief

ヒット数及びヒットしたエントリーのアクセッション番号のみを表示し、TITLEの表示は行わない。

(ｶ) |out=ユーザファイル名

ディスプレイへの表示を行う代わりに、検索結果を指定したユーザファイルに書き込む。

（c）検索式

置換前のアミノ酸及び置換後のアミノ酸をそれぞれアミノ酸に対する１文字記号で表わしそれをハイフンでつなぐ。この置換パターンを複数並べる時は間に１つ以上の空白を入れる。複数の置換パターンを並べた場合には、全ての置換パターンに合致したエントリーのみがヒットされる。

（５）著者名検索コマンド

コマンド　aut*hor

【機能】　与えられた検索式に合致する著者名を検索し、ヒット数、ヒットしたエントリーのアクセッション番号及びTITLEを表示する。また検索結果をカレントリストに登録する。

【書式】

author[|オプション]　検索式

（a）[ ]内は省略できる。

（b）|オプション

以下の6種類が用意されている。(ｱ)、(ｲ)、(ｳ)、(ｴ)同士の組み合わせは不可能であるが、これらのうちの1つと(ｵ)および(ｶ)、または(ｵ)あるいは(ｶ)との組み合わせは可能である。

(ｱ) |current

検索直前のカレントリスト内に検索範囲を限定して検索を行い、検索結果を新しいカレントリストとする。

(ｲ) |add

検索実行直前のカレントリストに検索の結果得られたリストを追加し、新しいカレントリストとする。

(ｳ) |subtract

検索実行直前のカレントリストから検索の結果得られたリストを削除し、新しいカレントリストとする。

(ｴ) |keep

検索の結果得られたリストを表示するが、カレントリストは検索以前のままに保つ。

(ｵ) |brief

ヒット数及びヒットしたエントリーのアクセッション番号のみを表示し、TITLEの表示は行わない。

(ｶ) |out=ユーザファイル名

ディスプレイへの表示を行う代わりに、検索結果を指定したユーザファイルに書き込む。

（c）検索式

3文字以上からなる語を1つ以上並べたものである。複数並べる場合には語と語の間に1つ以上の空白を入れる。複数の語を並べると、並べられた語全てを著者名の中に持つエントリーのみがヒットされる。

（6）雑誌名検索コマンド

コマンド　jou*rnal

【機能】　与えられた検索式に合致する雑誌名を検索し、ヒット数、ヒットしたエントリーのアクセッション番号及びTITLEを表示する。また検索結果をカレントリストに登録

する。

【書式】

|ournal[|オプション]　検索式

（a）[ ]内は省略できる。

（b）|オプション

以下の6種類が用意されている。(ｱ)、(ｲ)、(ｳ)、(ｴ)同士の組み合わせは不可能であるが、これらのうちの1つと(ｵ)および(ｶ)、または(ｵ)あるいは(ｶ)との組み合わせは可能である。

(ｱ) |current

検索直前のカレントリスト内に検索範囲を限定して検索を行い、検索結果を新しいカレントリストとする。

(ｲ) |add

検索実行直前のカレントリストに検索の結果得られたリストを追加し、新しいカレントリストとする。

(ｳ) |subtract

検索実行直前のカレントリストから検索の結果得られたリストを削除し、新しいカレントリストとする。

(ｴ) |keep

検索の結果得られたリストを表示するが、カレントリストは検索以前のままに保つ。

(ｵ) |brief

ヒット数及びヒットしたエントリーのアクセッション番号のみを表示し、TITLEの表示は行わない。

(ｶ) |out=ユーザファイル名

ディスプレイへの表示を行う代わりに、検索結果を指定したユーザファイルに書き込む。

（c）検索式

3文字以上からなる語を1つ以上並べたものである。複数並べる場合には語と語の間に1つ以上の空白を入れる。複数の語を並べると，並べられた語全てを雑誌名の中に持つエントリーのみがヒットされる。

（7）METHOD 検索コマンド

コマンド　　met*hod

【機能】　与えられた検索式に合致するmethodを検索し、ヒット数、ヒットしたエントリー

のアクセッション番号及びTITLEを表示する。また検索結果をカレントリストに登録する。

【書式】

method[|オプション]　検索式

（a）［ ］内は省略できる。

（b）|オプション

以下の6種類が用意されている。(ｱ)、(ｲ)、(ｳ)、(ｴ)同士の組み合わせは不可能であるが、これらのうちの1つと(ｵ)および(ｶ)、または(ｵ)あるいは(ｶ)との組み合わせは可能である。

(ｱ) |current

検索直前のカレントリスト内に検索範囲を限定して検索を行い、検索結果を新しいカレントリストとする。

(ｲ) |add

検索実行直前のカレントリストに検索の結果得られたリストを追加し、新しいカレントリストとする。

(ｳ) |subtract

検索実行直前のカレントリストから検索の結果得られたリストを削除し、新しいカレントリストとする。

(ｴ) |keep

検索の結果得られたリストを表示するが、カレントリストは検索以前のままに保つ。

(ｵ) |brief

ヒット数及びヒットしたエントリーのアクセッション番号のみを表示し、TITLEの表示は行わない。

(ｶ) |out=ユーザファイル名

ディスプレイへの表示を行う代わりに、検索結果を指定したユーザファイルに書き込む。

（c）検索式

3文字以上からなる語を1つ以上並べたものである。複数並べる場合には語と語の間に1つ以上の空白を入れる。複数の語を並べると、並べられた語全てをmethodの中に持つエントリーのみがヒットされる。

(8) EXPRESSION 検索コマンド

コマンド　exp*ression

【機能】　与えられた検索式に合致する expression を検索し、ヒット数、ヒットしたエントリーのアクセッション番号及び TITLE を表示する。また検索結果をカレントリストに登録する。

【書式】

expression[|オプション]　検索式

（a）[ ] 内は省略できる。

（b）|オプション

以下の 6 種類が用意されている。(ア)、(イ)、(ウ)、(エ) 同士の組み合わせは不可能であるが、これらのうちの 1 つと(オ) および (カ)、または (オ) あるいは (カ) との組み合わせは可能である。

(ア) |current

検索直前のカレントリスト内に検索範囲を限定して検索を行い、検索結果を新しいカレントリストとする。

(イ) |add

検索実行直前のカレントリストに検索の結果得られたリストを追加し、新しいカレントリストとする。

(ウ) |subtract

検索実行直前のカレントリストから検索の結果得られたリストを削除し、新しいカレントリストとする。

(エ) |keep

検索の結果得られたリストを表示するが、カレントリストは検索以前のままに保つ。

(オ) |brief

ヒット数及びヒットしたエントリーのアクセッション番号のみを表示し、TITLE の表示は行わない。

(カ) |out=ユーザファイル名

ディスプレイへの表示を行う代わりに、検索結果を指定したユーザファイルに書き込む。

（c）検索式

3 文字以上からなる語を 1 つ以上並べたものである。複数並べる場合には語と語の間に 1 つ以上の空白を入れる。複数の語を並べると、並べられた語全てを expression の中に持つエントリーのみがヒットされる。

(9) PURIFICATION 検索コマンド

コマンド　pur*ification

【機能】　与えられた検索式に合致するpurificationを検索し、ヒット数、ヒットしたエントリーのアクセッション番号及びTITLEを表示する。また検索結果をカレントリストに登録する。

【書式】

purification[|オプション]　検索式

（a）[ ]内は省略できる。

（b）|オプション

以下の6種類が用意されている。(ｱ)、(ｲ)、(ｳ)、(ｴ)同士の組み合わせは不可能であるが、これらのうちの1つと(ｵ)および(ｶ)、または(ｵ)あるいは(ｶ)との組み合わせは可能である。

(ｱ) |current

検索直前のカレントリスト内に検索範囲を限定して検索を行い、検索結果を新しいカレントリストとする。

(ｲ) |add

検索実行直前のカレントリストに検索の結果得られたリストを追加し、新しいカレントリストとする。

(ｳ) |subtract

検索実行直前のカレントリストから検索の結果得られたリストを削除し、新しいカレントリストとする。

(ｴ) |keep

検索の結果得られたリストを表示するが、カレントリストは検索以前のままに保つ。

(ｵ) |brief

ヒット数及びヒットしたエントリーのアクセッション番号のみを表示し、TITLEの表示は行わない。

(ｶ) |out=ユーザファイル名

ディスプレイへの表示を行う代わりに、検索結果を指定したユーザファイルに書き込む。

（c）検索式

3文字以上からなる語を1つ以上並べたものである。複数並べる場合には語と語の間に1つ以上の空白を入れる。複数の語を並べると、並べられた語全てをpurificationの中に持つエントリーのみがヒットされる。

(10) FUNCTION 検索コマンド

コマンド　fun*ction

【機能】　与えられた検索式に合致するfunctionを検索し、ヒット数、ヒットしたエントリーのアクセッション番号及びTITLEを表示する。また検索結果をカレントリストに登録する。

【書式】

function[|オプション]　検索式

（a）[ ]内は省略できる。

（b）|オプション

以下の6種類が用意されている。(ｱ)、(ｲ)、(ｳ)、(ｴ)同士の組み合わせは不可能であるが、これらのうちの1つと(ｵ)および(ｶ)、または(ｵ)あるいは(ｶ)との組み合わせは可能である。

(ｱ) |current

検索直前のカレントリスト内に検索範囲を限定して検索を行い、検索結果を新しいカレントリストとする。

(ｲ) |add

検索実行直前のカレントリストに検索の結果得られたリストを追加し、新しいカレントリストとする。

(ｳ) |subtract

検索実行直前のカレントリストから検索の結果得られたリストを削除し、新しいカレントリストとする。

(ｴ) |keep

検索の結果得られたリストを表示するが、カレントリストは検索以前のままに保つ。

(ｵ) |brief

ヒット数及びヒットしたエントリーのアクセッション番号のみを表示し、TITLEの表示は行わない。

(ｶ) |out=ユーザファイル名

ディスプレイへの表示を行う代わりに、検索結果を指定したユーザファイルに書き込む。

（c）検索式

3文字以上からなる語を1つ以上並べたものである。複数並べる場合には語と語の間に1つ以上の空白を入れる。複数の語を並べると、並べられた語全てをfunctionの中に持つエントリーのみがヒットされる。

(11) KEYWORD 検索コマンド

コマンド　key*word

【機能】　与えられた検索式に合致するkeywordを検索し、ヒット数、ヒットしたエントリーのアクセッション番号及びTITLEを表示する。また検索結果をカレントリストに登録する。

【書式】

keyword[|オプション]　検索式

（a）[ ]内は省略できる。

（b）|オプション

以下の6種類が用意されている。(ｱ)、(ｲ)、(ｳ)、(ｴ)同士の組み合わせは不可能であるが、これらのうちの1つと(ｵ)および(ｶ)、または(ｵ)あるいは(ｶ)との組み合わせは可能である。

(ｱ) |current

検索直前のカレントリスト内に検索範囲を限定して検索を行い、検索結果を新しいカレントリストとする。

(ｲ) |add

検索実行直前のカレントリストに検索の結果得られたリストを追加し、新しいカレントリストとする。

(ｳ) |subtract

検索実行直前のカレントリストから検索の結果得られたリストを削除し、新しいカレントリストとする。

(ｴ) |keep

検索の結果得られたリストを表示するが、カレントリストは検索以前のままに保つ。

(ｵ) |brief

ヒット数及びヒットしたエントリーのアクセッション番号のみを表示し、TITLEの表示は行わない。

(ｶ) |out=ユーザファイル名

ディスプレイへの表示を行う代わりに、検索結果を指定したユーザファイルに書き込む。

（c）検索式

3文字以上からなる語を1つ以上並べたものである。複数並べる場合には語と語の間に1つ以上の空白を入れる。複数の語を並べると，並べられた語全てをkeywordの中に持つエントリーのみがヒットされる。

(12) FEATURE 検索コマンド

コマンド　fea*ture

【機能】　与えられた検索式に合致するfeatureを検索し、ヒット数、ヒットしたエントリーのアクセッション番号及びTITLEを表示する。また検索結果をカレントリストに登録する。

【書式】

feature[|オプション]　検索式

(a) [ ] 内は省略できる。

(b) |オプション

以下の6種類が用意されている。(ｱ)、(ｲ)、(ｳ)、(ｴ) 同士の組み合わせは不可能であるが、これらのうちの1つと(ｵ) および (ｶ)、または (ｵ) あるいは (ｶ) との組み合わせは可能である。

(ｱ) |current

検索直前のカレントリスト内に検索範囲を限定して検索を行い、検索結果を新しいカレントリストとする。

(ｲ) |add

検索実行直前のカレントリストに検索の結果得られたリストを追加し、新しいカレントリストとする。

(ｳ) |subtract

検索実行直前のカレントリストから検索の結果得られたリストを削除し、新しいカレントリストとする。

(ｴ) |keep

検索の結果得られたリストを表示するが、カレントリストは検索以前のままに保つ。

(ｵ) |brief

ヒット数及びヒットしたエントリーのアクセッション番号のみを表示し、TITLEの表示は行わない。

(ｶ) |out=ユーザファイル名

ディスプレイへの表示を行う代わりに、検索結果を指定したユーザファイルに書き込む。

(c) 検索式

3文字以上からなる語を1つ以上並べたものである。複数並べる場合には語と語の間に1つ以上の空白を入れる。複数の語を並べると並べられた語全てをfeatureの中に持つエントリーのみがヒットされる。

(13) 検索結果一覧コマンド

コマンド　lis*t

【機能】　カレントリストに登録されているエントリーのアクセッション番号とTITLEを表示する。また現在のカレントリストの直前のカレントリストを復活させる。

【書式】

list[|オプション]

(a) [ ] 内は省略できる。

(b) |オプション

以下の4種類が用意されている。(ｲ) と他のオプションとの組み合わせは不可能であるが、(ｲ) 以外の3つのオプションの組み合わせは自由である。

(ｱ) |all

データベースに登録されている全エントリーのアクセッション番号とTITLEを表示する。

(ｲ) |restore

現在のカレントリストの直前のカレントリストを復活させる。

(ｳ) |code

カレントリストに登録されているエントリーのアクセッション番号のみを表示する。

(ｴ) |out=ユーザファイル名

ディスプレイへの表示を行う代わりに、検索結果を指定したユーザファイルに書き込む。

(14) シングル表示コマンド

コマンド　typ*e

【機能】　個々のエントリーの内容をディスプレイ表示する。またオプションによりユーザファイルに出力する。

【書式】

type[|オプション]　アクセッション番号

(a) [ ] 内は省略できる。

(b) |オプション

以下の6種類が用意されている。(ｳ)、(ｴ)、(ｵ) の組み合わせは不可能であるが、これらのうちの1つと (ｱ) および (ｲ) 、または (ｱ) あるいは (ｲ) との組み合わせは可能である。

(ｱ) |current

カレントリストに登録されている全エントリーの内容をエントリー順に表示する。この場合アクセッション番号を入力する必要はない。

(ｲ) |out=ユーザファイル名

ディスプレイへの表示を行う代わりに、検索結果を指定したユーザファイルに書き込む。

(ｳ) |sequence

エントリーの TITLE とアミノ酸配列のみを表示する。

(ｴ) |text

アミノ酸配列以外の内容を表示する。

(ｵ) |reference

エントリーの TITLE と REFERENCE のみを表示する。

(ｶ) |three_letter

通常アミノ酸配列の表示は１文字記号を使用するが、この代わりに３文字記号を使用する。アミノ酸配列を表示する場合にのみ有効。

(c) アクセッション番号

データベースに登録されているエントリーの内容を表示させる場合にはそのままアクセッション番号を入力する。ユーザファイルに保存されているエントリーを表示する場合には、

アクセッション番号＝ユーザファイル名

のように表示したい内容を含むユーザファイル名を指定する。このときのアクセッション番号はユーザファイルの中で使用しているアクセッション番号である。なおユーザファイルはデータベースシステムの copy コマンドで作られたものか、以下に示すユーザファイルフォーマットに従ったものでなければならない。

【ユーザファイルフォーマット】

">VA;" アクセッション番号　　注１

変異タンパク質名 " - " 生物種名

["N;Alternate names: " 別名 [";" 別名 ...]]　注２

"R; " 著者名 [{ ", " | ", and " } 著者名 ...]　注３

文献名

["A;Title: " 文献表題]

["A;" 文献中で特に触れるべきコメント]

"C;VARIATION"

"C;VARIATION"

"V;" code ":" 変異パターン "/" 変異の略省 "<" 天然タンパク質のアクセッション番号 ">"

["C;SOURCE:" 天然変異タンパク質の由来元]

["C;METHOD"]

["W;#Subtitution: " 置換方法]

["W;#Fusion: " 融合方法]

["C;EXPRESSION"]

["W;#Host: " 宿主名 [";" 宿主名 ...]]

["W;#Virus: " 培養に用いたウイルス名]

["W; {#Vector: | #Plasmid: } " 組み替え DNA を挿入した核外遺伝子]

["W;#Yield: " 生産された変異タンパク質の量]

["C; GROWTH-RATE"]

["W;----------"] 注4

["W;" 測定項目、条件]

["W;----------"]

["W;" 天然タンパク質の測定値]

["W;" 変異タンパク質の測定値]

["W;----------"]

["A;" 測定方法、表の値などに関する注意書き]

["C;PURIFICATION"]

["W;#Chromatography: " 変異タンパク質回収装置名]

["W;#Purity:" 回収した変異タンパク質の純度]

"C;FUNCTION"

"C;" 活性検査名

"W;----------"

"W;" 測定項目、条件

"W;----------"

"W;" 天然タンパク質の測定値

"W;" 変異タンパク質の測定値

"W;----------"

["A;" 測定方法、表の値などに関する注意書き]

["C;COMMENT: " コメント]

["C;KEYWORDS: " キーワード ["; " キーワード]

["F;" アミノ酸残基番号 "/" 配列部分の持つ特徴の分類名 ":" [配列部分の持つ特徴の詳細]]

["Z;<" クロスリファレンス ">"]

注１　" "で囲まれた内容を入力

注２　[ ]で囲まれた内容は省略可能

注３　{A | B}はＡまたはＢのどちらか一方を選択

注４　テーブルの外枠及び間仕切りを表している。ハイフンの数は任意

(15) マルチ表示コマンド

コマンド　　mty*pe

【機能】　元になる天然タンパク質が共通のエントリーを項目ごとにまとめて表示する。

【書式】

mtype[|オプション]　タンパク質名

（ａ）[ ]内は省略できる。

（ｂ）|オプション

以下の２種類が用意されている。(ｱ) と (ｲ) の組み合わせは可能である。

(ｱ) |out=ユーザファイル名

ディスプレイへの表示を行う代わりに、検索結果を指定したユーザファイルに書き込む。

(ｲ) |reference

エントリーの TITLE と REFERENCE のみを表示する。

（ｃ）タンパク質名

３文字以上からなる語を１つ以上並べたものである。複数並べる場合には語と語の間に１つ以上の空白を入れる。複数の語を並べると、並べられた語全てを TITLE 中に持つエントリーのみがヒットされる。

(16) コピーコマンド

コマンド　　cop*y

【機能】　エントリーの内容を入力用フォーマットにしたがってユーザファイルにコピーする。

【書式】

copy|out=ユーザファイル[|オプション]　アクセッション番号

（ａ）|out=ユーザファイル名

通常ユーザファイルはカレントディレクトリに作成されるが、ファイル名にディレクトリ名を含めて指定すると、指定されたディレクトリに作成される。

（b）[ ] 内は省略できる。

（c）|オプション

オプションには次の１種類が用意されている。

(7) |current

カレントリストに登録されている全エントリーの内容をエントリー順にコピーする。この場合アクセッション番号を入力する必要はない。

(17) サブデータベース設定コマンド

コマンド　bas*e

【機能】　クロスリンクして使用する他のデータベースを利用可能な状態にしたり、利用可能なデータベースを表示する。

【書式】

（a）base

利用可能なデータベースを表示する。

（b）base　データベース名

指定されたデータベースを利用可能にする。複数指定する場合にはカンマで区切る。リストアップされている全てのデータベースを利用可能にする場合にはアスタリスク"*"で代用できる。

５．２　検索用インデックスファイルの作成

データベースの対象となるデータの主要項目について、各項目別に検索用インデックスファイルを作成する。

(1) 基本インデックス (variant.inx) の作成

各エントリーのデータは全て結合されてvariant.refという１つのファイルにまとめられている。このファイルは個々のエントリーファイルを単に順番に結合した形となっており、１つのエントリーの内容が全て連続して収納されている。そこで基本インデックスでは、各エントリーのアクセッション番号とそのエントリーが始まるアドレスとを記録しておく。これにより、アクセッション番号を指定することにより、variant.refファイルを頭から検索せずに、その内容にダイレクトにアクセスでき、内容の表示に要する時間を短縮している。

(2) 各種検索用インデックスファイルの作成

アクセッション番号が分かれば、(1)で示したようにその内容を短時間で表示できる。このアクセッション番号を検索するために、エントリーの内容の主要な項目毎にインデックスファイルを作成する。インデックスを作成する項目と対応するインデックスファイル名及び検索コマンドを表

5－1に示す。

各項目には統制のない自由な言葉が使われており、またタンパク質名や実験方法など特殊な言葉も使われている。そのため検索方法としては最も単純な中間一致検索方法を採用した。一方では、タンパク質名や実験方法など殆ど同じフレーズが繰り返し現われる項目もある。そこでインデックスファイルを作成するにあたり、各項目を一行単位に分解し、その文を昇順に並べ替え、その行を含む全てのエントリーの番号を登録した。

表5－1　項目、インデックス名及び検索コマンド対応表

項目	インデックスファイル	検索コマンド
タイトル	.tti	find
著者	.aui	author
雑誌	.jri	journal
変異タンパク質由来元	.soi	source
変異作成方法	.mei	method
発現方法	.exi	expression
成長速度	.gri	growth
変異タンパク質回収方法	.pui	purification
測定機能	.fui	function
キーワード	.woi	keyword
配列特徴部位	.fti	feature

(3) クロス検索用インデックスファイルの作成

この変異タンパク質データベースを単独で用いても変異とそれが諸性質に及ぼす影響を知ることが出来る。しかしながら、既存の各種タンパク質に対するデータベース、特にPIR-Internationalのアミノ酸配列データベースや生物活性データベースとともに利用すると、天然タンパク質に関する様々な情報を補うことができる。この変異タンパク質データベースの各エントリーのフォーマットは、これらのデータベースで採用しているフォーマットに沿っており、(2)の各種インデックス作成用プログラムで同じようにインデックスファイルを作成できる。このため、容易にクロス検索用インデックスファイルを作成でき、同時に検索できるばかりではなく、この変異タンパク質データベースからそれらのデータベースの対応する内容を引用することも可能である。またこのクロス検索用インデックスでは、(2)で作成した項目の内容全てに対して3文字毎に分解してその文字列のアドレスを登録している。これにより、より高速な検索が可能となっている。

５．３　データベースの作成

文献単位で入力を行った後、個別エントリーファイルを作成し、各種インデックスを作成し、データベースの作成を行った。

図５－２にデータの作成からデータの提供までの本データベースの流れを示す。

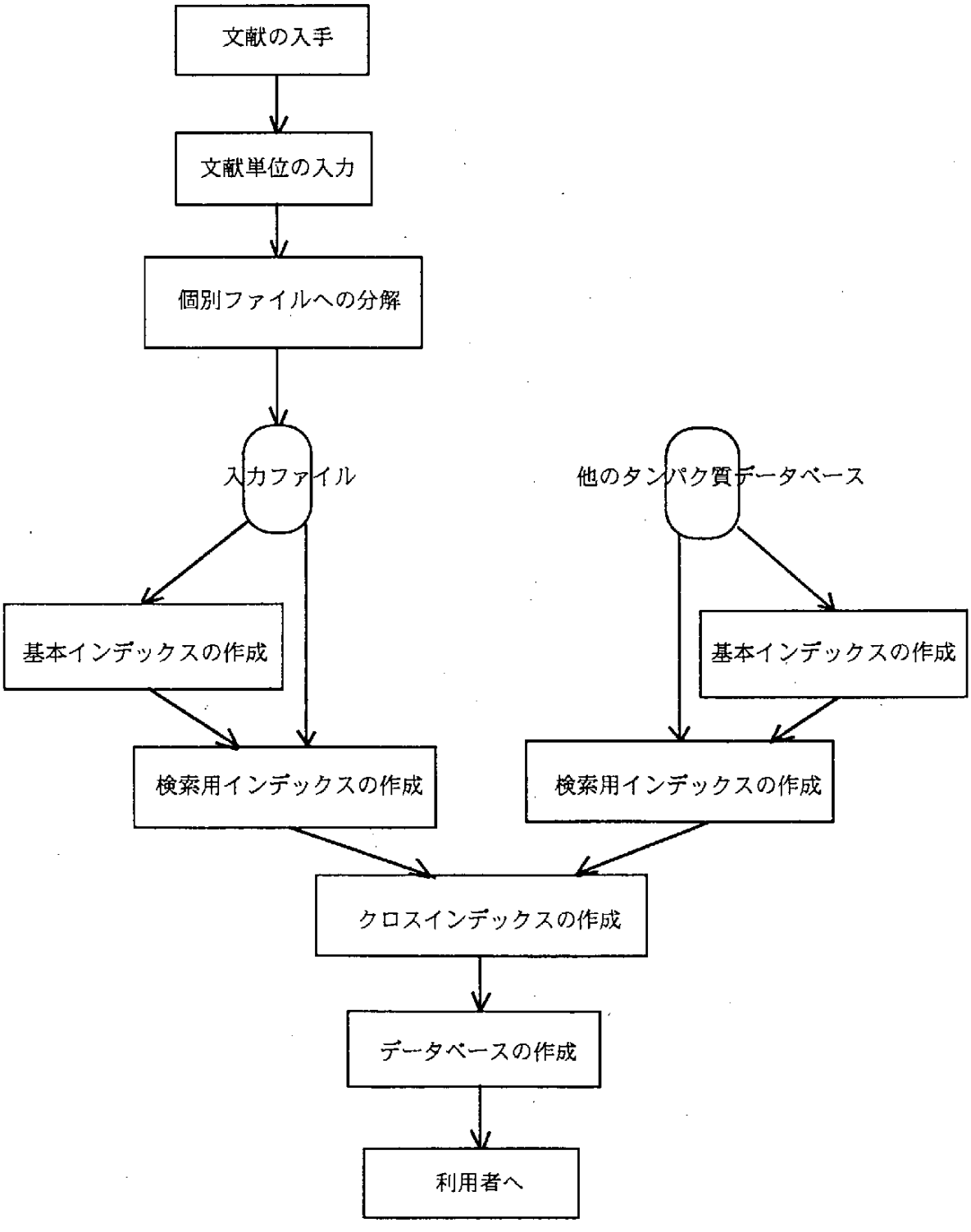


図５－２　本データベースシステムの流れ

５．４　ユーザデータベースの作成

本システムでは、各エントリーは第５.１節の「(14) シングル表示コマンド」で示したユーザファイルフォーマットを採用しており、このフォーマットにしたがって書かれたデータを次から次へと結合した形をしている。ユーザが独自のデータベースを作成する際もこのフォーマットにしたがえば、独自のデータベースを作成できる。ユーザデータベースの作成手順を以下に示す。

（１）ユーザファイルフォーマットにしたがって個々のエントリーを作成する。

（２）オペレーテイングシステムのコピーコマンドやエディタを利用して、データベース化したい全てのエントリーを１つのファイルに結合する。

（３）第５.２節（１）の基本インデックスを作成する。

（４）第５.２節（２）の各種検索用インデックスを作成する。

（５）第５.２節（３）で本データベースとのクロス検索用インデックスを作成する。

6．　データベースの検索

作成したデータベースが仕様どおり機能するかを調べるために、実際に検索されると思われる例題をいくつか想定し検索した。例題として、大きく分けてタンパク質名による検索、アミノ酸の置換パターンによる検索、著者名による検索、マルチディスプレイを取り上げる。

（以下、下線部分は入力を示す。）

6．1　タンパク質名による検索

最初にデータベースシステムを立ち上げると，プロンプトがATLAS> となる。

```
$ topiqus
ATLAS>
```

タンパク質名 super oxide dismutase （過酸化水素還元酵素）を検索することにする。タンパク質名は TITLE に書かれているので、検索コマンド find を使用する。

```
ATLAS> find super oxide dismu
Superoxide dismutase (Cu-Zn)(EC 1.15.1.1) C6A,C111S - Human
   var3:M10214    Superoxide dismutase (Cu-Zn)(EC 1.15.1.1) C6A,
                  C111S - Human
Superoxide dismutase (Cu-Zn)(EC 1.15.1.1) C6A,C111S,R114D - Human
   var3:M10215    Superoxide dismutase (Cu-Zn)(EC 1.15.1.1) C6A,
                  C111S,R114D - Human
Superoxide dismutase (Cu-Zn)(EC 1.15.1.1) fusion([1-152/C6A,
   C111S]=IgA1[102-118]=[C6A,C111S,R143D]) - Human
    var3:M10218   Superoxide dismutase (Cu-Zn)(EC 1.15.1.1)
                  fusion([1-152/C6A,C111S]=IgA1[102-118]=[C6A,
                  C111S,R143D]) - Human
Superoxide dismutase (Cu-Zn)(EC 1.15.1.1) fusion([1-152/C6A,
   C111S]=IgA1[102-118]=[C6A,C111S]) - Human
    var3:M10217   Superoxide dismutase (Cu-Zn)(EC 1.15.1.1)
                  fusion([1-152/C6A,C111S]=IgA1[102-118]=[C6A,
                  C111S]) - Human
Superoxide dismutase (Cu-Zn)(EC 1.15.1.1) fusion([1-152/C6A,
   C111S]=[C6A,C111S]) - Human
   var3:M10216    Superoxide dismutase (Cu-Zn)(EC 1.15.1.1)
                  fusion([1-152/C6A,C111S]=[C6A,C111S]) - Human
```

```
Superoxide dismutase (EC 1.15.1.1) C6A - Bovine
  var3:M03504    Superoxide dismutase (EC 1.15.1.1) C6A - Bovine
6 titles found
```

super oxide dismu という文字列を TITLE に持つエントリーが6件検索された。データベースには、super oxide dismu という文字列をもつエントリーは6件しかないので漏れなく検索されている。この6件のうちから、さらに2つのサブユニットを融合させたエントリーだけを取り出すことにする。そのためオプション | current を使用してカレントリストの中だけを検索する。

```
ATLAS> find|cur fusion
Superoxide dismutase (Cu-Zn)(EC 1.15.1.1) fusion([1-152/C6A,
  C111S]=IgA1[102-118]=[C6A,C111S,R143D]) - Human
   var3:M10218   Superoxide dismutase (Cu-Zn)(EC 1.15.1.1)
                 fusion([1-152/C6A,C111S]=IgA1[102-118]=[C6A,
                 C111S,R143D]) - Human
Superoxide dismutase (Cu-Zn)(EC 1.15.1.1) fusion([1-152/C6A,
  C111S]=IgA1[102-118]=[C6A,C111S]) - Human
   var3:M10217   Superoxide dismutase (Cu-Zn)(EC 1.15.1.1)
                 fusion([1-152/C6A,C111S]=IgA1[102-118]=[C6A,
                 C111S]) - Human
Superoxide dismutase (Cu-Zn)(EC 1.15.1.1) fusion([1-152/C6A,
  C111S]=[C6A,C111S]) - Human
  var3:M10216    Superoxide dismutase (Cu-Zn)(EC 1.15.1.1)
                 fusion([1-152/C6A,C111S]=[C6A,C111S]) - Human
3 titles found
```

カレントリストから TITLE にfusion という文字列を持つ3件のエントリーが検索され、新しいカレントリストとして登録された。データベース中には、この3件以外にも fusion という文字列を持つエントリーがあるので、この検索がカレントリストに限定されていることが確認できる。

次にこれら3件の内容を表示させる。シングル表示コマンド type でオプション | current を使用し、3件の内容を続けて表示させる。

```
ATLAS> type|current

var3:M10216
Superoxide dismutase (Cu-Zn)(EC 1.15.1.1) fusion([1-152/C6A,
```

C111S]=[C6A,C111S]) - Human

Hallewell, R.A., Laria, I., Tabrizi, A., Carlin, G., Getzoff, E.D., Tainer, J.A., Sousens, L.S., and Mullenbach, G.T., J. Biol. Chem. 264, 5260-5268, 1989
Title: Genetically engineered polymers of human CuZn superoxide dismutase.

Organization: 2;1-154<PIR1;DSHUCZ>

VARIATION: 2-6,'A',8-111,'S',113-*<DSHUCZ>,1-6,'A',8-111,'S',113-*<DSHUCZ>/fusion([1-152/C6A,C111S]=[C6A,C111S])

METHOD
Substitution: chemical synthesis of gene and cassette mutagenesis

EXPRESSION
Host: Escherichia coli strain sodAsodB
Plasmid: pNco5A

FUNCTION

Activity of cell lysates

Relative activity(percent)

WILD(clone): 100
M10216: 50

Assay: Pyrogallol method

DSHUCZ 2-6,'A',8-111,'S',113-* DSHUCZ 1-6,'A',8-111,'S',113-*

Composition of fragment

22 Ala A	6 Gln Q	18 Leu L	22 Ser S
8 Arg R	20 Glu E	22 Lys K	16 Thr T
14 Asn N	50 Gly G	1 Met M	2 Trp W

23 Asp D 16 His H 8 Phe F 0 Tyr Y

4 Cys C 18 Ile I 10 Pro P 28 Val V

Number of residues = 308

```
          5         10        15        20        25        30
 1 A T K A V A V L K G D G P V Q G I I N F E Q K E S N G P V K
31 V W G S I K G L T E G L H G F H V H E F G D N T A G C T S A
61 G P H F N P L S R K H G G P K D E E R H V G D L G N V T A D
91 K D G V A D V S I E D S V I S L S G D H S I I G R T L V V H
121 E K A D D L G K G G N E E S T K T G N A G S R L A C G V I G
151 I A Q M A T K A V A V L K G D G P V Q G I I N F E Q K E S N
181 G P V K V W G S I K G L T E G L H G F H V H E F G D N T A G
211 C T S A G P H F N P L S R K H G G P K D E E R H V G D L G N
241 V T A D K D G V A D V S I E D S V I S L S G D H S I I G R T
271 L V V H E K A D D L G K G G N E E S T K T G N A G S R L A C
301 G V I G I A Q D
```

var3:M10217

Superoxide dismutase (Cu-Zn)(EC 1.15.1.1) fusion([1-152/C6A, C111S]=IgA1[102-118]=[C6A,C111S]) - Human

Hallewell, R.A., Laria, I., Tabrizi, A., Carlin, G., Getzoff, E.D.,Tainer, J.A., Sousens, L.S., and Mullenbach, G.T., J. Biol. Chem. 264, 5260-5268, 1989

Title: Genetically engineered polymers of human CuZn superoxide dismutase.

Organization: 2;1-154<PIR1;DSHUCZ>

VARIATION: 2-6,'A',8-111,'S',113-*<DSHUCZ>,102-118<A1HU>,1-6, 'A',8-111,'S',113-*<DSHUCZ>/fusion([1-152/C6A,C111S]= IgA1[102-118]=[C6A,C111S])

METHOD

Substitution: chemical synthesis of gene and cassette mutagenesis

Fusion: two genes linked by human IgA1 hinge sequence

EXPRESSION

Host: Escherichia coli strain sodAsodB

Plasmid: pNco5A

EXPRESSION

Host: Yeast

Vector: pC1/1

Yeald: Abount 40 percent of total soluble cell protein

PURIFICATION: Heating at 65 C for 2 hr, DEAF-Sepharose chromatography.

FUNCTION

Activity of cell lysates

Relative activity(percent)	
WILD(clone):	100
M10217:	100

Assay: Pyrogallol method

DSHUCZ 2-6,'A',8-111,'S',113-* A1HU 102-118 DSHUCZ 1-6,'A', 8-111,'S',113-*

Composition of fragment

22 Ala A	6 Gln Q	18 Leu L	25 Ser S
8 Arg R	20 Glu E	22 Lys K	20 Thr T
14 Asn N	50 Gly G	1 Met M	2 Trp W
22 Asp D	16 His H	8 Phe F	0 Tyr Y
4 Cys C	18 Ile I	19 Pro P	29 Val V

Number of residues = 324

```
          5         10        15        20        25        30
 1 A T K A V A V L K G D G P V Q G I I N F E Q K E S N G P V K
```

31 V W G S I K G L T E G L H G F H V H E F G D N T A G C T S A

61 G P H F N P L S R K H G G P K D E E R H V G D L G N V T A D

91 K D G V A D V S I E D S V I S L S G D H S I I G R T L V V H

121 E K A D D L G K G G N E E S T K T G N A G S R L A C G V I G

151 I A Q P V P S T P P T P S P S T P P T P M A T K A V A V L K

181 G D G P V Q G I I N F E Q K E S N G P V K V W G S I K G L T

211 E G L H G F H V H E F G D N T A G C T S A G P H F N P L S R

241 K H G G P K D E E R H V G D L G N V T A D K D G V A D V S I

271 E D S V I S L S G D H S I I G R T L V V H E K A D D L G K G

301 G N E E S T K T G N A G S R L A C G V I G I A Q

var3:M10218

Superoxide dismutase (Cu-Zn)(EC 1.15.1.1) fusion([1-152/C6A, C111S]=IgA1[102-118]=[C6A,C111S,R143D]) - Human

Hallewell, R.A., Laria, I., Tabrizi, A., Carlin, G., Getzoff, E.D., Tainer, J.A., Sousens, L.S., and Mullenbach, G.T., J. Biol. Chem. 264, 5260-5268, 1989

Title: Genetically engineered polymers of human CuZn superoxide dismutase.

Organization: 2;1-154<PIR1;DSHUCZ>

VARIATION: 2-6,'A',8-111,'S',113-*<DSHUCZ>,102-118<A1HU>,1-6, 'A',8-111,'S',113-143,'D',145-*<DSHUCZ>/fusion([1-152/C6A, C111S]=IgA1[102-118]=[C6A,C111S,R143D])

METHOD

Substitution: chemical synthesis of gene and cassette mutagenesis

Fusion: two genes linked by human IgA1 hinge sequence

EXPRESSION

Host: Escherichia coli strain sodAsodB

Plasmid: pNco5A

EXPRESSION

Host: Yeast

Vector: pC1/1

Yeald: Abount 40 percent of total soluble cell protein

FUNCTION

Activity of cell lysates

Relative activity(percent)

WILD(clone): 100

M10218: 50

Assay: Pyrogallol method

DSHUCZ 2-6,'A',8-111,'S',113-* A1HU 102-118 DSHUCZ 1-6,'A', 8-111,'S',113-143,'D',145-*

Composition of fragment

22 Ala A	6 Gln Q	18 Leu L	25 Ser S
7 Arg R	20 Glu E	22 Lys K	20 Thr T
14 Asn N	50 Gly G	1 Met M	2 Trp W
23 Asp D	16 His H	8 Phe F	0 Tyr Y
4 Cys C	18 Ile I	19 Pro P	29 Val V

Number of residues = 324

```
          5         10        15        20        25        30
  1 A T K A V A V L K G D G P V Q G I I N F E Q K E S N G P V K
 31 V W G S I K G L T E G L H G F H V H E F G D N T A G C T S A
 61 G P H F N P L S R K H G G P K D E E R H V G D L G N V T A D
 91 K D G V A D V S I E D S V I S L S G D H S I I G R T L V V H
121 E K A D D L G K G G N E E S T K T G N A G S R L A C G V I G
151 I A Q P V P S T P P T P S P S T P P T P M A T K A V A V L K
181 G D G P V Q G I I N F E Q K E S N G P V K V W G S I K G L T
211 E G L H G F H V H E F G D N T A G C T S A G P H F N P L S R
241 K H G G P K D E E R H V G D L G N V T A D K D G V A D V S I
271 E D S V I S L S G D H S I I G R T L V V H E K A D D L G K G
301 G N E E S T K T G N A G S D L A C G V I G I A Q
```

6．2　アミノ酸置換パターンによる検索

あるアミノ酸が他のあるアミノ酸に置き換わっている変異タンパク質を検索してみる。この検索にはコマンドvariationを用いる。システインがアラニンに置き換わっている変異タンパク質の検索を試みる。

```
ATLAS> variation c-a
C-A
var3:M03504     Superoxide dismutase (EC 1.15.1.1) C6A - Bovine
var3:M10030     Alcohol dehydrogenase (EC 1.1.1.1) C135A -
                Drosophila melanogaster
var3:M10031     Alcohol dehydrogenase (EC 1.1.1.1) C218A -
                Drosophila melanogaster
var3:M10194     Asparagine synthase (glutamine-hydrolyzing)
                (EC 6.3.5.4) C1A  - Human
C-A,C-A
var3:M10032     Alcohol dehydrogenase (EC 1.1.1.1) C135A,C218A -
                Drosophila melanogaster
var3:M10216     Superoxide dismutase (Cu-Zn)(EC 1.15.1.1)
                fusion([1-152/C6A,C111S]=[C6A,C111S]) - Human
var3:M10217     Superoxide dismutase (Cu-Zn)(EC 1.15.1.1)
                fusion([1-152/C6A,C111S]=IgA1[102-118]=[C6A,
                C111S]) - Human
C-A,C-A,C-S
var3:M10218     Superoxide dismutase (Cu-Zn)(EC 1.15.1.1)
                fusion([1-152/C6A,C111S]=IgA1[102-118]=[C6A,
                C111S,R143D]) - Human
C-A,C-S
var3:M10214     Superoxide dismutase (Cu-Zn)(EC 1.15.1.1) C6A,
                C111S - Human
C-A,C-S,R-D
var3:M10215     Superoxide dismutase (Cu-Zn)(EC 1.15.1.1) C6A,
                C111S,R114D - Human
  5 variations found
```

システインからアラニンへの置換を含む５つの変異パターン、合計１０エントリー（１個のシステイン－アラニン置換を持つエントリーが４件、２個のシステイン－アラニン置換を持つエントリー

が３件、２個のシステイン－アラニン、１個のシステイン－セリン置換を持つエントリーが１件、１個のシステイン－アラニン、システイン－セリン置換を持つエントリーが１件、１個のシステイン－アラニン、システイン－セリン、アルギニン－アスパラギン置換を持つエントリーが１件）が漏れなく検索され、カレントリストに登録される。

６．３　著者名による検索

文献の著者名の中にsmithという名前があるエントリーを検索する。著者名を検索するコマンドauthourを用いる。

```
ATLAS> autho smith
Smith,F I
var3:M03550    Glucosylceramidase (EC 3.2.1.45) N388S - Human
var3:M03551    Glucosylceramidase (EC 3.2.1.45) R139Q,N388S -
   Human
var3:M10045    Glucosylceramidase (EC 3.2.1.45) T62K,N388S -
   Human
var3:M10046    Glucosylceramidase (EC 3.2.1.45) R139Q - Human
var3:M10047    Glucosylceramidase (EC 3.2.1.45) D376E - Human
var3:M10048    Glucosylceramidase (EC 3.2.1.45) N388S - Human
Smith,W L
var3:M10209    Prostaglandin-endoperoxide synthase
               (EC 1.14.99.1) Y230F - Sheep
var3:M10210    Prostaglandin-endoperoxide synthase
               (EC 1.14.99.1) Y238F - Sheep
var3:M10211    Prostaglandin-endoperoxide synthase
               (EC 1.14.99.1) Y331F - Sheep
var3:M10212    Prostaglandin-endoperoxide synthase
               (EC 1.14.99.1) Y361F - Sheep
var3:M10213    Prostaglandin-endoperoxide synthase
               (EC 1.14.99.1) Y393F - Sheep
2 authors found
```

Smith, F.I. と Smith, W.L. という２人の著者が見つかり、Smith, F.L. という著者に関するエントリーが６件、Smith, W.L. という著者に関するエントリーが５件、合計１１件のエントリーが漏れなく検索されカレントリストに登録される。

6．4　検索結果一覧表示

検索結果一覧コマンド list を使用して現在のカレントリストを表示して見る。6．3 の author コマンドに引き続き list を実行する。

```
ATLAS> list
 11 entries in the database

  var3:M03550   Glucosylceramidase (EC 3.2.1.45) N388S - Human
  var3:M03551   Glucosylceramidase (EC 3.2.1.45) R139Q,N388S -
                Human
  var3:M10045   Glucosylceramidase (EC 3.2.1.45) T62K,N388S -
                Human
  var3:M10046   Glucosylceramidase (EC 3.2.1.45) R139Q - Human
  var3:M10047   Glucosylceramidase (EC 3.2.1.45) D376E - Human
  var3:M10048   Glucosylceramidase (EC 3.2.1.45) N388S - Human
  var3:M10209   Prostaglandin-endoperoxide synthase
                (EC 1.14.99.1) Y230F - Sheep
  var3:M10210   Prostaglandin-endoperoxide synthase
                (EC 1.14.99.1) Y238F - Sheep
  var3:M10211   Prostaglandin-endoperoxide synthase
                (EC 1.14.99.1) Y331F - Sheep
  var3:M10212   Prostaglandin-endoperoxide synthase
                (EC 1.14.99.1) Y361F - Sheep
  var3:M10213   Prostaglandin-endoperoxide synthase
                (EC 1.14.99.1) Y393F - Sheep
```

6．3の結果検索された１１件のエントリーのアクセッション番号とエントリーの TITLE が漏れなく表示される。

次に直前のカレントリストの復活を試みる。カレントリストの中から TITLE に glucose という文字列を持つ検索を試みる。

```
ATLAS> find|curr glucose
   title: glucose
   No titles found
```

カレントリスト中のエントリーで TITLE に glucose という文字列を含むものは１件もないので

カレントリストは空になる。 list コマンドでカレントリストを表示して確認する。

```
ATLAS> list
   The current list is empty
```

ここでlistコマンドのオプション|restoreを利用して最後の検索コマンドを実行する直前のカレントリスト（6．3のauthorコマンドの結果作成されたカレントリスト）に戻す。

```
ATLAS> list|restore
    11 entries on the restored current list
```

listコマンドを使用してカレントリストを表示し、カレントリストが回復されたことを確認する。

```
ATLAS> list
   11 entries in the database

  var3:M03550   Glucosylceramidase (EC 3.2.1.45) N388S - Human
  var3:M03551   Glucosylceramidase (EC 3.2.1.45) R139Q,N388S -
                Human
  var3:M10045   Glucosylceramidase (EC 3.2.1.45) T62K,N388S -
                Human
  var3:M10046   Glucosylceramidase (EC 3.2.1.45) R139Q - Human
  var3:M10047   Glucosylceramidase (EC 3.2.1.45) D376E - Human
  var3:M10048   Glucosylceramidase (EC 3.2.1.45) N388S - Human
  var3:M10209   Prostaglandin-endoperoxide synthase
                (EC 1.14.99.1) Y230F - Sheep
  var3:M10210   Prostaglandin-endoperoxide synthase
                (EC 1.14.99.1) Y238F - Sheep
  var3:M10211   Prostaglandin-endoperoxide synthase
                (EC 1.14.99.1) Y331F - Sheep
  var3:M10212   Prostaglandin-endoperoxide synthase
                (EC 1.14.99.1) Y361F - Sheep
  var3:M10213   Prostaglandin-endoperoxide synthase
                (EC 1.14.99.1) Y393F - Sheep
```

6．5　マルチ表示

由来した天然タンパク質が同じ変異タンパク質について、変異タンパク質間の諸特性を比較するためにマルチ表示で出力する。ショウジョウバエ（Drosophila melanogaster）のアルコール脱水素酵素（alcohol dehydrogenase）に対する変異タンパク質についてマルチ表示を行う。まずショウジョウバエのアルコール脱水素酵素に対する変異タンパク質が何件登録されているかを確認するために find コマンドを使用して検索してみる。

```
ATLAS> find alcohol dehydrogenase drosophila melanogaster
Alcohol dehydrogenase (EC 1.1.1.1) C135A - Drosophila melanogaster
  var3:M10030    Alcohol dehydrogenase (EC 1.1.1.1) C135A -
                 Drosophila melanogaster
Alcohol dehydrogenase (EC 1.1.1.1) C135A,C218A - Drosophila
    melanogaster
  var3:M10032    Alcohol dehydrogenase (EC 1.1.1.1) C135A,C218A -
                 Drosophil melanogaster
Alcohol dehydrogenase (EC 1.1.1.1) C218A - Drosophila melanogaster
  var3:M10031    Alcohol dehydrogenase (EC 1.1.1.1) C218A -
                 Drosophila melanogaster
Alcohol dehydrogenase (EC 1.1.1.1) G14A - Drosophila melanogaster
  var3:M10029    Alcohol dehydrogenase (EC 1.1.1.1) G14A -
                 Drosophila melanogaster
Alcohol dehydrogenase (EC 1.1.1.1) G14V - Drosophila melanogaster
  var3:M10028    Alcohol dehydrogenase (EC 1.1.1.1) G14V -
                 Drosophila melanogaster
    5 title found
```

ショウジョウバエのアルコール脱水素酵素に対する変異タンパク質が５件検索された。そこでマルチ表示コマンド mtype を使用してこれら５件のエントリーを１つにまとめて表示する。

```
ATLAS> mtype alcohol dehydrogenase drosophila
Alcohol dehydrogenase (EC 1.1.1.1) - Drosophila melanogaster

Chen, Z., Lu, L., Shirley, M., Lee., W.R., and Chang, S.H.,
     Biochemistory 29, 1112-1118, 1990
    Title: Site-directed mutagenesis of glycine-14 and two
    "critical" cysteinyl residues in Drosophila alcohol
```

dehydrogenase.

VARIATION

M10028: 2-14,'V',16-*/G14V <DEFFA>

M10029: 2-14,'A',16-*/G14A <DEFFA>

M10030: 2-135,'A',137-*/C135A <DEFFA>

M10031: 2-218,'A',220-*/C218A <DEFFA>

M10032: 2-135,'A',137-218,'A',220-*/C135A,C218A <DEFFA>

METHOD <M10028><M10029><M10030><M10031><M10032>

Substitution: oligonucleotide-directed mutagenesis

EXPRESSION <M10028>

Host: Escherichia coli strain LC137

Vector: pPL2

PURIFICATION: The specific activity was too low to purify.<M10028>

PURIFICATION <M10030><M10031><M10032>

Chromatography: Sephadex G-100, Cibacron Blue 3GA-agarose

Purity: homogenious

FUNCTION

Dehydrogenation of 2-propanol

Relative activity	
WILD(clone):	1.00
M10029:	0.69
M10030:	1.17
M10031:	1.08
M10032:	1.45

	Km[NAD]	Km[2-propanol]	kcat	Ki[NADP]

```
                (mM)          (mM)       (1/s)     (mM)
-------------------------------------------------------
  WILD(clone):   0.12          0.61       2.8      0.944
  M10029:        0.42          0.56       1.9     11.9
-------------------------------------------------------
      Condition: pH 8.7, 25 C
```

アルコール脱水素酵素に対する５件の変異タンパク質は全て同じ文献に記載されていたので、REFERENCE は１件しか表示されない。変異タンパク質を作成した方法も全て共通なので METHOD も１件だけ表示される。 EXPRESSION については M10028 に対してのみ記載されている。 PURIFICATION については、 M10028 に対しての記載と M10020, M10031, M10032 に対する記載が異なっているので、２種類表示され、該当する記載にアクセッション番号が付いている。 FUNCTION では測定された特性ごとに表形式にまとめられる。 Relative activity は WILD（比較用天然タンパク質）, M10029, M10030, M10031, M10032 に対して測定されている。WILD と M10029 に対しては Km, kcat, Ki といった値に対しても測定されている。

６．６　ユーザファイルへの出力、ユーザファイルの表示

データベースからユーザファイルにエントリーの内容を出力する方法には２種類ある。 type コマンドでオプション |out=ユーザファイル名 を使用する方法と、 copy コマンドを使用する方法である。 type コマンドを使用するとディスプレイに表示される形式でコピーされ、copy コマンドの場合はデータが保存されているフォーマットでコピーされる。そのため、 copy コマンドで作成されたファイルをユーザが加工し、再び読み込み表示することができる。ここでは単純にデータを copy コマンドで出力し、そのファイルを表示させてみる。 m03504 というエントリーを my.file というファイルに出力させる。なお既に my.file というファイルが存在する場合にはデータが上書きされてしまう。

```
ATLAS> copy|out=my.file m03504
ATLAS>
```

データベースシステムを終了させてから、オペレーテイングシステムのコマンドを使用して、作成された my.file の内容を表示してみると以下のようになる。

```
>VA;M03504
Superoxide dismutase (EC 1.15.1.1) C6A - Bovine
R;McRee, D.E., Redford, S.M., Getzoff, E.D., Lepock, J.R.,
Hallewell, R.A., and Tainer, J.A.
```

J. Biol. Chem. 265, 14234-14241, 1990

A;Title: Changes in crystallographic structure and thermostability of a Cu,Zn Superoxide dismutase mutant resulting from the removal of a buried cysteine.

C;VARIATION

V;M03504: 1-5,'A',7-*/C6A <DSBOCZ>

C;METHOD

W;#Substitution: oligonucleotide-directed mutagenesis

C;EXPRESSION

W;#Host: Yeast strain 2150-2-3 Leu(-)

W;#Vector: pC1/1 (yeast shuttle vector)

W;#Post-translation: the N-termini are acetylated.

C;PHYSICOCHEMICAL PROPERTY

C;Thermotability of dimer form at 70 C for 3 hr

W;----------

W; Residual activity(percent)

W;----------

W; WILD(clone): 40

W; M03504: 70

W;----------

C;Thermodinamic parameters

W;----------

W; For major component

	Tm (C)	delta-H (kcal/mol)	delta-S (cal/K/mol)	delta-delta-G (kcal/mol)
WILD(clone):	89.5	161	445	0.0
M03504:	85.8	124	347	-1.3

W;----------

W;----------

W; For minor component

	Tm (C)	delta-H (kcal/mol)	delta-S (cal/K/mol)	delta-delta-G (kcal/mol)
WILD(clone):	82.8	113	318	0.0
M03504:	80.7	123	349	-0.75

W;----------

```
A;The parameters are calculated for a two-component reversible
denaturation model (with major and minor component) and defined as
temperature of half-completion for each transition.
C;KEYWORDS: Greek key beta-barrel motif, X-ray crystallography
Z;<KS1719>
```

この my.file の内容をデータベースシステムを利用して表示させる。この場合 type コマンドでアクセッション番号のあとにファイル名を付け加える。

```
ATLAS> type m03504=my.file

var3:M03504
Superoxide dismutase (EC 1.15.1.1) C6A - Bovine

McRee, D.E., Redford, S.M., Getzoff, E.D., Lepock, J.R.,
   Hallewell, R.A., and Tainer, J.A., J. Biol. Chem. 265, 14234-
   14241, 1990
   Title: Changes in crystallographic structure and
   thermostability of a Cu,Zn Superoxide dismutase mutant
   resulting from the removal of a buried cysteine.

Organization: 2;1-151<PIR1;DSBOCZ>

VARIATION: 1-5,'A',7-*/C6A <DSBOCZ>

METHOD
 Substitution: oligonucleotide-directed mutagenesis

EXPRESSION
 Host: Yeast strain 2150-2-3 Leu(-)
 Vector: pC1/1 (yeast shuttle vector)
 Post-translation: the N-termini are acetylated.

PHYSICOCHEMICAL PROPERTY
 Thermotability of dimer form at 70 C for 3 hr
-----------------------------------
```

Residual activity(percent)

WILD(clone):	40
M03504:	70

Thermodinamic parameters

For major component

	Tm (C)	delta-H (kcal/mol)	delta-S (cal/K/mol)	delta-delta-G (kcal/mol)
WILD(clone):	89.5	161	445	0.0
M03504:	85.8	124	347	-1.3

For minor component

	Tm (C)	delta-H (kcal/mol)	delta-S (cal/K/mol)	delta-delta-G (kcal/mol)
WILD(clone):	82.8	113	318	0.0
M03504:	80.7	123	349	-0.75

The parameters are calculated for a two-component reversible denaturation model (with major and minor component) and defined as temperature of half-completion for each transition.

KEYWORDS: Greek key beta-barrel motif, X-ray crystallography

DSBOCZ 1-5,'A',7-*/C6A

Composition of fragment

10 Ala A	3 Gln Q	8 Leu L	8 Ser S
4 Arg R	8 Glu E	10 Lys K	12 Thr T
6 Asn N	25 Gly G	1 Met M	0 Trp W
11 Asp D	8 His H	4 Phe F	1 Tyr Y

2 Cys C 9 Ile I 6 Pro P 15 Val V

Number of residues = 151

```
          5         10        15        20        25        30
  1 A T K A V A V L K G D G P V Q G T I H F E A K G D T V V V T
 31 G S I T G L T E G D H G F H V H Q F G D N T Q G C T S A G P
 61 H F N P L S K K H G G P K D E E R H V G D L G N V T A D K N
 91 G V A I V D I V D P L I S L S G E Y S I I G R T M V V H E K
121 P D D L G R G G N E E S T K T G N A G S R L A C G V I G I A
151 K
```

7．　展望

7．1　用語の統一

今回の方法では、タンパク質名、生物種名、変異の記載方法、著者名、雑誌名については基準にしたがって統一が取れているが、SOURCE, METHOD, EXPRESSION, PURIFICATION　及びFUNCTION に関しては、文献の記述にしたがっており、用語の統一が完全に取れているとはいえない。 またKEYWORDに関しては文献を詳細に読んだ人間の主観によるところが大きく、殆ど統一されていない。

KEYWORD についてはこれといった基準はないので、用語を統一するためには、以前のエントリーで用いられた用語をリスト化しておき、KEYWORD を決定する際に常にそのリストを参照する方法が最も現実的である。この方法を利用するためには、常にKEYWORD リストを更新しておく必要があり、 KEYWORD を自動的に切り出し、リスト化するシステムを作成する必要がある。

KEYWORD に関しては用語の統一以外に、その使用の統一も図る必要がある。由来した天然タンパク質が同じ変異タンパク質群について、その変異タンパク質群に共通した性質を示すKEYWORD を１つのエントリーに使用した場合、他のエントリーに対しても使用しなければならない。そのため、その用語をどのエントリーに対して使用したか対応表を作成しておく必要がある。これは結局 KEYWORD インデックスそのものである。

必要なKEYWORD が漏れなくエントリーに記載されているかのチェックは、その KEYWORD を含むエントリーのリストと、そのKEYWORD を含んでいておかしくないエントリーのリストとの比較で行える。後者のリストは、例えばタンパク質名などでの検索から作成する。もしこの２つのリストが一致していれば、必要なKEYWORD が漏れなくエントリーに記載されていることになる。もしKEYWORD インデックスのリストが少ない場合には、幾つかのエントリーでのKEYWORD の記入漏れが考えられる。逆に多い場合には、KEYWORD の記載が適切かどうかの判断をし、不適切な場合は削除する。このような修正が簡単なシステムに変更し、さらに毎日自動的に、各種インデックスの更新を行うようなメカニズムを付け加えれば、用語の統一が図れると予想される。

7．2　新規データの追加

変異タンパク質は今後もますます作成され、それらのデータも蓄積されていく。このデータベースは学術文献を分析する必要があり、また元となる文献（あるいはそのコピー）が簡単に入手できる環境が必要である。今後このデータベースをアップデイトしていくためには、大学等の公的機関に委託するなどの方法を取る必要があると考えられる。

7．3　他のデータベースとの整合性

今回作成した変異タンパク質データベースでは、タンパク質名、生物種名、著者名、雑誌名の記載方法はPIR-Internationalが採用している方法に準じた。これにより、この変異タンパク質データベースから、アミノ酸配列データベースや生物活性データベースの内容を引用することが可能となった。逆にこれら２つのデータベースが変異タンパク質データベースの内容を参照するためには、変異タンパク質独自の項目に対応し、クロスリファレンスとして変異タンパク質のアクセッション番号を登録することが必要となる。これにはPIR-Internationalとの協力が欠かせないが、もしこれが実現すれば、変異タンパク質データベースの利用範囲が広がり、世界的学術的貢献度も増すことになる。

7．4　表形式以外のデータ

この変異データベースでは、生物活性や物理化学的特性の変化を全て表形式で記載してある。しかし文献の中には特性の変化をグラフで示しているものも数多く見られる。このような場合、ある代表的な数点の情報しかデータベースに取り込んでいないことになる。ある簡単な式の組み合わせでそのグラフを表現できる場合は殆どないので、このようなデータは、画像データとして保存するのが最善であると思われる。今後ワークステーションでの画像を取り込んだデータベースに対する技術の発展が予想され、近い将来グラフデータ等を取り込むことが容易になると期待する。

謝　辞

本データベースの開発にあたって、データの収集及び本システムの評価などの点で、国際蛋白情報データベースに多大な協力をいただいた。ここに深く感謝の意を表する。

参考文献

1. George, D.G., Mewes, H.W., and Kihara H. (1987) Protein Seq. Data Anal. 1:27-39
2. Jone, C.S., Tsugita, A., Satake, K., Okibayashi, F., Imai, K., Yagi., T., Takahashi, K., and Yeh, L.-S. (1991) Protein Seq. Data Anal. 4:367-374
3. Ubasawa, A., Okibayashi, F., Jone, C.S., Ikehara, M., George, D.G., and Tsugita, A. (1991) Protein Seq. Data anal. 4:341-347

平成 5 年 3 月 発行

発　行　　財団法人　データベース振興センター
東京都港区浜松町二丁目4番1号
世界貿易センタービル7階
TEL 03—3459—8581

委託先　　日本電子計算株式会社
東京都江東区東陽2—4—24
TEL 03—5690—3202

印刷所　　大栄印刷産業株式会社
東京都荒川区東日暮里3—29—7
TEL 03—3806—2725