TM-0942

代謝反応データベース

田中　秀俊

July, 1990

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03)3456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology

代謝反応データベース

田中秀俊
(財) 新世代コンピュータ技術開発機構

代謝反応のデータベースを演繹オブジェクト指向言語で記述し、評価を行なった。これは、分子生物情報の分野のデータベース群を共通の知識表現言語で記述することにより、この分野の統合知識環境を実現するという大目標の一環である。

代謝反応は反応経路のネットワーク表現と、各ノード・アークにおける階層的な性質表現を必要とする。演繹ルールを前者に、複合オブジェクト表現を後者に利用できる点で、演繹オブジェクト指向言語は有効であると考えた。

実験で、反応連鎖はオブジェクト生成規則、物質はオブジェクトの継承関係と属性で記述できることを確認した。同時に考えるべき点も山積したのでそれを列挙し、問題提起とした。

Metabolic Reaction Database

Hidetoshi Tanaka

Institute for New Generation Computer Technolpgy (ICOT)

Mita-Kokusai Bldg. 21F, 4-28, Mita 1-chome, Minato-ku, Tokyo 108, Japan

A metabolic reaction database, written in deductive and object-oriented database language (Juan), is the first step towards an integrated database in molecular biology.

Metabolic reactions can be expressed by reaction networks, complex objects and nested attributes. The Juan is suitable to express the knowledge on reactions, because it has both deductive features which are suitable for networks, and object-oriented features which are suitable for hierarchical objects and attributes.

I will proceed with this work so that Juan will be evaluated further and will improve, as well as an integrated molecular biological database in Juan will be available.

1 はじめに

近年、分子生物情報の分野が情報処理の応用として注目され始めている [DCG] [PFDM] 。分子生物学で処理する情報は、遺伝子の核酸塩基配列やタンパク質のアミノ酸配列、立体構造、酵素機能など多岐にわたり、しかもいずれもデータ量が膨大であり、データベースはこの分野の研究にとって必須のものである。さらに最近は配列読みとりの技術の向上から遺伝子配列データベースの規模が年間倍増という爆発的な増大をみせ、この傾向は30億塩基といわれているヒトの全ゲノムの核酸塩基配列を解読する計画(ヒトゲノムプロジェクト)によりさらに加速されると見られている。

分子生物情報のデータベースは、現在はフラットなファイルをやりとりするのが主流である。ようやく一部では関係データベースへの移行が始まっている。しかし核酸やアミノ酸の配列データを蓄えるだけであれば関係データベースでもよいが、例えばその配列の特徴の記述を活用することを考えただけでも、関係データベースにさらに知識を扱えるような枠組を加えた、新しいデータベース技術が必要になる。

また、分子生物情報データベースは数多くの個別データベースからなっている。これによってデータの変換の手間、プログラムとのインターフェースの悪さ、それぞれのデータベースで用語・名称が異なるなどの弊害を生じており、解消の努力としてスキーマの統合や生物種を限った形でのデータベースの統合が試みられている。

ICOT では知識情報処理の観点から遺伝子情報処理を捉えている [FGC]。データベース技術に関しては、PSI 上に非正規関係モデルに基づく DBMS (Kappa) を開発し、さらに演繹オブジェクト指向の概念に基づく知識表現言語 (Juan) を試作中である [Juan]。今回は分子生物学の知識をこの言語で記述する実験を行ない、言語を評価すると同時に、分子生物情報データベースの統合的な利用環境の形態を考えたい。

2 分子生物情報データベース

2.1 遺伝子配列データベース

遺伝子配列データベースには GenBank (米・ロスアラモス国立研究所) EMBL (独・欧州分子生物学研究所) DDBJ (日・国立遺伝学研究所) の3つがあり、3者間での役割分担 (米、欧、日それぞれ自分の担当域で出された論文のチェックとデータ入力) およびデータ交換に関して既に協力体制が出来ている。

ただし、これらのデータベースはまだ文字列フラットファイルの形式での配布が主流である。年4回のデータのバージョンアップ時には全データを CD-ROM を含むオフラインのメディアを中心に配布する。なお、ロスアラモス国立研究所では昨年関係 DBMS Sybase を利用し始め、GenBank については RDBMS の利用を前提にした情報交換体制を運用し始めている [GBRDB]。フラットファイルへのアクセスにはこの遺伝子配列、タンパク質配列、立体構造の統合的な利用を可能にする IDEAS と呼ばれるソフトが広く使われている [IDEAS] 。

遺伝子配列データベースに関して ICOT では、PSI 上の DBMS Kappa に GenBank を試験的なスキーマで格納し、格納効率や利用形態の調査を進めている [Kappa]。

2.2 アミノ酸配列データベース

アミノ酸配列データベースに関しては国立バイオメディカル研究基金 (米・NBRF) が収集を呼びかけている PIR (Protein Identification Resource) が代表的である。こちらも主流はフラットファイルによる毎回全データ配布である。アクセスに関しては VMS のソフトが配布される。前述の IDEAS からも利用可能である。

2.3 分子生物情報データベース

分子生物情報データベースの統合に関してはCODATAがデータ交換の標準形式(統一スキーマ)の検討、提案を行なっている[CODATA]。また、生物種を限った上で核酸配列、遺伝子地図、タンパク質などのデータを統合的に扱えるようにしたデータベースが、既に大腸菌と酵母で作成されている[JIPID]。また、IDEASもある意味で統合環境を実現している。IDEASを使えばGenBank,PIR,PDBなどを同時に簡単に参照できる。

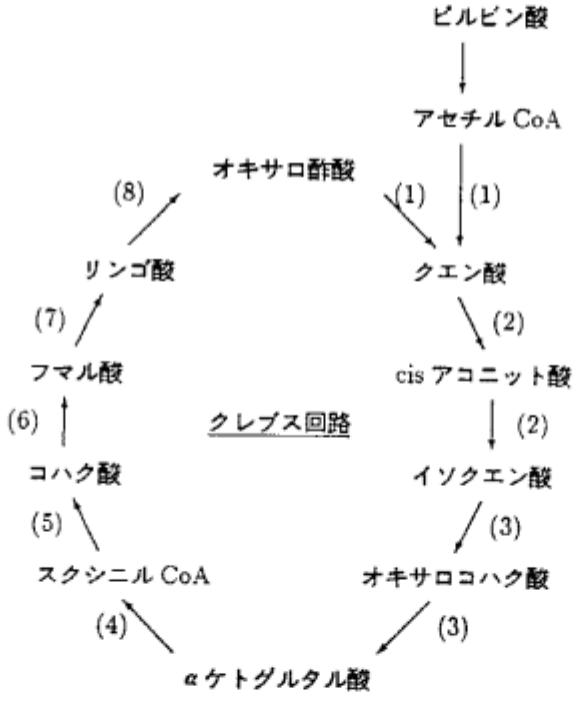
ICOTでは知識表現の観点から分子生物情報を見、表現の統一から自然に導かれる形での統合データベースの実現を目標にしている。既に遺伝子配列情報に関しては非正規関係モデルで表現を試みた[Kappa]。現在はタンパク質配列情報との非正規関係モデル上での融合と、利用形態の調査とを行なっている。今回報告するのは、流れとしてはその先にあたる、生化学の知識の表現方式の考察と記述実験についてである。

3 代謝反応データベース

3.1 代謝反応

今回記述実験の対象として選んだのはクレブス回路というアミノ酸、脂肪酸、炭水化物の共通の最終的代謝経路である。

クレブス回路の簡単な図を示しておく。[Krebs] [図1]



(1)	クエン酸合成酵素	— 縮合
(2)	アコニターゼ	— 脱水と水の付加
(3)	イソクエン酸デヒドロゲナーゼ	— 酸化と脱炭酸
(4)	αケトグルタル酸デヒドロゲナーゼ	— 酸化と脱炭酸
(5)	スクシニル CoA 合成酵素	— 基質レベルのリン酸化
(6)	コハク酸デヒドロゲナーゼ	— 酸化
(7)	フマラーゼ	— 水の付加
(8)	リンゴ酸デヒドロゲナーゼ	— 酸化

図1 クレブス回路(概観)

テーマとして代謝反応を選んだのは以下のような理由からである。

(1) データ構造の特異性

代謝反応は基本的にネットワークで表現できる。ただし反応にはループが存在し、そのループに対する入出力という概念がある。例えばクレブス回路というループの入力をピルビン酸と見た場合、出力は CO_2 と H_2O + エネルギーである。ループの構成物という概念ももちろんあって、同じクレブス回路を例にとればそれはアセチル CoA → クエン酸 →cis アコニット酸 ... という連続した化合物の列である。ループの例に限らなくても、一般に一連の反応に名前をつけて階層化 (ハイパーグラフ化) することはよく行なわれる。よって代謝反応を表現するにはこれを記述できる必要がある。

また、このネットワークは各生物ごとに異なるので、生物種の数だけネットワークが存在することになる。

さらに、同じ名前の化合物でも生物によって組成が異なる場合と同じ場合とがある。同じ名前である以上普通は機能が共通するので、何らかの形でまとめて表したい。すなわち、複数のネットワークのノード間にも関連性を表現する必要がある。

以上のような特異な構造を持つ知識をどのように表現できるか、また容易に表現するにはどのような条件を満たせば良いのか、を調べたいと考えた。

(2) 統合化の可能性の検証

遺伝子の配列データベースは、非正規関係モデルで効率良く表せた [Kappa]。また、ここでは詳しくは触れないが、DOOD 言語である F-Logic [F-Logic] で実験的に記述したところ、表現上の問題は少なかった。タンパク質立体構造に関してもオブジェクト指向言語 P/FDM で記述実験が既に報告されている [PFDM]。

分子生物関係の共通の知識表現手段を考えようとする時、もっとも条件の厳しいものは代謝反応になると考えている。逆に、これを問題なく表現できるなら、他の配列情報や立体構造は簡単に表すことができると考えられる。

3.2 DOOD の特徴との対応

知識表現言語としては ICOT で設計中の DOOD 言語 Juan を用い、その記述実験を行なった。DOOD を選んだのは、代謝反応ネットワークの物質 (ノード) と反応 (アーク) を同じ言語で表現するためである。DOOD 言語であれば、アーク部分には演繹的な性質が利用でき、ノードの部分には複合オブジェクト的な考え方が利用できる。

Juan は他にも以下のような特徴を持つ。

(1) 反応経路情報の表現

反応に関して質問をする場合、その反応の途中の条件や中間的に産生される物質が何か、といった経路上の情報も重要である。反応を記述するデータベースでは、入手が容易な化合物から必要な化合物 (薬品) を作る方法を問い合わせる手段だけでなく、その反応がその目標物まで連続し、そこできちんと停止する条件を得る手段を提供する必要がある。

よって反応の知識表現には、演繹的な性質に加えて経路情報を蓄える性質が必要となる。これに関しては DOOD 言語のひとつ、[F-Logic] でリストの導入例が挙げられており、これは Juan でも記述可能である。

(2) 部分情報の表現

Juanでは全てのファクトは部分情報である。これは分子生物学のような発展途上の学問分野の知識の実態に即している。つまり、新たな知識が次々に加わることを前提にした表現が可能。

(3) オブジェクトIDの表現

JuanではオブジェクトID内に属性を持つことが出来る。オブジェクトIDの名前空間が広く、内包的なオブジェクトの生成時にも意味のある、従って後で検索のしやすいネーミングをとれる。

4 記述実験

4.1 代謝反応の記述

以下は代謝反応ネットワークの記述例である。全て載せるのは繁雑なため、データに関しては図1の(1)および(2)の反応の部分だけを示してある。シンタクス、セマンティクスに関しては[Juan]を同時に参照されたい。

Juanで知識を表現するにあたって、その記述の指針として[F-Logic]を参考にし、束構造、データ(ファクト)、スキーマ、演繹ルールの4項目で表現した。

(1) 束構造

オブジェクトのis-a階層を宣言する。シンタクスとその意味は次の通り。
A : B -- A is a B. (A,Bはオブジェクト ID。)

(1.1) 環境：化合物

(1.2) タンパク質：化合物

(1.3) RNA：化合物

(1.4) 酵素：{タンパク質, RNA}

(2) ファクト

データに相当する部分である。物質に関するデータ(ノード情報)と反応に関するデータ(アーク情報)に分割して記述した。物質に関するデータには属性をほとんどつけていない。実際はさらに組成、立体構造、分子量、電荷といった属性が加わる。反応に関しては図1の(1)が(2.2.5)のK1に、(2)は(2.2.6)のK2と(2.2.7)のK3に対応する。これ以後の分は省略した。

シンタクスとその意味は次の通り。

A / [B = C]	— A の属性 B の値は C である
A / [B →C]	— A の属性 B の値のインスタンスには C もある
A / [B →{ C, D }]	— A の属性 B の値のインスタンスには C も D もある
A / [B →C]	— A の属性 B の値は C のインスタンス
A : D /[...]	— A is a D かつ A の属性記述は [...] の中

[1] ノード情報

酵素の属性として、どのような反応に関わるものかを記述するために、値として反応のオブジェクトIDをとる「反応名」を用意した。

(2.1.1) 水：環境 / [名称 = "水"]

(2.1.2) 酸素：環境 / [名称 = “酸素”]

(2.1.3) 二酸化炭素：環境 / [名称 = “二酸化炭素”]

(2.1.4) ATP：環境 / [名称 = “アデノシン 3 リン酸”]

(2.1.5) ADP：環境 / [名称 = “アデノシン 2 リン酸”]

(2.1.6) オキサロ酢酸：化合物 / [名称 = “オキサロ酢酸”]

(2.1.7) アセチル Co-A：化合物 / [名称 = “アセチル Co-A”]

(2.1.8) クエン酸：化合物 / [名称 = “クエン酸”]

(2.1.9) cis- アコニット酸：化合物 / [名称 = “cis- アコニット酸”]

(2.1.10) イソクエン酸：化合物 / [名称 = “イソクエン酸”]

(2.1.11) クエン酸合成酵素：酵素 / [名称 = “クエン酸合成酵素”, 反応名 ←{ K1 }]

(2.1.12) アコニターゼ：酵素 / [名称 = “アコニターゼ”, 反応名 ←{ K2, K3 }]

[2] アーク情報

(2.2.1) 縮合：反応種類 / [名称 = “縮合”]

(2.2.2) 脱水：反応種類 / [名称 = “脱水”]

(2.2.3) 加水：反応種類 / [名称 = “水の付加”]

(2.2.4) クレブス回路：反応 / [原料 ←{ ピルビン酸, 酸素, ADP },
産物 ←{ 二酸化炭素, 水, ATP },
場所 ←{ ミトコンドリア内 }, 要素反応 ←{ K1, K2, K3, ...}]

(2.2.5) K1：縮合, 反応 / [原料 ←{ オキサロ酢酸, アセチル Co-A }, 産物 ←{ クエン酸 },
使用酵素 ←{ クエン酸合成酵素 }]

(2.2.6) K2：脱水, 反応 / [原料 ←{ クエン酸 }, 産物 ←{ cis- アコニット酸, 水 },
使用酵素 ←{ アコニターゼ }]

(2.2.7) K3：加水, 反応 / [原料 ←{ cis- アコニット酸, 水 }, 産物 ←{ イソクエン酸 },
使用酵素 ←{ アコニターゼ }]

(3) スキーマ

データのスキーマを記述した。シンタクスとその意味は次の通り。

A / [B →C]　　— A の 属性 B の値は C のインスタンス

A / [B →{ C }]　　A の 属性 B の値は C のインスタンスの集合

(3.1) 化合物 / [名称 →string]

(3.2) 反応 / [原料 →{ 化合物 }, 産物 →{ 化合物 }, 使用酵素 →{ 酵素 }]

(3.3) 反応種類 / [名称 →string]

(3.4) 酵素 / [反応名 →{ 反応 }]

(4) 経路演繹ルール

経路を求めるルールを記述した。[1] は化合物 A から化合物 B が反応によって得られるかどうかの判定、[2] はそれに加えて経路情報を求めるものとした。

シンタクスとその意味は次の通り。

A [B = C] / [D = E]　— オブジェクト ID / 属性記述

A ⇐B, C　　オブジェクト B, C がある場合オブジェクト A を生成する。

[1] 反応可能性判定のみ

(4.1.1) 反応関係 [反応起点 = X, 反応終点 = Y] ⇐
W : 反応 / [原料 ←X : 化合物, 産物 ←Z : 化合物],
反応関係 [反応起点 = Z, 反応終点 = Y : 化合物]

(4.1.2) 反応関係 [反応起点 = X : 化合物, 反応終点 = X]

[2] 反応経路 (化合物) を調べる

(4.2.1) 反応関係 [反応起点 = X, 反応終点 = Y]
/ [経路 ←List [先頭 →Z, 残り →OldList] : 反応連鎖] ⇐
W : 反応 / [原料 ←X : 化合物, 産物 ←Z : 化合物],
反応関係 [反応起点 = Z, 反応終点 = Y : 化合物]
/ [経路 ←OldList : 反応連鎖]

(4.2.2) 反応関係 [反応起点 = X : 化合物, 反応終点 = X] / [経路 →nil]

4.2 表現方法の考察

Juan は表現力の強い言語である。このため、一つの知識の表現方法が複数存在するケースが多い。以下は表現方法を統一するための指針として考えたものである。

(1) 反応連鎖の表現について

反応連鎖の記述には 2 通りの方法が考えられる。この 2 通りは併用することとしたい。

- 上位オブジェクトで要素反応として書く。**(2.2.4)**
- 各ノードを経路演繹ルールでつなぐ。**(4.2.1),(4.2.2)**

前者だけでは全てを表現できないが、反応の表現にも階層的な部分は多く、この記述で見通しがよくなる。ただし、上位オブジェクトには順序を入れずに集合として記述し、順序の情報は下位オブジェクトを用いて演繹により求める方が良い。これは反応の枝分かれをこれ以外の方法では簡単に記述できないことによる。また、連鎖を表すために上位オブジェクトを設ける手法も考えられるが、現段階では生物学でコンセンサスのとれているオブジェクトに記述を限るべきだと考えた。

(2) オブジェクト ID の使用方法

Juan ではオブジェクト ID 内の構成要素として属性を持つことができる。ただし、この属性とその値の関係は集合的に一致という条件が付いている。(シンタクスで言うと、= しか使えない。) よって、この属性つきオブジェクト ID の使用方法はおのずと限定される。

今回の実験では以下のような指針で記述を行なった。

- ファクトではオブジェクト ID には属性をつけない。
- 内包的なオブジェクトには生成時に属性つきのオブジェクトを与える。

これは特に強い根拠があるわけではない。ファクトでは属性に — を用いるケースが多く、それと条件とが相容れなかったことと、内包的なオブジェクトはルールやファクトの変更で削除・更新されるべきものであるから、何か区別が必要だと感じたことが主な理由である。

(3) element_of の記述

Juan では element_of の関係を 2 通りに記述できる。一つは上位オブジェクトの属性として下位オブジェクトの集合を持つ形、いま一つはそれとちょうど逆の形である。今回は前者に記述を統一した。なお、あるオブジェクト A の element_of の意味での上位オブジェクトの集合は以下のように簡単に記述できる。

(2.2.8) X : 反応 / [要素反応 = A]

4.3 今後の課題

今回以下の部分は記述を省略した。今後の課題である。

- 無限木チェックの経路探索における利用方法

  Juan にはオブジェクト ID の属性部分に出現する無限ループを検出する仕組みが設けられる予定になっている。経路上にループがある場合、そのループを何回も回る解はこれをうまく利用すれば簡単に排除できるはずである。

- 階層を考えた経路探索方法

  ネットワークが階層化しているため、階層をまたがるような変なループを検索しないように、ルールの記述を工夫する必要がある。おそらく経路探索では全解探索ではなく必ず最上位階層を回答するようにルールを書く形が実現できれば最も自然と思われる。

- 定量情報の算出ルールの記述

  各反応の属性として、反応開始条件、停止条件といった定量的な知識を持たせる必要がある。定量的な質問に対して、計算ルールを適用して回答が出せるようにしたい。

- ベクトル、タプルの導入について

  今回は出なかったが、座標などはひとつのオブジェクトと考えてそこに名前 (Object-ID) をつけるべきか、複数の数字の組 = オブジェクトの組として別に考えるかを決めなければならない。

- 世界の記述の利用 (GenBank,PIR との interaction)

  Juan の「世界」の概念をまだ利用していない。おそらく代謝反応、PDB、GenBank、PIR などをそれぞれ Juan における「世界」として全体を構成するという利用の仕方が自然と思われる。

5 おわりに

今回の実験において、代謝反応の反応連鎖はオブジェクト生成規則 (演繹ルール)、物質はオブジェクトのクラス・インスタンス関係と属性で記述できることを確認した。今後は列挙した課題に順次取り組んでいくとともに、仕様上の要求を Juan の方にも反映させていくことも考えている。

謝辞

最後に本研究に関して多くの示唆を頂いた横田一正氏(ICOT)と、DOL会議、遺伝子情報WG、分子生物情報メーリングリストの方々に感謝致します。

参考文献

[DCG] Searls, D.B.: "Investigating the Linguistics of DNA with Definite Clause Grammars", NACLP 89, pp.189-208 (Oct 1989).

[PFDM] Gray, P.M.D. et al.: "An Object-oriented database for protein structure analysis", Protein Engineering, vol.3 no.4, pp.235-243 (1990).

[FGC] Uchida, S. and Yoshida, K.: "The Fifth Generation Computer Technology and Biological Sequencing", Proceedings of Workshop on Advanced Computer Technologies and Biological Sequencing (Nov 1988).

[Juan] Yokota, K.: "The Outline of a Deductive and Object-Oriented Language: Juan", 第78回データベースシステム研究会 (Jul 1990).

[GBRDB] Cinkosky, M.J. et al.:"GenBank/HGIR Technical Manual", LA-UR 88-3038 (LANL), (1988).

[IDEAS] Kanehisa, M.: "IDEAS User Manual", (1986).

[Kappa] Yokota, K. and Tanaka, H.: "GenBank in Nested Relation", Joint Japanese-American Workshop on Future Trends in Logic Programming 1989 (Oct 1989).

[CODATA] 沖林 他: 「蛋白質の属性データベース」, 情報学シンポジウム 1990, pp.73-77 (Jan 1990).

[JIPID] PIR-International (JIPID): PIR Newsletter no.3 (June 1990).

[Krebs] Watson, J.D. et al.: "Molecular Biology of the Gene (4th edition)", 「遺伝子の分子生物学」, 松原他訳, トッパン (1988).

[F-Logic] Kifer, M., Lausen, G. and Wu, J.: "Logical Foundations for Object-Oriented and Frame-Based Languages", Draft (1990).