CALIFORNIA STATE SCIENCE FAIR
2007 PROJECT SUMMARY

Name(s)

Dhruv Garg; Charles Xue

Project Number

S1304

Project Title

PrevCor: An Algorithm Using Amino Acid Properties and Protein Folding Patterns to Simulate the 3-D Folding of Proteins

Abstract

Objectives/Goals

The design and construction of a protein folding computer algorithm that maximizes accuracy for biotechnology researchers and scientists to perform flexible protein sequencing simulations based on genetic data.

Methods/Materials

We created a UI which converts user-inputted DNA sequences into an amino-acid sequence; this is stored in a lattice-modeled doubly-linked-list data structure, in which each node represents an amino acid. The structure between each node is part of a separate data structure overlaid on a wire frame, which provides the angular orientation and interaction type as enumerations. The interactions between individual nodes are based on probabilistic protein-folding equations derived from each amino acid's hydrophobic intensities, acidity properties, and relative frequencies of occurrence (alpha helixes, ß-sheets) in a polypeptide chain. We used Java and Java3D API in the Visual Studio Professional Edition development environment to construct this algorithm.

Results

The computer simulation was compared to the actual placement of amino acids in proteins such as insulin, hemoglobin, and pepsin. Our simulated proteins had an accuracy % of 98.039%, 87.805%, 81.308% respectively. Accuracy was determined by comparing the relative locations of amino acids to their actual lab-based models under a pH of 7.0. Under deviating pHs, due to the random instance algorithm used, accuracies ranged from between 2.0 x 1.9^|7.0-pH| - 4.0 x 1.9^|7.0-pH|. CPU usage and processing power in relation to the number of amino acids on a 2.0GHz Dual-Core processor are as follows (# of amino acids, CPU %): (10, 1.16%), (50, 2.11%), (100, 4.43%), (200, 19.64%), (300, 87.06%).

Conclusions/Discussion

As expected, larger protein sequences have a greater toll on the processor, which lowers accuracy. This is caused by the exponential increase in the number of interactions among the secondary and tertiary structures of the protein. An accuracy of >80% provides a protein with enough structural similarities to give a detailed analysis on its attributes, thus the simulation is effective up to 325 amino acids. Similarly, CPU usage is also based on an exponential curve and the CPU vs. # of amino acids curve closely fits the equation 1.015^A where A represents the # of amino acids. In the future, we plan to analyze much longer and data-intensive polypeptide chains by employing a supercomputer to manipulate data.

Summary Statement

A computer algorithm that predicts and simulates protein-folding using a lattice-model doubly-linked-list data structure and probabilistic equations derived from hydrophobic intensities, pK(a) data, and structural tendencies of amino acids.

Help Received

Mr. Bruce Kawanami (Monta Vista High School, Cupertino ROP Eng. Tech Teacher) was our project adviser. Everything else has been done by us independently. We also consulted various online resources and the details are provided in the bibliography of our documentation.