BRIEFINGS IN BIOINFORMATICS. VOL 10. NO 1. 1–10
Advance Access publication October 8, 2008
doi:10.1093/bib/bbn041

Bringing Web 2.0 to bioinformatics

Zhang Zhang, Kei-Hoi Cheung and Jeffrey P. Townsend

Submitted: 18th August 2008; Received (in revised form): 12th September 2008

Abstract

Enabling deft data integration from numerous, voluminous and heterogeneous data sources is a major bioinformatic challenge. Several approaches have been proposed to address this challenge, including data warehousing and federated databasing. Yet despite the rise of these approaches, integration of data from multiple sources remains problematic and toilsome. These two approaches follow a user-to-computer communication model for data exchange, and do not facilitate a broader concept of data sharing or collaboration among users. In this report, we discuss the potential of Web 2.0 technologies to transcend this model and enhance bioinformatics research. We propose a Web 2.0-based Scientific Social Community (SSC) model for the implementation of these technologies. By establishing a social, collective and collaborative platform for data creation, sharing and integration, we promote a web services-based pipeline featuring web services for computer-to-computer data exchange as users add value. This pipeline aims to simplify data integration and creation, to realize automatic analysis, and to facilitate reuse and sharing of data. SSC can foster collaboration and harness collective intelligence to create and discover new knowledge. In addition to its research potential, we also describe its potential role as an e-learning platform in education. We discuss lessons from information technology, predict the next generation of Web (Web 3.0), and describe its potential impact on the future of bioinformatics studies.

Keywords: *Web 2.0; bioinformatics; scientific social community; web service; pipelines*

INTRODUCTION

With the completion of many genome sequencing projects and the proliferation of genome-scale assays and analyses, bioinformatics research has become increasingly data-intensive. According to the 2007 update for the Bioinformatics Links Directory [1], there are nearly 1200 publicly web-accessible links including databases and web servers, that aim to collect, organize, visualize, integrate and analyze biological data. For a given task, researchers in the field of bioinformatics often need to consult numerous databases and web servers. However, the integration of heterogeneous datasets from disparate databases associated with multiple web servers is daunting for researchers [2]. It requires them to be proficient at computationally 'surfing' databases and web servers and algorithmically 'skimming' the requisite data. The challenge of decoding volumes of biological data from disparate sources underscores an imperative for greater data integration [3].

Toward this end, two major approaches to the integration of biological data from multiple heterogeneous databases have been widely adopted. These approaches have attempted to solve the problem in two divergent ways: centralization and decentralization. The centralized approach may be typified by the data warehouse and the federated database [4, 5]. The data warehouse approach brings all accessible data from various source databases to a local data warehouse, and then executes all queries on this

Corresponding author. Jeffrey P. Townsend, PhD, Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06520, USA. Tel: +1-203-432-4646; Fax: +1-203-432-5176; E-mail: jeffrey.townsend@yale.edu

Zhang Zhang, PhD, is a postdoctoral associate at the Department of Ecology and Evolutionary Biology, Yale University. His research is focused on bioinformatics, molecular evolution, evolutionary comparative genomics, data mining and data integration using Web 2.0.

Kei-Hoi Cheung, PhD, is an Associate Professor at Center for Medical Informatics, Department of Anesthesiology, Department of Computer Science and Department of Genetics, Yale University. His research interests include bioinformatics and the exploration of Semantic Web and Web 2.0 in biological database and tool interoperation.

Jeffrey P. Townsend, PhD, is an Assistant Professor at the Department of Ecology and Evolutionary Biology, Yale University. His research interests lie in bioinformatics, experimental functional genomics and analysis of DNA microarrays, phylogenetics and evolutionary biology.

local warehouse, rather than on the distributed sources. Consequently, the data warehouse approach improves query performance, such as fast response time [4], and provides both additional centralized control over data and tool sets that may be customized to meet users' needs [6]. However, it requires continuous updating to keep the data and the tools comprehensive of the evolution of the source data. Examples include BioWarehouse [7], an open source toolkit for constructing data warehouses; BIOZON [8], which integrates heterogeneous data types such as proteins, structures, domain families, protein–protein interactions and cellular pathways; and CFGP (Comparative Fungal Genomics Platform) [9], which incorporates fungal genomic data and several analysis tools into a data warehouse.

In contrast, the federated database approach translates a query against a federated database into a query against the many source databases, fetching the data from the source databases, parsing the results from disparate sources and then reformatting the data for its user base. The federated database approach is thus always up to date with the source databases. However, it generally has a poorer query performance [4]. Moreover, it requires continuous updating of the search agents that parse and reformat results whenever the source databases change their data structures (By some estimates in the field of bioinformatics, data structures roughly change twice a year [5]). Examples include DiscoveryLink [10], which provides users with a federated database and translates a query to access multiple data sources, and QIS (Query Integrator System [11]), which stores diverse queries for data integration from continuously changing heterogeneous data sources in the biosciences.

Data warehousing and federated databasing both build a centralized database, with their focuses on data translation and query translation, respectively. They confront problems stemming from storage facilities, frequent updates and high costs for data exchange and maintenance. For this reason, a decentralized approach has also been advanced, in which individual data providers agree to offer their data in standard formats, typified by BioMoby [12], Distributed Annotation System (DAS) [13, 14] and Taverna [15]. BioMoby is a system for interoperability between data providers, using web services for data exchange [12, 16, 17], but it adopts web services that only follow the Simple Object Access Protocol (SOAP) [18] and does not include other formats of web services described below. DAS is designed for distributed genome annotation, using a defined URL to transport data in the form of XML documents [13, 14, 19]. Compared with BioMoby, DAS uses well-defined URLs for commands to exchange data and fetches data with a precalculated time, so DAS does not support time-consuming computations [20], such as BLAST [21] analysis. Taverna, a part of MyGrid [22], aims to use and integrate the growing number of molecular biology tools and databases, and is a graphical workflow workbench only for desktop installation [15].

These piecemeal efforts at integration have only touched a fraction of web-based bioinformatic source data, so that for most complex queries, it remains challenging and laborious to integrate data present on multiple databases and/or to analyze data using tools located on different web servers. Cross-database communication through the World Wide Web (or Web [23]) may hold great promise for streamlining the resolution of these issues in bioinformatics studies. Web 2.0 has gained much attention as a revolutionary way of managing and remixing online data, enabling interoperability across heterogeneous data sources. In the following section, we introduce the Web 2.0 technologies to bioinformatics. We then explore major elements of Web 2.0 and propose a Web 2.0-based Scientific Social Community (SSC) model for bioinformatics. This model facilitates a collective, social and collaborative platform for data automatic analysis and data sharing as well as data integration. We present our perspectives on Web 2.0 and predict its impact on the future of bioinformatics studies, discussing its many potential roles in bioinformatic research.

BRINGING WEB 2.0 TO BIOINFORMATICS

What is Web 2.0?

We have only one World Wide Web. Web 2.0 is not a new Web, just a convenient term reflecting the evolution of the Web. In contrast to Web 1.0, Web 2.0 represents a shifted focus from working locally to working in a networked setting. In this shifted focus, the Web is seen as a social, collaborative and collective space. As defined by Tim O'Reilly, '*Web 2.0 is the business revolution in the computer industry caused by the move to the Internet as platform, and an attempt to understand the rules for success on that new platform*' [24].

Web 2.0 represents a revolutionary way of collecting and integrating online information and

knowledge repositories. A suite of novel approaches belonging to Web 2.0 may be tabulated and differentiated from Web 1.0 (Table 1). Several key elements characterizing Web 2.0 are:

(1) Social web: participation and communication can link people located anywhere with similar interests, forming a social network. The content provided by web sites in the Web 1.0 era may only be read, whereas users may easily generate and publish content on Web 2.0 sites. Examples of non-bioinformatic sites leading the Web 2.0 transformation include YouTube, blogging sites and wikis. Furthermore, users may adopt the Really Simple Syndication (RSS [25]) technology, subscribing to customized content provided by others. Such customization enables data distribution and sharing with greater convenience, rapidity and efficiency. This increase in efficiency of communication alone can facilitate data integration. Likewise, in order to discover knowledge from rapidly accumulating biological data through collective intelligence, a social web is also needed for bioinformatics to connect people with similar research interests.

(2) Web service: our web is not limited to any one personal computer platform, thus access to information deposited in the Web is pluralistic. Information presented by interlinked Web 1.0 sites (for example, HTML web pages), include data and layout, enabling humans to read and explore data in an easy and free-form exploratory manner, that is, user-to-computer communication. However, with Web 1.0 technology, it remains difficult to channel specific information between or among computers and between or among users. Web services have developed as a way to achieve software interoperability and to support computer-to-computer interaction through Web Application Programming Interface (Web API [26]) described in the Web Services Description Language (WSDL [27]). Web services use eXtensible Markup Language (XML; an open standard for describing data) for easy exchange of information between applications [28] and have several different formats, such as SOAP-based (a protocol for exchanging XML-based messages over computer networks [18]) web services, JSON-based (JavaScript Object Notation; a lightweight data-interchange format based on the object notation of the JavaScript language [29]) web services. Web services correspond to the programmatic interfaces, whereas web servers can be only accessed by browsers (such as Firefox, Safari and Internet Explorer). Thus, we denote the latter as browser-based tools, although web servers indeed provide a service too [6].

(3) Software as a Service (SaaS): software may be provided as a web service that is always on, and always improving in reply to users' latest needs. For instance, Google Documents [30], a service to create and share documents online, has successfully turned 'Office Software' into a web service that one may easily access through the browser instead of buying software, installing it and regularly upgrading it. 'Release early, release often' and 'the perpetual beta' are the motto for SaaS, and new features are added and updated frequently. For example, consider Gmail, offering free web-based email and archival services: it seems that a 'Beta' has been enclosed in the logo for years [24]. The major advantage of SaaS is that it removes the need for local installation and for communication across diverse platforms each with their own operating language. SaaS provides up-to-date web services to facilitate communication and collaboration over the Web. In the wake of accumulating of biological data, numerous new bioinformatics software and tools are also needed that provide SaaS.

(4) Users add value: Web 2.0 has been used successfully as a model in business. The principal reason for its value in successful Web 2.0 sites is the up-to-date information added by users and shared among users (such as, Craigslist, an online

Table 1: Differences between Web 1.0 and Web 2.0

	Web 1.0	Web 2.0
Alias	Hyperlink Web	Social Web
When	1994–2004 [71]	2004—now
Conception	Web as a medium	Web as a platform, software as a service
Information	Read only—receive information passively	Read and Write—create and receive information actively
Communication	User-to-computer	Computer-to-computer and user-to-user
Information discovery	Search and Browse	Publish and Subscribe

community featuring user-added classified advertisements [31], and YouTube, a video sharing website where users can upload, view and share video clips [32]). Due to this community enterprise, Web 2.0 sites are relatively cheap to maintain [33]. In comparison, existing methods for data integration (such as data warehousing) may also provide a unified portal to information over the Web. However, with centralized content control, they require significantly more time and money for maintenance. A poignant example of the issues of cost control for data warehouse is the Integrated Genome Database (IGD) [34]. The IGD used the data warehouse approach to integrate data from several databases, including GenBank, Genome Database, EMBL and SWISS-PROT. Although appealing to users in principle for its unified bioinformatics portal, the IGD survived only 1 year, due to frequent updates and high cost of maintenance in response to the changes of source databases [5]. In bioinformatics, the sword of Damocles hangs over every Web 1.0 database/web server, due to the threat of loss of financial support [35, 36]. With centralized updating of content, even a minor gap in funding support may hamstring and doom an otherwise cutting-edge Web 1.0 database, largely because in the Web 1.0 model, users are allowed only to retrieve information, but not to update it.

Scientific social community

The nature of Web 2.0 is social, collective and collaborative. It is an appropriate goal for our Web. Critical attention to Web 2.0 development for bioinformatics should be applied not merely to data integration itself, but also to data sharing and reuse by encouraging user participation, linking people with similar research interests, emphasizing collaboration and exchanging data via web services. To this end, therefore, we propose three goals for bioinformatics: data integration, automatic analysis and data sharing.

- When it comes to data integration, we need to explore the question, 'what is data in bioinformatics?' Data are not only sequences and other raw data, but also include analyzed results, methods, tools, algorithms, papers, knowledge (see [37] for a discussion about knowledge integration in biomedicine) and even connections among people. It is particularly the case in bioinformatics that many ways (including methods, tools and algorithms) of analyzing and integrating data have been created by researchers, but are buried in academic papers or by arcane coding, so that it is prohibitively difficult to share and reuse the efforts of predecessors.
- Automatic analysis should bear a greater role in analysis, since it is impossible to use nonautomated analysis methods to handle the increasing accumulation of biological data [38]. In order to realize this goal, therefore, data exchange should be based on computer-to-computer communication. That is, raw data should be accessible by computers via web services; and methods, tools and algorithms that are used to analyze data should be accessed as web services, too (that is, SaaS). This exchange can be defined as a pipeline with a combination of several web services (described in detail below).
- To facilitate data sharing, a pipeline involving raw data, analyzed methods, tools and algorithms can not only be shared among users, but can also be used as a medium to catch users' attention and to link people with similar research interests. Furthermore, these web services-based pipelines lower technological entrance barriers greatly for data integration, analysis and sharing.

Web 2.0 befits the rapid development of bioinformatics. To enact these three goals, we propose a Web 2.0-based SSC model for bioinformatics (Figure 1). In the SSC model, pipelines enable data integration and tool access through web services. Compared with existing efforts (such as, BioMoby [12] at http://biomoby.org/), web services in the proposed SSC model are not only limited to SOAP-based ones, but also include other formats of web services (e.g. JSON [29]). Web services-based pipelines provide users with a lightweight programming environment for easy data creation and sharing. Users may create pipelines (adding value), publish them online and subscribe to pipelines created by other users. In addition, users can blog about or even rank these pipelines. Consequently, pipelines may be widely shared, reused and even integrated into other pipelines. As in Web 1.0 sites, pipelines may also be searched by keywords or tags. As a result, communication and resource sharing among users can be greatly increased, making collaborations to discover knowledge through collective intelligence possible. In return, this success will encourage user participation and enhance/sustain SSC.

Every element of Web 2.0 is facilitated by participation in the SSC (Figure 1). Whereas Web 1.0

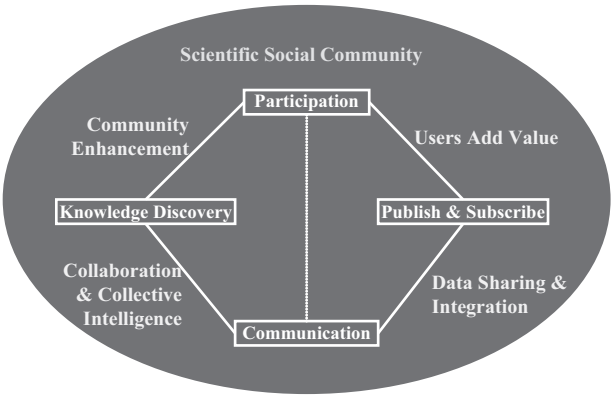

Figure 1: SSC: a Web 2.0-based model for bioinformatics. Web 2.0 encourages users' participation and users can publish and subscribe web services-based pipelines that facilitate data integration from multiple heterogeneous data sources. As a result, web services-based pipelines, regarded as user-added values, are widely shared among people. Consequently, people with similar research interests are linked together and communications among people increase, which makes knowledge discovery possible through collaboration and collective intelligence. In return, this encourages user participation to discover more knowledge and to enhance SSC. The dashed line between 'Participation' and 'Communication' means that some of users communicate with others without publishing and subscribing pipelines.

sites only provide information and do not involve users' participation, participation in Web 2.0 sites is enabled by the emergence of Web 2.0 technologies, such as blogging, tagging, RSS [25], AJAX (Asynchronous JavaScript And XML [39], a combination of technologies for creating highly interactive web applications), social networks, etc., that significantly simplify data provision and lower entrance barriers for user participation. The aim of SSC is to exploit the power of the community to achieve a goal. A successful example is Wikipedia [40], an online encyclopedia allowing any user to create and edit content. Efforts that accumulate current knowledge, like Wikipedia, have been extraordinary successes. Wikipedia features more content coverage than BBC (British Broadcasting Corporation) and CNN (Cable News Network) combined [41]. An attempted application of Wikipedia to genome reannotation was discussed recently and considered valuable [42]. Moreover, projects like Protein WikiProtein [43] and Gene Wiki [44] have been implemented to show the potential of bio-Wikis in action.

Communication and collaboration are of paramount importance to academic research; nevertheless, research activities at the cutting edge have been slow to adopt Web 2.0 technologies [45]. The proposed SSC model legitimizes the attempt to harness collective intelligence not just for knowledge deposition, but also for knowledge creation, maintenance and discovery. SSC is a generalized model for data integration/analysis/sharing and can be easily extended from field to field, so that while SSC is proposed for bioinformatics, its adoption may help catalyze paradigm-shifting advances in other fields of science as well. However, it is particularly appropriate for increasingly data-intensive and data-integrative bioinformatics studies. On the other hand, more SSCs from relevant academic fields are also needed for the furtherance of bioinformatics research due to its interdisciplinary nature. For example, a statistics SSC for estimating *P*-values, Fisher exact tests, correlation coefficients, etc., a numerical computation SSC for least-square fitting, linear regression, multidimensional minimization, etc. and a plotting SSC for different figures (scatter, line, bar, pie) with different formats (svg, eps, jpg, png), would all be endeavors that would enhance the research efforts of bioinformaticians as a collective.

Pipelines based on web services

As defined in the SSC model, a pipeline is based on web services which are accessed through Web APIs and executed on a remote computer hosting the requested services [46]. Due to the fast evolving data sources, web services-based pipelines aim to be automatable, reusable and repeatable [47–50]. A pipeline is a combination of 'widgets'—small, portable web applications that may be easily embedded into any web page, supporting lightweight programming and lightweight connections among web services. Therefore, SSC exploits the Service-Oriented Architecture (SOA) for web services-based pipelines, including three components: a service provider, a service management and a service requester. We present the architecture of web services and describe interactions between these components, which are defined by the WSDL (Figure 2).

To provide web services-based pipelines, the first step is to make Web APIs available that act as service providers, not browser-based web servers or databases. Hence, a fundamental requirement is that existing and upcoming databases and web servers expose their data and tools for reuse; that is, developing, describing and publishing their Web APIs (for example, using Service Composition and

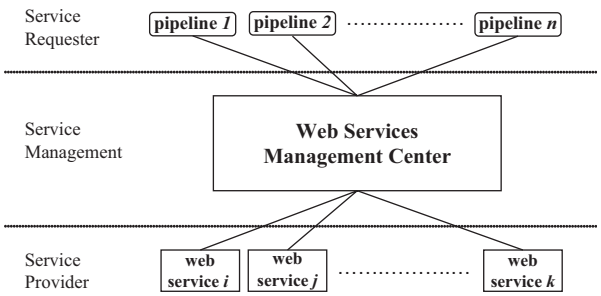

Figure 2: SOA for SSC that includes three components: (I) The service provider, who develops, describes and publishes web services, responding to service requests. (2) The service management, who registers web services and sends requests to service providers. (3) The service requester, who locates services defined in service management and submits service requests. All interactions between these three modules are described by WSDL.

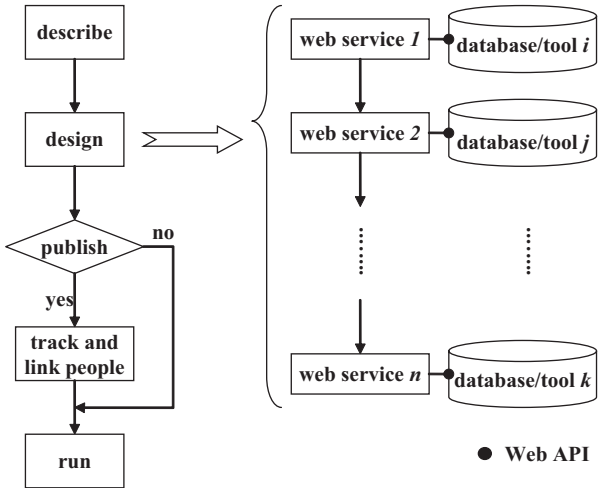

Figure 3: Flowchart of creating a web services-based pipeline. The flowchart mainly involves two parts. The left part is the overall procedure for creating a pipeline, including description, design, publication, connection with people through tracking and run. The right part is the details to design a web services-based pipeline that uses Web APIs (black circles) to access databases and tools.

Execution Tool [51] to develop web services) and responding to service requests. After setting up web services, the second step is to build the Web Services Management Center (WSMC) for service registrations and service requests. WSMC performs as a medium: accepting registrations sent from service providers and publishing them for service requesters. The third step involves the service requesters, who construct user-defined pipelines based on web services. Service requesters locate services in WSMC and send service requests to the service provider through the WSMC. As a result, similar pipelines are linked together, which can form a multiverse of SSCs involving diverse researchers with similar research interests.

A web services-based pipeline is a natural way to explore and manipulate data; users can easily construct pipelines involving several web services to solve complex biological tasks. In some cases existing efforts are similar to web services-based pipelines, while in other cases they are novel. One of the major features is data exchange through Web APIs. This exchange is computer-to-computer communication and, with suitable standardization, eases data integration from heterogeneous data sources. The essential point arises from the fact that information provided by browser-based databases or tools is for human use rather than for computer processing. In particular, with the rapid increase in volume and diversity of biological data, traditional user-to-computer communication for data integration requires that diverse data be presented for human 'consumption' in diverse human-readable ways. In contrast, web services are designed to enable computer-to-computer communication, and web services-based pipelines facilitate data integration in a fluid manner (Figure 3). To maximize the potential of future bioinformatics research efforts, existing and upcoming databases/tools should provide open data structures and Web APIs. As a result, web services-based pipelines would be able to access data and tools easily through Web APIs and there would be considerably diminished need to encode translations of data and queries or to spend time and money on procedural updates and maintenance efforts. Data integration may be achieved with far lower costs.

Based on the SSC model, web services-based pipelines can also facilitate data sharing and reuse. To create a new pipeline, users first need to provide the description and then to design the pipeline involving different web services (Figure 3). Users can set their pipelines as public or private, and track who uses them, leading to connections with people with similar interests. Web services-based pipelines maximize the scope for sharing. Accordingly, this pipelines-based sharing maximizes the Web potential for sharing (different from Taverna [15] that is only for a desktop environment). Yahoo Pipes [52] is an example of using the SSC model for sharing pipelines. Web services-based pipelines can be used

as 'Methods and Materials' in academic papers, so that they can be shared easily among researchers and consequently increase communications and collaborations among researchers. Web services provide more control for the tracking of who uses the service as well as how and why [24, 53]. Thus, researchers may also accurately evaluate the popularity and performance of tools through their web services. For instance, among several web services developed for performing 'BLAST' analysis, we may differentiate one with better performance, signified by its wider use and/or by its quality, as computed automatically [54]. Importantly, this evaluation may swiftly influence the direction of future development [55], as has CASP (Critical Assessment of Techniques for Protein Structure Prediction) [56], a community-wide competition to evaluate protein structure prediction algorithms.

CONCLUSION AND FURTHER PERSPECTIVES

Web 2.0 is a second generation of our web, emphasizing user participation and collaboration; it befits the rapid development of bioinformatics. Here, we have discussed the key elements of Web 2.0 for bioinformatics and proposed a SSC model, advocating establishment of platforms that encourage researcher participation and collaboration and harness collective intelligence for knowledge discovery. By comparison with existing related efforts, the proposed SSC allows researchers to do more than data retrieval, enabling data integration, automatic analysis and data sharing. Furthermore, it provides a medium for the creation and publication of bioinformatic pipelines based on web services. In the SSC, researchers with similar interests can be linked together by web services-based pipelines, and thus communication and collaboration among researchers can be greatly increased.

E-learning platform in education

Web 2.0 technologies may also aid education [57, 58], particularly by establishing a social community to increase students' participation and creativity. The SSC model proposed in this report may also serve as an 'e-learning' platform in which content is created, shared, remixed, repurposed and passed along by students in active ways. This model fits the paradigm of student-centered learning, which places the control of learning itself into the hands of the students. Moreover, web services-based pipelines may act as a medium for students to learn bioinformatics. Learning is then characterized not only by greater autonomy for the student, but also by a greater emphasis on practice, with creation, communication and participation playing key roles, and on a changed role for the teacher, at the extreme leading to a disappearance of the distinction between teacher and student altogether. The SSC model allows students to syndicate/aggregate content in creative ways. Such syndicated/aggregated content may then be fed forward to become fodder for other students' education and use.

Modularity and standardization

The SSC model proposed in this report would drive modularity and standardization in bioinformatics studies. The modularity would result naturally from the fact that for a given bioinformatics task, we need to use multiple web services that can be classified into different groups, such as a Sequence Retrieval Group including gene, coding and protein sequence retrievals with different species, an Alignment Group including pairwise and multiple sequences alignments, alignment visualization, a Structure Group including 2D and 3D structure predictions, 2D and 3D structure comparisons, 3D structure viewing, etc. To exchange data efficiently among web services, standardization is needed in the SSC model, such as nomenclature, data format, data model, interchange standards, etc. [59]. Rapid development of information technology has primarily been leveraged by the introductions of standards and protocols (for a key early example, consider Transmission Control Protocol/Internet Protocol (TCP/IP) developed in 1974 [60]). These standards lower social, legal and technical barriers for innovations. Likewise, modularity and standardization in bioinformatics would reduce net cost for computer-to-computer communication when running a pipeline, ease collaboration among researchers with different academic backgrounds and further promote the development of related life sciences.

Web 3.0 and bioinformatics

The Web will continue to evolve, and Web 2.0 is not the ultimate web, but an important stage to be achieved, since it brings with it new paradigms for social communication and collaboration [61]. In retrospect, Web 1.0 was 'read-only' (50K average band width, ABW) and Web 2.0 is

'read-write' (1 M ABW). With the increasing high performance/price ratio, many servers may be available to offer web services. What will this mean for Web 3.0 (10 M ABW)? Software in Web 2.0 is provided as a service, but it is possible that researchers will not be able to find the needed web services, especially for pipelines involving novel ideas and new methods. Therefore, an interface to upload user-defined software that is executed as a service is likely to be the not-too-distant future for bioinformatics studies. This uploaded software should be viewed not as a cost to the service provider, but as another way for users to add value. Hence, Web 3.0 will have the attributes of 'read-write-execute'. As a result, desktop programming will gradually be turned into web programming, and continued development of open Web APIs and protocols, open source software and open data (with enough eyeballs, all bugs are shadow [24]) will likely improve the qualities of communication and collaboration on the Web as well as increasing its volume (for example, the Bioperl project [62], an international open-source collaboration in life sciences). A further computationally appealing potential feature of Web 3.0 is the Semantic Web [63–66]. According to the statement of definition from the World Wide Web Consortium (W3C), the purpose of Semantic Web is to create a universal medium for the exchange of data, including several components, such as the Uniform Resource Identifier (URI), Resource Description Framework (RDF) core model, the RDF Schema language (RDFS) and the Web Ontology Language (OWL) [67]. A special interest group called 'Semantic Web for Health Care and Life Sciences Interest Group' (http://www.w3.org/2001/sw/hcls/) was established by W3C to explore the potential benefits of Semantic Web in the health care and life sciences domains [68]. Another is an interdisciplinary project named SWAN (Semantic Web Applications in Neuromedicine, evolving from the Alzheimer Research Forum [69]), aiming to use Semantic Web technologies to develop a practical, common, semantically structured, framework for scientific discourse [45]. SWAN, based on the scientific knowledge ecosystem, places attention on the social activity of the participants [70], which is one of key elements of Web 2.0.

Web 2.0 (and even Web 3.0) brings so much to bioinformatics and thus plays an increasingly important role. While the SSC proposed here aims for data exchange in bioinformatics, there is much to be done to enable the Web to reach its full potential, so that data over the Web may be reliably shared and processed by automated agents as well as by human users. To reach such an advanced state, we need to first shift the Web paradigm so that the Web becomes a platform and software runs as a service. The potential resulting benefits of Web 2.0 technologies for enhancement to current paradigms in bioinformatics research should not be underestimated, particularly with the successful establishment of a social, collective and collaborative platform, such as the SSC model proposed here.

Key Points

- Web 2.0 provides a revolutionary way of collecting and integrating biological data and knowledge repositories.
- Data in bioinformatics are not only limited to sequences, but also include methods, tools, algorithms, analyzed results, papers and even connections among people.
- SSC aims to exploit the collective intelligence for knowledge discovery, by encouraging users' participations, linking users with similar research interests and emphasizing communication and collaboration among users.
- Web services-based pipelines facilitate not merely data integration, but also automatic analysis and data sharing.

Acknowledgements

We thank two anonymous reviewers for their constructive comments on this article, Prof. Jun Yu, Zheng Wang, Francesc López, Aleksandra Adomas, Gina Wilpiszeski and Andrea Hodgins-Davis for valuable discussions on this article.

FUNDING

U.S. National Institutes of Health, awards (P01 DC04732 and U24 NS051869 to K.-H.C.); National Institute of General Medical Sciences (GM068087 to J.P.T.).

References

1. Fox JA, McMillan S, Ouellette BF. Conducting research on the web: 2007 update for the bioinformatics links directory. *Nucleic Acids Res* 2007;**35**:W3–5.
2. Editorial. Data, data everywhere. *Nat Struct Mol Biol* 2005;**12**:633.
3. Shadbolt N, Hall W, Berners-Lee T. The semantic Web revisited, intelligent systems. *IEEE* 2006;**21**:96–101.
4. Cheung KH, Smith AK, Yip KYL et al. Semantic Web approach to database integration in the life sciences. In: Baker CJO, Cheung KH (eds). *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*. New York: Springer, 2007, 11–30.
5. Stein LD. Integrating biological databases. *Nat Rev Genet* 2003;**4**:337–45.

6. Neerincx PB, Leunissen JA. Evolution of web services in bioinformatics. *Brief Bioinform* 2005;**6**:178–88.
7. Lee TJ, Pouliot Y, Wagner V, *et al*. BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics* 2006;**7**:170.
8. Birkland A, Yona G. BIOZON: a hub of heterogeneous biological data. *Nucleic Acids Res* 2006;**34**:D235–42.
9. Park J, Park B, Jung K, *et al*. CFGP: a web-based, comparative fungal genomics platform. *Nucleic Acids Res* 2008;**36**:D562–71.
10. Haas LM, Schwarz PM, Kodali P, *et al*. DiscoveryLink: a system for integrated access to life sciences data sources. *IBM Syst J* 2001;**40**:489–511.
11. Marenco L, Wang TY, Shepherd G, *et al*. QIS: a framework for biomedical database federation. *J Am Med Inform Assoc* 2004;**11**:523–34.
12. Wilkinson MD, Links M. BioMOBY: an open source biological web services proposal. *Brief Bioinform* 2002;**3**: 331–41.
13. Dowell RD, Jokerst RM, Day A, *et al*. The distributed annotation system. *BMC Bioinformatics* 2001;**2**:7.
14. BioDAS. http://www.biodas.org (25 March 2008, date last accessed).
15. Hull D, Wolstencroft K, Stevens R, *et al*. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res* 2006;**34**:W729–32.
16. Wilkinson MD, Senger M, Kawas E, *et al*. Interoperability with Moby 1.0—it's better than sharing your toothbrush!. *Brief Bioinform* 2008;**9**:220–31.
17. Kawas E, Senger M, Wilkinson MD. BioMoby extensions to the Taverna workflow management and enactment software. *BMC Bioinformatics* 2006;**7**:523.
18. Simple Object Access Protocol. http://en.wikipedia.org/wiki/Simple_Object_Access_Protocol (1 Feburary 2008, date last accessed).
19. Olason PI. Integrating protein annotation resources through the distributed annotation system. *Nucleic Acids Res* 2005;**33**: W468–70.
20. Prlic A, Down TA, Kulesha E, *et al*. Integrating sequence and structural biology with DAS. *BMC Bioinformatics* 2007; **8**:333.
21. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
22. Stevens RD, Robinson AJ, Goble CA. myGrid: personalised bioinformatics on the information grid. *Bioinformatics* 2003;**19**(Suppl. 1):i302–4.
23. World Wide Web. http://www.w3.org/WWW/ (1 February 2008, date last accessed).
24. O'Reilly T. What is Web 2.0: design patterns and business models for the next generation of software. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (1 Feburary 2008, date last accessed).
25. RSS. http://en.wikipedia.org/wiki/RSS (22 March 2008, date last accessed).
26. Shi X. Semantic web services: an unfulfilled promise. *IEEE IT Prof* 2007;**9**:42–5.
27. Web Services Description Language. http://en.wikipedia.org/wiki/Web_Services_Description_Language (1 Feburary 2008, date last accessed).
28. eXtensible Markup Language. http://en.wikipedia.org/wiki/Xml (1 Feburary 2008, date last accessed).
29. JSON. http://www.json.org/ (1 July 2008, date last accessed).
30. Google Documents. http://docs.google.com (1 February 2008, date last accessed).
31. Craigslist. http://www.craigslist.org/ (1 March 2008, date last accessed).
32. YouTube. http://en.wikipedia.org/wiki/YouTube (6 April 2008, date last accessed).
33. Greaves M. Semantic Web 2.0, intelligent systems. *IEEE* 2007;**22**:94–96.
34. Ritter O, Kocab P, Senger M, *et al*. Prototype Implementation of the Integrated Genomic Database. *Comput Biomed Res* 1994;**27**:97–115.
35. Thireou T, Spyrou G, Atlamazoglou V. A survey of the availability of primary bioinformatics web resources. *Genomics Proteomics Bioinformatics* 2007;**5**:70–6.
36. Merali Z, Giles J. Databases in peril. *Nature* 2005;**435**: 1010–11.
37. Clark T. Knowledge integration in biomedicine: technology and community. *Brief Bioinform* 2007;**8**:E1–3.
38. Sarkar IN, Egan MG, Coruzzi G, *et al*. Automated simultaneous analysis phylogenetics (ASAP): an enabling tool for phlyogenomics. *BMC Bioinformatics* 2008;**9**:103.
39. AJAX. http://en.wikipedia.org/wiki/AJAX (10 May 2008, date last accessed).
40. Wikipedia. http://www.wikipedia.com (1 March, 2008, date last accessed).
41. McLean R, Richards BH, Wardman JI. The effect of Web 2.0 on the future of medical practice and education: Darwikinian evolution or folksonomic revolution? *Med J Aust* 2007;**187**:174–7.
42. Salzberg SL. Genome re-annotation: a wiki solution? *Genome Biol* 2007;**8**:102.
43. Mons B, Ashburner M, Chichester C, *et al*. Calling on a million minds for community annotation in WikiProteins. *Genome Biol* 2008;**9**:R89.
44. Huss JW 3rd, Orozco C, Goodale J, *et al*. A gene wiki for community annotation of gene function. *PLoS Biol* 2008;**6**:e175.
45. Clark T, Kinoshita J. Alzforum and SWAN: the present and future of scientific web communities. *Brief Bioinform* 2007;**8**:163–71.
46. Web Services Activity Statement. http://www.w3.org/2002/ws/Activity.html (1 February 2008, date last accessed).
47. Potter SC, Clarke L, Curwen V, *et al*. The Ensembl analysis pipeline. *Genome Res* 2004;**14**:934–41.
48. Hoon S, Ratnapu KK, Chia JM, *et al*. Biopipe: a flexible framework for protocol-based bioinformatics analysis. *Genome Res* 2003;**13**:1904–15.
49. Bartocci E, Corradini F, Merelli E, *et al*. BioWMS: a web-based workflow management system for bioinformatics. *BMC Bioinformatics* 2007;**8**(Suppl. 1):S2.
50. Romano P. Automation of in-silico data analysis processes through workflow management systems. *Brief Bioinform* 2008;**9**:57–68.
51. Chandrasekaran S, Miller JA, Silver GS, *et al*. Performance analysis and simulation of composite web services. *Electronic Markets* 2003;**13**:120–32.

52. Yahoo Pipes. http://pipes.yahoo.com/pipes (12 May 2008, date last accessed).
53. Web 3.0: When Web Sites Become Web Services. http://alexiskold.wordpress.com/2007/03/19/web-30-when-web-sites-become-web-services/ (1 Feburary 2008, date last accessed).
54. Cardoso J, Sheth A, Miller J, *et al*. Quality of service for workflows and web service processes. *Web Semantics: Science, Services and Agents on the World Wide Web* 2004;**1**:281–308.
55. Editorial. Going for algorithm gold. *Nat Meth* 2008;**5**:659.
56. Critical Assessment of Techniques for Protein Structure Prediction. http://predictioncenter.gc.ucdavis.edu (1 July 2008, date last accessed).
57. Sandars J, Haythornthwaite C. New horizons for e-learning in medical education: ecological and Web 2.0 perspectives. *Med Teach* 2007;**29**:307–10.
58. Boulos MNK, Wheeler S. The emerging Web 2.0 social software: an enabling suite of sociable technologies in health and health care education. *Health Info Libr J* 2007;**24**:2–23.
59. Goble C, Stevens R. State of the nation in data integration for bioinformatics. *J Biomed Inform* 2008;**41**:687–693.
60. Cerf VG, Kahn RE. Protocol for packet network intercommunication. *IEEE Trans Commun* 1974;**Co22**:637–48.
61. Murugesan S. Understanding Web 2.0. *IEEE IT Prof* 2007;**9**:34–41.
62. Stajich JE, Block D, Boulez K, *et al*. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* 2002;**12**:1611–18.
63. Hendler J. Science and the semantic web. *Science* 2003;**299**:520–1.
64. Good BM, Wilkinson MD. The life sciences Semantic Web is full of creeps!. *Brief Bioinform* 2006;**7**:275–86.
65. Dibernardo M, Pottinger R, Wilkinson M. Semi-automatic web service composition for the life sciences using the BioMoby semantic web framework. *J Biomed Inform* 2008;**41**:752–65.
66. Lord P, Bechhofer S, Wilkinson MD, *et al*. Applying Semantic Web services to bioinformatics: experiences gained, lessons learnt. In: Mcilraith SA, Plexousakis D, Harmelen FV (eds). *Semantic Web - ISWC 2004*, Heidelberg: Springer/Berlin, 2004, 350–64.
67. Semantic Web Activity Statement. http://www.w3.org/2001/sw/Activity.html (1 Feburary 2008, date last accessed).
68. Cheung K-H, Yip KY, Townsend JP, *et al*. HCLS 2.0/3.0: health care and life sciences data mashup using Web 2.0/3.0. *J Biomed Inform* 2008;**41**:694–705.
69. Kinoshita J, Clark T. Alzforum. *Methods Mol Biol* 2007;**401**:365–81.
70. Ciccarese P, Wu E, Wong G, *et al*. The SWAN biomedical discourse ontology. *J Biomed Inform* 2008;**41**:739–751.
71. Web 1.0. http://en.wikipedia.org/wiki/Web_1.0 (1 February 2008, date last accessed).