データベースシステム 114－17
（1998. 1. 20）

マルチエージェントによる異種の分子生物学データベースの統合と検索の実現

三上 知親， 今井 孝， 吉川 孝伸， 粕川 雄也， 松田 秀雄， 中西 通雄， 橋本 昭洋
E-mail: {mikami, takashi, yosikawa, naka, matsuda, hasimoto}@ics.es.osaka-u.ac.jp
大阪大学 基礎工学部 情報工学科
〒560 大阪府豊中市待兼山町 1 番 3 号 TEL/FAX 06-850-6602

本稿では異種の分子生物学データベースを統合，検索する手法を提案する．分子生物学のデータは分野や目的ごとに異なったデータベースに格納される．それらのデータは相互に関連を持っているため，統合的に検索ができることが望ましい．本手法では，これらのデータベースを、ユーザーエージェントとデータベースエージェントという2種類のエージェントを用いて動的に統合し，検索することができる．ユーザーの要求により検索範囲を絞ることで，統合のコストを大幅に減らすことができる．また、分子生物学データを統合することにより，遺伝子情報を様々な角度から比較することができる．

キーワード：分子生物学データベース，統合データベース，エージェント

Integration of Heterogeneous Molecular Biology Databases and Their Retrieval System with Multiagents

Toshichika Mikami, Takashi Imai, Takanobu Yoshikawa, Takeya Kasukawa,
Hideo Matsuda, Michio Nakanishi, Akihiro Hashimoto
Dept. of Informatics and Mathematical Science, Graduate School of Engineering Science, Osaka University,
1-3 Machikaneyama, Toyonaka, Osaka, 560 Japan

In this paper, we propose a method for integrating heterogeneous molecular biology datebases, and for retrieving integrated data. Molecular biology data are distributed among multiple repositories that are constructed for defferent domains and different purposes. Since there are correlation among these data, it is desirable to retrieve from all these databases. In our method, the integration is dynamically carried out by two types of agents; database agent and user agent, which reside at a data repository and at a user, respectively, and retrieval is executed after integration. By limiting the search space with user's query, the cost of integration can be reduced considerably. Integrated data enable us to compare genes from various aspects.

Keywords: molecular biology database, database integration, agent

1 はじめに

近年の目覚しい遺伝子解析技術の発展により，さまざまな機関において分子生物学のデータの解析が行なわれ，その量は急激に増加している．

解析された分子生物学のデータは分野目的ごとに異なったデータベースに格納される．例えばDNA塩基配列データベース，タンパク質アミノ酸配列データベース，タンパク質立体構造データベースなどがあげられる．これらの情報はそれぞれDNA上での遺伝子情報，それが翻訳されてできるタンパク質の情報，そしてそれぞれのタンパク質が持つ特有の立体構造の情報を表している．同一遺伝子の情報であっても，様々な種類に分かれており，その種類ごとに分散してデータベースに格納されているわけである．

生物学においては，遺伝子情報を様々な角度から比較する必要がある．そのため生物学者の間で，分子生物学データベースを統合的に検索したいという要求がある．

しかしながら，分子生物学データベースを統合的に検索するにあたり，以下にあげるような幾つかの障害が存在する．

【1】各データベースは独立して運営されているため，スキーマやインターフェースが異なる．統合する際にはこれらの差異を吸収する必要がある．(分子生物学データベースの詳細については2章に記述する．)

【2】データが複雑である．例えば，DNA塩基配列データベースのGenome Sequence Database (GSDB)[2] はリレーショナルデータベースシステムで扱われており，32 の基底関係が存在し，それぞれ2から36の属性を持つ[3]．

【3】データ量が膨大であり，さらに増加し続けている．そのため，統合にかかるコストが大きい．例えばDNA 塩基配列データベースのGenBank のデータ量は，現在約4.6GB であり，ほぼ1年に2倍の勢いで増加している．[4, 5]

【4】データの更新だけでなく，スキーマの変更も頻繁に起こる．細かい変更も数えると，年に2，3回は変更がある．

【5】データを統合する基準を定めにくい．遺伝子情報を識別するものとして，遺伝子名などがあげられるが，同一の遺伝子であっても違う生物種の中にあれば異なった名前で呼ばれることもあり，一致の基準に用いることはできない．
遺伝子やタンパク質を最も正確に識別できる情報として，DNA 塩基配列またはタンパク質アミノ酸配列があるが，生物は進化するものであり，その過程でDNA やタンパク質は変異していく．進化の過程での変異はなだらかに起こるため，複数の遺伝子が同じであるかどうかの境界線を引くことは非常に難しい．

これらの障害から分かるように，分子生物学データベースを統合するためには，フォーマットの異なる大量のデータを統合しなければならない．

分子生物学データベースを統合するための手法が，現在まで幾つか提案されている．分類すると，

(1) 分散したデータベースを，マルチデータベースシステムを用いて統合する方法(CPL/Kleisli[6])
(2) 新たな単一のデータベースを構築し，データを再編集する方法 (Entrez[7])
(3) データベースのコピーをローカルに持ち，連邦型データベースシステムを用いて統合する方法 (IGD[8])
(4) WWWのハイパーテキストリンクなどを用いて関連するエントリどうしを結びつける方法 (Web DBget [9], SRSWWW [10])

のようになる．

しかし，(1) の手法ではユーザーは検索の対象となるすべてのデータベースのスキーマ，データ構造を知っていなければならない．つまり，各データベースのスキーマの差異を吸収していない (問題【1】【2】)．(2) や (3) の手法では少数のデータを統合することは比較的簡単であるが，データが大量になると，データを変換する際のコストが非常にかかってしまうし，各データベースのスキーマが頻繁に変更されるため，それに対応するためのコストも大きい (問題【3】【4】)．(4) の手法では，統合と言っても，エントリどうしをハイパーテキストリンクで結びつけているだけにすぎず，高度な検索もできない．このように，これらの手法が成功しているとは言い難い．

本稿では2種類のエージェントを用いた分子生物学データベースの統合手法ならびに検索手法を提案する．2種類のエージェントとはすなわち，ユーザー側に置かれるユーザーエージェント (以下UA) と各データベースに1つずつ置かれるデータベースエージェント (以下DBA) である．本システムにでは以下のように前述の問題を解決する．

UA はユーザーからの問い合わせ (以下 query) を受けとる機能と，データを統合する機能を持つ．DBA は各データベースを検索し，データを変換する機能を持つ．ユーザーはシステムで用いられるデータ構造だけを知っていれば良い (問題【1】【2】)．2種類のエージェントが通信を介してやりとりすることで，動的にデータを統合することができ，最新のデータに常にアクセスできる (問題【4】)．また，データを統合する部分を絞り込むことで，統合にかかるコストを，劇的に減少させることができる (問題【3】)．統合に際しては，各データベースにあるアミノ酸配列をもとに，ある程度幅を持たせた一致判定基準を採用している (問題【5】)．

TSIMMIS[11] でも類似した手法が使われているが，本手法では，特に分子生物学のデータを統合することに主眼を置いており，各分子生物学データベース特有の知識を活用している．

2 分子生物学データベース

分子生物学のデータは，計算機の扱いに精通していないユーザーにも利用しやすいように，多くのデータベースでテキストファイルの形式で格納されている．

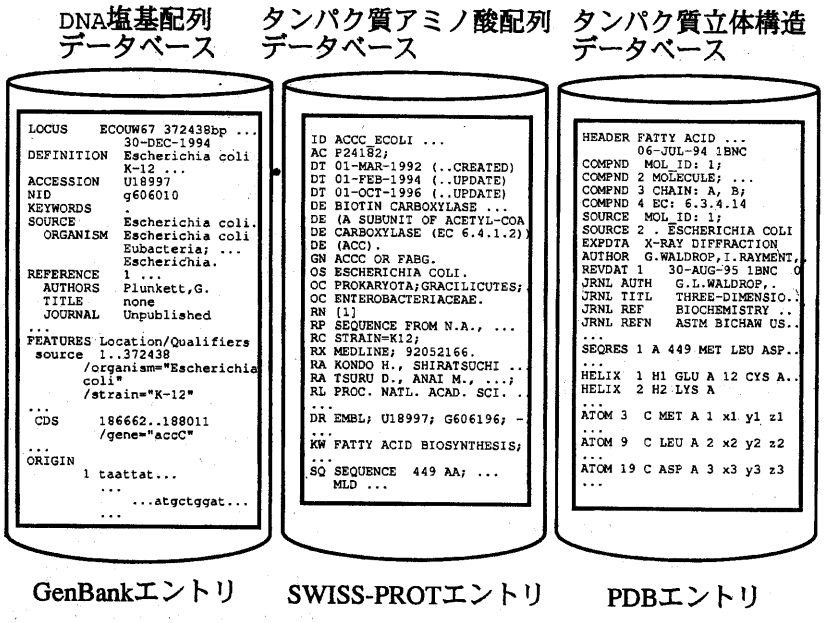

図 1: エントリの例

データは意味のあるまとまりごとに，多くは遺伝子ごとに分けられている．このまとまりはエントリと呼ばれている．

図 1にエントリの例を示す．図 1は左からDNA塩基配列データベースの GenBank，タンパク質アミノ酸配列データベースの SWISS-RPOT，タンパク質立体構造データベースの PDB のエントリの例である．この例では，GenBank のエントリは，大腸菌の accC という遺伝子に関してのデータであり，SWISS-PROT と PDB のエントリは，その遺伝子が翻訳されてできる AccC というタンパク質のデータを表している．同一遺伝子のデータであるにも関わらず，データベースによってまったくエントリの表現形式が異なっていることが分かる．

各エントリは，エントリの識別子を持つ．またさらに違う種類の識別子（アクセッション番号）をもつエントリもある．また関連するエントリの識別子やアクセッション番号をクロスリファレンス情報として持つエントリもある．ユーザーはこのクロスリファレンス情報を用いて，あるエントリから関連するエントリを検索することができる．この例では SWISS-PROT のエントリにある PDB へのクロスリファレンスを用いて，PDB のエントリを検索することができる．

3 統合検索システム

3.1 システム構成

システム全体の構成を図 2に示す．

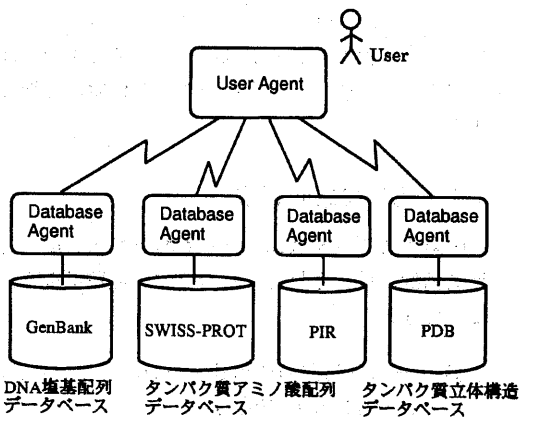

図 2: システム構成

UA はユーザーに置かれ，DBA は各データベースにつき 1 つ置かれる．UA と DBA はネットワークを通じて通信しあう．

3.2 エージェント

各エージェントの機能をまとめると以下のようになる．

- データベースエージェント
  - データベースの検索
  - 検索し抽出されたエントリのフォーマットの変換
- ユーザーエージェント
  - ユーザーからの検索要求の受けとり
  - 問い合わせ言語の解析
  - 得られた結果の統合

検索をする際に，UA がユーザーの query を受けとり，それを各 DBA に送ることで，ユーザーはデータベースの数に関わらず，ただ一度だけ query を UA に与えてやればよくなる．そのため，データが分散していることを意識する必要はなくなる．

データベースごとに DBA を置くことで，データベースのスキーマの変更の際に生じるコストを削減できる．なぜなら，スキーマの変更が起こった場合，そのデータベースの DBA のみを変更すればよく，他の部分は一切変更する必要がないためである．

また，一時的なデータベースの停止や，ネットワークの障害が起こった際にもその部分を切りはなすことで，他のデータベースへの検索は可能である．

3.3 統合検索の過程

検索するデータはユーザーが検索要求を出した時に動的に統合される．統合検索の過程は以下のようになる．

(1) ユーザーは問い合わせ言語 OQL によってシステムに query を出す．
(2) UA がユーザーからの query を受けとる．受けとった query を解析し，各 DBA に query の中のキーワードの部分を受け渡す．
(3) キーワードを受けとった DBA は，そのキーワードを含むようなエントリを各データベースから検索，抽出する．
(4) DBA は得られたエントリを本システムで使われる共通フォーマットに変換しそれを UA に送り返す．
(5) UA は受けとった結果を遺伝子ごとに統合し，統合オブジェクトにする．そしてその結果の中からクロスリファレンス情報を取り出す．それがすでに DBA から送られたエントリの情報であればなにもしないが，まだ送られていないエントリの情報であれば，クロスリファレンス情報を参照されている側のデータベースの DBA に送る．
(6) DBA は送られてきたクロスリファレンス情報によりエントリを抽出し，共通フォーマットに変換し UA に送りかえす．
(7) UA はまた送られてきた結果を統合し，クロスリファレンス情報を調べ，もしまだ送られていないエントリのクロスリファレンス情報があるなら (5) にもどり処理を繰り返す，なければ (8) に進む．
(8) UA は query により統合オブジェクトの比較を行ない，条件に合う結果をユーザーに返す．

データの統合がユーザーが検索要求を出した時に行なわれるので，常に最新のデータを検索することができる．また，統合するデータをキーワードにより絞り込むことで，データベース全体を統合することなく，ユーザーの要求するデータを洩れなく検索することができる．そのため，統合にかかるコストは大幅に削減される．

3.4 共通オブジェクトフォーマット

表 1は DBA が変換する共通フォーマットである．タグがついた単純なオブジェクトモデルを採用している．

データ量が膨大になるため立体構造情報はこの中には含めない．後の検索で必要になった時に，データベースから抽出する．

3.5 遺伝子一致の判定

1 章で触れたように，分子生物学データベースのデータを統合する際に，一致の基準を定めるのは容易ではない．

表 1: 共通オブジェクトフォーマット (部分)

タグ	内容
DB:	元々データがあったデータベース名
ID:	エントリ名
AC:	アクセッション番号
GN:	遺伝子名
SO:	生物種
KW:	キーワード
DR:	クロスリファレンス
PR:	タンパク質アミノ酸配列
SQ:	DNA 塩基配列

本システムでは，タンパク質アミノ酸配列の比較により，同一遺伝子のエントリであるかどうかを判定する．GenBank にはタンパク質アミノ酸配列を持たないエントリがあるが，必ず DNA 塩基配列を持つ．この DNA 塩基配列を翻訳し，タンパク質アミノ酸配列を得ることができる．

1 章で述べたように，同一の遺伝子であっても様々な要因により微妙にタンパク質アミノ酸配列は異なる．そのため，単純な文字列マッチングではなく，ある程度ミスマッチを許した文字列比較を行なう．また，各エントリに格納されたタンパク質アミノ酸配列は文字列の開始位置がずれていることもあるため，文字列の先頭位置を 1 文字ずつ移動させながら，マッチングを行なう．

3.6 エントリに関する問題

一つのデータベースの中の異なるエントリに同一遺伝子が現れる場合がある (図 3) この場合は，すべてのエントリを一致とする．

また，1 つのエントリ内に複数のコード領域 (DNA 塩基配列中のアミノ酸に翻訳される領域) が存在し，複数の遺伝子が現れる．(図 4) この場合は各遺伝子の領域をアミノ酸配列に翻訳し，文字列を比較，対応する領域のみを切り出す．

3.7 統合オブジェクト

同一遺伝子のエントリは統合され，統合オブジェクトとなる．

本システムで用いるスキーマ定義を付録 1 に掲載する．

リレーショナルデータベースでは複雑なデータの構造になるが，統合結果をオブジェクト指向技術を用いて管理することで，実世界のデータ構造に近い表現ができ，検索も直観的に行なえる．

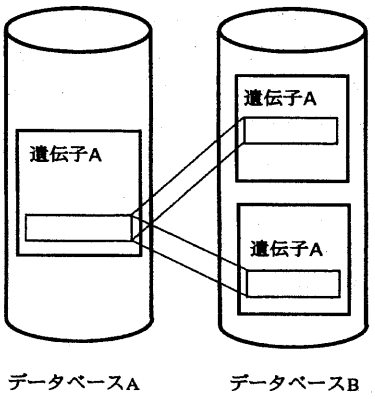

図 3: 遺伝子の重複

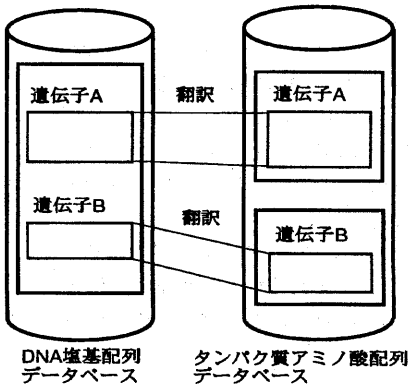

図 4: 複数のコード領域

本システムでは，DNA 塩基配列の類似度，タンパク質アミノ酸配列の類似度，タンパク質立体構造の類似度などを，各クラスのメソッドとして装備することにより様々な角度からデータを検索することが可能である．

3.8 問い合わせ言語

現在，問い合わせ言語 OQL が，オブジェクト指向データベースの問い合わせ言語の標準になりつつある．

本システムでは，OQL を拡張し，分子生物学データベースを検索するために適した，新たな問い合わせ言語を導入した．この言語の定義は付録に掲載する．

この問い合わせ言語を用いた検索条件の例を示す．

「生物Aのタンパク質Bに関係する遺伝子のうち，タンパク質アミノ酸配列が配列Cにx以上の類似度を持つものをすべて検索せよ．」という検索は，

```
DEFINE OBJECT1
SELECT X
  FROM X IN Gene
 WHERE KEYWORDS(A,B)
```

と

```
SELECT Y
  FROM Y IN OBJECT1
 WHERE Y.protein.seq.sim(:seq_C) >= x
```

によって得ることができる．

1つめの問い合わせは，データを統合しオブジェクトにするための定義問い合わせである．WHERE句に書かれたKEYWORD()に，引数としてキーワードの文字列を指定すると，そのキーワードにヒットするエントリが，各データベースから集められ統合される．

2つめの問い合わせは，統合されたオブジェクトどうしでの比較を行なう問い合わせである．あらかじめ変数seq_Cにアミノ酸配列が格納されていたとする．

4 結果と考察

実際にプロトタイプシステムを実装した．本研究では表2に示した4つのデータベースを用いて統合を行なった．プロトタイプシステムはLAN上のワークステーション(SPARCstation-20(SuperSPARC-II, clock 75MHz))を用い，データはそのローカルディスクに置き構築した．

表 2: システムで使用するデータベース

データベース	サイズ	エントリ数
GenBank	4.6GB	1274747
SWISS-PROT	221MB	59024
PIR	211MB	92175
PDB	2.1GB	10804

各エージェントどうしの通信プロトコルはMPI[12]を用いており，インターネット上でもこのシステムは動作する．

データの統合にかかる実行時間は表3のようになった．

表 3: 統合の時間

	KW1	KW2	KW3	KW4
抽出エントリ	19	72	232	351
統合オブジェクト	7	15	65	196
統合時間(秒)	12	58	247	528

キーワード
KW1 : tim Escherichia coli
KW2 : sugar transport periplasmic Escherichia coli
KW3 : serine Echerichia coli
KW4 : transport Haemophilus influenzae

検索条件のキーワードであるが，例えばキーワード2は

"retrieve all genes for periplasmic proteins related to sugar transport expressed in *Escherichia coli.*"

という検索要求を意図しており，現実的な検索の要求を果たしているといえる．

キーワード2によって検索され，実際に統合されたデータは表4のようになる．

遺伝子ごとに統合されたデータを用いて，様々な側面からデータを比較することができる．

図5は表4の統合オブジェクトの中の2つ(araFとrbsB)をDNA塩基配列，タンパク質アミノ酸配列そして2乗平均偏差を用いた立体構造の比較をした類似度を表している．

5 まとめ

エージェントを用いた，異種の分子生物学データベースを統合する手法を提案し，プロトタイプの実装を行なった．

エージェント技術を使うことにより，ユーザーとデータベースの間の，物理的および意味的距離を縮めることができ，また，オブジェクト指向技術を使うことにより，生物学の多様な側面を持った統合データに様々な方法でアクセスすることができる．

今後は，ユーザーインターフェースをより良くしていくことや，遺伝子どうしの比較や解析の機能をより充実させていくことがあげられる．

表 4: 統合結果

Gene	Definition	GenBank	SWISS-PROT	PIR	PDB
rbsB,rbsP, ...	d-ribose-binding periplasmic protein precursor.	ECORBSP, ...	RBSB_ECOLI	1DRJ, ...	1DRJ, ...
xylF,xylT	d-xylose-binding periplasmic protein percursor.	ECOUW76, ...	XYLF_ECOLI		
araG	l-arabinose transport atp-binding protein AraG.		ARAG_ECOLI	S01074	
araH	l-arabinose transport system permease protein AraH.		ARAH_ECOLI	S01075	
mglA	galactoside transport atp-binging protein MglA.		MGLA_ECOLI	B37277	
lamB,malB	maltoporin precursor (lambda receptor protein).	ECOLAMB, ...	LAMB_ECOLI	QRECL	
otsB	trehalose-phosphatase.	OTSB_ECOLI			
otsA	alpha, alpha-trehalose-phosphate synthase.		OTSA_ECOLI		
mglB	d-galactose-binding reriplasmic protein precursor.	ECOMGLB, ...	DGAL_ECOLI	1GLG	1GLG, ...
malF	maltose binding and maltose uptake proteins and the lambda receptor protein.	ECOMALB			
malE178	maltose-binding protein mutant			1MDPA	1MDP
malE322	maltose-binding protein mutant			1MDQ	1MDQ
araF	l-arabinose-binding periplasmic protein precursor.	ECARAFGH	ARAF_ECOLI	JGECA	1APB, ...
malE	maltose-binding periplasmic protein precursor.	ECOUW89	MALE_ECOLI	JGECM, ...	1DMB, ...
malM,molA	maltose operon periplasmic protein precursor.	ECMALM	MALM_ECOLI	BVECMM	

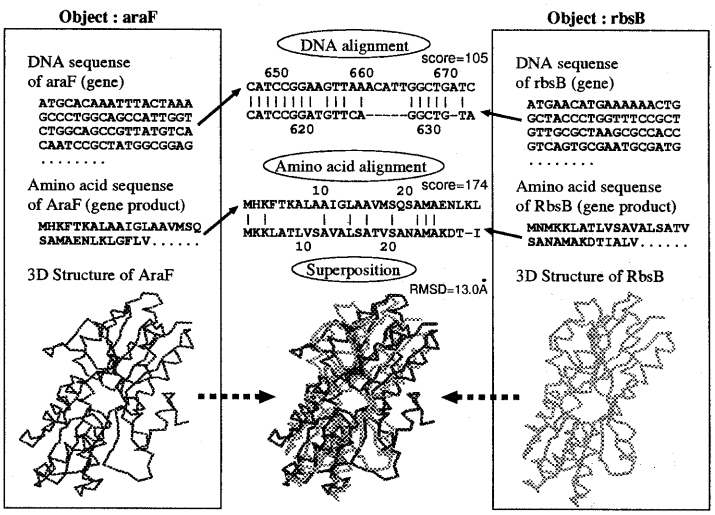

図 5: 統合された遺伝子のデータの比較の例. 遺伝子 araF と rbsB および翻訳されてできるタンパク質 AraF と RbsB

参考文献

[1] Imai, T., Matsuda, H., Sekihara, T., Nakanishi, M., Hashimoto, A.:Implementing an Integrated System for Heterogeneous Molecular Biology Databases with Intelligent Agents, Proceedings of IEEE Pacific Rim Conference on Communications, Computers, and Signal Processing, pp.807-810(1997).

[2] C.Harger,et al.,"The Genome Sequence Database Version 1.0 (GSDB) : from low pass sequences to complete genomes," *Nucleic Acids Research*,Vol.25, No.1,pp.18-23(1997).

[3] Genome Sequence Database Schema (GSDB) Version 1.0, The National Center for Genome Resources.
(Available at
`http://www.ncgr.org/gsdb/schema.html`).

[4] D.Benson, D.J.Lipman, and J.Ostell, "GenBank, " *Nucleic Acids Research*, Vol.24, No.1, pp.1-5, 1996.
(Available at
`http://www.ncbi.nlm.nih.gov/`).

[5] NCBI-Genank Flat File Release 100.0 - Distribution CD-ROM Release Notes, National Center for Biotechnology Information
(Available at
`ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt`).

[6] P.Buneman, S.b. Davidson, K. Hart, C. Overton "A Data Transformation System for Biological Data Sources".

[7] M.S.Boguski. Bioinformatics. *Current Opinion in Genetics and Development*, Vol. 4, pp. 383–388, 1994.

[8] O.Ritter. "The Integrated Genomic Database (IGD)."
Computational Methods In Genome Research, pp. 57–73, 1994.

[9] 秋山泰, 金久實. ゲノムネットのデータベース・サービス. 実験医学, Vol. 13, No. 10, pp. 1201–1205, 1995.

[10] T.Etzold and P.Argos. Transforming a set of biological flat file libraries to a fast access network. *CABIOS*, Vol. 9, pp. 59–64, 1993.

[11] S.Chawathe, H.Garcia-Molina, J.Hammer, K.Ireland, Y.Papakonstantinou, J.Ullman and J.Widon. The TSIMMIS project: integration of heterogeneous information sources. 情報処理学会研究会報告, Vol. 94-DBS-100, pp. 7–18, 1994.

[12] MPI:A Message-Passing Interface Standard 1.1 (Avalable at `http://www.mcs.anl.gov/mpi/index.html`).

A 付録1：スキーマ定義 (部分)

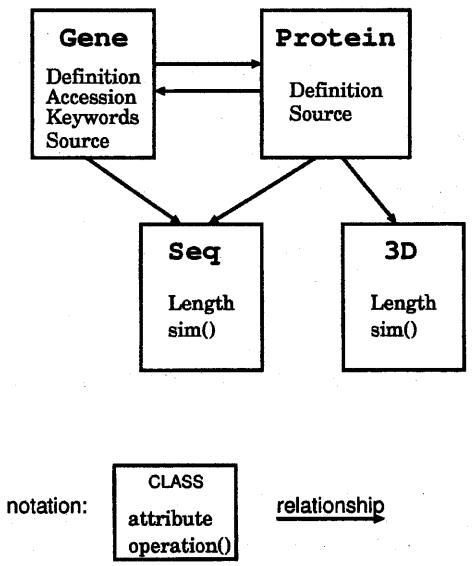

図 6: スキーマ定義

メソッド Sim() は，DNA 塩基配列とタンパク質アミノ酸配列の場合，配列アライメントを行ない，そのスコアを返す．立体構造の場合は2乗平均偏差のスコアを返す．

B 付録2：問い合わせ言語の文法 (部分)

query::= [**DEFINE** identifier **AS**]
SELECT [**DISTINCT**] expr {,expr}
  FROM identifier **IN** expr
  {,identifier **IN** expr }
  WHERE condition

condition ::= **KEYWORDS**(string{,string})
condition ::= compare_expr {**AND** compare_expr}
condition ::= compare_expr {**OR** compare_expr}
compare_expr ::= expr {relation expr}
expr ::= expr dot attribute_name
expr ::= expr dot operation_name(expr{,expr})
expr ::= number
expr ::= string
expr ::= identifier