3W－8

遺伝情報の統合処理支援モジュールの構築

田中 剛範† 山本 雅人† 三田村 保† 大内 東† 大柳 俊夫‡ 松嶋 範男‡
北海道大学工学部† 札幌医科大学保健医療学部‡

1. はじめに

本研究では、表現形式の異なる複数の遺伝配列情報データベース(DB)から抽出した情報を共通の形式に変換し格納する為の内部変数領域と、それらのデータの処理を行ない二次的な情報の抽出･生成を支援する為の基本操作関数との集合体であるオブジェクト指向型モジュールを構築する。任意の遺伝情報処理支援ツールを、本モジュール内に定義された操作関数の組合せで比較的容易に構成できる。

2. 概要

本モジュールは構成要素として以下のものを含む。(図１を参照)

2.1 基本情報の記憶領域 (basic elements)

入力として取り込んだDBエントリから基本情報(配列の実体、登録番号、名前、生物種、引用文献情報、キーワード、他DBとの関連、等)を抽出し、各要素ごとに分離した状態で内部変数上に格納する。

2.2 基本操作関数群

後に個々の目的に応じて作成する二次情報抽出や加工の為の関数の内部において、基礎的な部品として機能する。

入出力関数 (input functions/output function)

DBからの入力を受け付け、必要箇所のみを抽出してDB毎の表現形式の違いをなくし、モジュール内に格納する。また、加工後のデータを出力する。

[以下は基本情報を直接利用する関数である] (FF)

検索関数： 現在モジュール内に格納されているエントリが、パラメータとして与えられた条件(例：登録番号、生物種、文献情報、キーワード、配列パターン、等)を満たすかどうかを判断する。

性質判定関数： 配列要素である塩基やアミノ酸の持つ生物学･化学的性質を数値あるいは記号等で返す。

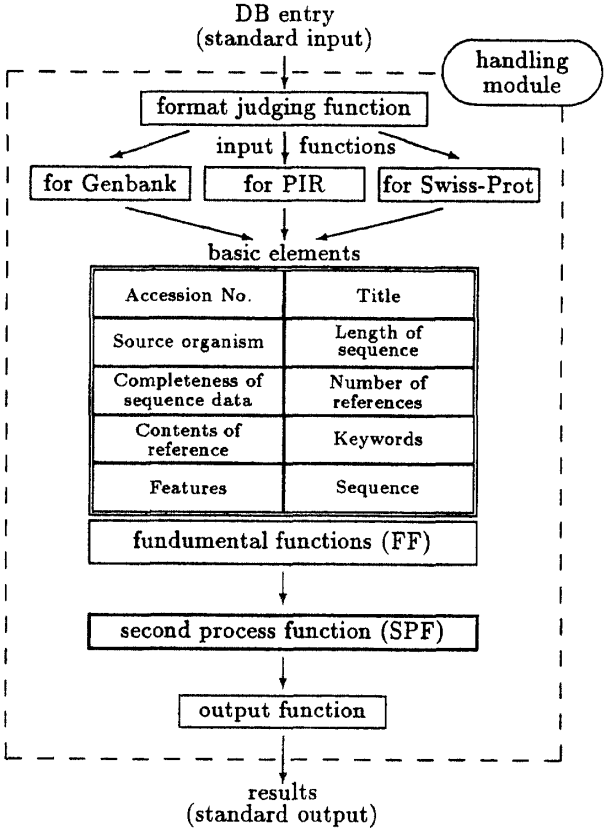

図 1: 統合処理支援モジュールの基本構成

配列結合関数(塩基配列)： １つまたは複数のエントリの配列からイントロン(非コード領域)を除きエクソン(コード領域)を繋ぎ会わせて、１本のアミノ酸配列に翻訳する為の準備を行う。

翻訳関数(塩基配列)： 指定された範囲を、または自動的に開始コドン及び終止コドンを探索しながら塩基配列をアミノ酸配列に翻訳する。

アラインメント関数： アミノ酸配列同士 or 塩基配列同士のアラインメントを行う。

2.3 二次処理関数 (SPF)

前述の基本関数群とモジュール内に取り込んだ基本情報の各要素その他を利用して、使用者が独自の目的に応じて構築する。従来のようにDBのエントリそのものから情報を引き出す手間を考えずに済み、取扱いが容易である。

Design and Implementation of Genomic Sequence Information Handling Module
†Takanori TANAKA, Masahito YAMAMOTO, Tamotsu MITAMURA, Azuma OHUCHI
‡Toshio OHYANAGI, Norio MATSUSHIMA
†Faculty of Engineering, Hokkaido University
‡School of Health Sciences, Sapporo Medical University

3. 実装

3.1 入出力方式

DBは通常大量のファイルの集合であるが、複数のエントリが1つのファイルにまとめられていたり、あるいは圧縮された形式で保存されてることもあり、モジュール側でそれぞれの形式全てに対応するのはコストが大きい。そこで、本モジュールへの一次データの入力は一貫して標準入力を通して行うものとし、DB上の情報は一度別のコマンドやプログラムによってその出力をモジュールの標準入力に流すものとする。

3.2 クラス定義

本モジュールでは1つのメインクラスと2つのサブクラスをを定義して使用している。

- モジュール主記憶クラス (メインクラス)

  2.1 節で提示した基本情報の格納領域および後述の2つのサブクラスに対するポインタを持ち、2.2 節で提示した各データ処理関数群を定義している。

- エントリデータ記憶クラス (図 2)

  このクラスは標準入力からのエントリデータを行単位のリスト構造に変換し、記憶する。これによりファイルと同様に行単位での識別作業が可能になり、しかもファイル操作より高速に処理を行える。

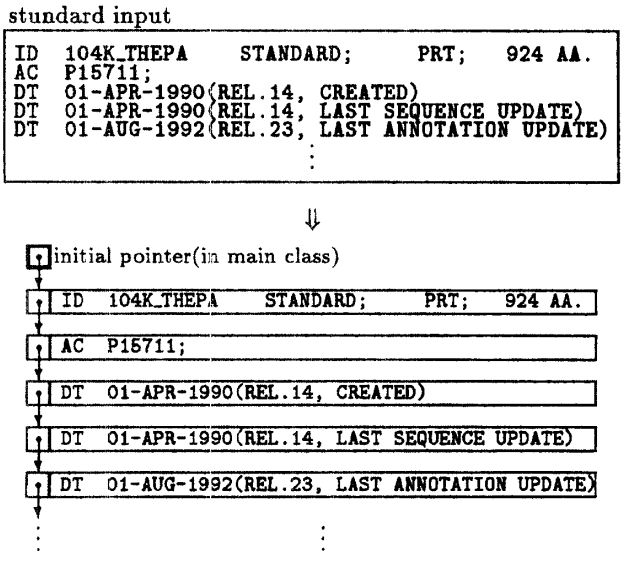

図 2: エントリデータ記憶クラスのデータ構造

- 文献情報記憶クラス (図 3)

  エントリ内の文献情報 (著者、題名、誌名、他) を記憶するためのクラス。動的メモリ管理に対応するためリスト構造をとる。

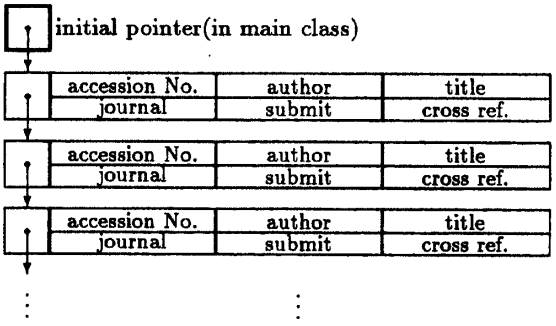

図 3: 文献情報記憶クラスのデータ構造

3.3 オブジェクトのメモリ管理

1件分のエントリデータを専用クラスに記憶した後、そこから各基本情報を抽出し、メインクラス内にあらかじめ用意された静的変数に格納するが、遺伝情報配列の実体や文献情報等の様にその情報量がエントリ毎に大きく異なるオブジェクトの記憶領域に関しては、動的メモリ管理を利用して1回の読込み毎に適切な容量を確保している。

概略を以下に示す。

1. 格納しようとするオブジェクトの大きさを求める
2. 必要な大きさのメモリを確保し、専用のポインタに結びつける
3. ポインタを通して実際のデータを代入する
4. 次のエントリを読み込む直前に、不要になったメモリは解放する

4. 評価

本モジュールに対する入力は一系統を念頭においており、DBからの入力に対するフィルタの役割を果たすようなツールならば、基本操作関数の組合せ+αによって比較的容易に構築することが可能である。筆者等はこのモジュールを利用して作成したツールを用いて、陽イオンチャンネルタンパク質にある種の特徴的なパターン配列が存在する事を見い出し、その解析を行ないモジュールの有効性を確かめた。[1]

5. おわりに

今後、実際のDB検索作業を通して改良の余地のある箇所を洗い出し、遺伝情報処理支援ツールとして現存する遺伝情報DBにない有意義な特色を出すこと、特に塩基-アミノ酸配列の対応や、引用文献に関する情報の取り扱いに重点を置きたい。

参考文献

[1] 田中 剛範, 山本 雅人, 三田村 保, 大内 東, 大柳 俊夫, 松嶋 範男: "陽イオンチャンネルタンパク質中の特徴配列の解析", 平成8年度 電気関係学会北海道支部連合大会講演論文集, p. 25 (1996).