オブジェクトDBMSによる 生体高分子構造情報データベースの構築*

稲田 稔, 黄 湘文, 牧之内顕文
九州大学大学院システム情報科学府知能システム学専攻
九州大学工学部電気情報工学科
九州大学大学院システム情報科学研究院知能システム学部門†

1 はじめに

今日, 生命科学分野では種々の生体高分子の配列・構造決定が進んでいる. 特にタンパク質で構造情報はその機能との関わりのため重要である. これら大量の生体高分子の情報を効率よく管理し, またそれらの情報を解析することにより生命活動のルールを得たい.

本論文では, 生体高分子情報の集積と解析のために, オブジェクトデータベースマネジメントシステム (DBMS) を用いて生体高分子の構造情報を扱うオブジェクトデータベースの構築について述べる. オブジェクトデータベースとして構築することにより, データ間の関係を自然に表現することができる.

2 Protein Data Bank(PDB)

DB を構築するにあたり, Protein Data Bank(PDB; http://www.rcsb.org/pdb/)[2] から得られる情報を利用した. PDB はアメリカ Rutgers 大学を中心とする Reserch Collaboratory for Structural Bioinfomatics(RCSB) によって運営される, タンパク質を中心とした生体高分子構造のアーカイブである. 世界中の研究者によって構造決定された生体高分子が登録され, その情報は Web 上で公開されている.

3 実装

3.1 環境

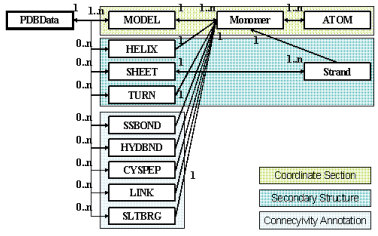
本システムはオブジェクト DBMS Jasmine[1] を用いて, SUN Ultra-5 ワークステーション上に実装した. プログラムは C と DBMS の問い合わせ言語である Object Database Query Language(ODQL) により記述した.

3.2 設計

PDB の情報を格納するために以下のようなクラスを設計した. 設計では, オブジェクト間の関係が自然界における原子・分子の関係を表現するように特に注意した.

- PDBData クラス
  主クラス. PDB の 1 エントリに対してこのクラスオブジェクトが 1 つ対応する.
- MODEL クラス
  分子の原子座標の組を記録する. 属性として Monomer オブジェクトの List を持ち, これによって配列情報と座標情報を保持する.
- Monomer クラス, AminoAcid クラス, Nucleotide クラス, HET クラス
  分子の構成残基の情報を記録. Monomer クラスは AminoAcid, Nucleotide, HET クラスの上位の抽象クラス.
- ATOM クラス, HETATM クラス
  分子を構成する各原子の名前や三次元座標情報を記録.
- HELIX クラス, SHEET クラス, TURN クラス
  二次構造情報を記録. 二次構造を構成する部分の開始残基, 終了残基, など.
- SSBOND クラス, LINK クラス, HYDBND クラス, SLTBRG クラス, CISPEP クラス
  分子内化学結合の情報を記録.
- COMPND クラス, SOURCE クラス
  分子の化学的出所や性質に関する情報を記録.
- JRNL クラス, REMARK クラス, REMARK2 クラス, REMARKn クラス, SITE クラス
  分子に関する文献情報・注釈情報を記録.
- CRYST1 クラス, ORIGX クラス, SCALE クラス, MTRIX クラス, TVECT クラス
  分子の座標変換行列を記録.

各クラス間の関係を図に表すと図 1 のようになる.



*Construction of a biomoleculer structural information database using object-oriented database management system

†Minoru INATA, Ng Hsiang Boon, Akifumi MAKINOUCHI (Department of Electrical Engineering and Computer Science, Kyushu University, Graduate School of Information Science and Electorical Engineering, Department of Intelligent Systems, Kyushu University)

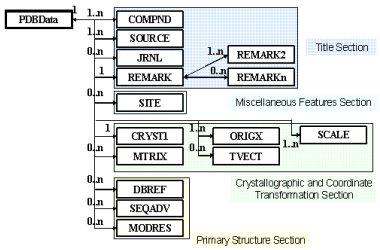


図 1: 各クラス間の関係

4 評価

4.1 PDB web site との検索機能比較

PDB は Web 上で以下のような検索を提供している.

- SearchLite
  エントリに対するキーワード検索. キーワードと一致するテキストを含んでいるエントリの情報を返す. 検索範囲はエントリ中の全データではなく, 一部に限られる. 例えば, データ投稿者, 分類, 注釈, など.

- SearchFields
  多数のフィールドを用いた検索. 各フィールドは検索範囲が決まっており, 記入されたフィールドのみが検索に影響し, 条件に合致するエントリの情報を返す. フィールドの組み合わせにより指定できる条件は SearchLite より広範である. 例えば, 分子の種類や構造決定手法についても指定できる.

このような検索が構築した DB でも可能であることを確かめた.

SearchLite については, DB で SearchLite の検索対象の情報が格納されているクラス・属性に対して検索を実行すればよい. ODQL では次のように記述できる. "kinase" というキーワードを含むデータを検索.

```
Bag<PDBData> result;
result = PDBData from PDBData
         where PDBData.author
               .hasElement("kinase")
         or PDBData.title
            .hasElement("kinase")
         or ...
```

SearchFields については, 各フィールドごとに検索範囲が決まっているので, where 句において各フィールドに対応する検索条件を and で結合していけばよい.
例) 分子の分類が"HYDROLASE (O-GLYCOSYL)" で, 構造解析手法が"X-RAY DIFFRACTION"

```
Bag<PDBData> result;
result = PDBData from PDBData
         where PDBData.classification
               == "HYDROLASE (O-GLYCOSYL)"
         and PDBData.expdta
             .hasElement("X-RAY DIFFRACTION");
```

また, SearchLite, SerachFields ではフォーム等に決められた条件の検索しかできないが, 本 DB ではユーザが自由に条件を設定した検索も可能である.
例) 複数の構造解析結果を持つ分子の種類別個数を求める.

```
Bag<[String type, Integer num]> result;
Bag<PDBData> tmp;
tmp = PDBData from PDBData
      where PDBData.model.count()>1;
result = group s in tmp by (s.type)
         with(partition().count());
```

4.2 データマイニング手法の適用

本 DB では先に述べたような検索だけでなく, データや問い合わせ結果に対して独自の条件で統計を取ったり, データマイニングを行うことが ODQL を用いて容易にできる. より複雑な処理をしたい場合には ODQL に C, C++等の外部言語を合わせて使用することもできる.

例)DB に含まれるタンパク質のヘリックス構造を構成するアミノ酸の種類別分布を見る.
対象オブジェクト数
HELIX オブジェクト:807, AminoAcid オブジェクト:9021
結果を図 2 に示す. [3] によれば, アラニン (ALA), グルタミン酸 (GLU), ロイシン (LEU), メチオニン (MET) などが α-ヘリックスに適したアミノ酸とされているので, この結果はおおむね正しいといえる.

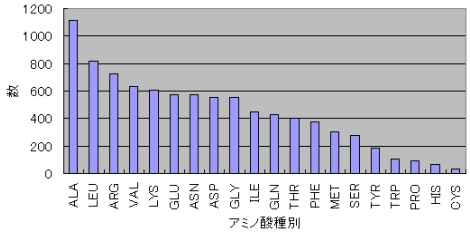


図 2: ヘリックス構成アミノ酸の分布

5 おわりに

本論文ではオブジェクト DBMS により生体高分子構造情報データベースを構築した. オブジェクト DBMS によって構築したことにより分子間の関係や分子と残基の関係などをオブジェクト間の関係として扱うことができる. また PDB の Web 上ではできない複雑な問い合わせや独自に統計を取ることも可能となる.

参考文献

[1] Setrag Khoshafian, Surapal Dasananda, Norayr Minassian,: "The Jasmine Object Database", Morgan Kaufmann Publishers, Inc., 1999

[2] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov and Philip E. Bourne,: "The Protein Data Bank", Nucleic Acids Research, Vol.28 235-242, 2000

[3] Carl Branden & John Tooze 著, 勝部幸輝 竹中章郎 福山恵一 松原 央 監訳,: " タンパク質の構造入門 第 2 版", Newton Press, 2000