The future prospects for *de novo* protein structure prediction from evolutionary information

Speaker: David Jones, University College London, United Kingdom

Despite great strides in *de novo* prediction of protein structure from amino acid sequence over the past decade, there seem to be rapidly diminishing returns in applying these methods to real problems. At the domain level, certainly, *de novo* prediction tends to be of very limited use, particularly for globular proteins. It's now rare to find interesting protein domains without any homologues of known 3-D structure, and even in cases where no templates can be found, template-free modelling results turn out to be little better than they were 10 years ago. More worryingly, there are no reliable ways of knowing if *de novo* methods have even been successful without actually solving the structure experimentally. In the last few years, developments in contact prediction based on evolutionary information have generated a lot of excitement, but it is still unclear as to how useful these methods will ultimately prove to be.

In this talk I will be describing some of our recent work in this area, where we have made use of Sparse Inverse Covariance Estimation methods to predict inter-residue contacts in proteins, and developed algorithms to derive useful 3-D models from this information. We have recently completed a large scale test of these methods which has highlighted some of the limitations of these approaches, which I will discuss. Overcoming these limitations will be vital if they are to be of significant practical use in the future.

Inference of past historical events using Approximate Bayesian Computation methods on population genetics data sets

Speaker: Frédéric Austerlitz, CNRS/Museum National d'Histoire Naturelle/Université Paris Diderot, France.

New computer-intensive estimation techniques such as Approximate Bayesian Computation (ABC) and Monte Carlo Markov chains (MCMC) allows inferring unknown parts of the history of species from contemporary population genetics data. I will illustrate these possibilities with several examples. First, I will talk about a set of human populations from Western Central Africa, consisting of hunter-gatherer Pygmy populations and neighbouring non-Pygmy populations, genotyped for several kinds of genetic markers. Using ABC techniques, we could infer the history of splitting and admixture between these different groups. We could also identify sex-specific demographic processes. The second example that I will mention is the harbour porpoise (Phocoena phocoena) population from the Black Sea. Using again ABC techniques, we showed that this population underwent a strong expansion around 5000 years ago, probably as a result of the reconnection of the Black Sea with the Mediterranean Sea, but that it underwent also a drastic decline around 50 years ago, which can be linked with the intensive hunting of cetaceans performed at that time. Finally I will talk about a study on worldwide human populations, in which by applying MCMC methods on a large set of populations with different lifestyles (farmers, herder and hunter-gatherers), we were able to show that these lifestyles strongly impacted the expansion patterns of these populations. These examples illustrate well how ABC and MCMC methods allow inferring precious information on the history of populations for which archeological records are not available.

Big Data
Speaker: Arnak Dalalyan, ENSAE/Université Paris-Est, France

In this talk, we begin by reviewing some results on popular sparse estimation methods based on L1-relaxation. These methods, such as the Lasso and the Dantzig selector, require the knowledge of the variance of the noise in order to properly tune the regularization parameter. This constitutes a major obstacle in applying these methods in several frameworks-such as time series, random fields, inverse problems-for which the noise is rarely homoscedastic and its level is hard to know in advance.

In the second part of the talk, we will present a new approach to the joint estimation of the conditional mean and the conditional variance in a high-dimensional regression setting with heteroscedastic noise. An attractive feature of the proposed estimator is that it is efficiently computable even for very large scale problems by solving a second-order cone program (SOCP). We will present theoretical analysis and numerical results assessing the performance of the proposed procedure.

Probabilistic approaches for detecting and locating whole genome duplications.
Speaker: Cécile Ané, University of Wisconsin – Madison, USA

Whole Genome Duplications (WGDs) can be difficult to detect when they are old and when synteny has been disrupted by genome rearrangements. To test the presence of WGDs on a species phylogeny, I will present two methods which do not require synteny information and build strength from the phylogenetic framework. They rely on a probability model for the evolution of gene families on a species tree with WGDs. Both methods use multiple gene families across multiple species. One method relies on aligned molecular sequences and the other simply uses information on gene counts. We assessed their performance with simulations and on a benchmark yeast dataset, where we recover strong evidence for a well-established WGD and a low retention rate of duplicated genes after this WGD.