ORIGINAL ARTICLE

UNRAVELING THE COMPLEXITIES OF LIFE SCIENCES DATA

Roger Higdon,[1–4] *Winston Haynes,*[1–4] *Larissa Stanberry,*[1–4] *Elizabeth Stewart,*[1,4] *Gregory Yandl,*[1,2,4] *Chris Howard,*[4,5] *William Broomall,*[2–4] *Natali Kolker,*[2–4] *and Eugene Kolker,*[1–4,6]

Abstract

The life sciences has entered into the realm of big data and data-enabled science, where data can either empower or overwhelm. These data bring with them the challenges of the 5 Vs of big data: volume, veracity, velocity, variety, and value. Both independently and through our involvement with DELSA Global (Data-Enabled Life Sciences Alliance International, DELSAglobal.org), the Kolker Lab is creating partnerships that identify data challenges and solve community needs. We specialize in solutions to complex biological data challenges, as exemplified by the community resource of MOPED (Model Organism Protein Expression Database, MOPED.proteinspire.org) and the analysis pipeline of SPIRE (Systematic Protein Investigative Research Environment, PROTEINSPIRE.org). Our work extends into the computationally intensive tasks of analysis and visualization of millions of protein sequences through innovative implementations of sequence alignment algorithms and creation of the Protein Sequence Universe tool (PSU). Pushing into the future, our lab is pursuing integration of multi-omics data and exploration of biological pathways, as well as assigning function to proteins and porting solutions to the cloud. Big data have come to the life sciences; discovering the knowledge in the data will bring breakthroughs.

Introduction

With the Completion of the Human Genome Project and the advent of high-throughput analysis technologies, 21st-century life sciences has entered the fourth paradigm of data-enabled science and the realm of big data.[1] These data will enable incredible possibilities for discovery, solutions, and even cures, yet they also bring with them the challenges of the 5 Vs of big data: volume, veracity, velocity, variety, and value. Life sciences' big data are generally challenging for their variety, value, and critical need for veracity. This differs from the situation in marketing or banking, where efforts for optimization find volume and velocity to be generally the biggest challenges. It is disheartening to realize that in this internet age, when information about pizza restaurants is at our fingertips, it can be a challenge to find crucial drug trial information and that precious resources (both people and funding) cannot be fully utilized due to inadequate cyberinfrastructure and organization.[2,3] Just as the life science community's tools and analyses need to be at their most robust to control the "data deluge," they are instead stagnating, (e.g. clusters of orthologous groups of proteins, COGs) or even ending, (e.g. Peptidome & The Arabidopsis Information Resource).[4–7]

The scale of biological data is exponentially increasing with sequencing technologies now producing data at a rate exceeding the growth in computing power predicted by Moore's

[1]Bioinformatics and High-throughput Analysis Laboratory, and [2]High-throughput Analysis Core, Center for Developmental Therapeutics, Seattle Children's Research Institute, Seattle, WA. [3]Predictive Analytics, Seattle Children's Hospital, Seattle, WA. [4]Data-Enabled Life Sciences Alliance International (DELSA Global), Seattle, WA. [5]Center for Developmental Therapeutics, Seattle Children's Research Institute, Seattle, WA. [6]Departments of Biomedical Informatics & Medical Education and Pediatrics, University of Washington, Seattle, WA.

Law (10,000-fold improvement in sequencing *vs.* 16-fold improvement in computing over Moore's Law).[8,9] In addition, the majority of research is generated in isolation and demonstrates only an 11% rate of reproducibility according to a recent study.[10] Moreover, 27% (+/-9%) of cancer cell lines are misidentified, one out of three proteins is unannotated, and according to one report, up to 85% of research efforts are wasted due to inadequate production and reporting practices.[10–13]

Beyond the obvious issues of scale and reproducibility, the complexity and diversity of these data poses the greatest challenge to unlocking knowledge and scientific discovery. Modern biological data spans a diverse collection of omics fields, including genomics, metagenomics, proteomics, transcriptomics, metabolomics, and lipidomics. These omics data are generated by various types of high-throughput technologies, including, for example, next-generation sequencing, mass spectrometry, imaging, arrays, liquid chromatography, and flow-cytometry. Relatively simple experiments generate data on the terabyte scale. Supporting storage and analysis of these data requires massive amounts of computational power linked to an endless array of databases, data formats, software packages, and pipelines. In addition to these requirements, we need comprehensive understanding of the metadata, experimental protocols, and standard operating procedures unique to each omics and high-throughput technology.

"BEYOND THE OBVIOUS ISSUES OF SCALE AND REPRODUCIBILITY, THE COMPLEXITY AND DIVERSITY OF THESE DATA POSES THE GREATEST CHALLENGE TO UNLOCKING KNOWLEDGE AND SCIENTIFIC DISCOVERY."

As the first step to unraveling the complexities of biology, it is necessary to understand the needs and challenges of the life sciences community with respect to data-enabled science. DELSA Global (Data-Enabled Life Sciences Alliance International, delsaglobal.org) was formed to accelerate the impact of data-enabled life sciences research on the pressing needs of the global society. DELSA Global was forged on the basis of the Data-Intensive Science Workshops (DISW-I and II) in 2010–2011, sponsored by the National Science Foundation (NSF) with matching support by Seattle Children's Research Institute (SCRI).[2,14–20] The workshops were attended by experts from many disciplines, representing academia, government, nonprofit research institutes, private enterprise, policy-making bodies, and media.

Currently, DELSA Global is building an ecosystem to provide a leading voice and coordinating framework for collective innovation in data-enabled science for the life sciences community. The alliance has endorsed eight high-impact projects that are poised to advance its mission and inspire new modes of business and innovation. As one of the founding members of DELSA Global, the Kolker Lab strives to understand the needs of the life sciences community in order to develop effective solutions for complex biological data challenges.

Surveying the Life Sciences Community

The Kolker Lab has carried out a number of initiatives to determine the needs of the life science community. These initiatives include (1) a survey of proteomics researchers in the United States by the University of Washington Business School to assess data and analysis needs, (2) organizing DISW-I and DISW-II, and (3) leading efforts to found and promote DELSA.

As part of the marketing evaluation plan under our current NSF project, University of Washington MBA students surveyed life scientists and proteomics experts. The survey indicated the immediate need for tools and resources to easily access publicly available proteomics experiments. In particular, for biomedical researchers unfamiliar with mass spectrometry technology, the important criteria included reliable data, statistically valid results, analysis tools with a user-friendly interface, transparent reporting of results, and the ability to share data. This survey led directly to the development of MOPED, the Model Organism Protein Expression Database (for details, see below).[21]

The DISW-I identified three top challenges and opportunities: (1) the research necessity of the life sciences community to integrate work across diverse domains and with computer and data experts, (2) a pressing need for reproducibility because of its critical importance toward scientific progress and the accelerated rate of raw data production, and (3) a perceived gap between the needs of the data-enabled life sciences and current funding initiatives and merit evaluation criteria.[15–20] The DISW-II proposed to establish a community alliance with the goals to (1) synergize research and educational efforts across the life sciences using contemporary computing approaches to comprehend large and diverse data, (2) become an integral part of the international and national developments to address the challenges and explore opportunities of data-enabled sciences, and (3) cohesively address the community needs through creation of the supporting ecosystem of federal agencies, foundations, academia, and industry.[2]

The opportunities and challenges of big data in life sciences research compelled the participants to found DELSA Global.[2,22,23] Through intense discussions and formal and informal surveys, the newly formed alliance has made significant efforts to identify the strategic needs of the life sciences community and ways to address them.

Solutions for Complex Biological Data

Integrated data resources

To simplify the comparison and sharing of proteomics data, enable knowledge discovery, and generate new hypotheses, the

MARY ANN LIEBERT, INC. • DOI: 10.1089/big.2012.1505

Kolker Lab developed MOPED (moped.proteinspire.org).[21] MOPED provides concise summaries of protein identification, relative and absolute (concentration, ng/mL) expression, and other quantitative data from standardized analysis of model organism studies. MOPED supports querying, browsing, and visualizing data across organisms, tissues, conditions, and pathways (Fig. 1). It also links to protein and pathway databases, including Entrez, GeneCards, KEGG (Kyoto Encyclopedia of Genes and Genomes), Reactome, and UniProt.[24–28]

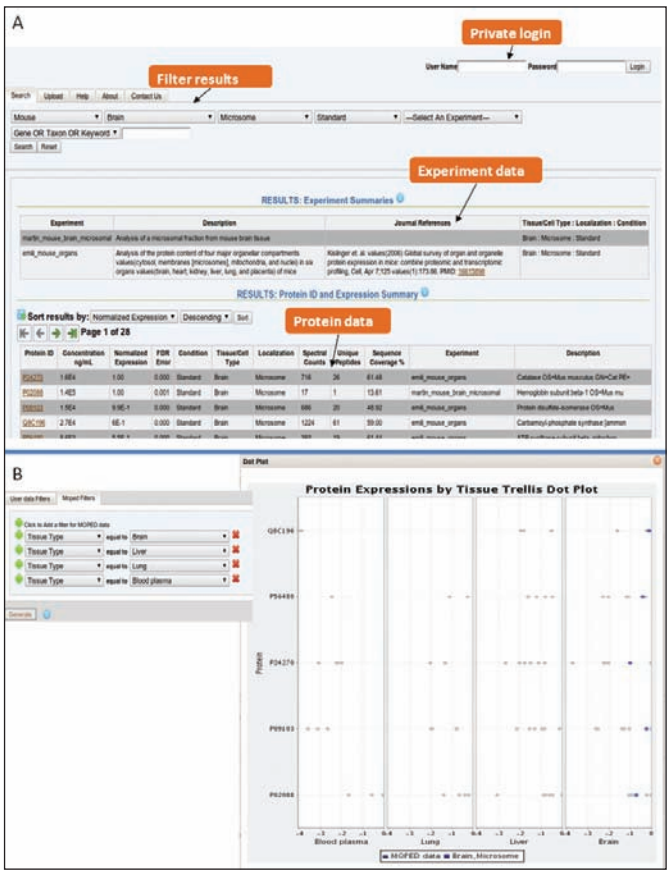

FIG. 1. Model Organism Protein Expression Database (MOPED; moped.proteinspire.org). (A) MOPED provides querying of protein identification and expression data across different organisms, tissues, and localizations for different experiments. (B) MOPED provides tools for visualizing individual experiment data relative to the data within MOPED.

Currently, MOPED 2.1 contains over 43,000 proteins with at least one spectral match and more than 11 million high-certainty spectra. MOPED is continuously updated and enhanced with the next major release scheduled for November 2012. In addition to public access, MOPED provides a private entry that allows users to share and explore their data prior to publication. According to Google Analytics, within the past year MOPED has had nearly 20,000 unique users from over 90 countries.

Biological data pipelines

Complex high-throughput biological data typically require analytical pipelines to process, integrate, and analyze data. With NSF support, the Kolker Lab has developed the Systematic Protein Investigative Research Environment (SPIRE, proteinspire.org).[29,30] SPIRE was designed in response to the community need for a reliable and simple yet powerful and flexible proteomics analysis pipeline. SPIRE integrates the best open-source search tools and data analysis methods for mass spectrometry proteomics analysis, such as X!Tandem, Open Mass Spectrometry Search Algorithm (OMSSA), and a composite search. Novel analysis methods implemented in SPIRE produce a 50 to 85% increase in protein IDs over other current combinations of scoring and single search engines, while also providing accurate multifaceted error estimation. Through MOPED, SPIRE combines analysis results with data on protein function, pathways, and protein expression from model organisms and also connects results to publicly available proteomics data. SPIRE is used as a standardized environment for the processing and analysis of proteomics data for MOPED.[21]

Determining the function of protein sequences

Functional annotation of newly sequenced genomes and metagenomes is one of the principal challenges in the life sciences. Rapidly advancing sequencing technologies are exponentially expanding the protein sequence universe (PSU).[31] Without updated methods in functional and comparative genomics, comprehensive approaches for assigning functional annotation to genes/proteins could not keep up with the ever-expanding size of the sequence universe (e.g., the prominent COG database).[4] There has never been a greater need for a scalable and efficient computational resource to visualize, explore, and assign biological meaning to new proteins.

All-versus-all sequence alignments

Our laboratory completed the first of a kind all-versus-all sequence alignment for 9.9 million proteins in the UniRef100 database.[32,33] The alignment was done on the Microsoft Windows Azure cloud system[34] with 475 eight-core virtual machines that produced over 3 billion filtered records in six days. Using the normalized alignment score, we have assigned 68% of 5.1 million bacterial proteins into clusters from the COG database.[32] The remaining proteins were classified into functional groups using an innovative implementation of a single-linkage algorithm on a Hadoop computing cluster using Hive and the MapReduce paradigm.[35,36] This implementation significantly reduced the run time for nonindexed queries and optimized clustering performance.[32] Consequently, nearly 2 million proteins were combined into half a million functional groups. Similarly, the eukaryotic database was expanded by over 1 million proteins with unclustered proteins classified into 100,000 new functional groups.[32] (Fig. 2).

The UniRef100 clustering project showcased both the promise and the challenges of large biological data. The project took the considerable efforts of an unusually diverse group of researchers along with multiple cloud systems to successfully complete the task. Publicly available cluster resources are struggling to cope with

influx of data and, as a result, are either no longer supported[37–39] or provide limited interactive and analytic capabilities.[40,41] These problems highlight the pressing need in the biological community for a scalable and efficient computational approach to visualize, explore, and analyze large-scale biological data.

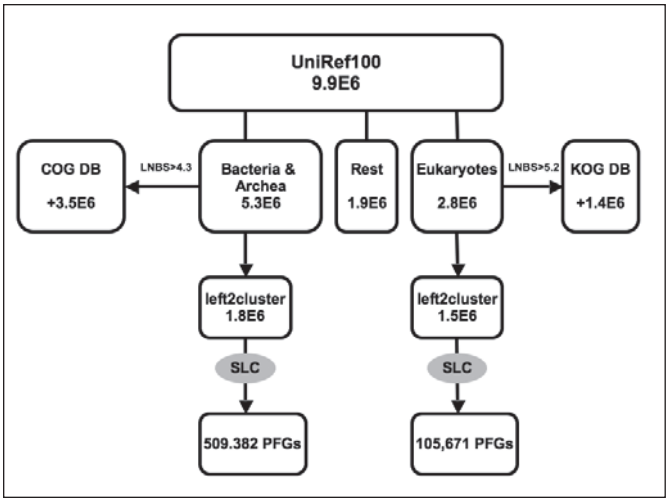

FIG 2. Protein classification paradigm. A total of 5 out of 10 million proteins were assigned to different COG/KOG functional categories. The remaining proteins were clustered-based all-versus-all BLAST alignments.

Visualizing complex biological data

The PSU visualization platform enables exploration, analysis, and annotation of the continually expanding universe of protein sequences. The platform uses a multidimensional scaling of sequence alignment scores to create a three-dimensional representation of the protein universe. The PSU preserves important grouping features such as relative proximity of functionally similar clusters and clear structural separation between protein groups of specific and general functions. The PSU is scalable, integrates different similarity measures with functional and experimental data, and facilitates sequence annotation. It will enable users to analyze new data in the context of the existing knowledge of protein sequences using a library of scientific tools.[31,42] An example comparing sequence similarity among different COG functional classifications is shown in Figure 3.

Community outreach and education

The public and many scientists lack information on the role of bioinformatics, statistics, and high-throughput technologies in biological knowledge discovery. This lack of information creates a great need for education and outreach in data-enabled sciences. One of the principal missions of DELSA Global is outreach and education programs. To fulfill this mission, DELSA Global has endorsed two projects: *Social Networking Platform for Tool Brokering/Community Building and Matchmaking and Training Data Scientists*.

In addition to the involvement in DELSA Global activities, the Kolker Lab has developed a number of educational resources, including instructional videos on proteomics data analysis in SPIRE;[29,30] a series of articles on statistical and bioinformatics concepts for the *Encyclopedia of Systems Biology*;[43] and interactive exhibits on proteomics for grades K–12, demonstrating the scientific principles of mass spectrometry (MS) and chromatography.[44] In addition, the lab has led working groups on education and outreach at a DELSA Global meeting and DISWs.[20]

Future Challenges for Complex Biological Data

Multi-omics integration

The biological functions of organisms depend on complex and highly interactive systems of biomolecules, including DNA, RNA, proteins, metabolites, and lipids. These biomolecules are rapidly being characterized by new high-throughput multi-omics data from genomics, metagenomics, transcriptomics, proteomics, metabolomics, and lipidomics experiments. Future data-enabled biological discoveries will require high-throughput data to be integrated and analyzed jointly across multi-omics experiments. However, current public databases and analysis tools typically focus on a single omics (principally genomics), biomolecule, or organism, overlooking the complex interrelationships of systems biology. Development of valuable integrated resources is challenging due to the 5 Vs of big data. The scale of the data and the complexity of the technologies, formats, ontologies, and methodologies come together in a whirlpool of potentially useful, but often bewildering, cross-references. To meet these challenges, we will transform our current single-omics resources into a new Multi-Omics Profiling Expression Database that integrates data across omics experiments. With this resource, meta-analysis studies, such as was achieved once for yeast, will be more easily performed, and similar approaches can be expanded to other organisms.[45]

> "IT IS CLEAR THAT THE LIFE SCIENCES HAVE BECOME BIG DATA AND DATA-ENABLED SCIENCES."

Pathway analysis

Most analysis tools for complex biological data do not take into account the wealth of available biological knowledge. For example, current pathway analysis models largely ignore the underlying graph structure of a pathway and the catalytic/inhibitory relationships it implies. Discarding this information reduces the power of the analysis and prevents testing the correct hypothesis of the pathway effect on expression levels. To address this limitation, the Kolker Lab has developed Differential Expression Analysis for Pathways (DEAP). The new pathway analysis approach utilizes

MARY ANN LIEBERT, INC. • DOI: 10.1089/big.2012.1505

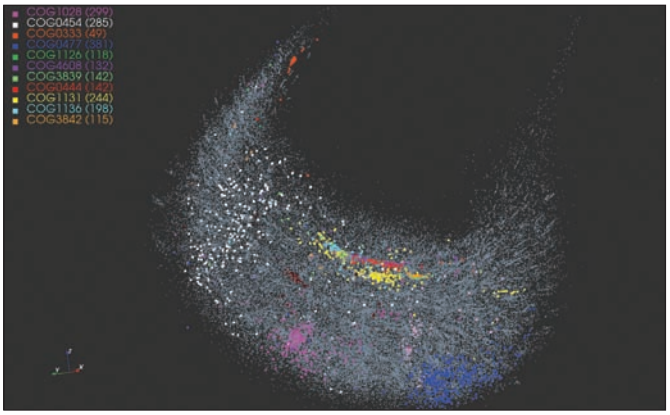

FIG. 3. The Protein Sequence Universe (PSU). The PSU tool provides visualization of complex protein sequence data. In this example, COG functional category is overlaid across sequence similarity data. The tightly clustered COG categories at the center of the graphic represent very similar functions (ABC-type ATP-ases), while the more dispersed COG categories represent disparate and more general functional categories.

information on pathway structure to test the enrichment hypothesis and to compare expression patterns across the conditions.

Annotation resources

Sequence orthology has long been utilized to denote functional similarity; as such, clusters of orthologous sequences are used to extend functional annotation of genes and proteins. One prominent example of such a resource is the COG database, whose papers have been cited more than 4,500 times (REF). Currently, the prokaryotic COG database contains over 190 thousand proteins grouped into 4,873 clusters. However, despite the high number of citations and user volume, high sustainability and maintenance costs have forced the resource to become stagnant. As it stands, the COG database has not been updated since 2006. This is another example of unsustained resources, a pressing issue recently reviewed in "The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection."[6]

With the rapid increase in data volumes, maintaining and enhancing a resource like COG requires new powerful technologies and always-increasing computing resources. This requirement was highlighted by all-versus-all sequence alignment project that was made possible through combined efforts with Microsoft Research and Amazon Web Services.[32] The skills and lessons acquired in this project have inspired Kolker Lab to further our collaborative work on revitalizing and expanding the COG database.

Cloud computing

Cloud technologies offer a viable solution to data-intensive science through scalable computing capabilities and large data storage.[8,46–48] In addition, the large scale of data intensifies the need for easy and efficient access to the analysis software and bioinformatics tools. The informal DELSA Global community survey showed a strong appeal to adopting the successful industry framework of apps stores (e.g., Amazon, Apple, Google, Microsoft) to allow for better dissemination and adaptation of biological and multi-omics research tools. Finally, prize competitions are a powerful and underutilized way to accelerate and deepen scientific discoveries and drive development of new tools and applications.[49,50]

Conclusions

It is clear that the life sciences have become big data and data-enabled sciences. Data-enabled science may have at its core the generation of data in the lab, but transforming the data to knowledge and then action goes far beyond the lab. The transformation will require massive resources and a transdisciplinary effort put forth by the scientific community to solve the challenges of big data. The need is urgent and growing, given the issues of data generation outstripping computing power and the lack of reproducibility of research. Organizations like DELSA Global can inform the life sciences community, lead the way for groups like the Kolker Lab to put forth new solutions to big data challenges, and create a new paradigm in the life sciences of cooperation, collaboration, and sharing at every level.

Acknowledgments

We would like to thank the following members of the University of Washington MBA Field Study program under the direction of Gordon Neumiller, who were instrumental in identifying the community need for an integrated resource such as MOPED: Jennifer Bragg, Lucas Donigian, Laura Kay, Natalia Perez, and Adam Ware. We sincerely appreciate stimulating discussions with Chris Moss, Randy Salamon, Evelyne Kolker, Maggie Lackey, Courtney MacNealy-Koch, Geoffrey Fox, Vural Ozdemir, Phil Bourne, Peter Arzberger, Corinna Gries, Dan Atkins, Doron Lancet, Rob Arnold, Jack Faris, Michael Galperin, Skip Smith, Tom Hansen, Jim Hendricks, and Chris Mentzel. Research reported in this publication was supported by the National Science Foundation under the Division of Biological Infrastructure award 0969929, National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under awards U01-DK-089571 and U01-DK-072473, The Robert B. McMillen Foundation award, and The Gordon and Betty Moore Foundation award to E.K. This support is very much appreciated. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation, National Institutes of Health, The McMillen Foundation, or The Moore Foundation.

Disclosure Statement

The authors declare no competing financial interests exist.

References

1. Hey T., Tansley S., and Tolle, K. The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA: Microsoft Research 2009.

2. Kolker, E., Stewart, E., and Ozdemir, V. Opportunities and challenges for the life sciences community. OMICS: A Journal of Integrative Biology 16, 138–147 (2012).
3. Kolker, E., and Stewart, E. Data to knowledge to action. Scientist April 25 (2012).
4. Natale, D. A., et al. Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). Genome Biol 1, 1-19 (2000).
5. Slotta, D. J., Barrett, T., and Edgar, R. NCBI Peptidome: a new public repository for mass spectrometry peptide identifications. Nat Biotechnol 27, 600–601 (2009).
6. Galperin, M. Y., and Fernández-Suárez, X. M. The 2012 nucleic acids research database issue and the online molecular biology database collection. Nucleic Acids Res 40, D1–D8 (2012).
7. Rhee, S. Y. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Research 31, 224–228 (2003).
8. Stein, L. D. The case for cloud computing in genome informatics. Genome Biology 11, 207 (2010).
9. Moore, G. E. Cramming more components onto integrated circuits. Electronics 38 (1965).<AQ6>
10. Begley, C. G., and Ellis, L. M. Drug development: raise standards for preclinical cancer research. Nature 483, 531–533 (2012).
11. Chalmers, I., and Glasziou, P. Avoidable waste in the production and reporting of research evidence. The Lancet 374, 86–89 (2009).
12. Bork, P. Powers and pitfalls in sequence analysis: the 70% hurdle. Genome Res. 10, 398–400 (2000).
13. Zerbino, D. R., Paten, B., and Haussler, D. Integrating Genomes. Science 336, 179–182 (2012).
14. Kolker, E. Special issue on data-intensive science. OMICS: A Journal of Integrative Biology 15, 197–198 (2011).
15. Barga, R., et al. Bioinformatics and data-intensive scientific discovery in the beginning of the 21st century. OMICS: A Journal of Integrative Biology 15, 199–201 (2011).
16. Bernstein, P. A., et al. Technology and data-intensive science in the beginning of the 21st century. OMICS: A Journal of Integrative Biology 15, 203–207 (2011).
17. Faris, J., et al. Communication and data-intensive science in the beginning of the 21st century. OMICS: A Journal of Integrative Biology 15, 213–215 (2011).
18. Ozdemir, V., et al. Policy and data-intensive scientific discovery in the beginning of the 21st century. OMICS: A Journal of Integrative Biology 15, 221–225 (2011).
19. Smith, A., et al. Biology and data-intensive scientific discovery in the beginning of the 21st century. OMICS: A Journal of Integrative Biology 15, 209–212 (2011).
20. Wolf, F., et al. Education and data-intensive science in the beginning of the 21st century. OMICS: A Journal of Integrative Biology 15, 217–219 (2011).
21. Kolker, E., et al. MOPED: Model Organism Protein Expression Database. Nucleic Acids Res 40, D1093–1099 (2012).
22. Kolker, E., Stewart, E., and Özdemir, V. DELSA Global for 'Big Data' and the bioeconomy: catalyzing collective innovation. Industrial Biotechnology 8, 176–178 (2012).
23. Ozdemir, V., et al. Towards an ecology of collective innovation: Human Variome Project (HVP), Rare Disease Consortium for Autosomal Loci (RaDiCAL) and Data-Enabled Life Sciences Alliance (DELSA). Current Pharmacogenomics and Personalized Medicine 9, 243–251 (2011).
24. Maglott, D. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research 33, D54–D58 (2004).
25. Stelzer, G., et al. In-silico human genomics with GeneCards. Hum Genomics 5, 709–717 (2011).
26. Ogata, H., et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 27, 29–34 (1999).
27. Joshi-Tope, G., et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 33, D428–432 (2005).
28. Bairoch, A., et al. The Universal Protein Resource (UniProt). Nucleic Acids Res. 33, D154–159 (2005).
29. Kolker, E., et al. SPIRE: Systematic protein investigative research environment. J Proteomics 75, 122–126 (2011).
30. Kolker, E., et al. Corrigendum to 'SPIRE: Systematic Protein Investigative Research Environment' [J Proteomics 75, 122–126 (2011)]. J Proteomics 75, 3789 (2012).
31. Stanberry, L., et al. Visualizing the Protein Sequence Universe. HPDC '12 Proceedings of the 21st ACM International Symposium on High Performance Distributed Computing (In press.).
32. Kolker, N., et al. Classifying Proteins into Functional Groups Based on All-versus-All BLAST of 10 Million Proteins. OMICS: A Journal of Integrative Biology 15, 513–521 (2011).
33. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23, 1282–1288 (2007).
34. Barga R. Windows Azure at http://www.windowsazure.com/ October 5 (2012)
35. Shvachko, K., Kuang, H., Radia, S., and Chansler, R. The Hadoop Distributed File System. Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), 1–10 (2010); doi:10.1109/MSST.2010.5496972.
36. Dean, J., and Ghemawat, S. MapReduce. Communications of the ACM 51, 107 (2008).
37. Tatusov, R. L., et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41 (2003).
38. Kriventseva, E. V., Fleischmann, W., Zdobnov, E. M., and Apweiler, R. CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins. Nucleic Acids Res. 29, 33–36 (2001).
39. Krause, A., Stoye, J., and Vingron, M. The SYSTERS protein sequence cluster set. Nucleic Acids Res. 28, 270–272 (2000).
40. Jensen, L. J., et al. eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res. 36, D250–254 (2008).
41. Klimke, W., et al. The National Center for Biotechnology In-

MARY ANN LIEBERT, INC. • DOI: 10.1089/big.2012.1505

formation's Protein Clusters Database. Nucleic Acids Res. 37, D216–223 (2009).
42. Fox G. The COG Project at http://manxcatcogblog.blogspot.com/ October 5 (2012)
43. Stanberry, L., Haynes, W., Higdon, R., and Kolker, E. Statistical Methods in Systems Biology (contributing authors). In: Encyclopedia of Systems Biology, Eds. Dubitzky W, Wolkenhauer O, Yokota H, Cho KH, Springer Publishing, New York City, NY, In Press (2013).
44. Haynes, W., et al. Basic Principles of Proteomics: Mass Spectrometry and Column Chromatography Explained. Presented at Annual Mobile Laboratory Coalition Conference, Seattle, June 22-25 (2012).
45. Higdon, R., Haynes, W., and Kolker, E. Meta-analysis for protein identification: a case study on yeast data. OMICS: A Journal of Integrative Biology 14, 309–314 (2010).
46. Bateman, A., and Wood, M. Cloud computing. Bioinformatics 25, 1475 (2009).
47. Langmead, B., Schatz, M. C., Lin, J., Pop, M., and Salzberg, S. L. Searching for SNPs with cloud computing. Genome Biol. 10, R134 (2009).
48. Stein, D. P., et al. Cloud computing for comparative genomics. BMC Bioinformatics 11, 259 (2010).
49. Carpenter, J. May the best analyst win. Science 331, 698–699 (2011).
50. Allio, R. J. CEO interview: the InnoCentive model of open innovation. Strategy & Leadership 32, 4–9 (2004).

Address correspondence to:

Eugene Kolker
Seattle Children's Research Institute
1900 9th Avenue
Seattle, WA 98105

eugene.kolker@seattlechildrens.org