MASSP: A hybrid genetic-neural system for predicting protein secondary structure

Armano G.(1), Mancosu G.(2), Orro A.(1), Saba M.(1), and Vargiu E.(1)

(1) DIEE - Dept. of Electrical and Electronic Engineering, University of Cagliari Piazza d'Armi, I-09123 Cagliari, Italy (2) Shardna Life Sciences, Piazza Deffenu 4, I-09121 Cagliari, Italy

Motivation

Being the prediction of protein structure a very complex task, most methodologies concentrate on the simplified task of predicting secondary structures. In this paper, we illustrate a technique based on multiple experts, aimed at predicting protein secondary structures. The prediction activity results from the interaction of a population of experts, each integrating genetic and neural technologies. Roughly speaking, an expert of this kind embodies a genetic classifier designed to control the activation of a feedforward artificial neural network for performing a locally-scoped prediction activity. Genetic and neural components (i.e., guard and embedded predictor, respectively) are devoted to perform different tasks and are supplied with different information: Each guard is aimed at (soft-)partitioning the input space, insomuch assuring both the diversity and the specialization of the corresponding embedded predictor, which in turn is devoted to perform the actual prediction.

Methods

The technique described in this paper is an implementation of the generic NXCS architecture (standing for Neural XCS), which has been explicitly customized for predicting protein secondary structure. The most relevant component of an NXCS system is its population P of experts, each being able to interact with its operating environment. Experts interact with: (i) a selector, (ii) a combination manager, (iii) a rewarding manager, and a (iv) a creation manager. The selector is devoted to collect all experts whose guard covers the given input, thus forming the match set. The combination manager is entrusted with combining the outputs of experts belonging to the match set, so that a suitable voting policy can be enforced on them. The main task of the rewarding manager is forcing all experts in the match set to update their fitness, according to the reward obtained by the external environment. The creation manager is responsible for creating experts, when needed. In its basic form, each NXCS expert E can be represented by a triple , where: (i) g is a binary function that selects inputs according to the value of some relevant features, (ii) h is an embedded expert whose activation depends on g(x), and (iii) w is a weighting function used to perform output combination. Hence, the output of E coincides with h(x) for any input x that matches g(x), otherwise it is not defined. In the case E contributes to the final prediction (together with other experts), its output is modulated by the value w(x) of its weighting function, which represents the expert strength in the voting mechanism. This value may depend on several features, including the input x, the overall fitness of the corresponding expert, and the reliability of the prediction made by the embedded expert. Typically, the guard g of a generic NXCS classifier is implemented by an XCS-like classifier, able to match inputs according to a set of selected features deemed relevant for the given application, whereas the embedded classifier h consists of a feed forward ANN, trained and activated on the inputs acknowledged by the corresponding guard. In the task of predicting secondary structures, genetic guards are entrusted with processing some biological features deemed relevant for the given task. The "AAindex" database has been used for retrieving information about

hydrophobicity, dimension, charge and other features required for evaluating the given metrics. In the current implementation of the system eight domain-specific metrics have been devised and implemented. A sample metrics is: "Check whether hydrophobic amino acids occur in a window of predefined length according to a clear periodicity", whose underlying rationale is that sometimes hydrophobic amino acids are regularly distributed along alpha-helices.

Results

To perform experiments, we used a subset of the PDBSelect database. Embedded predictors, implemented by MLPs, are characterized by a unique hidden layer of 10-25 neurons, depending on the amount of inputs that can be selected by the corresponding guards. Each MLP is able to encode 15 contiguous residues (centered on the residue to be predicted), and outputs three values that represent pseudo-probabilities associated with the adopted secondary structure labels (i.e., alpha helix, beta sheet, and coil). Experiments have been devised to evaluate the impact of genetic selection in the performance of the system. The overall population has been set to 600 experts, with an average of about 20 experts involved in the match set. Experimental results (74.8%) are comparable with other state-of-the-art systems, also taking into account that no post-processing is performed on the predicted secondary structure and that the quality of the currently-adopted metrics has not been assessed by a biologist.

Contact email: giuliano.armano@diee.unica.it