WHEN
Friday, 16 September 2016 from 10:00 to 17:30 (BST)

WHERE
Malet Place Engineering Building 1.03, UCL - 2 Malet Place, London, WC1E 7

REGISTRATION
bit.ly/phylogroupX

PROGRAMME

10:00 - 10:30		**Coffee and registration**
10:30 - 11:00	Keneth Hoehn *University of Oxford*	*Modeling the decay of hotspot motifs in broadly neutralizing antibody lineages*
11:00 - 11:30	Joe O'Reilly *University of Bristol*	*The Mk Model for Total Evidence Dating: Fit for Purpose?*
11:30 - 12:00	Jose Barba-Monotya *University College London*	*Constraining uncertainty in timescale of angiosperm evolution and the veracity of a Cretaceous Terrestrial Revolution*
12:00 - 12:30	Greg Slodkowicz *European Bioinformatics Institute*	*Patterns of adaptive evolution: a structural perspective*
12:30 - 14:00		**Lunch**
14:00 - 14:30	Kris Parag *University of Oxford*	*Optimal Point Process Filtering for the Coalescent and Birth-death Processes*
14:30 - 15:00	Fabricia Nascimento *University of Oxford*	*Stochastic models of endogenous retrovirus evolution and proliferation*
15:00 - 15:30	Iain Moal *European Bioinformatics Institute*	*Modelling the selective forces driving the evolution of the phosphotyrosine / SH2 interaction network*
15:30 - 16:00		**Coffee**
16:00 - 16:30	Asif Tamuri *European Bioinformatics Institute*	*Studying natural selection using mutation-selection codon models*
16:30 - 17:00	Mario dos Reis *Queen Mary University of London*	*The impact of ancestral population size and incomplete lineage sorting on Bayesian estimation of species divergence times*
15:00 - 17:30	Nick Goldman *European Bioinformatics Institute*	*A measure of effective sequence number*

Modeling the decay of hotspot motifs in broadly neutralizing antibody lineages
Kenneth Hoehn – 10:30-11:00

The adaptive immune system in humans is organized around the production of antibodies from B cell receptors. An important part of this process is somatic hypermutation and clonal selection, in which activated B cells undergo rapid mutation and selection for antigen binding. Some B cell lineages produced by this process are capable of neutralizing and co-evolving with chronic viruses such as HIV. Phylogenetic methods have shown great promise in understanding the development of these broadly neutralizing antibody lineages (bNAbs). However, mutational process for generating these lineages is highly biased by tri- and dimer hotspot motifs, which violates important assumptions in most phylogenetic substitution models. Here, we develop a modified GY94-type substitution model that partially accounts for this context-dependency while preserving independence of sites in calculation. This model is a non-reversible, mean field approximation that weights substitutions by the probability that they occurred within a hotspot motif, according to all possible flanking codons. We show that this model is a substantially better fit to three well-characterized bNAb lineages than the standard GY94 model. We implement this model in a flexible framework which, in addition to more accurate phylogenetic analyses, may be used to test hypotheses regarding the degree and type of bias introduced by different hot and coldspot motifs.

The Mk Model for Total Evidence Dating: Fit for Purpose?
Joe O'Reilly, Phil Donoghue, Ziheng Yang and Mario dos Reis – 11:00-11:30

The resurgence of totalevidence methods for divergencetime estimation has created a dependence on the Mkmodel of morphological evolution. This model attempts to encapsulate the process of morphological change in a simplistic manner, assuming that a single rate of change is applicable for all possible character state transitions, and that there is an absence of covariance between characters. The assumptions of the Mkmodel appear to be an oversimplification of the process of morphological evolution, potentially affecting the accuracy of estimated divergence times and evolutionary rates. To test the ability of the Mkmodel to handle varying levels of betweencharacter covariance we simulated data using an alternative model that produces morphological matrices with a predetermined level of betweencharacter covariance. The Mkmodel was then used to analyse these simulated matrices, with the resulting age estimates being compared to the true ages of clades in the tree on which the matrices were simulated. We demonstrate that the Mkmodel has a propensity to underestimate the rate of morphological change, leading to overly ancient age estimates irrespective of the presence of covarying characters. We also show that betweencharacter covariance negatively influences the accuracy of divergence time estimates, particularly in smaller datasets. Our results suggest that the phenomenon of betweencharacter covariance must be taken into account for divergencetime estimation analyses utilising morphological data, both when constructing scored character matrices and when modelling the state transitions that each character undergoes.

Constraining uncertainty in timescale of angiosperm evolution and the veracity of a Cretaceous Terrestrial Revolution

Jose Barba-Montoya, Harald Schneider, Mario dos Reis, Phil Donoghue and Ziheng Yang – 11:30-12:00

Knowledge of divergences among angiosperm lineages is fundamental to understand the process of time evolution, as it allows us to place the rise of angiosperms as part of the Cretaceous Terrestrial Revolution and its consequences for extant environments. Molecular dating approaches invariably yield age estimates that predate the oldest unequivocal fossil record of angiosperms which dates back around ~ 126 million years ago (Ma). Estimates as old as the Triassic and Jurassic were found in independent studies using different methods, calibration strategies and increasing genomic and taxonomic coverage. Arguably, some of these studies use methods that do not adequately accommodate rate variation or they fail to adequately justify their calibrations in terms of phylogenetic affinity or their geologic age. Our study overcomes these shortcomings, while also attempting to control for other sources of error. Our molecular clock analyses imply that crown angiosperms originated during the Triassic to the Jurassic interval. We demonstrate that even though many sources of uncertainty are explored, attempts to control for these factors still do not bring clock estimates and earliest confident fossil occurrences into agreement. We reject a post- Jurassic origin of angiosperms, therefore there must be a cryptic early history to angiosperms.

Patterns of adaptive evolution: a structural perspective

Greg Slodkowicz, Nick Goldman – 12:00-12:30

Protein structure is a major cause of site-to-site evolutionary rate variation. Many structural features such as solvent accessibility, local packing density and proximity to active sites or interfaces have been shown to modulate the evolutionary rate. It is, however, not well understood how these features affect the prevalence of adaptive evolution. Most codon-based models, which are commonly applied for detecting sites under positive selection, do not incorporate any information about the protein structure. In this study, we attempted to form a better view of adaptation on molecular level by asking whether residues under positive selection are close to each other on the protein structure. We generated a large dataset of trees and alignments for 39 mammalian species (covering over 80% of human genes) and calculated sitewise values of selective constraint (dN/dS). We then mapped positively-selected sites onto available crystal structures and analysed whether they tend to be co-located by statistically assessing the distribution of pairwise distances between them. We find that positively-selected sites frequently form tight clusters on protein structures and that this conclusion is robust to low alignment quality and other technical issues. Identified clusters can be assigned into one of several categories: we find that groups of positively-selected residues can surround active sites, occur in binding regions, and

form small, linear clusters in the N-termini of proteins. To our knowledge, the last of these findings has not been previously reported. Additionally, the prevalence of clustering varies in different enzyme classes, with oxidoreductases exhibiting the most evidence for clustering.

Optimal Point Process Filtering for the Coalescent and Birth-death Processes
Kris Parag – 14:00-14:30

The coalescent and birth-death processes are widely used stochastic models for estimating the dynamics of biological populations from samples of genetic diversity. Several well-developed methods already exist for performing population inference using both of these models. However, these methods often trade accuracy and sophistication for flexibility and ease of use. We introduce the Snyder filter, a Bayesian method from electrical and control engineering, as an alternative inference scheme. The Snyder filter is exact, easily adaptable and makes use of the Poisson process nature of birth-death and coalescent models.

Stochastic models of endogenous retrovirus evolution and proliferation
Fabrícia F. Nascimento, Aris Katzourakis and Allen G. Rodrigo – 14:30-15:00

Endogenous retroviruses (ERVs) are presented in all vertebrate genomes analysed to date, including the human genome, and are presented in multiple copies that are passed to the offspring. ERVs are also classified as transposable elements (TEs) because of their ability to move to different loci in a host genome. However, ERVs result from ancient infectious viruses and are often referred to as viral fossils. The vast majority of known ERVs are inactive, but some studies suggest the association of ERVs in host diseases and their expression in early development of human embryo. To understand how ERVs proliferate and evolve in host genomes, a simple stochastic model was developed consisting of a single parameter – birth rate – and phylogenetic trees were simulated based on previous models described for TEs: the strict master and transposon model. In the strict master model only a single copy of a given lineage is able to replicate, and all other genomic copies of TEs are derived from that master copy. In the transposon model, any element of a given lineage is able to replicate in the host genome. Results from this simple model indicate that visual inspection of phylogenetic trees alone can be misleading. However, if a set of statistical summaries is calculated, we are able to distinguish between these two extremes models with high accuracy by using a data mining algorithm. Following this simple model, a more complex model was developed with three additional parameters – inactivation and deletion rates and a probability p that the newly generated element is an inactive element. This last parameter was introduced to take in consideration the intermediate model for TEs, in which only a proportion of elements of a given lineage is able to replicate in a host genome. Limited information is available on how ERVs proliferate and evolve. To overcome this limitation, the approximate Bayesian computation (ABC)

framework was used to predict model parameters. Statistical analyses carried out on simulated data suggested that the ABC method is robust: Even though we were unable to predict deletion rate, we were able to predict relatively well the other parameters. The ABC framework performed exceptionally well for the prediction of the current proportion of inactive elements in a phylogenetic tree.

Modelling the selective forces driving the evolution of the phosphotyrosine / SH2 interaction network

Ian Moal – 15:00-15:30

I aim to build an sequence interdependent codon substitution model in which the probablity of substitution depends upon how the change in sequence affects the binding affinity of protein pairs. I will start by focussing on the interactions between phosphotyrosine (pY) containing proteins and proteins containing SH2 domains, where binding affinity is determined by the sequence of the residues flanking the pY and the sequence of the SH2 domain, and for which affinity can be calculated using an inexpensive energy function. I aim to apply this model to the phosphotyrosine signalling network, in order to uncover how evolutionary forces have been acting across the whole network.

Studying natural selection using mutation-selection codon models

Asif Tamuri and Mario dos Reis – 16:00-16:30

Mutation-selection models of codon evolution, first introduced by Halpern & Bruno, have seen renewed interest in recent years. Large amounts of sequence data coupled with computational advances have made these mechanistic models of sequence change tractable despite the large number of parameters. We explore a couple of variations of the model to characterise natural selection, in particular frequency-dependent selection. We apply these models to influenza protein-coding sequences and contrast the frequency-dependent selection models with selective-shift models developed previously.

The impact of ancestral population size and incomplete lineage sorting on Bayesian estimation of species divergence times

Mario dos Reis – 16:30-17:00

In this talk I will discuss our results from two recent papers (Angelis and dos Reis, 2015, Curr. Zool. and Yoder et al. 2016, PNAS) where we study how estimates of species divergence times are affected by ignoring the coalescent process of genes in populations, and how we can fix the problem by calibrating coalescent trees to geological time (thus obtaining robust time estimates). An updated example of

calibrating the hominoid phylogeny using next-generation sequencing estimates of the mutation rate is provided.

A measure of effective sequence number

Asif Tamuri, Jakub Truszkowski and **Nick Goldman** – 17:00-17:30

Molecular sequences are not independent observations of the evolved state of an ancestral genome. Rather, they have a correlation structure determined by the evolutionary tree relating them. A consequence of this is that once you have some observations, you already know something about their relatives; particularly if they are close relatives. This has led us to the idea of 'effective sequence number', a measure of how many 'independent' observations an observed, phylogenetically related, sample is equivalent to. In this talk I will introduce the key concepts of our measure of effective sequence number, propose a few applications, ask if anyone has any more ideas to help us out, and try to gauge whether this is interesting enough to pursue further.

PhyloGroup X participants

First Name	Surname	Email Address
A S Md Mukarram	Hossain	asmmh2@cam.ac.uk
Asif	Tamuri	tamuri@gmail.com
Chengmin	Shi	c.shi@ucl.ac.uk
Fabricia	Nascimento	fabricia.nascimento@zoo.ox.ac.uk
Greg	Slodkowicz	gregs@ebi.ac.uk
Iain	Moal	moal@ebi.ac.uk
Ivana	Pilizota	ivana.pilizota.13@ucl.ac.uk
Jakub	Truszkowski	jakubt@ebi.ac.uk
Jan	Janouskovec	janjan.cz@gmail.com
Joe	O'Reilly	jo13931@bristol.ac.uk
Jose-Antonio	Barba-Montoya	jose.montoya.12@ucl.ac.uk
Karina	Zile	karina.zile@gmail.com
Kenneth	Hoehn	kenneth.hoehn@oriel.ox.ac.uk
Kris	Parag	kris.parag@zoo.ox.ac.uk
Mario	dos Reis	m.dosreisbarros@qmul.ac.uk
Mitra	Kabir	mitra.kabir@postgrad.manchester.ac.uk
Nick	Goldman	goldman@ebi.ac.uk
Richard	Nichols	r.a.nichols@qmul.ac.uk
Taoyang	Wu	taoyang.wu@uea.ac.uk
Yang	Ziheng	z.yang@ucl.ac.uk
Yuttapong	Thawornwattana	yuttapong.thawornwattana.09@ucl.ac.uk