Summary

Recent years have seen an explosion of availability of protein sequence data. However, the vast majority of these data are unlabeled, that is, the sequences are not accompanied by supplementary information about their functional or structural properties. In this perspective, the development of statistical methods which are able to leverage this huge availability of sequence data to try to unveil the sequence function/structure relation represents an interesting chance for scientists, and especially for biophysicists and computational biologists.

Among the statistical methods developed to tackle sequence data, a relevant role has been played by statistical physics inspired strategies, such as the generalized Potts model, where protein sequences are interpreted as vectors of q-states *spin* variables, to which a scalar *energy* function is associated. In this framework, the general idea is to use the sequence data to determine the model constituent parameters, as in an inverse-problem of statistical physics. Such techniques have proven to be particularly effective in the context of *multiple sequence alignments* (MSA), for the determination of structural properties and in predicting mutational effects. A fundamental requirement for these models to be highly predictive is that they have to be global, or alternatively stated, epistatic. The minimal choice to achieve such feature is considering pairwise interactions between the protein residues, as it happens in the case of the *Direct Coupling Analysis* approach.

In this thesis, we present some novel unsupervised inference methods which are inspired by the DCA approach, but with the aim to extend them to protein sequence data which are produced by laboratory experiments. Considering the short time scales that characterize these experiments, especially when compared to the natural evolution process, such data turn out to be inherently out of equilibrium. We believe that incorporating (at the least effectively) this dynamical information into the statistical model might be beneficial to infer more efficiently and accurately the fine-grain struc- ture of the fitness landscape, i.e. the functional (and structural) properties of the protein sequences in the vicinity of the ones tested experimentally.

The thesis outline goes as follows. In Ch. 1 we give a general biological introduction, describing what proteins are and why they are so important for living organisms. Then, we will introduce what an MSA is, and what information we can extract from this data structure. Finally, we will review the experimental techniques on which we will apply the proposed inference methods.

These are treated in Ch. 3 and 4, and go under the names of *Annealed Mutational approximated Landscape* (AMaLa) and betaDCA respectively. The former was specifically conceived to be applied to sequence data generated from Directed Evolution experiments, whereas the second was meant as a more general model that could be applied to a wide variety of experimental settings. The distinctive feature of both methods is that they do not require accurate population information to infer meaningful models.

Another statistical physics inspired model which has recently sparked attention in the context of protein sequence data is represented by *Resitricted Boltzmann machines* (RBM). In Ch. 5 of this thesis, we investigate the chance to employ *Expectation Propagation*, an iterative algorithm for approximating intractable probability distributions, to infer the constituent parameters of an RBM. The work related to this problem is still on- going, and we present here the results obtained so far, postponing to future manuscripts further analysis.