ICOT Technical Memorandum: TM-1269

TM-1269

知識処理を使った遺伝子情報の解析システム

広沢　誠（日立）、田中　令子（情数研）、
石川　幹人

ICOT
Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan
(03)3456-3191～5

Institute for New Generation Computer Technology

知識処理を使った遺伝子情報の解析システム

Genetic Information Processing using Knowledge Engineering

広沢 誠＋*　　田中 令子＋＋
HIROSAWA Makoto　TANAKA Reiko
石川 幹人＋
ISHIKAWA Masato

＋（財）新世代コンピュータ（ＩＣＯＴ）　＋＋（株）情報数理研究所
ICOT　IMS

We developed a system that extracts information from genetic sequences. The analysis of genetic sequences is an important technology for the research of cancer, AIDS and on. One of the important genetic information to be extracted is motif information. With motif information we can infer the function of genetic sequences identified in experiments. The system we developed extracts genetic information like motif based on a rule base method. The information in the rule base was extracted from biologists. The system also has a biological knowledge base that is described in deductive object-oriented database language QUIXOTE. In this paper the structure of our system and effectiveness of the language in the system are described.

1　概要

生物学のデータベースを活用し生物の遺伝子を解析するシステムのプロトタイプを作成した [Hirosawa *et al.*1993]。システムは、解析のためのルースベースと、解析に必要である生物学の知識ベースから構成されている。この知識ベースは、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] を用いて記述した。

そして、この解析システムを簡単な実問題に適用した。また、その結果を解析し、システムが演繹オブジェクト指向データベースの特徴をいかした遺伝子の解析を行なうことができることを確認した。

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、ウイルス学ではエイズの治療法の研究などにも必須の研究分野である。

遺伝子解析は、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。DNA としてコードされている生物の遺伝子は、生体内で mRNA 等によりアミノ酸の鎖に変換される。これが最終的に、蛋白質となり生体内で部品として機能している。

使用されるアミノ酸は２０種類あり、各々をアルファベットを用いて表現することができる。従って、蛋白質をアルファベットの配列で表すことができる。例えば、"GIVEQCCTSICSLYQLENYCN" 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、Ｉはイソロイ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、この遺伝子が表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含むアミノ酸の配列パターンである。この配列パターンを該当する蛋白質のモチーフと呼ぶ。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が判明しているものもある。例えば、あるクラスの蛋白質が持つモチーフに対応する部位は、その蛋白質 DNA に結合するとわかっている。この時、モチーフに対応する遺伝子に変異が生ずるとその蛋白質は、DNA に結合できなくなる。

モチーフを発見する手法として *multiple alignment* がある。例えば、第１図（上）に示した複数の蛋白質のモチーフを探すことを考える。モチーフを発見するためには、すべての配列に共通な部分の対応づけを行なう必要がある。この対応づけを行なうこと、そして、行なった結果を multiple alignment という（第１図（下））。図では、類似しているアミノ酸を同じカラムに揃えるためにアミノ酸配列の適切な場所に '-' が挿入してある。

第１図の蛋白質では、'#' で示した部分と 'VAIKT' で示した部分にモチーフを発見することができる。このように、モチーフを発見する場合には multiple alignment を作成する必要がある。しかし、現在、品質の良

*現　(株) 日立製作所

```
seq1:KLGQFGEVWMGTWNGTTRVAIKTLKPGTMSPE
seq2:IGFGVYRGTLRLPSQDCKTVAIKTLKDTSPGGQWW
seq3:SQGGFGMVYEGNARDIIKGETRVAIKTVNESASLRERI
seq4:GQGAFVTVYKGLWIPEGEKIPVAIKTLREATSPKANK

seq1:KLGQFGEVWMGTWNGTTR-------VAIKTLKPGTMSPE--
seq2:-IG-FG-VYRGTLRLPSQ---DCKTVAIKTLKDTSPGGQWW
seq3:--SQGGFGMVYEGNARDIIKGET-RVAIKTVNESASLRERI
seq4:--GQGAFVTVYKGLWIPEGE-KI-PVAIKTLREATSPKANK
       ######                  VAIKT
```

Figure 1

い multiple alignment を完全自動的に作成する手段はないので、品質の良い モチーフを自動的に発見することも困難である。

3 遺伝子の解析システム

本システム[Hirosawa *et al.* 1993] 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である。また、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２図に示す。以下、その構成と動作を説明する。

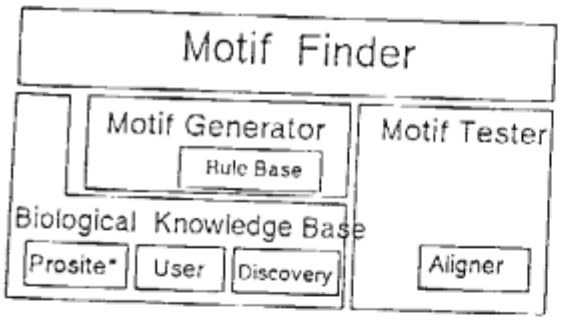

Figure 2

3.1 システムの構成

まず、Aligner は、類似部分の対応づけを簡易な方法で行なう。この multiple alignment を *Motif Finder* は解析し、配列の共通部分を検出する。第１図の例であれば、'VAIKT' というパターンと、類似部分として '#' で示されている部分を検出する。

Motif Genarator は、Motif Finder の解析結果を基に、他に含まれている可能性のあるモチーフを可能性の高いものから順に生成していく。可能性の高いものをいうのは Biological Knowledge Base に含まれるモチーフの情報に基づくものであり、可能性の低いものはこれに基づかないものである。

具体的には、Motif Genarator は、*Motif Rule Base* に含まれる ルールを、ルールに割り当てられている優先度にしたがって実行し、他に含まれている可能性があるモチーフの候補を生成する。現在は、１０個のルールが登録されている。ルールは、必要であれば *Biological Knowledge Base* の *Prosite** と *User* を参照する。

Motif Tester は、生成されたモチーフに当てはまる各部分配列を対応させるという制約の基に multiple alignment を作成し、これが生物学的統計的基準 [Dayhoff *et al.* 1978] を満たしているかを判断する。モチーフとして容認された場合には、これに対応する multiple alignment が Motif Generator に送られ、新たなモチーフ検出サイクルが始まる。なお、発見されたモチーフは Biological Knowledge Base の *Discovery* に登録される。

3.2 Biological Knowledge Base

システムは、Biological Knowledge Base を、モチーフを検出する時に参照し、また、これに、発見したモチーフを格納する。Biological Knowledge Base には、モチーフに関する生物学的知識が登録されている。この知識は、依頼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。

Biological Knowlege Base は、外部からは、モチーフに関する３つのサブデータベースに分かれているように見える。各々は、Prosite*、User、Discovery と名付けられている。Prosite* には、モチーフデータベースとして代表的な Prosite[Bairoch 1991] という既存のデータベース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。このクラス分けは蛋白質の分類法などに基づくものであり、Prosite で採用しているものを基本にしている。しかしながら、このクラス分けでは十分ではないので、生物学的な知識にしたがい再分類を行なっている。

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。

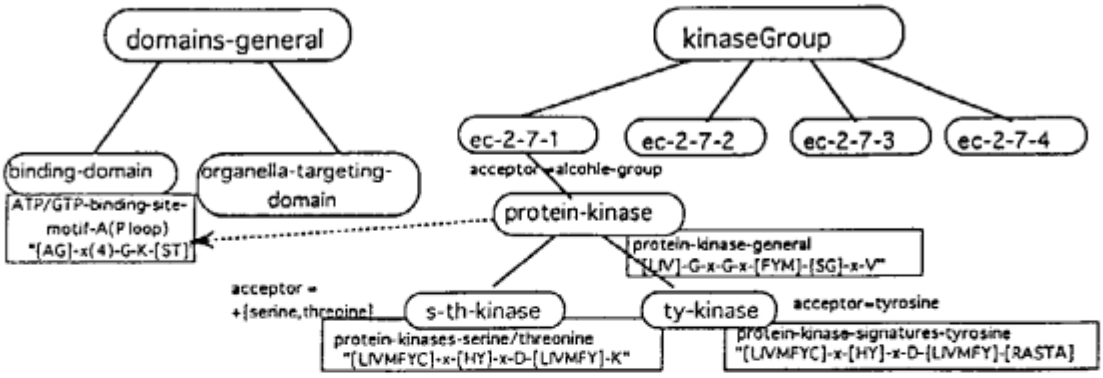

Figure 3

Figure 3 に Prosite* の一部を示す。図では、キナーゼというリン酸基を転移する蛋白質のグループ（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。そして、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、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）。また、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。

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モチーフの他に、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）。

Figure 3 に示した階層構造の *QUIXOTE* を用いた記述を Figure 4 に、プロテイン・キナーゼに登録されているモチーフを Figure 5 に示す。

```
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 4

Figure 5 の上の 3 つのモチーフは Prosite* に属するものであり、一番下のモチーフは 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。

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。そのためのメソッドも、*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。

4 適応例

システムが、生物学的知識を用いてモチーフを発見する例を以下に示す。

Figure 6(上) は、7 本の配列の類似部分の対応づけを、Motif Generator に備えられている Aligner が簡易に行なった結果である。配列の前半部のみを示してある。この結果から Motif Generator は 'G-x-G-x-F-G' という共通パターンを抽出する。

Motif Generator は、このパターンを知識ベース (Prosite*) で検索し、プロテイン・キナーゼが持つ protein-kinase-general というモチーフであることを認識する。なお、Figure に示されていない部分であるが、プロテイン・キナーゼが持つ ATP/GTP binding site というモチーフがあることも認識している。これにより、全ての蛋白質が少なくてもプロテイン・キナーゼであることと、ATP/GTP に結合する能力があることが判明する。

次に、さらに蛋白質の種類を特定できるかを調べる。全ての蛋白質がプロテイン・キナーゼのサブクラスのどちらか一方のみに所属しているかを調べる。しかし、チロシン・キナーゼが持つモチーフも、セリン／スレオニン・キナーゼが持つモチーフも共通には

```
kinaseGroup::protein-kinase[name = "Protein kinases general"]/
   [pattern = "[LIV]-G-x-G-x-[FYM]-[SG]-x-V",
    comment  = "involved in  ATP binding ",
    source   = prosite
    otherMotif = domains_general:binding_domain
       [name = "ATP/GTP-binding site motif A (P-loop)"]
   ];;
protein_kinase::ty_kinase[name = "Protein kinases signatures tyrosine"]/
   [pattern = "[LIVMFYC]-x-[HY]-x-D-[LIVMFY]-[RSTA]-x(2)-N-[LIVMFYC](3)",
    source  = user,
    acceptor = tyrosine,
   ];;
protein_kinase::s_th_kinase[name = "Protein kinases serine/threonine"]/
   [pattern = "[LIVMFYC]-x-[HY]-x-D-[LIVMFY]-K-x(2)-N-[LIVMFYC](3)",
    source  = prosite,
    acceptor +={serine,threonine},
   ];;
kinaseGroup::protein-kinase[name = "Protein kinases ATP-bind"]/
   [pattern = "A-x-K",
    source   = user
   ];;
```

Figure 5

```
                                  # #                  #
seq1:KLGQGCFGEVWMGTWNGTTR-------VAIKTLKPGTMSPE--AFLQEAQVMKKL---RHEKLVQLYAVVSE-EPIYIVTEYMSKGSLLDFLK
seq2:-IGEGEFGEVYRGTLRLPSQ---DCKTVAIKTLKDTSPGGQWWNFLREATIMGQF---SHPHILHLEGVVTKRKPIMIITEFMENGA------
seq3:--GQGSFGMVYEGNARDIIKGEAET-RVAVKTVNESASLRERIEFLNEASVMKGF---TCHHVVRLLGVVSKGQPTLVVMELMAHG-------
seq4:--GSGAFGTVYKGLWIPEGE-KVKI-PVAIKELREATSPKANKEILDEAYVMASV---DNPHVCRLLGICLT-STVQLITQLMPFGCL-----
seq5:LLGKGTFGQVYQVKKKDTQR---IY-AMKVLSKKVIVKKNEIAHTIG-ERNILVTTASKSSPFIVGLKFSFQTPTD-LYLVTDYMS-------
seq6:VLGKGSFGKVMLADRKGTEE---LY-AIKILKKDVVIQDDDVECT-MVEKRVLALL--DKPPFLTQLHSCFQTVDR-LYFVMEYVNGG-----
seq7:TLGTGSFGRVMLVKHKETGN---HY-AMKILDKQKVVKLKQIEHTLNEKRILQAV---NFPFLVKLEFSFKDNSN--LYMVMEYVPGGE----
       * * **                    # #
seq1:KLGQGCFGEVWMGTWNGTTR-------V-AIKTLKPGTM-SPE--AF---LQEAQVM---KKL---RHEKLVQL-YA-VVSE-EPIYIVTEYMSKGSLLDFLK
seq2:-IGEGEFGEVYRGTLRLPSQ---DCKTV-AIKTLKDTSP-GGQWWNF---LREATIM---GQF---SHPHILHL-EG-VVTKRKPIMIITEFMENGA------
seq3:--GQGSFGMVYEGNARDIIKGEAET-RV-AVKTVNESAS-LRERIEF---LNEASVM---KGF---TCHHVVRL-LG-VVSKGQPTLVVMELMAHG-------
seq4:--GSGAFGTVYKGLWIPEGE-KVKI-PV-AIKELREATS-PKANKEI---LDEAYVM---ASV---DNPHVCRL-LG-ICLT-STVQLITQLMPFGCL-----
seq5:LLGKGTFGQVYQVKKKDTQR-------IYAMKVLSKKVI-VKKN-EIAHTIGERNILVTTASK---SSPFIVGLKFS-FQTP-TDLYLVTDYMS---------
seq6:VLGKGSFGKVMLADRKGTEE-------LYAIKILKKDVV-IQDD-DVECTMVEKRVL---ALL--DKPPFLTQL-HSCFQTV-DRLYFVMEYVNGG-------
seq7:TLGTGSFGRVMLVKHKETGN-------HYAMKILDKQKV-VKLK-QIEHTLNEKRIL---QAV---NFPFLVKLEFS-FKDN-SNLYMVMEYVPGGE------
       * * **                      * *                  *
```

Figure 6

持っていないので、一般的なプロテイン・キナーゼであることが判明する。

さらに、知識ベース (User) に登録されているモチーフが配列に存在しないかを調べる。この結果、'A-x-K' というパターンが配列に含まれる可能性が示唆されるので、'A-x-K' というパターンを揃えるという制約のもとで Aligner は、Figure 6(下) を作る。そして、この multiple alignment は、生物学的統計基準 [Dayhoff *et al.* 1978] から妥当であることを Motif Finder は確認する。

最後に、この multiple alignment を調べると、'[LIM]-x-E-[ARK]' いうパターンが存在しているので、これをプロテイン・キナーゼのモチーフとして知識ベース (Discovery) に登録する。このように、本システムでは遺伝子のモチーフを発見することができる。

5 まとめ

モチーフに関する生物学的知識を、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。この記述したものを知識ベースとするモチーフ発見システムを構築した。そして、このシステムでプロテイン・キナーゼという蛋白質の新たなモチーフを発見することができた。発見過程では、演繹オブジェクト指向データベースの機能を利用していることを確認し、演繹オブジェクト指向データベースが遺伝子解析にも有効であることを示した。

参照文献

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

[Dayhoff,O. *et al.* 1978] Dayhoff,M.O., Schwatz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. In Dayhoff,M.O.(ed), *Atlas of Protein Sequence and Structure* Vol.5, *Suppl.3*, Nat. Biomed. Res. Found., Washington, D. C., 363-373.

[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, H.Tsuda, and K.Yokota. Objects, Properties, and Modules in Quixite. *Proc. of FGCS92, pp89-112, 1992*