Experimental Motif Extraction System

Abstract

This system proves the effectiveness of the Parallel Inference Machine(PIM) for the motif extraction problem, which extracts common patterns(motifs) from s protein database. This uses the minimum description length principle and genetic algorithms.

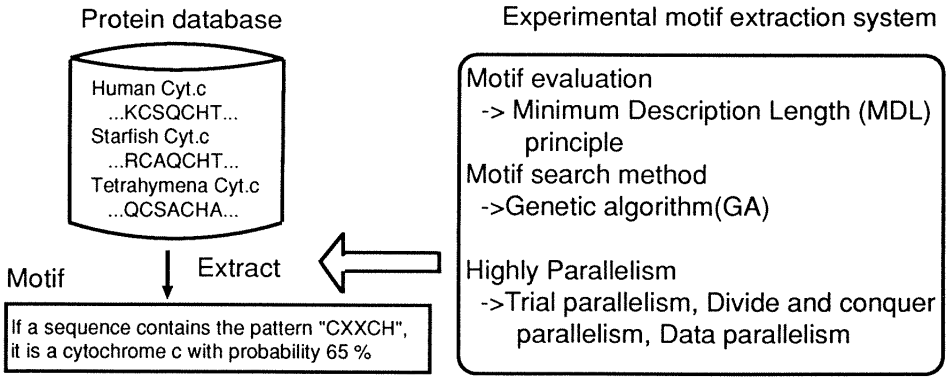
Features

The experimental motif extraction system automatically extracts common patterns in some protein categories, such as cytochrome c. The system regards a motif as a stochastic rule to deal with exceptions to the classification of proteins.

1) The Minimum Description Length (MDL) principle was adopted as a criterion for motif evaluation to avoid motif's overfitting to sample data.

2) Genetic Algorithms (GA) were employed as a motif search method to reduce the effects of the combinatorial explosion and to reduce search time.

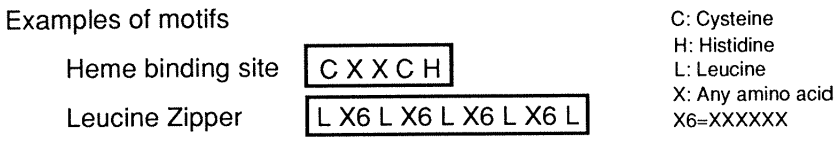
3) Highly parallelism on the PIM was achieved by exploiting trial, divide-and-conquer and data parallelism.



Configuration of Experimental Motif Extraction System

1 Motif extraction problem

Motif extraction is one of the important problems in genetic information processing. It extracts common patterns(motifs) from amino acid sequences of the same protein category, which are conserved in the evolution process and characterize the function/structure of proteins.

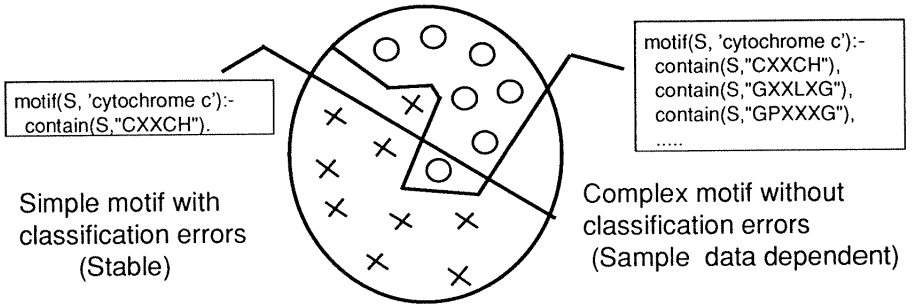


2 Motif evaluation by the MDL principle

As a criterion for motif evaluation, the minimum description length(MDL) principle was adopted. The MDL principle selects a motif with the shortest description length defined below.

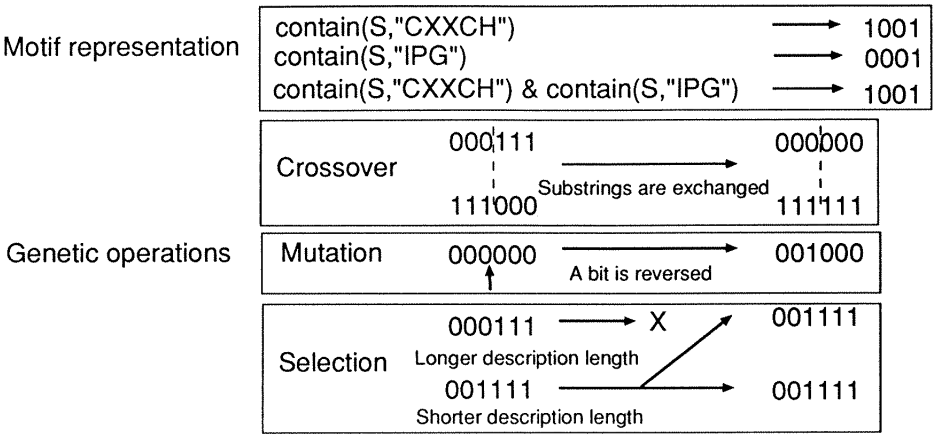
Description length = Complexity of motif + Classification error rate

The MDL principle enables us to compare a simple motif with exceptions and a complex motif without exceptions as illustrated below.



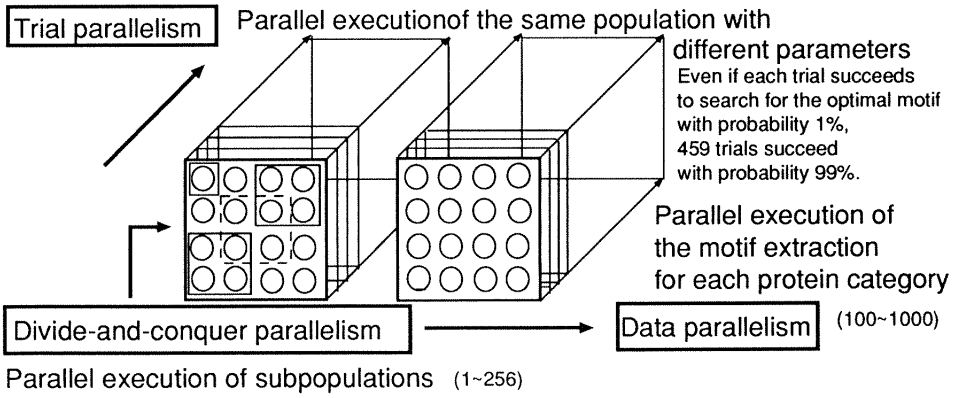
3 Motif search by genetic algorithm

As a motif search method, genetic algorithms (GA) were adopted. GA realizes probabilistic search by applying genetic operations to a population of motif candidates represented by binary strings. The genetic operations consist of crossover, mutation and selection. The MDL principle plays an essential role in selecting good motif candidates (shorter is better).



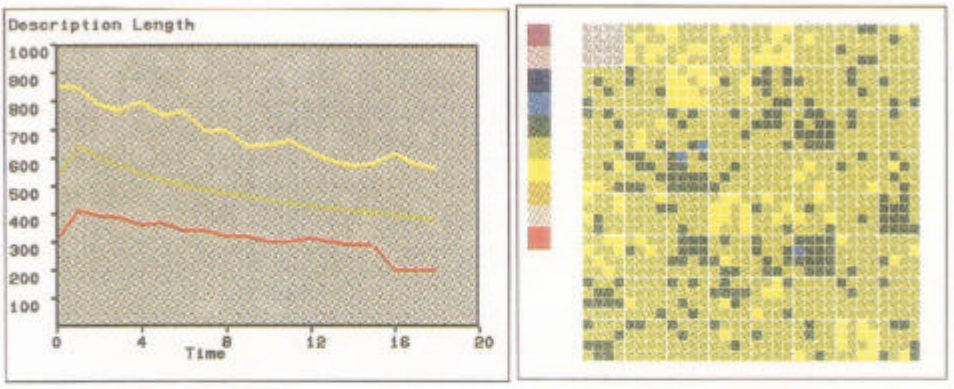
4 Parallelism in the motif extraction system

Three kinds of parallelism can be exploited in the motif extraction system; trial, divide-and-conquer and data parallelism.



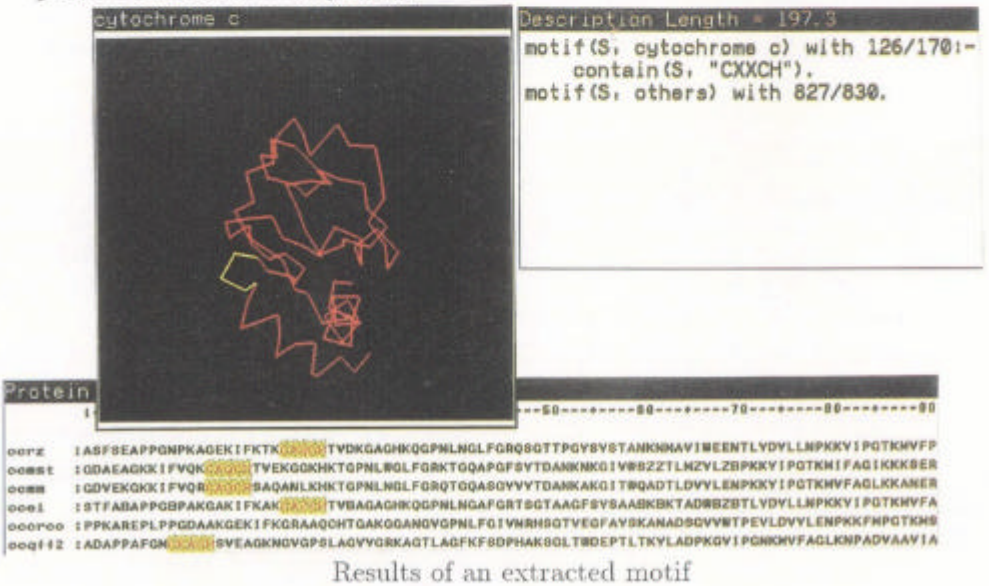
5 Outline of Demonstration

To enable monitoring of the runtime behavior of our motif extraction system, an on-line monitor and an on-line tracer are provided. The monitor displays the current description length of each motif candidate in the population. The tracer displays the lowest description length, the average description length and the highest description length of each generation.



Snapshot of the on-line tracer(left) and the on-line monitor(right)

The extracted motif can be shown by a form of stochastic decision predicates, the positions in protein sequences and the position of three dimensional protein structure, if they exist.



Results of an extracted motif