

Protein identification automation: a script to search multiple sequences using a single program

Oliveira, C.C.S.; Rodrigues, H.S.; Castro, M.S.; Fontes, W.

Laboratório de Bioquímica e Química de Proteínas, Depto. de Biologia Celular, Universidade de Brasília, DF.

Protein identification can be time consuming for the researcher. Often the user starts from such a large number of sequences that only the data submission by manually filling web forms could take an entire day. Some search engines allow the submission of files containing multiple sequences; however they are few and limited. With the aid of a simple program, this search can be automated, allowing for users with no programming knowledge to improve their searches.

The goal of this script is to allow the user of a search mechanism to submit many sequences automatically, demanding the minimum of time from the user.

The core of the program is a small PHP script, inserted in a HTML search form. Search parameters are normally fixed: the researcher uses the same parameters for a set of searches. Thus, the only field that varies corresponds to the sequence. The first function of this script is to store the sequences, obtained from a text file, in a small array for posterior submission, where each element of this array corresponds to a sequence.

With the form filled, a JavaScript function directs the search order to the corresponding service, and a repetition loop allows the sending of all sequences supplied by the user in the text file. The JS function is in a hidden frame, controlling the sending of the data in another frame, visible to user, containing the search form.

The result is a working script, running as both server-side and client-side that allows researchers to perform searches using a single program, based on multiple sequences. A search of 1000 sequences can take up to 50 hours to be carried through using the non-automated way. This script allows the automation of this task, demanding only 5 minutes of the user.

This is a pilot system that integrates a part of a broader project of protein identification automation based on searches for sequence similarities. Combining PHP and JS, a very practical result has been obtained, so that researchers, even without knowing programming languages, have the possibility to get the results quickly.

Supported by UnB, PIBIC, CNPq and FINATEC.