UQU APS UQU ASSOCIATION OF POSTGRADUATE STUDENTS

PitchMyResearch.com

Competition:

UQUAPS 2016 "Pitching Research" Competition

Submission id:	Date submitted:
UQUAPS-2016-039	**12 Sep 2016 at 09:57 AWST**

Faculty or Institute:	School:
UQ Science	**Chemistry and Molecular Biosciences**

Programme:	Load:	Level:
PhD	**Full-time**	**7-9 months**

Name:

Gabriel Foley

(A) Working Title:

Leveraging uncertainty in ancestral sequence reconstruction using partial order graphs.

Word count: **1000 words**

(A) Working Title	Leveraging uncertainty in ancestral sequence reconstruction using partial order graphs.
(B) Basic Research Question	Can we link individual amino acid predictions together to help rank ancestral proteins and improve our ability to engineer novel proteins?
(C) Key paper(s)	Bar-Rogovsky, H. et al. Assessing the prediction fidelity of ancestral reconstruction by a library approach. Protein Engineering, Design and Selection 28, 507-518 (2015). Löytynoja, A., Vilella, A. J. & Goldman, N. Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. Bioinformatics 28, 1684-1691 (2012).
(D) Motivation / Puzzle	We can already predict the ancestors of protein sequences and use these predictions to get a head start on which amino acid positions to mutate in order to engineer new proteins. However, the number of potential ancestors to explore quickly becomes impossible to experimentally investigate in a lab and we are forced to synthesise only a small number of potential variants. Instead, we could use computer-based models to better inform which positions to mutate and to rank the variants based on their probabilities. This means we could tailor the actual synthesis towards proteins identified as likely to be able to function.
THREE	**Three** core aspects of any empirical research project i.e. the "IDioTs" guide
(E) Idea	Pharmaceutical companies target existing proteins with desirable functions and attempt to improve their ability to be used on a mass scale. One way of doing this is to predict the ancestral sequence that a set of proteins evolved from. This identifies the key amino acids that must be retained and the amino acids that can be mutated to create novel proteins. Three problems exist with this workflow - 1) We often predict ancestors with individual positions of equally likely amino acid probabilities - leading to lots of potential ancestral variants, 2) Attempting to experimentally create lots of ancestor variants means we are forced to sample only a tiny set of the ancestral space, and 3) Current methods typically consider amino acid sites independently, when we know that amino acid sites do not evolve independently. The idea being proposed is that when we have ancestor variants we don't sample them in a laboratory but first screen them computationally. And to rank the variants we link the predictions of individual amino acids together. As we align protein sequences to construct the ancestral sequence, we create a partial order alignment graph representing the probabilities of moving from each amino acid at one position to each amino acid at any subsequent position. Rather than creating a single template with possible variable sites, we use this model to generate a large-scale number of predictions - mimicking the protein synthesis stage normally performed in the physical world. We use the results of this analysis to inform us of amino acid pairs that we rarely

	see in the existing data and limit our protein synthesis to common variants more likely to be functional.
(F) Data	Data to be used are publicly available protein sequences that have been hand-curated to ensure sampling from a wide range of species. Starting with proteins from the cytochrome P450 families, due to the expert knowledge possessed by collaborators and to pre-existing relationships with partners interested in their industrial applications. There will likely be issues with trying to align the proteins in a way that best represents their shared history while capturing enough diversity so that our ancestral variants exhibit novel functions. Testing of methods is to be performed on simulated data from existing open source benchmarking libraries such as BAliBase 3. Experimental validation will be by performing protein synthesis on the identified set of likely variants.
(G) Tools	The primary tool is to be constructed within this project - a probabilistic graphical model capable of aligning sequences and generating libraries of potential ancestors. Alignment of protein sequences will occur by extending existing Hidden Markov Model alignment frameworks for aligning partial order graphs. Various methods for graph sampling are to be investigated. Theory is well documented but will take considerable time to implement. All of the software to implement and test the tool are open source and available. The tool will be coded in Java with a graphical user interface in JavaScript developed concurrently.
TWO	**Two** key questions
(H) What's New?	The simple piece of additional information in the probabilities linking amino acids allows us to create graphs that let us generate and rank huge sets of potential proteins. We can visualise this as the overlapping of three areas - Ancestral reconstruction, linking probabilities using graphical models, and using computational models to generate large sets of data. The intersection of these areas represents the novelty in this idea.
(I) So What?	Pharmaceutical companies are interested in cytochrome P450 proteins as they are the key enzymes in drug metabolism - responsible for about 75% of clinical drug metabolism. This idea would have massive run on effects for cost effectiveness of drug manufacturing. We know that ancestral reconstruction is capable of working, but if we can improve the success rate of proteins that actually fold we can save significant amounts of time and money.
ONE	**One** bottom line
(J) Contribution?	A novel method of generating and ranking large libraries of potential ancestors packaged in a tool with a graphical user interface to allow users to easily perform this analysis.
	There is an existing collaboration between my bioinformatics lab and a biochemistry lab knowledgeable about cytochrome P450s and capable of

(K) Other Considerations	synthesising and assessing protein function. The most recent paper (Bar-Rogovsky et al., 2015) discussing ancestral libraries was published in Protein Engineering, Design and Selection. The most recent partial order graph application (Loytynjoja & Goldman, 2012) was published in the higher impact-factor journal Bioinformatics. Given the novel nature of the current work, these journals are reasonable lower and upper bounds. There is a risk that this method doesn't deliver substantially better predictive power. We can mitigate the chance of failure by applying the underlying models to other areas within bioinformatics. A tool that takes sets of linked predictions and generates and ranks potential arrangements would be useful in areas such as mapping sequence reads to genomes. The project is substantial and the modularisation of ideas and the ability to apply the concepts in other research domains means we can dynamically focus or expand the scope as the project progresses.