演繹オブジェクト指向データベースの分子生物学への適用

田中秀俊

(財) 新世代コンピュータ技術開発機構

1 はじめに

分子生物学の分野ではゲノムやタンパク質に関する莫大なデータが未整理状態で累積されてきており、生物学データに適したデータモデルとDBMS、未整理データを整理して知識化する作業を支援する方法、知識の格納と利用の方法、などを整備していく必要がある。

ICOTでは知識表現言語として演繹オブジェクト指向データベース(DOOD)の概念を採用した *Quixote* [1]を設計開発している。この言語の分子生物学への適用、特に知識の格納および利用方法の整備への適用を通じ、データの知識化の基盤となるデータベース/知識ベースの実現を考えている。また、同じくICOTで設計開発した非正規関係DBMS Kappa-P[2]を併用し、知識ベース内の定型的な知識に関する検索効率向上を図る。

本稿では、既存の公共のタンパク質データベースを効率的に内包し、生物学データからの知識抽出の基盤となるような個人用タンパク質知識ベースの試作について、その要求機能や実現可能性について述べる。

2 タンパク質知識ベースの構成と用途

タンパク質知識ベースの当面の用途は、タンパク質の配列や構造と機能との関係を知識として蓄え、それをタンパク質の生成や機能未知タンパク質の機能予測に役立てることにある。そのための手法として現在有力なのが、モチーフ(特定の機能や構造に共通に現れる配列パターン)を手掛りにする方法である。

モチーフを得るには配列のデータベースから同じ構造や機能の配列、もしくは配列の一部を抜きだして、アミノ酸同士の類似性も考慮しながら配列の共通パターンを求めるという作業(Multiple Alignment[3])を行う。モチーフの表現は、確率を導入するような試みもあるが[4]、まだ正規表現的な形式で文字列パターンとしてのみ表されることが多い。公共的なデータベースとして提供されているProSite[5]は、正規表現的な記述を自然言語で補足したものを採用している。数は1992年6月で689種類とまだ少ない。モチーフ抽出には「同じ構造」や「類似機能」といったものの定義もまた重要で、それには一般的な生物学的知識が必要となる。

従ってタンパク質知識ベースに必要な内容を考えると、まず表1に示すような既存の公共データベース群[6]、それらを同時に効率良く扱うための技術、機能や構造の記述方式と利用方法、そして機能の類似の定義に使えるような生物学知識の表現方法の開発、などがあげられる。さらに、科学のデータベースに共通する要求である、公共的な知識と個人的なデータや仮説との共存方法も併せて考える必要がある。

データベース群はDBMS Kappa-Pの管理下に置き、格納・検索効率の向上を図るとともに簡単な検索に関する直接のアクセスを許す。他の知識は *Quixote* で記述し、質問処理も *Quixote* で行う。データベース群に対しても *Quixote* からのアクセス方法を用意する。

表1: 主要公共タンパク質関連データベース

名称	主な内容	国内問合せ先
GenBank	DNA/RNA配列	遺伝研
EMBL	DNA/RNA配列	遺伝研
PIR	タンパク質配列	東理大(JIPID)
Swiss-Prot	タンパク質配列	遺伝研
PDB	タンパク質立体構造	阪大蛋白研, 化情協
ProSite	モチーフ	遺伝研
REBASE	制限酵素	遺伝研
Enzyme	酵素	遺伝研

3 DOODによる機能モチーフの表現と利用

Quixote は演繹データベース言語にオブジェクト指向を導入した形で拡張が施されており、表現力が要求されるタンパク質データやその関連知識の記述に向いている[7][8][9]。ここではタンパク質知識ベースの機能と用途の例として機能モチーフの表現と利用を考える。

3.1 モチーフの表現

Quixote によるモチーフ表現例を式1に示す。

式1:
```
(1) zinc_finger /
    [function="nucleic acid-binding structure"];;
(2) zinc_finger / [pattern = Y];;
(2') zinc_finger [subname = X] / [pattern = Y];;
(3) zinc_finger [function = "exception_1"] / [...];;
```

zinc fingerと呼ばれるモチーフについて、*Quixote* ではその機能(function)とアミノ酸パターン(pattern)とをそれぞれ(1), (2)のように分割して記述できる。これはオブジェクト項('/'の左側)が一致するものを同じオブジェクトとみなし、属性部分('/'の右側)をまとめてしまう機能による。zinc fingerの種類に細分化が起きた、など、研究の進展によって予期せぬ属性が増える場合にも、(2)から(2')のように、オブジェクト項に属

Molecular Biological Databases in DOOD
Hidetoshi Tanaka
Institute for New Generation Computer Technology

性を付加するだけで対応できる柔軟性を持つ。さらに、ここでは暗黙に順序関係とその間の継承関係が定義されていて、(1) の属性部分に記述された属性 (function) は自動的に (2') の形式のオブジェクト全てに継承される。(2') の具体例を式 2 に示す [5]。例外は (3) のようなオブジェクトで、オブジェクト項に書かれた属性値が採用される。

式 2 : ProSite 9版の zinc finger

```
zinc_finger[subname="C2H2"] /
  [pattern="C-x(2,4)-C-x(12)-H-x(3,5)-H"];;
zinc_finger[subname="C3EC4"] /
  [pattern="C-x-H-x-[LIVMFY]-C-x(2)-C-[LIVM]"];;
zinc_finger[subname="GATA"] /
  [pattern="C-x-N-C-x(4)-T-x-L-W-R-R-x(3)-G-x(3)-
            C-N-A-C"];;
zinc_finger[subname="Poly(ADP-ribose) polymerase"] /
  [pattern="C-K-x-C-x-[EQ]-x(3)-K-x(3)-R-x(16,18)-
            W-[YH]-H-x(2)-C"];;
```

3.2 機能記述

モチーフの機能は ProSite のようにモチーフ毎の記述が入手可能な場合と、文献 DB の形で間接的に格納して内容検索で対応する場合とがある。式 3 にそれぞれの簡単な記述形式を示す。式中、属性名末尾の '+' は集合値をとる属性であることを示す。(1),(2) とも、function や abstract のなどの値 (自然言語による長文) に対する内容検索は、システム側に付加した方がいい機能だが、現段階では未実装である。*Quixote* ではこの他に化学反応式を格納して機能記述の一助とすることもできる [7][8][9]。

式 3 :

```
(1) zinc_finger / [function="'Zinc finger' domains
 [1-5] are nucleic acid-binding protein structures
 first identified in the Xenopus transcription factor
 TFIIIA. These domains have since been found in
 numerous nucleic acid-binding proteins...."];;
(2) reference [id="TWXL3", no=1] /
      [keywords+ = {"...", ..},
       authors+ = {"Ginsberg A.M.", ...},
       journal = "Cell (1984) 39:479-489",
       abstract = ".....", ... ];;
```

3.3 モチーフの利用

公共データベースは Kappa に格納され、*Quixote* からアクセスする際はデータベースごとのモジュールの形で表現される (*Quixote* - Kappa 間のアクセスは他言語インタフェースの実装後提供される予定)。式 4 はモジュールをまたがる質問の例で、PIR の mammal 由来の配列に ProSite に登録された dna_binding モチーフがあるかどうか、という質問を意図している。':' の左側がモジュール名である。'||' の右には変数への制約を記述できる。例ではモチーフ、配列とも公共データベースのモジュール (pir,prosite) を指定している。個人的データベースを別モジュールに管理してある場合は、公共・個人両方を含むようなモジュールを定義する等により容易に同様な質問をすることができる。

original_method にあると仮定されている include というオブジェクトはパターンの包含判定プログラムである。C-x(2,4)-C-x(12)-H-x(3,5)-H のようなパターン記述を解釈し、アミノ酸配列にサーチをかける。このようなオブジェクトは *Quixote* の他言語インタフェースの実装後の実現を予定している。

式 4 :

```
?- prosite : X / [pattern = Xp],
   pir : Y / [hostname = Yh, sequence = Ys],
   orginal_method:include[seq=Ys,pat=Xp]
  ||{X =< dna_binding, Yh =< mammal}.
```

4 おわりに

タンパク質解析の基盤となるような個人用知識ベースについて、モチーフの知識ベースを例にその利用形態と *Quixote* による表現方式を述べた。

課題としては、知識ベースシステムとしての情報検索技術のサポートが挙げられる。例えば、前章で内容検索の必要性に言及している。また独自に解読したアミノ酸配列など、新規配列が既存のモチーフを含むかどうかのチェックを考える時、質問の方を格納しておきデータ追加の際に処理してしまう方法 [10] が有効であると思われる。

今後はタンパク質の知識ベースの構築を通じて、このような分野特有の問題のうち一般化できる部分や、逆に一般的な技術で分野特有の問題に適用可能なものがないかといった所を、引き続き検討して行きたい。

参考文献

[1] Yasukawa, H., Tsuda, H. and Yokota, K.: "Objects, Properties, and Modules in *Quixote*", *FGCS 92*, (Jun 1992).

[2] Kawamura, M., Sato, H., Naganuma, K. and Yokota, K.: "Parallel Database Management System : Kappa-P", *FGCS 92*, (Jun 1992).

[3] Ishikawa, M., Hoshida, M., Hirosawa, M., Toya, T., Onizuka, K. and Nitta, K.: "Protein Sequence Analysis by Parallel Inference Machine" *FGCS 92*, (Jun 1992).

[4] 小長谷：「確率的モチーフ：現状と課題」, 第2回 公開ワークショップ・ヒトゲノム計画と情報処理技術, (Dec 1991).

[5] Bairoch, A.: "ProSite: A Dictionary of Protein Sites and Patterns", *User Manual, Release 9.0*, (Jun 1992).

[6] Lesk, A.M. (ed.): *Computational Molecular Biology, Sources and Methods for Sequence Analysis*, Oxford Univ. Press, (1988).

[7] 田中:「分子生物学のデータベース」, 情報処理学会研究報告 91-DBS-84-23, (Jul 1991).

[8] Tanaka, H.: "Protein Function Database as a Deductive and Object-Oriented Database", *Database and Expert Systems Applications*, Springer-Verlag, (Aug 1991).

[9] Tanaka, H.: "Integrated System for Protein Information Processing", *FGCS 92*, (Jun 1992).

[10] Terry, D., Goldberg, D., Nichols, D. and Oki, B.: "Continuous Queries over Append-Only Databases", *SIGMOD 92*, (Jun 1992).

演繹オブジェクト指向データベースの分子生物学への適用

田中秀俊

(財) 新世代コンピュータ技術開発機構

1 はじめに

分子生物学の分野ではゲノムやタンパク質に関する莫大なデータが未整理状態で累積されてきており、生物学データに適したデータモデルとDBMS、未整理データを整理して知識化する作業を支援する方法、知識の格納と利用の方法、などを整備していく必要がある。

ICOTでは知識表現言語として演繹オブジェクト指向データベース(DOOD)の概念を採用した *Quixote* [1]を設計開発している。この言語の分子生物学への適用、特に知識の格納および利用方法の整備への適用を通じ、データの知識化の基盤となるデータベース/知識ベースの実現を考えている。また、同じくICOTで設計開発した非正規関係DBMS Kappa-P[2]を併用し、知識ベース内の定型的な知識に関する検索効率向上を図る。

本稿では、既存の公共のタンパク質データベースを効率的に内包し、生物学データからの知識抽出の基盤となるような個人用タンパク質知識ベースの試作について、その要求機能や実現可能性について述べる。

2 タンパク質知識ベースの構成と用途

タンパク質知識ベースの当面の用途は、タンパク質の配列や構造と機能との関係を知識として蓄え、それをタンパク質の生成や機能未知タンパク質の機能予測に役立てることにある。そのための手法として現在有力なのが、モチーフ(特定の機能や構造に共通に現れる配列パターン)を手掛りにする方法である。

モチーフを得るには配列のデータベースから同じ構造や機能の配列、もしくは配列の一部を抜きだして、アミノ酸同士の類似性も考慮しながら配列の共通パターンを求めるという作業(Multiple Alignment[3])を行う。モチーフの表現は、確率を導入するような試みもあるが[4]、まだ正規表現的な形式で文字列パターンとしてのみ表されることが多い。公共的なデータベースとして提供されているProSite[5]は、正規表現的な記述を自然言語で補足したものを採用している。数は1992年6月で689種類とまだ少ない。モチーフ抽出には「同じ構造」や「類似機能」といったものの定義もまた重要で、それには一般的な生物学的知識が必要となる。

従ってタンパク質知識ベースに必要な内容を考えると、まず表1に示すような既存の公共データベース群[6]、それらを同時に効率良く扱うための技術、機能や構造の記述方式と利用方法、そして機能の類似の定義に使えるような生物学知識の表現方法の開発、などがあげられる。さらに、科学のデータベースに共通する要求である、公共的な知識と個人的なデータや仮説との共存方法も併せて考える必要がある。

データベース群はDBMS Kappa-Pの管理下に置き、格納・検索効率の向上を図るとともに簡単な検索に関する直接のアクセスを許す。他の知識は *Quixote* で記述し、質問処理も *Quixote* で行う。データベース群に対しても *Quixote* からのアクセス方法を用意する。

表1: 主要公共タンパク質関連データベース

名称	主な内容	国内問合せ先
GenBank	DNA/RNA配列	遺伝研
EMBL	DNA/RNA配列	遺伝研
PIR	タンパク質配列	東理大(JIPID)
Swiss-Prot	タンパク質配列	遺伝研
PDB	タンパク質立体構造	阪大蛋白研, 化情協
ProSite	モチーフ	遺伝研
REBASE	制限酵素	遺伝研
Enzyme	酵素	遺伝研

3 DOODによる機能モチーフの表現と利用

Quixote は演繹データベース言語にオブジェクト指向を導入した形で拡張が施されており、表現力が要求されるタンパク質データやその関連知識の記述に向いている[7][8][9]。ここではタンパク質知識ベースの機能と用途の例として機能モチーフの表現と利用を考える。

3.1 モチーフの表現

Quixote によるモチーフ表現例を式1に示す。

式1:
```
(1) zinc_finger /
    [function="nucleic acid-binding structure"].;
(2) zinc_finger / [pattern = Y].;
(2') zinc_finger [subname = X] / [pattern = Y].;
(3) zinc_finger [function = "exception_1"] / [...].;
```

zinc fingerと呼ばれるモチーフについて、*Quixote* ではその機能(function)とアミノ酸パターン(pattern)とをそれぞれ(1), (2)のように分割して記述できる。これはオブジェクト項('/'の左側)が一致するものを同じオブジェクトとみなし、属性部分('/'の右側)をまとめてしまう機能による。zinc fingerの種類に細分化が起きた、など、研究の進展によって予期せぬ属性が増える場合にも、(2)から(2')のように、オブジェクト項に属

Molecular Biological Databases in DOOD
Hidetoshi Tanaka
Institute for New Generation Computer Technology

性を付加するだけで対応できる柔軟性を持つ。さらに、ここでは暗黙に順序関係とその間の継承関係が定義されていて、(1) の属性部分に記述された属性 (function) は自動的に (2') の形式のオブジェクト全てに継承される。(2') の具体例を式 2 に示す [5]。例外は (3) のようなオブジェクトで、オブジェクト項に書かれた属性値が採用される。

式 2 : ProSite 9 版の zinc finger

```
zinc_finger[subname="C2H2"] /
  [pattern="C-x(2,4)-C-x(12)-H-x(3,5)-H"];;
zinc_finger[subname="C3HC4"] /
  [pattern="C-x-H-x-[LIVMFY]-C-x(2)-C-[LIVM]"];;
zinc_finger[subname="GATA"] /
  [pattern="C-x-W-C-x(4)-T-x-L-W-R-R-x(3)-G-x(3)-
            C-W-A-C"];;
zinc_finger[subname="Poly(ADP-ribose) polymerase"] /
  [pattern="C-K-x-C-x-[EQ]-x(3)-K-x(3)-R-x(16,18)-
            W-[YH]-H-x(2)-C"];;
```

3.2 機能記述

モチーフの機能は ProSite のようにモチーフ毎の記述が入手可能な場合と、文献 DB の形で間接的に格納して内容検索で対応する場合とがある。式 3 にそれぞれの簡単な記述形式を示す。式中、属性名末尾の '+' は集合値をとる属性であることを示す。(1),(2) とも、function や abstract のなどの値 (自然言語による長文) に対する内容検索は、システム側に付加した方がいい機能だが、現段階では未実装である。*Quixote* ではこの他に化学反応式を格納して機能記述の一助とすることもできる [7][8][9]。

式 3 :

```
(1) zinc_finger / [function="'Zinc finger' domains
 [1-6] are nucleic acid-binding protein structures
 first identified in the Xenopus transcription factor
 TFIIIA. These domains have since been found in
 numerous nucleic acid-binding proteins...."];;
(2) reference [id="TWXL3", no=1] /
      [keywords+ = {"...", ..},
       authors+ = {"Ginsberg A.M.", ...},
       journal = "Cell (1984) 39:479-489",
       abstract = ".....", ... ];;
```

3.3 モチーフの利用

公共データベースは Kappa に格納され、*Quixote* からアクセスする際はデータベースごとのモジュールの形で表現される (*Quixote* - Kappa 間のアクセスは他言語インタフェースの実装後提供される予定)。式 4 はモジュールをまたがる質問の例で、PIR の mammal 由来の配列に ProSite に登録された dna_binding モチーフがあるかどうか、という質問を意図している。':' の左側がモジュール名である。'||' の右には変数への制約を記述できる。例ではモチーフ、配列とも公共データベースのモジュール (pir,prosite) を指定している。個人的データベースを別モジュールに管理してある場合は、公共・個人両方を含むようなモジュールを定義する等により容易に同様な質問をすることができる。

original_method にあると仮定されている include というオブジェクトはパターンの包含判定プログラムである。C-x(2,4)-C-x(12)-H-x(3,5)-H のようなパターン記述を解釈し、アミノ酸配列にサーチをかける。このようなオブジェクトは *Quixote* の他言語インタフェースの実装後の実現を予定している。

式 4 :

```
?- prosite : X / [pattern = Xp],
   pir : Y / [hostname = Yh, sequence = Ys],
   orginal_method:include[seq=Ys,pat=Xp]
  ||{X =< dna_binding, Yh =< mammal}.
```

4 おわりに

タンパク質解析の基盤となるような個人用知識ベースについて、モチーフの知識ベースを例にその利用形態と *Quixote* による表現方式を述べた。

課題としては、知識ベースシステムとしての情報検索技術のサポートが挙げられる。例えば、前章で内容検索の必要性に言及している。また独自に解読したアミノ酸配列など、新規配列が既存のモチーフを含むかどうかのチェックを考える時、質問の方を格納しておきデータ追加の際に処理してしまう方法 [10] が有効であると思われる。

今後はタンパク質の知識ベースの構築を通じて、このような分野特有の問題のうち一般化できる部分や、逆に一般的な技術で分野特有の問題に適用可能なものがないかといった所を、引き続き検討して行きたい。

参考文献

[1] Yasukawa, H., Tsuda, H. and Yokota, K.: "Objects, Properties, and Modules in *Quixote*", *FGCS 92*, (Jun 1992).

[2] Kawamura, M., Sato, H., Naganuma, K. and Yokota, K.: "Parallel Database Management System : Kappa-P", *FGCS 92*, (Jun 1992).

[3] Ishikawa, M., Hoshida, M., Hirosawa, M., Toya, T., Onizuka, K. and Nitta, K.: "Protein Sequence Analysis by Parallel Inference Machine" *FGCS 92*, (Jun 1992).

[4] 小長谷：「簡単的モチーフ：現状と課題」, 第 2 回 公開ワークショップ・ヒトゲノム計画と情報解析技術, (Dec 1991).

[5] Bairoch, A.: "ProSite: A Dictionary of Protein Sites and Patterns", *User Manual, Release 9.0*, (Jun 1992).

[6] Lesk, A.M. (ed.): *Computational Molecular Biology, Sources and Methods for Sequence Analysis*, Oxford Univ. Press, (1988).

[7] 田中:「分子生物学のデータベース」, 情報処理学会研究報告 91-DBS-84-23, (Jul 1991).

[8] Tanaka, H.: "Protein Function Database as a Deductive and Object-Oriented Database", *Database and Expert Systems Applications*, Springer-Verlag, (Aug 1991).

[9] Tanaka, H.: "Integrated System for Protein Information Processing", *FGCS 92*, (Jun 1992).

[10] Terry, D., Goldberg, D., Nichols, D. and Oki, B.: "Continuous Queries over Append-Only Databases", *SIGMOD 92*, (Jun 1992).