修 士 論 文

ベクトル空間法を用いてゲノムデータベース全体から関連性を抽出する手法に関する研究

指導教官　佐藤 賢二 助教授

北陸先端科学技術大学院大学
知識科学研究科 知識システム基礎学専攻

片岡 孝雄

2000 年 2 月 15 日

目 次

図 目 次

表 目 次

第 1 章

序論

1.1 研究の背景と問題点

全ての生物は細胞から構成され、ひとつひとつの細胞の中に格納されている DNA には生物を形作るのに必要な数十万個から数百万個の遺伝子の遺伝子が入っている。この遺伝子セットはゲノムと呼ばれ、生命の設計図に相当する [8]。ワトソンおよびクリックにより DNA の二重螺旋構造が明らかになって以来、遺伝子レベル・分子レベルで生命の姿を明らかにするための研究が爆発的な勢いで進展した。これらは分子生物学やゲノム解析学と呼ばれる学問分野に成長し、情報科学と並んで 21 世紀の発展がますます期待されている。また、生物学的な実験手法や実験技術の進化により、遺伝子や蛋白質分子について明らかになったデータは加速度的に増え続けている。これを格納したデータベースは一般にゲノムデータベースと呼ばれ、それぞれ膨大な量のデータを保有している。例えば米国 NCBI で集積／再配布されている遺伝子データベースである GenBank には、現在約 535 万個の遺伝子データが登録されているし、米国ブルックヘブン研究所でスタートした蛋白質立体構造データベースである PDB の場合は、登録数こそ 1 万余りと少ないものの、そのデータ量は 5GB にもなる。これらゲノムデータベースに登録されているデータをよりよく利用するためには、従来から計算機および計算機科学者の協力が必要とされ、ゲノム情報学という新しい学問分野を形成するに到っている。その成果により、今日の生物学者や医学者はネットワーク経由でゲノムデータベースのキーワード検索や類似した遺伝子配列の検索（ホモロジー検索）を行なうことが当たり前になり、計算機を用いた遺伝子や蛋白質の解析は、実験室で行なわれる生物学的実験と並んで必須のものとなった。

一方、Web ブラウザの登場以来、WWW は飛躍的な発展を遂げ [3]、インターネット上で公開されている無数の情報源を日常的に利用する事が、我々の生活や文化を変えよう

としている。また、世界中に散らばる数億数十億の Web ページを効果的に知識へと変換するために、計算機科学で研究されていた全文検索技術と高速な検索を可能にするインデクシング技術を利用したいわゆるサーチエンジンが目覚しい発達を遂げた。その結果、商用の製品と並んでフリーウェアのサーチエンジンも登場し、比較的中小規模のサイト検索によく利用されている。サーチエンジンは、検索対象となるテキストデータから網羅的に単語を切り出して頻度計算を行ない、ある単語を含むテキスト集合を高速に選び出すためのインデクスを作成しておく事により、利用者が指定したキーワード列を多く含む文書を検索結果として利用者に返す。さらに最近では、質問としてキーワードを与えるのではなく、文書そのものを与えるタイプのサーチエンジンも登場している（Justsystem の ConceptBase[6][7] など）。これは文書を頻度つきキーワードベクトルとして抽象的に表現し、質問文書のベクトルと類似したベクトルを持つものを選び出す手法で、通常のサーチエンジンがキーワードから文書への対応関係をインデクスとして保持しているのと逆に、文書からキーワードベクトルへの対応関係を保持している。これを利用すれば、従来ゲノムデータベースであまり効果的に利用できなかった、データエントリ中の自然言語情報（物質名やその定義など、各種のアノテーション）を使って、類似したデータエントリを検索したり、類似度に基づいてデータベース全体を自動分類することが可能になる。

しかしながら、現在の類似文書検索手法は、新聞記事などのように一般的な自然言語で書かれた文章が主要コンテンツであるようなデータを対象として発展してきた。そのため、科学技術データベースの一種であるゲノムデータベースに対し、そのまま適用するにはいくつか問題点がある。まず、ゲノムデータベースのデータエントリには遺伝子配列や蛋白質立体構造を記述するための文字列や数値が多く含まれる。これらは動詞や名詞など普通の意味での「単語」ではなく、しかもデータベース中には非常に多種類のこのような「非単語」がある。また、単語に含まれる特殊記号や、物質名をいくつかつなげた造語の問題、同義語の問題なども、単語の正しい認識や切り出しを阻む厄介な問題である。さらに、ゲノムデータベースは一種の科学技術データベースであるため、必然的に専門用語を多く含んでいる。このため、新聞記事の文章などとは単語の頻度分布や各単語の重要度が大きく異なっている。ここでは専門用語でない通常の英単語は重要ではなく、むしろ正しい類似度計算を撹乱する阻害要因となることが予想される。

以上述べたようにいくつかの問題点はあるものの、本研究ではまず標準的なサーチエンジンソフトウェアをゲノムデータベースに対して適用することで単語の切り出しと頻度計算を行ない、インデクスに格納されたそれらの情報を逆転させて類似文書検索用のキーワードベクトル生成を行なうことにより、どの程度データエントリの分類が行なえるかどうかを調べる。このアプローチでは、自然言語処理分野で研究されてきたような品詞解析

や未知語検出、複合語の扱いなどの深い自然言語処理を行なわない代わりに、流通しているサーチエンジンを使った高速なインデクス生成処理が利用できることや、通常のキーワード検索サービスの提供と同時並行的に類似文書検索サービスや分類サービスを提供できる可能性があることなど、利点も多い。ゲノム情報処理の分野でも、サーチエンジンのインデクスから類似度計算を行なう研究は世界的にも例がなく、実用性の高い新しい研究であると言える。

1.2 本論文の構成

本論文は、10章から構成されている。第 2章、第 3章、第 4章は本研究の説明をする上での予備知識を説明している。第 2章は、本研究で使用した全文検索エンジン Namazu について説明する。第 3章は、情報検索技術について説明しており、特に通常使用されているキーワード検索の仕組みと、本研究で使用するベクトル空間法について説明する。第 4章は、本研究で対象とするゲノムデータベースについて説明する。第 5章は関連研究について説明している。第 6章は、ゲノムデータベースに Namazu を使用することにより、キーワードから文書への対応関係を保持するインデクスを生成する方法を説明する。第 7章は、第 6章のインデクスを使って文書からキーワードへの対応を持つベクトルを生成する方法を説明する。第 8章は、インデクシングされたキーワードを区別するために作成した辞書について説明する。第 9章は本研究の主題であり、エントリの分類実験について記述している。最後に、第 10章で結論を述べる。

第 2 章

全文検索エンジン Namazu

2.1 はじめに

近年の Web の発達は目覚ましいものがあり、社会的に注目されている。コンピュータ関連の本や雑誌はもちろん、一般の新聞や雑誌にも「インターネット」の文字は頻繁に登場する状態と化している。それにともない、インターネットはどんどん普及し、広がっている。それは即ち、情報空間がどんどん拡大しているということであり、インターネット上に次々に大量の情報が蓄積され続けていることを示している。

インターネットに膨大な情報が格納されるにつれて、その中から欲しいデータを見つけ出すための検索技術に高い関心が集まるようになった。インターネット上にある情報というのは、誰か管理人のような人が存在していて、その人がきちんと整理して総合的・網羅的に提供しているようなものは極めて少なく、どちらかといえば、細かな情報の断片が大量に存在するとでも言えるような性質のものである。しかもその情報の質は玉石混淆の状態であり、その結果として、どこにどのような情報があるかを完全な形で把握するのは事実上不可能になってしまっている。無秩序と混沌が優勢をしめる情報空間の中に存在するインターネット上の情報を対象とした検索を行うために、もともと関係がなかった情報を、キーワードで関連づけ、クリック１つで実際にハイパーリンクで結びつけてしまおうという方法がとられている。これがいわゆる「サーチエンジン」である。

サーチエンジンには大きく分けて、ディレクトリ型とロボット型の２つがある。

ディレクトリ型のサーチエンジンとは、たくさんのページへのリンクを主題別に分類して提供するサービスである。例えば、芸術、ビジネス、教育のように分野別に整理して並べてある。

一方、必要な情報を発見するための方法として、古くから研究されている全文検索技術

を使って、キーワードによる検索サービスを提供する方法も広く普及している。こちらをロボット型のサーチエンジンと呼ぶ。Web ブラウザの一種である「ロボット」と呼ばれるソフトウエアを使って、インターネット上の Web ページを定期的に収集し、それらに対するインデクスの作成を自動的に行うことにより、検索機能を提供するサービスである。利用者は、主題に関するキーワードを自分で考え、検索フォームに入力する。システムの側では、入力されたキーワードを基に全文検索を行い検索結果を利用者に返す。検索処理を行うエンジン部分のことをが全文検索システムという。

全文検索システムは、特定の規則にのっとった複雑なキーワードの登録などを必要とせず、テキストデータのまま必要な情報を取り出すことができる。また一般に、検索対象の文書数の多寡にあまり大きく影響されず、高速であるという利点がある。検索結果を得るまでの応答速度は、文書数の対数（文書数を n とすれば、$\log(n)$）に比例する。従って、文書数が大量になるほど、全文検索システムの性能が効果的に現れる。

2.2 Namazu の概要

Namazu[1][2] は、日本で最も多く利用されている全文検索システムである（製作者は愛知大学の高木哲氏)。CGI として動作させることにより、小中規模の WWW 全文検索システムを構築することができるほか、ハードディスク内のファイルを対象としたパーソナルな用途にも使えるように設計されている。配布条件に関しては、UNIX 系プラットホームのフリーのソフトウエアで、一般的なライセンス方式である GNU 一般公有使用許諾書のバージョン２（GPL もしくは GPL2 と表記されることが多い）に従うフリーソフトウエアである。

Namazu には、フリーソフトウエアとしては画期的な多くの特徴がある。例えば、

- 各種プラットホームに対応
- 高速な検索を実現
- 多くの検索方法をサポート
- 検索結果表示の柔軟性
- 正確な HTML の取り扱い
- 複数インデクスからの検索をサポート
- 多彩な検索クライアント

などが挙げられる。

Namazu はインターネット上における様々な分野の様々なシーンで広く利用されている。例えば、Namazu を導入して全文検索サービスを行っている Web サイトには、民主党のような政党、京都府や沖縄県などの官公庁、国立国会図書館や国土地理院のような公共機関、全国の大学や各種研究機関、新聞社や出版社のような大量のデータを扱うマスコミ報道機関、一般企業などがあり、またフリーソフトである Namazu はソースコードが公開されているため、専門的な研究にも応用されている。

2.3 Namazu のシステム構成

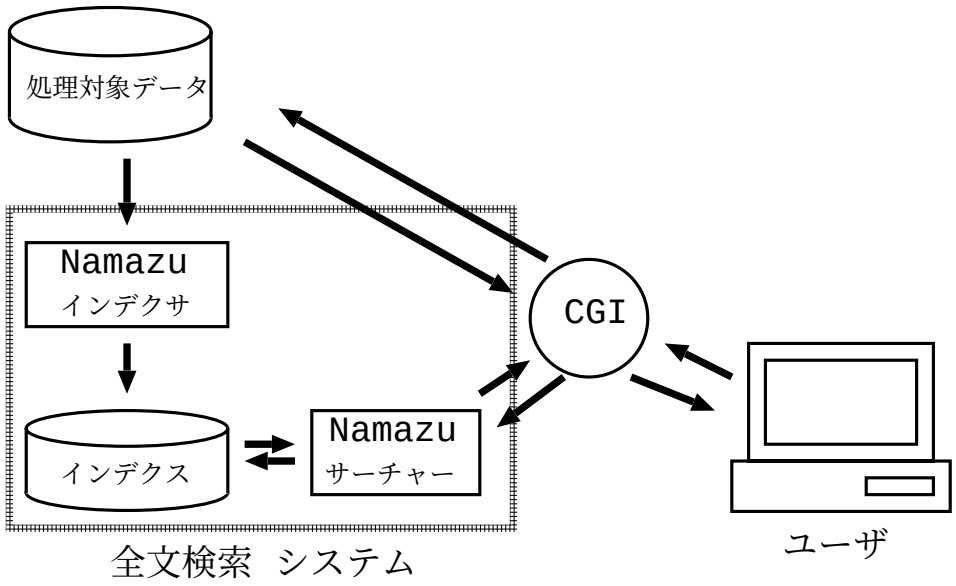


図 2.1: Namazu の構造

Namazu システム全体は以下のような要素から成る。（図 2.1）.

- 検索対象となるデータ（文書）
- 検索対象となる文書から単語を切出し、単語から文書への索引を作り出す部分（インデクサ）
- インデクサにより作り出された索引（インデクス）
- ユーザからの検索要求を受取り、インデクスを使用して、実際に検索を行う部分（サーチャー）

- ユーザから検索要求を受取り、それをサーチャーに手渡す仕事をし、またサーチャーからの検索結果を受取り、その検索結果と実際の検索対象となったデータとを結び付けて、ユーザに結果として表示する CGI
- 検索要求を出し、検索結果を受け取るユーザ

インデクスとは、英語では書籍の索引のことであり、本の巻末についているものである。Namazu のインデクスも、書籍の索引と同じ機能を持っている。Namazu の場合、インデクス作成（以後インデクシングと呼ぶ）でつくられるインデクスは、どの単語がどの文書（Web ページ等）に含まれているかを示すものである。通常の本の場合には、重要と思われる単語についてのみ著者がいくつか抜き出し、その単語についてのインデクスだけが作成される。一方、Namazu の場合は、文書に含まれる全ての単語についてのインデクスを作成することが通常行われる。検索処理自体は一般的なデータベースと同様に、ハッシュ法を用いたインデクスを使用するため、多少無駄な単語がインデクスに含まれていたとしても、それはインデクスの容量を増やすことにはなるが、そのことが検索スピードを遅くする要因にはあまりならない。

2.4 Namazu によるインデクスの作成と検索

前章で述べたように、Namazu のシステムはインデクサと検索部分とに大きく分けることができる。Namazu には、それぞれに対応するものとして、**mknmz** と **namazu** という 2 つのコマンドがある。mknmz は、インデクスを作成するプログラム（インデクサ）である。一方、namazu は、mknmz で作成したインデクスを使って実際に検索するコマンドである。

2.4.1 Namazu によるインデクスの作成

mknmz を実行してインデクスを作成するには、引数として、検索対象の文書が入ったディレクトリを指定する。例えば、
/home/db110/warehouse/entry/aaindex に検索対象の文書群があるとすれば、UNIX システムでは以下のようなコマンドを実行する。

%mknmz /home/db110/warehouse/entry/aaindex

これにより、/home/db110/warehouse/entry/aaindex 以下に含まれる全てのファイルを再帰的に探索してリストアップし、インデクシングが行われる。その結果、カレントディレクトリにインデクスファイルが作成される。

また、mknmz には多くのオプションが用意さてれいる。ここでそれらのオプションを紹介することにする。

-a	すべてのファイルを対象とする
-c	日本語の単語のわかち書きに ChaSen を用いる
-e	ロボットよけされているファイルを除外する
-h	Mail/News のヘッダ部分をそれなりに処理する
-k	日本語の単語のわかち書きに KAKASI を用いる
-m	ChaSen の形態素解析の品詞情報を利用する (名詞のみ登録)
-q	インデクス処理の最中にメッセージを表示しない
-r	man のファイルを処理する
-u	uuencode と BinHex の部分を無視する
-x	HTML のヘディングによる要約作成を行わない (文書の先頭から作成)
-D	Date:, From: といったヘッダを要約につけない (ディフォルトではつける)
-E	単語の両端の記号は削除する (ディフォルトでは含める)
-G	送り仮名を削除する (ディフォルトでは含める)
-H	平仮名のみの単語は登録しない (ディフォルトでは登録を行う)
-K	記号はすべて削除する (ディフォルトでは登録を行う)
-L	行頭・行末の調整処理を行わない (ディフォルトでは調整を行う)
-M	MHonArc で作成された HTML の処理を行わない (ディフォルトでは行う)

-P	フレーズ検索用のインデクスを作成しない (ディフォルトでは作成する)
-R	正規表現検索用のインデクスを作成しない (ディフォルトでは作成する)
-U	URL の encode を行わない (ディフォルトでは行う)
-W	日付によるソート用のインデクス作らない (ディフォルトでは作成する)
-X	フィールド検索用のインデクスを作らない (ディフォルトでは作成する)
-Y	削除された文書の検出を行わない (ディフォルトでは行う)
-Z	文書の更新/削除を反映しない (ディフォルトでは行う)
-A	.htaccess で制限されたファイルを除外する
-l（lang）	言語を設定する ('en' or 'ja')
-F（file）	インデクス対象のファイルのリストを読み込む
-I（file）	ユーザ定義のファイルをインクルードする
-O（dir）	インデクスファイルの出力先を指定する
-T（dir）	NMZ.head,foot,body.* のディレクトリを指定する
-t（regex）	対象ファイルの正規表現を指定する

インデクシングの結果、ディフォルトではカレントディレクトリに NMZ.i や NMZ.f といった名前のたくさんの種類のインデクスファイル群が生成される。各インデクスファイルを表 2.1として示す。

ファイル名	内容
NMZ.i	インデクスファイル (転置ファイル, inverted ファイル)
NMZ.ii	インデクスファイル seek 用インデクス
NMZ.f	文書のリストのファイル (各文書の情報を記録)
NMZ.fi	文書のリストのファイル seek 用インデクス
NMZ.h	キーワードの先頭 2 byte 用のハッシュテーブル
NMZ.r	インデクスに登録されているファイルのリスト
NMZ.head.en	検索結果出力用ヘッダファイル （英語）
NMZ.head.ja	検索結果出力用ヘッダファイル （日本語）
NMZ.foot.en	検索結果出力用フッタファイル （英語）
NMZ.foot.ja	検索結果出力用フッタファイル （日本語）
NMZ.body.en	キーワードが与えられなかったときのメッセージ （英語）
NMZ.body.ja	キーワードが与えられなかったときのメッセージ （日本語）
NMZ.msg.en	ロック時のメッセージ用ファイル（英語）
NMZ.msg.ja	ロック時のメッセージ用ファイル（日本語）
NMZ.log	インデクスの更新ログ
NMZ.slog	検索されたキーワードのログ
NMZ.lock	検索時のロックファイル
NMZ.lock2	インデクス作成時のロックファイル
NMZ.le	little-endian なインデクスのときに存在
NMZ.be	big-endian なインデクスのときに存在
NMZ.w	正規表現/中間/後方一致用の単語表
NMZ.p	フレイズ検索用のインデクス
NMZ.pi	フレイズ検索用のインデクスのインデクス
NMZ.t	文書のタイムスタンプ、欠番の情報を記録
NMZ.field.subject,from,date,message-id,...	フィールド検索用のインデクスのインデクス

表 2.1: インデクスファイル群（NMZ.*）の説明

2.4.2 Namazu を使ったキーワード検索

Namazu では、先ほどインデクサの mknmz で作成したインデクス（NMZ.*）を使って検索を行う。コマンド名は namazu である。namazu の実行方法は、１つ目の引数として、検索したいキーワードをダブルクオート” ” で囲んで指定する。そして、２つ目の引数として、インデクスが置かれているディレクトリを指定する。例えば、···/genomeDB/omim/index に検索したい対象のデータのインデクスがあり、”alzheimer” のキーワードで検索したい場合は次のようになる。

%namazu “alzheimer” ···/genomeDB/omim/index

namazu にも mknmz 同様、多くのオプションが用意されている。ここでそれらのオプションを紹介する。

-n（num）	一度に表示する件数
-w（num）	表示するリストの先頭番号
-s	短いフォーマットで出力
-S	もっと短いフォーマット (リスト表示) で出力
-v	usage を表示する (この表示)
-f（file）	namazu.conf を指定する
-h	HTML で出力する
-l	新しい順にソートする
-e	古い順にソートする
-a	検索結果をすべて表示する
-r	参考ヒット数を表示しない
-o（file）	指定したファイルに検索結果を出力する
-C	コンフィギュレーション内容を表示する
-H	先の検索結果へのリンクを表示する

-F　　⟨FORM⟩ ... ⟨/FORM⟩ の部分を強制的に表示する

-R　　URL の置き換えを行わない

-U　　plain text で出力する時に URL encode の復元を行わない

-L（lang）　　メッセージの言語を設定する ja または en

（この章の記述は参考文献 [1]、[2] に拠る。）

第 3 章

ベクトル空間法

前述のように、近年の著しい Web の隆盛により、大量の情報の中から必要なものを選びだすための情報検索技術が注目されるようになった。ここで、その情報検索技術について説明することにする。

3.1 情報検索の概念

多数の情報の中から求める情報を探し出すことを、情報検索（information retrieval）と呼び、とくにその対象がテキスト情報である場合はテキスト検索（text retrieval）と呼ぶ。情報検索の過程を図 3.1として示す。

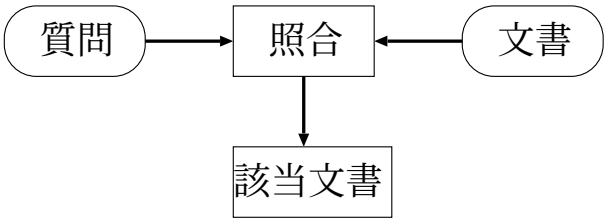


図 3.1: 情報検索の概念図

検索者は調べたい事柄をもっている。例えば「自然言語処理に関する書籍を探したい」と考えている。これを検索意図と呼ぶ。一方、検索の対象として、書籍や論文といった文書群や Web ページ等がある。情報検索システムはユーザの検索意図に適合する文書群を検索し、検索結果として出力する。

ここで問題となるのが、検索意図と文書の照合である。例えば先の検索意図の場合、この照合において各文書が「自然言語処理に関する書籍」かどうかを判定しなければならない。しかし、図 3.1のままで実現することは難しい。第一に検索意図に関する問題である。「自然言語処理に関する書籍を探したい」というように自然言語の文として質問するならば、検索システムはその文を理解し、検索意図を把握することが必要である。第 2 の問題は文書群の問題である。こちらも同様に自然言語の文章として与えられるため、その文章内容をコンピュータが理解することは非常に困難である。

これらの問題を解決するために、検索意図と文書をそのままの形で照合するのではなく、処理しやすい中間的な媒体を設定し、それを用いて照合を行う。この中間的な媒体として語（単語、複合語）の集合を用いるのが標準的である。その情報検索の過程を図 3.2 として示す。

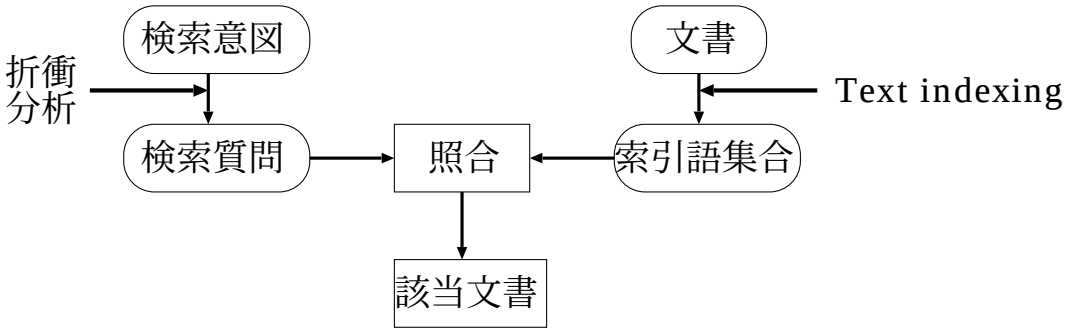


図 3.2: 索引語を用いた情報検索

すなわち、

(1) 各文書に対してその文書の内容を表す索引語（index term）を割り当てる。これをテキストに対する索引語付け（indexing）と呼ぶ。

(2) 検索者は検索意図を索引語とその組み合わせによって表現する。こうして表現された検索意図を検索質問（query）と呼ぶ。

(3) 各文書に付けられた索引語集合と検索者から与えられた検索質問の照合を行い、該当文書を見つける。

この方法の背景となっているのは、「テキストが表現している内容を語の集合という形で近似的に表現できる」という考え方である。機械処理という観点から考えるとき、テキ

ストの内容というものは捉え難いものであり、これを直接処理することは非常に困難である。しかし、これを語の集合によって近似的に表現すれば、かなり色々な処理が可能になり、近似表現上での照合により求める内容が書かれているテキストを見つける処理が実現できる。

検索意図を索引語とその組み合わせによって表現する方法として、ブーリアン演算子を使う方法がある。これは、ブーリアン演算子を使用することにより、質問の単語間の関係を表す方法である。

(1) **or 演算子**　事実上、同義語として２つの単語が取り扱われる。「単語１ or 単語２」と質問が与えられたとき、どちらか一個が存在すれば良い。

(2) **and 演算子**　両方兼ね備えたものを検索する際に使用する。「単語１ and 単語２」というような質問の時、両方の単語を指し示す。

(3) **not 演算子**　not 演算子は制限するもの、単語の範囲を狭くするものである。普通、連言と一緒に使い、特定の単語の該当範囲を制限するのに用いられる。「単語１ and not 単語２」のような検索質問から導き出される結果は、単語１が該当するすべての文書から単語２が含まれる文書を除いたものになる。

3.2　インデクス

情報要求が索引語の束であろうと、ブーリアン演算子と索引語の組み合わせであろうと、探索操作はその索引語の組み合わせが存在するかどうか、蓄えられた文書群にたいして確認しなくてはいけない。様々な探索方法、データ構造が使えるが、高速ファイルをアクセスし、図 3.2の情報検索の方法を実現するため、転置インデクスを作成する方法がとられる。文書、検索要求の内容を反映していると考えられる索引語（index term）を抽出し、その集合をデータ構造とするものである。

検索対象となる文書群にはそれぞれの文書ごとにその文書の内容を代表する索引語の集合が割り当てられる。そして、キーワード検索、索引語による検索を行う際は、ある索引語がどの文書にあるのかを知りたい。そこで、どの索引語がどの文書にあるか分かるように、ある索引語が、どの文書に登場するのかを示すインデクスを作成する。インデクスには、ある索引語が出現する文書と、その出現頻度を記録しておく。このようなインデクスを転置インデクスと呼ぶ。前述した全文検索システム Namazu も、インデクサ mknmz

により、この形式のインデクスを作成する。これにより、検索質問との照合計算を高速に行うことを実現する。転置インデクスを表3.1として示す。

	文書１	文書２	文書３	文書４
索引語１	2	0	5	0
索引語２	5	3	0	0
索引語３	6	1	5	2
索引語４	0	8	2	1

表 3.1: 転置インデクス

このような転置インデクスを用いると、例えば「索引語１ and 索引語２」に該当する文書を求めることは、次のようにしてそれぞれの索引語に対する行ベクトルの論理積を求めることになる。

索引語１	１０１０ ＝	{文書１, 文書３}
索引語２	１１００ ＝	{文書１, 文書２}
索引語１ and 索引語２	１０００ ＝	{文書１}

この方法により、検索質問に出現する索引語に対する行ベクトルだけから、検索質問に適合する該当文書を決定することができる。上記の例では索引語数、文書数ともに少数であるが、実際の検索ではどちらも非常に大きな数になる。このため、転置インデクスはかなり大きな表になるが、検索質問に対する該当文書を求める際にはそのほんの一部しか参照する必要はない。このため非常に高速な検索が可能となる。

3.3 索引語の重み

転置インデクス法だけでは索引語がどの程度テキスト中で重要かを表現していない。また、該当文書にどの文書がより検索意図に近いかの順位付けをつけることができない。その問題の解決法として一般に、文書、質問ともに重要度に応じて索引語に重みづけが行われる。これにより、他の文書の中から目的の文書を検索するために、より重要な索引語と、より重要でない索引語を識別することができる。例えば、表3.2のような重みが与え

られたとする。この時、{ 索引語２, 索引語３ } のように複数の索引語からなる索引語からなる検索質問に対する答えは表 3.3のように計算される。

	索引語１	索引語２	索引語３	索引語４
文書１	0.2	0.5	0.6	0
文書２	0	0.3	0.1	0.8
文書３	0.5	0	0.5	0.2
文書４	0	0	0.3	0.3

表 3.2: 文書と重み付づけられた索引語

	索引語２		索引語３			順位
文書１	0.5	+	0.6	=	1.1	1
文書２	0.3	+	0.1	=	0.4	3
文書３	0	+	0.5	=	0.5	2
文書４	0	+	0.3	=	0.3	4

表 3.3: 重み付けられた索引語を利用した検索

索引語の重要度を表す重みは、重要度の高いものには大きな値を、低いものには小さな値を与える。

3.4 ベクトル空間法

3.1節、3.2節、3.3節では、現在一般的に使われている情報検索システムの考え方、作成されるインデクスについて説明した。これは、Namazu にも使われている考え方であり、Namazu で作成されるインデクスは転置インデクスである。

現在、通常使われている全文検索システムは、複数のキーワードを AND や OR 等の演算子を組み合わせて検索するものである。しかし、本章で説明するベクトル空間法ではキーワードの代わりに文章を索引語として利用できるものである。また、索引語を直接含んでいない文章でも概念が似ていると思われればヒットする（類似情報抽出）という特徴を持つ。

ベクトル空間法（vector-space model）とは、文書と検索質問の両方をある統一的な表現によって表し、この間に類似度（similarity）を定義することによって似ている文書を探し出す方法である。

文書と検索質問の両方に使用するある統一的な表現とは、文書の内容を、その文書に出現する索引語と、その出現頻度のベクトルで表現するものである。これは、「テキストが表現している内容を語の集合という形で近似的に表現できる」という考え方に基づく。つまり、検索対象となる文書群に、それぞれの文書ごとにその文書の内容を代表する索引語の集合が割り当てられ、その出現頻度で、その文書の内容を表すのである。即ち、表 3.4 のような文書の内容を表すようなベクトルを各文書ごとに作成することによって各文書の内容を表す。このベクトルを文書ベクトルと呼ぶことにする。ベクトル空間法による情報検索を行う際には、この形式のインデクスを作成する必要がある。

	索引語 1	索引語 2	索引語 3	索引語 4
文書 1	2	5	6	0
文書 2	0	3	1	8
文書 3	5	0	5	2
文書 4	0	0	2	1

表 3.4: 文書の内容を表したベクトル

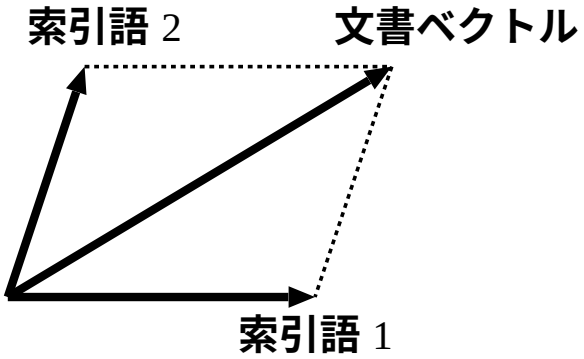


図 3.3: ベクトル空間法における文書の表現

この文書ベクトルを使用し、ベクトル演算をすることにより、文書間の類似度を計算する。ベクトル空間法は t 個の索引語のそれぞれを意味の基本単位として考え、それらの線

形結合として文書や検索質問の意味内容を表現しようという方法である。

ベクトル空間において、２つのベクトルの類似度はいろいろな形で定義できる。ここでは、本研究で採用した内積を用いる方法について説明する。文書ベクトルを V_1 と V_2 とすると、文書ベクトル V_1 と V_2 の間の類似度 $sim(V_1, V_2)$ は次のように表す。

$$sim(V_1, V_2) = \frac{V_1 \cdot V_2}{|V_1||V_2|} \tag{3.1}$$

$$= \frac{\sum_{i,j=1}^{t} a_i V_i b_j V_j}{\sqrt{\sum_{i=1}^{t} V_i^2}\sqrt{\sum_{j=1}^{t} V_j^2}} \tag{3.2}$$

V_1, V_2 ： 文書ベクトル

V_i ： 文書ベクトル V_1を構成する索引語 i の出現頻度

V_j ： 文書ベクトル V_2を構成する索引語 j の出現頻度

a_i ： 文書ベクトル V_1に対する索引語 i の重要度

b_j ： 文書ベクトル V_2に対する索引語 j の重要度

この式を用いることにより、２つの文書間の類似度を測り、２つの文書の内容が似ているか似ていないかを判断する。

（この章の記述は参考文献 [4]、[5] に拠る）

第 4 章

ゲノムデータベース

4.1　ゲノムネットの概要

ワトソンとクリックによる DNA の 2 重らせん構造の発見 [10]（1953 年）以来、遺伝現象の物理的理解が急速に進展し、生物学の最前線はさまざまな遺伝子の発見競争の場となった。その延長で、1985 年頃からヒトのもつ全遺伝情報を決定してしまおうという機運が米国などで高まり、ゲノムプロジェクトが開始された。このプロジェクトを思いついたのは、1984 年当時カリフォルニア大学サンタクルーズ校の学長であった R. シンスハイマー博士である。[12] そして、1985 年 5 月、サンタクルーズで最初の会議が開かれた。1988 年、技術調査局からヒトゲノムプロジェクトの見通しに関するレポートが、アメリカ議会に提出された。そして 1989 年、1 億ドルの予算でスタートされた。同様に英・仏・日・独といった国々でも始まり、国際共同研究となった。また同時に、バクテリアから高等動植物まで、様々な生物種でも、全遺伝情報の決定も行われている。[12][8]

遺伝情報とは、親から子に伝えられる情報であり、体を形づくる細胞がすべて備えているものである。[9] その情報は DNA（デオキシリボ核酸）という化学物質の分子構造、A（アデニン：Adenine）, T（チミン：Thymine）, G（グアニン：Guanine）, C（シトシン：Cytosine）という DNA の塩基の並び（塩基配列）によって決定される。ある生物がもつ DNA の塩基配列の全体あるいはその一部は、ゲノム情報と呼ばれる。

現在、プロジェクトはおおむね計画通りに進んでおり、すでに大腸菌、酵母、線虫などの情報は決定済みである。国際共同研究として始まったゲノムプロジェクトは、いまや国際戦争であり、各国政府が投入する膨大なゲノム関連予算に対し、民間企業もさらに大規模な投資を行い、早く情報をおさえ知的所有権を獲得するために、ゲノムの塩基配列決定（シークエンシング）競争に駆り立てられている。[11] そのような状況のなか、ヒトの情

名前	機関	アドレス
ゲノムネット	京都大学化学研究所	www.geome.ad.jp
NCBI	米国バイオテクノロジー情報センター	www.ncbi.nlm.nih.gov
EBI	欧州バイオインフォマィックス研究所	www.ebi.ac.uk
SIB	スイスバイオインフォマティックス研究所	www.expasy.ch

表 4.1: 世界の主要なゲノム関連サーバ

報も 2000 年春には 90%が決定される見通しである。この解読作業には最新の情報科学の成果が応用され、巨大で複雑なゲノム情報を扱うことで、情報科学自体にも新たな発見が促される可能性があると期待されている。

このようなゲノムプロジェクトは、加速度的に大量の配列データを産み出すことになった。それに伴い分子・細胞レベルでの生命現象に関する基礎データから、病気の診断・治療への可能性を示す応用データまで、さまざまな生物学的なデータが急速に蓄積されつつある。生物学はかつて直面したことのない大量情報の時代に入ったのである。これら、大量のデータを利用するためにデータベース化する必要がでてきた。こうして作られたデータベースが**ゲノムデータベース**である。ゲノムデータベースには、DNA の塩基配列以外の情報、例えば、たんぱく質の情報なども含まれている。

1991 年度より開始された文部省ヒトゲノムプログラムの下で、京都大学化学研究所と東京大学医科学研究所ヒトゲノム解析センターの研究スタッフは、生物学の新展開に対応するためには情報インフラストラクチャーの整備が不可欠であるとの認識から、**ゲノムネット（GenomeNet）**と名付けたコンピュータネットワークの構築と運用を行ってきた。ゲノムネットは世界中に存在する生物学・医学関連の多用なデータベースを各研究者のデスクトップで統合して利用できる環境をつくり、世界に先駆けて次の新しいデータベースを構築することを目的として構築されている。ゲノムネットは世界的にみても主要なサーバの１つである。世界の主要なゲノム関連サーバを表 4.1として示す [13]。

4.2 ゲノムネットに含まれるゲノムデータベースの種類と量

ゲノムネットは、生物学の分野では文献情報の他に、ゲノムの地図と塩基配列、タンパク質のアミノ酸配列と立体構造、代謝系や制御系の分子ネットワーク、神経系や免疫系における細胞のネットワーク、そして発生・分化・老化や疾病に関する個体レベルのデータ

データベース名	データの内容	エントリ数
AAindex	アミノ酸指標	500
BRITE	分子間相互作用	278
COMPOUND	代謝化合物	5628
EMBL	核酸塩基配列	3952878
ENZYME	酵素反応	3705
EPD	プロモータ配列	1356
Genes	遺伝子カタログ	107118
GenBank	核酸塩基配列	4028171
LITDB	文献	298877
OMIM	遺伝病	10573
PDB	立体構造予測	9254
PDBSTR	アミノ酸配列	15639
PIR	アミノ酸配列	122810
PMD	変異たんぱく質	7078
PRF	アミノ酸配列	118187
PRINTS	配列モチーフ	865
PROSITE	配列モチーフ	1374
SwissProt	アミノ酸配列	80000
TRANSFAC	転写因子	9737

表 4.2: ゲノムネットが提供するデータベース

など、多種多様なデータがデータベース化されている。これらは相互に深く関連し、また頻繁に更新されている。これら多種多様なデータは、それぞれ独自のフィロソフィーで作成され、一つ一つ独立したデータベースとして存在する。 ここに、ゲノムネットに存在するデータベースを表 4.2として示す。各々のデータベースでは、一つ一つの遺伝情報は、エントリと呼ばれる単純なファイル（フラットファイル）に一つ一つ格納されており、表 4.2に書かれているエントリ数とは、その数である。

4.3　ゲノムデータベースのデータの様式

4.2節にも登場したが、ゲノムネットに含まれるゲノムデータベースでは、一つ一つの遺伝情報はエントリと呼ばれる単位で取り扱われている。各エントリにはエントリ名（またはアクセッション番号）と呼ばれるデータベース内でユニークな名前がつけられ、取り扱われている。従って、

データベース名：エントリ名

の組を与えてやりさえすれば、世界中に存在する数多くのデータベースを統合的に参照することが可能である。

ここで、エントリの中の様子について説明する。GenBank というデータベースのエントリの中の一つを図 4.1として示す。

```
LOCUS       MABRRC       1332 bp    rRNA            BCT       26-JUL-1993
DEFINITION  Marine Eubacterial sp. (FL5) PCR generated ribosomal RNA fragment.
ACCESSION   L10936
VERSION     L10936.1  GI:308917
KEYWORDS    ribosomal RNA.
SOURCE      Marine Eubacterial sp. rRNA.
  ORGANISM  unidentified marine eubacterium
            Bacteria; environmental samples.
REFERENCE   1  (bases 1 to 1332)
  AUTHORS   Delong,E.F., Franks,D.G. and Alldredge,A.L.
  TITLE     Diversity of aggregate-attached versus free-living marine bacterial
            assemblages
  JOURNAL   Limnol. Oceanog. (1993) In press
FEATURES             Location/Qualifiers
     source          1..1332
                     /organism="unidentified marine eubacterium"
                     /db_xref="taxon:28248"
     rRNA            <1..>1332
                     /gene="rRNA"
                     /product="ribosomal RNA"
     gene            1..1332
                     /gene="rRNA"
BASE COUNT      358 a    287 c    399 g    284 t      4 others
//
```

図 4.1: GenBank のエントリ

図 4.1のように、ゲノムデータベースに保管されている遺伝情報は、“LOCUS”, “DEFINITION”, “ACCESSION” 等のように各フィールドごとに分けて格納されている。また、ゲノムデータベース内のデータは、自然言語で書かれている部分と、DNA の塩基配列情報のような、ただの数値や文字列情報である部分とがある。ゲノムデータベースのコンテンツの主体は、この数値や文字列情報であり、自然言語情報はあまり積極的に使われることはない。ここで、本研究で対象となったゲノムデータベースについて、それらのフィールドについて一部例を挙げ、その書いている内容について説明し、そのフィールドについての自然言語情報の多寡を評価する。自然言語情報の多寡は ♣♣♣♣♣ と、5 段階評価で示す。

また、付録に、本研究で対象となった全てのゲノムデータベースのフィールドについて示す。

PROSITE

DR

自然言語情報：♣

```
DR   P33395, RRF2_DESVH, T; Q48660, Y156_LACLA, T; O68025, Y166_RHOCA, T;
DR   Q55433, Y846_SYNY3, T; O07465, YB01_RHOPA, T; Q10613, YC87_MYCTU, T;
DR   P77484, YFHP_ECOLI, T; P44675, YFHP_HAEIN, T; O07573, YHDE_BACSU, T;
DR   P21498, YJEB_ECOLI, T; P40610, YJEB_VIBPA, T; Q51134, YLDA_NEIME, T;
DR   O69219, YOR2_AZOVI, T; O34527, YRZC_BACSU, T; P71047, YWGB_BACSU, T;
```

PROSITE の DR フィールド（UPF0074 エントリ）

フィールド内容：DR = Cross-reference to SWISS-PROT
1 エントリに 0 個以上記述 [14]。

RU

自然言語情報：♣♣♣

```
RU   Additional rules:
RU   (1) The cysteine must be between positions 15 and 35 of the sequence in
RU       consideration.
RU   (2) There must be at least one charged residue (Lys or Arg) in the first
RU       seven residues of the sequence.
```

PROSITE の RU フィールド（PROKAR_LIPOPROTEIN エントリ）

フィールド内容：RU = Rule
1 エントリに 0 個以上記述。

DE

自然言語情報：♣♣

```
DE   Neutral zinc metallopeptidases, zinc-binding region signature.
```

PROSITE の DE フィールド（ZINC_PROTEASE エントリ）

フィールド内容：DE = Short description
1 エントリに 1 つ記述。

第 5 章

関連研究

5.1 はじめに

本研究の関連研究として、大阪大学で行われた**「分子生物学データベースにおける単語の共起性と出現位置による関連エントリ探索手法」**[16] について説明する。この研究は異なる種類のデータベース間の関連するエントリを統合的に検索することを目的として行われた。そこで、検索手法として、3.4節で説明したベクトル空間法を用いており、エントリそのものが持つ単語の出現位置と出現頻度により、複数のデータベースのエントリ中に共通して出現する単語を用いてエントリ間の類似度を定義し、関連するエントリを探索する手法を提案している。また、この研究では GenBank と SwissProt の 2 つのゲノムデータベースでのみ、対象として研究を行っている。

5.1.1 検索手法

5.1.2 抽出対象の絞り込み

分子生物学で用いるデータベースのエントリは、一般にサイズが非常に大きく、DNA やアミノ酸配列そのものも含まれているため、そのまま扱うには無駄が多いとの理由により、まず最初に、あらかじめ単語の抽出対象となるフィールドのみを抜き出したテキストファイルを作成し、実際の処理はそのファイルに対して行っている。

5.1.3 単語の抽出

前述のテキストファイルから、区切り記号で区切られた単語を直接抽出している。この研究で 定義する区切り記号とは、以下の通りである。

{“ ”,”!”, “”, ”#”, ”$”, ”%”, ”&”, ”(“, ”)”, “[”, “]”, “{“, “}”, “*”, “+”, “-”, “,”, “/”, “:”, “;”, “<”, “=”, “>”, “?”, “@”, “\”, “|”}

単語の重要度の算出

抽出された単語に対し、その出現頻度によって重要度を付けている。この大阪大学の研究では Inverse Document Frequency 法（以後、IDF 法と呼称する）を用いている。IDF 法を用いると、データベース DB_n 中に N_{DB_n} 個のエントリがあるとき、c_i 個のエントリに含まれる単語 k_i の重要度 w_i は式 5.1のように定義される。

$$w_i = \log\left(\frac{N_{DB_n}}{c_i}\right) \tag{5.1}$$

ベクトル空間法を用いた関連エントリの探索

関連エントリを探索する方法としてベクトル空間法を用いている。

エントリ E_x をベクトル形式で表すと、(5.2) のように表すことができる。

$$E_x = (w_{x1}, w_{x2}, w_{x3}, \cdots, w_{xk}, \cdots, w_{xM_x}) \tag{5.2}$$

但し、w_{xk} はエントリ E_x 中に単語 k が存在すれば k の重要度は式 5.1で与えられ、存在しなければ 0 である。また、M_x とはエントリ E_x 中に出現する索引語の数である。

次に、2 つのデータベース DB_i、DB_j があるとする。M を DB_i、DB_j のどちらか、あるいは両方に出現する単語数であると定める。(5.2) のベクトルは、DB_x に出現しない単語 k に対して $w_{xk} = 0$ を与え、どちらも索引語の数は M_x とし、M_x 次元のベクトルに拡張する。

これを用いて、DB_i 中にあるエントリ E_i と DB_j 中にあるエントリ E_j の類似度 $Sim(E_i, E_j)$ を、ベクトル E_i と E_j がなす角 θ としたとき、$\cos\theta$ と定める。即ち、下記のような式で表せる。

$$Sim\,(E_i, E_j) \;=\; \frac{(E_i \cdot E_j)}{|E_i| \cdot |E_j|} \tag{5.3}$$

$$= \frac{\sum_{k=1}^{M} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{M} (w_{ik})^2 \sum_{k=1}^{M} (w_{jk})^2}} \tag{5.4}$$

5.2 関連研究が提案する探索精度向上手法

5.2.1 数値を含む複合語への対応

データベースから抽出した単語には、数字も多く含まれている。これらの数字は単独で意味を成すことは少なく、多くの場合は前後の語とつながった複合語の形で意味を成す場合が多い。また、一方のデータベースでは隣接語と完全にくっついて 1 語になっているのに対し、他方ではスペースやハイフン等の区切り記号によって 2 語に分けられているケースも存在する。

従って、隣接する 2 語を 1 つの単位、複合語として考える。そして、その複合語がどれくらいエントリに出現するか頻度統計をとり、その 2 語が実際に隣接して出現した回数以上出現する確率を求める。その確率が著しく低いものは、統計的に見て、たまたま隣り合っているとは考えにくく、前後の単語に何らかの関係があるとみなし、複合語として取り扱う。

5.2.2 同義語への対応

語尾活用が起こっただけで同一の意味を表す語を、異なる単語と認識してしまうのを防ぐ。語の構成として活用変化の起こる部分を語尾、語尾のつく基幹部を語幹とすると、語尾により品詞を変えながらも同じ意味の語として機能する場合は多い。そこで、以下のように同義語を定め、取り扱う。

1. 接尾語辞書を用意し、全エントリ中の全ての単語について単語の末尾と辞書を比較する。複数の接尾語と一致する場合は、もっとも長い接尾語を使用する。(例:activation の場合、-ation と -ion の 2 つの接尾語とマッチするが、長い-ation を採用する)

2. ある単語から接尾語を取り除いたものを語幹とする。但し、一致する接尾語がない単語に関しては、単語自身を語幹とする。

3. 語幹と全エントリ中の全ての単語を比較し、先頭が一致するものを語幹ごとにリスト化する。リスト内には、同じ語幹を持つものが並んでいる。ここで、同一リスト内に含まれる単語を「仮の同義語」であるとする。この時点では必ず最短の語幹にマッチしてしまうので、例えば reg-ion と regul-ate が同じリストに入っている。

4. 仮の同義語の中で、その語の実際の語幹に注目し、お互いの語幹を語の先頭に持つ語のみを真の同義語と判定する。

この実装に対しては、一般に検索時に同義表現をまとめて加味するか、登録時に同義のものをまとめるかに大別されるが、ここでは個別の検索時の時間を優先させるため、事前に同義語を加味した単語ベクトルを登録している。

フィールドごとの対応

実験に際しては、エントリに登場する全てのフィールドを使用している訳ではない。5.1.2節の抽出対象の絞り込みの段階で、採用するフィールドを絞ってある。その採用するフィールドの中で、より平文に近い部分を採用した際の冗長性がもたらす弊害を緩和するために、データベース全体をキーワード列挙部分とコメント部分に分け、それぞれについて別々に単語の重要度を計算し、ベクトル空間法を適用する際、どちらの部分に含まれたかによって、ベクトルの要素に重みづけをして類似度を算出している。

5.3 本研究との相違点

大阪大学での研究は、異なる種類のデータベース間の関連するエントリを統合的に検索することを目的として行われた。エントリを検索するということは、似たエントリ群、関連するエントリ群を集めてくるということである。それは、本研究の目的とする関連性を抽出する手法を探る研究に通じるものがあると考え、この大阪大学での研究を、本研究を進める上で、参考にした。しかしながら、大阪大学の研究と、本研究の間には以下の相違点がある。

- 大阪大学での研究は、ゲノムデータベースを Genbank と SwissProt の 2 つに限定して研究を行っている。しかし、本研究ではゲノムデータベースから関連性を抽出する手法を探ることが目的であり、2 つのゲノムデータベースで行うのは不十分であると考えた。そこで本研究では、12 種類のゲノムデータベースについて、エントリ間の類似度を調べている。

- 大阪大学での研究は、5.1.2節の抽出対象の絞り込みで述べたように、エントリ中の全てのフィールドを使用してエントリのベクトル化を行っている訳ではない。しかし、本研究では、6.4節で述べたような例外はあるものの、基本的に全てのフィールドを使用し、エントリをベクトル表現し、類似度計算を行っている。

- 大阪大学での研究では 5.1.2節のフィールドごとの対応で述べたように、データベース全体をキーワード列挙部分とコメント部分に大きく 2 つに大別している。しかし、フィールドには遺伝子名を記したものもあれば、キーワードを記したものもあり、フィールドごとに書かれている情報は異なる。フィールドを 2 種類に大別してしまっては、エントリを特徴づける重要な情報が記されたフィールドと、さほど重要でないフィールドとが同列に扱われてしまい、各フィールドのエントリに対する意味が曖昧になってしまうと考えた。そこで、本研究では、データベース全体をフィールドごとに分けて取り扱い、フィールドごとに類似度計算をすることも行っている。

第 6 章

ゲノムデータベースへの Namazu の適用

6.1 はじめに

3.4節で述べたように、ベクトル空間法を使用してエントリ間の類似度を計算するには、各エントリを、エントリに含まれる索引語と、その出現頻度のベクトルで表す必要がある。そのために、ゲノムデータベースに Namazu を適用し、インデクスを作成させる。そして、Namazu が作成したインデクスから、エントリの内容を表すベクトルを作成することを試みる。これにより、Namazu に単語の切出しをまかせることができる。

6.2 ゲノムデータベースに Namazu を適用する際の問題点

ゲノムデータベースのデータは、一つ一つの遺伝情報ごとにエントリと呼ばれる単純なファイル（フラットファイル）に格納されている。このことは、4章でも述べた。ところがエントリは、1 つのエントリが 1 つのテキストファイルに収められて格納されている訳ではない。ゲノムネットで提供される各々のデータベースにおいて、各々のデータベースに属するエントリが 1 つのテキストファイルに全て連続して書き連ねてられて、格納されている。すなわち、1 つのデータベースにおける全てのエントリが、1 つのテキストファイルに保管されているのである。

ゲノムデータベースに Namazu を適用する際、このことが問題となる。Namazu は 2 章、3章で説明したように mknmz で転置インデクスを作成することにより、高速検索を可能にしている。転置インデクスは表 3.1に示しているように、ある索引語がどの文書にいくつ含まれているか書かれたものである。よって、いくつかの検索したい文書が必要となる（文書が 1 つしかない場合、検索する必要がなくなる）。Namazu は、1 つのファイ

ルを 1 つの文書として認識する（どこにどんなファイルが存在するのか検索するシステムである）。従って、全てのエントリが収められたファイルをインデクシングしてもあまり意味がない。このようなことを行っても、全てのエントリが収められたファイルがどんなキーワードをいくつ持つか調べることができるだけである。本研究では、一つ一つのエントリの間に関連性があるかないかを調べる必要がある。すなわち、一つ一つのエントリ間の類似度を調べる必要がある。よって、各エントリで、どんな索引語が、いくつ存在するのかを調べる必要がある。ゆえに、各データベースごとに全エントリが記されたファイルから、各エントリごとに、そのエントリのみについて記された 1 つのファイルを作成せねばならない。すなわち、各データベースごとに全エントリが記されたファイルを、エントリごとに切り分ける必要がある。エントリを検索する上でもこの作業は必要となる。

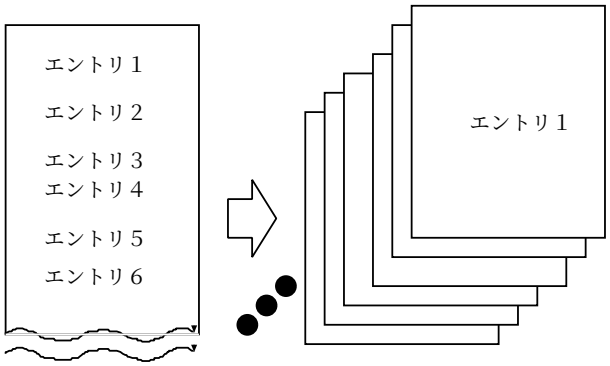


図 6.1: エントリのスプリット

6.3 エントリ及びフィールドの切り出し

6.2節で話した理由により、全てのエントリが記されたファイルを、各エントリごとに1つのファイルに格納するように切り分ける（本研究ではエントリのスプリッティングと呼ぶことにする）。この作業には研究室で別途制作された entry-splitter.pl というプログラムを使用して行った。次に entry-splitter.pl の機能について説明する。

entry-splitter.pl は、全てのエントリが収められたファイルから、1つのエントリごとに1つのファイルに切り分けて収めることができる。本研究では、各ゲノムデータベースごとに、全てのエントリが収められたファイルから、同じく各ゲノムデータベースごとに、1つのエントリごとに1つのファイルに切り分けて収めている。その結果を図6.2に示し、そのデータ構造を図6.3に示す。

```
[tkataoka@ks27e1u10] 56 % pwd
/home/db111/warehouse/entry/aaindex
[tkataoka@ks27e1u10] 57 % ls
ALTS910101    FASG760102    KOSJ950115    OOBM850103    RADA880101
ANDN920101    FASG760103    KRIW710101    OOBM850104    RADA880102
ARGP820101    FASG760104    KRIW790101    OOBM850105    RADA880103
ARGP820102    FASG760105    KRIW790102    OVEJ920101    RADA880104

⋮
⋮
```

図 6.2: AAindex における 1 ファイル 1 エントリに分割かれたエントリ群

また、entry-splitter.pl は各ゲノムデータベースごとに、そのゲノムデータベースのエントリに記載されているフィールドの内容を、エントリごとにその対象のフィールドだけを抜き出して、1つのファイルとして分割することができる。その結果を図6.4、図6.5に示す。また、その階層構造を図6.6に示す。

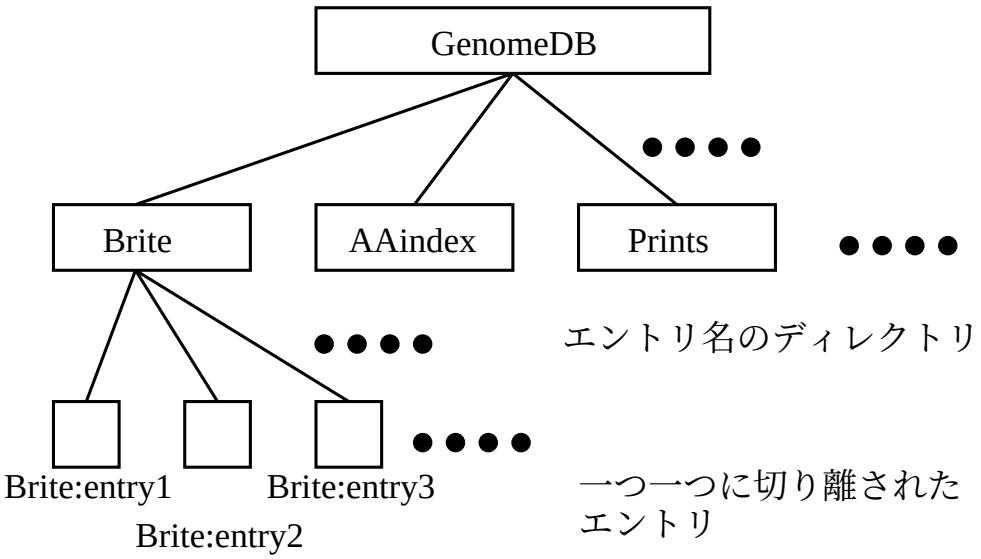


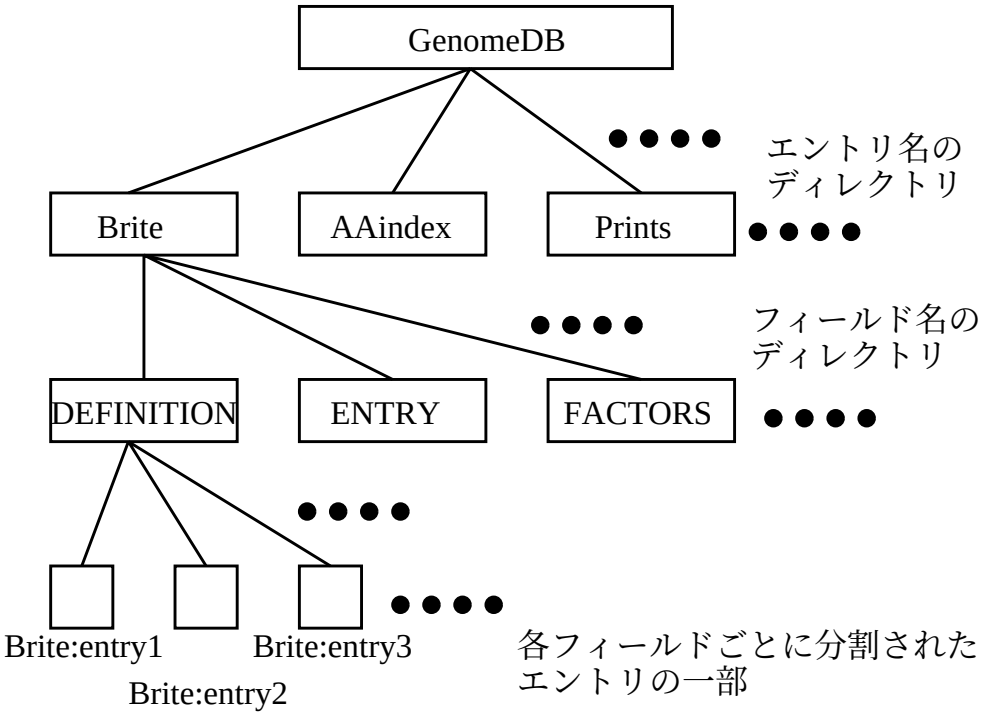
図 6.3: エントリごとに分割されたエントリのデータ構造

```
[tkataoka@ks27e1u10] 85 % pwd
/home/db110/warehouse/field/aaindex
[tkataoka@ks27e1u10] 86 % ls
A/  C/  D/  H/  I/  J/  R/  T/
```

図 6.4: AAindex におけるフィールドごとに分割したエントリ群を収める、各フィールド名のディレクトリの様子

```
[tkataoka@ks27e1u10] 89 % pwd
/home/db110/warehouse/field/aaindex/A
[tkataoka@ks27e1u10] 91 % ls
ALTS910101   FASG760102   KOSJ950115   OOBM850103   RADA880101
ANDN920101   FASG760103   KRIW710101   OOBM850104   RADA880102
ARGP820101   FASG760104   KRIW790101   OOBM850105   RADA880103
ARGP820102   FASG760105   KRIW790102   OVEJ920101   RADA880104
                                    ⋮
                                    ⋮
```

図 6.5: AAindex におけるフィールドごとに分割したエントリ群



6.4　エントリの切り出しを行う際における例外的な処理

本研究での例外的処置として、2 つのことを行った。その例外的処理について、説明する。

6.4.1　PDB の切り出しを行う際における例外的な処理

第一に、ゲノムデータベースの一つである PDB に対して行った例外的処理について説明する。PDB は、立体構造について記されたデータベースであるが、そのエントリには、大量の数値データが含まれている。例として、PDB のエントリの一部を図 6.7として示す。

図 6.7で示す通り、PDB のエントリには大量の数値データが含まれている。この大量の数値データが Namazu でインデクシングを行う際に問題となる。なぜならば、Namazu がインデクシングを行う際、これら大量の数値データも Namazu は索引語として認識し、インデクシングを行おうとする。その際、例えば、11.470 という数値と、11.406 という数値は異なる索引語と Namazu は認識する。そのため、インデクスには大量の数値の索引語が含まれることになる。しかし、PDB エントリはあまりにも大量の数値データが含まれており、Namazu がインデクシングしきれない。また、11.470、11.406、10.241 という数値データには、それ自体にはたいした意味はない。その為、エントリという文書の意味内容を表すためには、あまり重要ではないと考えられる。これを回避するために、本研究では PDB のエントリをスプリッティングを行う際に、大量の数値データを含む ATOM や HETATM 等のフィールドを除いて、スプリッティングを行っている。

```
ATOM   1600  N   ALA  202       4.998  15.171  54.910  1.00 30.76           N
ATOM   1601  CA  ALA  202       3.900  16.042  55.301  1.00 29.94           C
ATOM   1602  C   ALA  202       2.695  15.817  54.399  1.00 31.26           C
ATOM   1603  O   ALA  202       1.569  15.718  54.876  1.00 30.84           O
ATOM   1604  CB  ALA  202       4.338  17.513  55.257  1.00 26.93           C
ATOM   1605  N   LYS  203       2.942  15.694  53.098  1.00 32.87           N
ATOM   1606  CA  LYS  203       1.869  15.505  52.114  1.00 36.42           C
ATOM   1607  C   LYS  203       1.477  14.069  51.778  1.00 38.62           C
ATOM   1608  O   LYS  203       0.470  13.857  51.113  1.00 41.91           O
ATOM   1609  CB  LYS  203       2.205  16.242  50.813  1.00 30.48           C
ATOM   1610  CG  LYS  203       2.290  17.749  50.978  1.00 24.70           C
ATOM   1611  CD  LYS  203       2.464  18.453  49.666  1.00 18.61           C
ATOM   1612  CE  LYS  203       2.206  19.911  49.875  1.00 21.92           C
ATOM   1613  NZ  LYS  203       2.139  20.677  48.620  1.00 34.69           N
ATOM   1614  N   GLY  204       2.270  13.093  52.207  1.00 40.55           N
ATOM   1615  CA  GLY  204       1.957  11.701  51.913  1.00 43.50           C
ATOM   1616  C   GLY  204       2.200  11.282  50.469  1.00 46.05           C
ATOM   1617  O   GLY  204       1.402  10.553  49.876  1.00 47.73           O
ATOM   1618  N   LEU  205       3.297  11.766  49.903  1.00 47.27           N
ATOM   1619  CA  LEU  205       3.685  11.470  48.528  1.00 48.90           C
ATOM   1620  C   LEU  205       4.592  10.245  48.394  1.00 49.71           C
ATOM   1621  O   LEU  205       5.038   9.719  49.437  1.00 49.38           O
ATOM   1622  CB  LEU  205       4.386  12.692  47.933  1.00 48.17           C
ATOM   1623  CG  LEU  205       3.524  13.798  47.324  1.00 48.48           C
ATOM   1624  CD1 LEU  205       2.091  13.733  47.834  1.00 44.53           C
ATOM   1625  CD2 LEU  205       4.179  15.147  47.600  1.00 42.74           C
TER    1626        LEU  205
HETATM 1627  P   PO4  300      11.470  25.915  63.056  1.00 41.12           P
HETATM 1628  O1  PO4  300      11.406  24.762  62.081  1.00 48.86           O
HETATM 1629  O2  PO4  300      10.241  26.712  63.337  1.00 53.42           O
HETATM 1630  O3  PO4  300      11.771  25.113  64.258  1.00 51.08           O
```

図 6.7: PDB エントリの様子（1NSJ エントリの一部）

6.4.2 OMIM の切り出しを行う際における例外的な処理

OMIM には ^ で始まるエントリ名のエントリが存在する。このエントリは obsolete entry と呼ばれるエントリで、今では使われなくなった古いエントリ名を表す。例として図 6.8として示す。

```
OMIM
MIM Entry: ^102550
Title:
  ^102550 MOVED TO 102630
Text:
  This entry was incorporated into entry 102630 on 10 April 1997.

Contributors:
  Mark H. Paalman - edited: 4/10/1997
Creation Date:
  Victor A. McKusick: 6/4/1986
Edit Dates:
  jenny: 04/15/1997
  jenny: 4/10/1997
  supermim: 3/16/1992
  supermim: 3/20/1990
  ddp: 10/26/1989
  marie: 3/25/1988
  reenie: 6/4/1986
```

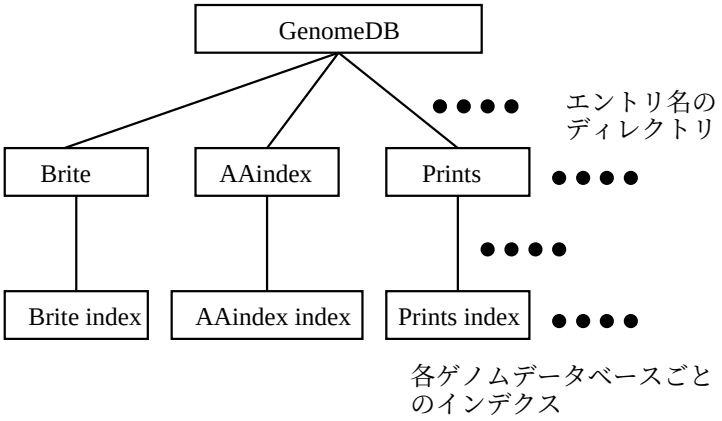
図 6.8: ^ で始まる OMIM の obsolete エントリ（ ^ 102550 エントリ）

図 6.8を見ても分かるように、 ^ で始まる obsolete エントリは、そのエントリが今ではどのエントリに移ったのか必ず記されている。

そのため、現在使われている新しいエントリさえ有れば良いと判断し、この ^ で始まる obsolete エントリは本研究では省くことにした。従って、エントリのスプリッティングを行う際、この obsolete エントリは対象外とした。

6.5　ゲノムデータベースのインデクシング

ゲノムデータベースにNamazuを適用する前に、6.2節、6.3節、6.4節でお話してきたような前処理的なことを行った後、ゲノムデータベースにNamazuを適用し、mknmzにより、インデクシングを行った。インデクシングは、各ゲノムデータベースごとに、エントリ単位とフィールド単位、双方について行っている。Namazuを適用し、インデクスを作成した時の、インデクスファイルのデータ構造を図6.9、図6.10として示す。



またここで、各ゲノムデータベースに関して、エントリ単位で mknmz によりインデクシングを行った場合のエントリ数、インデクスされた索引語数、生成したインデクスファイルのサイズ（MB)、生成に要した時間、インデクスを生成した日時をまとめ、表 6.1として示す。また、この時使用した計算機は、Sun Microsystems 社製の Enterprise 3000 であり、2 台でインデクスの作成を行った。この 2 台の計算機のホスト名は db1 と db2 である。db1 と db2 の詳しいマシンのスペックを表 6.2として示す。

データベース名	entry 数	索引語数	Index file size [KB]	生成時間	生成日時
BRITE	278	4641	301	2:16.30	Fri Aug 27 02:22:23 1999
AAindex	500	6333	500	3:52.50	Fri Aug 27 02:23:40 1999
PRINTS	865	162018	18510	26:47.64	Fri Aug 27 03:57:10 1999
EPD	1356	41799	1943	19:09.6	Fri Sep 10 18:39:49 1999
PROSITE	1374	110817	2923	26:33.2	Sun Sep 12 20:15:46 1999
ENZYME	3705	48554	3438	33:20.49	Fri Aug 27 02:57:54 1999
COMPOUND	5628	24803	1080	29:36.33	Fri Aug 27 02:50:16 1999
PMD	7078	92104	23114	1:51:09.03	Fri Aug 27 04:36:40 1999
TRANSFAC	9737	98796	10321	2:04:13.33	Fri Aug 27 05:39:42 1999
PDB	9254	319183	197304	5:35:16.34	Fri Aug 27 08:14:29 1999
OMIM	10573	251119	57185	4:07:39.69	Fri Aug 27 06:43:24 1999
PDBSTR	15639	205763	312719	8:33:29.75	Wed Sep 22 23:21:23 1999
SwissProt	80000	3598743	227311	25:33:00.6	Mon Sep 13 21:48:50 1999
PIR	122810	1309280	292739	32:17:06.52	Sat Aug 28 11:02:14 1999
Genes	107118	2856675	215953	24:06:55.6	Sat Sep 11 18:46:47 1999
PRF	118187	3899584	127988	25:02:21.2	Sun Sep 12 19:49:12 1999
LITDB	298877	1961512	161244	27:58:02.40	Sat Aug 28 07:19:09 1999

表 6.1: 各ゲノムデータベースごとにエントリ単位でインデクスを作成した結果

ホスト名	db1	db2
制作会社	Sun Microsystems	Sun Microsystems
システムモデル	Enterprise 3000	Enterprise 3000
メインメモリ	4.0GB	1024MB
CPU 数	4	4
CPU タイプ	sparc	sparc
OS 名	SunOS Version 5.6	SunOS Version 5.6

表 6.2: インデクス作成に使用した計算機のスペック

なお、Namazu によりインデクスの作成を行う際、mknmz のオプションとして、”-E” を使用した。これにより、mknmz による単語の切出しの際、単語の両端に記号がある場合、これを削除する。即ち、単語に区切り記号などが隣接している場合は、隣接している区切り記号などを削除して、インデクスを作成している。

実際のインデクス作成の際、mknmz は、以下のように使用した。

%mknmz -a -E -O インデクスを格納するディレクトリ　インデクス作成の対象となるエントリ群があるディレクトリ

第 7 章

ベクトル化

6章で作成したインデクスを使い、3.4節で説明した表 3.4のように、エントリ中に含まれる索引語と、その索引語の出現頻度のベクトルでエントリを表現する。

7.1 バイナリファイルからテキストファイルへの変換

Namazu によりインデクシングされ、作成されたインデクスファイルは 2.4.1節で説明したように、NMZ.*の形で保存される。数ある NMZ ファイルの中で、3.2節で説明したような転置インデクスが実際に記録されているのは、NMZ.i ファイルである。しかし、NMZ.i ファイルはバイナリファイルであり、この状態では NMZ.i ファイルの中の情報を読むことができない。そこで、バイナリファイルから、テキストファイルに変換する必要がある。

NMZ.i ファイルをバイナリファイルから、テキストファイルに変換するため、nmztxt.pl という Namazu のデータベースとテキストの双方向変換サブルーチン群を使って変換を行った。

変換結果を図 7.1として示す。

⋮
⋮
balzano
1022 1
3945 1
4143 1
5516 1
6032 1
9026 1

balzaretti
764 1
2231 1

balzo
6234 1

bam22
2518 3
6968 7
8833 1
8923 1

bamatter
4619 2
5084 3

bamba
4274 1

bamberger
1592 4
⋮
⋮

図 7.1: NMZ.i をテキストファイルに変換した結果（OMIM のエントリ単位でインデクシングした NMZ.i）

```
112     acid:1 act:1 alpha:1 alpha-carbon:1 amino:1 and:3 biology:1 carbon:1
chain:1 charton:1 chemical:1 correlation:1 fauchere:3 for:1 int:1 kier:1 laut
erwein:1 lit:1 original:1 parameters:1 peptide:1 pharmacology:1 pliska:1 pmid
:1 protein:1 quant:1 reference:1 rel:1 res:1 shift:1 side:1 struct:1 studies:
1 verloop:1
113     acid:1 amino:1 and:3 biol:1 biology:1 chain:1 charton:3 correlation:1
 effect:1 electrical:1 fauchere:2 for:1 int:1 kier:1 lit:1 localized:1 missin
g:1 original:1 parameters:1 peptide:1 pharmacology:1 pliska:1 pmid:1 pro:1 pr
otein:1 reference:1 res:1 side:1 studies:1 theor:1 verloop:1
114     acid:1 amino:1 and:2 biochem:1 biochemical:1 biology:1 bond:1 chain:1
 charton:1 commission:1 correlation:1 data:1 donors:1 eur:1 fauchere:2 for:1
hydrogen:1 int:1 iub:1 iupac:1 iupac-iub:1 joint:1 kier:1 lit:1 nomenclature:
1 number:1 original:1 parameters:1 peptide:1 pharmacology:1 pliska:1 pmid:1 p
rotein:1 reference:1 res:1 side:1 studies:1 these:1 two:1 verloop:1
```

図 7.2: ベクトル（AAindex をエントリごとにインデクシングしたものを変換）

7.2 転置インデクスからベクトルに変換

7.2節で読み出し可能となった転置インデクスを、ベクトル空間法で類似度計算ができるように、ベクトルに変換する。変換した結果を図 7.2として示す。

この転置インデクスからベクトルへの変換の際、Namazu がインデクシングした索引語の中で、意味があるとは思えないものは除外した。即ち、抽出した索引語の絞り込みを行った。以下にどのような語を除外したか示す。

- 文字が含まれていない語。即ち、数字、記号のみの語
- 語の中に数字が 4 文字以上含まれるもの。（例：ad8754protein）
- 文字/単語または、文字. 単語の形をとっているもの (例：s/protein, t.alzheimer)
- 語の頭が数字で始まるもの

第 8 章

辞書を用いたキーワードのフィルタリング

8.1 辞書の必要性

Namazu の場合、日本語文書に対してはわかち書きの結果分解された文字列を、英語文書に対しては空白で区切られた文字列を、基本的に全てキーワードとしてインデクシングする。対象となる文書が新聞記事や電子メールなどのように自然言語情報を豊富に含んでいる場合はこれでもサーチエンジンとして実用上十分な性能を発揮できるが、ゲノムデータベースのような科学技術データベースに対してそのまま Namazu を適用すると、各種の問題が生じる。中でも、本研究のようにインデクシングの結果得られる単語およびその頻度情報を用いてベクトル空間法を使おうとする場合、エントリの特徴をベクトルとして表現するのにふさわしくない単語までインデクシングされていることが問題となる。最も顕著な例は、エントリ中でフィールドの種別を表すフィールド識別子（フィールド名）である。これは、フィールドの種類によっては殆んど全てのエントリに出現するため、このようなものがベクトル中に多数含まれると、エントリの特徴が曖昧になってしまう可能性がある。多少異なる例としては、エントリ中に埋め込まれているクロスリファレンスがある。これはあるエントリが別のエントリと関連を持っていることを示しているが、データベースやフィールドによってはこのようなクロスリファレンスとしてのエントリ名が多数出現し、通常の文章に現れる重要な単語（特徴を表す専門用語など）よりも支配的にベクトルを形成してしまう可能性がある。さらに、このようなものを全て排除したとしても、専門用語と一般英単語に関する問題がある。専門用語に比べて一般英単語は、エントリの特徴を表している可能性が低く、かつ、頻度的には専門用語を上回ってしまう可能性がある。このような状態で生成したベクトルでは、やはりエントリの特徴が曖昧になってしまう危険がある。以上の理由で、本研究では、生物医学系の専門用語、一般英単語、および

フィールド名のそれぞれについて辞書を作成した。ベクトル計算を行なう前にこれを用いてキーワードのフィルタリングを行なうことにより、分類精度が高まることが期待される（詳細については9章で述べる）。

8.2 辞書作成

専門用語辞書については以下のデータソースを用いた。

- 医学文献データベース Medline で論文アブストラクトの分類に用いられている MeSH Term。
- LSD(Life Science Dictionary) プロジェクトによるライフサイエンス辞書のバージョン３。

これらのデータソースから単語を抽出し、必要に応じて加工を行なう事で辞書を作成した。一般にこのような辞書は、シングルキーワード以外にも熟語などを含んでいる。一方、Namazu が生成するインデックスには、英文の場合シングルキーワードしか出現しない（空白を頼りに単語の認識を行なうため）。よって、フィルタリングを行なうためには辞書側もシングルキーワードで構成する必要がある。このような方針で、データソースに出現した熟語をシングルキーワードに分解した後、sort や uniq などの簡単な Unix コマンドやシェルスクリプトを用いて半手動的に不要なゴミを除去し、専門用語辞書を生成した。

次に、一般英単語辞書については以下のデータソースを用いた。辞書生成については専門用語辞書と同様に、半手動的に行なった。

- Unix の標準辞書（/usr/dict/words）。
- フリーウェアの和英辞書 edict。

フィールド名については、6.3節で説明した entry-splitter.pl をゲノムネットの各種データベースに適用することにより自動的に作られるディレクトリ名（＝フィールド名）を使用して辞書作成を行なった。データベースの種類によって、エントリに出現するフィールド名の一覧が異なるので、データベースごとに別々のフィールド名辞書を作成した。

8.3 辞書のデータ量と重複

作成した辞書は、辞書同士で重なる部分も持っている。その状況を表 8.1に示す。

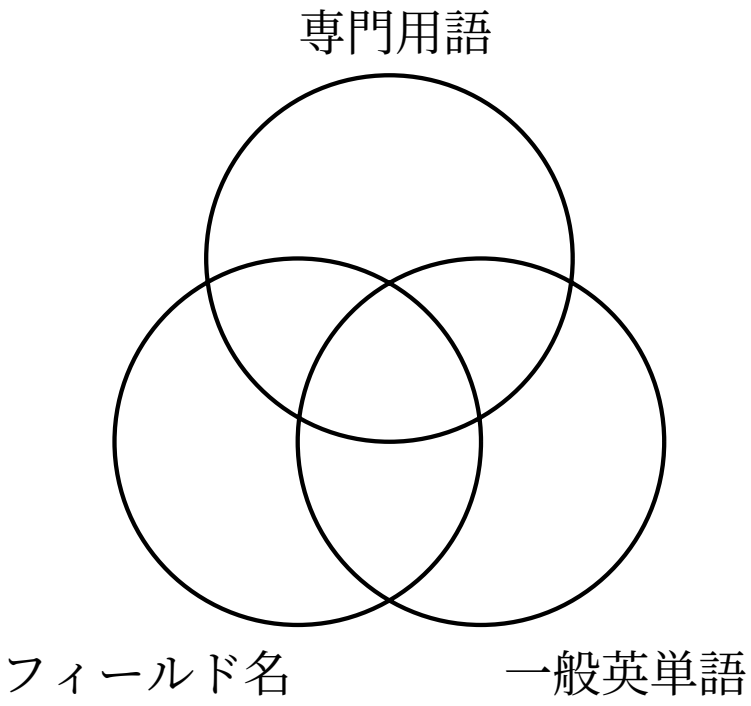


図 8.1: 一般英単語辞書と専門用語辞書とフィールド名辞書の位置関係

データ領域	単語数
専門用語	29062
一般英単語	35034
フィールド名	295
専門用語 ∧ 一般英単語	8021
専門用語 ∧ フィールド名	0
一般英単語 ∧ フィールド名	1
専門用語 ∧ 一般英単語 ∧ フィールド名	0

表 8.1: 各辞書の単語数と、その重なり状況

第 9 章

エントリ分類の実験

9.1 類似度計算

7章で作成したベクトルを使用して、類似度計算を行う。類似度の計算式は 3.4節で説明した通り、式 (3.2) を使用した。

$$\begin{aligned} sim(V_1, V_2) &= \frac{V_1 \cdot V_2}{|V_1||V_2|} \\ &= \frac{\sum_{i,j=1}^{t} a_i V_i b_j V_j}{\sqrt{\sum_{i=1}^{t} V_i^2}\sqrt{\sum_{j=1}^{t} V_j^2}} \end{aligned} \tag{3.2}$$

V_1, V_2 : 文書ベクトル

V_i : 文書ベクトル V_1を構成する索引語 i の出現頻度

V_j : 文書ベクトル V_2を構成する索引語 j の出現頻度

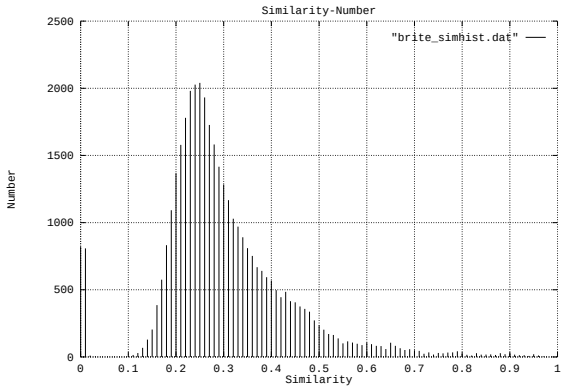
a_i : 文書ベクトル V_1に対する索引語 i の重要度

b_j : 文書ベクトル V_2に対する索引語 j の重要度

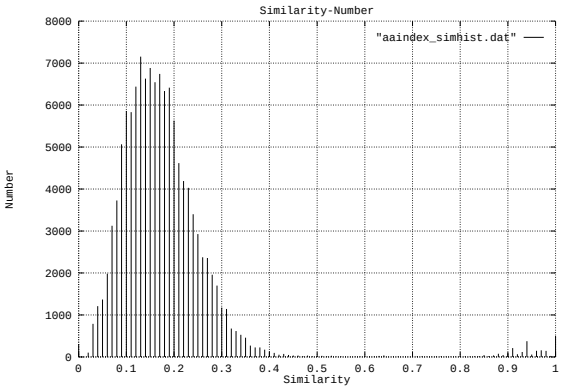
各データベースのエントリごとの類似度の計算を行った。各データベースごとに、全てのエントリの組み合わせにおいて類似度の計算を行い、類似度 0.01 刻みのヒストグラムとしてその結果を図 9.1、図 9.2として表示する。横軸が類似度であり、刻みは 0.01 である。縦軸は、その類似度であったエントリの組み合わせの件数である。

類似度計算の対象としたゲノムデータベースは BRITE、AAindex、PRINTS、EPD、PROSITE、ENZYME、COMPOUND、PMD、TRANSFAC、PDB、OMIM、PDBSTR

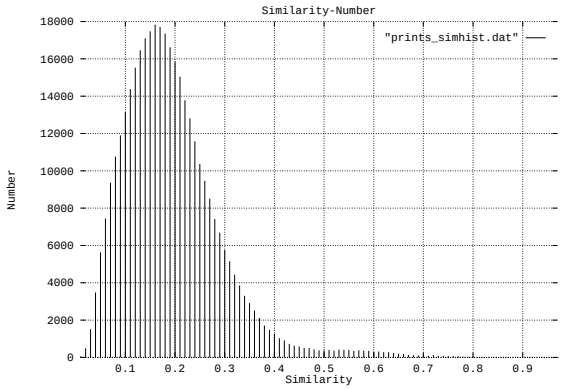
の 12 種類とした。他のゲノムネットに存在するゲノムデータデースについては、データ量が多くて計算することができなかった。これは、文書間の類似度計算が基本的に文書数の二乗の計算量（n(n-1)/2）になることが理由である。



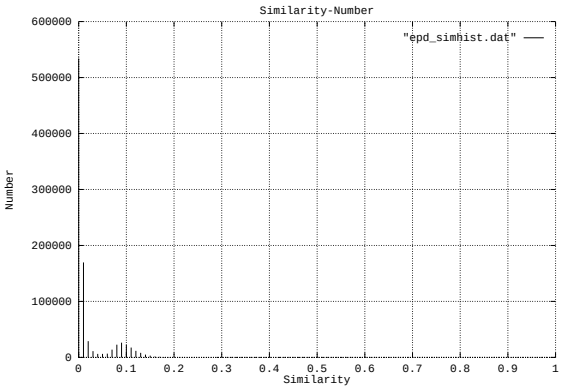
BRITE



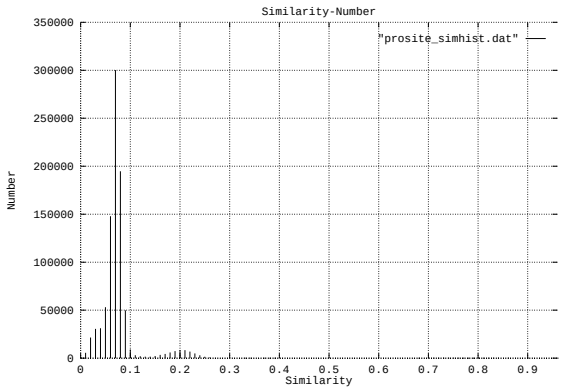
AAindex



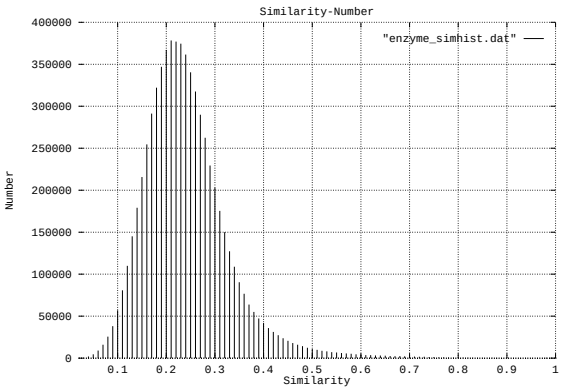
PRINTS



EPD

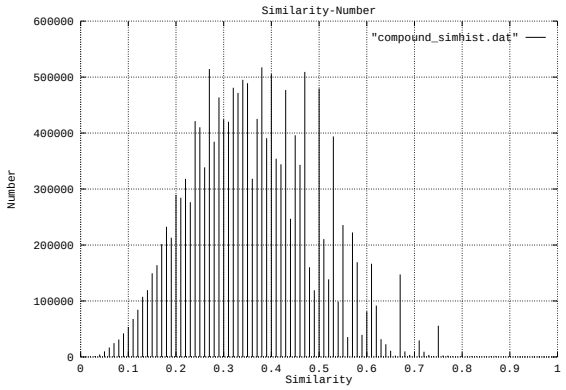


PROSITE

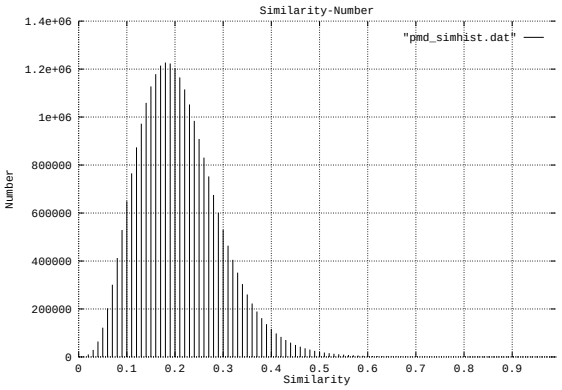


ENZYME

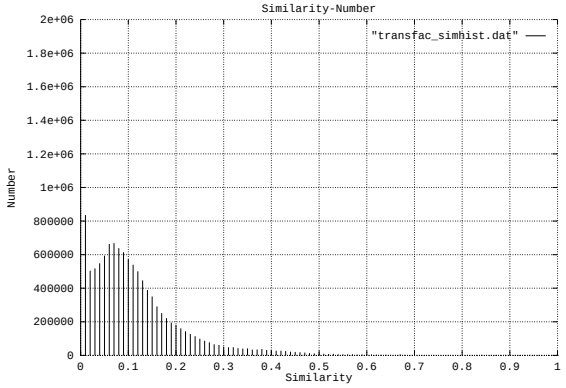
図 9.1: エントリ間の類似度分布



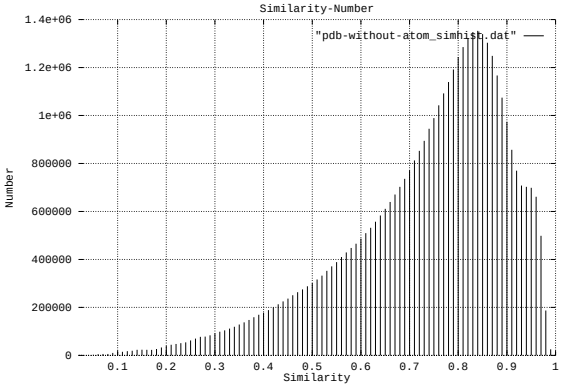
COMPOUND



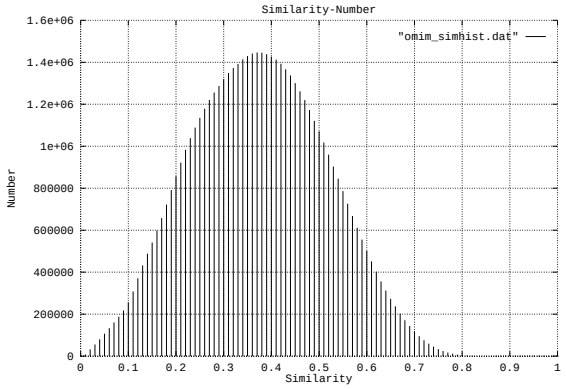
PMD



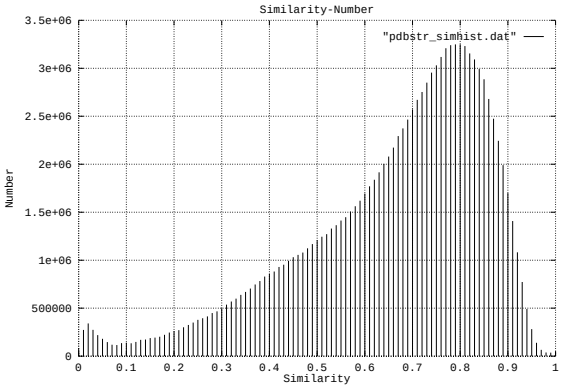
TRANSFAC



PDB



OMIM



PDBSTR

図 9.2: エントリ間の類似度分布

次に、各エントリに含まれている索引語を、専門用語、一般英単語、フィールド名とに分けて、類似度の計算を行う。理由は8章で述べたように、これらの語を全て同列に扱うのは得策ではないと考えるからである。

エントリ間の類似度を決定づける上で、専門用語は重要な働きをすることが予想される。ゲノムデータベースは科学データベースであることもあり、専門用語がエントリ中に数多く出現する。その為、専門用語の重みを変化させることで、各ゲノムデータベースごとのエントリ間の類似度分布が変化することが考えられる。

ここでは、8章で作成した辞書を使用して、専門用語を抽出し、専門用語の重みを 0 とした。即ち、専門用語を除いて類似度計算を行わせ、その結果、どう類似度分布が変化するかを見てみた。その場合の類似度計算の対象となる索引語集合を図 9.3として示す。

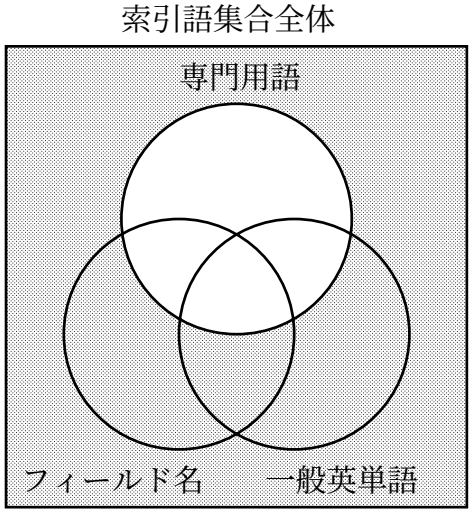


図 9.3: 専門用語を除いた際における類似度計算の対象範囲

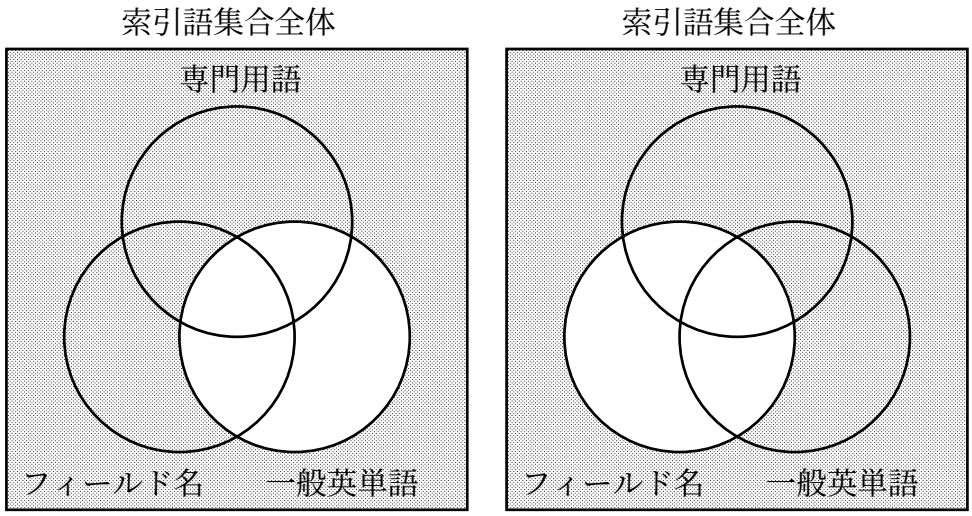
また、その結果を図 9.5及び図 9.6として示す。

同様に一般英単語とフィールド名についても、8章で作成した辞書を使用して、抽出し、重みを 0 として、類似度計算を行った。その場合の類似度計算の対象となる索引語集合を図 9.4として示す。

また、その結果を、一般英単語を除いた場合については図 9.7及び図 9.8として、フィールド名を除いた場合については図 9.9及び図 9.10として示す。

一般英単語は、専門用語と同様、エントリ中に数多く出現する。そのためエントリ間の類似度を全体的に大幅に高くする要因となっているであろうことが予想された。

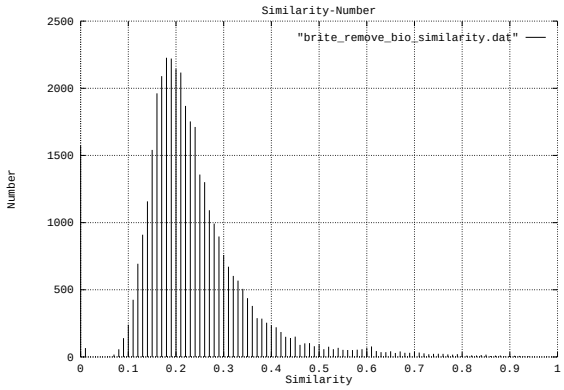
また、各ゲノムデータベースのエントリの記述方式にもよるが、フィールド名も相当の



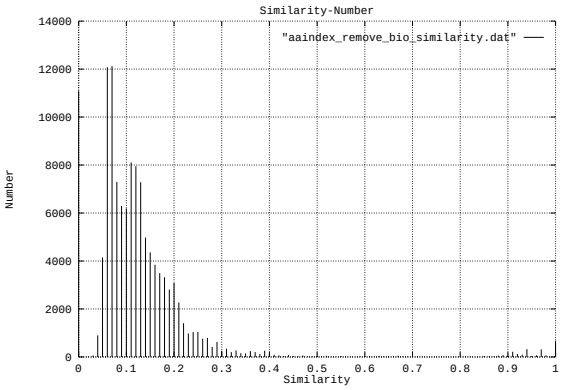
一般英単語を除いたとき　　フィールド名を除いたとき

図 9.4: 類似度計算の対象範囲

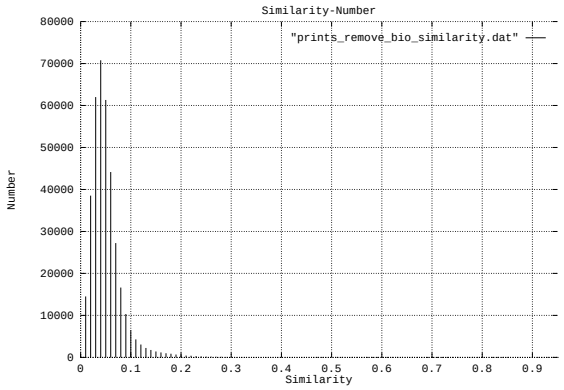
数出現している。従って、ゲノムデータベースの種類によっては大幅にエントリ間の類似度を高くしている要因となっていると予想された。



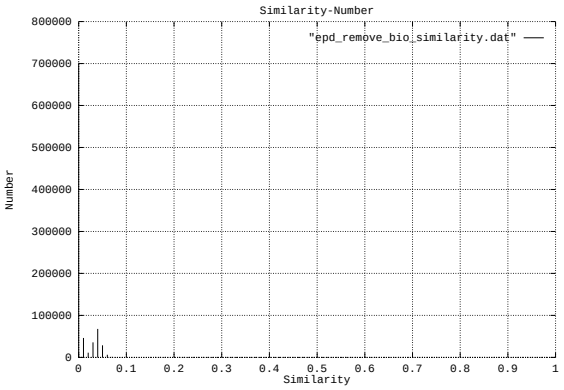
BRITE



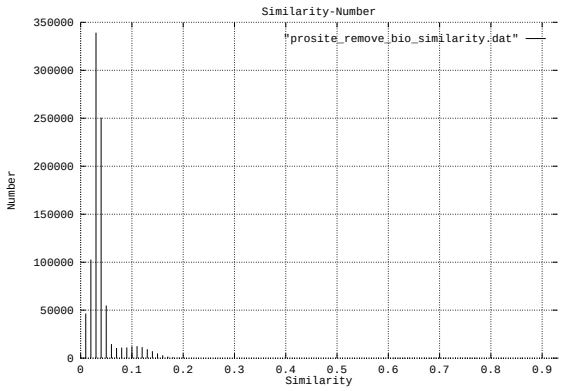
AAindex



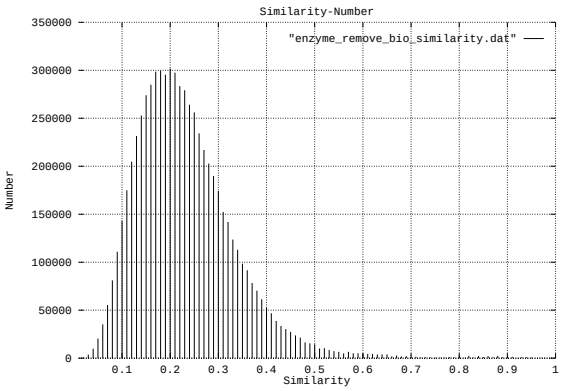
PRINTS



EPD

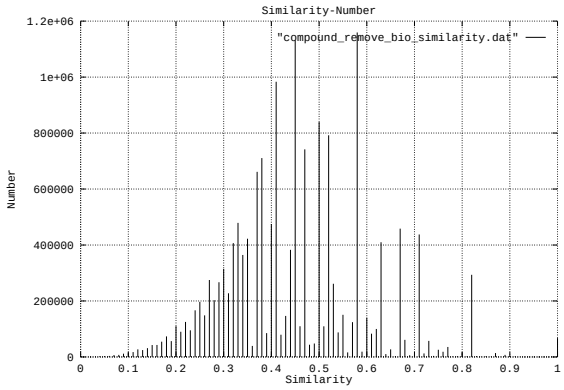


PROSITE

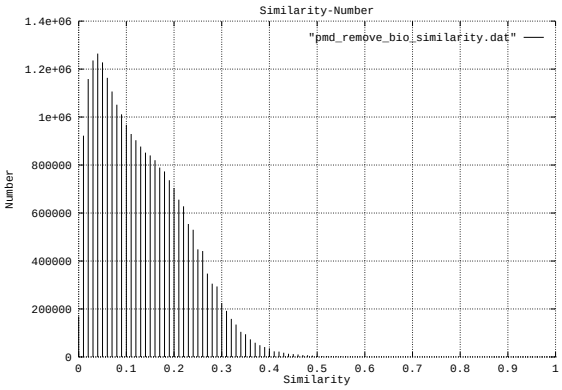


ENZYME

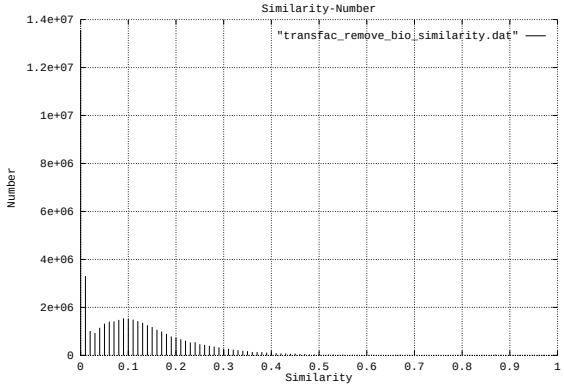
図 9.5: 専門用語を除いたときの類似度分布



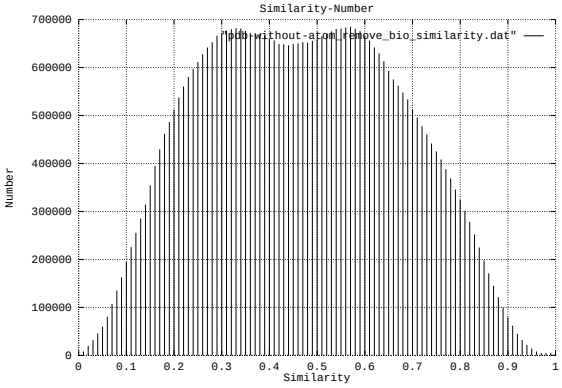
COMPOUND



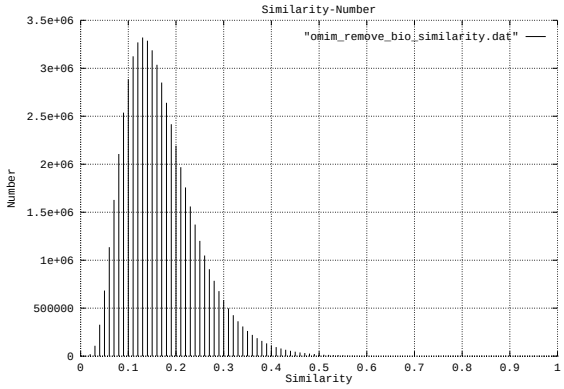
PMD



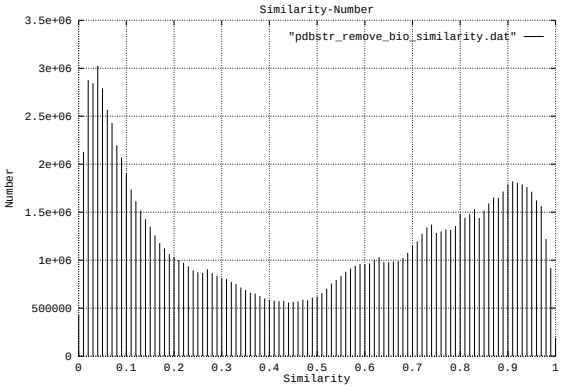
TRANSFAC



PDB

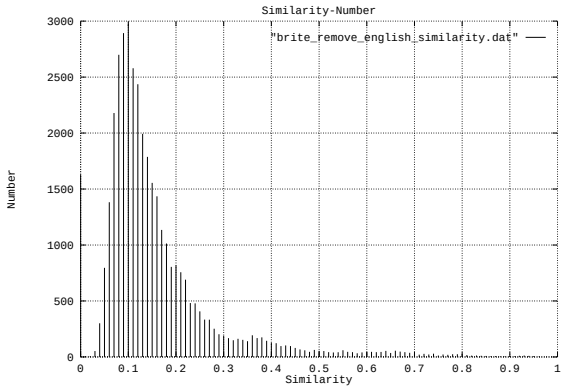


OMIM

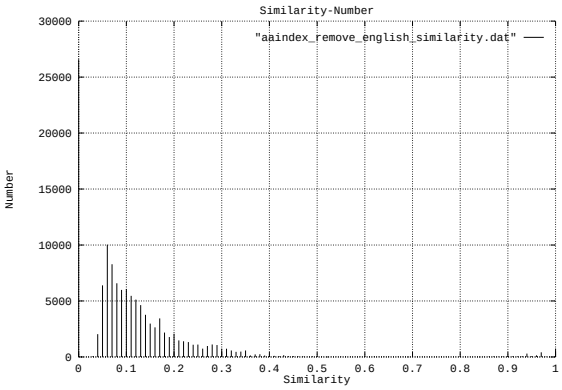


PDBSTR

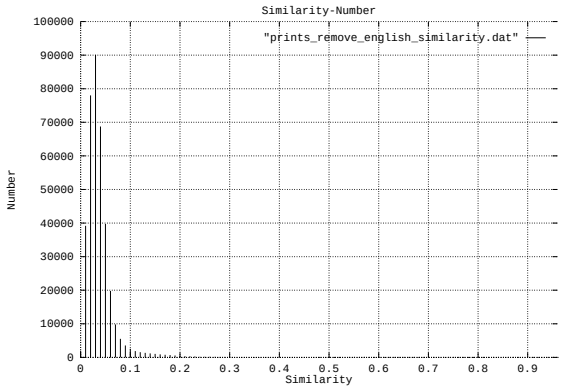
図 9.6: 専門用語を除いたときの類似度分布



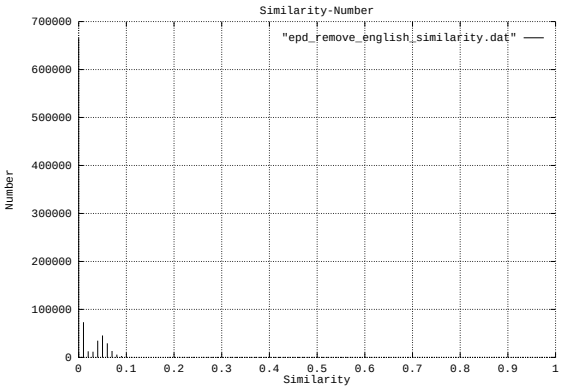
BRITE



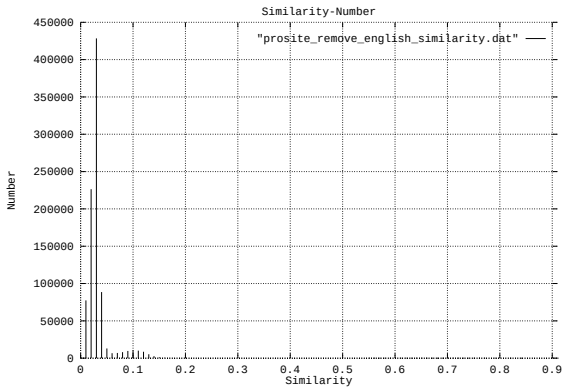
AAindex



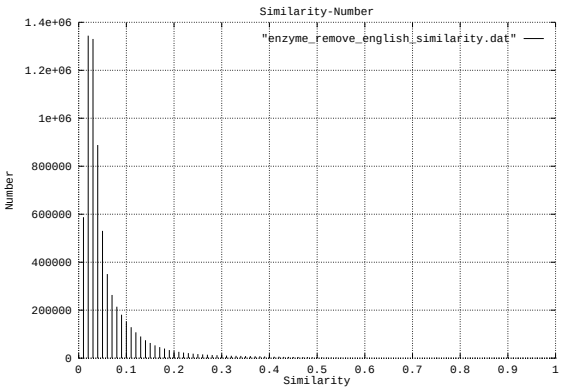
PRINTS



EPD

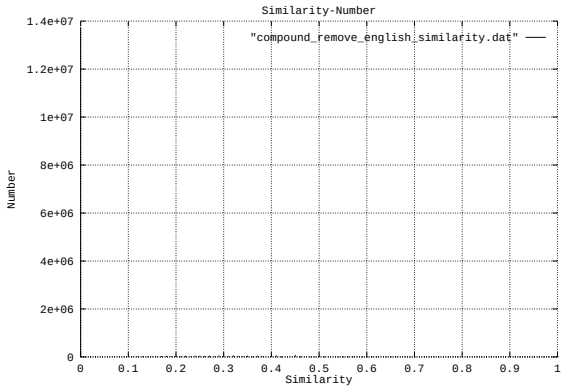


PROSITE

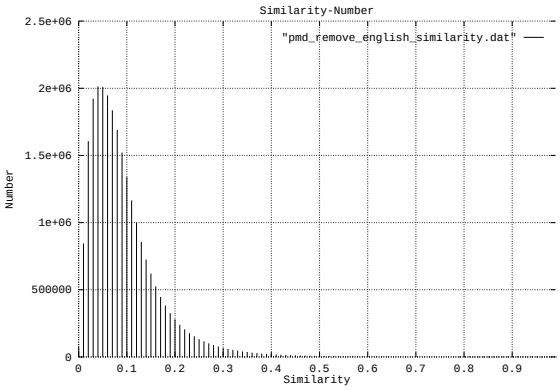


ENZYME

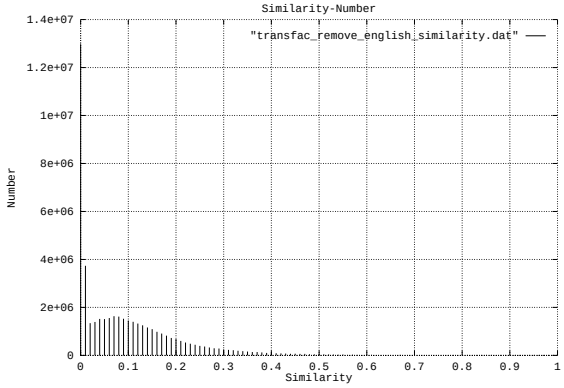
図 9.7: 一般英単語を除いたときの類似度分布



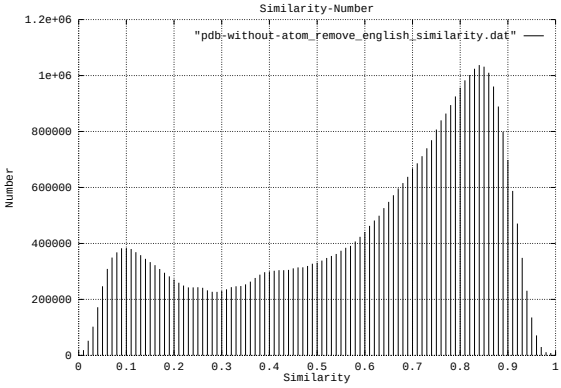
COMPOUND



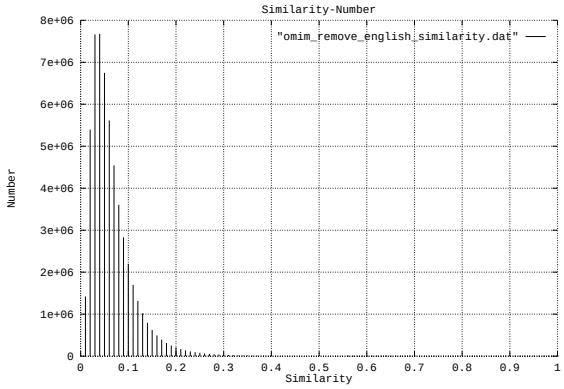
PMD



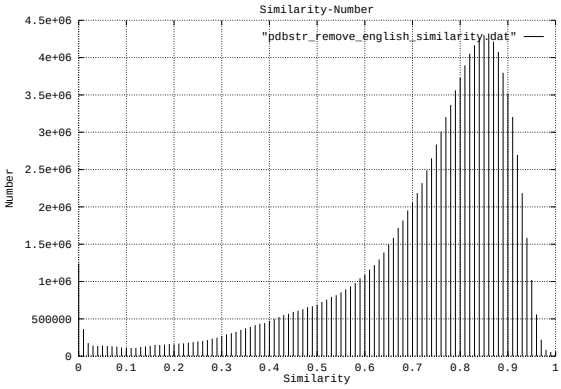
TRANSFAC



PDB

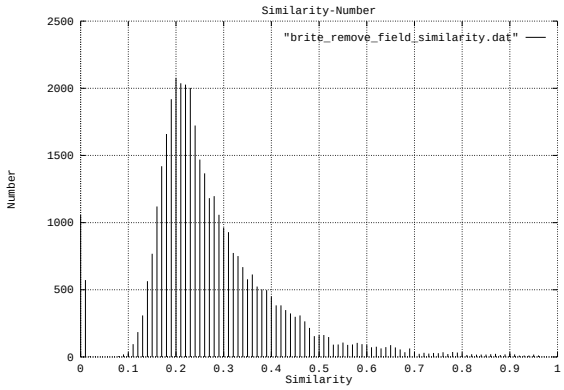


OMIM

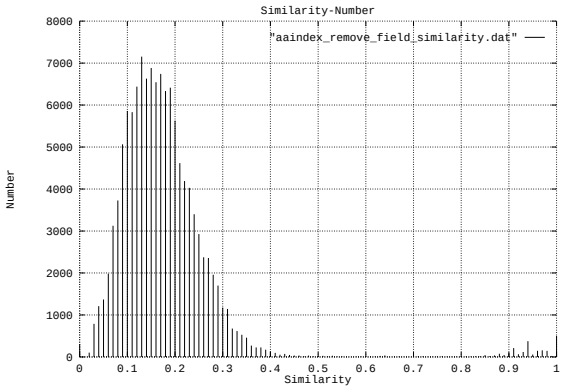


PDBSTR

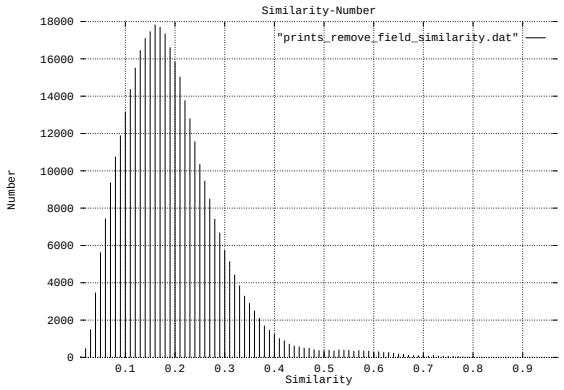
図 9.8: 一般英単語を除いたときの類似度分布



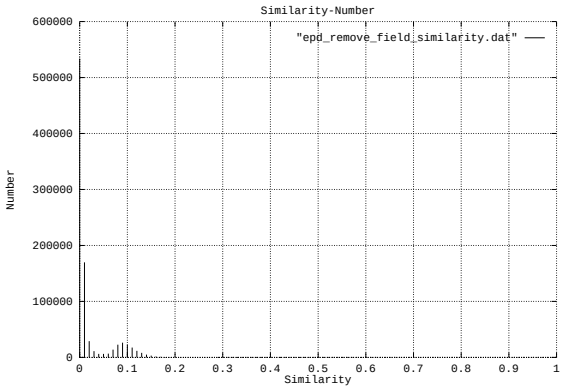
BRITE



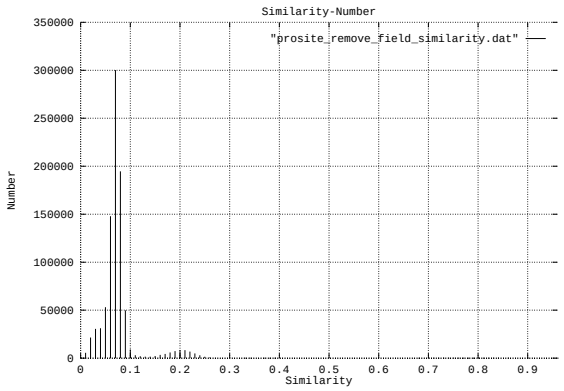
AAindex



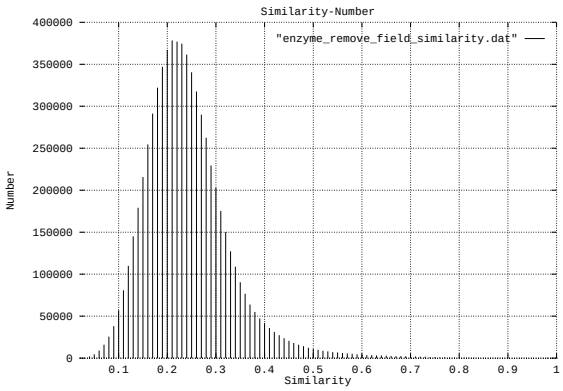
PRINTS



EPD

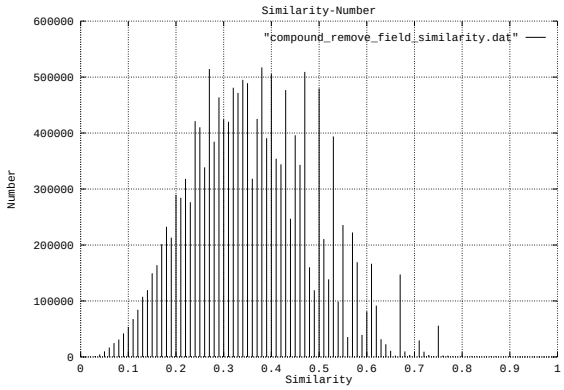


PROSITE

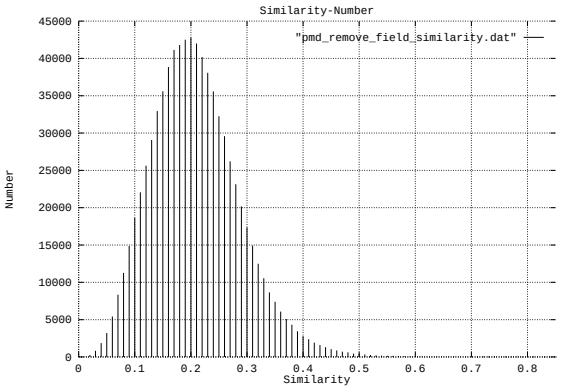


ENZYME

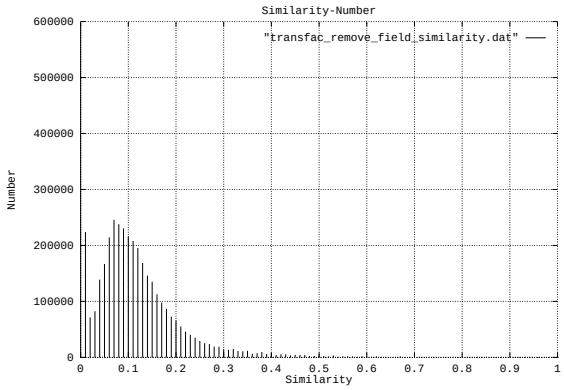
図 9.9: フィールド名を除いたときの類似度分布



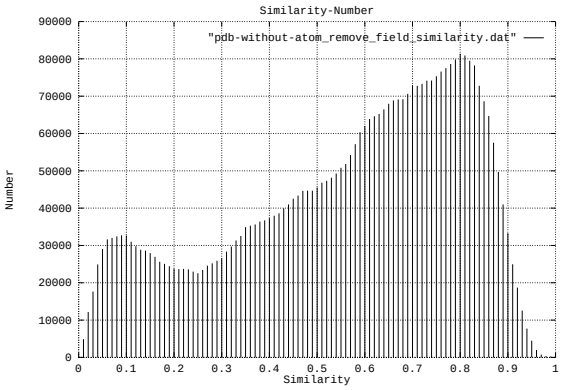
COMPOUND



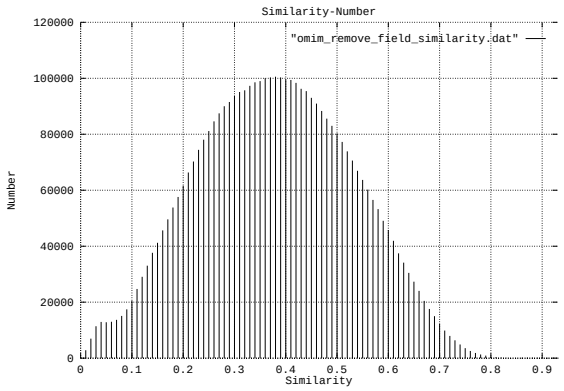
PMD



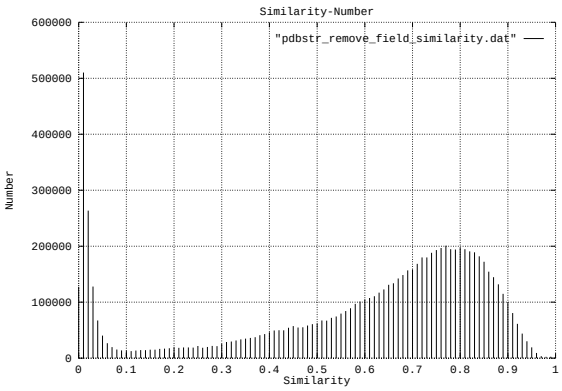
TRANSFAC



PDB



OMIM



PDBSTR

図 9.10: フィールド名を除いたときの類似度分布

処理を加えずにエントリ間の類似度を計算した時の類似度分布（図 9.1及び図 9.2）と、専門用語を除いた類似度分布（図 9.5及び図 9.6）を見比べると、専門用語を除いた類似度分布の方は、全体的に左寄りの分布を示しており、類似度が低くなっているのが分かる。専門用語が一致することにより、高くなっていた類似度の部分だけ低くなったのである。

BRITE と ENZYME は他のデータベースに比べ、専門用語を除いた際の影響が少ない。これは、BRITE は他のデータベースに比べ、一つ一つのエントリに記述されている情報が少ないということ、自然言語で書かれている情報が、REFERENCE 情報であり、生物的な専門用語とは違う種類の情報であることが原因と思われる。ENZYME は、エントリ中に記述されているデータ量に比して、自然言語で書かれた情報が少ないことが原因と考えられる。（付録参照）

一般英単語を除いた類似度分布（図 9.7及び図 9.8）を見ると、一般英単語を除くことにより、ほぼどのデータベースも専門用語を除いた時以上に、全体的に類似度分布が、類似度が低い方に移動している。一般英単語の方が専門用語よりエントリ中に数多く記述されているためと思われる。ただ、PDB 及び PDBSTR は、専門用語を除いたときほどの影響を受けていないのが特徴的である。

フィールド名を除いた類似度分布（図 9.9及び図 9.10）を見ると、ほぼどのデータベースも、専門用語や一般英単語を除いた時程には影響を受けていないようである。しかしながら、PDB 及び PDBSTR は他のデータベースに比べて、著しく影響を受けている。PDB は、付録に記述されているように、フィールド名がエントリ中に数多く出現する。そのため、フィールド名を除くことによる影響が強く出ていると思われる。TRANSFAC 及び PMD についても、PDB や PDBSTR 程ではないが、フィールド名を除くことによる影響が見て取れる。これも同様の理由によるものであると考える。（付録参照）

同様に、専門用語、一般英単語、及びフィールド名を全て除いた場合について類似度の分布を調べてみた。その場合の類似度計算の対象となる索引語集合を図 9.11として示す。また、その結果を図 9.13及び図 9.14として示す。

また、専門用語はエントリ間の類似度を決定づける上で重要な働きをすることが予想されることから、専門用語のみを類似度計算の対象とした場合、さらに、専門用語からフィールド名との重複を除いた場合についての類似度の分布も調べてみた。その場合の類似度計算の対象となる索引語集合を図 9.12として示す。また、その結果を専門用語のみの場合は図 9.15及び図 9.16として、専門用語からフィールド名との重複を除いた場合については、図 9.17 及び図 9.18として示す。

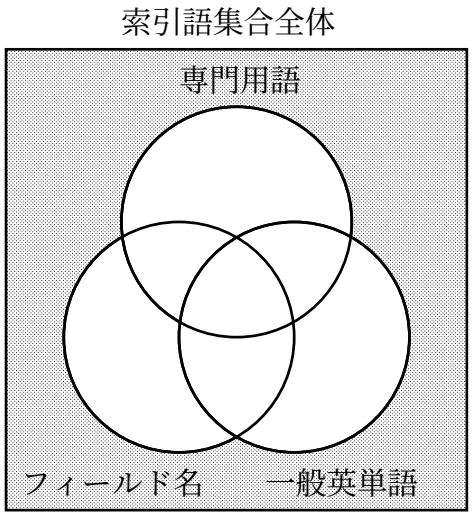
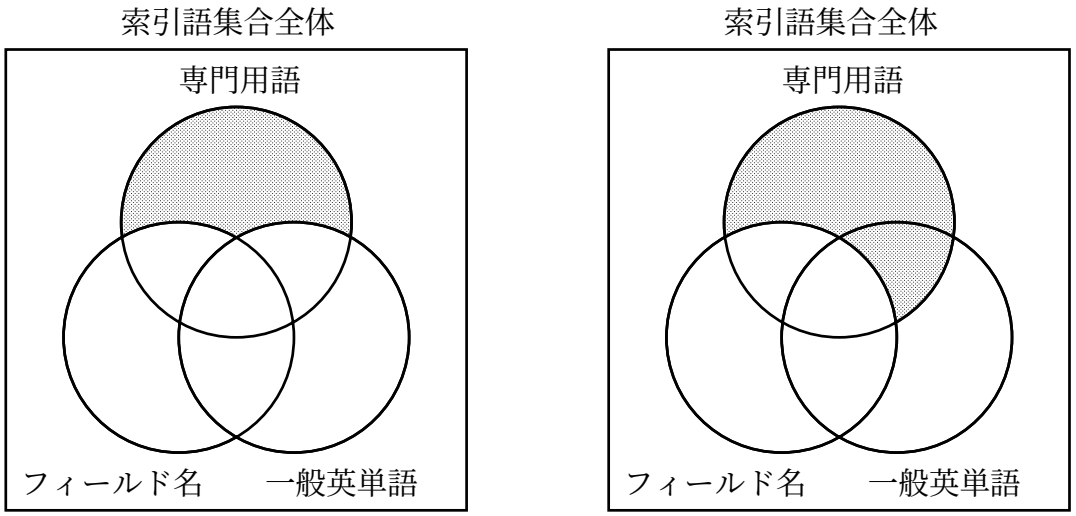
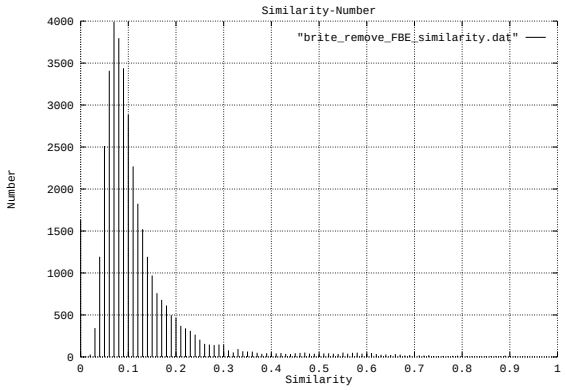


図 9.11: 専門用語、一般英単語、フィールド名を除いた際における類似度計算の対象範囲

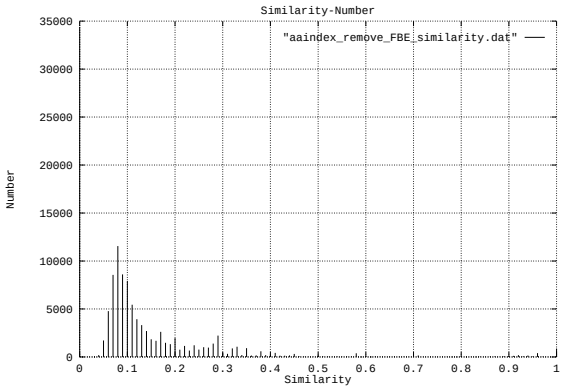


専門用語のみ　　専門用語からフィールド名との重複を除いたもの

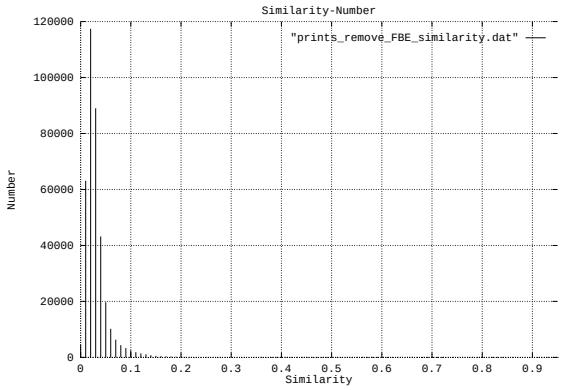
図 9.12: 類似度計算の対象範囲



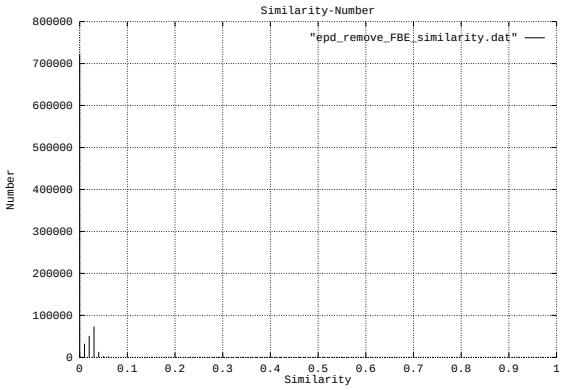
BRITE



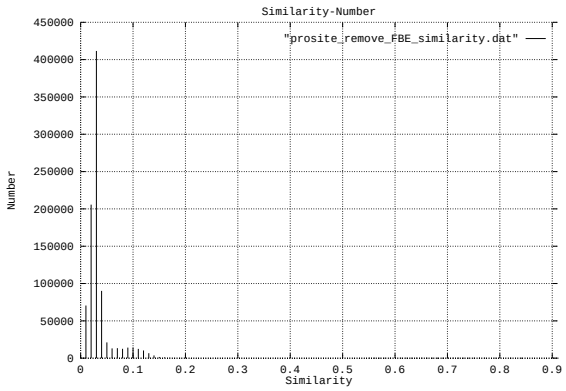
AAindex



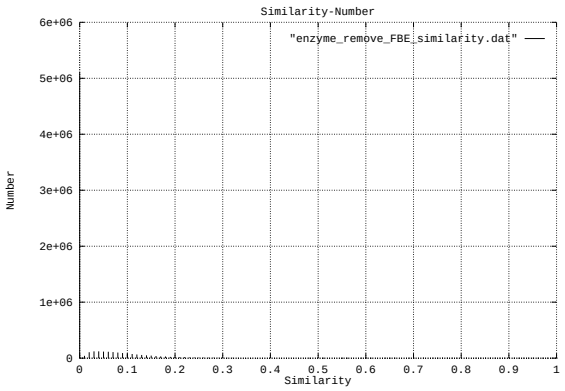
PRINTS



EPD

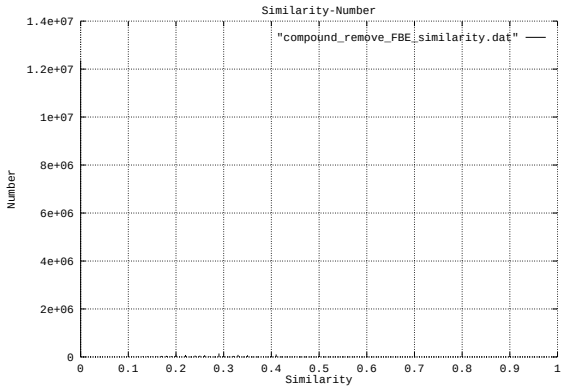


PROSITE

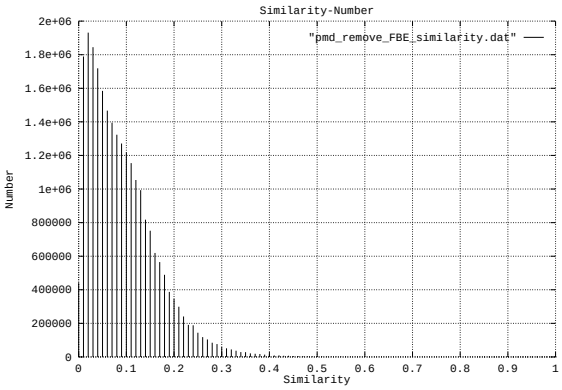


ENZYME

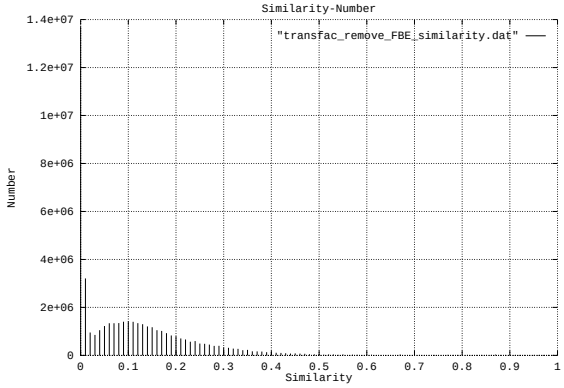
図 9.13: 専門用語・一般英単語・フィールド名共に除いたときの類似度分布



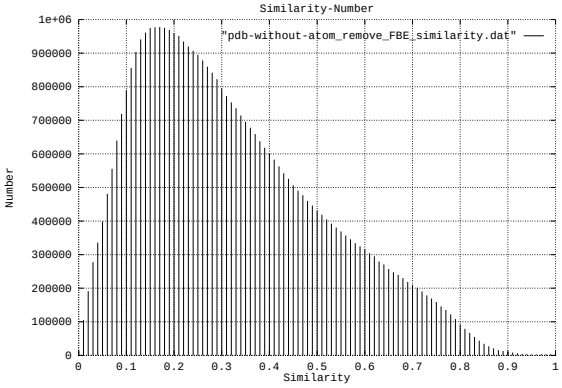
COMPOUND



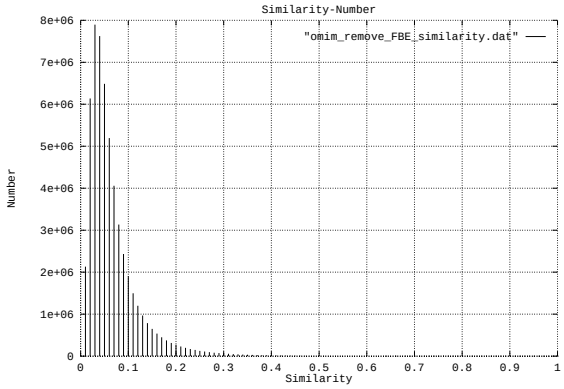
PMD



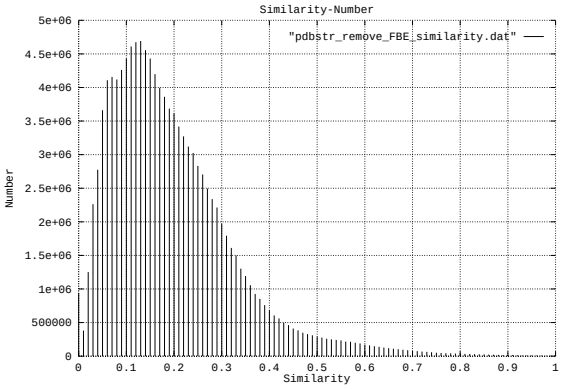
TRANSFAC



PDB

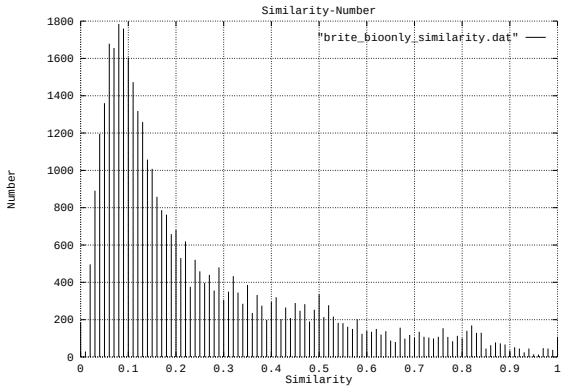


OMIM

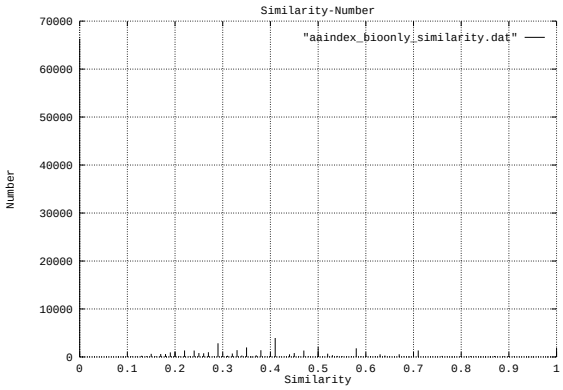


PDBSTR

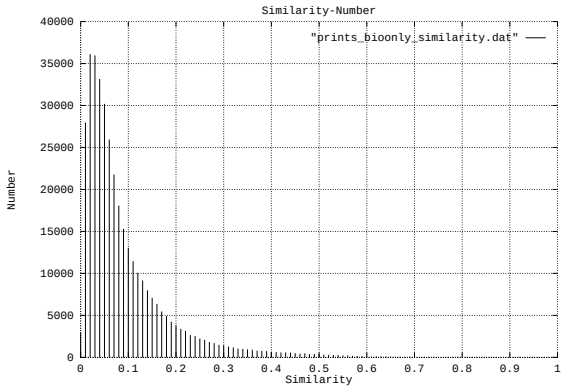
図 9.14: 専門用語・一般英単語・フィールド名共に除いたときの類似度分布



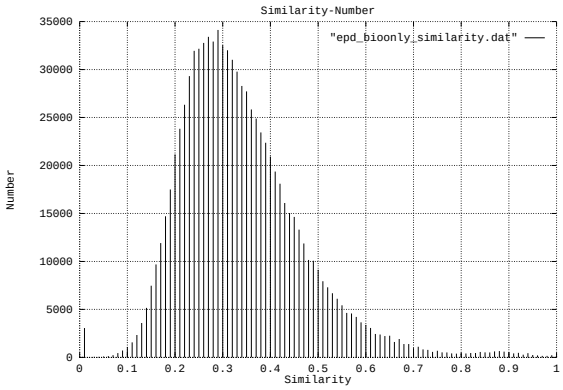
BRITE



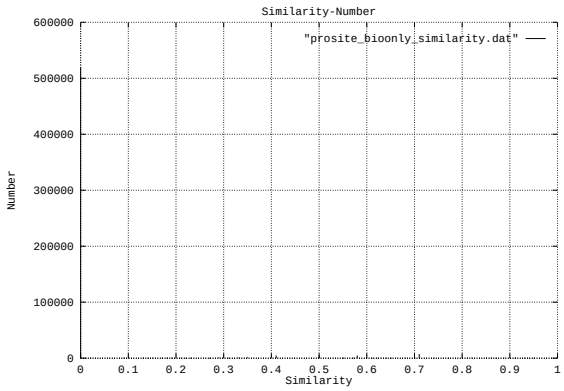
AAindex



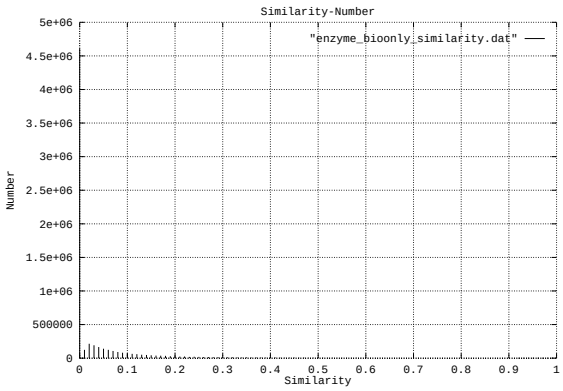
PRINTS



EPD

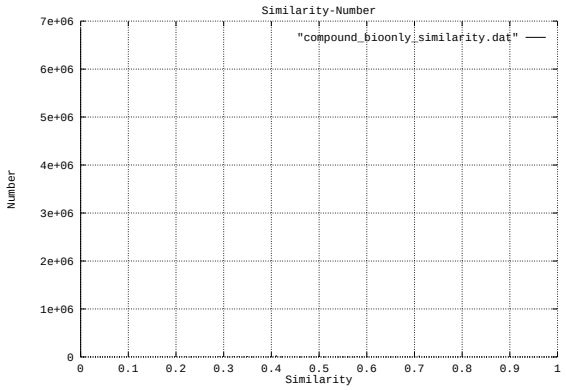


PROSITE

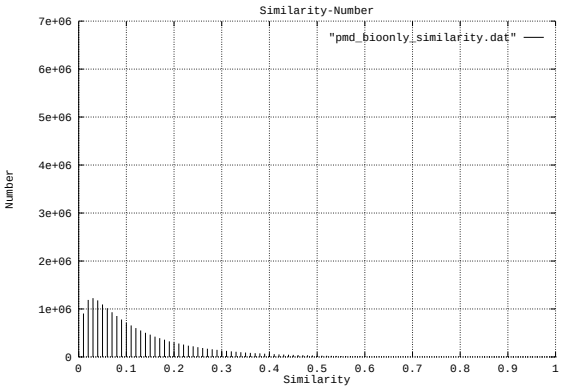


ENZYME

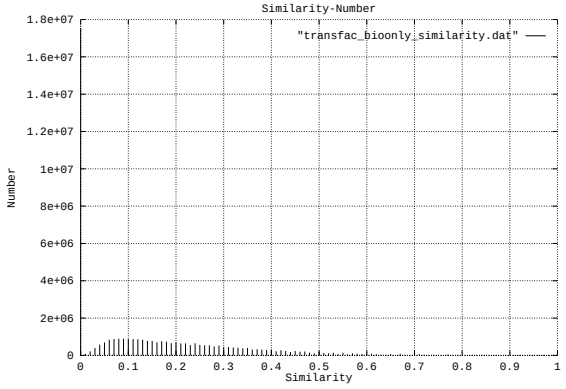
図 9.15: 専門用語のみの時の類似度分布



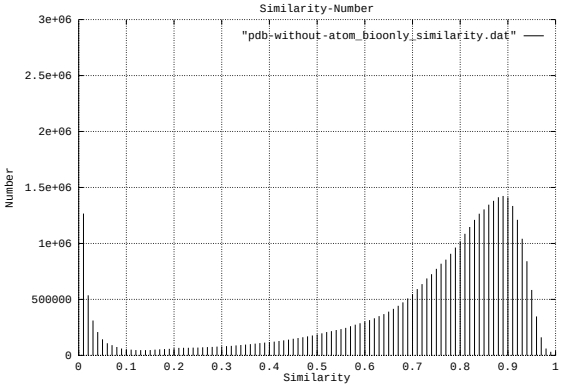
COMPOUND



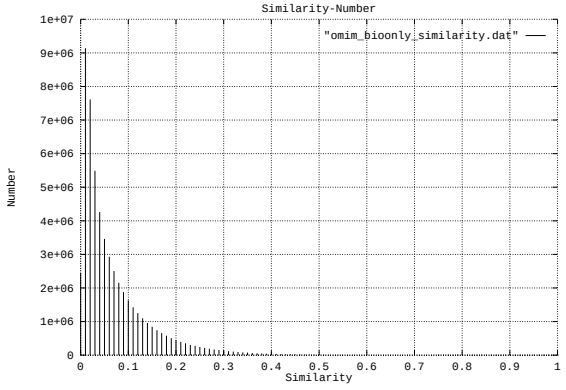
PMD



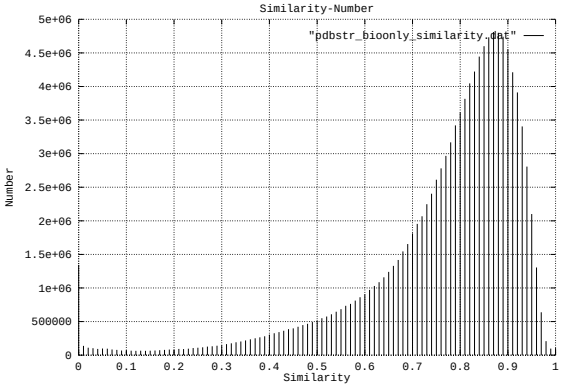
TRANSFAC



PDB

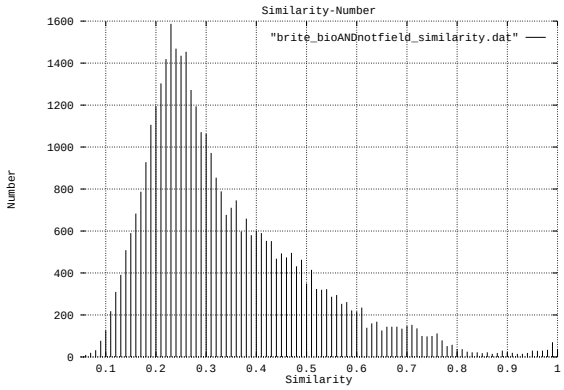


OMIM

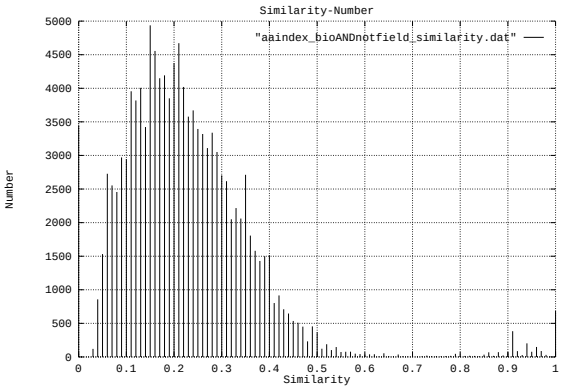


PDBSTR

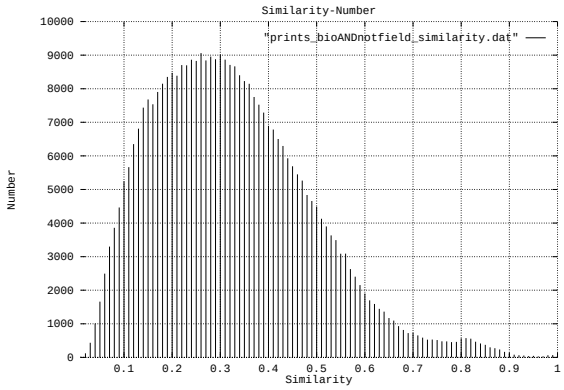
図 9.16: 専門用語のみの時の類似度分布



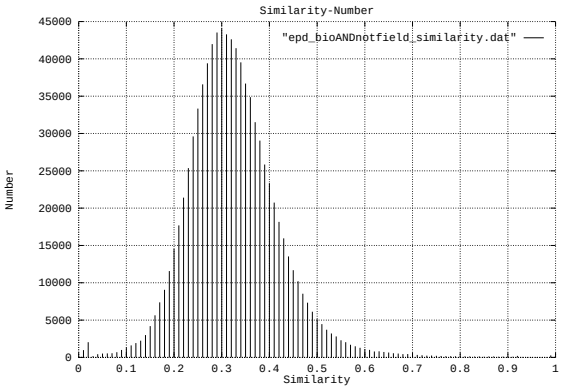
BRITE



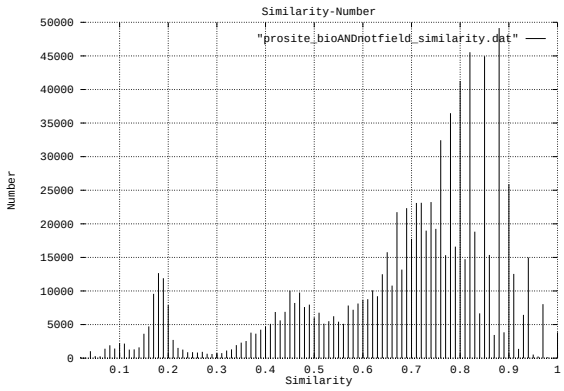
AAindex



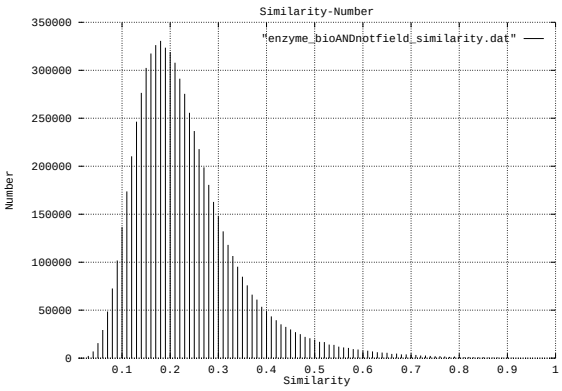
PRINTS



EPD

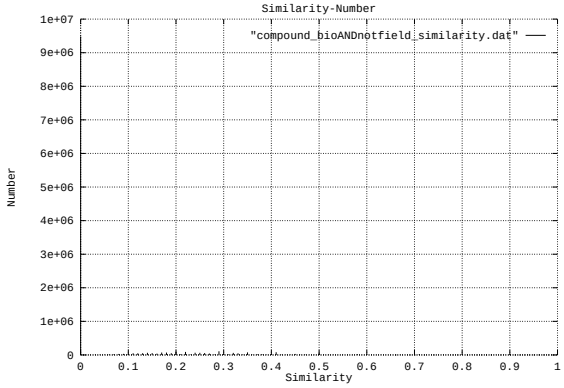


PROSITE

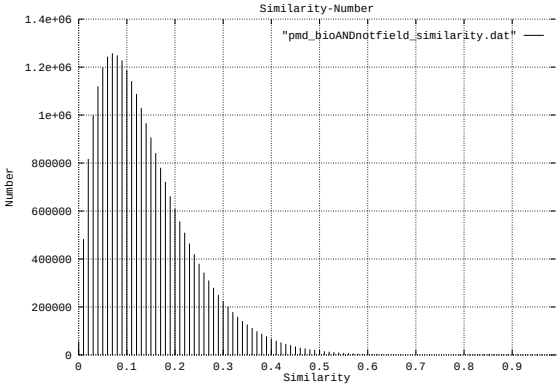


ENZYME

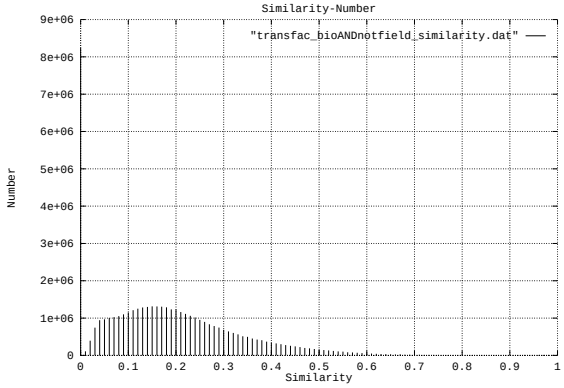
図 9.17: 専門用語からフィールド名との重複を除いた時の類似度分布



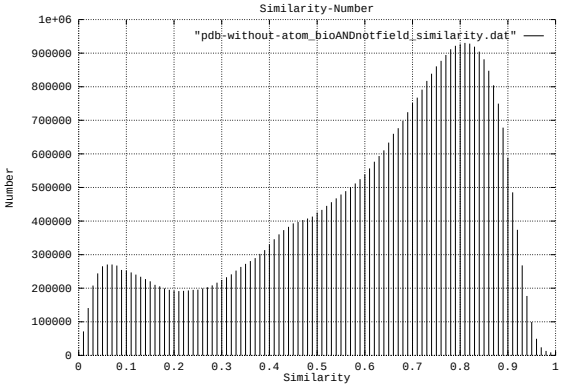
COMPOUND



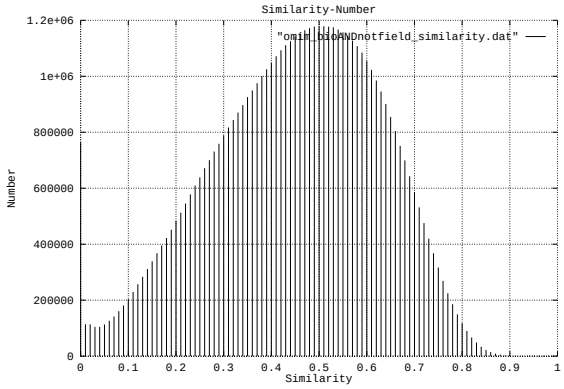
PMD



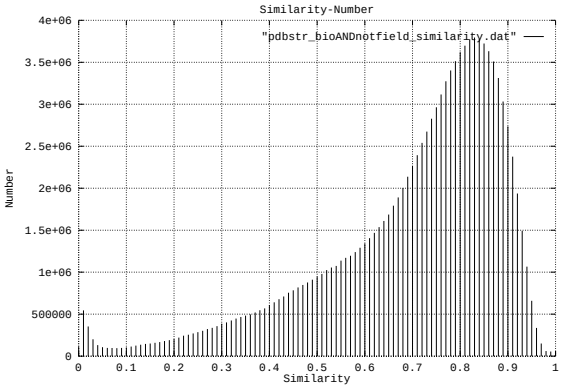
TRANSFAC



PDB



OMIM



PDBSTR

図 9.18: 専門用語からフィールド名との重複を除いた時の類似度分布

専門用語、一般英単語、及びフィールド名を全て除いた場合の類似度分布（図 9.13及び図 9.14）を見ると、全てのデータベースが類似度が低い方（左の方）に偏って分布していることが分かる。EPD、ENZYME、COMPOUND は類似度が 0 のところに分布が集中している。専門用語、一般英単語、及びフィールドを全て除いた場合、残るのはそれ自身ではあまり意味をもたないゴミとなる索引語であると考える。PMD、PDB、PDBSTR では依然高い類似度での分布を示しているが、PDB については、SEQRES、SHEET、HELIX、TURN、HET、MODRES、CISPEP、HETSYM、HYDBND、LINK、SEQADV、SIGATM、SIGUIJ、SITE、SLTBRG、SSBOND、TER 等で多く見られる ALA、CYS、ASP、TYR 等というような 3 文字の記号表記（アミノ酸の 3 文字表記）があり、これが影響を及ぼしていると考えられる。（付録参照）

専門用語のみで計算された場合の類似度分布（図 9.15及び図 9.16）を見ると、AAindex、ENZYME、PROSITE、COMPOUND は類似度分布が類似度 0 のところに集中しており、これらのデータベースでは専門用語のみの索引語が、他のエントリの索引語となかなか一致しないことが分かる。逆に、EPD は専門用語のみで計算された場合の方が、類似度の分布が、より鮮明に表示される。

専門用語であり、フィールド名ではない索引語を対象に類似度計算された場合の類似度分布については、専門用語のみで計算された場合と大きくは違わないであろうと予測していたが、多くのデータベースで激しい変化が見られた（図 9.17および図 9.18）。

9.2 ENZYME に関する分類分け

ENZYME データベースを対象として、分類分けを試みる。ENZYME データベースは、エントリ名が数字 1. 数字 2. 数字 3. 数字 4 という形になっており、数字 1 が 1 番大きな大分類グループを示し、数字 4 が一番小さな小分類グループという形になっている。即ち、ENZYME データベースはエントリ名によって分類分けがされており、どのように分類分けがされているのか答えが分かっている。このため、実験を行い、その分類分け手法が有効なのか否かの評価が行え、実験対象データベースとしては最適であると考えた。しかし、各分類に含まれるエントリ数には偏りがあることや、ENZYME データベース全体のエントリ数が 3700 もあり、クラスタリングを行なって分類の系統樹を描いた時に読めなくなることを考えて、今回の実験では一番小さな小分類グループ番号（上記の数字 4）として 1 を持つエントリだけを対象とした。これにより、エントリ数は 207 個になり、偏りの問題もある程度軽減される。

実験には統計解析パッケージ S を用いた。分類方法としては階層型クラスタリングを用い、要素間の距離定義としてはユークリッド距離を、クラスタ間の距離定義としては群平均法を用いた。

まず、ENZYME データベースにおけるエントリ全体に対して類似度計算を行った際の分類分けの結果を図 9.19として示す。この段階でもある程度の分類性能があるようだが、系統樹の形はまだ判然とせず、右肩下がりというバランスの悪い状態であることが分かる。

次に、自然言語を多く含むフィールド（CLASS、COFACTOR、COMMENT、EFFECTOR、INHIBITOR、NAME、PRODUCT、REACTION、SUBSTRATE、SYSNAME）だけを持つファイルを使用した場合と、逆に自然言語をあまり含まないフィールド（DBLINKS、DISEASE、ENTRY、GENES、MOTIF、PATHWAY、STRUCTURES）だけを持つファイルを使用した場合について、同様に類似度計算を行なった。特定のフィールドだけを持つファイルの生成には entry-splitter.pl を用いた。それぞれの分類分けの結果を図 9.20および図 9.21に示す。自然言語を多く含む場合とそうでない場合を比較してみると、後者の場合、すなわち単語の殆んどがクロスリファレンスとしてのエントリ名である場合は、エントリ全体の場合と比べても明らかに系統樹の形が崩れている。また、極端な右肩下がりになっていることから、さきほどバランスを崩していた大きな原因が後者にあることが分かる。よって、以後の実験では前者の場合だけを考える。

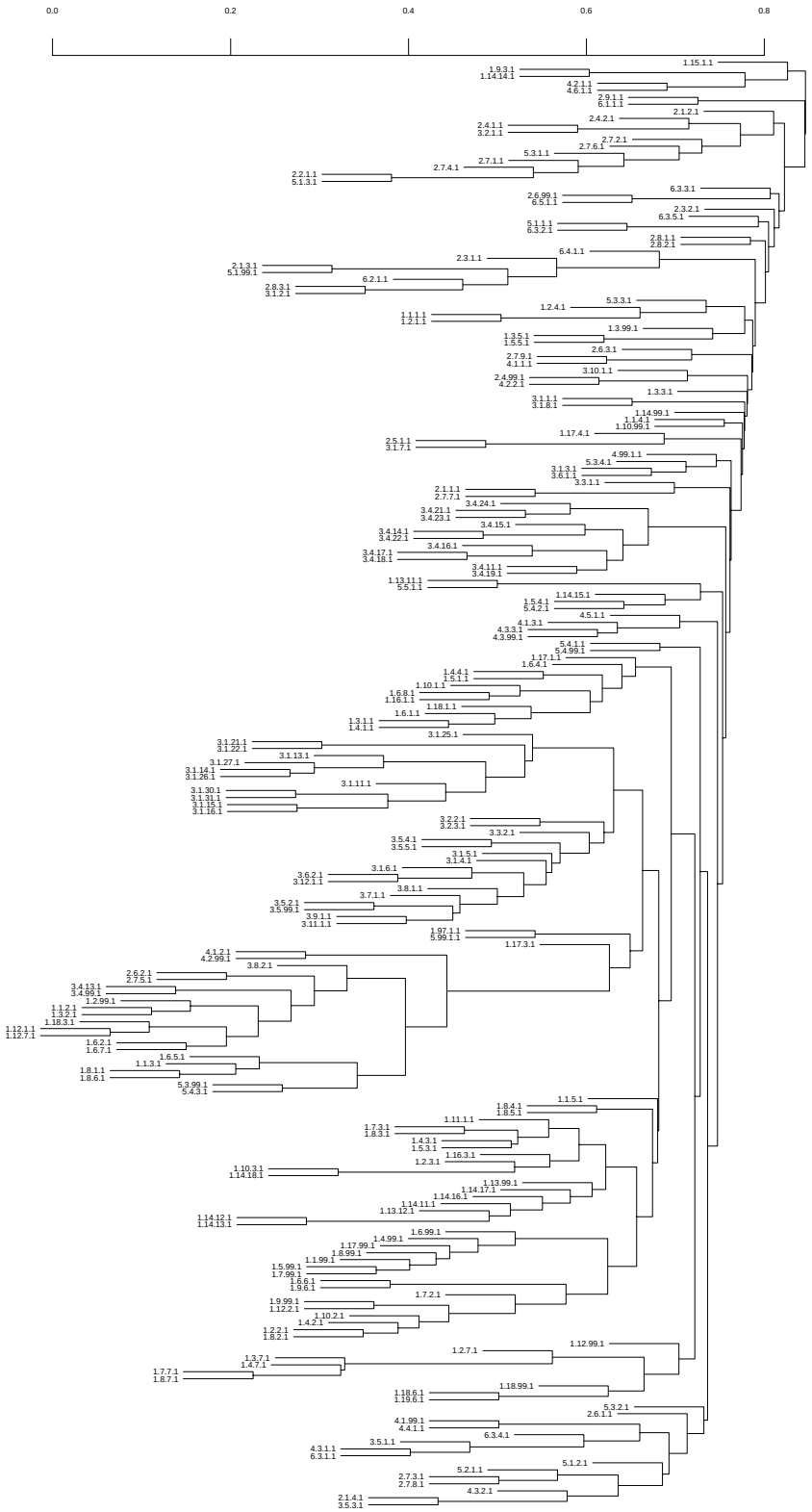


図 9.19: ENZYME:エントリ全体を類似度計算の対象としたときの分類分け

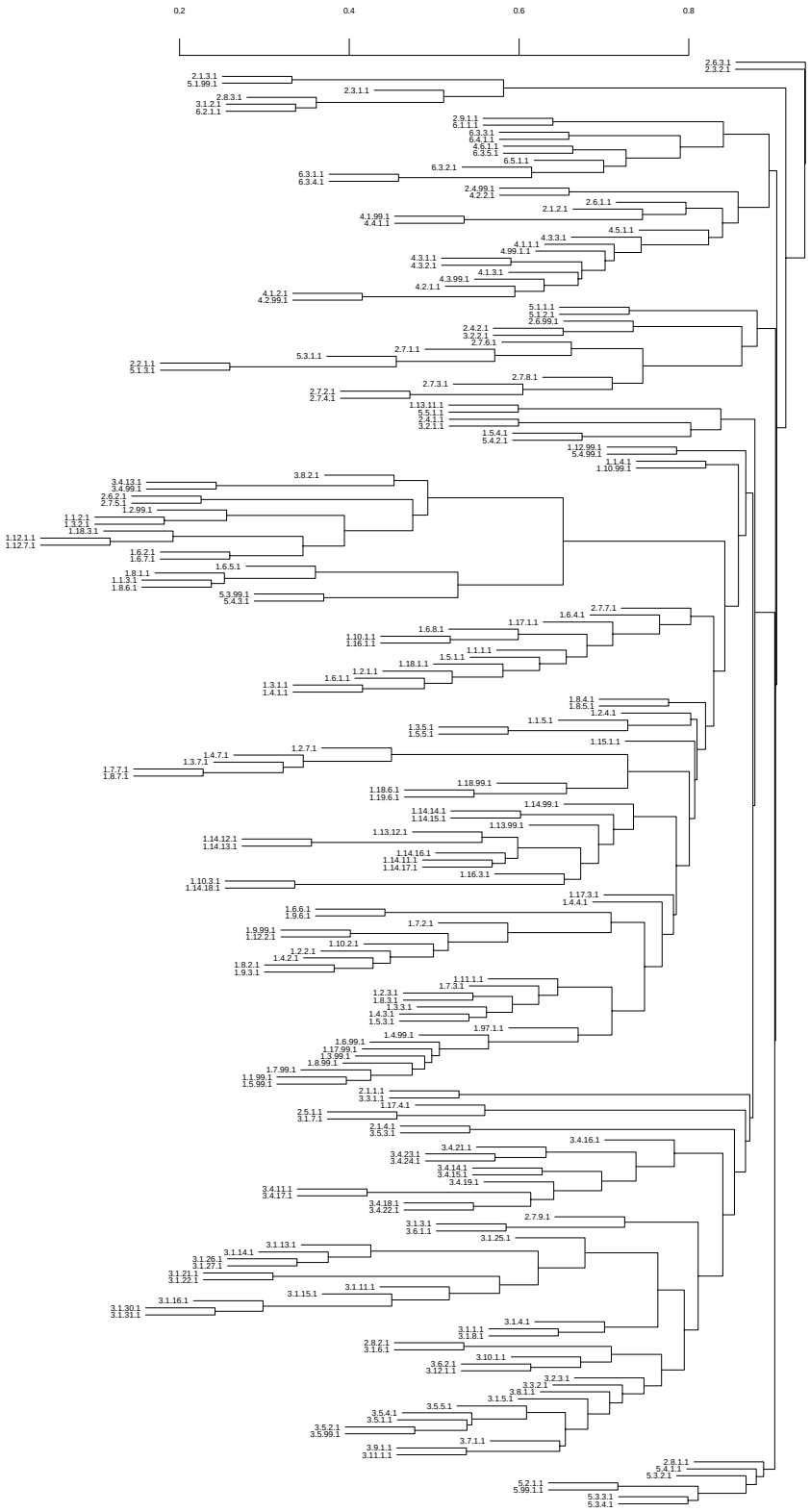


図 9.20: ENZYME:自然言語を多く含むフィールドを類似度計算の対象としたときの分類分け

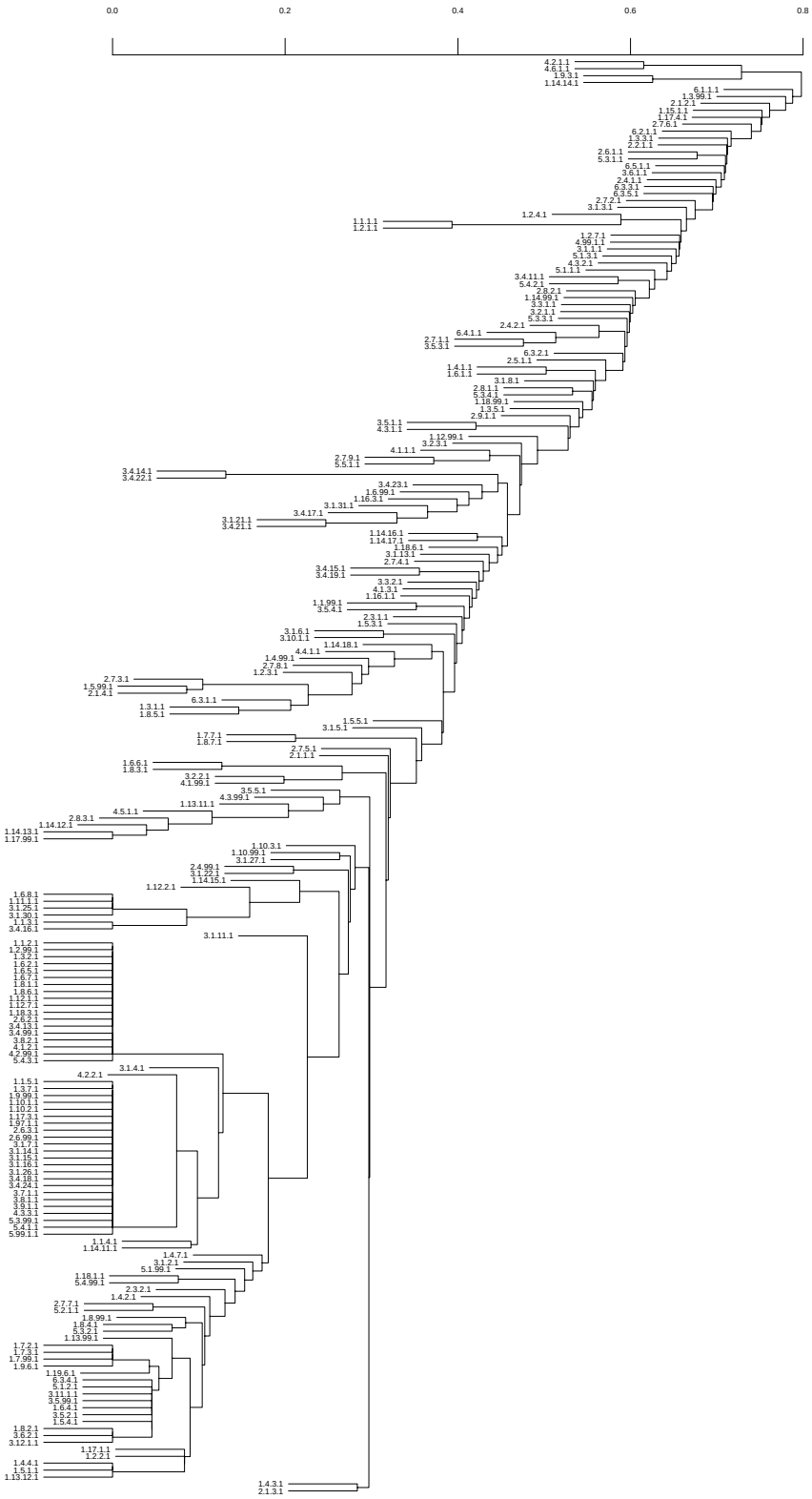


図 9.21: ENZYME:自然言語をあまり含まないフィールドを類似度計算の対象としたときの分類分け

次に、自然言語を多く含むフィールドに対し、専門用語辞書を使ってフィルタリングした単語だけを使ってベクトルを計算し、分類を行なった（図 9.22)。図 9.20に比べると部分木の形がより鮮明になり、全体にバランスが取れてきている。しかし、細かく見ていくとクラスタが融合したり境界が曖昧になっている印象がある。

次に、専門用語辞書と他の辞書（一般英単語辞書およびフィールド名辞書）の共通部分が与えている影響を調べるために、これらの共通部分を除いた場合について分類を行なった。フィールド名辞書との共通部分を抜いた結果を図 9.23に、一般英単語辞書との共通部分を抜いた結果を図 9.24に、それぞれ示す。図 9.22と比較すると、前者に関しても後者に関してもかなり影響がでている。特に後者の場合は、系統樹の形が非常に鮮明になっている。このことから、フィールド名の除去と一般英単語の除去がともに効果的であることが分かる。

また、フィールド名と一般英単語の両方を除去した場合についてもやってみたが、結果的には一般英単語を除去した場合と殆んど変わらなかった（図 9.25)。実は自然言語を多く含むフィールドのフィールド名（CLASS、COFACTOR、COMMENT、EFFECTOR、INHIBITOR、NAME、PRODUCT、REACTION、SUBSTRATE、SYSNAME）については、殆んどが一般英単語辞書にも載っており、この実験の場合はフィールド名の除去が一般英単語の除去に包含されてしまったことが原因であった。しかし、一般論としてこれが成立するわけではない（aaindex の 1 文字フィールド名などは一般英単語辞書には載らない)。

最後に、本手法が他のデータベースにも有効かどうかを見る目的で、PROSITE のサブセット（名前が D から G までのアルファベットで始まるエントリ：251 個）を対象に、自然言語を多く含むフィールド（DE、RU）だけを切り出し、さらに専門用語辞書からフィールド名辞書および一般英単語辞書との重なりを除去したものを使って単語をフィルタリングし、分類を行なった（図 9.26)。この実験ではフィールドに含まれる単語数が少なく、あまり系統樹らしくはならなかったが、それでも部分木が出来ている個所についてはそれなりに高い精度で分類されていることが分かる（PROSITE はエントリ名からある程度どういう内容か予想がつく)。

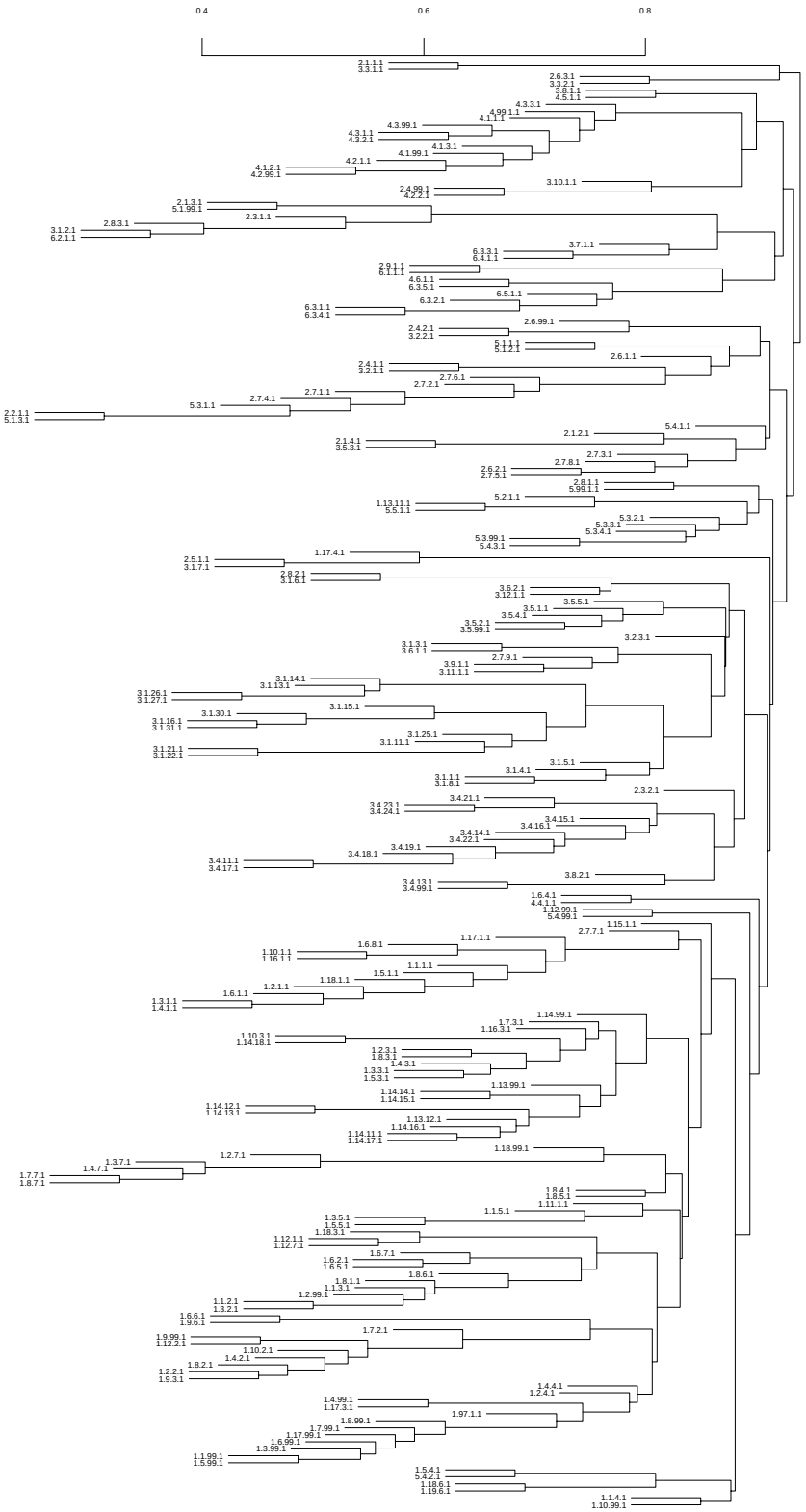


図 9.22: ENZYME:自然言語を多く含むフィールドで専門用語を類似度計算の対象としたときの分類分け

図 9.23: ENZYME:自然言語を多く含むフィールドで、専門用語でありフィールド名ではない索引語を類似度計算の対象としたときの分類分け

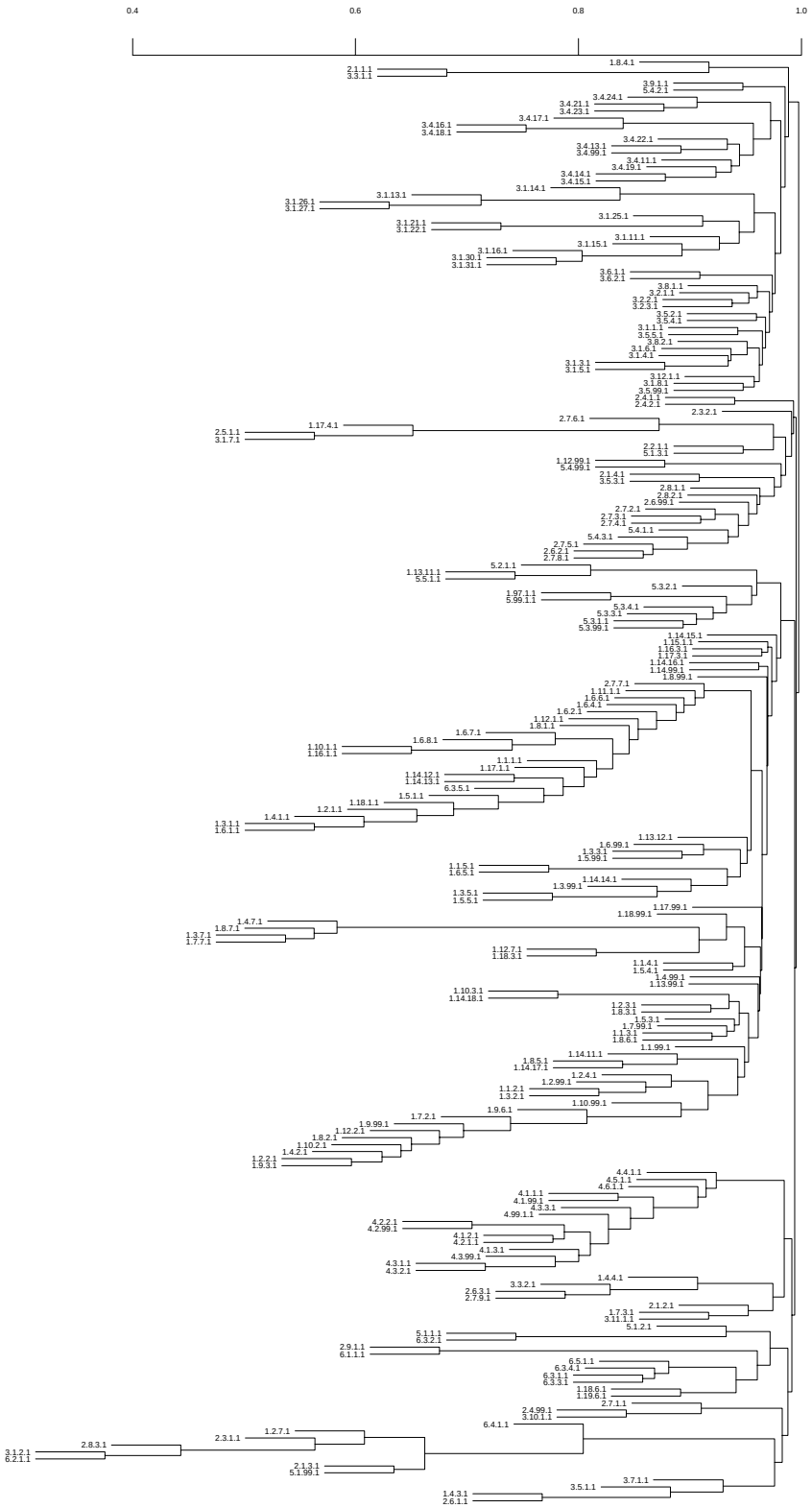


図 9.24: ENZYME:自然言語を多く含むフィールドで、専門用語であり一般英単語ではない索引語を類似度計算の対象としたときの分類分け

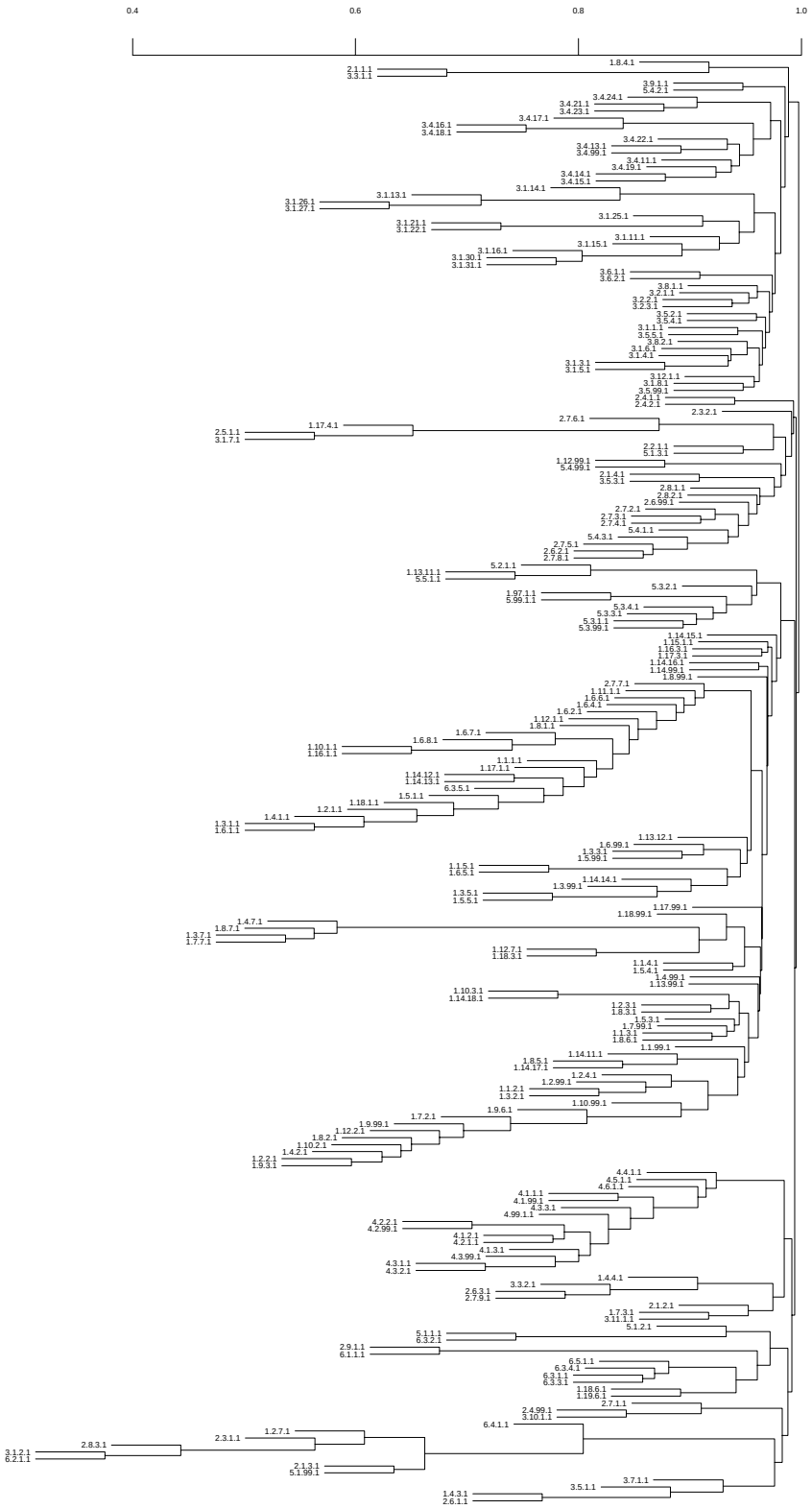


図 9.25: ENZYME:自然言語を多く含むフィールドで専門用語のみを類似度計算の対象としたときの分類分け

図 9.26: PROSITE:自然言語を多く含むフィールドで専門用語のみを類似度計算の対象としたときの分類分け

第 10 章

結論

本研究では、ゲノムネット上で更新されているゲノムデータベースを対象に、広く流通しているサーチエンジンソフトウェアである Namazu を適用し、そのインデクスからエントリ間の類似度を抽出する可能性について、検討と実験を行なった。まず、エントリ内には自然言語情報を多く含むフィールドとそうでないフィールドがあり、後者は類似度計算に有害であることが予想されるため、各フィールドが類似度計算に使えそうかどうかを網羅的に調べた（結果は付録を参照)。また、専門用語、一般英単語、およびそれ以外のゴミを識別するため、公的に配布されている幾つかの辞書を用いて、専門用語と一般英単語の辞書を作成した。

計算機実験としては、まず各種のゲノムデータベースをエントリ単位でファイルに切り出し、次に Namazu を使ってこれをインデクシングする一方で、ベクトル空間法に基づいて作成した類似度計算プログラムを用いてエントリ間の類似度計算を行なった。その結果、各データベースで類似度の分布傾向に違いが出ることが分かった。これは各データベースが自然言語（すなわち専門用語と一般英単語を合わせた、普通の意味での「単語」）を多く含むか、クロスリファレンス情報（すなわち別のエントリ名）などの「準単語」を多く含むか、それともそれ以外のゴミ（配列断片や数値など）を多く含むかの違いによると思われる。9 章でグラフを使って端的に示した通り、単語の種類を識別しない状態では一見すると滑らかな類似度分布が、実は専門用語以外の一般単語や準単語やゴミの出現頻度によって支配されていること、データベースの種類にもよるが専門用語に限定することで多くの場合正しい類似度分布が得られそうなことが分かった。

さらに、適切な分類を行なうための正しい類似度計算が本当にできているかについて、計算機実験を行なった。実験では ENZYME データベースのサブセットを利用し、自然言語を多く含むフィールドとそうでないフィールド（この場合はクロスリファレンス情報を

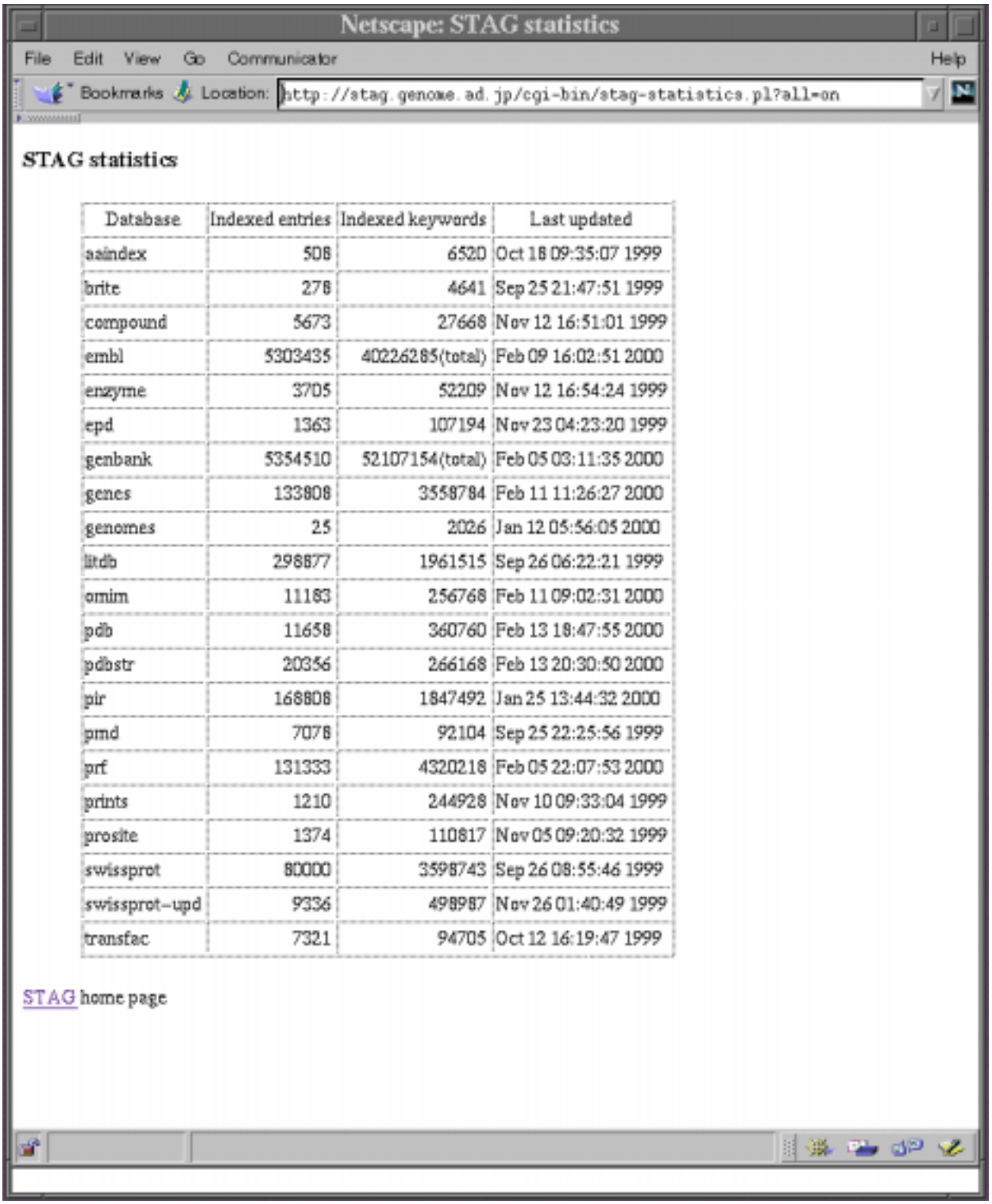
主に含むフィールド）の重みづけを変え、専門用語と一般英単語の識別も行なうことで、各種の条件のもとで分類がどのように変化するかを調べた。結果として、準単語や一般英単語は正しい類似度計算に有害であり、これを排除することで分類がより正しく行なわれることが分かった。この条件で別のデータベース（PROSITE のサブセット）を分類した結果、本手法がデータベースの種類によらずある程度有効であることが端的に示された。これは予想を裏付けており、シングルキーワードのみを使った単語の頻度分布を用いてゲノムデータベースエントリの類似度計算を行なう場合でも、適切な辞書を用意してやることでかなり正確に分類を行なえることが分かった。

今後の課題としては、複数の単語からなる専門用語の認識などがある（現在は mknmz で単語を切り出しているので、シングルワード単位の頻度しか計算できない)。また、同義語や複数形、変化形の認識、シソーラスの利用などの問題もある。これらは自然言語処理手法として既に情報科学分野では確立している技術であるため、割合い簡単に適用できる可能性が高い。さらに進んだ今後の課題としては、抽出した分類情報と他の情報との相関関係の提示（データマイニング）が考えられる。これはテキストマイニングとして近年活発に研究が行なわれている分野であり、一層の成果が期待できる。

最後に、本研究から発展したサービスとして、Namazu を用いたゲノムデータベースの全文検索サービス STAG（図 10.1および図 10.2）が 1999 年 11 月から東京大学医科学研究所ヒトゲノム解析センターで開始され、検索の高速性やヒット率の高さなどの点で、高く評価されていることを付記しておく。このサービスは以下の URL でアクセス可能であり、世界中から利用されている。

http://stag.genome.ad.jp/

図 10.1: STAG のトップページ



Database	Indexed entries	Indexed keywords	Last updated
aaindex	508	6520	Oct 18 09:35:07 1999
brite	278	4641	Sep 25 21:47:51 1999
compound	5673	27668	Nov 12 16:51:01 1999
embl	5303435	40226285(total)	Feb 09 16:02:51 2000
enzyme	3705	52209	Nov 12 16:54:24 1999
epd	1363	107194	Nov 23 04:23:20 1999
genbank	5354510	52107154(total)	Feb 05 03:11:35 2000
genes	133808	3558784	Feb 11 11:26:27 2000
genomes	25	2026	Jan 12 05:56:05 2000
litdb	298877	1961515	Sep 26 06:22:21 1999
omim	11183	256768	Feb 11 09:02:31 2000
pdb	11658	360760	Feb 13 18:47:55 2000
pdbstr	20356	266168	Feb 13 20:30:50 2000
pir	168808	1847492	Jan 25 13:44:32 2000
pmd	7076	92104	Sep 25 22:25:56 1999
prf	131333	4320218	Feb 05 22:07:53 2000
prints	1210	244928	Nov 10 09:33:04 1999
prosite	1374	110817	Nov 05 09:20:32 1999
swissprot	80000	3598743	Sep 26 08:55:46 1999
swissprot-upd	9336	498987	Nov 26 01:40:49 1999
transfac	7321	94705	Oct 12 16:19:47 1999

図 10.2: STAG の statistics

謝辞

本研究を進めるにあたり、終始熱心な御指導を賜わりました佐藤 賢二助教授に心から御礼申し上げます。また、終始貴重なご助言を賜わりました小長谷 明彦教授に心から御礼申し上げます。本研究で大いに参考になった情報検索技術について副テーマにおいて御世話になりました藤波 努助教授に心から御礼申し上げます。また、暖かい御助言を賜わりました奥村 学助教授に心から感謝致します。

計算機を快く御貸し下さいました東京大学医科学研究所の皆様に心から御礼申し上げます。

本研究を様々な側面から御援助下さいました佐藤研究室、小長谷研究室の皆様に感謝致します。

最後に、多くの方々の暖かい御協力により本研究を行うことができたことを心から感謝致します。

参考文献

[1] 馬場 肇, 日本語全文検索システムの構築と活用, ソフトバンク株式会社, 1998.

[2] NamazuHP, http://openlab.ring.gr.jp/namazu/

[3] Peter van der Linden, not just JAVA, 株式会社アスキー, 1998.

[4] 長尾 真, 佐藤 理史, 黒橋 禎夫, 角田 達彦, 自然言語処理, 岩波書店, 1996.

[5] G.salton, Automatic Text Procecessing, Addison-Wesley,1989.

[6] 道本 健二, ジャストの命運を握る新検索技術の実力を見る 日経バイト 12 月号, p232-238 1997.

[7] ConceptBaseHP, http://www.justsystem.co.jp/cb/

[8] 小長谷 明彦, 情報フロンティアシリーズ 23 遺伝子とコンピュータ ー生命設計図をひもとく, 共立出版株式会社, 2000.

[9] 中井謙太, ゲノム解析入門 コンピュータサイエンス誌 bit ６月号, 共立出版株式会社, p3-7, 1999.

[10] 高木利久, ゲノムデータベース コンピュータサイエンス誌 bit ７月号, 共立出版株式会社, p16-22, 1999.

[11] 金久 實, ゲノムネット コンピュータサイエンス誌 bit ８月号, 共立出版株式会社, p28-32, 1999.

[12] 生田 哲, 入門ビジュアルサイエンス ヒト遺伝子の仕組み, 日本実業出版社, 1995.

[13] 高木利久, 金久 實, ゲノムネットのデータベース利用法 [第２版], 共立出版株式会社, 1998.

[14] 配列データベース, http://www.dna.affrc.go.jp/htdocs/mm/db/

[15] GenomeNet WWW server, http://www.jaist.genome.ad.jp/

[16] 日浦 太, 橋本 昭洋, 修士学位論文 分子生物学データベースにおける単語の共起性と出現位置による関連エントリ探索手法, 大阪大学 大学院基礎工学研究科情報数理系専攻 計算機科学分野, 1998.

研究業績

[1] Takao Kataoka and Kenji Satou: A Full-Text Search System Covering the Whole GenomeNet , Genome Informatics 1999, UNIVERSAL ACADEMY PRESS,INC. TOKY-O, JAPAN.

付録

フィールドにおける自然言語情報の多寡

本研究で対象となったゲノムデータベースのフィールドについて、自然言語情報の多寡を評価する。自然言語情報の多寡は ♣♣♣♣♣ と、5 段階評価で示す。

BRITE データベース

DEFINITION

自然言語情報：♣♣♣

```
DEFINITION  Two-hybrid system was used to identify protein-protein
            interactions that occur in the pheromone response pathway. Ste4,
            Ste5, Ste7, Ste11, Ste12, Ste20, Fus3 and Kss1 were tested in
            all pairwise combinations.
```

BRITE の DEFINITION フィールド（CC_SCE00075 エントリ）

ENTRY

自然言語情報：♣

```
ENTRY       DEV_DME00001
```

BRITE の ENTRY フィールド（DEV_DME00001 エントリ）

FACTORS

自然言語情報：♣♣

```
FACTORS
  NAME      activin beta B subunit
  SYMBOL
```

```
  SYNONYM
  CROSS_REF [PIR:I51199]
            [GB:S61773]
  NAME      activin receptor IIB
  SYMBOL    XActRIIB
  SYNONYM   XAR1
  CROSS_REF [SP:P27041]
            [GB:M88594]
            [PS:PS00107]
            [PS:PS00108]
            [PS:PS50011]
  NAME      activin receptor-like kinase 4
  SYMBOL    XALK4
  SYNONYM   XALK4
  CROSS_REF [GB:U60643]
```

BRITE の FACTORS フィールド（DEV_XLA00009 エントリ）

INTERACTION

自然言語情報：♣♣

```
INTERACTION
  STAGE     syncytial blastoderm
  SPACE     embryo ventral side
```

BRITE の INTERACTION フィールド（DEV_DME00018 エントリ）

REFERENSE

自然言語情報：♣♣♣♣

```
REFERENCE
  MEDLINE   93204953
  AUTHORS   Zhou Z, Gartner A, Cade R, Ammerer G, Errede B
  TITLE     Pheromone-induced signal transduction in Saccharomyces
            cerevisiae requires the sequential function of three protein
            kinases.
  JOURNAL   Mol Cell Biol 13, 2069-2080, (1993)
```

BRITE の REFERENSE フィールド（エントリ）

AAindex データベース

A

自然言語情報：♣♣

```
A Vihinen, M., Torkkila, E. and Riikonen, P.
```

AAindex のフィールド（VINM940101 エントリ）

C

自然言語情報：♣

```
C NOZY710101    0.917  EISD860101    0.912  MEEJ800102    0.900
  CIDH920102    0.862  CIDH920105    0.861  ZIMJ680105    0.851
  FAUJ830101    0.846  PLIV810101    0.845  VENT840101    0.826
  MIYS850101    0.824  SWER830101    0.820  CIDH920103    0.819
  CIDH920104    0.817  MEEJ810102    0.813  BROC820101    0.811
  MEIH800101   -0.816  VHEG790101   -0.818  LEVM760101   -0.838
  BULH740101   -0.856  HOPT810101   -0.859  WOLS870101   -0.873
  PARJ860101   -0.883
```

AAindex の C フィールド（RADA880102 エントリ）

D

自然言語情報：♣♣

```
D Average relative fractional occurrence in E0(i-1) (Rackovsky-Scheraga, 1982)
```

AAindex の D フィールド（RACS820111 エントリ）

H

自然言語情報：♣

```
H RACS820107
```

AAindex の H フィールド（RACS820107 エントリ）

I

自然言語情報：♣

```
I   A/L    R/K    N/M    D/F    C/P    Q/S    E/T    G/W    H/Y    I/V
    6.5   -0.9   -5.1    0.5   -1.3    1.0    7.8   -8.6    1.2    0.6
    3.2    2.3    5.3    1.6   -7.7   -3.9   -2.6    1.2   -4.5    1.4
```

AAindex の I フィールド（ROBB760101 エントリ）

J

自然言語情報：♣

```
J Biochemistry 27, 1664-1670 (1988)
* (Pro Cys Asp missing)
```

AAindex の J フィールド（RADA880102 エントリ）

R

自然言語情報：♣

```
R LIT:1505154 PMID:3221397
```

AAindex の R フィールド（RISJ880101 エントリ）

T

自然言語情報：♣♣♣

```
T Comparing the polarities of the amino acids: Side-chain distribution
  coefficients between the vapor phase, cyclohexane, 1-octanol, and
  neutral aqueous solution
```

AAindex の T フィールド（RADA880101 エントリ）

PRINTS データベース

GR

自然言語情報：♣♣♣

```
GR; POLIKARPOV, I., NORTH, A.C.T. AND SAWYER, L.
GR; CAVAGGIONI, A., FINDLAY, J.B.C. AND NORTH, A.C.T.
```

PRINTS の GR フィールド（BETALACGMUP エントリ）

bb

自然言語情報：♣

```
bb;
bb;
bb;
bb;
bb;
```

```
bb;
bb;
bb;
bb;
bb;
```

PRINTS の bb フィールドの一部（UBIQUITIN エントリ）

cd

自然言語情報：♣

```
cd;  3|  16   16   16
cd;  2|   0    0    0
cd; --+----------------
cd;   |   1    2    3
```

PRINTS の cd フィールド（TEADOMAIN エントリ）

ci

自然言語情報：♣

```
ci; COMPOSITE FINGERPRINT INDEX
ci; --------------------------
```

PRINTS の ci フィールド（TAUTRANSPORT エントリ）

cr

自然言語情報：♣

```
cr:
```

PRINTS の cr フィールド（TASKCHANNEL エントリ）

dn

自然言語情報：♣

```
dn; OWL11_1     3   148 NSINGLE
dn; OWL17_1     1   231 NSINGLE
dn; OWL18_0     1   300 NSINGLE
dn; OWL19_1     1   325 NSINGLE
dn; OWL26_0     1   510 NSINGLE
dn; SPTR37_9f   8   500 NSINGLE
```

PRINTS の dn フィールド（TYRKINASE エントリ）

fc

自然言語情報：♣

```
fc; THYMDSNTHASE1
fc; THYMDSNTHASE2
fc; THYMDSNTHASE3
fc; THYMDSNTHASE4
fc; THYMDSNTHASE5
```

PRINTS の fc フィールド（THYMDSNTHASE エントリ）

fd

自然言語情報：♣

```
fd; QAVKDKEAPMKVAQTRL                    A46170    356   356
fd; QAIKDKENPLKLAMTRL                 SUSMRNASEQ    275   275
fd; RTHRPNVELCRDPAQYRMVQE                A46170    375     2
fd; RTYRPNVELCRDNAQYGLVNE             SUSMRNASEQ    294     2
fd; EIQDSIDKLQQKLAESK                    A46170    398     2
fd; EIQDSIKALEKKLQDAH                 SUSMRNASEQ    317     2
fd; KEIALKKNSIFVD                        A46170    429    14
fd; KDLELKNNSLDLD                     SUSMRNASEQ    348    14
```

PRINTS の fd フィールド（TEKTIN エントリ）

fl

自然言語情報：♣

```
fl; 15
fl; 12
fl; 13
fl; 18
fl; 17
fl; 24
```

PRINTS の fl フィールド（VNEBNERGLAND エントリ）

fm

自然言語情報：♣

```
fm; FINAL MOTIF-SETS
fm; ----------------
```

PRINTS の fm フィールド（TBOX エントリ）

ft

自然言語情報：♣

```
ft; Tektin motif I - 2
ft; Tektin motif II - 2
ft; Tektin motif III - 2
ft; Tektin motif IV - 2
```

PRINTS の ft フィールド（TEKTIN エントリ）

ga

自然言語情報：♣

```
ga; 14-NOV-1994; UPDATE 23-JUN-1999
```

PRINTS の ga フィールド（TCOMPLEXTCP1 エントリ）

gc

自然言語情報：♣

```
gc; THERMOPTASE
```

PRINTS の gc フィールド（THERMOPTASE エントリ）

gd

自然言語情報：♣♣♣♣♣

```
gd; Tafazzins [1] are expressed in high levels in cardiac and skeletal muscle.
gd; As many as 10 isoforms can be present in different amounts in different
gd; tissues. Isoforms with hydrophobic N-termini are thought to be membrane
gd; anchored, while shorter forms, lacking the hydrophobic stretch, may be
gd; cytoplasmic (these latter are found in leukocytes and fibroblasts, but not
gd; in heart and skeletal muscle). A central hydrophilic domain may serve as
gd; an exposed loop that interacts with other proteins.
gd;
gd; Defects in taz are the cause of Barth syndrome, a severe inherited disorder,
gd; often fatal in childhood [1]. The disease is characterised by cardiac and
```

```
gd; skeletal myopathy, short stature and neutropenia [1].
gd;
gd; TAFAZZIN is a 6-element fingerprint that provides a signature for
gd; tafazzins. The fingerprint was derived from an intitial alignment of 3
gd; sequences: the motifs were drawn from short conserved regions spanning
gd; the central portion of the alignment - motif 4 lies in the hydrophilic
gd; domain. A single iteration on OWL30.2 was required to reach convergence,
gd; no further sequences being identified beyond the starting set.
gd;
gd; An update on SPTR37_9f identified a true set of 3 sequences.
```

PRINTS の gd フィールド（TAFAZZIN エントリ）

gn

自然言語情報：♣

```
gn; COMPOUND(11)
```

PRINTS の gn フィールド（TASKCHANNEL エントリ）

gp

自然言語情報：♣♣♣

```
gp; PROSITE; PS00316 THAUMATIN
gp; BLOCKS; BL00316
gp; PFAM; PF00314 thaumatin
gp; PDB; 1THI
gp; SCOP; 1THI
```

PRINTS の gp フィールド（THAUMATIN エントリ）

gr

自然言語情報：♣♣♣♣♣

```
gr; 1. WAGNER, R.L., APRILETTI, J.W., MCGRATH, M.E., WEST, B.L., BAXTER, J.D.
gr; AND FLETTERICK, R.J.
gr; A structural role for hormone in the thyroid hormone receptor.
gr; NATURE 378 690-697 (1995).
gr;
gr; 2. CHEN, J.D. AND EVANS, R.M.
gr; A transcriptional co-repressor that interacts with nuclear hormone
```

```
gr; receptors.
gr; NATURE 377 454-457 (1995).
```

PRINTS の gr フィールド（THYROIDHORMR エントリ）

gt

自然言語情報：♣♣

```
gt; Xeroderma pigmentosum group G/yeast RAD superfamily signature
```

PRINTS の gt フィールド（XPGRADSUPER エントリ）

gx

自然言語情報：♣

```
gx; PR00752
```

PRINTS の gx フィールド（VASOPRSNV1AR エントリ）

ic

自然言語情報：♣

```
ic; VWFADOMAIN1
ic; VWFADOMAIN2
ic; VWFADOMAIN3
```

PRINTS の ic フィールド（VWFADOMAIN エントリ）

id

自然言語情報：♣

```
id; MGSDVRDLNALLPAVSS                WT1_MOUSE      1     1
id; MGSDVRDLNALLPAVPS                WT1_HUMAN      1     1
id; GGGGGCGLPVSGARQWA                WT1_MOUSE     20     2
id; GGGGGCALPVSGAAQWA                WT1_HUMAN     19     1
id; VLDFAPPGASAYGSL                  WT1_MOUSE     38     1
id; VLDFAPPGASAYGSL                  WT1_HUMAN     37     1
id; GGPAPPPAPPPPPPP                  WT1_MOUSE     53     0
id; GGPAPPPAPPPPPPP                  WT1_HUMAN     52     0
```

PRINTS の id フィールド（WILMSTUMOUR エントリ）

il

自然言語情報：♣

```
il; 14
il; 25
il; 25
il; 25
il; 25
il; 27
```

PRINTS の il フィールド（TOGAVIRIN エントリ）

im

自然言語情報：♣

```
im; INITIAL MOTIF-SETS
im; ------------------
```

PRINTS の im フィールド（TPI2FAMILY エントリ）

it

自然言語情報：♣

```
it; Ubiquitin motif I - 1
it; Ubiquitin motif II - 1
it; Ubiquitin motif III - 1
```

PRINTS の it フィールド（UBIQUITIN エントリ）

sd

自然言語情報：♣

```
sd;   21 codes involving  5 elements
sd;    1 codes involving  4 elements
sd;    2 codes involving  3 elements
sd;    0 codes involving  2 elements
```

PRINTS の sd フィールド（URICASE エントリ）

sh

自然言語情報：♣

```
sh; SCAN HISTORY
sh; ------------
```

PRINTS の sh フィールド（TETREPRESSOR エントリ）

si

自然言語情報：♣

```
si; SUMMARY INFORMATION
si; -------------------
```

PRINTS の si フィールド（TMPROTEINSRA エントリ）

sn

自然言語情報：♣

```
sn; Codes involving 4 elements
sn; Codes involving 3 elements
sn; Codes involving 2 elements
```

PRINTS の sn フィールド（S21N4MTFRASE エントリ）

st

自然言語情報：♣

```
st; SCXA_MESMA     SCX1_ORTSC     SCX2_TITSE     SCX1_TITBA
st; SCXE_BUTOC     SCX2_TITBA     SCXL_ANDAU     SCX3_BUTOM
st; SCX3_ORTSC     SIXE_BUTJU
st; SCX1_ANDAU     SCX3_ANDAU     SCX4_ANDAU     SIX4_ANDAU
st; O77091         SCXV_TITSE     SIX1_ANDAU     SIX1_MESMA
```

PRINTS の st フィールド（SCORPNTOXIN エントリ）

tbb

自然言語情報：♣

```
tbb;
```

PRINTS の tbb フィールド（TMPROTEINSRA エントリ）

tp

自然言語情報：♣

```
tp; VA5_VESPE      VA5_VESMC      VA5_VESVU      VA5_VESFL
tp; VA5_VESGE      VA5_VESVI      VA52_VESCR     VA51_VESCR
tp; VA5_DOLAR      VA5_VESSQ      VA5_POLFU      VA5_POLEX
tp; VA52_DOLMA     VA3_SOLRI      O16135         VA3_SOLIN
```

PRINTS の tp フィールド（V5ALLERGEN エントリ）

tt

自然言語情報：♣

```
tt; AMPT_THETH   AMINOPEPTIDASE T (EC 3.4.11.-) (AP-T) (HEAT STABLE AMINOPEPTIDASE) - THERMUS AQU
tt; AMPT_THEAQ   AMINOPEPTIDASE T (EC 3.4.11.-) (AP-T) (HEAT STABLE AMINOPEPTIDASE) - THERMUS AQU
tt; AMP2_BACST   AMINOPEPTIDASE II (EC 3.4.11.-) (AP-II) - BACILLUS STEAROTHERMOPHILUS.
tt; AMPS_BACSU   AMINOPEPTIDASE AMPS (EC 3.4.11.-) - BACILLUS SUBTILIS.
tt;
tt; O51096       AMINOPEPTIDASE II - BORRELIA BURGDORFERI (LYME DISEASE SPIROCHETE).
```

PRINTS の tt フィールド（THERMOPTASE エントリ）

EPD データベース

RT

自然言語情報：♣♣♣♣

```
RT   "Sequence homologies in the region preceding the transcription
RT   initiation site of the liver estrogen-responsive vitellogenin and
RT   apo-VLDLII genes";
RT   "Vitellogenin genes A1 and B1 are linked in the Xenopus laevis
RT   genome.";
```

EPD の RT フィールド（XL_VTB1 エントリ）

フィールド内容：RT = Reference Title.[?]

ME

自然言語情報：♣♣♣

```
ME   Nuclease protection with homologous sequence ladder [1].
ME   Nuclease protection; injected amphibian oocytes [3].
```
EPD の ME フィールド（SP_H2BL エントリ）

フィールド内容：ME = Methods.

RA

自然言語情報：♣♣♣

```
RA   Guyader M., Emerman M., Sonigo P., Clavel F., Montagnier L.,
RA   Alizon M.;
RA   Franchini G., Gurgo C., Guo H.G., Gallo R.C., Collalti E.,
RA   Fargnoli K.A., Hall L.F., Wong-staal F., Reitz M.S.;
RA   Kornfeld H., Riedel N., Viglianti G.A., Hirsch V., Mullins J.I.;
```
EPD の RA フィールド（SRV1_LTR エントリ）

フィールド内容： RA = Reference Authors.

DE

自然言語情報：♣♣

```
DE   Class I transplantation antigens of major histocompatibility
DE   complex PD1
```
EPD の DE フィールド（SS_MHCP エントリ）

フィールド内容： DE = Description

DO

自然言語情報：♣♣

```
DO        Experimental evidence: 3
DO        Expression/Regulation: embryo(oral ectoderm),.;+heavy Me++
```
EPD の DO フィールド（SP_MTB エントリ）

フィールド内容： DO = Documentation

DR

自然言語情報：♣

```
DR   EMBL; X03712.1; XLAGA1G; [-1563, 129].
DR   SWISS-PROT; P02012; HBA1_XENLA.
```

EPD の DR フィールド（XL_HBA1 エントリ）

フィールド内容： DR = Database cross-References

KW

自然言語情報：♣♣

```
KW   Chlorophyll, Photosynthesis, Multigene famiy, Transmembrane,
KW   Light-harvesting complex, chlorophyll binding protein.
```

EPD の KW フィールド（NP_CABE エントリ）

AP

自然言語情報：♣

```
AP   Alternative promoter #1 of 2; exon 1; site 1; major promoter.
```

EPD の AP フィールド（ZM_PML1_1 エントリ）

フィールド内容： AP = Alternative Promoter.

DT

自然言語情報：♣

```
DT   ??-JUN-1993 (Rel. 35, created)
DT   07-JUN-1999 (Rel. 59, Last annotation update).
```

EPD の DT フィールド（SS_CPT7 エントリ）

フィールド内容： DT = Date

FP

自然言語情報：♣

```
FP   Zm zein 19K C    P1 :-S  EM:X53582.1    1+    1500; 58014.001 057*1
```

EPD の FP フィールド（ZM_ZEAC_1 エントリ）

フィールド内容： FP = Functional Position.

HG

自然言語情報：♣♣

```
HG   Homology group 189; Sea urchin ectoderm enriched type 2 RNAs
```

EPD の HG フィールド（SP_SP2A エントリ）

フィールド内容： HG = Homology Group.

ID

自然言語情報：♣

```
ID   XL_VTA2     standard; single; VRT.
```

EPD の ID フィールド（XL_VTA2 エントリ）

フィールド内容： ID = Identification.

OS

自然言語情報：♣♣

```
OS   Xenopus tropicalis (western clawed frog)
```

EPD の OS フィールド（XT_ACT2 エントリ）

フィールド内容： OS = Organism Species.

RF

自然言語情報：♣♣

```
RF        Nat279:737     PNAS77:1265    PNAS85:507
```

EPD の RF フィールド（SP_H2BE エントリ）

フィールド内容： RF = literature Reference.

SQ

自然言語情報：♣♣

```
SQ   Sequence 600 BP; 159 A; 160 C; 134 G; 147 T; 0 other;
     tcgatccctt ctttccagcg gcttgtgctt tggcaggcat gatgagttac tctacgtgat
     aaacgatgag aatgaactgc caagcgaatc cacttctatt tatacagcga gcgaggattt
     aacggttata cgcttatgaa aatgagtccg actgcacgcg agaaacacca atcactgcaa
     gccattcagc gcggtttcgc tccgtgtacg agagaacgac ggcccctgaa ttaattcatt
     attcatgagg tccgaatgta cgtttgagcc accaatcaca cagagcgctc tacgtaaata
     cgcagggccc cgctgttcgg gcgacacatt tgcatacacc cgtgcaaaag catgcgtaca
     ctcgcacgta tatgcaaata atagtgtgtt cgcttgccgt tactcatcgg ccccgcatct
     gattggctcc cattggatcc tcgctgtgcg tttcgatcct ccacagacgt ataaatacct
     agctcgcacc aatttggaag catacagcga ttctcatctt acttgccaaa gcgtaaccaa
     atctatcaaa tcatcatgtc tggacgtggt aaaggagcag gaaaggcccg tgctaaggcc
```

EPD の SQ フィールド（SP_H2AL エントリ）

フィールド内容： SQ = Sequence.

SE

自然言語情報：♣♣

```
SE   aagcacacaaattttggtatgtatgtccaatcgtgtatccatcacctatAATATTTTGAG
```

EPD の SE フィールド（ZM_PML1_1 エントリ）

AC

自然言語情報：♣♣

```
AC   EP11006;
```

EPD の AC フィールド（ZM_ADH1 エントリ）

フィールド内容： AC = Accession number(s).

RL

自然言語情報：♣♣♣

```
RL   Nucleic Acids Res. 18:111-117(1990).
RL   Gene 80:249-257(1989).
RL   Eur. J. Cell Biol. 42:161-170(1986).
RL   Cell 29:1015-1026(1982).
```

EPD の RL フィールド（ZM_ZEAC_1 エントリ）

フィールド内容： RL = Reference Location.

RN

自然言語情報：♣

```
RN   [1]
RN   [2]
RN   [3]
RN   [4]
```

EPD の RN フィールド（ZM_ZEAC_1 エントリ）

フィールド内容： RN = Reference Number.

RX

自然言語情報：♣

```
RX   MEDLINE; 87173056.
RX   MEDLINE; 87287229.
RX   MEDLINE; 87173040.
```

EPD の RX フィールド（SRV1_LTR エントリ）

フィールド内容： RX = Reference cross-references.

TX

自然言語情報：♣♣♣

```
TX   5. Echinoderm promoters
TX   5.1. Chromosomal genes
TX   5.1.6. Unclassified
TX   5.1.6.1. Genes defined by regulatory properties
TX   5.1.6.1.1. Spec1/Spec2 family of calmodulin-related proteins
```

EPD の TX フィールド（エントリ）

フィールド内容： TX = Taxonomy.

XX

自然言語情報：♣

```
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
```

EPD の XX フィールド（XL_U2 エントリ）

PROSITE データベース

DR

自然言語情報：♣

```
DR   P33395, RRF2_DESVH, T; Q48660, Y156_LACLA, T; O68025, Y166_RHOCA, T;
DR   Q55433, Y846_SYNY3, T; O07465, YB01_RHOPA, T; Q10613, YC87_MYCTU, T;
DR   P77484, YFHP_ECOLI, T; P44675, YFHP_HAEIN, T; O07573, YHDE_BACSU, T;
```

```
DR   P21498, YJEB_ECOLI, T; P40610, YJEB_VIBPA, T; Q51134, YLDA_NEIME, T;
DR   O69219, YOR2_AZOVI, T; O34527, YRZC_BACSU, T; P71047, YWGB_BACSU, T;
```

PROSITE の DR フィールド（UPF0074 エントリ）

フィールド内容：DR = Cross-reference to SWISS-PROT
1 エントリに 0 個以上記述。

RU

自然言語情報：♣♣♣

```
RU   Additional rules:
RU   (1) The cysteine must be between positions 15 and 35 of the sequence in
RU       consideration.
RU   (2) There must be at least one charged residue (Lys or Arg) in the first
RU       seven residues of the sequence.
```

PROSITE の RU フィールド（PROKAR_LIPOPROTEIN エントリ）

フィールド内容：RU = Rule
1 エントリに 0 個以上記述。

DE

自然言語情報：♣♣

```
DE   Neutral zinc metallopeptidases, zinc-binding region signature.
```

PROSITE の DE フィールド（ZINC_PROTEASE エントリ）

フィールド内容：DE = Short description
1 エントリに 1 つ記述。

DT

自然言語情報：♣

```
DT   JUN-1994 (CREATED); JUN-1994 (DATA UPDATE); JUL-1998 (INFO UPDATE).
```

PROSITE の DT フィールド（TRANSPOSASE_MUTATOR エントリ）

フィールド内容：DT = Date
1 エントリに 1 つ記述。例のフィールドは、1994 年 6 月に登録され、最後に更新されたのは 1998 年 7 月であることを示している。

MA

自然言語情報：♣

```
MA   /GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=20;
MA   /DISJOINT: DEFINITION=PROTECT; N1=3; N2=18;
MA   /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=-0.3853; R2=0.01803082; TEXT='NScore';
MA   /CUT_OFF: LEVEL=0; SCORE=492; N_SCORE=8.5; MODE=1;
MA   /CUT_OFF: LEVEL=-1; SCORE=381; N_SCORE=6.5; MODE=1;
MA   /DEFAULT: D=-20; I=-20; B1=-50; E1=0; MI=-105; MD=-105; IM=-105; DM=-105;
MA   /I: B1=0; BI=-105; BD=-105;
MA   /M: SY='N'; M=-5,6,-21,-6,-11,-13,-14,-9,5,-13,-12,-6,20,-18,-7,-13,6,1,-4,-33,-13,-11;
MA   /M: SY='R'; M=-16,-5,-30,-3,15,-25,-20,2,-28,23,-20,-10,0,-15,20,47,-6,-10,-24,-22,-12,14;
MA   /M: SY='E'; M=-9,12,-30,21,22,-32,14,-7,-36,-3,-26,-22,5,-9,1,-9,0,-14,-30,-29,-24,12;
```

PROSITE の MA フィールドの一部（BH4_2 エントリ）

フィールド内容：MA = Matrix/profile
1 エントリに 1 つ記述。

NR

自然言語情報：♣

```
NR   /RELEASE=38,80000;
NR   /TOTAL=10(10); /POSITIVE=10(10); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR   /FALSE_NEG=0; /PARTIAL=2;
```

PROSITE の NR フィールド（VINCULIN_1 エントリ）

フィールド内容：NR = Numerical results
1 エントリに 0 個以上記述。

CC

自然言語情報：♣

```
CC   /TAXO-RANGE=??E??; /MAX-REPEAT=1;
CC   /SITE=2,zinc; /SITE=4,zinc; /SITE=9,zinc; /SITE=11,zinc; /SITE=13,zinc;
CC   /SITE=15,zinc;
```

PROSITE の CC フィールド（ZN2_CY6_FUNGAL_1 エントリ）

フィールド内容： CC = Comments
1 エントリに 0 個以上記述。

ID

自然言語情報：♣

```
ID   XYLOSE_ISOMERASE_1; PATTERN.
```

PROSITE の ID フィールド（XYLOSE_ISOMERASE_1 エントリ）

フィールド内容： ID = Identification
1 エントリに 1 つ記述。エントリの一番最初に記述。Prosite の ID が示されている。この例では、XYLOSE_ISOMERASE_1 という名称であることを示している。また、PATTERN. により、配列の規則性を示している。（法則の場合は RULE と記述している。）

AC

自然言語情報：♣

```
AC   PS00802;
```

PROSITE の AC フィールド（TRANSKETOLASE_2 エントリ）

フィールド内容： AC = Accession number
1 エントリに 1 つ記述。

DO

自然言語情報：♣

```
DO   PDOC00290;
```

PROSITE の DO フィールド（TROPOMYOSIN エントリ）

フィールド内容： DO = Pointer to the documentation file
1 エントリに 1 つ記述。DO フィールドに記されているドキュメンテーションファイルに関連事項が掲載されていることが示されている。

PA

自然言語情報：♣

```
PA   Y-x(2)-F-[LIVMA](2)-x-L-x(4)-G-x(2)-F-[EQ]-[LIVMF]-P-[LIVM].
```

PROSITE の PA フィールド（UPF0032 エントリ）

フィールド内容： PA = Pattern
1 エントリに 0 個以上記述。

ENZYME データベース

DBLINKS

自然言語情報：♣♣♣♣

```
DBLINKS     University of Geneva ENZYME DATA BANK: 6.3.5.1
            WIT (What Is There) Metabolic Reconstruction: 6.3.5.1
            BRENDA, the Enzyme Database: 6.3.5.1
            SCOP (Structural Classification of Proteins): 6.3.5.1
```

ENZYME の DBLINKS フィールド（6.3.5.1 エントリ）

COMMENT

自然言語情報：♣♣♣♣

```
COMMENT     Transforms glutathione and spermidine into glutathionylspermidine.
            Involved in the synthesis of trypanothione in trypanosomatids.
            The enzyme from E.coli is bifunctional and also catalyses the
            glutathionylspermidine amidase (EC 3.5.1.78) reaction, resulting in
            a net hydrolysis of ATP.
```

ENZYME の COMMENT フィールド（6.3.1.8 エントリ）

CLASS

自然言語情報：♣♣♣

```
CLASS       Ligases
            Forming carbon-oxygen bonds
            Ligases forming aminoacyl-tRNA and related compounds
```

ENZYME の CLASS フィールド（6.1.1.21 エントリ）

PRODUCT

自然言語情報：♣♣♣

```
PRODUCT     ADP
            Orthophosphate
            1-(5-Phosphoribosyl)imidazole-4-acetate
            Pyrophosphate
```

ENZYME のフィールド（6.3.4.8 エントリ）

SYSNAME

自然言語情報：♣♣♣

```
SYSNAME     Biotin:apo-[3-methylcrotonoyl-CoA:carbon-dioxide ligase
            $ (ADP-forming)] ligase (AMP-forming)
```

ENZYME の SYSNAME フィールド（6.3.4.11 エントリ）

DISEASE

自然言語情報：♣♣♣

```
DISEASE     MIM: 261515  Enoyl-Coenzyme A, hydratase/3-hydroxyacyl Coenzyme A
                         dehydrogenase; Peroxisomal bifunctional enzyme
                         deficiency (1)
            MIM: 600305  Dodecenoyl-Coenzyme A delta isomerase (3,2
                         trans-enoyl-Coenzyme A
```

ENZYME の DISEASE フィールド（5.3.3.8 エントリ）

INHIBITOR

自然言語情報：♣♣♣

```
INHIBITOR   Aspirin
            Ibuprofen
            Diflunisal
            Diclofenac
            Fenprofen calcium
            Indomethacin
            Ketoprofen
            Meclofenamate sodium
            Mefenamic acid
            Naproxen
            Piroxicam
            Sulindac
            Tolmetin sodium
```

ENZYME の INHIBITOR フィールド（1.14.99.1 エントリ）

NAME

自然言語情報：♣♣♣

```
NAME        Acetate--CoA ligase
            Acetyl-CoA synthetase
            Acetyl activating enzyme
            Acetate thiokinase
            Acyl-activating enzyme
```

ENZYME の NAME フィールド（6.2.1.1 エントリ）

PATHWAY

自然言語情報：♣♣♣

```
PATHWAY     PATH: MAP00252  Alanine and aspartate metabolism
            PATH: MAP00330  Arginine and proline metabolism
            PATH: MAP00340  Histidine metabolism
            PATH: MAP00410  beta-Alanine metabolism
```

ENZYME の PATHWAY フィールド（6.3.2.11 エントリ）

REACTION

自然言語情報：♣♣

```
REACTION    ATP + Biotin + Apo-[propionyl-CoA:carbon-dioxide ligase
            $ (ADP-forming)] = AMP + Pyrophosphate +
            [Propionyl-CoA:carbon-dioxide ligase (ADP-forming)]
```

ENZYME の REACTION フィールド（6.3.4.10 エントリ）

SYSNAME

自然言語情報：♣♣

```
SYSNAME     3alpha,7alpha,12alpha-Trihydroxy-5beta-cholestanate:CoA ligase
            $ (AMP-forming)
```

ENZYME の SYSNAME フィールド（6.2.1.29 エントリ）

GENES

自然言語情報：♣

```
GENES       BSU: pycA(pycA)
            MTU: Rv2967c(pca)
            AAE: aq_1517(pycB)
            SCE: YBR218C(PYC2) YGL062W(PYC1)
```

```
            CEL: D2023.2
            MMU: 97520(Pcx)
            HSA: 5091
```
ENZYME の GENES フィールド（6.4.1.1 エントリ）

SUBSTRATE

自然言語情報：♣♣

```
SUBSTRATE   ATP
            Benzoate
            CoA
            2-Fluorobenzoate
            3-Fluorobenzoate
            4-Fluorobenzoate
```
ENZYME の SUBSTRATE フィールド（6.2.1.25 エントリ）

COFACTOR

自然言語情報：♣

```
COFACTOR    D-Ribose 1,5-bisphosphate
            2-Deoxy-D-ribose 1,5-bisphosphate
            alpha-D-Glucose 1,6-bisphosphate
```
ENZYME の COFACTOR フィールド（5.4.2.7 エントリ）

EFFECTOR

自然言語情報：♣

```
EFFECTOR    Phospholipid
            Calcium
```
ENZYME の EFFECTOR フィールド（3.4.21.60 エントリ）

STRUCTURES

自然言語情報：♣

```
STRUCTURES  PDB: 1AUS  1BUR  1BWV  1BXN  1RBA  1RBL  1RBO  1RCO  1RCX  1RLC
                 1RLD  1RSC  1RUS  1AA1  2RUS  3RUB  4RUB  5RUB  8RUC  9RUB
                 1RXO
```
ENZYME の STRUCTURES フィールド（4.1.1.39 エントリ）

MOTIF

自然言語情報：♣

```
MOTIF       PS: PS00452  G-V-[LIVM]-x(0,1)-G-x(5)-[FY]-x-[LIVM]-[FYW]-[GS]-
                         [DNTHKW]-[DNT]-[IV]-[DNTA]-x(5)-[DE]
            PS: PS00458  G-P-x-C-x-Y-x-A-A-x-V-x-R-x(3)-H-W
            PS: PS50011  Protein kinase domain profile
```

ENZYME の MOTIF フィールド（4.6.1.2 エントリ）

ENTRY

自然言語情報：♣

```
ENTRY       EC 6.1.1.20
```

ENZYME の ENTRY フィールド（6.1.1.20 エントリ）

COMPOUND データベース

PATHWAY

自然言語情報：♣♣

```
PATHWAY     PATH: MAP00860  Porphyrin and chlorophyll metabolism
```

COMPOUND の PATHWAY フィールド（C06319 エントリ）

NAME

自然言語情報：♣♣

```
NAME        Hydrogenobyrinate a,c diamide
            Hydrogenobyrinic acid a,c diamide
            Hydrogenobyrinate diamide
```

COMPOUND の NAME フィールド（C06503 エントリ）

DBLINK

自然言語情報：♣

```
DBLINKS     CAS: 304-21-2
```

COMPOUND の DBLINK フィールド（C06536 エントリ）

ENTRY

自然言語情報：♣

```
ENTRY       C06530
```

COMPOUND の ENTRY フィールド（C06530 エントリ）

ENZYME

自然言語情報：♣

```
ENZYME      3.5.3.21 (R)    3.5.3.21 (R)
```

COMPOUND の ENZYME フィールド（C06382 エントリ）

FORMULA

自然言語情報：♣

```
FORMULA     C58H88CoN16O14P
```

COMPOUND の FORMULA フィールド（C06509 エントリ）

STRUCTURES

自然言語情報：♣

```
STRUCTURES  PDB: 1AG7  1AS5  1AV3  1BKV  1CGD  1GIB  1JOH  1KCP  1P1P  1VIB
                 1WCT  2CCO
```

COMPOUND の STRUCTURES フィールド（C06509 エントリ）

PMD データベース

FUNCTION

自然言語情報：♣♣

```
FUNCTION        Insulin-binding [=]: Tyrosine kinase activity [0]:Insulin-
                  stimulated uptake of glucose.
FUNCTION        Insulin-binding [=]: Tyrosine kinase activity [0]:Insulin-
                  stimulated uptake of glucose.
FUNCTION        Insulin-binding [=]: Tyrosine kinase activity [0]:Insulin-
                  stimulated uptake of glucose.
```

PMD の FUNCTION フィールド（R870188 エントリ）

REFERENCE

自然言語情報：♣♣♣♣

```
REFERENCE    1. Chaleff R.S. & Mauvais C.J. Science (1984) 224, 1443-1445
             2. Chaleff R.S. & Ray T.B. Science (1984) 223, 1148-1151
             3. Chaleff R.S., Sebastian S.A., Creason G.L., Mazur B.J.,
                Falco S.C., Ray T.B., Mauvais C.J. & Yadav N.B. (1986)
                In arntzen,C.J. & Ryan,C.A. (eds), UCLA symposium on
                Molecular and Cellular Biology. Alan R.Liss, NY, New
                series, Vol. 48, pp415-425
```

PMD の REFERENCE フィールド（A88023 エントリ）

PURPOSE

自然言語情報：♣♣♣

```
PURPOSE          Systematic alteration of extra-membrane loops by insertion of
                   an epitope.
```

PMD の PURPOSE フィールド（A930064 エントリ）

フィールド内容： PURPOSE
研究の目的

TITLE

自然言語情報：♣♣♣

```
TITLE            Protein engineering and the study of structure-function
                   relationships in receptors.
```

PMD の TITLE フィールド（R900735 エントリ）

フィールド内容： TITLE
題名。

CHANGE

自然言語情報：♣

```
CHANGE           See ENTRY A870291,A890295 and A900470.
CHANGE           See ENTRY A890291.
CHANGE           See ENTRy A900143.
```

```
CHANGE          See ENTRY A870233.
CHANGE          See ENTRY A890397.
CHANGE          See ENTRY A880454.
```
PMD の CHANGE フィールド（R921441 エントリ）

フィールド内容： CHANGE
突然変異体のポジション、種類。

AUTHOR

自然言語情報：♣♣

```
AUTHORS         Bell G.I., Froguel P., Nishi S., Pilkis S.J., StoffelM.,
                  Takeda J., Vionnet N. & Yasuda K.
```
PMD の AUTHOR フィールド（R931331 エントリ）

フィールド内容： AUTHOR
著者。

COMMENT

自然言語情報：♣♣♣

```
COMMENT         Ther have determined the conformation of an analogue of LamB
                  signal peptide inserted into a model membrane using the
                  transferred nuclear Overhauser effect NMR technique.
```
PMD の COMMENT フィールド（A931495 エントリ）

フィールド内容： COMMENT
コメント。

TRANSPORT

自然言語情報：♣

```
TRANSPORT       Secretion [- -]: It remained in ER.
TRANSPORT       Secretion [=]
TRANSPORT       Secretion [- -]
TRANSPORT       Secretion [0]
```
PMD の TRANSPORT フィールド（A921417 エントリ）

PROTEIN

自然言語情報：♣♣

```
PROTEIN         Plasminogen
                #Length 791AAs
PROTEIN         Fibrinogen alpha chain
PROTEIN         Fibrinogen gamma chain
PROTEIN         Prothrombin
PROTEIN         Coagulation factor IX
PROTEIN         Coagulation factor XII
```

PMD の PROTEIN フィールド（R931123 エントリ）

MATURATION

自然言語情報：♣♣

```
MATURATION      Cleavage at Arg145-Ala146 by factor XIa [=]: Cleavage at
                  Arg180-Val181 by factor XIa [-]
```

PMD の MATURATION フィールド（A931116 エントリ）

STABILITY

自然言語情報：♣♣

```
STABILITY       Thermal stability in the absence of Fru1,6P2 [- -], and in the
                  presence of Fru1,6P2 [=]: Equilibrium unfolding profile in
                  guanidinium chloride differs with respect to the second
                  transitional change.
STABILITY       Thermal stability in the absence of Fru1,6P2 [- -], and in the
                  presence of Fru1,6P2 [=]
```

PMD の STABILITY フィールド（A931498 エントリ）

STRUCTURE

自然言語情報：♣♣

```
STRUCTURE       Recognition by the anti-CD11b monoclonal antibodies [=]:
                  Heterodimer formation [=]
STRUCTURE       Recognition by the anti-CD11b monoclonal antibodies [=]:
                  Heterodimer formation [=]
STRUCTURE       Recognition by the anti-CD11b monoclonal antibodies [=]:
```

```
Heterodimer formation [=]
```
PMD の STRUCTURE フィールド（A931444 エントリ）

VARIANT

自然言語情報：♣♣

```
VARIANT         Glu-Glu 244-245 Lys-Lys (apoE3, No., apoE7)
```
PMD の VARIANT フィールド（A890043 エントリ）

VARIATION

自然言語情報：♣♣

```
VARIATION       Glu-Leu 286-287 Gly-Lys (BGPa, No., W233)
                Ser 288 (termiantion)   (BGPa, No., W233)
VARIATION       Domain A2 is frameshifted.
VARIATION       Tyr-Asn 382-383 Cys-Lys (BGPa, No., W211)
                Ala 384 (termiantion)   (BGPa, No., W211)
```
PMD の VARIATION フィールド（A910035 エントリ）

CROSS

自然言語情報：♣

```
CROSS-REFERENCE OWHU
CROSS-REFERENCE S03407
CROSS-REFERENCE S03407
```
PMD の CROSS フィールド（R900706 エントリ）

DISEASE

自然言語情報：♣

```
DISEASE         Duchenne muscular dystrophy
DISEASE         Duchenne muscular dystrophy
DISEASE         Duchenne muscular dystrophy
DISEASE         Becker muscular dystrophy
```
PMD の DISEASE フィールド（A931268 エントリ）

ENTRY

自然言語情報：♣

```
ENTRY           R900663 - Variant                       1611073
```

PMD の ENTRY フィールド（R900663 エントリ）

EXPRESSION

自然言語情報：♣

```
EXPRESSION      Intracellular concentration [-]
EXPRESSION      Intracellular concentration [-]
EXPRESSION      Intracellular concentration [-]
EXPRESSION      Intracellular concentration [=]
EXPRESSION      Intracellular concentration [=]
```

PMD の EXPRESSION フィールド（R830050 エントリ）

JOURNAL

自然言語情報：♣

```
JOURNAL         Jikken Igaku (1992) 10(8), 947-953(Jpn)
```

PMD の JOURNAL フィールド（R920650 エントリ）

N

自然言語情報：♣

```
N-TERMINAL      MRKEA
N-TERMINAL      VAGAS
N-TERMINAL
```

PMD の N フィールド（R920752 エントリ）

SOURCE

自然言語情報：♣

```
SOURCE          Human
SOURCE          Sparce-fur mouse
SOURCE          Sparce-fur mouse with abnormal skin and hair
```

PMD の SOURCE フィールド（エントリ）

TRANSFAC データベース

RT

自然言語情報：♣♣♣♣

```
RT  Cloning by recognition site screening of two novel
RT  GT box binding proteins: a familiy of Sp1 related genes
RT  Members of the Sp transcription factor family control
RT  transcription from the uteroglobin promoter
RT  Different members of the Sp1 multigene family exert
RT  opposite transcriptional regulation of the long terminal repeat of HIV-1
```

TRANSFAC の RT フィールド（T02338 エントリ）

FT

自然言語情報：♣♣♣♣

```
FT       67     67   serine, potential PKC phosphorylation site
FT       79     79   serine, PKA phosphorylation site
FT       86     90   DLSSD motif
FT       88    111   repressor domain
FT      102    102   serine, potential CKII phosphorylation site
FT      122    199   glutamine-rich region Q2 (15/78)
FT      230    255   basic region
FT      257    278   leucine zipper (L4)
```

TRANSFAC の FT フィールド（T02361 エントリ）

MM

自然言語情報：♣♣♣

```
MM  southwestern blotting
MM  functional analysis
MM  gel shift competition
MM  direct gel shift
```

TRANSFAC の MM フィールド（R04878 エントリ）

OC

自然言語情報：♣♣♣

```
OC  eukaryota; animalia; metazoa; chordata; vertebrata;
OC  tetrapoda; mammalia; eutheria; rodentia; myomorpha; muridae; murinae
```

TRANSFAC の OC フィールド（T02403 エントリ）

BF

自然言語情報：♣♣♣

```
BF  T00900; WT1 I -KTS;Quality: 2; Species: human, Homo sapiens.
BF  T01839; WT1 -KTS;Quality: 2; Species: human, Homo sapiens.
```

TRANSFAC の BF フィールド（R04865 エントリ）

BS

自然言語情報：♣♣♣

```
BS  R03367; RABBIT$UG_16; Quality: 2; rabbit, Oryctolagus cuniculus
BS  R03373; RABBIT$UG_22; Quality: 2; rabbit, Oryctolagus cuniculus
BS  R04839; HIV1$HIV1_27; Quality: 3; HIV-1, human immunodeficiency virus type 1
BS  R04930; HS$GROA_02; Quality: 3; human, Homo sapiens
```

TRANSFAC の BS フィールド（T02338 エントリ）

CC

自然言語情報：♣♣♣

```
CC  binding affinity of Ni(II)Sp1 to this sequence is
CC  half less than that of native Zn(II)Sp1 [1]
```

TRANSFAC の CC フィールド（R04813 エントリ）

CP

自然言語情報：♣♣♣

```
CP  embryo: mesenchymal cells in cranofacial, pericardial,
CP  primitive dermal, prevertebral, and genital structures
CP  [2], visceral arch, limb bud [1]; adult: heart, skeletal muscle, uterus [1];
```

TRANSFAC の CP フィールド（R04813 エントリ）

DE

自然言語情報：♣♣

```
DE  tPA (tissue-type plasminogen activator); Gene: G001122.
```

TRANSFAC の DE フィールド（R04884 エントリ）

FF

自然言語情報：♣♣♣

```
FF  reduced trans-activation and transformation capability
FF  (1-10%) compared with c-Myc [5], [6];
FF  gene amplification causes small-cell lung carcinoma [8];
FF  expression controlled by transcriptional attenuation [7];
```

TRANSFAC の FF フィールド（T02385 エントリ）

IN

自然言語情報：♣♣♣

```
IN  T00722; Rb; human, Homo sapiens.
IN  T00797; TBP; fruit fly, Drosophila melanogaster.
IN  T00795; TBP; fission yeast, Schizosaccharomyces pombe.
IN  T00794; TBP; human, Homo sapiens.
IN  T00796; TBP; mouse, Mus musculus.
IN  T01412; NF-EM5; mouse, Mus musculus.
```

TRANSFAC の IN フィールド（T02068 エントリ）

OS

自然言語情報：♣♣

```
OS  golden hamster, Mesocricetus auratus
```

TRANSFAC の OS フィールド（T02314 エントリ）

RA

自然言語情報：♣♣♣

```
RA  Kuras L., Cherest H., Surdin-Kerjan Y., Thomas D.
RA  Mountain H. A., Bystroem A. S., Korch C.
RA  Thomas D., Jacquemin I., Surdin-Kerjan Y.
RA  Saiz J. E., Buitrago M. J., Soler-Mira A., del Rey F., Revuelta J. L.
```

TRANSFAC の RA フィールド（T02310 エントリ）

CL

自然言語情報：♣

```
CL  C0012; bHLH-ZIP.
```

TRANSFAC の CL フィールド（T02387 エントリ）

CN

自然言語情報：♣♣

```
CN  brain, kidney, lung, pancreas, spleen, testis [1];
```

TRANSFAC の CN フィールド（エントリ）

DR

自然言語情報：♣

```
DR  EMBL: X03294; HSNMYC2(g).
DR  EMBL: X03295; HSNMYC3A(g).
DR  EMBL: M13241; HSNMYC01(g).
DR  EMBL: M13228; HSNMCY1A(r).
DR  EMBL: X02363; HSNMYC3(g).
DR  EMBL: Y00664; HSNMYC(g).
DR  SwissProt: P04198; MYCN_HUMAN.
DR  PIR: A25744; TVHUM2.
DR  PIR: A01355; TVHUMC.
DR  PIR: A22937; A22937.
DR  PIR: S02249; S02249.
```

TRANSFAC の DR フィールド（T02379 エントリ）

DT

自然言語情報：♣

```
DT  01.12.97 13:05:52 (created); ewi.
DT  01.12.97 (updated); ewi.
```

TRANSFAC の DT フィールド（T02308 エントリ）

FA

自然言語情報：♣

```
FA  PEBP2alphaA/til-1
```

TRANSFAC の FA フィールド（T02318 エントリ）

HO

自然言語情報：♣♣

```
HO  TFIIA-alpha/beta precursor (human), TOA1 (yeast)
```

TRANSFAC の HO フィールド（T02225 エントリ）

RE

自然言語情報：♣

```
RE  3' to P2 promoter
```

TRANSFAC の RE フィールド（R04520 エントリ）

S1

自然言語情報：♣

```
S1  ATG
```

TRANSFAC の S1 フィールド（R03854 エントリ）

SC

自然言語情報：♣

```
SC  translated from EMBL #AF005936
```

TRANSFAC の SC フィールド（T02318 エントリ）

SO

自然言語情報：♣♣

```
SO  0339; rec(human-vaccinia virus-HeLa)
SO  0152; MEL
```

TRANSFAC の SO フィールド（R04934 エントリ）

SY

自然言語情報：♣♣

```
SY  TAF-135; TAF-130; TBP-associated factor 135; TBP-associated factor 130;
```

TRANSFAC の SY フィールド（T02328 エントリ）

AC

自然言語情報：♣

```
AC  T02308
```

TRANSFAC の AC フィールド（T02308 エントリ）

EL

自然言語情報：♣

```
EL  Fp promoter
```

TRANSFAC の EL フィールド（R04933 エントリ）

ID

自然言語情報：♣

```
ID  T02306
```

TRANSFAC の ID フィールド（T02306 エントリ）

MX

自然言語情報：♣

```
MX  M00199; V$AP1_C
MX  M00173; V$AP1_Q2
```

TRANSFAC の MX フィールド（T01115 エントリ）

RN

自然言語情報：♣

```
RN  [1]
RN  [2]
RN  [3]
```

TRANSFAC の RN フィールド（T02410 エントリ）

SF

自然言語情報：♣♣♣♣

```
SF  alternative splice product is L-Myc (long form) lacking
SF  the whole DNA-binding bHLH-ZIP domain <T02385>
SF  the short form only comprises the trans-activation domain of L-Myc [2];
```

TRANSFAC の SF フィールド（T02386 エントリ）

SQ

自然言語情報：♣

```
SQ  LESTDGERLVKSPQCSNPGLCVQPHHIGVSVKELDLYLAYFVHAADSSQSESPSQPSDAD
SQ  IKDQPENGHLGFQDSFVTSGVFSVTELVRVSQTPIAAGTGPNFSLSDLESSSYYSMSPGA
SQ  MRRSLPSTSSTSSTKRLKSVEDEMDSPGEEPFYTGQGRSPGSGSQSSGWHEVEPGMPSPT
SQ  TLKKSEKSGFSSPSPSQTSSLGTAFTQHHRPVITGPRASPHATPSTLHFPTSPIIQQPGP
SQ  YFSHPAIRYHPQETLKEFVQLVCPDAGQQAGQVGFLNPNGSSQGKVHNPFLPTPMLPPPP
SQ  PPPMARPVPLPMPDTKPPTTSTEGGAASPTSPILVPGIKVAASHHPPDRPPDPFSTL
```

TRANSFAC の SQ フィールド（T02299 エントリ）

ST

自然言語情報：♣

```
ST  -133
```

TRANSFAC の ST フィールド（R04890 エントリ）

SZ

自然言語情報：♣

```
SZ  464 AA; 49.6 kDa (cDNA), 62-64 kDa (SDS) [2], 58 KDa, 60 kDa (SDS) [4];
```

TRANSFAC の SZ フィールド（T02379 エントリ）

TY

自然言語情報：♣

```
TY  D
```

TRANSFAC の TY フィールド（R04945 エントリ）

XX

自然言語情報：♣

```
XX
XX
XX
XX
XX
XX
```

TRANSFAC の XX フィールド（T02373 エントリ）

PDB データベース

REMARK

自然言語情報：♣♣♣♣♣

```
REMARK   1                                                          9LYZ  14
REMARK   2                                                          9LYZ  15
REMARK   2 RESOLUTION. 2.5 ANGSTROMS.                               9LYZ  16
REMARK   3                                                          9LYZ  17
REMARK   3 REFINEMENT. NONE.                                        9LYZ  18
REMARK   4                                                          9LYZ  19
REMARK   4 THESE COORDINATES WERE DETERMINED FROM MODEL FITTING TO A    9LYZ  20
REMARK   4 DIFFERENCE ELECTRON DENSITY MAP AT 2.5 ANGSTROMS         9LYZ  21
REMARK   4 RESOLUTION.  THE PHASES FOR THE REFLECTIONS OF NATIVE    9LYZ  22
REMARK   4 LYSOZYME EMPLOYED IN CALCULATION OF THE DIFFERENCE MAP WERE  9LYZ  23
REMARK   4 THOSE FROM THE REFINED TETRAGONAL STRUCTURE WHICH HAD A  9LYZ  24
REMARK   4 CONVENTIONAL R OF 0.23 (REFINED PHASE ANGLES SUPPLIED BY 9LYZ  25
REMARK   4 MS. D. GRACE).                                           9LYZ  26
REMARK   5                                                          9LYZA  3
REMARK   5 CORRECTION. INSERT REVDAT RECORDS. 30-SEP-83.            9LYZA  4
```

PDB の REMARK フィールド（9LYZ エントリ）

フィールド内容：実験の詳細、ただし他のところでは扱っていない情報に限る。

COMPND

自然言語情報：♣♣♣♣

```
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: GLUTATHIONE S-TRANSFERASE P1-1;
COMPND   3 CHAIN: A, B;
COMPND   4 FRAGMENT: TWO INTACT MONOMERS;
COMPND   5 SYNONYM: GSTP1-1;
COMPND   6 EC: 2.5.1.18;
COMPND   7 ENGINEERED: YES;
COMPND   8 BIOLOGICAL_UNIT: DIMER
```

PDB の COMPND フィールド（9GSS エントリ）

フィールド内容：巨大分子の概要

SOURCE

自然言語情報：♣♣♣♣

```
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 ORGANISM_COMMON: HUMAN;
SOURCE   4 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   5 MOL_ID: 2;
SOURCE   6 SYNTHETIC: YES
```

PDB の SOURCE フィールド（9GSS エントリ）

フィールド内容：生物、化学の分子生物のエントリ

JRNL

自然言語情報：♣♣♣♣

v

```
JRNL        AUTH   R.BONE,A.FUJISHIGE,C.A.KETTNER,D.A.AGARD              9LPR   9
JRNL        TITL   STRUCTURAL BASIS FOR BROAD SPECIFICITY IN             9LPR  10
JRNL        TITL 2 ALPHA-LYTIC PROTEASE MUTANTS                          9LPR  11
JRNL        REF    BIOCHEMISTRY                  V.  30 10388 1991      9LPR  12
JRNL        REFN   ASTM BICHAW  US ISSN 0006-2960                 033   9LPR  13
```

PDB の JRNL フィールド（9LPR エントリ）

フィールド内容：文献引用状態

SEQRES

自然言語情報：♣

```
SEQRES   1    104  ALA CYS ASP TYR THR CYS GLY SER ASN CYS TYR SER SER  9RNT  71
SEQRES   2    104  SER ASP VAL SER THR ALA GLN ALA ALA GLY TYR LYS LEU  9RNT  72
SEQRES   3    104  HIS GLU ASP GLY GLU THR VAL GLY SER ASN SER TYR PRO  9RNT  73
SEQRES   4    104  HIS LYS TYR ASN ASN TYR GLU GLY PHE ASP PHE SER VAL  9RNT  74
SEQRES   5    104  SER SER PRO TYR TYR GLU TRP PRO ILE LEU SER SER GLY  9RNT  75
SEQRES   6    104  ASP VAL TYR SER GLY GLY SER PRO GLY ALA ASP ARG VAL  9RNT  76
SEQRES   7    104  VAL PHE ASN GLU ASN ASN GLN LEU ALA GLY VAL ILE THR  9RNT  77
SEQRES   8    104  HIS THR GLY ALA SER GLY ASN ASN PHE VAL GLU CYS THR  9RNT  78
```

PDB の SEQRES フィールド（9RNT エントリ）

フィールド内容：分子の各チェーン、アミノ酸、核酸配列の残基情報

SHEET

自然言語情報：♣

```
SHEET    1   A 5 ILE A 174  CYS A 178  0
SHEET    2   A 5 MET A 191  THR A 196 -1  N  THR A 196   O  ILE A 174
SHEET    3   A 5 ARG A 253  ARG A 258  1  N  ASP A 256   O  MET A 191
SHEET    4   A 5 LYS A 234  CYS A 239 -1  N  CYS A 239   O  ARG A 253
SHEET    5   A 5 ILE A 224  THR A 227 -1  N  ASP A 226   O  VAL A 238
SHEET    1   B 2 PHE A 291  ASN A 294  0
SHEET    2   B 2 THR A 297  PRO A 300 -1  N  ARG A 299   O  THR A 292
```

PDB の SHEET フィールド（9ICQ エントリ）

フィールド内容：分子における螺旋シートの位置識別

KEYWORD

自然言語情報：♣♣♣

```
KEYWDS    DNA-DIRECTED DNA POLYMERASE, DNA REPLICATION, DNA REPAIR,
KEYWDS   2 NUCLEOTIDYLTRANSFERASE,
KEYWDS   3 COMPLEX (NUCLEOTIDYLTRANSFERASE/DNA)
```

PDB の KEYWORD フィールド（9ICQ エントリ）

フィールド内容：エントリ中での関連用語

HELIX

自然言語情報：♣

```
HELIX    1  L1 SER     24  GLY     43  1 DISRUPTION IN THE CENTER       9PAP 112
HELIX    2  L2 GLU     50  ASP     57  1                                9PAP 113
HELIX    3  L3 TYR     67  TYR     78  1                                9PAP 114
HELIX    4  R1 ASN    117  GLN    128  1                                9PAP 115
HELIX    5  R2 GLY    138  LEU    143  1                                9PAP 116
```

PDB の HELIX フィールド（9PAP エントリ）

フィールド内容：分子における螺旋の位置識別

TURN

自然言語情報：♣

```
TURN      1 T1A SER A  11  ASP A  14     TYPE III                       8AT1 411
TURN      2 T2A ASP A 129  ASN A 132     TYPE I                         8AT1 412
TURN      3 T3A PRO A 189  LEU A 192     TYPE I                         8AT1 413
TURN      4 T1C SER C  11  ASP C  14     TYPE III                       8AT1 414
TURN      5 T2C ASP C 129  ASN C 132     TYPE I                         8AT1 415
TURN      6 T3C PRO C 189  LEU C 192     TYPE I                         8AT1 416
```

PDB の TURN フィールド（8AT1 エントリ）

FINOTE

自然言語情報：♣♣♣

```
FTNOTE   1                                                               9INS  82
FTNOTE   1 WATER 24 IS PROBABLY A PARTIALLY OCCUPIED SODIUM ION.         9INS  83
FTNOTE   1 ANOTHER SODIUM SITE WITH VERY LOW OCCUPANCY OVERLAPS THE      9INS  84
FTNOTE   1 MAJOR CONFORMATION OF HIS B10 (SEE REF 3).                    9INS  85
FTNOTE   2                                                               9INS  86
FTNOTE   2 SEE REMARK 5.                                                 9INS  87
```

PDB の FINOTE フィールド（9INS エントリ）

HET

自然言語情報：♣♣♣

```
HET     MG  A 500       1     MAGNESIUM ++                               9RUB 176
HET    CBX  A 191       3     CARBOXYLIC GROUP                           9RUB 177
HET    RUB  A 600      18     RIBULOSE-1,5-BISPHOSPHATE                  9RUB 178
HET     MG  A 500       1     MAGNESIUM ++                               9RUB 179
HET    CBX  B 191       3     CARBOXYLIC GROUP                           9RUB 180
HET    RUB  B 700      18     RIBULOSE-1,5-BISPHOSPHATE                  9RUB 181
```

PDB の HET フィールド（9RUB エントリ）

フィールド内容：特殊な残基の登録

HETNAM

自然言語情報：♣♣♣

```
HETNAM     DTP 2'-DEOXYADENOSINE 5'-TRIPHOSPHATE
HETNAM      MN MANGANESE (II) ION
HETNAM      NA SODIUM ION
```
PDB の HETNAM フィールド（9ICQ エントリ）

フィールド内容：HETID によって与えられた化合物の化学的命名

MODRES

自然言語情報：♣♣

```
MODRES 5GDS CHG I    1  GLY  CYCLOHEXYL GLYCINE
MODRES 5GDS NAL I    3  ALA  NAPHTHYL ALANINE
MODRES 5GDS HAC I   23  ALA  CYCLOHEXYL ALANINE
MODRES 5GDS DGL I   26  GLU  D-GLU
```
PDB の MODRES フィールド（9INS エントリ）

TITLE

自然言語情報：♣♣♣

```
TITLE     DNA POLYMERASE BETA (POL B) (E.C.2.7.7.7) COMPLEXED WITH
TITLE    2 SEVEN BASE PAIRS OF DNA; SOAKED IN THE PRESENCE OF
TITLE    3 OF DATP (0.1 MILLIMOLAR) AND MNCL2 (0.5 MILLIMOLAR)
```
PDB の TITLE フィールド（8ICQ エントリ）

フィールド内容：タイトル

AUTOR

自然言語情報：♣♣

```
AUTHOR    M.BACHELIN,G.HESSLER,G.KURZ,J.G.HACIA,P.B.DERVAN,H.KESSLER
```
PDB の AUTOR フィールド（8DRH エントリ）

フィールド内容：著者

CAVEAT

自然言語情報：♣♣

```
CAVEAT     1BMU    THERE ARE CHIRALITY ERRORS IN C-ALPHA CENTERS
```
PDB の CAVEAT フィールド（1BMU エントリ）

CISPEP

自然言語情報：♣

```
CISPEP   1 GLU     40     PRO      41           0         -0.37
```

PDB の CISPEP フィールド（5TMP エントリ）

EXPDTA

自然言語情報：♣♣

```
EXPDTA    X-RAY DIFFRACTION
```

PDB の EXPDTA フィールド（6UPJ エントリ）

フィールド内容：実験情報

HEADER

自然言語情報：♣♣

```
HEADER    OXIDOREDUCTASE(OXYGENASE)                18-MAY-90   8CPP      8CPP   2
```

PDB の HEADER フィールド（8CPP エントリ）

フィールド内容：登録している名前

HETSYM

自然言語情報：♣♣

```
HETSYN     GLC GLUCOSE
```

PDB の HETSYM フィールド（2BQP エントリ）
フィールド内容：異性体情報

HYDBND

自然言語情報：♣

```
HYDBND       N   GLY A  148                O   PHE B   41
HYDBND       N   GLY A  148                O   PHE B   41
HYDBND       OG  SER A  147                O   HIS B   57
HYDBND       OG  SER A  147                O   HIS B   57
```

PDB の HYDBND フィールド（1AMH エントリ）

JNRL

自然言語情報：♣

```
JNRL        REFN   ASTM EJBCAI  IX ISSN 0014-2956                0262  166DA  3
```

PDB の JNRL フィールド（166D エントリ）

フィールド内容：文献引用状態

LINK

自然言語情報：♣

```
LINK        FE   HEM A 153                NE2 HIS A 101
LINK        FE   HEM A 153                C   CMO A 154
LINK        FE   HEM B 153                NE2 HIS B 101
LINK        FE   HEM B 153                C   CMO B 154
```

PDB の LINK フィールド（5HBI エントリ）

SEQADV

自然言語情報：♣♣

```
SEQADV 4MSI ALA      1  SWS  P19614    ASN     1 CLONING ARTIFACT
SEQADV 4MSI THR     16  SWS  P19614    ALA    16 ENGINEERED
SEQADV 4MSI ALA     64  SWS  P19614    PRO    64 SEE REMARK 999
SEQADV 4MSI ALA     65  SWS  P19614    PRO    65 SEE REMARK 999
```

PDB の SEQADV フィールド（4MSI エントリ）

SIGATM

自然言語情報：♣

```
SIGATM    1  C1  MPR     5       0.013   0.014   0.013  0.00  0.00      1ETL 170
SIGATM    2  O   MPR     5       0.008   0.009   0.008  0.00  0.00      1ETL 174
SIGATM    3  C2  MPR     5       0.016   0.016   0.017  0.00  0.00      1ETL 178
SIGATM    4  C3  MPR     5       0.019   0.016   0.017  0.00  0.00      1ETL 182
SIGATM    5  S3  MPR     5       0.004   0.000   0.004  0.00  0.00      1ETL 186
SIGATM    6 1H2  MPR     5       0.151   0.170   0.155  0.00  0.05      1ETL 190
```

PDB の SIGATM フィールドの一部（1ETL エントリ）

SIGUIJ

自然言語情報：♣

```
SIGUIJ    1  C1  MPR     5        4     15      5     13      8     15  1ETL 172
SIGUIJ    2  O   MPR     5        2      9      4      8      5     10  1ETL 176
SIGUIJ    3  C2  MPR     5        5     18      6     15     10     19  1ETL 180
SIGUIJ    4  C3  MPR     5        6     18      6     17     11     18  1ETL 184
SIGUIJ    5  S3  MPR     5        1      3      2      4      3      4  1ETL 188
```

PDB の SIGUIJ フィールドの一部（1ETL エントリ）

SITE

自然言語情報：♣

```
SITE     1  P1  3 HIS    12  HIS   119  LYS    41
SITE     1  P2  2 LYS     7  ARG    10
SITE     1  B1  2 THR    45  ASP    83
SITE     1  B2  2 ASN    71  GLU   111
```

PDB の SITE フィールド（4RSD エントリ）

SLTBRG

自然言語情報：♣

```
SLTBRG       ND1 HIS B  13                OE2 GLU A  17
SLTBRG       NE2 HIS B  13                OE2 GLU A  10
SLTBRG       CZ  ARG A  23                OE1 GLU B  28
```

PDB の SLTBRG フィールド（1A93 エントリ）

SPRSDE

自然言語情報：♣

```
SPRSDE     28-JAN-98 364D      356D
```

PDB の SPRSDE フィールド（364D エントリ）

SSBOND

自然言語情報：♣

```
SSBOND   1 CYS     26    CYS     84                                      6RSA 153
SSBOND   2 CYS     40    CYS     95                                      6RSA 154
SSBOND   3 CYS     58    CYS    110                                      6RSA 155
SSBOND   4 CYS     65    CYS     72                                      6RSA 156
```

PDB の SSBOND フィールド（6RSA エントリ）

TER

自然言語情報：♣

```
TER    1545      PRO A 208                                               8FAB1728
TER    3181      SER B 223                                               8FAB3364
TER    4726      PRO C 208                                               8FAB4909
TER    6409      LYS D 222                                               8FAB6592
```

PDB の TER フィールド（8FAB エントリ）

TVECT

自然言語情報：♣

```
TVECT        0.00000   0.00000  19.64000                                 2C4S  67
```

PDB の TVECT フィールド（2C4S エントリ）

OMIM データベース

Text

自然言語情報：♣♣♣♣♣

```
Text:
  Kozlowski et al. (1988) reported an Algerian family in which 5 members
  had a seemingly 'new' form of spondylometaphyseal dysplasia. It was
  considered possible that the patient reported by Schmidt et al. (1963)
  had this disorder; see 184250. None of the spondylometaphyseal
  dysplasias show a combination of such severe metaphyseal changes and
  severe genu valgum. Among the metaphyseal dysplasias, only the Jansen
  type (156400) shows severe metaphyseal changes, but in that disorder
  genu varus deformity is found and sclerosis of the skull is a common
  finding in older patients (Holthusen et al., 1975).
```

OMIM の Text フィールド（184253 エントリ）

MINI-MIM

自然言語情報：♣♣♣♣♣

MINI-MIM:
The Langer-Giedion syndrome (LGS) has similarities to the trichorhinophalangeal syndrome type I (TRPS1; 190350; see also 275500), particularly with regard to facies, bulbous nose, sparse hair, and cone-shaped epiphyses. Distinguishing features in LGS are multiple exostoses, mental retardation, microcephaly, and redundant skin. Less consistent features include hyperextensible joints, recurrent upper respiratory tract infections, hearing loss, delayed speech development (Langer et al., 1984), and genitourinary anomalies (Partington et al., 1991). Most cases of LGS are sporadic, but it has been reported as concordant in monozygotic twins and recurrent in a few families (Brenholz et al., 1989). It now appears that the Langer-Giedion syndrome is a 'contiguous gene syndrome' (Hou et al., 1995) due to deletions in chromosome 8q24.1 that result in loss-of-functional copies of the genes for TRPS1 and multiple exostoses (EXT1; 133700).

OMIM の MINI-MIM フィールド（150230 エントリ）

References

自然言語情報：♣♣♣♣♣

v

References:
1. Borochowitz, Z.; Soudry, M.; Mendes, D. G.: Familial recurrent dislocation of patella with autosomal dominant mode of inheritance. Clin. Genet. 33: 1-4, 1988.

2. Carter, C.; Sweetnam, R.: Recurrent dislocation of the patella and of the shoulder: their association with familial joint laxity. J. Bone Joint Surg. 42B: 721-727, 1960.

3. Miller, G. F.: Familial recurrent dislocation of the patella. J. Bone Joint Surg. 60B: 203-204, 1978.

OMIM の References フィールド（169000 エントリ）

Clinical Synopsis

自然言語情報：♣♣♣♣

```
Clinical Synopsis:

  Skin:
     Epidermolysis bullosa involving hands and feet only

  Misc:
     Skin cooling with ice before friction prevents lesions

  Inheritance:
     Autosomal dominant mutation of keratin 5 (KRT5;
     148040) or keratin 14 gene (KRT14;
     148066)
```

OMIM の Clinical Synopsis フィールド（131800 エントリ）

Allelic Variants

自然言語情報：♣♣♣♣♣

```
Allelic Variants:
  .0001
  RHIZOMELIC CHONDRODYSPLASIA PUNCTATA, TYPE 3
  ALKYLDIHYDROXYACETONEPHOSPHATE SYNTHASE DEFICIENCY
  AGPS, ARG419HIS
  In a patient with isolated alkyl-DHAP synthase deficiency (Wanders et
  al., 1994; see 600121), De Vet et al. (1998) detected a G-to-A
  transition at nucleotide 1256 of the alkyl-DHAP synthase gene, resulting
  in an arg419-to-his substitution.
```

OMIM の Allelic Variants フィールド（603051 エントリ）

See Also

自然言語情報：♣♣

```
See Also:
  Matsuda et al. (1979); Sase et al. (1985)
```

OMIM の See Also フィールド（603471 エントリ）

Creation Data

自然言語情報：♣♣

```
Creation Date:
  Victor A. McKusick: 8/23/1999
```

OMIM の Creation Data フィールド（603471 エントリ）

MIM Entry

自然言語情報：♣

```
MIM Entry: 604159
```

OMIM の MIM Entry フィールド（603471 エントリ）

PDBSTR データベース

DEFINITION

自然言語情報：♣♣♣♣

```
DEFINITION  TERNARY STRUCTURE OF HHAI METHYLTRANSFERASE WITH ADOHCY AND
            DNA CONTAINING A G:ABASIC MISMATCH AT THE TARGET BASE PAIR
            MOL_ID: 1;
            MOLECULE: CYTOSINE-SPECIFIC METHYLTRANSFERASE HHAI;
            CHAIN: A;
            EC: 2.1.1.73;
            ENGINEERED: YES;
            BIOLOGICAL_UNIT: MONOMER;
            MOL_ID: 2;
            MOLECULE: DNA;
            CHAIN: C, D;
            BIOLOGICAL_UNIT: DOUBLE STRANDED DNA
```

PDBSTR の DEFINITION フィールド（9MHTA エントリ）

SOURCE

自然言語情報：♣♣♣♣

```
SOURCE      MOL_ID: 1;
            ORGANISM_SCIENTIFIC: HOMO SAPIENS;
            ORGANISM_COMMON: HUMAN;
```

```
EXPRESSION_SYSTEM: ESCHERICHIA COLI;
MOL_ID: 2;
SYNTHETIC: YES
```

PDBSTR の SOURCE フィールド（9ICA エントリ）

DEPOSITOR

自然言語情報：♣♣

```
DEPOSITOR    C.R.DUNN,J.J.HOLBROOK,H.MUIRHEAD
```

PDBSTR の DEPOSITOR フィールド（9MHTA エントリ）

MEMBER

自然言語情報：♣♣

```
MEMBER       9ICYA          335    PROTEIN             96/10/24   96/11/15
```

PDBSTR の MEMBER フィールド（9ICYA エントリ）

FEATURE

自然言語情報：♣♣

FEATURES		FROM	TO	DESCRIPTION
HELIX	1	13	28	A
HELIX	1	33	48	B
HELIX	1	55	61	C
HELIX	1	67	79	D
HELIX	1	83	90	E
HELIX	1	92	102	F
HELIX	1	108	118	G
HELIX	1	122	131	H
HELIX	1	134	147	I
HELIX	1	152	169	J
HELIX	1	179	184	K
HELIX	5	208	220	L
HELIX	5	262	273	M
HELIX	1	276	288	N
HELIX	1	316	323	O
HELIX	5	330	332	P
SHEET	0	174	178	A

```
  SHEET   -1    191      196    A  N  THR A 196   O  ILE A 174
  SHEET    1    253      258    A  N  ASP A 256   O  MET A 191
  SHEET   -1    234      239    A  N  CYS A 239   O  ARG A 253
  SHEET   -1    224      227    A  N  ASP A 226   O  VAL A 238
  SHEET    0    291      294    B
  SHEET   -1    297      300    B  N  ARG A 299   O  THR A 292
```

PDBSTR の FEATURE フィールド（9ICIA エントリ）

RESOLUTION

自然言語情報：♣♣

```
RESOLUTION  RANGE      20.0 - 2.2  ANGSTROMS
```

PDBSTR の RESOLUTION フィールド（9LDBA エントリ）

MODEL

自然言語情報：♣♣

```
RESOLUTION  RANGE      20.0 - 2.2  ANGSTROMS
```

PDBSTR の MODEL フィールド（3EZBA エントリ）

SEQENCE

自然言語情報：♣

```
SEQUENCE
  SVQATREDKF SFGLWTVGWQ ARDAFGDATR TALDPVEAVH KLAEIGAYGI TFHDDDLVPF
  GSDAQTRDGI IAGFKKALDE TGLIVPMVTT NLFTHPVFKD GGFTSNDRSV RRYAIRKVLR
  QMDLGAELGA KTLVLWGGRE GAEYDSAKDV SAALDRYREA LNLLAQYSED RGYGLRFAIE
  PKPNQPRGDI LLPTAGHAIA FVQELERPEL FGINPETGHE QMSNLNFTQG IAQALWHKKL
  FHIDLNGQHG PKFDQDLVFG HGDLLNAFSL VDLLENGPDG APAYDGPRHF DYKPSRTEDY
  DGVWESAKAN IRMYLLLKER AKAFRADPEV QEALAASKVA ELKTPTLNPG EGYAELLADR
  SAFEDYDADA VGAKGFGFVK LNQLAIEHLL GAR
```

PDBSTR の SEQENCE フィールド（3EZBA エントリ）

STRUCTURE

自然言語情報：♣

```
STRUCTURE        X-CA    Y-CA    Z-CA   PHI   PSI    X-CB    Y-CB    Z-CB
  206 GLU  76     0.73   29.61      27.34 -999 124   -0.68   29.03   27.23      0
  207 PRO  77     3.58   27.08      28.01  -43 171    3.76   26.76   29.49      0
  208 GLN  78     3.86   23.69      26.34  -98 110    5.04   23.40   25.41      0
  209 ASN  79     2.02   20.89      28.14  -96  76    0.62   21.08   28.74      0
  210 CYS  80     3.54   18.09      26.10 -106 153    4.67   18.41   25.12      0
  211 SER  81     4.10   14.46      26.89  -65 152    3.56   13.32   26.02      0
  212 ALA  82     7.77   13.79      27.80  -63 112    8.28   13.05   29.05      0
  213 THR  83     9.59   13.45      24.46  -97-999    9.38   14.50   23.37      0
```

PDBSTR の STRUCTURE フィールド（8PCHP エントリ）

SEGMENT

自然言語情報：♣

```
SEGMENT      1 of  3  ( A of A,B,I )
```

PDBSTR の SEGMENT フィールド（9HVPA エントリ）