アミノ酸配列研究用データベースの試作

猪股 優†、○長内 隆‡、後藤 智範†、山本 晴彦†

Experimental Construction of Amino acid Sequence Databases

Masaru INOMATA, ○Takashi OSANAI, Tomonori GOTOH, Haruhiko YAMAMOTO

There are typical three kinds of of Amino acid Sequence Databases, PDB, pdb-aa, nr-aa. Originally these databases have their own format. However, each format seems as records of data, is not effective for advanced research such as homological search for amino acid sequences of proteins, ORF prediction. Experimental construction of XML database have been under construction to these databases for the purpose of efficient usage in related area. This paper reports the characteristics and problems on these databases.

1．はじめに

近年、計算機の発達スピードは凄まじい。これを受けて計算機を利用した、特に医療への応用を目的とした遺伝子や蛋白質の配列情報解析は、各種ゲノムプロジェクトの広がりとそれに続く構造プロテオミクスプロジェクトの進展から注目されている。この分野ではかねてより遺伝子のORF予測、アミノ酸配列の相同性検索・機能推定、タンパク質の二次構造予測、ホモロジーモデリングなど計算機の利用は見受けられてきた。しかし、その基礎データを納めたフォーマット、例えば遺伝子配列・アミノ酸配列を納めたFASTA形式やタンパク質立体構造を納めたPDB形式は「データの記録として」と言った趣が強く、保存形式を見るに意味的に有用な情報の単位とした構造化はあまりされていない。この形態では今後の計算機による２次利用という点において、いくらか不便に思える。今回、当該分野の代表的なデータを対象にXML化を指向した形式でデータを再構築し、その後の利用法までを考えてみる。

2. 対象データ

今回取り扱ったデータは下記の２フォーマット３種類である。以後の節で個々について説明する。

(1) 蛋白質立体構造データを納めたPDBフォーマットデータ

(2) アミノ酸配列データを納めたFASTAフォーマットデータ(pdb－aa、 nr・aa)

2.1 蛋白質立体構造データ

蛋白質立体構造データは通常PDBと呼称される。PDBとは*Protein Data Bank*の略称であり、これは、RCSB(*Research Collaboratory for Structural Bioinformatics*／http:// www.rcsb.org/pdb/)等のサイトで配布されている。１ファイル１レコードのデータで、その形式は図１のとおり

† 神奈川大学 理学部

‡ 理化学研究所 タンパク質構造機能研究グループ

である。

```
HEADER    ACTIN-BINDING PROTEIN                    29-SEP-95   1SOL      1SOL   2
TITLE     A PIP2 AND F-ACTIN-BINDING SITE OF GELSOLIN, RESIDUE             1SOL   3
TITLE    2 150-169 (NMR, AVERAGED STRUCTURE)                               1SOL   4
COMPND    MOL_ID: 1;                                                       1SOL   5
COMPND   2 MOLECULE: GELSOLIN (150-169);                                   1SOL   6
COMPND   3 CHAIN: NULL                                                     1SOL   7
SOURCE    MOL_ID: 1;                                                       1SOL   8
SOURCE   2 SYNTHETIC: YES                                                  1SOL   9
EXPDTA    NMR, MINIMIZED AVERAGE STRUCTURE                                 1SOL  10
```

図1：PDB フォーマット（1SOL.pdb より一部抜粋）

データはフラットファイルであり、人間の可視化性を強めたものである。個々のデータは各行先頭から6文字以内で記された単語(以後識別子と呼称)で区別するようになっている。

表1：PDB 識別子一覧

ANISOU	HETNAM	REVDAT
ATOM	HETSYN	SCALE1
AUTHOR	HYDBND	SCALE2
CAVEAT	JRNL	SCALE3
CISPEP	KEYWDS	SEQADV
COMPND	LINK	SEQRES
CONECT	MASTER	SHEET
CRYST1	MODEL	SIGATM
DBREF	MODRES	SIGUIJ
END	MTRIX1	SITE
ENDMDL	MTRIX2	SLTBRG
EXPDTA	MTRIX3	SOURCE
FORMUL	OBSLTE	SPRSDE
FTNOTE(*1)	ONHOLD(*2)	SSBOND
HEADER	ORIGX1	TER
HELIX	ORIGX2	TITLE
HET	ORIGX3	TURN
HETATM	REMARK	TVECT

*1：Footnotes (FTNOTE) have been

*2：解説テキストに存在しない識別子

また、1行80文字という制約があるため、それを超えるものは複数行にわたって記述される。図1では識別子[TITLE]、[COMPND]、[SOURCE]の部分がそれにあたる。これら識別子は PDB 内に 54 種類存在するが、PDB フォーマットを解説しているテキストでは、すでに扱わなくなっている項目やテキスト中に存在しない物もある。

2.2 アミノ酸配列データ

アミノ酸配列データを納めた FASTA 形式がある。これもまたタンパク質立体構造データと同じくフラットファイルである。内容は単純で、先頭行に">配列 ID"となる識別子を付けて改行後にその配列の実際の配列(遺伝子配列、タンパク質アミノ酸配列:図中の下線)を記述する形式である。一つのファイルに一つだけの配列エントリーを入れただけでも構わないが、先頭行に再度">配列 ID*"を追加する事により、複数の配列情報を一つのファイルにまとめる事が出来る。この配列 ID の由来を選ぶ事により、例えば、既に立体構造が解明されたアミノ酸配列を集めた pdb-aa なるものや、日々報告されるアミノ酸配列を集め、その冗長性をなくした NR(*Non-redundant*)-aa なるものが作成可能である。これらは既に NCBI(*National Center for Biotechnology Information,* http:// www.ncbi.nlm.nih.gov/)等のサイトで入手可能である。その他にも固有の生物種が持つ配列を集めた物など多様

に存在する。

NR データは、複数のデータベースを元に生成されたという経緯があるために、同一のアミノ酸配列でありながら異なる ID を複数持つものもある。そのような場合は区切りコード（図中の^A）を用いて個々のレコードを表し、アミノ酸配列を連結するという形式をとっている。

```
>gi|229728|pdb|1C2R|A Chain A, Cytochrome c2^Agi|229729|pdb|1C2R|B
Chain B, Cytochrome c2
GDAAKGEKEFNKCKTCHSIIAPDGTEIVKGAKTGPNLYGVVGRTAGTYPEFKY
KDSIVALGASGFAWTEEDIATYVKDPGAFLKEKLDDKKAKTGMAFKLAKGGE
DVAAYLASVVK
```

図２：NR データのフォーマット(pdb-aa より抜粋)

3. XML - DB 化

第２節において説明したデータに対してタグをつけてゆくことで XML-DB 化を目指す。この作業を行うことにより、ユーザは ID やタイトルといったものでレコードをダイレクトに指定・操作が可能で、またある特定のデータ、例えば PDB に収録された原子の x 座標のみを指定し、新たなデータベースを生成することも容易になる。

3.1 タグ PDB フォーマット

PDB のタグ付けに関しては図３のような構成になっている。<REC></REC>は１レコード単位を意味するタグである。その他は見ての通り、入れ子の構造を持つ物もあれば、フラットな状態のままの物も存在する。

```
<REC>
        <HEADER>
                <CLASS></CLASS>
                <DATE></DATE>
                <ID></ID>
        </HEADER>
        <REVDAT>
                <mod num="?">
                        <modDT></modDT>
                        <modID></modID>
                        <modTP></modTP>
                        <modRD></modRD>
                </mod>
        </REVDAT>
        <KYEWDS></KEYWDS>
          . . .
</REC>
```

図３：PDB タグフォーマット(一部抜粋)

```
<REC>

        <G-ID></G-ID>

        <DB-NAME></DB-NAME>

        <DB-ID></DB-ID>

        <NAME></NAME>

        <SEQ-LEN></SEQ-LEN>

        <SEQUENCE></SEQUENCE>

</REC>
```

図４：NR タグフォーマット

3.2 タグ NR データフォーマット

NR データに対するタグ付けは、元来データ構成が単純なため、図４を見てもわかるように簡素である。今回タグをつけるにあたり、オリジナルのデータでは存在しなかったアミノ酸配列部の長さを<SEQ-LEN></SEQ-LEN>で括り追加を施している。これにより、配列長指定によるデータ操作が可能となる。また、異なる ID が１レコード中に存在する場合は、レコードを分割し「１

ID／１アミノ酸」という形式でデータを保存している。

3.3 タグ付け結果

データサイズの変動を表２に示す。

表２：タグ付け前後のデータサイズの変動（単位：MB）

	PDB	pdb・aa	nr・aa
タグ付け前	6,765	3.13	260
タグ付け後	4,665	5.39	486

４．考察

4.1 データサイズの変動

表２の通り、NR データの２種類に関しては単純にデータ量が増加している。これは追加されたタグの分がそのまま増加量につながったものである。逆に PDB に関しては、そのサイズが減少している。これは現状のタグ付けプログラムでは、タグの種類が少ないのに加えて、１行 80 文字という制約のために存在した意味を持たない空白文字を省略した結果によるものと考えることができる。タグの構造設計を煮詰め、その種類を増加させることで、さらに細かくデータを扱うことが可能となるが、データサイズが大幅に増加することが予想されるため、この問題に関しては今後も検討していく余地があると思われる。

4.2 利用方法

本研究は、当初からサーバ・クライアント型システムによる NR データ／PDB 連動検索、および座標計算等が研究できる環境を最終形態とし進めてきたが、XML 化によりその幅は広がると思われる。図５は、XSL を使用し PDB の HEADER 部分を抜き出した形でブラウザにリストとして表示をさせている例である。今後はある特定の条件で検索をかけ、図５のようにリストとして表示し、その中から必要なデータだけをダウンロードする事や、ブラウザ上から直接解析ツールに移行し結果を得ることも容易になると思われる。

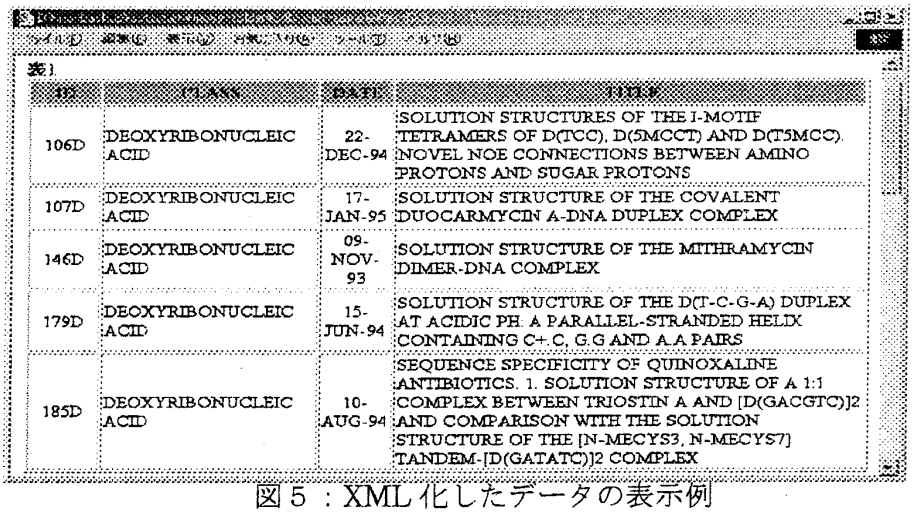

表1

ID	CLASS	DATE	TITLE
106D	DEOXYRIBONUCLEIC ACID	22-DEC-94	SOLUTION STRUCTURES OF THE I-MOTIF TETRAMERS OF D(TCC), D(5MCCT) AND D(T5MCC). NOVEL NOE CONNECTIONS BETWEEN AMINO PROTONS AND SUGAR PROTONS
107D	DEOXYRIBONUCLEIC ACID	17-JAN-95	SOLUTION STRUCTURE OF THE COVALENT DUOCARMYCIN A-DNA DUPLEX COMPLEX
146D	DEOXYRIBONUCLEIC ACID	09-NOV-93	SOLUTION STRUCTURE OF THE MITHRAMYCIN DIMER-DNA COMPLEX
179D	DEOXYRIBONUCLEIC ACID	15-JUN-94	SOLUTION STRUCTURE OF THE D(T-C-G-A) DUPLEX AT ACIDIC PH. A PARALLEL-STRANDED HELIX CONTAINING C+.C, G.G AND A.A PAIRS
185D	DEOXYRIBONUCLEIC ACID	10-AUG-94	SEQUENCE SPECIFICITY OF QUINOXALINE ANTIBIOTICS. 1. SOLUTION STRUCTURE OF A 1:1 COMPLEX BETWEEN TRIOSTIN A AND [D(GACGTC)]2 AND COMPARISON WITH THE SOLUTION STRUCTURE OF THE [N-MECYS3, N-MECYS7] TANDEM-[D(GATATC)]2 COMPLEX

図５：XML 化したデータの表示例

5. 今後の課題

XML化の一番の問題点はファイルサイズの増加である。今回はまだ発展途上の段階であり、タグ付けに関してもまだまだ検討する余地が残されている。本来ならば存在する個々のデーター一つ一つが識別できるようタグを付けることが望ましいが、研究用プログラムにおいて必須でないデータまでにタグをつけ、データ量を増大させることが計算機にとって有効でないことは明らかである。記憶装置の大容量化が進み、ブロードバンド時代を迎えたとはいえ、扱うデータ量はよりコンパクトであるべきである。よって、この点について今後検討し続け、適宜プログラムを改良していく必要があると思われる。

また、去る3月に国立遺伝子研究所で行われた研究会において、大阪大学蛋白質研究所もPDBデータのXML化という、同等な研究を行っていることが発表された。そちらの動向も注目しつつ、本研究を継続してゆく予定である。

参考文献およびサイト

[1] 金久 寛：「ゲノム情報への招待」、共立出版(1996)

[2] 高木 利久, 金久 寛：「ゲノムネットのデータベース利用法」、共立出版(1999)

[3] Russell F. Doolittle：「Methods in ENZYMOLOGY Volume 266 Computer Methods for Macromolecular Sequence Analysis」、ACADEMIC PRESS(1996)

[4] NIAS DNA Bank,、http://www.dna.affrc.go.jp/

[5] The RCSB Protein Data Bank,、http://www.rcsb.org/

[6] NCBI Home Page、http://www.ncbi.nlm.nih.gov/