Editorial

Computer Science, Biology and Biomedical Informatics Academy: Outcomes from 5 Years of Immersing High-school Students into Informatics Research

Andrew J. King[1], Arielle M. Fisher[1], Michael J. Becich[1], David N. Boone[1]

[1]Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA

Received: 09 September 2016 Accepted: 03 December 2016 Published: 28 February 2017

Abstract

The University of Pittsburgh's Department of Biomedical Informatics and Division of Pathology Informatics created a Science, Technology, Engineering, and Mathematics (STEM) pipeline in 2011 dedicated to providing cutting-edge informatics research and career preparatory experiences to a diverse group of highly motivated high-school students. In this third editorial installment describing the program, we provide a brief overview of the pipeline, report on achievements of the past scholars, and present results from self-reported assessments by the 2015 cohort of scholars. The pipeline continues to expand with the 2015 addition of the innovation internship, and the introduction of a program in 2016 aimed at offering first-time research experiences to undergraduates who are underrepresented in pathology and biomedical informatics. Achievements of program scholars include authorship of journal articles, symposium and summit presentations, and attendance at top 25 universities. All of our alumni matriculated into higher education and 90% remain in STEM majors. The 2015 high-school program had ten participating scholars who self-reported gains in confidence in their research abilities and understanding of what it means to be a scientist.

Keywords: Bioinformatics, biology, biomedical informatics, computer science, engineering, math, pathology informatics, science, technology

Introduction

The Agency for Healthcare Research and Quality and the National Library of Medicine (NLM) recognize the need for the education of researchers interested in working in biomedical informatics and related fields.[1,2] The current lack of skilled workers and researchers trained in informatics is confounded by the increase in the utilization of health information technology[3] and deficiencies in the current educational programs.[4-7] To meet the plethora of career opportunities in this developing discipline, we must try to understand how best to educate its workforce.[5] Magana *et al*. emphasize the importance of preparing professionals at early stages including developing a pipeline "starting at or even before the high-school level."[4] In addition, others suggest that to engage the best and brightest students, "there needs to be a fundamental change in how computer science is taught in our schools."[6] Research indicates that a critical component of increasing STEM persistence is providing early access to mentors, which opens pathways to the achievement of young people through encouragement and guidance.[8]

Several studies demonstrate that benefits resulting from early mentorship in STEM fields can last a lifetime and help improve achievement from students in underserved communities.[8-10] Many studies highlight the importance of mentorship for undergraduate student development to address academic and social needs, increase focus and motivation toward achieving learning goals, and ultimately facilitate their retention in STEM majors.[11] Mentoring relationships are particularly important for students from underrepresented populations.[9,11] While it

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/2153-3539.201110

Address for correspondence: Dr. David N. Boone, 5607 Baum Boulevard, Room 433, Pittsburgh, PA, 15206. E-mail: dnb14@pitt.edu

How to cite this article: King AJ, Fisher AM, Becich MJ, Boone DN. Computer science, biology and biomedical informatics academy: outcomes from 5 years of immersing high-school students into informatics research. J Pathol Inform 2017;8:2.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2017/8/1/2/201110

is clear that undergraduate mentorship increases persistence in STEM fields, we believe that introducing mentors at an earlier age will improve students' confidence – especially those from underrepresented or disadvantaged backgrounds – in their ability to perform research and will encourage students to choose and maintain a STEM major as undergraduates. A panel at the U.S. News STEM Solutions Conference emphasized the importance of mentoring high-school students as many of them may be unaware of their aptitude and ability to achieve in STEM.[8] In addition to the academic benefits, mentoring high-school students also leads to decrease in absenteeism, tardiness, and bullying.[10]

The pipeline at the University of Pittsburgh's Department of Biomedical Informatics is designed to provide a diverse group of students the opportunity to develop their skills through faculty mentorship and impactful authentic research in an innovative field. Individual mentorship is tailored to students based on their relative interests in the discipline and includes the provision of career, social, and emotional support in a setting intended to promote self-exploration. We aim to engage students as early as high school and continue to foster and modify this relationship as they plan and progress through their undergraduate studies. Our goal is to increase interest and persistence in STEM fields – especially in biology, biomedical, and pathology informatics.

Overview of Expanding Pipeline

Computer science, biology and biomedical informatics summer academy

All facets of the pipeline are directed by a full-time faculty member and a team of graduate students led by a graduate teaching fellow. The pipeline [Figure 1] initiates at the high-school level with the Computer Science, Biology and Biomedical Informatics (CoSBBI) Summer Academy. CoSBBI is a part of the University of Pittsburgh Cancer Institute (UPCI) Academy (http://www.upci.upmc.edu/UPCIAcademy/index.cfm#1). The mission of the UPCI Academy is to provide cutting-edge research and career preparatory experiences to a diverse group of motivated high-school students who are pursuing higher education and careers in STEM fields, especially research and medicine. A direct goal of the UPCI Academy and CoSBBI is to increase the number of students prepared for STEM fields as undergraduates from underrepresented populations, as defined by the National Institutes of Health (NIH) (http://www.ninds.nih.gov/diversity_programs/definitions.htm) to include underrepresented racial/ethnic groups, individuals with disabilities, and individuals from disadvantaged backgrounds based on income or demonstrably prohibitive social, cultural, or educational environments. While striving to meet the UPCI Academy mission, the CoSBBI program – as described previously[12,13] – introduces high-school scholars of all backgrounds to their first informatics research experiences. Each scholar is paired with a faculty research mentor and works on an individual research project designed by the mentor and in line with their research interests. Scholars are individually matched to mentors based on their interests as described in application essays. In addition to the authentic research, CoSBBI scholars are exposed to various academic and career development experiences. For example, the scholars receive career spotlight discussions from faculty members and outside speakers detailing their unique careers and career paths. The students also receive workshops intended to enhance diversity

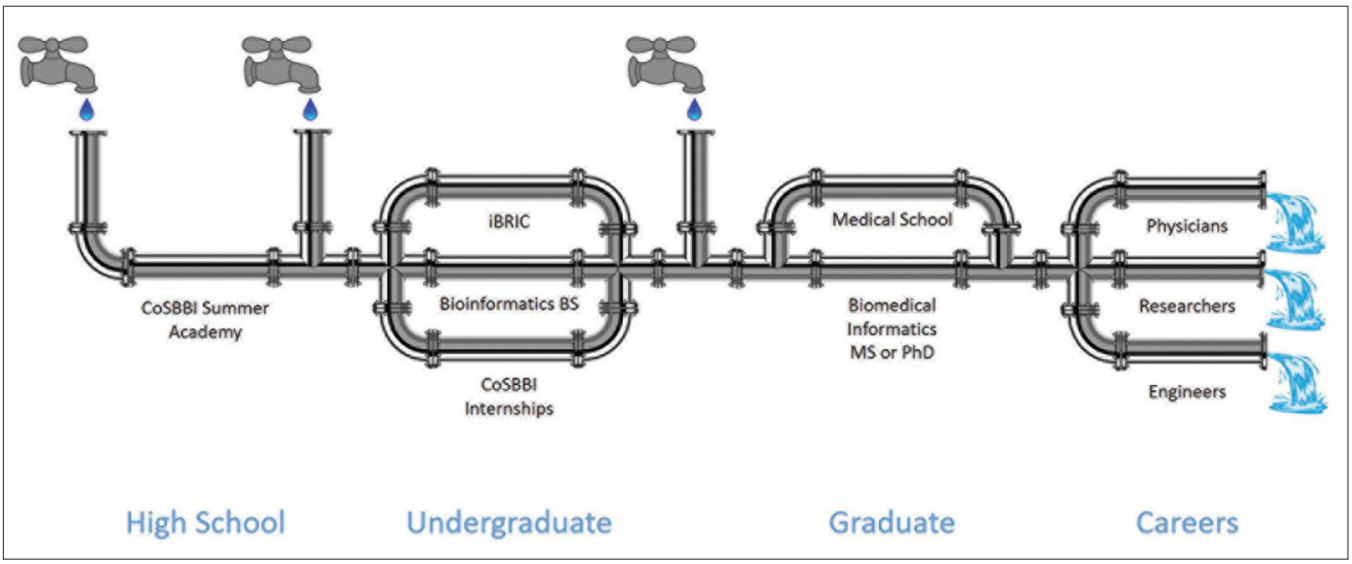

Figure 1: The Science, Technology, Engineering, and Mathematics educational pipeline to careers in biomedical and pathology informatics. The pipeline initiates at the high-school level with the Computer Science, Biology and Biomedical Informatics program, which provides a potential entry point for young scholars. Students can perpetuate through the pipeline as a Computer Science, Biology and Biomedical Informatics intern and/or their undergraduate major of study (i.e. bioinformatics), which provides another potential pipeline entry point for college students. Underrepresented undergraduate students in Science, Technology, Engineering, and Mathematics fields are offered another potential entry point into the Science, Technology, Engineering, and Mathematics pipeline through our Biomedical Research, Informatics, and Computer Science program

awareness, leadership, scientific communication, and more. Between 2011 and 2015, CoSBBI hosted 42 high-school scholars utilizing 31 unique faculty and staff mentors, in addition to numerous graduate student and postdoctoral fellow co-mentors.

Computer science, biology and biomedical informatics internship

After the Summer Academy, CoSBBI alumni have the option to return for CoSBBI internships. These paid summer positions are available while a scholar is still in high school and after a scholar begins their undergraduate education. Between 2014 and 2015, eight students returned as CoSBBI interns, two of whom conducted the evaluation of the 2014 CoSBBI program and were lead authors in the last year's program update.[13] Additionally, in 2015 in partnership with the Pittsburgh Health Data Alliance between the University of Pittsburgh Medical Center (UPMC), the University of Pittsburgh, and Carnegie Mellon University (CMU) and the Center for Commercial Applications of Healthcare Data, we launched an innovation internship. Innovation interns are marketed toward CoSBBI undergraduate alumni with the goal of exposing young students to innovation and entrepreneurship.

Internship in biomedical research, informatics, and computer science

For those interested in pursuing informatics as a career or academic endeavor, we recommend bioinformatics or a similar major because of the interdisciplinary nature of the field.[4] Students of bioinformatics, which is offered at the University of Pittsburgh and other schools, gain a core background in biology, chemistry, computer science, and statistics. However, many universities do not offer bioinformatics as a major. This is especially true of minority-serving institutions (MSIs). Underrepresented undergraduate students from two MSIs – Lincoln University and the University of Puerto Rico-Rio Piedras – may enter our informatics education pipeline through a new program named Internships in Biomedical Research, Informatics, and Computer Science (iBRIC). iBRIC is supported by the Pennsylvania's Department of Health CURE program's Big Data for Better Health initiative, NIH's Big Data to Knowledge program, and the NLM. iBRIC is a longer (10 weeks) and deeper immersion into informatics research that is similar to the CoSBBI Academy, but with additional professional and academic development activities tailored to the education and experience level of participating scholars. The program culminates with presentations at the Duquesne University Summer Undergraduate Research Symposium (http://www.duq.edu/academics/schools/natural-and-environmental-sciences/undergraduate-research/summer-research-symposium).

Common standards

For both CoSBBI and iBRIC, we hold participating scholars to the standards of graduate-level work. The American Medical Informatics Association lists eight fundamental scientific skills, in which graduate students of biomedical informatics should become competent; these skills are listed and defined in Table 1 and in quotes below.[14] The skill of "acquiring professional perspective" is addressed although the didactic secessions that provide an overview of the many disciplines covered under the umbrella of biomedical and pathology informatics. The same graduate students and faculty members who serve as mentors and coordinators lead these didactic sessions. A focus of CoSBBI is to help the young scholars transition from closely monitored high school work to a less structured university environment, and a focus of iBRIC is to further strengthen academic independence. As part of these objectives, mentors help their mentees learn to "analyze problems," "produce solutions," and "articulate the rationale" behind the solutions on their own. This includes "implementing, evaluating, and refining" the project methods as the summer progresses.

Table 1: Core competencies in biomedical informatics

Competency	Description
Acquire professional perspective	Summarize and explain the history and values of the discipline and its relationship to related fields while demonstrating an ability to read, interpret, and critique the core literature
Analyze problems	Analyze, understand, abstract, and model a specific biomedical problem in terms of data, information, and knowledge components
Produce solutions	Use the problem analysis to identify and understand the space of possible solutions and generate designs that capture essential aspects of solutions and their components
Articulate the rationale	Defend the specific solution and its advantage over competing options
Implement, evaluate, and refine	Demonstrate an ability to carry out the solution, to assess its validity, and iteratively improve its design
Innovate	Create new theories, typologies, frameworks, representations, methods, and processes to address biomedical and informatics problems
Work collaboratively	Demonstrate the ability to team effectively with partners from diverse disciplines
Disseminate and discuss	Communicate effectively to audiences in multiple disciplines in persuasive written and oral form

Through their individual projects, the CoSBBI and iBRIC students have opportunities to "work collaboratively" with graduate students in the same laboratory, which helps companion scholars navigate research roadblocks. Scholars get a sampling of "innovation" across different domains within informatics through department research colloquiums presented by both local and national guest speakers. Finally, the weekly journal clubs and end of academy presentations offer scholars a chance to "educate, disseminate, and discuss" their work. Furthermore, we encourage submission of projects to other venues and meetings that provide larger audiences and feedback.

We continue to foster relationships with CoSBBI students/interns and iBRIC scholars as they transition into

their undergraduate studies and beyond. Mentors and program coordinators regularly interact with CoSBBI/iBRIC alumni and offer recommendation letters and other means of support as they apply to college and progress through their career. While we promote and encourage biology, biomedical, and pathology informatics, we support all career and academic choices. The vast majority of our past scholars are on track to eventually become physicians, researchers, computer scientists, and engineers. We discuss these outcomes in greater detail below.

Demonstrations of Success of the Pipeline

During the first 5 years of the CoSBBI Summer Academy, we hosted 42 high-school students. Of these scholars, nearly one-third of them are from backgrounds that are underrepresented in the sciences, as defined by the NIH [Figure 2]. The NIH does not consider females to be underrepresented in the biomedical sciences, but <20% of computer science degrees in the US are awarded to women. The percentage of our alumni who are either female or from underrepresented groups is over 50% [Figure 2]. Of the 37 CoSBBI alumni who graduated high school, and for whom we have data, all were enrolled in higher education at the time of last contact. Most of our scholars are enrolled at regional schools ($n = 12$) such as the University of Pittsburgh, CMU, and Pennsylvania State University, or at the US News ranked top 25 universities (http://www.colleges.usnews.rankingsandreviews.com/best-colleges/rankings/national-universities) such as Princeton, Yale, Harvard, and Stanford ($n = 19$; CMU is both regional and top 25). Encouragingly, 90% ($n = 28$ of 31) of CoSBBI alumni with known majors matriculated into STEM majors, including 32.2% ($n = 10$) who declared a major as computer science or bioinformatics [Figure 3]. The vast majority of the remaining scholars are enrolled in pathology or informatics-related fields such as mathematics, biology, and engineering [Table 2], suggesting that the pipeline is initiating or perpetuating students' interests in informatics and related sciences. In addition, all nine underrepresented students in college, as well as eight out of nine female alumni with known majors, are in STEM fields. This demonstrates that early mentorship and research experiences can potentially increase the diversity of students in the emerging sciences as they have the traditional sciences.

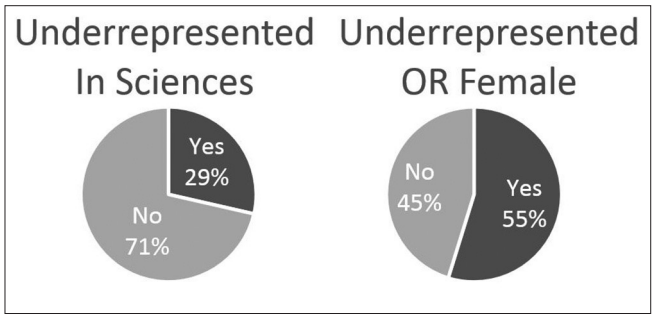

Figure 2: The diversity of Computer Science, Biology and Biomedical Informatics scholars. The total number of CoSBBI scholars from 2011 to 2015 is 42. (Left) Percentage of CoSBBI participants considered underrepresented in biomedical sciences based on the National Institutes of Health definition. (Right) Percentage of CoSBBI scholars who are either female or underrepresented. Examined due to the gender bias that exists in computer science

Evaluation of the 2015 Computer Science, Biology and Biomedical Informatics Academy

Publications, honors, and presentations

The 2015 CoSBBI Summer Academy included ten scholars. Several of these students leveraged the research and mentorship from their internship into publications, awards, and presentations. One scholar independently published her research on the discovery of an overexpressed gene in an aggressive breast cancer subtype in the International High School Journal of Science (http://www.ihsjs.com/magazines/IHSJS-Dec-2015/). Another scholar attended the next-generation high school summit at the White House to present his analysis of the effect of fusion genes on the development of breast cancer (http://www.triblive.com/news/allegheny/9413966-74/cancer-kurukulasuriya-allderdice?sf42404517 = 1#axzz3r5QbbSZB). At the American Medical Informatics Association 2015 High School Scholars Program (https://www.amia.org/amia2015/high-school-scholars), three of five oral presentations were from our pipeline program. Two of the students were 1st year CoSBBI scholars, and the third presenter

Table 2: Declared majors of computer science, biology, and biomedical informatics alumni

Declared majors of computer science, biology, and biomedical informatics alumni (count)
Engineering (10)
Math (6)
Computer Science (4)
Biology and Chemistry (4)
Healthcare (2)
Business (2)
Political science (1)
Undecided (2)
Unknown/unreported/in high school (11)

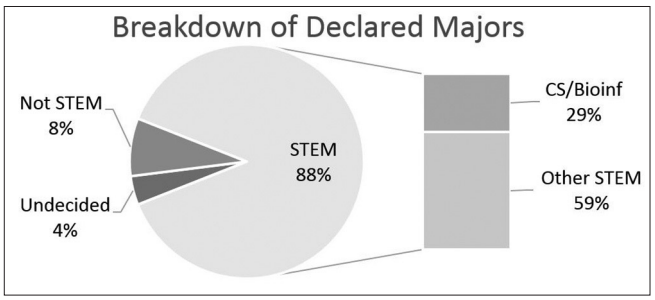

Figure 3: Computer Science, Biology and Biomedical Informatics scholars matriculate into Science, Technology, Engineering, and Mathematics fields. Breakdown of declared undergraduate majors for CoSBBI alumni ($n = 31$). Longitudinal data were collected through E-mail and social media to determine the matriculation status of the past scholars

was an alumnus of the program who later became a CoSBBI intern in a pathology informatics laboratory. This intern became interested in pathology after gaining initial exposure to the field during his 1st year in the CoSBBI Academy.

2015 Computer science, biology and biomedical informatics students self-reported gains from anonymous surveys

Annual evaluation by anonymous surveys demonstrated that scholar interest in research increased as a result of their summer research project. Nine and eight scholars voluntarily and anonymously completed both pre- and post-academy surveys, respectively. The survey questions were adapted from the SURE Survey[15] and a survey used previously by the School for Science and Math at Vanderbilt.[16] Between pre- and post-surveys, the number of scholars who planned to pursue a doctoral degree in a science-related field increased from one to five (data not shown). Six of the eight students indicated that they are very likely to continue doing research in high school or as an undergraduate, suggesting an initiation or deepening of their interest in research (data not shown). In addition, all but one scholar felt that their idea of a career scientist became clearer over the course of the summer. Not surprisingly, the scholars felt that one-on-one time with their mentors was of most value, a notion that is echoed by all of the mentors (data not shown). Other self-reported gains on the posttest included "tolerance for obstacles faced in the research process," "understanding that scientific assertions require supporting evidence," and "readiness for more demanding research." Finally, three scholars who reported that they were uncomfortable with computer programming at the start of the summer no longer felt that way at the completion of the academy (data not shown).

Another goal of our STEM pipeline is to increase the communication skills of our scholars so that they can effectively communicate scientific literature in addition to their own research. We address scientific communication by offering various opportunities to practice presentation skills during the program. These opportunities include weekly journal clubs that allow scholars to present a peer-reviewed article to the rest of the group, periodic research update sessions, in which scholars report their progress to their peers, and at the conclusion of the academy, each scholar creates and presents a poster in addition to an oral presentation. We assessed scholars' self-perceived gains in communication and other areas through a modified survey from the SSMV. This survey prompted the scholars to rate themselves on a Likert scale (from 1 - I have no skills to 10 - I am an expert) on identical questions before and after the academy. Further, on the posttest, the scholars were asked to rate their perception of their skills in each category before the academy – we refer to this as the post-pre-test [Figure 4]. This was a strategy used by the SSMV with the assumption that the students' ideas of skills will change after the experience and they may have rated themselves too high before the academy. Although there was a trend, we did not find a significant decrease in self-assessment between pre- and post-pre-analyses. The scholars' ability to "make use of primary research literature to understand current advances in a scientific field" had the greatest self-evaluated improvement (*t*-test $P < 0.05$), with a three-point increase on a ten-point scale [Figure 4]. Likewise, "communicating with

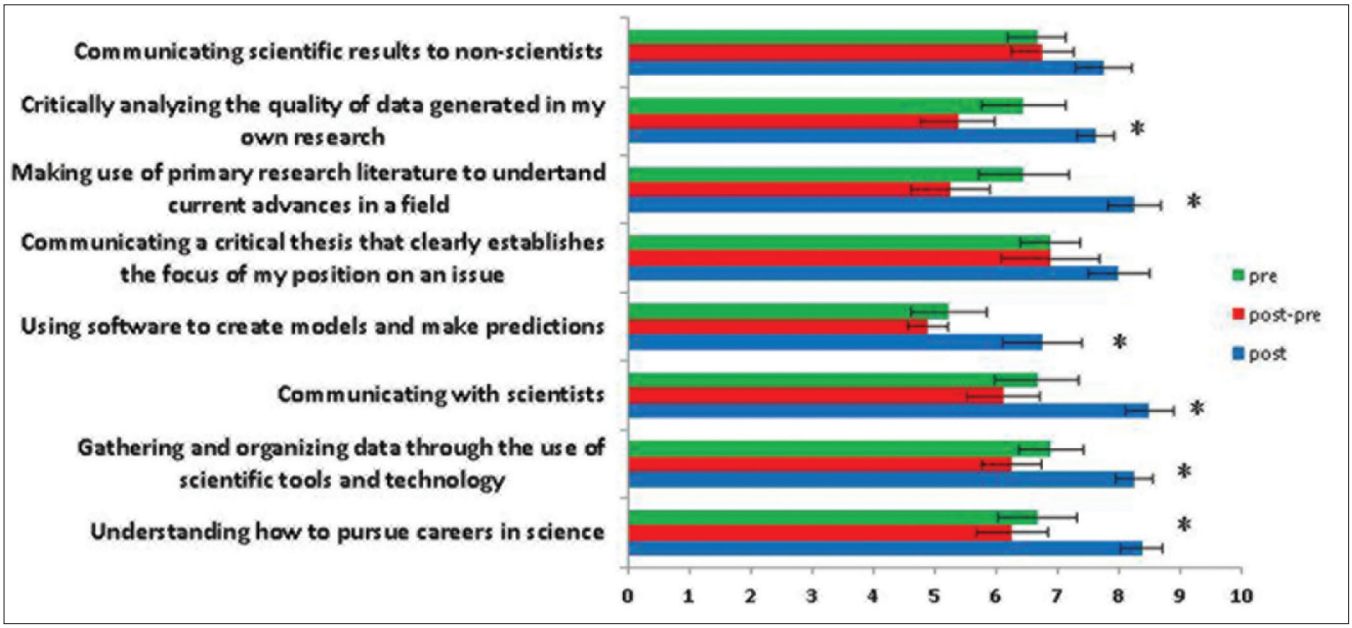

Figure 4: 2015 Computer Science, Biology and Biomedical Informatics scholar self-reported gains. Anonymous and voluntary surveys were administered for the students to rate themselves before (pre) and after (post) the academy on a Likert scale from 1 to 10 (1 - I know little to nothing; 5 - I am developing skills; 10 - I am an expert). On the survey administered after the Academy the students were also asked to reflect on their skills before the academy (post-pre). $N = 9$ for the presurvey and $n = 8$ for the postsurvey. *Indicates $P < 0.05$; unpaired *t*-test between pre and post

scientists" significantly increased from the pre- to post-test. These results suggest that engaging students in the practice/feedback cycle in a safe learning environment (i.e., a dedicated high-school summer program) promotes personal development of scientific communication skills. Other skills with significant self-reported increases include "critically analyzing the quality of data generated in my own research," "using software to create models and make predictions," "gathering and organizing data through the use of scientific tools and technology," and "understanding how to pursue careers in science" [Figure 4]. The students did not report a significant gain in communicating results to nonscientists. Likely, this is because the CoSBBI Academy is a full immersion into a scientific research environment that mostly trains students to communicate their research to other scientists. This is an area we hope to improve by including a workshop on how to communicate science to nonscientists.

Early Outcomes of Graduates from the University of Pittsburgh's Bioinformatics Undergraduate Program

The University of Pittsburgh created a Bioinformatics Bachelor's program in the fall of 2009 and saw its first graduates of this degree in 2013. Through its first 3 years of graduating classes, sixteen students acquired a bioinformatics degree. In addition, three students graduated with similar training before the major becoming official. Of these nineteen graduates, we have the postgradation paths of all but two of them. Nine graduates pursued informatics or computational graduate programs. Seven of these nine transitioned directly into PhD programs. Two graduates became software developers for research universities. The remaining graduates went into industry with about half joining start-up companies and the other half joining the research divisions of established corporations. Finally, two graduates started their careers as scientific information technology consultants. The trajectory of scholars who exit the undergraduate level exemplifies the opportunities available to the young scholars in our pipeline.

Conclusions

In this editorial, we highlight students who have progressed through various stages in a biomedical and pathology informatics pipeline. The pipeline introduces young scholars to informatics as early as their sophomore year of high school, continues scholar engagement throughout undergraduate years of study, and will aid in the transition to medical and graduate programs. CoSBBI alumni have coauthored peer-reviewed manuscripts, presented at scientific and professional meetings, won awards, and remain in informatics and STEM-related fields as undergraduates. The high number of CoSBBI alumni that matriculate into STEM majors, specifically computer science and informatics-related fields, demonstrates that early access to mentorship and authentic research has the potential to increase the number and diversity of students prepared for the emerging fields of informatics as it has for other STEM fields.

As the fields of informatics grow, pipeline and outreach programs of multiple varieties are starting to emerge to reach the next generation of informaticians. CoSBBI expanded from and still is a part of the existing UPCI Academy that offers traditional wet-laboratory biomedical research experiences throughout the UPCI. Programs like these exist at universities across the country. Growing from an existing program is a possible path of minimal resistance to increase the education of high-school students on informatics research and careers. Having a larger organization has helped the CoSBBI Academy with recruiting, administrative support, organization, circumventing barriers for working with minors at University facilities and has provided a wide exposure to academic and career options to our students. In fact, our results are not unique to our portion of the broader academy as ~90% of all UPCI Academy alumni are currently enrolled in STEM fields. However, the number of students in bioinformatics or computer science is significantly higher among the CoSBBI alumni as compared to all other UPCI Academy alumni, suggesting that the type of research students are exposed to impacts their future academic careers.

As discussed more in depth in our previous editorial, the CoSBBI model has been finessed over time to circumvent issues as they arise. Although we find it to be beneficial, a particular challenge of being part of a larger and broader program is defining expectations to applicants of what a research project in pathology, biology, or biomedical informatics entails. This has improved with time as knowledge of informatics slowly spreads. However, misconceptions still exist about the nature of computational research. We find that it is necessary to explicitly state on our application and website that most CoSBBI scholars will work exclusively in a dry laboratory setting and will not work in a traditional wet-laboratory. This has helped set expectations and enrich for students with a keen interest in informatics and computational research – a key we find for successful mentorship pairings.

As the first CoSBBI alumni begin to graduate from college over the next few years, we hope to gain insight into the long-term impact of the program on our scholars' career choices. In the meantime, we will continue to offer mentored research experiences to high-school and college students hoping that early exposure to pathology, biology, and biomedical informatics research will encourage diverse and high-performing students to remain in the expanding fields after graduation and that scholars of all backgrounds and means are able to pursue their ambitions in STEM.

Acknowledgments

We acknowledge Eric Polinko, PhD, for supplying the data on the bioinformatics graduates of the University of Pittsburgh, Lindsay Surmacz for CoSBBI program coordination with the main UPCI Academy, Dr. Michael T. Lotze, Dr. Steffi Oesterreich, and Dr. Nancy Davidson for their leadership and devotion to

the UPCI Academy, Victoria Khersonsky for her technical support for CoSBBI scholars, and DBMI graduate students and faculty/staff mentors for their unwavering support and enthusiasm, and all mentors who donated their time throughout the University of Pittsburgh. CoSBBI is supported through grants to the UPCI Academy from the Doris Duke Foundation-Clinical Research Experiences for High School Students at the University of Pittsburgh (Grant No. 2014154), the National Cancer Institute CURE Program (3P30CA047904-27S2), and support from the Jack Kent Cook Foundation and donations from UPMC and grateful parents and patients. In addition, we would like to thank the team from the DBMI NLM Training Program Grant in Biomedical Informatics (T15 LM007059), the UPCI Cancer Center Support Grant for the Cancer Bioinformatics Service (P30 CA47904), the Clinical and Translational Science Institute Biomedical Informatics Core (UL1 RR024153).

Financial support and sponsorship

The study was supported by the National Cancer Institute CURE Program (3P30CA047904-27S2), Jack Kent Cook Foundation, Doris Duke Charitable Foundation-Clinical Research Experiences for High School Students at the University of Pittsburgh (Grant No. 2014154).

Conflicts of interest

There are no conflicts of interest.

References

1. Corn M, Rudzinski KA, Cahn MA. Bridging the gap in medical informatics and health services research: Workshop results and next steps. J Am Med Inform Assoc 2002;9:140-3.
2. NLM's University-based Biomedical Informatics Research Training Programs; 2016. Available from: https://www.nlm.nih.gov/ep/GrantTrainInstitute.html. [Last accessed on 2016 May 20].
3. The Office of the National Coordinator for Health Information Technology (ONC). Report to Congress; October, 2014. Available from: https://www.healthit.gov/sites/default/files/rtc_adoption_and_exchange9302014.pdf. [Last accessed on 2016 Jul 01].
4. Magana AJ, Taleyarkhan M, Alvarado DR, Kane M, Springer J, Clase K. A survey of scholarly literature describing the field of bioinformatics education and bioinformatics educational research. CBE Life Sci Educ 2014;13:607-23.
5. Hersh W. Health and biomedical informatics: Opportunities and challenges for a twenty-first century profession and its education. IMIA Yearb Med Inform 2008;47:138-45.
6. Chatterjee S, LeRouge CM, Chiarini Tremblay M. Educating students in healthcare information technology: Is community barriers, challenges, and paths forward. Commun Assoc Inf Sys 2013;33:1-14.
7. Williams MS, Ritchie MD, Payne PR. Interdisciplinary training to build an informatics workforce for precision medicine. Appl Transl Genom 2015;6:28-30.
8. Williams JP. The Crucial Role of Mentors in STEM. U.S. News and World Report; 2014. Available from: http://www.usnews.com/news/stem-solutions/articles/2014/04/24/the-crucial-role-of-mentors-in-stem. [Last accessed on 2016 Jul 01].
9. Griffin KA, Perez D, Holmes AP, Mayo CE. Investing in the future: The importance of faculty mentoring in the development of students of color in STEM. New Dir Inst Res 2010;148:95-103.
10. Coles A. The role of mentoring in college access and success. Research to practice brief. Washington, DC: Institute for Higher Education Policy. Retrieved from ERIC database. (ED520415) 2011;1-10.
11. Adams HG. Mentoring: An Essential Factor in the Doctoral Process for Minority Students. Notre Dame, IN: National Consortium for Graduate Degrees for Minorities in Engineering and Science, Inc. Retrieved from ERIC database. (ED358769) 1992;1-8.
12. Dutta-Moscato J, Gopalakrishnan V, Lotze MT, Becich MJ. Creating a pipeline of talent for informatics: STEM initiative for high school students in computer science, biology, and biomedical informatics. J Pathol Inform 2014;5:12.
13. Uppal R, Mandava G, Romagnoli KM, King AJ, Draper AJ, Handen AL, *et al.* How can we improve science, technology, engineering, and math education to encourage careers in biomedical and pathology informatics? J Pathol Inform 2016;7:2.
14. Kulikowski Ca, Shortliffe Eh, Currie Lm, Elkin Pl, Hunter Le, Johnson Tr, *et al.* Amia Board White Paper: Definition Of Biomedical Informatics And Specification Of Core Competencies For Graduate Education In The Discipline. J Am Med Inform Assoc 2012;19:931-8.
15. Lopatto D. Survey of undergraduate research experiences (SURE): First findings. Cell Biol Educ 2004;3:270-7.
16. Eeds A, Vanags C, Creamer J, Loveless M, Dixon A, Sperling H, *et al.* The school for science and math at Vanderbilt: An innovative research-based program for high school students. CBE Life Sci Educ 2014;13:297-310.

Abstracts

The Logic of Cancer

Ishan Levy[1], Roger Day[2]

[1]Department of Biomedical Informatics, University of Pittsburgh Cancer Institute, Computer Science, Biology and Biomedical Informatics, Summer Academy, [2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA.
E-mail: day@upci.pitt.edu

Introduction

There is a rapidly growing area of research looking into "driver genes." One definition is genes which in some tumors contain mutations or changes in expression that are selected for. The concept of driver genes is guiding computational efforts to understand cancer diversity, but several examples show that it is not well defined and provides an incomplete view of cancer. Critical consideration and modification of the concept may promote better identification of effective prevention and treatment strategies.

Methods

We catalog examples suggesting clarifications and extensions. We replace the concept of a driver gene with a system of Boolean equations of states, which can include genomic mutations and other aberrations. We then develop a method of finding all possible therapeutic strategies using the equations. Relaxing the assumption of acyclicity allows distinction between cancer initiation events and continued requirements.

Results

An open source web application was developed to input knowledge about many aspects of cancer, including the "hallmark" cancer capabilities, molecular networks, immune evasion, and tumor initiation/promotion. The output is a set of therapeutic strategies. We find that initiating genomic alterations may appear as driving genes while being useless as targets for therapies.

Conclusions

The model provides a systematic way of understanding cancer and its components and clarifies the driver gene concept.

Future Direction

The model can be extended by assigning to the states quantitative values in the unit interval and modifying the logical operators accordingly. Input of some biological knowledge could be automated.

Using Differential Gene Expression to Detect Genomic Alterations and their Functional Differences

Albert Kim[1], Joyeeta Dutta-Moscato[2], Xinghua Lu[2]

[1]Department of Biomedical Informatics, University of Pittsburgh Cancer Institute, Computer Science, Biology and Biomedical Informatics, Summer Academy, [2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA.
E-mail: jod30@pitt.edu

Introduction

The phosphoinositide 3-kinase (PI3K) signaling pathway induces cell growth/proliferation. A significant mutation in pathway protein can change expression of downstream genes. Certain amino acids in PI3K, called hotspots, are mutated more frequently than others in cancerous cells. Mutations in distinct hotspots could affect different aspects of protein function, resulting in distinguishable changes in differentially expressed genes (DEGs). We examined the hotspots in tumor samples where we believe that differential gene expression was driven by PIK3CA mutations.

Methods

The tumor-specific driver identification algorithm uses causal inference algorithms with data from The Cancer Genome Atlas to generate a dataset of tumors, its driver genes, and the target DEGs of each driver gene. Lists of differentiated genes with significant roles for three unique hotspots were analyzed. Gene ontology allowed discovery of biological functions that were affected by each set of DEGs.

Results

Three hotspots examined at amino acid positions 1046, 544, and 344 (±3). Of 185 mutations, 171 (92.4%) mutations at AA 1046, 227/246 (92.2%) at AA 544, and 35/36 (97.2%) at AA 344 called as driver mutations and had 97, 16, and 38 unique target DEGs, respectively, sharing 310. Thirteen functions shared by the three hotspots, three unique to AA position 1046, and one shared by position 1046 and 544, but not 344.

Conclusions

Algorithm is a potentially useful tool. Consistent DEGs can imply potential biomarkers for certain events. Analysis of the functions of the unique and common DEGs of each hotspot provides insight on the different roles of separate locations of protein mutations.

Classifying Electroencephalography Swallowing Signals by Liquid Viscosity

Olaoluwa Owoputi[1], Ervin Sejdic[2]

[1]Department of Biomedical Informatics, University of Pittsburgh Cancer Institute, Computer Science, Biology and Biomedical Informatics, Summer Academy, [2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA.
E-mail: esejdic@pitt.edu

Introduction

Dysphagia is a swallowing disorder caused by damage to the brain. It is also known that there is more evident when swallowing

liquids of relatively low viscosity (e.g., water). However, the exact neurological mechanism that is responsible for the disorder is neither well known nor known how liquid viscosity affects swallowing difficulty. However, electroencephalography (EEG) technology, which is used to measure brain activity in the cortex, can shed light on the role of the brain in dysphagia.

Methods

EEG data were obtained from 53 healthy adults as they swallowed five boluses each of water, nectar-thick, and honey-thick in both neutral and chin tucked head positions. The data were then processed to calculate the wave entropy, peak frequency, bandwidth, and centroid of each electrode in each participant. The processed data were used with a random subspace K-nearest neighbor algorithm to classify each swallow based on the type of liquid that was swallowed.

Results

Using the wave entropy and bandwidth of each electrode, the classifier was able to distinguish between water and honey-thick swallows in both head positions. The success rates for the neutral and chin tucked head positions were 75.7% and 72.7%, respectively.

Conclusions

The fact that the classifier was able to compare water and honey-thick, two liquids with different viscosities, reflects the difference in neurological activity when swallowing liquids of different viscosities.

Future Directions

An additional experiment could be conducted, in which the EEG swallowing signals of healthy people were compared with those of dysphagic people. The results could then be used to more precisely determine the effect of dysphagia on neurological activity while swallowing.

Comparison of Query Performance of a Research Data Warehouse Stored in a Relational Star Schema Database versus in a NoSQL Document-store Database

Mit Patel[1], Bill Shirey[2], Shyam Visweswaran[2]

[1]Department of Biomedical Informatics, University of Pittsburgh Cancer Institute, Computer Science, Biology and Biomedical Informatics, Summer Academy, [2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA.
E-mail: shirey@pitt.edu

Introduction

Two methods are compared for storing and querying data in a single-node research data warehouse: a standard data warehousing relational star schema Oracle database and a NoSQL document store MongoDB database. When Oracle stores extensive multidimensional databases, queries become drastically sluggish. MongoDB's document data structure, on the other hand, could potentially allow for reduced query execution times, making it an efficient alternate to star schema databases when storing and querying Big Data. If so, MongoDB would make a practical replacement for Oracle as an i2b2 data warehouse.

Methods

A test dataset of medical data was first imported into both Oracle and MongoDB. Queries of increasing complexity were then constructed and incorporated into a software framework that executed the queries against both databases on the same computer. To ensure equal resource allocation, Oracle queries were executed while MongoDB was turned off and vice versa, ensuring that an equal amount of processing resources would be dedicated to each database.

Results

MongoDB only executed two out of the seven total queries faster than Oracle; of the three patient count queries, Oracle outperformed MongoDB by an average of 97% on two while MongoDB performed 98% faster than Oracle on the third. Of the three queries that returned lists of patients, Oracle again outperformed MongoDB by an average of 95% on two while MongoDB performed 63% faster than Oracle on the third. Furthermore, since MongoDB does not utilize indexes for exclusion operators in its queries, Oracle outperformed MongoDB by an average of 99% when executing a real-life query involving negation logic.

Conclusions

MongoDB's slight improvement in query execution performance over Oracle along with its lack of indexing support for queries involving negation logic, which are ubiquitously executed on a wide range of datasets, ultimately makes a single node of MongoDB an impractical replacement for a single node of star schema Oracle as a research data warehouse.

Future Directions

Additional NoSQL storage types, such as key-value stores, can be compared to see if they make better candidates than MongoDB for replacing star schema-based data warehouses. Multiple storage nodes along with distributed queries utilizing map-reduce can also be included to better model typical usage of NoSQL technology.

Using Natural Language Processing to Improve the Prediction of Relevant Data in Electronic Medical Records

Arushi Bandi[1], Harry Hochheiser[2], Andrew J. King[2]

[1]Department of Biomedical Informatics, University of Pittsburgh Cancer Institute, Computer Science, Biology and Biomedical Informatics, Summer Academy,

[2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA.
E-mail: andrew.king@pitt.edu

Introduction

Physicians using electronic medical records (EMRs) often struggle with information overload due to a large amount of irrelevant data. Researchers suggested the idea of learning EMRs that use machine-learning algorithms to predict information a physician would deem relevant to specific patient cases. This project aimed to explore whether features extracted from the natural language processing (NLP) of patient progress notes can improve these predictions.

Methods

We used a collection of laboratory tests as the target, for which a physician determined relevance, and features extracted from patients' records were used to predict patient-specific relevant laboratories. The NLP software Apache cTAKES™ was used to extract clinical mentions of features. Then, a Boolean occurrence table was created with columns indicating clinical terms and rows indicating the terms' mention per patient. The scikit-learn implementation of logistic regression classification was used to predict which laboratory tests a physician would deem relevant.

Results

The relevance of 17 laboratory tests was predicted for 38 patient cases in three sets per patient: unison of all records (580 files); most recent files (38 files); most recently more than 1 day (36 files). Approximately seventy clinical terms were extracted from each document. There was an average of 1.7% increase in precision and 0.7% increase in recall when using NLP-extracted features and clinical values compared against clinical values alone.

Conclusions

Features extracted from free text show a small improvement over existing models. Future research will determine if the improvement is significant and if the models can be improved further.

Identifying Transcription Factor Binding Motifs: A Convolutional Neural Network Approach

Lukas Schmit[1], Joyeeta Dutta-Moscato[2], Xinghua Lu[2]

[1]Department of Biomedical Informatics, University of Pittsburgh Cancer Institute, Computer Science, Biology and Biomedical Informatics, Summer Academy, [2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA.
E-mail: jod30@pitt.edu

Introduction

Transcription factors are regulatory proteins that bind to specific DNA sequences (motifs) in order to regulate transcription and ultimately gene expression. Identifying transcription factor binding motifs (TFBS) could lead to novel drugs that target transcription factors. Here, we present a deep learning approach to classifying motifs from chromatin immunoprecipitation followed by sequencing (ChIP-Seq), which achieves state of the art accuracy for several transcription factors.

Methods

We trained a convolutional neural network with stochastic gradient descent. We obtained data from the NHGRI ENCODE project. The model was architecturally optimized on ChIP-Seq peaks enriched in MafK-binding sites and random sequences from the human genome for control. We assessed performance based on classification accuracy and then compared the MAFK-optimized model's performance on 65 other proteins with known functionality in cancer.

Results

The best model optimized to recognize MAFK-binding events achieved 96.6% classification accuracy (sensitivity 95.8%, specificity 97.4%). When the same architecture was retrained on various proteins, it converged to accuracies ranging from 56.4% to 97.4%. Our results compared favorably to a motif recognition method using standard position weight matrices. We also observed better results when we used a majority of the peak sequence in our input to the model, although due to high variability in peak length, this involved a tradeoff in size of training data.

Conclusions

Convolutional neural networks can provide a superior classification for identification of TFBS from ChIP-Seq data.

Future Directions

We plan on using the network to isolate binding motifs from ChIP-Seq data. It may also be useful to zero-pad shorter sequences to increase model input size while maximizing potential presence of motifs in the input.

Investigating Fusion Genes as Mediators of Breast Cancer Metastasis

Jahnik Kurukulasuriya[1], Nolan Priedigkeit[2], David N. Boone[2,3], Adrian V. Lee[2]

[1]Department of Biomedical Informatics, University of Pittsburgh Cancer Institute, Computer Science, Biology and Biomedical Informatics, Summer Academy, [2]Magee-Women's Cancer Research Center, University of Pittsburgh Cancer Institute, [3]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA.
E-mail: priedigkeit.nolan@medstudent.pitt.edu

Introduction

Fusion genes are drivers of several types of cancers and are effective, cancer-specific targets for treatment. They have been extensively studied as cancer-initiating events, yet the role of fusion genes in cancer progression is largely unknown. Given

the inherently unstable breast cancer genome, we hypothesize that fusion genes may mediate breast cancer metastasis.

Methods

RNA-Seq and clinical data from seven patient-matched primary and metastatic breast cancer pairs were acquired from The Cancer Genome Atlas. Sequencing reads were aligned with STAR, differential expression was performed using DESeq2, and gene fusions were called with FusionCatcher. The same pipeline was performed in 52 breast cancer cell lines.

Results

Metastatic-specific fusions were found in patient samples, a few of which were also present in cell lines (CACNG4 – CACNG1, M1EN – GRB7, GOLT1A – KISS1…). On average, metastatic tumors harbored nine acquired fusions compared to their patient-matched primaries (range 2–14). GOLT1A–KISS1 is present in a metastatic patient sample and three of the cell lines (MDAMB175, 21PT, AU565). KISS1 is a known metastasis suppressor gene, making the fusion a prime candidate for further investigation.

Conclusions

We found that fusions genes may be acquired in metastatic breast cancer, representing a potentially novel mediator of metastasis. Future studies will focus on increasing the sample size of patient-matched pairs, filtering fusion candidates based on recurrence rates, and assessing for metastatic phenotypes (i.e. invasion, migration) within *in vitro* models.

SNHG15 is an Insulin-like Growth Factor 1-regulated Long Noncoding RNA that is Overexpressed in Aggressive Breast Cancer Subtypes and is Necessary for Cell Proliferation

Sreeroopa Som[1], Adrian V. Lee[2,3], David N. Boone[2,3]

[1]Department of Biomedical Informatics, University of Pittsburgh Cancer Institute, Computer Science, Biology and Biomedical Informatics, Summer Academy, [2]Department of Biomedical Informatics, University of Pittsburgh, [3]Magee-Women's Cancer Research Center, University of Pittsburgh Cancer Institute, Pittsburgh, PA, USA. E-mail: priedigkeit.nolan@medstudent.pitt.edu

Introduction

Approximately 80% of the genome is transcribed, yet only 2% is translated into proteins. About thirty-thousand long noncoding RNAs (lncRNAs) with largely unknown functions were newly identified through large next-generation sequencing projects. Insulin-like growth factor 1 (IGF1) signaling is involved in initiation and progression of a subset of breast cancer. It was recently discovered that IGF1 regulates a subset of lncRNAs including SNHG7, which regulates proliferation through a negative feedback loop controlling IGF signaling. It is unclear whether other IGF1-regulated lncRNAs are important for cell proliferation. Hypothesis: Mining The Cancer Genome Atlas (TCGA) data can aid in identifying other IGF1-regulated lncRNAs that play an important role in breast cancer cell proliferation.

Methods

Breast cancer expression data of lncRNAs were explored using the cBio and TCGA data portals. Our top target, SNHG15, was knocked down in MCF-7 (ER+) and MDA-MB-231 (TNBC) cells using small interfering RNA (siRNA), and reduced expression was confirmed through quantitative polymerase chain reaction. Proliferation assays were conducted following siSNHG15 treatment.

Results

Analysis of annotated lncRNAs in TCGA data through the cBio Portal revealed twelve IGF-regulated lncRNAs dysregulated in breast cancer. One such lncRNA, SHNG15, is a family member of SNHG7 and is overexpressed in 7% of patients. Patients with overexpressed SNHG15 levels are highly enriched for aggressive basal-like subtype. Reduction of SNHG15 through RNAi in both MCF-7 and MDA-MB-231 cells significantly decreased proliferation. siRNA treatment in MDA-MB-231 cells yielded 90% knockdown of SNHG15 but did not alter expression of small nucleolar RNA in its intron, suggesting that SNHG15 itself is necessary for full proliferation.

Conclusions

Utilizing publicly available data highlights functions of highly expressed lncRNAs in specific breast cancer subtypes. SNHG15-mediated cell proliferation will be investigated by exploring SNHG15 regulation of IGF1-regulated genes.

Enabling Large-scale Annotation of Drug–Drug Interactions in Product Labels to Create a Drug Interaction Knowledge Database

Joshua Le[1], Richard D. Boyce[2]

[1]Department of Biomedical Informatics, University of Pittsburgh Cancer Institute, Computer Science, Biology and Biomedical Informatics, Summer Academy, [2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA.
E-mail: rdb20@pitt.edu

Introduction

When patients take multiple drugs simultaneously, interactions known as drug–drug interactions (DDIs) may occur and annually account for between 0.02% and 0.17%

of 130 million emergency room visits. Although potential interactions are documented in product labeling, they are not recorded in a structured format that would allow for identification by a computer. Such formalization would help regulatory agencies, and drug safety scientists retrieve and maintain potential DDI evidence. During a pilot study, it was found that crowdsourcing is a reasonable route to pursue due to lower costs and greater efficiency. Previous studies showed that it was feasible to automatically and accurately preannotate possible DDIs in product labels. Humans will later refine these annotations through a graphical user interface. This project focused on developing a "bridge" in the project; in other words, a script that generates computable input for a named entity recognizer (NER) algorithm that would find only possible DDIs inside of this collection of structured product labels. The goal was to preannotate as many drug mentions as possible to ease human work.

Methods

HTML tables that contained possible DDIs were generated from a pool of about 50,000 product labels, which included all tables in the DDI sections for all product labels as of November 2013. Further processing yielded 1057 tables that may include drug mentions and DDIs. These data were the input for a script that found the basic statistics of the tables. Then, its output was placed into the NER that only found drug mentions in each table.

Results

Basic statistics of the sample of 1057 tables include 2182 total headers and 350 unique headers. Of these tables, the NER found that there were 79,941 drug mentions which may be DDIs as there may be more than one drug mention per data cell. However, a manual sampling of tables revealed that there appeared to be a general trend of a smaller number of possible DDIs than the count of drug mentions per table. In terms of the number of drug mentions, the tables had a minimum of 0, a maximum of 801, a median of 34, and a mean of 74.71.

Conclusions

This script enables the computation of drug mentions in each product label through the NER, which produces output that allows for the further processing and preannotation

before the handoff toward crowdsourcing to find the final DDIs in each product label.

Future Directions

Further development includes (1) examining where this heuristic approach fails and attempting to address it and (2) developing heuristics for extracting DDI mentions from other kinds of columns.

Aiding Chemotherapy Decisions for Breast Cancer Patients by Combining Oncotype DX Recurrence Scores with Visualization of "Number Needed to Treat" and Adverse Events

Xi (Lily) Xu[1], Roger S. Day[2]

[1]Department of Biomedical Informatics, University of Pittsburgh Cancer Institute, Computer Science, Biology and Biomedical Informatics, Summer Academy, [2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA.
E-mail: day@upci.pitt.edu

Introduction

Biomarker development is highly active in medical research; however, in the scientific literature, the impact of biomarkers on individual patients receives surprisingly little attention. A good example is the oncotype DX recurrence score (RS) for patients with estrogen receptor + early breast cancer that predicts the chance of recurrence and the benefit from chemotherapy. For clinicians, visualizing consequences for groups of patients has been shown to be more effective than communicating with probabilities. Therefore, we developed a visualization based on number needed to treat, to assist chemotherapy decisions based on a patient's RS.

Methods

An interactive web application, ShinyAE, was built with the shiny package in the R programming language. The probability of recurrence with or without chemotherapy as a function of RS is from Paik 2006. The distribution of adverse event grades is from Fisher 1997.

Results

ShinyAE contains two plots. One is a graph of number needed to treat relative to RS. The area under curve is broken into regions of different colors, each representing a grade of chemotherapy adverse event. The other produces a stack of boxes visualizing the number of patients with different outcomes, standardized to one patient who benefits. The user can specify RS, and the two plots respond dynamically.

Conclusions

Compared to traditional receiver operating characteristic plots, ShinyAE plots are directly useful for informing patient treatment decisions.

Future Directions

A more developed model predicting chemotherapy adverse events based on individual measurements contributing to RS could be beneficial. This application can be adapted for use with other biomarkers and diseases.