Identifying relevant positions in proteins by Critical Variable Selection

(S. Grigolon, S. Franz, M. Marsili, arXiv:1503.03815)

Silvia Grigolon [1,†]

[1] *LPTMS - Laboratoire de Physique Théorique et Modèles Statistiques UMR CNRS 8626, Univ Paris - Sud, 15, Rue Georges Clémenceau, 91405 Orsay CEDEX, France**

In the last decades development of better inference methods on data samples has been one of the central points of interest for the wide research community focused on complex systems.

Normally, the main idea is that of assuming a model to be the best one describing a system and exploiting correlation functions – defined in the most suitable way for a given dataset – to find out a sort of hierarchy in the interaction network arising out of a given data sample. The aim is then that of finding out whether there are and who are the key–players and finally what kind of interactions are set among them.

Here we present a natural application of a new method [M. Marsili et al., *J. Stat. Mech.*, P09003, 2013], renamed Critical Variable Selection, which does not take directly into account two-points correlations but is able to select the best ways for a set of variables to cluster, in the particular case maximizing an information amount defined on the clustering itself. In protein sequences, this leads for instance to the selection of a subset of sites turning out to be key–points not only along the primary but also along the tertiary structure of the protein itself.

The results obtained by the study of two protein families, Che-Y receptors and the ion channels, compared with those turning out from the application of Statistical Coupling Analysis, being a generalization of the Principal Component Analysis [V. Plerou et al., *Phys. Rev. Lett.*, **83**:7, 1471-74, 1999, N. Halabi et al., *Cell*, **138**:4, 774-86, 2009] show interesting overlaps between the two methods, being the new one also able to identify most of the tightest contacts in the proteins tertiary structures [F. Morcos et al., *PNAS*, **108**:49, E12093-301, 2011].

* † silvia.grigolon@gmail.com