演繹オブジェクト指向データベース言語を用いた遺伝子知識ベースの記述 *

5C－1

広沢 誠 †　田中 令子　石川 幹人
（株）日立製作所 ‡　IMS　ICOT

1 はじめに

生物の遺伝子を解析することは、生命現象を解き明かすために必須の技術である。現在、バイオテクノロジーの進歩により、すでに検出され、解析されるべき遺伝子は急激に増加している。この大量の遺伝子を自動的に解析し、生物学的情報を抽出するためには、高速の解析アルゴリズムと共に、知識処理の導入が必要である。しかし、現在、遺伝子関係のデータは、知識処理に利用することを念頭に構築されていないため、高度の知識処理を行なうシステムを作ることは困難である。

我々は、生物学的知識を表現する方式について研究してきた。今回は、遺伝子のモチーフという情報を発見するという課題を念頭において知識表現の研究をすることにした。我々は、既にモチーフ発見システムを開発している [Hirosawa *et al.*1993]。この研究では、暫定的な知識表現を用いていたので、今回は、知識ベースを演繹オブジェクト指向データベースの言語である *QUIXOTE* [Yasukawa *et al.*1992] を用いて作成し、この知識ベースを基にシステムを再構築した。この結果、システムの性能が向上した。今回は、オブジェクト指向的な表現により生物学的概念の適切な表現が可能になったことを示す。

2 遺伝子の解析とは

生物が持つ遺伝子の解析は、生物学の分野だけではなく、医学では癌のメカニズムの解析にも必須である。遺伝子解析とは、新たな遺伝子が実験により同定された時、この遺伝子が生体内で担っている機能を推定することである。生物の遺伝子は、細胞の核の中にDNAとしてコードされ格納されている。この DNA は必要に応じ mRNA に転写され、tRNA の等の働きによりアミノ酸の鎖に変換される。このアミノ酸の鎖は、適切な形に成形され蛋白質として機能するようになる。

蛋白質の鎖のユニットとして使用されるアミノ酸は２０種類ある。各々のアミノ酸を一つのアルファベットを用いて表わすと、蛋白質をアルファベットの配列で表現することができる。例えば、“GIVEQCCTSICSLYQL” は、インシュリンの一部を示したアミノ酸の配列である。ここで、G はグリシンというアミノ酸である。このような表現方法をとると、遺伝子の解析とは、アルファベットの配列として表現されている遺伝子のアミノ酸の配列から、この遺伝子がコードしている蛋白質の機能を推測することとなる。

遺伝子の解析のために必要な情報としてモチーフがある。モチーフとは、ある種類の蛋白質が共通に含むアミノ酸の配列パターンである。モチーフの中には、それに対応する蛋白質の部位が果たす機能（ex. DNA に結合する）が判明しているものもある。このように、モチーフは蛋白質の機能を予測するのに重要な情報である。

3 遺伝子の解析システム

我々は、すでに発見されている遺伝子とそれに対応する蛋白質が持つモチーフが判明している時、これを事例として、入力された蛋白質の持つモチーフを発見するシステムを開発した。これは、入力された蛋白質の機能も推測する。システムの構成を第１図に示す。以下、その構成と動作を簡単に説明する。

まず、*Motif Finder* は、Aligner が入力配列に対して行なった 類似性解析結果を解析し、配列の共通部分を検出する。*Motif Generator* は、この解析結果を基に、モチーフの候補を可能性の高いものから順に生成していく。可能性の高いものというものは Biological Knowledge Base に含まれるモチーフの情報に基づくいて生成したモチーフ候補である。

具体的には、Motif Generator は、*Motif Rule Base* に含まれる ルールを、優先度にしたがって実行し、モチーフ候補を生成する。（詳しい記述は、[Hirosawa *et al.*1993] にある）。ルールは、必要であれば *Biological Knowledge Base* の *Prosite** と *User* を参照する。これらの知識は、演繹オブジェクト指向データベース言語である *QUIXOTE* により記述されている。*Motif Tester* は、生成されたモチーフが生物学的統計的基準を満たしているかを調べる。Motif Finder でモチーフとして容認された場合には、これを Biological Knowledge Base の *Discovery* に登録される。

4 Biological Knowledge Base

Biological Knowlege Base は、３つのサブデータベースに分かれている。各々は,Prosite*,User, Discovery である。Prosite* には、既存のモチーフデータベースとして代表的な Prosite[Bairoch 1991] に含まれているモチーフなどが階層的に表現されている。Prosite* におけるクラス分けは、Prosite で採用しているものを基本にしているが、これでは十分ではないため、生物学的な知識にしたがい再分類を行なっている（以前の分類方法に復元することも可能である）。

User は、ユーザーが文献などから得たモチーフを登録する場所である。また、Discovery は、システムが Prosite* や User に登録されている知識を利用して発見したモチーフを登録する場所である。User や Discovery が用いているクラス構造は Prosite* で用いているものと同じものである。

Figure 2 に Prosite* の一部を視覚化したものを示す。図では、キナーゼというリン酸基を転移する蛋白質 (kinaseGroup) は、４つのサブグループに再分類され、その１つがプロテイン・キナーゼ (protein-kinase) を含んでいる。それは、さらに、チロシン・キナーゼ (ty-kinase) とセリン／スレオニン・キナーゼ (s-th-kinase) とに分類されている。

各階層には、対応する蛋白質が持つモチーフが登録されている。プロテイン・キナーゼは、protein-kinase-general というモチーフのエントリーを持っており、

*Description of Genetic Knowledge Base using Deductive Object-Oriented Database Language

†Makoto Hirosawa *et al.*

‡Hitachi System Development Lab. 1099 Ohzenji Asao Kawasaki 215

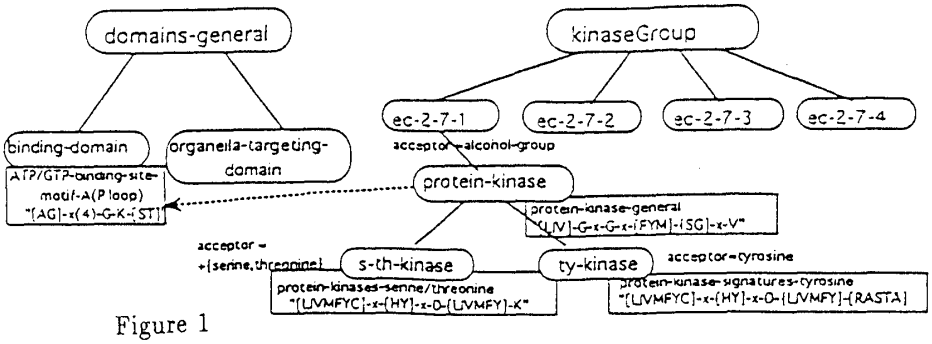

Figure 1

```
kinaseGroup >= ec-2-7-1 ;;
kinaseGroup >= ec-2-7-2 ;;
kinaseGroup >= ec-2-7-3 ;;
kinaseGroup >= ec-2-7-4 ;;
    ec-2-7-1 >= protein-kinase ;;
protein-kinase >= s-th-kinase;;
        protein-kinase >= ty-kinase;;
```

Figure 3

```
kinaseGroup::protein-kinase[name = "Protein kinases general"]/
      [pattern = "[LIV]-G-x-G-x-[FYM]-[SG]-x-V",
       source   = prosite*
       otherMotif = domains_general:binding_domain
         [name = "ATP/GTP-binding site motif A (P-loop)"]    ];;
kinaseGroup::protein-kinase[name = "Protein kinases ATP-bind"]/
            [pattern = "A-x-K",
             source   = user      ];;
```

Figure 4

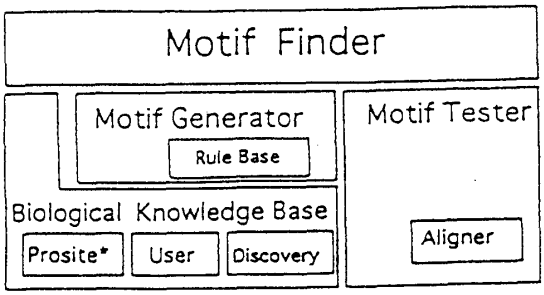

Figure 2

このパターンは "[LIV]-G-x-G-x-[FYM]-[SG]-x-V" である（ここで、[SG] は S と G のどちらかという意味であり、x はどのアミノ酸でもよいという意味である）。チロシン・キナーゼは、protein-kinase-signatures-tyrosine というモチーフのエントリーを持っている。また、チロシン・キナーゼ のモチーフとして、それ自体に登録されているものの他に、上位概念のプロテイン・キナーゼが持つモチーフも参照することができる。これは、*Quixote* のメソッドを介して可能となる。

また、プロテイン・キナーゼは、Prosite では binding-domain というクラスに属している '[AG]-x(4)-G-K-[ST]' というモチーフパターンも持っている（Prosite では、この情報は自然言語で書かれていたために、計算機では読みとれなかったが、我々はこの情報をデータベースに陽に記述し計算機で読みとれるようにした）。

Quixote を用いた記述を Figure 3（知識の階層構造）と Figure 4（モチーフ）を記述したものを に示す。Figure 4 には、2 つのモチーフのエントリーがある。この内、上のモチーフは Prosite* に属するものであり、下のモチーフは User に属するものである。どちらに属するかは source の属性値により記述する。source の属性値 としては、prosite*、user、discovery のどれかをとることができる。

Prosite* の プロテイン・キナーゼ（一番上のエントリー）には otherMotif という属性があるが、これは、Prosite では、他のクラス（domains-general:binding-domain）に含まれているモチーフを、システムがプロテイン・キナーゼのモチーフとしても参照可能にするためのものである。これを可能にするためのメソッドも、*Quixote* で記述されている。このような多重継承は、演繹オブジェクト指向データベースの一機能である。

チロシン・キナーゼとは、acceptor の属性値として、tyrosine を持つ。この二つのクラスの上位クラスであるプロテイン・キナーゼは、acceptor の属性値として、さらに上のクラスである ec-2-7-1 から、alcohol-group を継承している。したがって、蛋白質が発見された時、この蛋白質がプロテイン・キナーゼとのみ判明している場合には acceptor の属性値が alcohol-group であると推論する。しかし、さらに蛋白質がチロシン・キナーゼであると特定される場合には、システムはさらに正確な情報を提示することができる（tyrosine は、alcohol-group の一員である）。これも、演繹オブジェクト指向データベースの一機能である。

5 まとめ

生物学的情報であるモチーフに関する知識を、演繹オブジェクト指向データベース言語 *Quixote* を用いて記述した。オブジェクト指向的な表現方法により生物学的な概念が有効に表現されることを示した。また、演繹オブジェクト指向データベースの演繹的側面が、論理型言語で記述したシステムと相性が良いので演繹オブジェクト指向データベースが知識処理を用いる遺伝子解析にも有効であることが判明した。

参照文献

[Bairoch 1991] Bairoch,A. Prosite : A dictionary of Protein site and pattern : User manual Release 7.00, May 1991.

[Hirosawa *et al.* 1993] Hirosawa,M. *et al.* Protein Multiple Sequence Alignment using Knowledge. *Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences*,Vol.1 pp803-812,1993.

[Yasukawa *et al.* 1992] H.Yasukawa *et al.* Objects, Properties, and Modules in Quixite. *Proc. of FGCS92, pp89-112, 1992*