Bioinformatics Needs Assessment

Prepared by Courtney Crummett with Erja Kajosalo, Michael Noga, and Howard Silver of the Interdisciplinary Biosciences Group, MIT Libraries.
April 2014

Executive Summary

An assessment of the Bioinformatics Program at MIT Libraries was conducted using quantitative and qualitative data collection methods during FY13-14. Interviews were conducted to gain insight about bioinformatics researcher's needs and behaviors and insight about the bioinformatics support offered by the MIT Libraries. Data was collected from various services of the bioinformatics program as well as from other library services. The assessment found that the bioinformatics community is interdisciplinary and crosses traditional life science departmental boundaries. The bioinformatics community takes a collaborative do-it-yourself (DIY) approach to computational skills and analytical tools –if they don't know something or have something to use, they find someone who does or they build it themselves. Themes from the assessment emerged such as computational skills, tools, data, instruction and interdisciplinarity. The bioinformatics community has a desire for computational skills and modular training. The MIT Libraries bioinformatics training sessions are well attended; training sessions taught by experts are popular. Recommendations for the Bioinformatics Program at MIT Libraries include being more aware of open source software tools used by the community, attempting to expand the use of commercial tools in courses, and expanding outreach and advocacy regarding bioinformatics to the entire MIT community.

The initial learning objectives of the bioinformatics assessment are included below. Responses learned from the assessment have been added as sub-bullet points.

- What elements of the Bioinformatics Program at MIT Libraries should be continued, expanded or discontinued?
    - Instruction sessions and training should be continued.
    - Instructions sessions focusing on commercial tools embedded in courses should be expanded.
    - Outreach regarding bioinformatics support should be expanded to all MIT departments.
    - In-house developed instruction in advanced bioinformatics should not be expanded at this time.
- What services should be added to the Bioinformatics Program at MIT Libraries to meet the needs of the bioinformatics community members?
    - Data management assistance
- Do bioinformatics community members think of the library as a place for bioinformatics support?
    - Instructions statistics and usage of bioinformatics commercial tools suggest that the MIT Libraries are providing vital services to the bioinformatics community. Additionally, some interviewee responses agreed that these services are valuable.

Bioinformatics Program at the MIT Libraries

MIT Libraries has been committed to understanding the research needs of the unique bioinformatics community at MIT since the early 2000s. Based on recommendations from the B-Team Report to the Engineering and Science Libraries shared in April 2006, the Interdisciplinary Bioscience Group (IBG) committee was established to support and investigate the need for academic library support in interdisciplinary life sciences, specifically bioinformatics, and a formal Bioinformatics Program of library services and resources was also established. In 2007, MIT Libraries hosted Courtney Crummett, a National Library of Medicine Associate Fellow. In September of 2008, Crummett began an appointment as the Bioinformatics Librarian, a position that is included in the MIT Libraries Specialized Content and Services Department (SCS). Currently, the Bioinformatics Program at the MIT Libraries provides instruction and training, access to bioinformatics resources and tools and a forum for collaborative science.

In order to maintain MIT Libraries' commitment to understanding the research needs of the bioinformatics community and to update information and develop new recommendations from the B-Team Report, an assessment of the Bioinformatics Program at MIT Libraries was planned for FY13-14. The purpose of the assessment was to determine:

- What elements of the Bioinformatics Program at MIT Libraries should be continued, expanded, or discontinued?
- What services should be added to the Bioinformatics Program at MIT Libraries to meet the needs of the bioinformatics community?
- Does the bioinformatics community think of the library as a place for bioinformatics support?

The assessment included a mix method approach of qualitative interviews coupled with various data pieces mined from March through October of 2013.

Qualitative Methods: Interviews

Interviews were conducted to gain a greater understanding of the needs, perceptions, behaviors and characteristics of the MIT community members that the Bioinformatics Program at MIT Libraries serves. From March through June of 2013, IBG members conducted 7 interviews of MIT faculty, staff, and students from various departments (Table 1). Interview participants were identified by using the snowball method: IBG members requested interviews from a list of potential interviewees; for those interviews that were completed, the interviewee was asked for suggestions of other people to interview. Interview questions were developed by IBG members in consultation with members of the Assessment Team (Lisa Horowitz

and Jennie Murack). A list of questions can be found in Appendix 1. Questions focused on computational skills, software tools, and data management. Not all questions were asked at every interview. The majority of interviews were conducted in teams of one interviewer and one note taker.

Table 1 Interviewee demographics

Interviewees	Department	Status
1	Biological Engineering	Graduate Student
2	Chemical Engineering/Koch Institute	Faculty
3	Biology	Research Scientist
4	Materials Science Engineering/ Koch Institute	Postdoctoral Fellow
5	Biology	Graduate Student
6	Brain and Cognitive Sciences	Faculty
7	Biology/ Koch Institute	Graduate Student

After the interviews were completed, IBG members reviewed the transcripts and developed a list of common themes for coding analysis through group discussion. IBG members coded the interviews in groups of 2 and compared results. Disagreements in coding analysis were discussed and agreed upon. The final coding can be seen in Appendix 2.

Quantitative Methods: Data Mining

From March to November of 2013, various data was gathered to complement the interview findings and offer a variety of perspectives of bioinformatics activity at MIT. Data was chosen based on whether it directly represented services offered by the Bioinformatics Program, its availability and ease of access, and the possibility that it may show an interesting perspective of how bioinformatics is reflected in the greater MIT context.

Use Statistics for Commercial Bioinformatics Software Tools

Usage statistics of three commercial software analysis tools, BIOBASE, Ingenuity Pathways Analysis (IPA) and GeneGo MetaCore were reviewed. These tools are licensed by the Bioinformatics Program through a collaborative funding model in which stakeholders around MIT (departments, labs, centers, or institutes) share the cost of the license with MIT Libraries and MIT Libraries handles the licensing, access, and other logistical details.

Use of BIOBASE has been declining steadily while use for IPA and GeneGo has been increasing. Complete use statistics for each tool are reported in Appendix 3. BIOBASE has used IP-filtered access since 2009. The number of unique visitors and number of visits have declined over time from 466 unique visitors and 1195 visits in 2009 to 320 unique visitors and 657 visits in 2012. IPA and GeneGo have user account access models. The use of these tools has increased by both number of

active users and length of use. In the 4 years of the IPA license, use has increased from 91 active users to 125 active users and from 801 hours to 1461 session logins from 22 Departments, Labs and Centers. In the 4 years of the GeneGo license, use has increased from 37 active accounts to 60 active accounts, although the largest increase was seen in the first year of the license. These accounts come from 20 Departments, Labs, and Centers.

MIT Courses Related to Bioinformatics

The MIT 2013/14 course catalog was examined to estimate the number of courses related to bioinformatics offered and the number of departments that offer a course related to bioinformatics. Keywords search included:
sequencing
systems biology
bioinformatics
computation/al biology
genomic*
proteom*

47 courses related to bioinformatics come from 13 different departments in the 2013-14 course offerings (see Table 2). Of the 47 courses 38 are unique course as many are cross-listed among two and sometimes three departments. For example, HST.506J Computational Systems Biology (Same subject as 6.874J) is counted for both HST and Biology. A complete list of courses found can be seen in Appendix 4.

Table 2 Departments offering courses related to bioinformatics

Department	Number of Courses
Anthropology	1
Biological Engineering	4
Biology	11
Brain and Cognitive Sciences	3
Business	1
Chemical Engineering	3
Civil and Environmental Engineering	1
Computational and Systems Biology	4
Electrical Engineering and Computer Science	7
Health Sciences and Technology	9
Mathematics	1
Physics	1
Science, Technology and Society	1
Total Unique Courses	38

Bioinformatics Instruction Statistics

Instruction statistics for bioinformatics classes taught or hosted by the Bioinformatics Program since April 2005 through February 2014 were collected to examine trends in types of classes taught and attendance. Bioinformatics classes include sessions taught during IAP, special workshops hosted at MIT Libraries and embedded instruction. Over this period of time, 14 different types of classes have been offered 63 times to over 1161 attendants (See appendix 5). Retrospective statistics were collected using archived instruction statistics.

MIT Theses related to Bioinformatics

MIT theses were investigated to determine what departments issue theses related to bioinformatics. Keyword searching was completed in the Barton Theses Collection and bibliographic information for each result was taken from the Data Warehouse. Theses published from 2008-2012 were examined. Keywords used include:

next generation sequencing
systems biology
bioinformatics
computation* biology
genomic*
proteom*

214 theses related to bioinformatics were published from 2008-2012 and came from 21 departments (See Appendix 6).

Analysis

IBG members examined the 7 interviews alongside the data mined from 4 sources and identified common themes and then grouped findings and discussions points according to those themes. Those themes are: computational skills, tools, data, interdisciplinarity, and instruction. Below are findings organized according to these themes. Coded themes from the interview analysis reported in Appendix 2 are represented parenthetically.

Computational Skills

- The bioinformatics community takes a collaborative DIY approach to computational skills. Most interviewees are self taught (6/7 self taught computational skills) - they teach themselves and/or use their own abilities and/or find someone who can help them (5/7 lab members with diverse skills and use the KI or BioMicroCenter cores and 6/7 access to computational skills).

- What are the risks and needs associated with DIY approaches to both computational skills and tools. What are they missing? How can the libraries support or facilitate this DIY approach?
- These researchers leverage their micro communities (labs, research groups, institutes) for computer skills they lack and training needs. Sometimes, they take advantage of services provided to them by a core facility (5/7 lab members with diverse skills and use the KI or BioMicroCenter cores and 6/7 access to computational skills).
- The micro communities are purposely designed to bring together members with various levels of research and computational skills.
- Not everyone enters MIT with the same computational background or skills. Some are not completely prepared when they get here (2/7 Biology students do not gain bioinformatics skills).
- They know what skills they need (2/7 desire for statistical/mathematical skills and 3/7 desire for modular training) and if they don't have them they get the help (5/7 lab members with diverse skills and use the KI or BioMicroCenter cores and 6/7 access to computational skills).

Tools

- Much like computational skills, the community takes a DIY approach to analysis tools. If they can't find something that meets their needs, they build it or find someone that has already built something (4/7 in house development for lab, 5/7 in house develop tool users, and 3/7 in house development shared). The tools section of the interviews was the most enlightening part of the assessment. Over 50 tools were mentioned in the interviews, the majority of which were in-house developed (4/7 in house development for lab) or open source (7/7 open source tool users) and for research specific computational or programming tasks. The large majority of these tools are one-of tools developed to complete a single and specific analytical task. For a complete list, see Appendix 7.
- Programming knowledge is standard. All interviewees mentioned one of the 3 main programming languages—R, Python or MATLAB (3/7 MATLAB, 4/7 R and 4/7 Python).
- Commercial tools don't meet all their needs and they need to supplement with in-house developed or open source tools (4/7 in house development for lab, 5/7 in house develop tool users, 7/7 open source tool users, 2/7 not using commercial tools, and 3/7 dissatisfaction with commercial tools).
- Some still are using commercial tools (4/7 using commercial tools). GeneGo and IPA use has increased over time; BIOBASE use has decreased over time. For both GeneGo and IPA, there are power users and there are one-time users and the same power users appear year after year. '
- One response to the interview question "how do the MIT Libraries support your bioinformatics research and how you learn bioinformatics skills": "Support by the Libraries is useful, especially IPA Pathways."

Data

- Bioinformatics community members need streamlined, large and free data storage solutions, yet our interviewees were not aware they have data management needs. They store data in multiple places and ways (3/7 multiple ways, 3/7 servers, 2/7 Broad servers, 4/7 KI and BioMicroCenter servers, 4/7 laptop, etc). Very few interviewees expressed data storage support needs or issues (1/7 data storage support and 2/7 data storage issue) when asked about it.
- Many bioinformatics data types were mentioned in interviews: sequencing, genomic, imaging, proteomics, transtone, metagenomics, ocean survey data, expression data, secretoprotein, molecular interactions, cellular function, patient data.
- Interviewees share data (6/7 shares data) and they use repositories for retrieving source data (5/7 uses repositories for source data). Data repositories mentioned by name when asked about the use of repositories for source data during the interviews include Encode, Broad Institute Datasets, Cancer Cell Line Encyclopedia, Achilles Project, Immport, ProPortal, NCBI, GEO, Nucleotide, The Cancer Genome Atlas, Ensemble, and UCSC Genome Browser.
- There may be large datasets associated with 214 theses related to bioinformatics published from 2008-2012.
- Some interviewees are interested in data visualization tools (3/7 interested in data visualization tools and support). The topic of data visualization came up during the interviews without being prompted by a specific question about data visualization.

Interdisciplinarity

- Interviewees expressed that bioinformatics is interdisciplinary (6/7 bioinformatics at MIT is happening everywhere).
- All interviewees (7/7 Collaborative research approach within MIT) collaborate on research with other MIT researchers and almost all (5/7 collaborative research approach beyond MIT) collaborate with researchers outside MIT.
- Even though the Broad Institute split from MIT, some MIT bioinformatics community members (3/7) are still affiliated with the Broad Institute and have many advantages like server access for data storage and networking and collaboration capabilities.
- The breadth of subjects represented by the department affiliations of the commercial tool users, departments offering courses related to bioinformatics and issuing MIT theses related to bioinformatics show that bioinformatics is interdisciplinary, from anthropology to physics to biology to civil engineering, and is a campus wide need, not just a need of the life science research community at MIT.

- Many courses related to bioinformatics are cross-listed across 2 or 3 departments and are being offered by unexpected departments like Anthropology and Science Technology & Society.
- Theses published from 2008-2013 containing bioinformatics keywords came from 21 departments; the top five departments producing theses containing bioinformatics keywords include Biology, Electrical Engineering & Computer Science, Mechanical Engineering, Biological Engineering and Health Sciences & Technology

Instruction

- Training is important (2/7 desire for statistical/mathematical skills and 3/7 desire for modular training), although interviewees did not necessarily look for it from the libraries (1/7 use MIT Libraries bioinformatics support and 3/7 have attended MIT Libraries' IAP sessions).
- Some responses to the interview question "how do the MIT Libraries support your bioinformatics research and how you learn bioinformatics skills" include: "My first year, I took an IAP class on R and that was helpful by exposing me to code with someone there to answer questions. I think that format is best, when you have an assignment and there is someone there to help you and answer questions, somewhat self-taught. But the class was helpful as an introduction" and "I appreciated Courtney's classes on PubMed and NCBI – how to look at nucleotide & protein sequences, that was very helpful in general."
- The 38 unique MIT courses related to bioinformatics represent various types of courses: methodology, application, and theory.
- Since 2005, the Bioinformatics Program at MIT Libraries has offered 63 classes averaging 18 attendants per session.
- Instruction statistics from the Bioinformatics Program at MIT libraries show a high attendance for classes taught by outside experts (NCBI Discovery).
- "Bioinformatics for Beginners," the hallmark instruction session of the Bioinformatics Program at MIT Libraries developed and presented by library staff has an average of 14 attendants per session.
- Of the top five attended classes, only one is developed and presented by library staff (Bioinformatics for Beginners).

Recommendations & Implications

Recommendation and implications for the Bioinformatics Program and MIT Libraries are listed below grouped into themes.

Computation Skills

- The Bioinformatics Librarian should investigate the possibility of purchasing a modular training package, like Open Helix, to meet the desire for modular training expressed during the interviews.

- What are the risks and needs associated with DIY approaches to both computational skills and tools. What are they missing? How can the libraries support or facilitate this DIY approach?

Tools

- Based on the interviews and use statistics of commercial bioinformatics tools licensed by the MIT Libraries, there is a definite need in the community for commercial tools and the Libraries have developed a role to support this need. The Bioinformatics Librarian should continue to focus energy, knowledge building and collection efforts on commercial software.
- There may be a role for the Libraries regarding open source tools used by the community. The bioinformatics librarian should be aware of popular and frequently used open source tools at MIT and investigate possible roles for the Libraries to help this community grow and gain knowledge across campus.
- Given the high number of MIT courses related to "bioinformatics", there might be an opportunity for (more) embedded instruction with the commercial software tools.
- Marketing of MIT Libraries' statistics support and classes is relevant to this community and should be continued.

Data

- The Bioinformatics Librarian should continue to be a member of the Research Data Service Working Group and should investigate research data needs and solutions specific to the bioinformatics community through informal conversations, independent study or formal project work.
- Given the large number of theses related to bioinformatics (appendix 6) and the general knowledge of data associated with bioinformatics research, there may be service opportunities for MIT Libraries to offer such as helping these students with data storage through DSpace@MIT and sharing knowledge about data management through the Research Data Services Working Group.
- Library administration should investigate the possibility of providing or facilitating services and/or support in the area of data visualization. This topic was brought up during an interview and was not prompted by an interview question. There may be opportunities to develop services for the life science community and other departments around campus.

Interdisciplinarity

- When advocating on behalf of the bioinformatics community through outreach, instruction and materials selection, the Bioinformatics Librarian should consider the needs of all departments, not just the traditional life sciences departments. Additionally, the Libraries should consider what role they could play in creating a community that crosses departmental boundaries.

- The interviews were a great reminder that there is so much you can learn by just getting out and talking with the community. The Bioinformatics Librarian is encouraged to make efforts to communicate more with the community and to attend relevant group meetings of Boston based bioinformaticians, researchers and service providers that gather weekly at MIT.

Instruction

- Bioinformatics instruction should be a continued part of the Bioinformatics Program at MIT Libraries.
- Considering the current capacity and subject expertise of the Bioinformatics Librarian, developing new curriculum may not be the best use of time; outsourcing to research experts and purchasing a modular training package might be a better option.
- Possible future instruction or training topics gleaned from the interviews include: programming class, data management, open source tools, data visualization, and bioinformatics data repositories.

Appendices

Appendix 1: Interview Questions

General

How/where did you gain computational skills?

Where do you go when you need help with bioinformatics?

Where and how do people you work with that don't have computational skills get those? (undergrad versus grad?)

What are your obstacles or barriers to using/learning/completing research in bioinformatics?

How do you think these barriers or obstacles can be overcome?

How do the MIT Libraries support your bioinformatics research or how you learn bioinformatics skills?

Tools

What bioinformatics tools are you using in your research?

What commercial, for-fee bioinformatics tools are you using in your research?

What non-commercial, open source developed tools are you using in your research?

What in-house developed bioinformatics tools are you using in your research? Who was it developed by?

Are you developing tools or programs yourself?

What types of bioinformatics tools or software do you wish you had access to? Why can't you use those tools?

What types of bioinformatics analysis do you wish you could conduct in your research? Why are you unable?

Have you used commercial or open source or in-house developed tools on the same research project?

Data

Can you describe the data you generate?

What do you do with the data you produce after you have used it?

Tell me about where you store this data.

Tell me about how you share this data.

How do your incorporate data from repositories or other sources (other than data you created yourself)? How do use this data? How you decide what data to use?

Wrap up

Where is bioinformatics happening at MIT?

Do you collaborate with other departments on/in bioinformatics research? Outside MIT?

Is there anything you want to share with me regarding bioinformatics at MIT?

We would like to speak to more people, could you suggest anyone that would be good to speak with?

Appendix 2: Interview Coded Themes

Themes	1	2	3	4	5	6	7	Totals	% Total
Computational Skills									
Self-taught computational skills	■		■	■	■	■	■	6	86
Biology students do not gain bioinformatics skills						■	■	2	29
Desire for modular training		■			■	■		3	43
Has access to computational skills	■	■	■	■		■	■	6	86
Lab members have diverse skills	■	■		■		■	■	5	71
Uses the KI Core & BioMicroCenter	■	■	■		■		■	5	71
Has used MIT Libraries bioinformatics support							■	1	14
Has attended MIT Libraries IAP	■	■			■			3	43
Desire for biomedical journal collection and faster ILB				■		■		2	29
Desire for statistical/mathematical skills					■		■	2	29
Tools									
Tools: in house development for lab		■	■			■	■	4	57
Tools: in house development shared			■	■		■		3	43
Tools: in house developed tools users	■	■	■			■	■	5	71
Tools: open source tool users	■	■	■	■	■	■	■	7	100
Tools: commercial tool users		■	■	■		■		4	57
Tools: not using commercial tools	■						■	2	29
Tools: software versioning issues	■							1	14
Desire for cost sharing of tools			■	■				2	29
Dissatisfaction with commercial tools	■		■	■				3	43
MATLAB	■	■		■		■		4	57
R	■	■		■			■	4	57
Python	■				■	■	■	4	57
Data									
Data stored in multiple ways		■	■	■				3	43
Data stored on servers		■	■	■				3	43
Data stored on Broad servers		■				■		2	29
Data stored on KI & BioMicroCenter servers		■	■	■			■	4	57
Data stored on laptop, etc		■		■	■		■	4	57
Looking for data storage support		■						1	14
Expresses data storage issue	■	■						2	29
Shares data	■	■	■	■		■	■	6	86
Not sharing data					■			1	14
Uses repositories for source data	■		■		■	■	■	5	71
Interested in data visualization tools and support		■		■	■			3	43

Wrap Up									
Themes	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Totals**	**% Totals**
Bioinformatics at MIT is happening in life science depts								4	57
Bioinformatics at MIT is happening everywhere								6	86
Collaborative research approach beyond MIT								5	71
Collaborative research approach within MIT								7	100
Affiliated with Broad								3	43

Appendix 3 Use Statistics for Commercial Bioinformatics Software Tools

BIOBASE

BIOBASE Use Statistics from 2009-2012

	2012	2011	2010	2009
Unique Visitors	320	436	503	466
Number of Visits	657	841	1031	1195

Ingenuity Pathways Analysis (IPA)

Summary of IPA use over 4 years of license

	2013	2012	2011	2010
Total length of all session logins (hrs)	1258	1461	1053	801
Total number of session logins	323	315	*	*
Total users	125	103	90	91

Session login is defined as any time any active account has logged into the tool. MIT Department affiliation was taken from the MIT Data Warehouse using email accounts associated with each IPA account.

License year 4: June 2013-2014

May 1 2013-April 30 2014

DLCs	Student	Researcher	Faculty	Affiliate	Length of Session Logins (hrs)	Number of Session Logins	Total Number of Active Users
Biological Engineering	6	4	0	0	50	20	10
Biology	33	7	1	0	369	69	41
Brain & Cognitive Sciences	1	0	1	0	61	5	2
Chemical Engineering	1	1	0	0	14	4	2
Civil & Environmental Engineering	1	0	0	0	4	1	1
Computational & Systems Biology Initiative	3	0	0	0	47	9	3
CSAIL	0	2	0	0	26	6	2
Electrical Engineering & Computer Science	4	0	0	0	26	7	4
Health Sciences & Technology	4	0	0	0	39	18	4
Institute for Medical Engineering & Sciences	0	1	0	0	20	4	1
Koch Institute	0	23	0	0	269	87	23
McGovern Institute for Brain Research	0	1	0	0	93	10	1
Media Arts & Sciences	1	0	0	0	3	1	1
Unknown	0	0	0	11	121	30	11
Picower Institute for Learning & Memory	0	7	0	0	71	27	7
Research Laboratory for Electronics	0	1	0	0	5	3	1

Whitehead Institute	0	11	0	0	40	22	11
Totals	54	58	2	11	1258	323	125

IPA Department Level Use Statistics License Year 3 2012

April 2012-April 2013

DLCs	Student	Researcher	Faculty	Length of Session Logins (hrs)	Session Logins	Total Number of Active Users
Biological Engineering	8	9	0	199	53	17
Biology	20	8	2	324	82	30
Brain & Cognitive Sciences	1	0	2	5	5	3
Chemical Engineering	1	1	0	13	6	2
Computational & Systems Biology Initiative	2	0	0	5	3	2
CSAIL	0	2	0	20	6	2
Electrical Engineering & Computer Science	3	0	0	21	7	3
Health Sciences & Technology	3	0	0	14	3	3
Koch Institute	0	19	0	331	71	19
Materials Processing Center	0	1	0	4	1	1
Materials Science & Engineering	0	1	0	15	5	1
McGovern Institute for Brain Research	0	1	0	147	9	1
Physics	1	0	0	12	1	1
Picower Institute for Learning & Memory	0	8	0	229	31	8
Research Laboratory of Electronics	0	1	0	5	2	1
Whitehead Institute	0	9	0	117	30	9
Totals	39	60	4	1461	315	103

IPA Department Level Use Statistics License Year 2 2011

April 2011- April 2012

DLCs	Student	Researcher	Faculty	Affiliate	Length of Session Logins (hrs)	Total Number of Active Users
Biological Engineering	9	8	0	1	254	18
Biology	18	9	2	0	433	29
Chemistry	1	0	0	0	7	1
Computational & Systems Biology Initiative	2	0	0	0	25	2
Electrical Engineering & Computer Science	2	0	0	0	3	2
Health Sciences & Technology	6	1	0	0	19	7
Koch Institute	0	16	0	0	170	16
McGovern Institute for Brain Research	0	2	0	0	10	2
Picower Institute for Learning & Memory	0	4	0	0	49	4
Whitehead Institute	0	10	0	0	83	10
Total	38	50	2	1	1053	91

IPA Department Level Use Statistics License Year 1 2010

May 2010-April 2011

Department	Student	Researcher	Faculty	Other	Total Length of all Session Logins	Total Number of Active Users
Biological Engineering	10	6	1	1	167	18
Biology	16	6	1	0	170	23
Brain & Cognitive Sciences	1	0	0	0	1	1
Center for Environmental Health Sciences	1	1	0	0	4	2
Center for Global Change Science	0	1	0	0	2	1
Chemical Engineering	0	1	0	0	9	1
Chemistry	0	1	0	0	3	1
Civil & Environmental Engineering	1	0	0	0	2	1
Computational Biology & Systems Initiative	4	0	0	0	70	4
Electrical Engineering and Computer Science	1	0	0	0	1	1
Health Sciences & Technology	8	1	0	0	62	9
Koch Institute	0	14	0	0	163	14
Microfluids Lab	0	1	0	0	4	1
Picower Institute for Learning & Memory	0	3	0	0	41	3
Whitehead Institute	0	11	0	0	102	11
Totals	42	46	2	2	801	91

GeneGo

GeneGo Usage Statistics Summary from 2009-2012

	2012	2011	2010	2009
Length of Time Logged In (h:min:secs)	281:04:51	264:12:09	407:29:54	169:94:08
Total Number of Users	60	55	54	37

Active account is defined as any account that has been logged into for longer than 1 minute. Department and status of individual accounts were determined by using the MIT Roles database and the MIT People directory.

GeneGo Department Level Use Statistics License Year 4 2012

September 2011-October 2012

DLCs	Grad	UG	Post Doc	Research Sci./Affil.	Faculty	Employee	Affil.	Total Number of Users	Login in time (h:min:sec)
Koch Institute	0	0	11	6	0	1	1	19	101:15:21
Whitehead Institute	0	0	5	1	0	7	1	14	75:36:05
Biology	3	1	2	0	1	0	0	7	14:02:57
Biological Engineering	3	1	3	1	1	0	0	9	22:24:40
Health Sciences & Technology	1		0	0	0	0	0	1	1:24:44
Mathematics	0	1	0	0	0	0	0	1	12:24:56
Picower Institute for Learning & Memory	0	0	1	0	0	0	0	1	16:40:12
Chemical Engineering	1	0	0	0	0	0	0	1	5:15:32
Mechanical Engineering	1	0	0	0	0	0	0	1	0:52:01
Civil & Environmental Engineering	0	1	0	0	0	0	0	1	27:21:00
Physics	1	0	0	0	0	0	0	1	0:02:00
McGovern	0	0	0	0	0	1	0	1	0:03:58
Other	0	0	0	0	0	0	3	3	3:41:25
Totals	10	4	22	8	2	9	5	60	281:04:51

GeneGo Department Level Use Statistics License Year 3 2011

MIT Usage January 2011-September 2011

DLCs	Grad	UG	Post Doc	Research Sci/Affil.	Faculty	Employee	Affil.	Total Number of Active Users	Login in time (h:min:sec)
Koch Institute	0	0	4	6	0	0	0	10	88:39:46
Whitehead Institute	0	0	5	1	0	7	0	13	64:53:57
Biology	5		2		0	0	0	7	38:18:23
Biological Engineering	3	1	4	2	0	0	0	10	26:35:20
Health Sciences & Technology	2		1	0	0	0	0	3	15:17:07
Mathematics	0	1	0	0	0	0	0	1	10:31:20
Picower Institute for Learning & Memory	0	0	1	0	0	0	0	1	7:29:41
Chemical Engineering	1	0	0	0	0	0	0	1	4:05:47
Division of Comparative Medicine	0	0	1	0	0	0	0	1	2:39:03
MIT Portugal Program	0	0	0	1	0	0	0	1	2:16:31
Civil & Environmental Engineering	0	1	0	0	0	0	0	1	1:35:48
NA	0	0	0	2	0	0	0	2	1:15:30
Lincoln Lab	0	0	0	0	0	1	0	1	0:27:52
Computational & Systems Biology	2	0	0	0	0	0	0	2	0:03:29
Center for Environmental Health	0	0	0	0	0	1	0	1	0:02:35
Totals	13	3	18	12	0	9	0	55	264:12:09

GeneGo Department Level Use Statistics License Year 2 2010

January 2010-September 2010

DLCs	Grad	UG	Post Doc	Research Sci./Affil.	Faculty	Employee	Affil.	Total Number of Active Users	Login in time (h:min:sec)
Koch Institute	0	0	9	4	0	0	1	14	179:43:36
Whitehead Institute	0	0	2	0	0	14	1	17	105:21:02
Biological Engineering	3	1	4	0	1	0	0	9	51:20:11
Biology	3	1	1	0	1	0	0	6	45:22:11
N/A		0	0	0	0	0	1	1	12:03:40
Health Sciences & Technology	2	0	0	0	0	0	2	4	11:58:41
Brain & Cognitive Sciences	1	0	0	0	0	0	0	1	1:11:36
Chemistry	1	0	0	0	0	0	0	1	0:16:01
Lincoln Lab	0	0	0	0	0	1	0	1	0:12:56
Totals	10	2	16	4	2	15	5	54	407:29:54

GeneGo Department Level Use Statistics License Year 1 2009

January 2009-October 2009

DLCs	Grad	UG	Post Doc	Research Sci/Affil.	Faculty	Employee	Affil.	Total Number of Active Users	Login in time (h:min:sec)
Whitehead Institute	0	0	6	0	0	6	0	12	61:01:59
Koch Institute	0	0	3	5	0	0	1	9	56:12:19
Biological Engineering	1	0	1	0	0	0	0	3	28:16:39
Health Sciences & Technology	0	0	0	1	0	0	1	2	16:00:02
Biology	3	0	0	1	1	0	0	5	6:51:25
Brain & Cognitive Sciences	1	0	0	3	0	0	0	4	0:51:26
Chemical Engineering	1	0	0	0	0	0	0	1	0:15:56
Electrical Engineering & Computer Science	1	0	0	0	0	0	0	1	0:14:22
Totals	7	0	10	10	1	6	2	37	169:44:08

Appendix 4 MIT Courses Related to Bioinformatics
Collected from the MIT 2013/14 course catalog

1.S82 Special Problems in Environmental Microbiology and Chemistry
6.047 Computational Biology: Genomes, Networks, Evolution
6.503 Foundations of Algorithms and Computational Techniques in Systems Biology
6.581J Foundations of Algorithms and Computational Techniques in Systems Biology
6.802 Computational Systems Biology
6.872J Biomedical Computing
6.874J Computational Systems Biology
6.878J Advanced Computational Biology: Genomes, Networks, Evolution
7.13 Experimental Microbial Genetics
7.31 Current Topics in Mammalian Biology: Medical Implications
7.32 Systems Biology
7.33J Evolutionary Biology: Concepts, Models and Computation
7.36J Foundations of Computational and Systems Biology
7.493J Microbial Genetics and Evolution
7.57 Quantitative Biology for Graduate Students
7.81J Systems Biology (Same subject as 8.591J)
7.82 Topics of Mammalian Development and Genetics
7.89J Topics in Computational and Systems Biology (Same subject as CSB.100J)
7.91J Foundations of Computational and Systems Biology
8.591J Systems Biology
9.015J Molecular and Cellular Neuroscience Core
9.26J Principles and Applications of Genetic Engineering for Biotechnology and Neuroscience
9.520 Statistical Learning Theory and Applications
10.548J Tumor Pathophysiology and Transport Phenomena: A Systems Biology Approach (Same subject as HST.525J)
10.555J Bioinformatics: Principles, Methods and Applications (Same subject as HST.940J)
10.965 Seminar in Biosystems Engineering
15.077J Statistical Learning and Data Mining
18.418 Topics in Computational Molecular Biology
20.213 DNA Damage and Genomic Instability
20.390J Foundations of Computational and Systems Biology (Same subject as 7.36J)
20.482J Foundations of Algorithms and Computational Techniques in Systems Biology (Same subject as 6.581J)
20.490J Foundations of Computational and Systems Biology (Same subject as 7.91J)
21A.303J The Anthropology of Biology
CSB.100J Topics in Computational and Systems Biology
CSB.110 Research Rotations in Computational and Systems Biology
CSB.190 Research Problems in Computational and Systems Biology

CSB.199 Teaching Experience in Computational Systems Biology
HST.160/161 Molecular Biology and Genetics in Modern Medicine (Subject meets with HST.160)
HST.506J Computational Systems Biology (Same subject as 6.874J)
HST.507J Advanced Computational Biology: Genomes, Networks, Evolution (Same subject as 6.878J)
HST.508 Quantitative Genomics
HST.509 Computational and Functional Genomics
HST.510 Genomics and Computational Biology
HST.525J Tumor Pathophysiology and Transport Phenomena: A Systems Biology Approach
HST.527 Blood Vessels and Endothelial Phenotypes in Health and Disease
HST.940J Bioinformatics: Principles, Methods and Applications(Same subject as 10.555J)
STS.034 Science Communication: A Practical Guide

Appendix 5 Bioinformatics Instruction Statistics

Collected from April 2005 through February 2014

Class	Times Offered	Attendance	Average Attendance
NCBI Discovery Workshop	3	254	85
Bioinformatics for Beginners	15	217	14
Practically Genomic	6	156	26
Gene Pattern Workshop	2	116	58
GeneGo	8	91	11
IPA	5	62	12
BIOBASE	7	51	7
Biology Dept Technology Seminar	2	44	22
Introduction to Genome and Protein Sequence Analysis	2	37	19
7.16-NCBI Training	4	36	9
7.16-IPA training	3	34	11
An Introduction to Analysis Methods for Microarray Data	1	27	27
BLAST	4	23	6
Ensemble Workshop	1	13	13
Totals	63	1161	18

Appendix 6 MIT Theses Related to Bioinformatics

Departments	2008	2009	2010	2011	2012	Totals
Aeronautics & Astronautics	1	0	0	0	1	2
Biological Engineering	6	4	4	2	3	19
Biology	11	6	12	11	8	48
Chemical Engineering	3	3	4	2	0	12
Chemistry	1	1	2	4	2	10
Civil Engineering	2	0	0	2	0	4
Computation & Systems Biology	0	1	6	4	4	15
Computation for Design and Optimization	0	1	1	0	0	2
Earth Atmospheric and Planetary Sciences	1	0	3	2	2	8
Electrical Engineering & Computer Science	6	7	8	5	5	31
Engineering Systems Division	0	2	2	1	0	5
Health Sciences & Technology	6	6	2	2	3	19
Humanities	0	0	0	1	0	1
Materials Science	0	0	1	2	0	3
Mathematics	0	1	1	1	0	3
Mechanical Engineering	4	10	1	4	1	20
Media Arts and Science	2	0	0	2	0	4
Operations Research Center	0	0	0	0	1	1
Physics	1	2	1	1	0	5
Science Technology and Society	0	0	1	0	0	1
Sloan	0	0	0	0	1	1
Totals	44	44	49	46	31	214

Appendix 7 List of Tools Mentioned in Interviews

Blast
HMMer
Muscle
SMART
PFAM
ClustalW
T-COFFEE
SOAL
BGI's algorithm
Trinity
Bowtie
BWA
Tophat
Cufflinks
cummeBUND
DESeq
COG cluster pipeline
RNA-Seq pipeline
ChIP-Seq pipeline
Galaxy
CLC Workbench
Bioconductor R
cMonkey
DAVID
GOStat
GSEA
GenePattern
SpotFire
Ingenuity
FastTree
iTOL
PhyML
Artemis
DART
SAMtool
PyCogent
Genetic Interaction Tool
NCBI
Reactome
MatLab
Integrated Genome Viewer
Pasa
Immport
Genious
Lasergene
Python
RNA Fold
Vienna RNA package
Sci-py
Mole-pie
UCSC Genome Browser
Ensemble
PCLamp
GPS Macs
TRANSFAC
Encode Project
Psicquic
KEGG
Cytoscape
Cancer Genome Atlas
PRISM
Histone Modifaction
MACS
MEME
GEO