

Exposing bioinformatic programs as Web Services

1. What research is being done by the researcher(s)

Protein-protein interaction motifs

Richard Edwards' primary research interest is exploring the functional importance of short, linear motifs (SLiMs) in proteins through molecular evolutionary genetics. He is heavily involved in developing better bioinformatics approaches and tools for detecting both known and novel protein motifs in protein datasets, particularly those concerned with protein-protein interactions. A major component of his work is developing methods and pipelines to link and analyse the results from high throughput studies using these tools. To date, his work has concentrated on the human and yeast proteomes, with an emphasis on the role of SLiMs in human disease, particularly cardiovascular disease and cancer.

As an evolutionary biologist, he is interested in the evolutionary dynamics of SLiMs and the protein interactions they mediate. The evolutionary plasticity of SLiMs has not yet been determined and has important implications for automated functional annotation based on sequence features. It is hoped that a better understanding of evolutionary patterns will also help improve *in silico* methods for identifying novel motifs. The evolutionary stability of protein interaction networks is of great importance when making cross-species inferences about interactions or functions.

Integrated Bioinformatics.

Beyond his main research, Richard has a broad interest in most aspects of evolutionary biology and has been involved in many collaborative projects, including: predicting functional specificity in protein families, protein identification through tandem mass spectrometry of proteins, processing protein datasets from proteomics studies, bioinformatics prediction of tyrosine phosphorylation motifs, 3D analysis of protein c-termini, investigating protein-coding tandem repeats in the human genome and identifying functionally important human SNPs as candidates in heart disease.

Currently, Richard is using a combination of applications, including MUSCLE, CLUSTALW, BLAST and SLiMfinder (his own application) in his work. He uses publicly accessible data which he downloads and uses locally.

He would like to expose some of the software he has written as web services, making them accessible to the wider academic community, and enabling their integration with tools such as Taverna. The resulting workflows would be hosted on MyExperiment (<http://www.myexperiment.org>). The software is designed to work on multiple datasets (the latest software is single dataset friendly) and has high throughput involving large tables of numbers. He downloads separate databases and puts them into custom databases. For visualization he uses a network visualisation tool called Cytoscape (<http://www.cytoscape.org/>).

2. What are the issues that ENGAGE could address?

Primarily Richard Edwards wants to produce multiple Taverna workflows that he can make available to his students and the wider community. To accomplish this involves the following:

- 1) The same protein sequence has to be looked up in different databases (approx. 6) and then it is mapped between them. However, the biggest stumbling block is that four of these six databases do not have a web services interface. For these databases, where the service is not always documented, there are three options:

- 1) persuade the service provider to build one, 2) produce a screen scraping script, or 3) automatically download the data every so often.
- 2) SLiMFinder (Short Linear Motif Finder) looks for short amino acid sequences in proteins and is an application developed in Python by Richard Edwards. This needs to be wrapped using SOAPLab2 so that it can be exposed to Taverna. As the protein sequences from the databases in stage (1) are input into this application, this work depends on stage (1) being completed. Feasibility work can be undertaken prior to stage (1) being complete by testing with static protein sequence files.
- 3) Once SLiMFinder has been wrapped the next stage is to wrap the following applications in the same way: GOPHER, GABLAM, RJE_SEQ and RJE_UNIPROT.

3. Future benefit to user community

In the short term, the workflow(s) produced for Taverna will not only aid Richard Edwards in his research, but will also be used in teaching, 3rd year projects, etc.

In the longer term, the workflows and the wrapped applications would be made available to the wider bioinformatics community. It is envisaged that the workflows will be added to the myExperiment (<http://www.myexperiment.org>) website for sharing. The work accomplished here will show the community what can be achieved by wrapping their applications and creating workflows in Taverna. This has potential benefits to other domains.