A Structural Sequence Representation of Protein: Secondary Structure Element Sequence (SSES)

Sung-Hee Park  Chan-Yong Park  Dae-Hee Kim  Seon-Hee Park  
{sunghee, cypark, dhkim98, shp}@etri.re.kr

Bioinformatics Research Team, Electronic and Telecommunication Research Institute, Daejeon, Korea

Keywords: protein structure comparison, sequence secondary structure

1 Introduction

Many protein representations and descriptors have been proposed to be used to describe proteins. There are largely two representations: sequence and structure. Sequence representation has no structural information as structure representation has no sequential information. We’d like to represent sequential and structure information of protein. So, this paper proposes a sequential and structural protein representation: secondary structure element sequence (SSES).

On the early days, the representation of protein structure consists of positions of all “atoms” consisting of protein and their intra-molecule distance. This representation had critical weak points in terms of computational complexity and sensitivity of errors. The representation is simplified to solve the weaknesses.

This representation is improved by considering the position of alpha carbons instead of calculating the position of all atoms. Furthermore, recently there was a research that had improved the weak points of sensitivity of errors and computational complexity. The research divides a protein to several segments uniformly and then compares two protein structures with the average positions of the alpha carbons consisting of a segment. And the other research approach is to represent a protein structure to vectors of protein secondary structures. Then it measures a similarity with the vector representation.

As there has been simplification of representation in structural representation of protein, this simplification may be applied to amino acid sequence. Secondary structure elements (SSEs) are a simplification of protein structure. Protein amino acid sequence can be reduced to sequence of SSE. In this paper, we represent proteins as sequences of SSEs, and then measure their similarity to classify them.

2 Method and Results

In this paper, to represent protein structure, we use a protein structural sequence representation SSES (Secondary Structure Elements Sequence) that means a sequence of protein secondary structure. This representation is an abstraction of amino sequence as SSE is a structural abstraction of protein. And we designed and implemented a protein structure classification system using its sequence

2.1 Protein Structure Comparison Model using a sequence of protein secondary structures

The flowchart of protein structure retrieval system is given in Figure 1.
First, SSE is found to simplify structural information of protein. SSES is formed by sequencing SSE. In calculating protein similarity, this system can easily use sequence alignment algorithms, for an example, an edit distance with dynamic programming. For clustering, general unsupervised classifiers like K-means or Fuzzy C-means can be used in the classification process.

2.2 classification Result
An example of retrieval result is shown in Figure 2. As you can see the figure, the 4 proteins are similar to each other for their SSESs.

![Protein Structure Data (PDB files)](image)

The result of classification

3 Conclusions
In this paper, to represent protein structure, we have used SSES. This representation is an abstraction of amino sequence as SSE is a structural abstraction of protein. And the representation is also a sequential and structural protein representation. We have shown this usefulness of this representation by applying an application of classification. Because the protein structure classification system using SSES shows fast rough classification with relatively precise result, it can be used as high level classification criteria or measure in hierarchical classification scheme.

References