モチーフ抽出実験システム

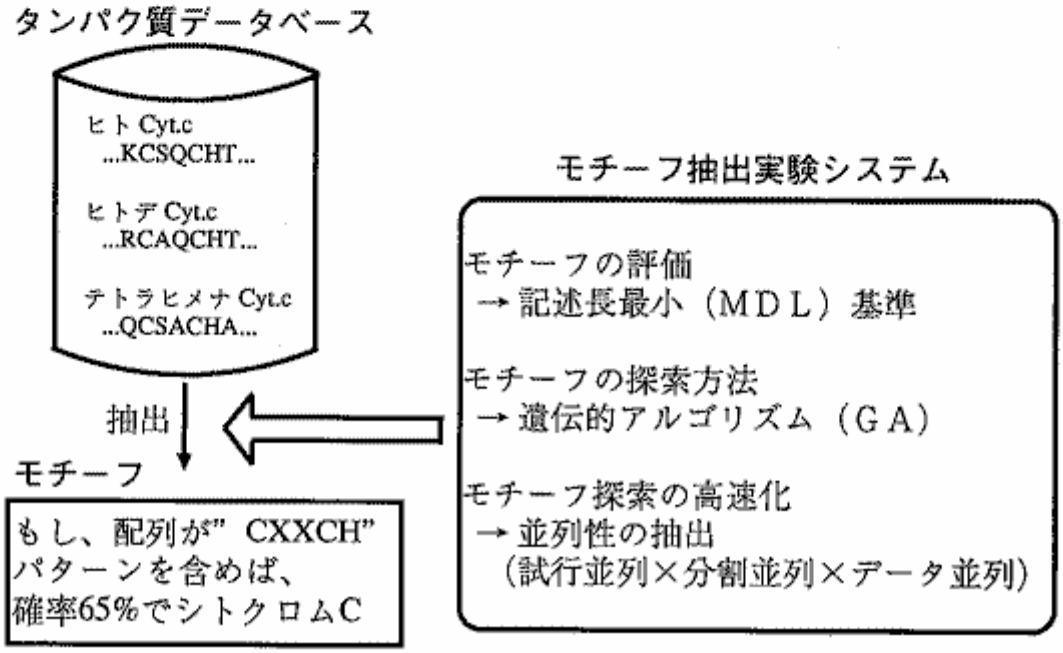
概要

分子生物学の分野の重要なテーマであるタンパク質の配列モチーフ抽出を行なう実験システムをＰＩＭ上に構築した。大規模な探索空間を持つタンパク質モチーフ抽出問題において、記述長最小基準と遺伝的アルゴリズムを用いた確率的探索手法が高い並列性を有し、ＰＩＭ上で効率的に実行できることを示す。

特徴

モチーフ抽出実験システムは、タンパク質データベースＰＩＲに含まれるタンパク質を対象に、シトクロムＣなどの特定のタンパク質を識別するためのモチーフを自動的に抽出する実験システムである。本システムでは、データに混在するエラーや分類エラーの問題を解決するために、モチーフを例外事象を含む確率的規則として扱う。モチーフ抽出実験システムの特徴は次の通りである。

1）モチーフ評価に記述長最小（ＭＤＬ）基準を採用し、抽出されるルールの現データベースへの過剰適合を防止。
2）ルールの学習に確率的探索アルゴリズムである遺伝的アルゴリズム（ＧＡ）を利用し、計算時間の組み合わせ的爆発を回避。
3）試行並列、分割並列、データ並列の３種類の並列性を持つ並列ＧＡにより、ＰＩＭの高並列性をフルに活用。



モチーフ抽出実験システムの構成

1 モチーフ抽出問題

モチーフとは、同種のタンパク質に共通して見られるアミノ酸配列パターンである。モチーフはタンパク質の機能、構造を特徴づけ、進化の過程でも保存されてきたと考えられている。代表的なモチーフの例をあげる。

ヘム結合部位　　CXXCH

ロイシンジッパー（ヘリックス同士の結合）　　LX6LX6LX6LX6L

C：システイン　H：ヒスチジン　L：ロイシン

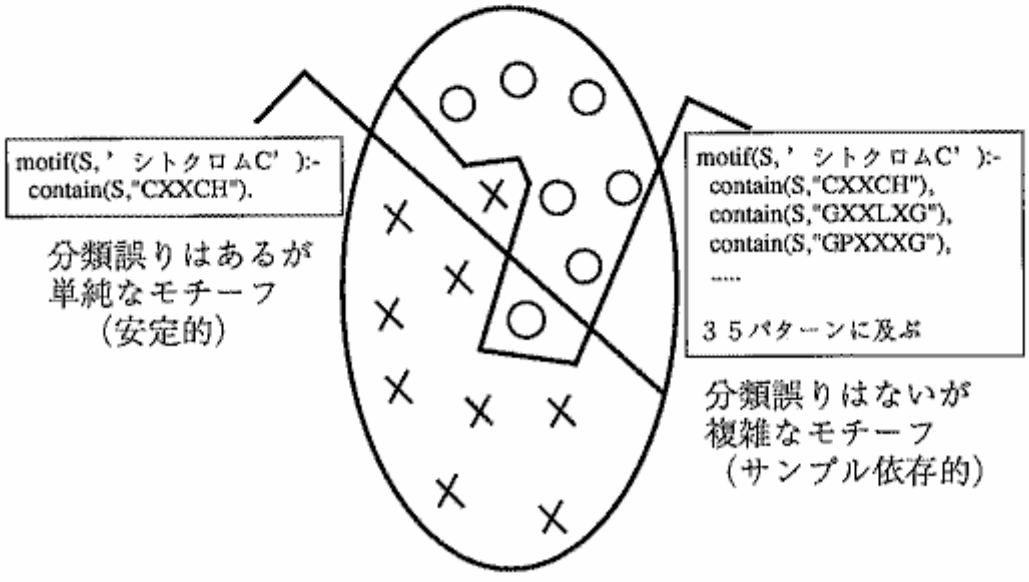
X：任意のアミノ酸　X6：XXXXXX

2 記述長最小基準（MDL）によるモチーフ評価

モチーフの評価基準として、MDL基準を採用した。MDL基準は、次式で与えられる記述長が小さいモチーフをより良いモチーフと考える基準である。

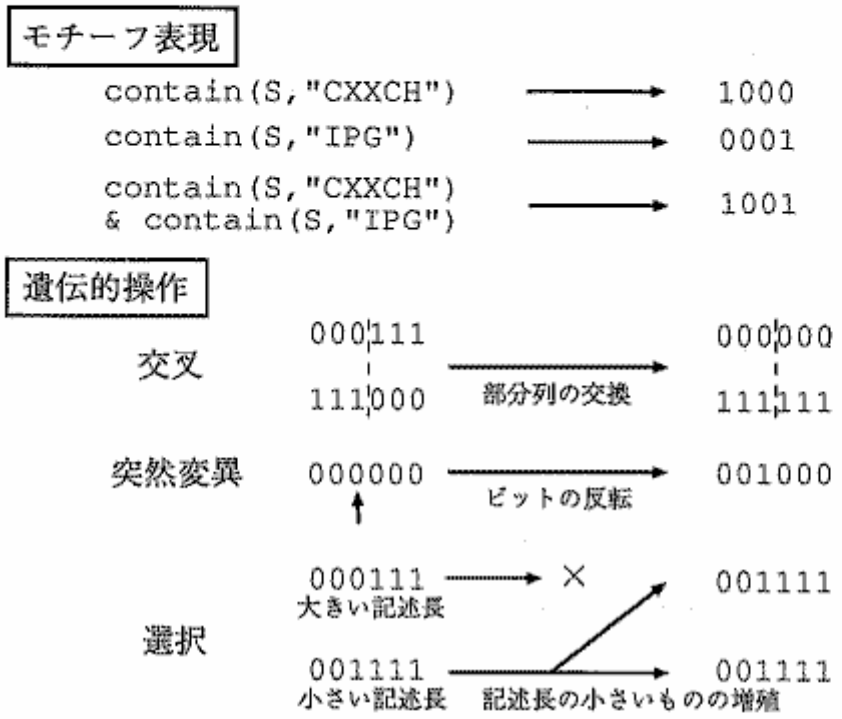
記述長 ＝ モチーフの複雑さ ＋ 分類誤りの程度

MDL基準は、下図のような、分類誤りはあるが単純なモチーフと分類誤りはないが複雑なモチーフとの比較基準を与える。



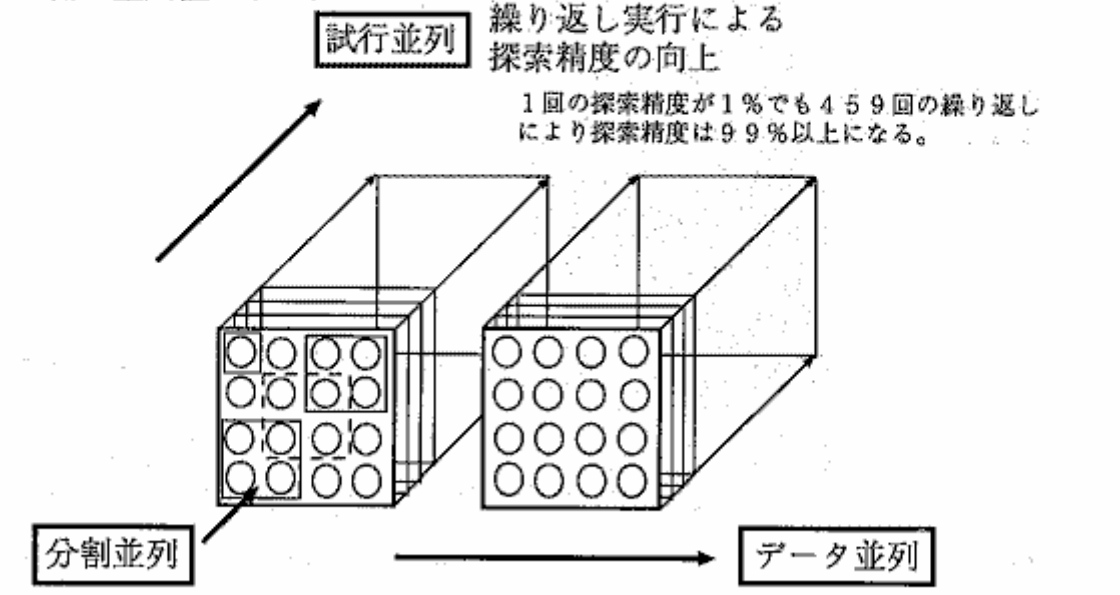
3 遺伝的アルゴリズムによるモチーフ探索

モチーフの探索手法として、生物の進化過程をヒントに考案された確率的探索アルゴリズムである遺伝的アルゴリズム（ＧＡ）を採用した。ビット列で表現されたモチーフに対して、交叉、突然変異、選択の３種類の遺伝的操作を繰り返し適用し、ＭＤＬ基準の意味で良いモチーフを抽出する。



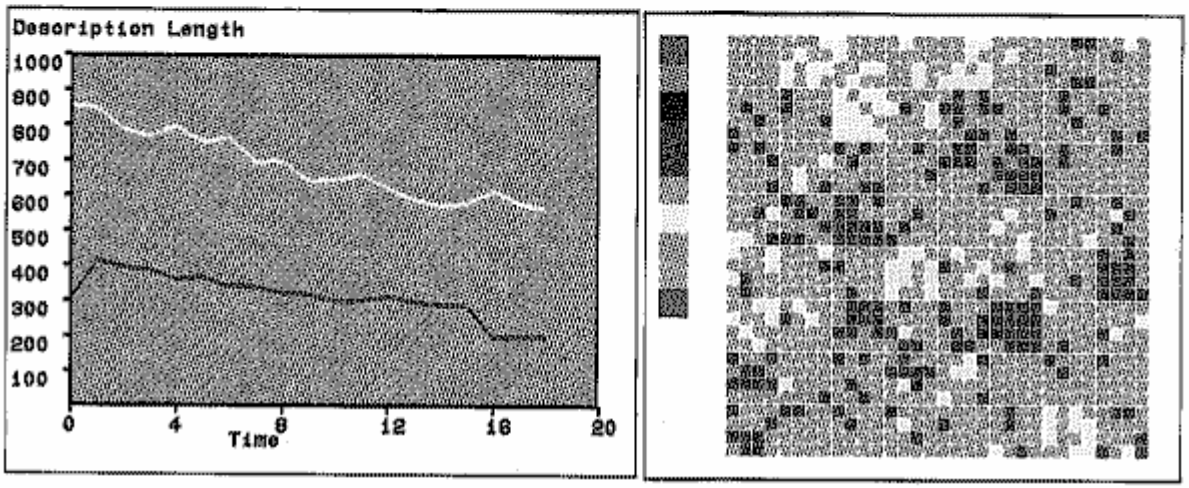
4 遺伝的アルゴリズムによるモチーフ抽出の並列性

遺伝的アルゴリズムによるモチーフ抽出には、試行並列、分割並列、データ並列の３種の並列性がある。

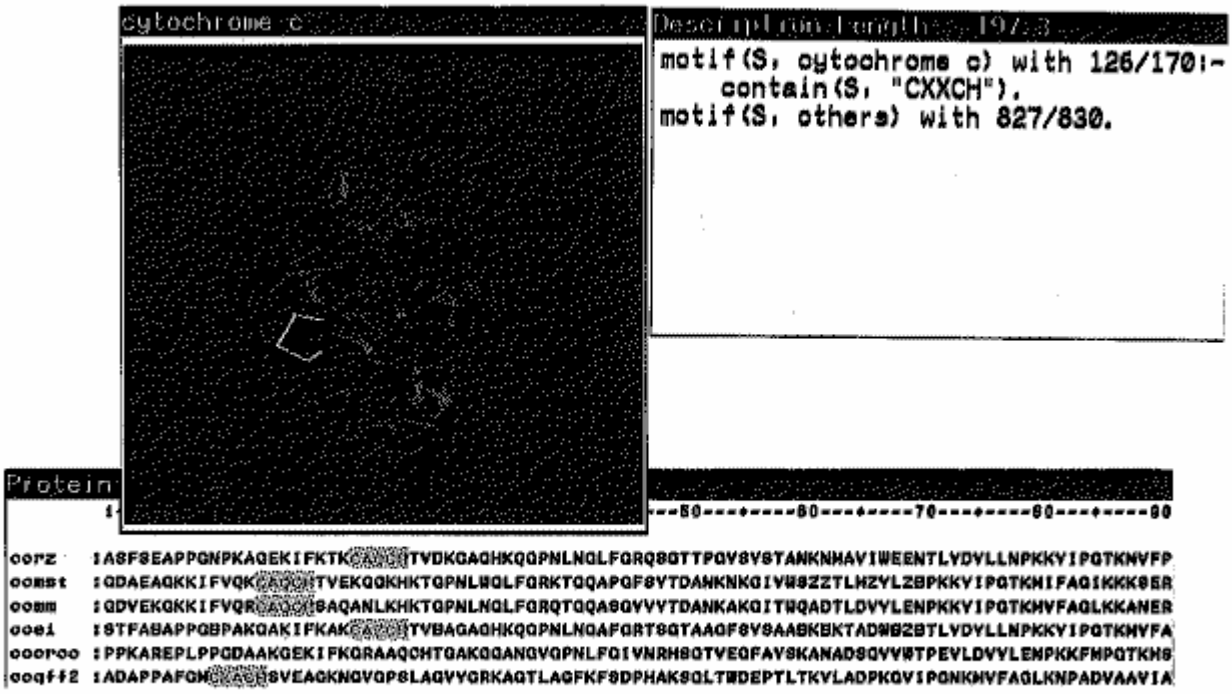


5 デモ概要

モチーフの抽出条件を設定し、ＰＩＭ上で実際にモチーフ抽出実験を行なう。モチーフ抽出の実行中には、ＰＩＭ上の各プロセッサ上に存在するモチーフ候補の記述長が時間とともに減少していく様子を、モザイク状のカラー表示と折れ線グラフ表示でリアルタイムに示す。モチーフ抽出の結果は、確率的決定述語の形で表示する。この際、抽出されたモチーフのアミノ酸配列上での位置、タンパク質の立体構造中での位置も合わせて表示する。モチーフ抽出の途中経過と抽出結果の画面例を以下に示す。



画面例１：モチーフ抽出の途中経過



画面例２：モチーフ抽出結果