Computers in Biology and Medicine 43 (2013) 2028–2035

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Accelerating *in silico* research with workflows: A lesson in *Simp*licity

Paul Walsh [a], John Carroll [a], Roy D. Sleator [b,*]

[a] *nSilico LifeSciences, Ltd., Melbourne Building, Bishopstown, Cork, Ireland*
[b] *Department of Biological Sciences, Cork Institute of Technology, Rossa Avenue, Bishopstown, Cork, Ireland*

ARTICLE INFO

Article history:
Received 3 December 2012
Accepted 12 September 2013

Keywords:
Bioinformatics
Computational biology workflow
annotation
Usability
Genomics
*Simp*licity

ABSTRACT

Bioinformatics is the application of computer science and related disciplines to the field of molecular biology. While there are currently several web based and desktop tools available for biologists to perform routine bioinformatics tasks, these tools often require users to manually and repeatedly co-ordinate multiple applications before reaching a result. In an effort to reduce time and error, workflow tools have been developed to automate these tasks. However, many of these tools require expert knowledge of the techniques and supporting databases which more often than not lies outside the scope of most biologists. Herein, we describe the development of sequence *i*nformation management platform (Simplicity), a workflow-based bioinformatics management tool, which allows non-bioinformaticians to rapidly annotate large amounts of DNA and protein sequence data.

1. Background

1.1. Introduction

Some of the most common goals in molecular biology, particularly in the post-genomics era, areto quickly and accurately identify genes in a genome/metagenome [1], to ascribe a putative role for each gene [2], determine the structure of the encoded protein, and ultimately to ascertain the function of the predicted protein [3], often leading to the discovery and development of new or improved therapeutics. Bioinformatics has become a popular alternative to expensive and laborious wet lab activities, allowing biologists to quickly form a testable hypothesis about what a protein may be and/or what role it likely plays [4].

However, bioinformatics tasks often require biologists to manually and repeatedly co-ordinate multiple tools to produce a result as outlined in Fig. 1. Data transfer, between applications, is either by manual cut-and-paste or in more advanced cases by 'screen-scraping' web pages using scripting languages like PERL, often with additional data 'massaging' (*e.g.*, small alterations in formatting, selections of subsets, and simple local transformations such as DNA-to-protein translation) [5,6]. Furthermore, a majority of contemporary bioinformatics tools fail to provide a reliable record – results from one website are simply copied and pasted into another, with no record of important parameters such as algorithm settings, time stamps or database versions used [7].

Typically, the first step in a program of research performed by a biologist is to use a gene prediction tool to find the open reading frames (ORF) in a genome/metagenome [2]. Once the ORFs are determined a biologist must perform a sequence similarity check for each ORF found [8]. A sequence similarity tool compares the sequence being analyzed against a database of known DNA or protein sequences (such as GenBank for DNA or the UniProtKB/Swiss-Prot protein database). One of the most popular sequence similarity tools is BLASTX [9]. The biologist must copy each ORF into BLASTX and select the databases to search and the scoring matrix or algorithm to use. A process which must be repeated for every ORF found. Some protein functions can be quickly identified due to their high sequence similarity to proteins whose function has already been identified and experimentally proven in a 'wet' lab – a process known as homology based transfer. A high sequence similarity suggests that the protein sequence being analyzed (the query sequence) is similar to a known sequence (the subject), and since structure informs function, a putative function can be ascribed to the query sequence [10].

ORFs whose functions are not identified by sequence similarity against the primary databases must undergo further analysis using various methods in order to determine the function of the protein being analysed. Motif searching (patterns and profiles) using tools such as Prosite will help to identify highly conserved signature sequences which may provide clues as to the protein's evolutionary origins [11]. If no sequence homologies exist, genomic context or expression based systems such as Phydbac2 may be used. Furthermore, at least in the case of proteins, structure based approaches such as FATCAT, VAST and FAST for full 3D structure or PROCAT for 3D structure motifs can be used [10]. In essence, the task of ascribing a function to each gene in the genome/metagenome involves a multistep workflow which unnecessarily ties-up the

* Corresponding author. Tel.: +353 21 4335405; fax: +353 21 4326851.
E-mail address: roy.sleator@cit.ie (R.D. Sleator).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.09.011

biologist – distracting them from the wet lab experimentation for which they are properly trained [12].

1.2. Motivation - The need for Bioinformatics Management Tools

Workflows, such as the one described above can be labour intensive, error prone, untraceable and often result in the generation of significant amounts of data which the biologist must organise themselves [13,14]. Annotating sequence sets manually using the available online tools can therefore be quite labour intensive and, depending on the genome/metagenome size, may take several hundred man hours to complete [15].

Early attempts at automating bioinformatics tasks included using screen scraping scripts to create pipelines [16], however Hyper-text mark-up language) (HTML) based bioinformatic web interfaces were for the most part designed to be used by humans not scripts. This technique proved troublesome as web pages are occasionally redesigned forcing the programmer to rewrite the script to enable it to work on the new web page. Furthermore, most early bioinformatics workflow tools were developed for specialist bioinformaticians, often based on a UNIX platform using command line software [17]. While this approach remains popular, it is often a difficult transition for non-computer savvy biologists venturing into the bioinformatics arena for the first time, and as such, requires a significant time investment.

During the late 1990 and the early 2000 many bioinformatics workflow developers turned to developing GUI (Graphical User Interface) (GUI) applications, while others began developing web based tools using recently developed WEB 2.0 technologies [18]. These applications, or tools, allow researchers to visually build a workflow by selecting several components, each of which performs a separate task in the workflow. Complex and powerful workflows can be created saving the researcher both time and effort.

1.3. Bioinformatics Workflow Tools

Several bioinformatic workflow tools, both open source and commercial, are currently available (Table 1). While some are desktop applications that employ local resources and web services for workflow execution in the application, there are also web based workflow tools built using Web 2.0 technologies. In this case the workflow execution is performed on a server while a browser is used to create workflows and to review workflow results. Relatively few rich internet application (RIA) bioinformatics workflow tools have been developed including Microsoft's general workflow tool, Popfly, which has been discontinued and Calvin [19]. RIA are embedded in an internet browser and have similar behaviour to desktop applications enabling sophisticated user interactions, client-side processing, asynchronous communications, and multimedia [20].

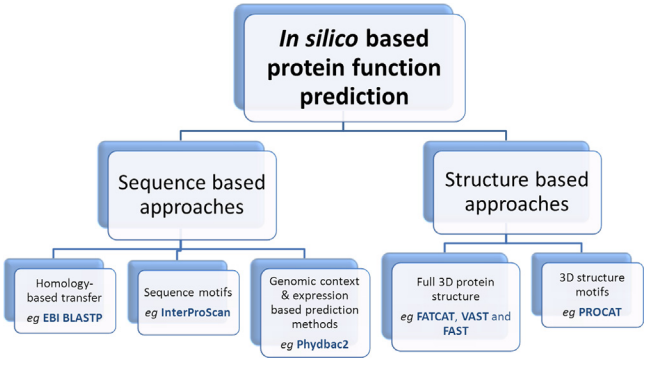

Fig. 1. Overview of the typical workflow which a Biologist– might use to predict the function of a protein (modified from [27]).

In general, existing bioinformatics workflow tools can be grouped into one of three categories: those designed for (i) the expert bioinformatician, (ii) biologists with some bioinformatics expertise and (iii) biologists with little or no bioinformatics acumen. Both Taverna and Biobike are aimed at the first group; users experienced in writing workflow software in Perl, Phyton or Lisp, yet desirous of an easier and faster way to create workflows without the need to program. This group would be aware of the web services available, the different data formats used and how results need to be transformed to suit the input of another web service. Galaxy and Ugene are aimed at the second group, users who have experience using bioinformatics tools but don't have the technical knowledge to write workflows in Perl, Phyton or other relevant languages. These users are aware of the different online tools available and of some of the data formats but would rather use an automated tool to handle the data transformation. Finally, GenomeQuest, Bioextract and Weblab are aimed at all levels of expertise, from the novice to the expert user. GenomeQuest allows bioinformaticians to write workflows in a scripting language called smarty, while biologists who have no experience of writing workflows can simply request GenomeQuest to generate the workflows (see Table 2 for an overview of each of the seven workflow tools mentioned above).

2. *Simp*licity architecture

*Simp*licity is a bespoke bioinformatics management system that allows biologists, with little or no computing background, to manage and analyse information generated from large scale genomic/metagenomic sequencing projects [1]. Currently *Simp*licity incorporates a number of the most common bioinformatics tools, including Gene Prediction (Glimmer, EMBOSS GetOrf), Similarity Searching (EBI NCBI Blastp, RSCB PDB Blast Search, Interproscan, Pfam and CATH), Multiple Sequence Alignment (ClustalW2), Phylogenetics (PHYLIP fProtdist, PHYLIP fNeighbor, PHYLIP fproml and PHYLIP fprotpars), Primer prediction (EPRIMER3) and Genome Mapping (Gview). More tools are currently being added to meet individual user requirements as well as being tailored for specific projects. *Simp*licity was developed using an evolutionary prototyping software development model [21]. This approach implements only confirmed requirements from biologists. Evolutionary prototyping involves implementing well understood requirements in a rigorous fashion and writing the code in a way that is easily modifiable. The prototype then evolves as unknown

Table 1
The most commonly used tools for workflow design and execution.

Desktop applications	Web based tools
Taverna – www.Taverna.org.uk	WebLab – http://weblab.cbi.pku.edu.cn
Ugene – http://ugene.unipro.ru	Bioextract – http://bioextract.org
Wildfire – http://wildfire.bii.a-star.edu.sg/index.php	Galaxy – http://galaxy.psu.edu
Triana – www.trianacode.org	Biobike – www.biobike.org
Kepler – http://kepler-project.org	Ergatis – http://ergatis.sourceforge.net
Pipeline Pilot – http://accelrys.com/products/pipeline-pilot (commercial)	Genomequest – www.genomequest.com (commercial)