Jenny Molloy <jcmcoppice12@gmail.com>

Form Submission

synbio-webmaster@lists.cam.ac.uk <synbio-webmaster@lists.cam.ac.uk> Thu, Apr 9, 2015 at 2:11 PM
To: Jenny Molloy <jcm80@cam.ac.uk>

Your name
David Lea-Smith
Your E-Mail Address
djl63@cam.ac.uk
The Idea

The idea is to develop a computational system and corresponding web site to couple information on metabolic and biochemical networks within bacteria (focussing initially on cyanobacteria) with networks of protein evolutionary history and homology. The system is designed to be easily used by biologists with minimal computing experience. The main principle is that the variability in the enzymatic components of pathways across a large sample of different genomes provides a valuable resource for the understanding and manipulation of biosynthesis.

We aim to create a web-based tool to easily analyse the gene and protein families involved in the complete multi-organism network of biosynthetic pathways across hundreds of genomes. This will allow points of interest such as conserved, specialist and missing biosynthetic steps to be quickly and easily identified from within a clade of organisms. Thus the synthetic pathways within individual species can be better understood, and underpinned with concrete data relating to genes and homologues etc. Importantly, the genome specific differences will enable identification of useful pathway components that can be recombined in novel ways to introduce foreign biosynthetic pathways into an organism of interest.

This project involves the interconnection of two networks of information 1) the traditional biochemical network of enzyme-driven metabolic pathways and 2) the evolutionary connection between homologous genes and proteins across a wide range of organisms. These networks can readily be described by a computational system that models each enzyme as a "node" with connections to other nodes. In the first case the connections will correspond to metabolite (product-substrate) links between genes, and in the second case the connections represent significant sequence similarity, from which homology can be inferred. These networks will be computed and/or curated and stored in a database that will then underpin the web-based tool. Gene families not involved in metabolic pathways will also be collated and stored in the same system, and though not presented in exactly the same way, they will be equally accessible.

The web site will display the information in a manner that aims to illustrate how different biosynthetic pathways differ between genomes. The presentation of data will be primarily visual, anchoring it to a metabolic pathway map; either a complete multi-genome "superset" map or just a subset, focussing on a particular pathway. The display of gene presence/absence and conservation across the genomes will also be graphical, including for genes not involved in a metabolic pathway. The user will then be able to investigate the full depth of sequence based information, connect to external bioinformatics databases (GenBank, UniProt etc.) and extract any family trees and alignments for any and all points of interest. All data will be downloadable as spreadsheets and in a variety of popular bioinformatics formats. The metabolic map may also be superimposed with other, orthogonal data (e.g. transcriptomics, ChIP-seq, metabolomics) that can be anchored to the genes, so the information can be displayed and analysed in a pathway-wide context, rather than as the more common linear genome or array-based representations. Free, open-source software allowing mapping of genomics data to a complete metabolic gene map (and also to genes not involved in metabolism) is currently not available and would make many biological analyses simpler.

David Lea-Smith will be primarily responsible for the creation of a holistic biosynthetic pathway map from a wide-ranging review of databases and published literature relating to metabolism and biosynthetic pathways. This will be completed for cyanobacteria in the initial instance. Eschirichia coli, the best annotated bacterium and the model cyanobacterium, Synechocystis sp. PCC 6803 and Nostoc sp. PCC 7120, will be used to create the anchor points for a consistent annotation of homologous gene clusters. David will also be involved in the preliminary testing of the web site and

provide feedback to ensure that the eventual outcome suits the needs of the biological community. The clustering of genes into families naturally aims to make it clear what the homology relationships are between different proteins, e.g. to identify orthologues. Where the traditional naming of genes and proteins are either inconsistent or missing, combining knowledge about orthology and where a protein is likely to act in a particular pathway make the identity of any component unambiguous. Thus as a necessary side-effect, the system will be to generate a single consistent nomenclature for all of the enzymes in the network of pathways, across all the organisms of interest. In the future this could provide a basis for automatically annotating new genome sequences within the same scheme.

Tim Stevens will be responsible for the bioinformatics analyses and the display of the results within the pathway maps as a website. This work involves several sequential steps: comparison of all protein sequences from all genomes under study to generate a matrix of detectable similarities; the hierarchical clustering sequences into family groups; the detection of remote homology and identification of missing genes; the connection of clusters, and thus also individual genes, to anchor points on the biosynthetic pathway map. All of this information will be stored in an SQL database (as is standard) and be presented as an interactive, searchable, graphically-oriented web page. A hierarchical approach will be used for the clustering of protein sequences into family groups because the amount of conservation within a family can vary substantially from case to case. Also, by moving up and down a familial hierarchy a user of the website will be able to see how specialisation of function arises as species and sequences diverge. This will allow sequence variation (or absence) to be related to metabolic capabilities.

Initial work will focus on cyanobacteria because it is a mainstay of the Howe lab and because a large amount of analysis on cyanobacterial synthetic pathways has already been performed, with a practical application towards generating biofuels. This group of bacteria will also serve as a test bed for the system, fixing any problems and refining the web site before the project is opened-up to bacteria at large. This project must be of limited scope to be achievable within a limited time period, and so will focus on a subset of Bacteria, but the technology would naturally be expandable to further clades, including those from Archaea and Eukarya.

Who we are

David Lea-Smith (djl63@cam): Postdoctoral associate in the Department of Biochemistry (Chris Howe lab).
David specialises in microbial genetics and biochemistry with a focus on photosynthetic organisms (cyanobacteria and purple bacteria). He uses synthetic biology, genomics and genetic tools to understand bacterial physiology and metabolism and develops strains with increase biofuel or electrical output using energy derived from either photosynthesis or degradation of waste products.

Tim Stevens (tjs23@cam): Senior Investigator Scientist in the Munro Lab at the MRC Laboratory of Molecular Biology in Cambridge. Tim provides computational biology oversight, development and training within the LMB's Cell Biology Division. He researches several bioinformatics areas with a primary focus on protein homology detection methods, 3D genome structure and interactions (PMID:24067610), transmembrane proteins (PMID:20603021) and is author of the book "Python Programming for Biology" (ISBN-13: 978-0521895835).

Sign team up to SRI site?

1

Implementation

The funding will be used primarily for the purchase of a rack-mount computer server (e.g. Dell PowerEdge R730) that will run the website and database, with a capability for on-the-fly informatics searches/analysis and provision for the database to be expanded as needed to encompass further information for more species. It will be housed within the Department of Biochemistry and top-level administration will be provided by the Bioinformatics and Computational Biology Service (http://computing.bio.cam.ac.uk/index.html), but all of the set-up and day-to-day running of the server and web site will be handled by the primary collaborators. Mid-way though the term of the project, a test version of the web site running on the server will be made available locally to the Synthetic Biology SRI and full public deployment will occur by the end of the term. The funding will also cover the costs of publishing this work in an appropriate peer reviewed journal.

Benefits and outcomes

This project idea is underpinned by the collaboration of a University laboratory that specialises in understanding molecular evolution, metabolism and biochemistry of photosynthetic micro-organisms, with a specialist in computational biology from the LMB who can provide all of the necessary expertise in bioinformatics and the computer programming required to create on-line services and databases.

The eventual outcome of this project will be the creation of a tool to enable any researcher within the field of synthetic biology to investigate synthetic outcomes and plan genetic modifications in the context

of a complete biochemical network. Syntheses of interest will be presented in an accessible manner based on a metabolic pathway map, while the full depth of information will be made available at every enzymatic point, for example to show sequence alignments and trees to illustrate how homologues have arisen, diverged and disappeared.

This tool will be built using open, web-based technologies and thus the final product will be readily shareable with the entire biological community. Coupled with this, all of the computational code that underpins the system will be made available as open source software, with the potential for future collaborative development with other groups.

The project would provide the synthetic biology field, and wider disciplines, with a validated set of protein families and a consistent nomenclature (which in isolation is not a trivial task). This could then naturally pervade newly sequence genomes, providing a robust means by which many gene sequences can be automatically and consistently annotated.

Much of the preliminary proof-of-concept work for this project has already been achieved and so it is entirely realistic to complete the project, and present it as a functional web resource, within six months. An initial database covering cyanobacterial genomes and proteome sequences has already been assembled. Also, all of the computational code required to perform exhaustive all-versus-all sequence based homology searches, and the clustering and hierarchical sub-clustering thereof, has been written.

Given our preliminary work on cyanobacterial genomes, this clade will be the first to be presented in a prototype web site. This prototype will be made available locally to the Synthetic Biology SRI, for testing and appraisal by anyone who wishes to try the system, thus promoting interaction within the entire SRI. In due course the situation will be expanded to cover further bacterial clades and then released publicly.

When completed, this work will be published in an appropriate peer reviewed journal.

Sponsor for the research and cost centre

Prof. Chris Howe, Department of Biochemistry ch26@cam.ac.uk

Sponsor support confirmed?

1

Budget

Rack-mount server unit to run website and perform on-the-fly analyses: Dell PowerEdge R730 CPU:Intel® Xeon® E5-2630 @2.4Ghz, RAM:16GB RDIMM, 2133MT/s, HDD (main):200GB Solid State Drive, HDD (backup):300GB 15K RPM SAS. Cost: £2,457.05 Publication costs, estimated at approximately £1500.