Samuel Bosch
Mentors: prof. Paolo De Los Rios [paolo.delosrios@epfl.ch]
Dr. Stefano Zamuner [stefano.zamuner@epfl.ch]

Inverse statistical methods and pseudolikelihood approximation

Inverse problems in statistical physics are motivated by the challenges of "big data" in different fields, in particular experiments in biology. In inverse problems, the usual procedure of statistical physics needs to be reversed: Instead of calculating the observables over time on the basis of model parameters, we seek to infer parameters of a model based on observations.

In this project, I will focus on the inverse of the Potts model. The Potts model, a generalization of the Ising model, is a model of interacting spins. The reason of this is theoretical (maximum entropy) and practical: it is the simplest model that can explain two-point statistics in data. Also, it is widely used in the community. For example, the reconstruction of neural and gene regulatory networks, and the determination of the 3D structure of proteins [1].

What is the aim of the project?

a) Simulate a Potts model with pairwise and triplet interactions
b) Code an inference system based on Pott's model from scratch and test it.
c) The test will be done by inferring the parameter of a Pott's model, starting from an equilibrium ensemble of configurations, simulated using Monte Carlo techniques.
d) Other tests may involve inferring pairwise couplings of a protein family: Direct Coupling applications (DCA) [2]
e) In prof. De Los Rios' group, a new technique has been devised to infer higher order interactions from data. For this reason, the code should be devised in such a way that it is easy to extend to this more general case.
f) If there is enough time, we would test the new techniques to infer the parameters of a simulated Pott's model in which triplet interactions are present.

New skills which will be obtained through the project:

a) C++ (so far, I only have experience in Python, Matlab and some basics in C)
b) Implementation of a simple Metropolis algorithm
c) Block analysis (or other techniques) to check the convergence of the simulation
d) Theory of inverse Pott's model and maximum likelihood modeling
e) Code for function minimization/maximization
f) Code and theory of neural networks with several hidden layers

Further details:
Tremendous efforts have been made to determine the three-dimensional structure of proteins. A linear amino acid chain folds into a convoluted shape [3,4], the folded protein, thus bringing amino acids into close physical proximity that are separated by a long distance along the linear sequence. The three-dimensional structure of a protein determines its physical and chemical properties, and how it interacts with other cellular components: broadly, the shape of a protein determines many aspects of its function. Protein structure determination relies on crystallizing proteins and analyzing the X-ray diffraction pattern of the resulting solid. Given the experimental effort required, the determination of a protein's structure from its sequence alone has also been key challenge to computational biology for several decades. The computational approach models the forces between amino acids in order to find the low-energy structure a protein in solution will fold into. Depending on the level of detail, this approach requires extensive computational resources.

[1] Nguyen, H. Chau, Riccardo Zecchina, and Johannes Berg. "Inverse statistical problems: from the inverse Ising problem to data science." Advances in Physics 66.3 (2017): 197-261.
[2] Ekeberg, Magnus, Tuomo Hartonen, and Erik Aurell. "Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences." Journal of Computational Physics 276 (2014): 341-356.
[3] Dill, Ken A., and Justin L. MacCallum. "The protein-folding problem, 50 years on." science 338.6110 (2012): 1042-1046.
[4] De Juan, David, Florencio Pazos, and Alfonso Valencia. "Emerging methods in protein co-evolution." Nature Reviews Genetics 14.4 (2013): nrg3414.