A Graphical Tool for Protein Sequences Analysis

Armano G, Saba M, Vargiu E

DIEE - University of Cagliari, Piazza d'Armi, I-09123 Cagliari, Italy

Motivation

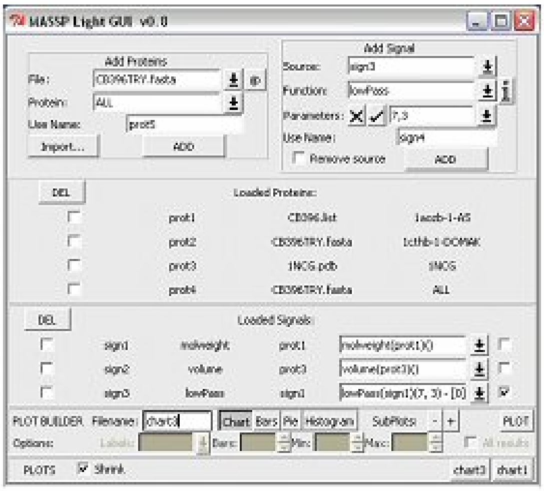
Nowadays, one of the most relevant problems in bioinformatics is how to manage the increasing amount of data, empirically produced by researchers involved in life sciences. In fact, in the last few years, an enormous quantity of information has been produced in these fields although the amount of structural information is much more greater than the functional one. To support researchers in discovering the organizational principles and the functional relationships that characterize biological systems, several tools and systems has been devised and implemented. In this paper, a graphical tool for protein sequences analysis is presented, aimed at supporting the user in performing statistical analysis of proteins, in particular at highlighting the relationship between primary and secondary structure.

Methods

Our work is driven by the underlying assumption that the information about a protein sequence can be spread over a set of numerical signals all originating from the given sequence. In this paper we present a graphical tool that allows the user to devise and test (a pipeline of) suitable filters aimed at highlighting specific properties deemed relevant by the user. The results of this filtering activity can be easily plotted by resorting to standard display facilities (charts, histograms, etc..). It is worth pointing out that the proposed tool provides a specific support for studying the relationship between primary and secondary structure in terms of relevant features such as hydrophobicity, aromaticity, dimension, electrical charge, and so on. The system can analyze single proteins or a set of proteins. In the former case, given a protein P, the system supports two functionalities: (i) converting each amino acid in P according to a selected numeric property (e.g. electrical charge), and (ii) filtering P using a window flowing along it (e.g. counting for each amino acid in P the number of Glycines within the current window). Once that P has been converted into a numerical signal, any standard or user-defined filter can be applied to it (e.g. low-pass filtering), possibly giving rise to a pipeline of filters that progressively modify the sequence according to the user needs. The signal obtained after processing P can also be compared to the "profile" of a specific secondary structure, while attempting to highlight a correlation between the two signals. In the latter case, the system supports two functionalities: (a) a given pipelined operation can be applied to all proteins in the given set, possibly yielding a statistics about its capability of predicting a selected secondary structure, and (b) a specific patterns can be searched for, yielding a statistics about its occurrence in the given set of proteins, possibly focusing on its capability of predicting a specific secondary structure. Figure 1.1. shows the graphical interface of the proposed tool. Written in Python, it is completely "open", meaning that additional "ad-hoc" filtering operations can be integrated in a simple way.

Results

The system is currently under testing. Preliminary results highlight that it is particularly suitable for validating or not hypotheses about potential correlations between primary and secondary structure. For the sake of simplicity, we present a case study focusing on the capability of finding a specific pattern: the zinc finger motif. In this case, firstly the user gives as input the set of proteins to be analyzed (in FASTA, PDB, or XML format), and the pattern to be searched for, using regular expressions. Then, the system returns the match set using a textual and/or a graphical format (in this case a chart). Subsequently, among the overall set of proteins belonging to the match set, the user may choose a protein to be further investigated. For instance, the user could be interested in displaying the profile of the selected protein. The chart corresponding to this query is depicted in figure 1.2.



MASSP Light GUI v0.0

Add Proteins

File: CB396TRY.fasta

Protein: ALL

Use Name: prot5

Add Signal

Source: sign3

Function: lowPass

Parameters: 7,3

Use Name: sign4

☐ Remove source

Loaded Proteins:

☐	prot1	CB396.list	1aozb-1-AS
☐	prot2	CB396TRY.fasta	1cthb-1-DOMAK
☐	prot3	1NCG.pdb	1NCG
☐	prot4	CB396TRY.fasta	ALL

Loaded Signals:

☐	sign1	molweight	prot1	molweight(prot1)()	☐
☐	sign2	volume	prot3	volume(prot3)()	☐
☐	sign3	lowPass	sign1	lowPass(sign1)(7, 3) - [0]	☑

PLOT BUILDER Filename: chart3

Options:

PLOTS ☑ Shrink

Fig. 1.1 The system graphical interface

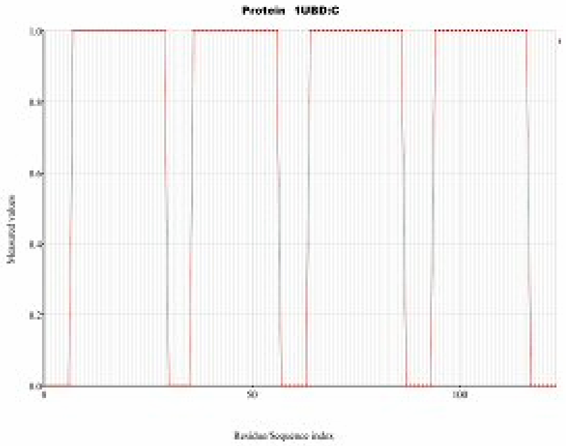


Fig. 1.2 The profile of the 1UBD:C protein

Contact email: vargiu@diee.unica.it