P5: Using Genetic Programming to Design a New System that Enables Discovery of Novel Caspase3 Substrates

Ala' Jabari[1], Hashem Tamimi, Yaqoub Ashhab

[1]Biotechnology Research Center, Palestine Polytechnic University, P.O.Box 198, Hebron, Palestine

Introduction

Caspase3 is a protein that plays a critical role in programmed cell death process (apoptosis). The key role of caspase3 is to cleave proteins that are critical for cellular integrity. Caspase3 substrate proteins are normally cleaved after aspartic acid residues (D). There is a growing interest in studying and discovering the molecular mechanisms that regulate apoptosis due to its significant impact in understanding various diseases such as cancer, AIDS and neurodegenerative disorder. Despite the availability of enormous amount of biological data about most mammalian proteomes, little information is known about the global profile of caspase3 cellular substrates. Accordingly, development of accurate and cost-effective technique is needed to perform high-throughput screening on whole proteomes to physico-chemical properties discover novel substrates.

Genetic Programming (GP) is an emerging machine learning techniques that is considered as a very promising method to generate software. It aims at finding computer programs that perform a user-defined task. The attractiveness of Genetic Programming is in its ability to tell the computer what to do rather than how to do. The use of GP in computational biology field has demonstrated a very strong capability to perform automatic feature selection of the biological sequences. In this paper Genetic Programming has been applied to predict the presence and location of caspase 3 cleavage sites in protein sequences.

Material and Methods

In a previous work, we have collected one hundred and fifty caspase 3 substrates with experimentally mapped cleavage sites [2]. The full protein sequences of these substrates were downloaded from UniProt [4]. The cleaved motifs were extracted as subsequences of 14 amino acid length, where 8 amino acids before and 5 after the cleaved aspartic acid residue were included in the obtained peptide. Negative data (non cleaved Ds) were generated by extracting protein subsequences outside the cleavage site where the ninth amino acid is also an aspartic acid residue.

Two different GP experiments were conducted. In the first experiment, each amino acid was converted into its equivalent ASCII numerical value and a set of logical functions, F1={and, or, equals, if-else} were used as a set of operations on the amino acids. In the second experiment, numerical indices of different amino acid physico-chemical properties were used to represent each amino acid. The physico-chemical properties were retrieved from APDBase [1]. Another set of functions F2={ plus, minus, times, cosine, sine, log} is used as operations on the physico-chemical properties. In both experiments, we used the same GP parameters in order to conduct fair comparison between the experiments. So we used 1000 individuals and 50 generations. The sum of absolute differences between the obtained and expected result was used as fitness function for both experiments.

Because the data size was relatively small, we used leave one out as a cross validation to estimate the performance of our model, where one sample is taken out each time from the whole set of data to be used for testing purposes. The process is repeated until all samples are used as testing data. We calculated the success rate by dividing the sum of true positive and true negative over the whole data set.

Results

Here we show the use of two different systems; ASCII-based index and physic-chemical properties-based indices tested on the positive and negative data using the leave-one-out

approach for performance validation. The results obtained using the two approaches are summarized in Table 1.

Experiment	ASCII-based index	Physico-chemical properties indices
Accuracy rate (min-max)	87.12% - 89.43%.	80.01% - 83.87%.

Table 1: Results obtained using leave one out cross-validation

From the result obtained we can see that there is a slightly higher success rate using the first method. Our results showed that GP-based classification method can provide a platform to develop cost-effective and powerful prediction tools that can be used for a wide range of pattern recognition applications in biological sciences.

References

Amino acid Physical-chemical property Database [online], Available: www.rfdn.org/bioinfo/APDbase/APDbase.php.

Ayyash M. and Ashhab Y. (2008) Developing a Powerful Bioinformatics Tool for Prediction of Caspase 3 Substrate: Preliminary Analysis of the Human Proteome. International Biotechnology Symposium, Sfax, Tunisia, 4-8, 2008.

Song J., and et al. (2010). Cascleave: towards more accurate prediction of caspase substrate cleavage sites. Bioinformatics (26): 752–760

UniProt, [online]. Available: www.uniprot.org.