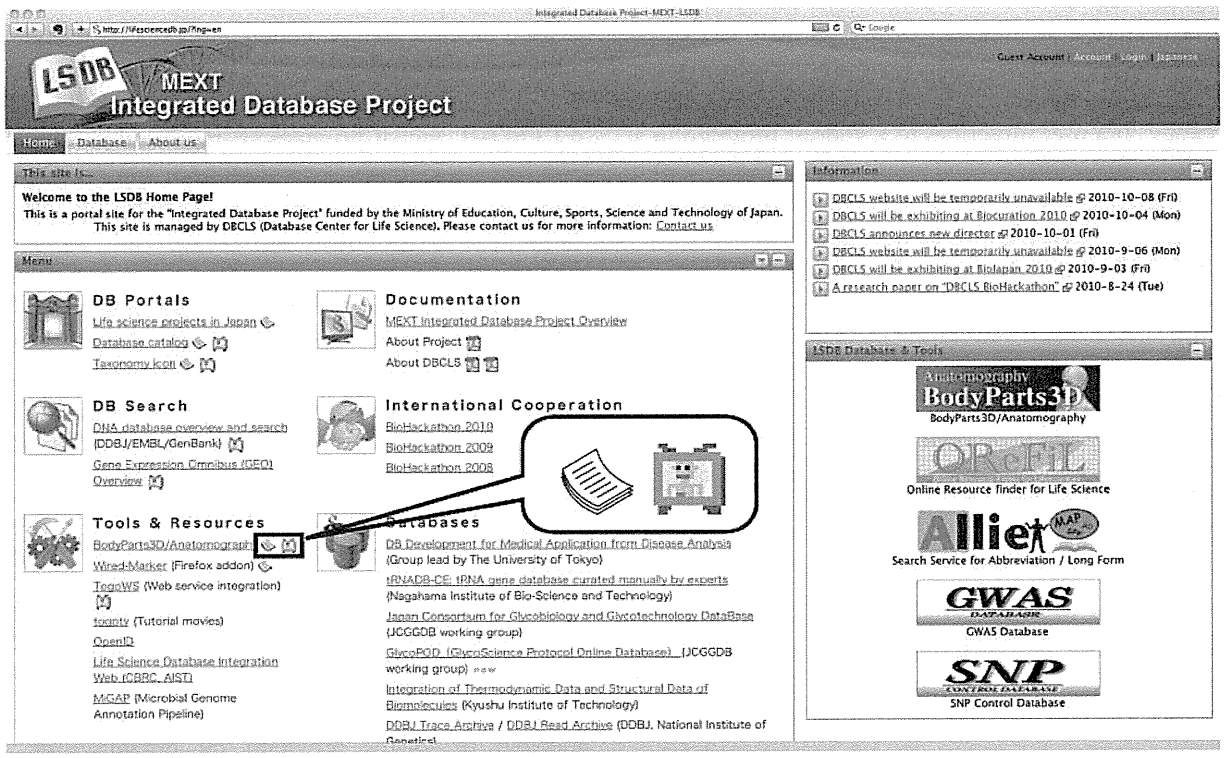

Figure 1: Screenshot of the MEXT Integrated Database Project portal (http://lifesciencedb.jp/en/). Paper icons and TV icons zoomed in the call-out following service names are linked to PDF documents and tutorial videos, respectively. MEXT, Ministry of Education, Culture, Sports, Science and Technology of Japan.

as TogoWS [56], which provides an integrated SOAP and REST APIs for interoperable bioinformatics Web services and OReFiL [57], which is an online resource finder for life science. We plan to expand our own English contents so as to enable our service to be used all over the world.

Making videos

Currently, two types of video are provided: (i) tutorial videos of databases and tools (screencasts) and (ii) lecture videos of symposiums and workshops (live action). For screencast videos, a screen where the database or tools were operated was captured and edited using screencast software equipped with a caption-adding function, such as Camtasia Studio (TechSmith Corporation, Okemos, MI, USA) for Windows and DesktopToMovie (Pencil Software, Okinawa, Japan; only a Japanese-language version is available) for Mac [58–60]. Recently, Camtasia:Mac (TechSmith Corporation) has been released, and we recommend its use rather than DesktopToMovie. For live-action videos, a lecture was recorded using a digital video camera or voice recorder, and then the source media was edited or embedded with presentation slides using tools such as Final Cut Pro (Apple Inc.) or iMovie (Apple Inc.). It is also possible to output presentations in Keynote (Apple Inc.) to videos. After capturing and editing, the source media was encoded in QuickTime format (.mov) and MPEG-4 format (.m4v) for distribution via websites and vodcast, respectively. The video compression type was set to H.264, and the sound format was specified to AAC if an audio track was included. For encoding in the QuickTime format, the 'Prepare for Internet Streaming' option was set to 'Fast Start' rather than 'Fast Start—Compressed Header' because the compressed header file format is impossible to play on Flash players. Other useful software packages for screencasting are listed in the article of Wikipedia entitled 'Comparison of screencasting software' [61].

To create user-friendly and high-quality tutorials, we suggest the following points: plan the tutorial; do a run-through before recording; edit adequately; pause at essential points; make the duration as short as possible and keep effects to a minimum.

To capture a video smoothly, it is important to create a plan and run through it before recording. Editing costs may increase considerably if these preparatory steps are skipped. Here, editing involves deleting unnecessary frames and loading animation frames, thus ultimately reducing video downloading time, user viewing time and also file size. At key operating points, it is necessary to pause the animation; viewers need time to understand and absorb the information. In TogoTV videos, we insert a pause of about 5–10 s, depending on the situation. We also recommend that the video duration be made as short as possible and that animation effects be suppressed to a minimum. Most TogoTV contents fit in a 5-min video, except for lectures. Excessive production not only increases the production cost but also conceals the essence of the video. In general, since a dynamic video tends to increase file size, suppression of excessive animation will reduce the file size. Most video repositories have upload limitations based on video length and file size.

The most important thing when creating a video is to create a high-quality video that would be useful to viewers. When one creates video easily without any consideration of the quality of the product, it would be a waste of viewer's time and content creator's time and would add to the already overwhelming 'noise' of available training materials. Because both creating and viewing video are time consuming, one needs to create a video carefully. In a case of TogoTV, we have adopted an internal review in order to ensure quality. From planning to drafting, reviewing and publishing takes about a week in our case.

Video distribution

We used a blog to distribute videos. This allows the video creation date to be clear to viewers since the web service interface changes often. In addition, it is possible to easily implement a comment-posting system and an update-notification system via RSS. We selected tDiary for the original TogoTV site [62]. It is a blog kit written in the Ruby programming language with the ability to easily add functions using plug-in programs such as update notification (makerss.rb) via RSS feeds. In addition to plug-in programs included with tDiary, we activated third-party plug-in programs, e.g. opening/closing caption texts that appear in the video (netabare.rb). We also developed plug-in programs to display the view-count ranking, query word trends by keyword search from Google and video recommendations (similar to the Amazon recommendation system). We chose WordPress for English interface of TogoTV [63]. It has smart interface compared with that of tDiary and can also be easily extended using plug-in programs.

We used JW Player (LongTail Video, New York, NY, USA), an open source software program for playing Flash video on web browsers [64]. The latest version of JW Player supports HTML5 video elements; thus, a user can view a video directly on an iPod, iPhone or iPad. The H264 Streaming Module for Apache was installed on the server for video streaming on the hypertext transfer protocol (HTTP) instead of setting up a real-time messaging protocol (RTMP) server [65]. This module provides the ability to start a video at any specified point.

A screenshot of the TogoTV home page is shown in Figure 2 [54]. There are links for video categories, most-watched videos and discrete tutorial videos sorted by creation date, similar to a blog. Every video has category tags to enable the searching of a video by categories, such as 'Genome analysis', 'Visualization' and 'Sequence analysis'. By clicking a date or thumbnail image, an individual video screen page will open (Figure 3). In the video screen page, there are description of the video, summary text of the story, and all captions or slides in the video in addition to the video player. Clicking the thumbnail on the video player plays the video from the beginning, and clicking a summary text or the thumbnail of a slide plays the video from the specific time at which the caption or slide appears.

We also distribute videos as vodcasts, so anyone can download and watch the videos on their own computer or handheld device, such as an iPod, iPhone or iPad via iTunes by subscribing to TogoTV vodcast. The vodcast distribution website was constructed using iWeb, which was a part of the iLife package (Apple Inc.). YouTube videos are also provided to maximize the reach of the videos, and it is possible to watch the videos even when the original TogoTV server is down for maintenance.

When the TogoTV service began, YouTube restricted upload file size and duration time up to 1 GB and 10 min, respectively. Above all, the video quality was extremely low, for example, it was difficult to read captions. Thus, we decided to use our own distribution server. There are many features that have made possible by preparing our own server such as recommendations, the ability to play

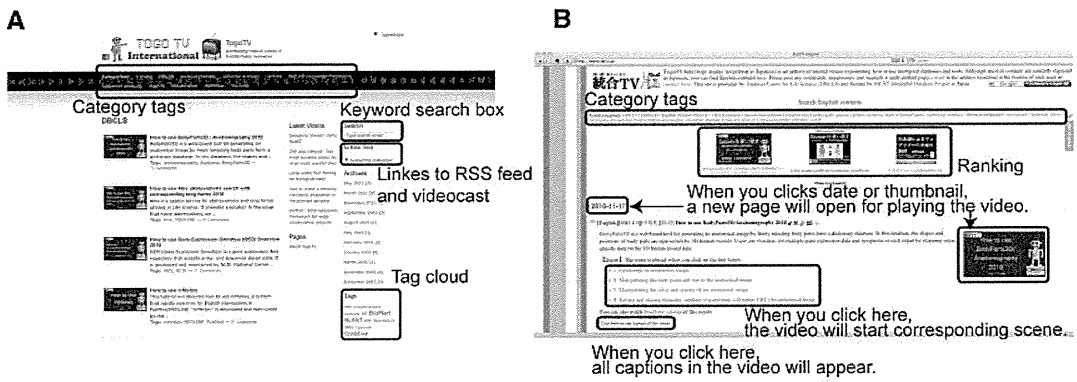

Figure 2: Screenshot of the (**A**) international version of and (**B**) original version of TogoTV website. All contents are sorted by publishing date. There are list of category tags, daily/weekly/monthly ranking of playing video (only for the original site), publishing date, description, thumbnail, summary of the story and link to all captions in the video. When you click a date or thumbnail, a new page will open for playing the video (Figure 3A). When you click a sentence of summary, the video will start corresponding scene.

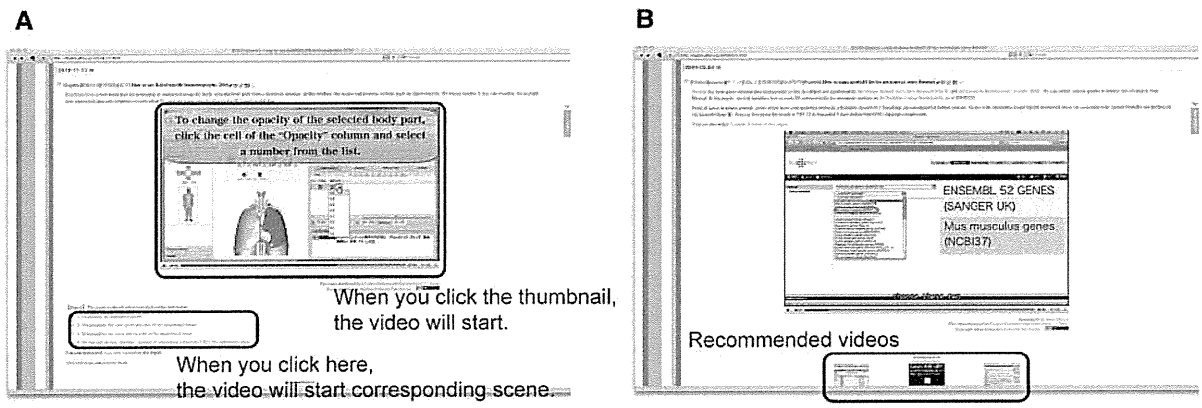

Figure 3: Screenshot of TogoTV's tutorial programs. Clicking the thumbnail in the center of the page will start the video. When the summary text of a video is clicked, the video will start the scene corresponding to the summary. There are links to recommended videos that are related playing video at the bottom of the page. (**A**) This video is entitled 'How to use BodyParts3D/Anatomography 2010'. (**B**) This video is entitled 'How to make probeID list for microarray using BioMart'.

from specific time and updating a video without changing URL. In addition, search of YouTube video hits more noise because the YouTube contains various genres, not limited to educational videos. However, YouTube now allows high-definition quality videos, and limitations of file size and duration time increased up to 2 GB and 15 min, respectively [66]. In addition, the time limitation of upload videos can be derestricted if users have complied with the YouTube community guidelines and copyright rules [67]. For these reasons, when distributing tutorial videos, we recommend using YouTube rather than a private server. Indeed, YouTube contains many useful tutorial videos. When we searched YouTube 'biomart' as a query, 42 videos were found. In the search result, 22 videos were noise that was unconcerned to biology, and 13 videos including one lecture video were uploaded by us (account: togotv). The rest included five videos provided by one of the service provider (account: EnsemblHelpdesk) and two videos provided by BioMart users. YouTube provides video-embedding code, so anyone can embed a video on any website. Clicking the embed button below a video shows the embedding code as well as options and skins for embedding.

When we publish a new video, an announcement is distributed using RSS and Twitter [68, 69]. The Twitter notification is a reposting of the RSS feed using Twitterfeed [70].

Update of content

Since the databases and tools are constantly evolving, it is important to keep up with updating the corresponding tutorial videos. Speedy update of the content is one of advantages of TogoTV; it cannot do in a paper device. Update of the resources is constantly monitored using a news release, RSS and DBCLS Database Catalog, which is a service developed in the Integrated Database Project [71]. An entry in this catalog contains 'Last Modified' date information that is automatically collected by a crawling program. When a resource is updated significantly or changed its interface, we replace an old tutorial video with a recreated video as soon as possible. In May 2011, we have 310 tutorial videos in TogoTV (excluding lecture videos), and a total of 54 videos are updated ones.

Examples of tutorial videos in TogoTV

How to use BodyParts3D/anatomography

Anatomography is a 3D rendering tool for human anatomy and has been developed as part of the Integrated Database Project [72]. A user can generate anatomical images by selecting body parts stored in the BodyParts3D database and setting their opacities, colors and viewpoint [73]. The image is useful for communication between physicians and patients, and it can be generated as a heat map of the human body based on an organ name and a numeric value such as organ-specific gene expression data and cancer mortality. Figure 3A shows a screenshot of the BodyParts3D/anatomography tutorial video [74]. This video describes how to build a 3D image, how to manipulate viewpoint and size, how to set opacities and colors and how to output to an image file. Videos of other services provided by the Integrated Database Project are also available at the project page [55]. A TV icon after the service name (Figure 1) provides a link to a tutorial video.

How to use BioMart

We provide how-to videos of not only our own services but also useful tools all over the world. BioMart, a query-oriented data management system, is one of the most important tools in genome science [75]. Users can submit various queries to retrieve lists of interest from BioMart. A screenshot of the tutorial video entitled 'How to make probeID list for microarray using BioMart' is shown in Figure 3B [76]. In this video, the process for creating an ID conversion list for microarray analysis on the BioMart central portal website is introduced [77]. From all genes in the mouse genome, genes that have corresponding entries in the Affymetrix mouse430 2 GeneChip are considered for further analysis. Genes with the Affymetrix GeneChip ID mentioned above are associated with the Agilent ProbeID and RefSeq ID via the Ensembl Gene ID. The results are downloadable in the tab separated value format with GNU-zip (.gzip) compression.

Lectures

Lectures are also distributed from TogoTV. Currently, we broadcast eight lectures in English. One is a video of a lecture about this service (TogoTV) held at the 2007 Annual Conference of the Japanese Society for Bioinformatics (JSBi2007, Figure 4A) [78]. The second lecture is about Gendoo [79], a functional profiling tool for gene and disease features using the Mesh vocabulary, held at JSBi2008 [80]. The third is about copyright and data sharing in science entitled 'Copyright in the Digital Age and Its Impact on Scientific Data Sharing' by Professor Lawrence Lessig from Harvard Law School from the 'Balancing Intellectual Property Protection and Data Sharing in Science' symposium (Figure 4B) [81]. The others are about processing of large genomic data from 'Workshop on Parallel and Distributed Processing of Large Genome Data' [82].

In addition to hosting TogoTV, we also organize workshops, which is called AJACS (All Japan Annotators/Curators/System DB administrators) workshop, for Japanese users of biological databases and tools for educational purposes [83]. We have published a lecture video with an online handout for participants to review and for interested parties who could not attend. Currently, there are 63 lecture videos in Japanese from 13 workshops.

Statistics and user feedback

In May 2011, we provided over 450 videos, most of which are in Japanese. TogoTV is accessed 20 000 times per month from 5000 unique IP addresses and videos are played 4000 times per month. According to the analysis of IP addresses, accesses

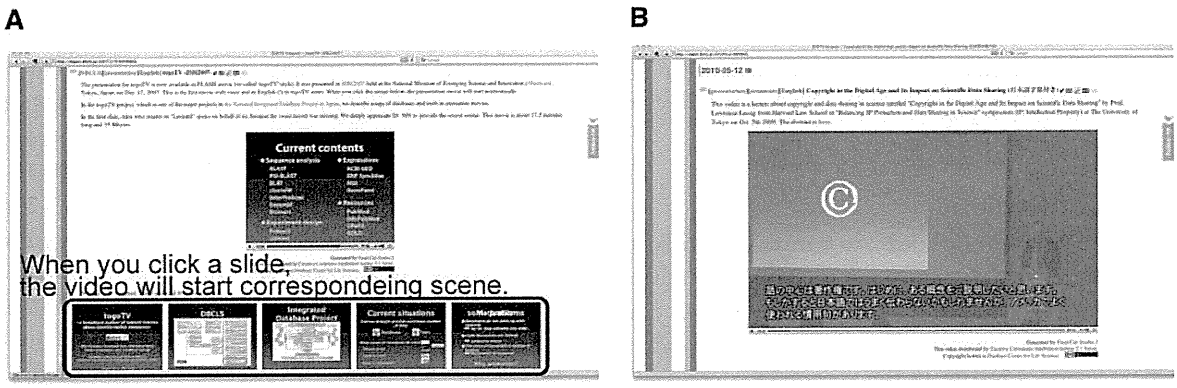

Figure 4: Screenshot of TogoTV's lecture programs. There are the slides appeared in the presentation at the bottom of the page. When you click a slide, the video will start corresponding scene. (**A**) Video of voice-over on slides. (**B**) Live-action video.

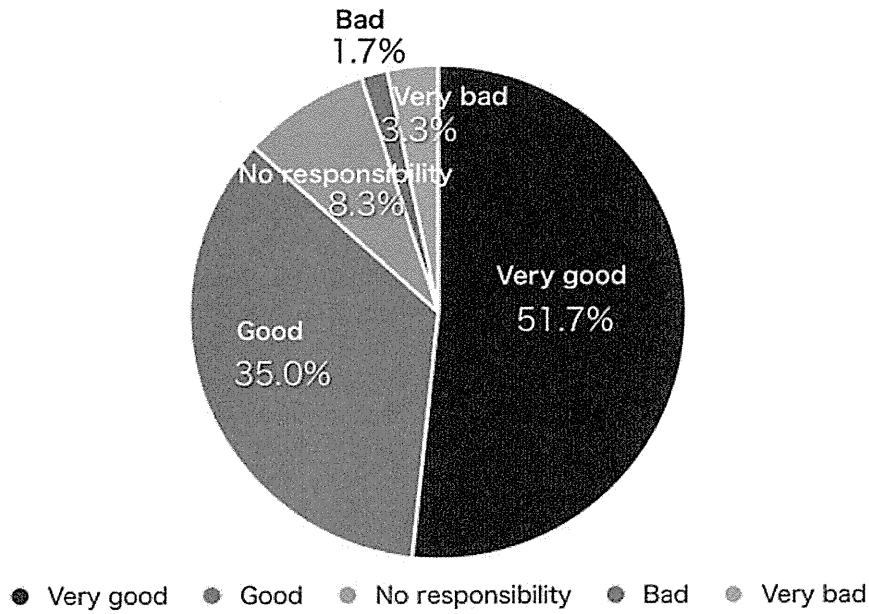

Figure 5: User feedback about TogoTV. External evaluation about the services provided by the Integrated Database Project was carried out via web from September to October 2009 [89]. There were 117 respondents including principal investigators, corporate researchers, postdocs and PhD students in total, and 60 peoples evaluated TogoTV. The question was 'For this service, please give us feedback on how to proceed in the future'. Respondents chose from five options: 'Because it seems pretty useful, I suggest it should be driven forward' (Very good); 'Because it seems useful in its own way, I suggest it may be driven forward' (Good); 'I do not have responsibility to assess it' (No responsibility); 'Because it seems unuseful, I suggest it need not be driven forward' (Bad); and 'Because it seems completely unuseful, it should be dropped' (Very bad).

from '.jp' and '.com' domain accounted for about 33% and 31%, respectively. The rest included unknown domain (28%), '.net' (5%) and '.edu' (1%) domains. Although contents of TogoTV are mainly Japanese, there are accesses from outside of Japan. The analysis of access was carried out using AWStats, and we excluded the bot programs from the analysis [84]. Videos accessed on YouTube measured through the YouTube API are played about 5000 times per month. Since TogoTV was started in July 2007, videos have been watched >250 000 times. The tutorial videos of basic tools

and databases such as primer3, reactome, blast, clustalw, and Ensembl are the popular contents in TogoTV [36, 85–88].

Reactions of most viewers are positive judging from the responses received during interviews at lectures and exhibitions. According to some instructors who used our videos for their bioinformatics lectures, their students' understanding increased considerably. Figure 5 shows a part of the results of the Integrated Database Project user feedback investigation in FY2008–2009. More than 85% of reviewers reacted positively to our service.

CONCLUSION

It is possible to provide tutorials in the video format because the continued development of computer hardware and software and the Internet. Various service developers have created and distributed tutorial videos via their own server or YouTube; our contribution was the launching of TogoTV, which provides over 450 videos and is one of the most active site collecting and maintaining tutorial videos. As pointed out by Williams *et al.* [3], tutorial videos are not always in depth enough to provide a full understanding of the resource to users. But it is effective for them to touch on use of web resources. Further development of computer and web infrastructure will accelerate this movement in the future.

It would be helpful if database and tool developers publish how-to videos as well as documents to encourage greater use. Even if one is just a user of web resources and not a service developer or provider, the creation and distribution of a tutorial video based on experience is useful for numerous researchers. The most important thing is to create a high-quality video that would be useful to viewers. To show summary text, captions and dialogues of a video as well as description is also useful in determining whether or not to watch the video. The creation of tutorial videos in a community is particularly useful for sharing and standardizing the annotation and curation process.

We believe that providing tutorial videos created by database and tool providers as well as users will promote research activities and help to distribute the knowledge of database and tool handling in research communities. Thus, we propose that everyone who produces and uses web resources create tutorial videos and share them.

Key Points

- Improvements of computer technology and the Internet enable creation and distribution of a tutorial video easily.
- Some major database and tool developers provided tutorial videos via their website and/or YouTube.
- We developed TogoTV, a website where tutorial videos of bioinformatics databases, tools and lectures are distributed, and this attempt acquired a good reputation.
- Let us create and share tutorial videos all together.

Acknowledgements

The authors would like to thank Yoko Ohmura, Yoko Yamaguchi, Kouichi Takewaka, Atsuko Chiba, Takuya Hashimoto, Takao Yokoyama, Yuki Okuda, Tatsuro Ohta, Haruko Hirukawa, Shoichiro Ohishi, Tomohiro Ono, Hayato Sakata, Tamayo Suizu, Hideyuki Takeda, Tetsuya Negoro, Maori Ito, Natsuki Kubo and Fumio Takahashi for creating videos, Yusuke Kumagae for the plug-in development and Takeru Nakazato for supervising web design. All contributors are presented on the TogoTV just like 'end-title roll' at http://togotv.dbcls.jp/movie/endroll_for_togotv.mov. They also thank Dr Ayumi Koso and Dr Mari Minowa for internationalization of the contents and summarization of user feedback.

FUNDING

This work was supported by the Integrated Database Project of Ministry of Education, Culture, Sports, Science and Technology in Japan [07046015, 10100921].

References

1. Galperin MY, Cochrane GR. The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res* 2011;**39**: D1–6.
2. Brazas MD, Yamada JT, Ouellette BF. Providing web servers and training in Bioinformatics: 2010 update on the Bioinformatics Links Directory. *Nucleic Acids Res* 2010;**38**: W3–6.
3. Williams JM, Mangan ME, Perreault-Micale C, *et al.* OpenHelix: bioinformatics education outside of a different box. *Brief Bioinform* 2010;**11**:598–609.
4. Cassie JM, Collins GF, Daggett CJ. The use of videotapes to improve clinical teaching. *J Med Educ* 1977;**52**:353–4.
5. Ruiz JG, Mintzer MJ, Issenberg SB. Learning objects in medical education. *Med Teach* 2006;**28**:599–605.
6. Vigeant D, Lefebvre H, Reidy M. The use of video as a pedagogic tool for the training of perioperative nurses: a literature review. *Can Oper Room Nurs J* 2008;**26**:8–9, 14–5, 17–20.
7. Repanos C, Anderson D, Earnshaw J, *et al.* Manipulation of nasal fractures with local anaesthetic: a 'how to do it' with online video tutorial. *Emerg Med Australas* 2010;**22**:236–9.
8. YouTube – Broadcast Yourself. http://www.youtube.com/ (24 May 2011, date last accessed).

9. Dailymotion – Online Videos, Music, and Movies. Watch a Video Today! http://www.dailymotion.com/ (24 May 2011, date last accessed).
10. Vimeo, Video Sharing For You. http://www.vimeo.com/ (24 May 2011, date last accessed).
11. List of video hosting services – Wikipedia the free encyclopedia. http://en.wikipedia.org/wiki/List_of_video_hosting_services (24 May 2011, date last accessed).
12. Ustream.tv: You're On. http://www.ustream.tv/ (24 May 2011, date last accessed).
13. Justin.tv – Streaming live video broadcasts for everyone. http://www.justin.tv/ (24 May 2011, date last accessed).
14. Stickam – The Live Community, Live Streaming, Video Chat. http://www.stickam.com/ (24 May 2011, date last accessed).
15. JoVE: Journal of Visualized Experiments – a Video Journal for Biological and Medical Research. http://www.jove.com/ (24 May 2011, date last accessed).
16. SciVee | Making Science Visible. http://www.scivee.tv/ (24 May 2011, date last accessed).
17. DnaTube.com – Scientific Video Site. http://www.dnatube.com/ (24 May 2011, date last accessed).
18. Free Online Course Materials | MIT OpenCourseWare. http://ocw.mit.edu/ (24 May 2011, date last accessed).
19. MIT World | Distributed Intelligence. http://mitworld.mit.edu/ (24 May 2011, date last accessed).
20. UT OpenCourseWare. http://ocw.u-tokyo.ac.jp/english (24 May 2011, date last accessed).
21. Academic Earth | Online Courses | Academic Video Lectures. http://academicearth.org/ (24 May 2011, date last accessed).
22. YouTube – Education – YouTube EDU. http://www.youtube.com/edu (24 May 2011, date last accessed).
23. TED: Ideas worth spreading. http://www.ted.com/ (24 May 2011, date last accessed).
24. OCW Consortium – OpenCourseWare Websites. http://www.ocwconsortium.org/courses/ocwsites (24 May 2011, date last accessed).
25. List of educational video websites – Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/List_of_educational_video_websites (24 May 2011, date last accessed).
26. Boulos MN, Maramba I, Wheeler S. Wikis, blogs and podcasts: a new generation of Web-based tools for virtual collaborative clinical practice and education. *BMC Med Educ* 2006;**6**:41.
27. Jham BC, Duraes GV, Strassler HE, *et al.* Joining the podcast revolution. *J Dent Educ* 2008;**72**:278–81.
28. Thapa MM, Richardson ML. Dissemination of radiological information using enhanced podcasts. *Acad Radiol* 2010;**17**:387–91.
29. Apple – iTunes U – Learn anything, anywhere, anytime. http://www.apple.com/education/itunes-u/ (24 May 2011, date last accessed).
30. Cooper PS, Lipshultz D, Matten WT, *et al.* Education resources of the National Center for Biotechnology Information. *Brief Bioinform* 2010;**11**:563–9.
31. Tutorials. http://www.ncbi.nlm.nih.gov/education/tutorials/ (24 May 2011, date last accessed).
32. YouTube – NCBINLM's Channel. http://www.youtube.com/ncbinlm (24 May 2011, date last accessed).
33. Zhang H, Morrison MA, Dewan A, *et al.* The NEI/NCBI dbGAP database: genotypes and haplotypes that may specifically predispose to risk of neovascular age-related macular degeneration. *BMC Med Genet* 2008;**9**:51.
34. dbGaP Tutorial. http://www.ncbi.nlm.nih.gov/projects/gap/tutorial/dbGaP_demo_1.htm (24 May 2011, date last accessed).
35. PubMed Online Training. http://www.nlm.nih.gov/bsd/disted/pubmed.html (24 May 2011, date last accessed).
36. Flicek P, Amode MR, Barrell D, *et al.* Ensembl 2011. *Nucleic Acids Res* 2011;**39**:D800–6.
37. YouTube – EnsemblHelpdesk's Channel. http://www.youtube.com/EnsemblHelpdesk (24 May 2011, date last accessed).
38. Huntley RP, Binns D, Dimmer E, *et al.* QuickGO: a user tutorial for the web-based Gene Ontology browser. *Database* 2009; doi:10.1093/database/bap010 (29 September 2009, date last accessed).
39. Video tutorials on the use of QuickGO. http://www.ebi.ac.uk/QuickGO/tutorial.html (24 May 2011, date last accessed).
40. Barrell D, Dimmer E, Huntley RP, *et al.* The GOA database in 2009-an integrated Gene Ontology Annotation resource. *Nucleic Acids Res* 2009;**37**:D396–403.
41. Gene Ontology Annotation (UniProtKB-GOA) | Tutorial | EBI. http://www.ebi.ac.uk/GOA/annotationexample.html (24 May 2011, date last accessed).
42. Goecks J, Nekrutenko A, Taylor J, *et al.* Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010;**11**:R86.
43. Galaxy | Published Page | Screencasts. http://main.g2.bx.psu.edu/screencast (24 May 2011, date last accessed).
44. Hull D, Wolstencroft K, Stevens R, *et al.* Taverna: a tool for building and running workflows of services. *Nucleic Acids Res* 2006;**34**:W729–32.
45. Videos | Taverna. http://prototype.taverna.org.uk/documentation/taverna-2-x/videos/ (24 May 2011, date last accessed).
46. Obayashi T, Nishida K, Kasahara K, *et al.* ATTED-II updates: condition-specific gene coexpression to extend coexpression analyses and applications to a broad range of flowering plants. *Plant Cell Physiol* 2011;**52**:213–9.
47. YouTube – ATTEDable's Channel. http://www.youtube.com/user/ATTEDable (24 May 2011, date last accessed).
48. BioInformatics Tutorial Series (BITS) – Research Guides at MIT Libraries. http://libguides.mit.edu/bits (24 May 2011, date last accessed).
49. Video Tutorials | Countway Library of Medicine. https://www.countway.harvard.edu/menuNavigation/libraryServices/classes/videoTutorials.html (24 May 2011, date last accessed).
50. Tutorials | BIREC. http://birec.org/tutorials (24 May 2011, date last accessed).
51. OpenHelix: List of available tutorials. http://www.openhelix.com/cgi/tutorials.cgi (24 May 2011, date last accessed).
52. Tip Of The Week | The OpenHelix Blog. http://blog.openhelix.eu/?category_name=tip-of-the-week (24 May 2011, date last accessed).
53. TogoTV (in Japanese). http://togotv.dbcls.jp/ (24 May 2011, date last accessed).

54. TogoTV | distributing tutorial videos of bioinformatics resources. http://togotv.dbcls.jp/en/ (24 May 2011, date last accessed).
55. Integrated Database Project – MEXT-LSDB. http://lifesciencedb.jp/en/ (24 May 2011, date last accessed).
56. Katayama T, Nakao M, Takagi T. TogoWS: integrated SOAP and REST APIs for interoperable bioinformatics Web services. *Nucleic Acids Res* 2010;**38**:W706–11.
57. Yamamoto Y, Takagi T. OReFiL: an online resource finder for life sciences. *BMC Bioinformatics* 2007;**8**:287.
58. Schumacher G, Thurkettle MA. Camtasia studio: a teaching tool for academicians and clinicians. *Comput Inform Nurs* 2007;**25**:130–4.
59. TechSmith | Camtasia Screen Recorder Software, Home. http://www.techsmith.com/camtasia/ (24 May 2011, date last accessed).
60. DesktopToMovie (in Japanese). http://pencilsoftware.com/dtm.html (24 May 2011, date last accessed).
61. Comparison of screencasting software – Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Comparison_of_screencasting_software (24 May 2011, date last accessed).
62. tDiary | Download tDiary software for free at SourceForge.net. http://sourceforge.net/projects/tdiary/ (24 May 2011, date last accessed).
63. WordPress > Blog Tool and Publishing Platform. http://wordpress.org/ (24 May 2011, date last accessed).
64. LongTail Video | Home of the JW Player. http://www.longtailvideo.com/ (24 May 2011, date last accessed).
65. H264–Trac. http://h264.code-shop.com/trac/wiki (24 May 2011, date last accessed).
66. YouTube Blog: Upload limit increases to 15 minutes for all users. http://youtube-global.blogspot.com/2010/07/upload-limit-increases-to-15-minutes.html (24 May 2011, date last accessed).
67. YouTube Blog: Up, Up and Away – Long videos for more users. http://youtube-global.blogspot.com/2010/12/up-up-and-away-long-videos-for-more.html (24 May 2011, date last accessed).
68. TogoTV RSS feed. http://togotv.dbcls.jp/en/feed/ (24 May 2011, date last accessed).
69. TogoTV on Twitter. http://twitter.com/#!/togotv (24 May 2011, date last accessed).
70. Twitterfeed.com : feed your blog to twitter. http://twitterfeed.com/ (24 May 2011, date last accessed).
71. Integrated Database Project-MEXT-LSDB. http://lifesciencedb.jp/lsdb.cgi?gg=dbcatalog&lng=en (24 May 2011, date last accessed).
72. BodyParts3D. http://lifesciencedb.jp/bp3d/?lng=en (24 May 2011, date last accessed).
73. Mitsuhashi N, Fujieda K, Tamura T, *et al.* BodyParts3D: 3D structure database for anatomical concepts. *Nucleic Acids Res* 2009;**37**:D782–5.
74. TogoTV – How to use BodyParts3D/Anatomography 2010. http://togotv.dbcls.jp/20101117.html (24 May 2011, date last accessed).
75. Haider S, Ballester B, Smedley D, *et al.* BioMart Central Portal-unified access to biological data. *Nucleic Acids Res* 2009;**37**:W23–7.
76. TogoTV – How to make probeID list for microarray using Biomart. http://togotv.dbcls.jp/20090304.html (24 May 2011, date last accessed).
77. BioMart – MartView. http://www.biomart.org/biomart/martview/ (24 May 2011, date last accessed).
78. TogoTV – togoTV – JSBi2007. http://togotv.dbcls.jp/20071219.html (24 May 2011, date last accessed).
79. Nakazato T, Bono H, Matsuda H, *et al.* Gendoo: functional profiling of gene and disease features using MeSH vocabulary. *Nucleic Acids Res* 2009;**37**:W166–9.
80. TogoTV – Functional profiling of OMIM data using MeSH vocabulary – JSBi2008. http://togotv.dbcls.jp/20081219.html (24 May 2011, date last accessed).
81. TogoTV – Copyright in the Digital Age and Its Impact on Scientific Data Sharing. http://togotv.dbcls.jp/20100512.html (24 May 2011, date last accessed).
82. Workshop on Parallel and Distributed Processing of Large Genome Data. http://mlab.cb.k.u-tokyo.ac.jp/en/events/lgd/ (24 May 2011, date last accessed).
83. AJACS – MotDB (in Japanese). http://motdb.dbcls.jp/?AJACS (24 May 2011, date last accessed).
84. AWStats - Free advanced log file analyzer for web, ftp or mail statistics (GNU GPL). http://awstats.sourceforge.net/ (24 May 2011, date last accessed).
85. Untergasser A, Nijveen H, Rao X, *et al.* Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res* 2007;**35**:W71–4.
86. Croft D, O'Kelly G, Wu G, *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* 2011;**39**:D691–7.
87. Altschul SF, Madden TL, Schäffer AA, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
88. Chenna R, Sugawara H, Koike T, *et al.* Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* 2003;**31**:3497–500.
89. Result – LSDB Evaluation Questionnaire (in Japanese). http://lifesciencedb.jp/feedback/hyouka20/?result#i760c229 (24 May 2011, date last accessed).

DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2011, Article ID bar013, doi:10.1093/database/bar013

Original article

Allie: a database and a search service of abbreviations and long forms

Yasunori Yamamoto[1,*], Atsuko Yamaguchi[1], Hidemasa Bono[1] and Toshihisa Takagi[2]

[1]Database Center for Life Science, Bunkyo-ku, Tokyo and [2]Department of Computational Biology, University of Tokyo, Kashiwa, Chiba, Japan

***Corresponding author:** Tel: +81 (0)3 5841 0251; Fax: +81 (0)3 5841 8090; Email: yy@dbcls.rois.ac.jp

Submitted 25 November 2010; Revised 25 March 2011; Accepted 28 March 2011

Many abbreviations are used in the literature especially in the life sciences, and polysemous abbreviations appear frequently, making it difficult to read and understand scientific papers that are outside of a reader's expertise. Thus, we have developed Allie, a database and a search service of abbreviations and their long forms (a.k.a. full forms or definitions). Allie searches for abbreviations and their corresponding long forms in a database that we have generated based on all titles and abstracts in MEDLINE. When a user query matches an abbreviation, Allie returns all potential long forms of the query along with their bibliographic data (i.e. title and publication year). In addition, for each candidate, co-occurring abbreviations and a research field in which it frequently appears in the MEDLINE data are displayed. This function helps users learn about the context in which an abbreviation appears. To deal with synonymous long forms, we use a dictionary called GENA that contains domain-specific terms such as gene, protein or disease names along with their synonymic information. Conceptually identical domain-specific terms are regarded as one term, and then conceptually identical abbreviation-long form pairs are grouped taking into account their appearance in MEDLINE. To keep up with new abbreviations that are continuously introduced, Allie has an automatic update system. In addition, the database of abbreviations and their long forms with their corresponding PubMed IDs is constructed and updated weekly.

Database URL: The Allie service is available at http://allie.dbcls.jp/.

Introduction

With the fast pace of progress in the life sciences and the increase of accompanying literature, new domain-specific terms such as gene, protein, chemical compound or disease names are routinely introduced. These terms often consist of multiple words, and many researchers create or use abbreviations for them in their articles. Chang *et al.* (1) reported on average one new abbreviation appears in every five to ten abstracts, and our survey showed that MEDLINE entries have increased by about 650 000 per year on average from 2004 to 2009. Existing dictionaries cannot keep up with this situation. As a result, the clarity of articles decreases (2) and polysemy or synonymy issues arise. Another study (3) reported that 81.2% of abbreviations are ambiguous and have an average of 16.6 meanings. For example, the abbreviation SPF may stand for any one of 'specific pathogen-free', 'S-phase fraction', 'sun protection factor' and more. Here, we call these terms that have abbreviations 'long forms'. In addition, several long forms have lexical variants. For example, 'acute myeloid leukemia' and 'acute myeloid leukaemia' share identical concepts, and both are abbreviated as *AML*. Both of these long forms frequently appear (5652 and 1270) in the MEDLINE data.

A significant problem is that not all abbreviations in the MEDLINE data appear with their corresponding long forms (4). This situation can make it difficult for researchers to understand articles, especially when these are outside of their fields of expertise. This circumstance often happens with the emergence of new high-throughput technologies such as microarrays. Moreover, document search systems such as PubMed would return many non-relevant entries when a polysemous abbreviation is used as a query.

To help researchers learn domain-specific abbreviations easily, we have developed a system called Allie that looks

Database, Vol. 2011, Article ID bar013, doi:10.1093/database/bar013

up abbreviation-long form pairs from the entire MEDLINE database. Allie displays either long forms or abbreviations that correspond to a query consisting of either an abbreviation or a long form, respectively. Thus, if 'SPF' is given as a query, Allie displays a list of its corresponding long forms mentioned in the above example (i.e. 'specific pathogen--free', etc.).

In addition, for each hit pair, Allie returns the research field in which it frequently appears along with other abbreviations that co-occur with it to help users quickly learn about the context of its appearance. For example, 'Dermatology' is the research field of 'sun protection factor', which is usually abbreviated as SPF, and UV, UVR or MED are some of the abbreviations that co-occur with the pair. This novel functionality thus provides users with a way of disambiguating polysemous abbreviations in addition to indicating PubMed/MEDLINE data in which a target pair appears. This information can also be used to narrow down a set of hit pairs or PubMed/MEDLINE entries; thus Allie can be used to find articles that contain a particular target pair in a contextual manner. Moreover, for those who want to use Allie from their own programs or web servers, Allie also implements Simple Object Access Protocol (SOAP) and Representational State Transfer (REST) interfaces.

As mentioned above, new abbreviations are introduced rapidly, and we take this issue seriously since Allie's target users are actively working researchers including database annotators and curators in life sciences. To update Allie periodically, we built an automatic update system that extracts pairs from newly added MEDLINE data and reflects them in Allie. Although there have already been several abbreviation search systems (1, 4–9), some do not exist any more or have not been updated for a long time (more than a year). Thus, to our knowledge, we can claim that Allie is the only system of its kind that is updated periodically.

Methods

Database construction

The database used by Allie is constructed in advance. The construction process consists of the following six consecutive tasks: (i) splitting MEDLINE data into sentences, (ii) extracting abbreviation-long form pairs from the sentences, (iii) merging lexical variants, (iv) applying a domain-specific dictionary to identify conceptually identical terms, (v) forming groups of conceptually identical pairs considering their appearances in MEDLINE and (vi) for each group, choosing representatives of the abbreviations and their corresponding long forms.

MEDLINE titles and abstracts are split into sentences by the tool sptoolkit, and pairs are extracted from the sentences by ALICE (10). ALICE achieved a recall of 95% and a precision of 97% on randomly selected MEDLINE data, and so Allie inherits this performance.

After obtaining a list of pairs, Allie merges some lexical variants in the long forms using UMLS SPECIALIST Lexicon (11). More precisely, Allie uses the 'Agreement and Inflection' file to map a term to its basic form. If there is a basic form whose inflectional form exactly matches a long form, it is replaced with the basic form only if the basic form is used as a long form elsewhere and it appears more frequently than the original one. If it appears less, all the long forms that exactly match the basic form are replaced with that original one (i.e. the inflectional form). In addition, if a subset of a long form that includes its last word exactly matches an inflectional form, Allie processes it similarly to cope with those long forms which consist of adjectives and an inflectional form such as 'Acute lymphatic leukaemia' (Acute + lymphatic leukaemia). In more detail, when a long form consists of n words, we express it as '$w_1\ w_2 \ldots w_n$', where w_1 is the first word and the w_n is the last. In this situation, if a term '$w_i \ldots w_n$' ($1 < i < n$) or 'w_n' exactly matches an inflectional form, Allie processes it in the same way.

Next, Allie normalizes those terms that are conceptually identical but that have different expressions using GENA (12) by applying an identifier to terms having identical concepts. This normalization is at a more conceptual level than the previous process is. For example, GENA returns the same concept ID to both 'premature atrial contractions' and 'premature atrial complexes', both of which are abbreviated as 'PAC', but these are not in the file mentioned above. In developing our database, we used a customized version of GENA, developed by its creator, which can identify not only gene names, but also chemical compounds or disease names. The method of identifying named entities and normalizing them is described in (12), which states that a trie-based algorithm with several heuristics is used to recognize entities in text. In addition, UMLS Metathesaurus (13) is used to identify and normalize chemical compounds and disease names.

For each concept ID, taking all pairs of abbreviations and long forms, if those pairs whose long forms share the same ID are treated as one pair, the synonymy problem can be solved to some extent. However, it introduces another problem. Here, we take two terms 'mitotic index' and 'S-phase fraction' as an example. In MEDLINE, 'S-phase fraction' is usually abbreviated as 'SPF', and 'mitotic index' is always abbreviated as either 'MI' or 'IM'. Therefore, Allie should not display 'mitotic index' when a user searches for long forms that correspond to 'SPF'. However, it is displayed if Allie identifies pairs by only applying concept IDs added by GENA because GENA does not consider relationships between abbreviations and their long forms. In this example,

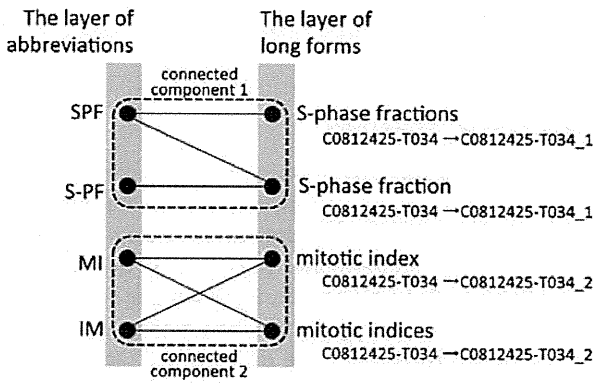

Figure 1. A part of a bipartite graph used in Allie. GENA gives the concept ID 'C0812425-T034' to 'S-phase fractions', 'S-phase fraction', 'mitotic index' and 'mitotic indices'. Allie changes the ID to one that corresponds to each cluster using connected components of the graph.

GENA would apply the same concept ID to 'mitotic index' and 'S-phase fraction'.

To take care of this problem, Allie changes the concept IDs of long forms after the application of GENA as follows. Allie constructs a bipartite graph with one layer representing a set of abbreviations and the other layer representing a set of long forms. Each edge denotes the existence of a pair in MEDLINE, and each long form is labeled by its concept ID given by GENA. Next, Allie computes the connected components of the graph. Then, Allie changes the concept ID given by GENA to a new one by concatenating a connected component ID as a suffix to the original one. For example, as shown in Figure 1, GENA gives the same concept ID 'C0812425-T034' to 'S-phase fractions', 'S-phase fraction', 'mitotic index' and 'mitotic indices'. Since the connected component that includes 'S-phase fraction' is different from the one that includes 'mitotic index', Allie appends the connected component ID to each concept ID, such as 'C0812425-T034_1' or 'C0812425-T034_2'. In other words, Allie divides each group of pairs with a same concept ID into subgroups by generating intersections of groups and connected components. Allie thus obtains the final groups of conceptually identical pairs by using the new concept IDs. Finally, for each group, the pair appearing most frequently in MEDLINE is selected as the representative of the group.

To obtain the research field of each pair, we use Journal Subject Terms, which are assigned by National Library of Medicine (NLM) to MEDLINE journals to describe the journals' overall scope. When multiple research fields are added to a pair, the most frequently added one is chosen.

Database update

The database update mainly consists of two parts. The first part performs tasks (1) and (2) (i.e. splitting MEDLINE data into sentences, extracting abbreviation-long form pairs from the sentences) and the second part consists of the remaining tasks (3) through (6). Since MEDLINE is usually updated every weekday, and the first part takes relatively a short amount of time to complete the task, it is run weekday. Since the second part takes more time, it is run once a month. Therefore, this update is reflected in the Allie search service monthly. We make notice here that the daily MEDLINE update data are obtained from NLM under a license agreement between NLM and DBCLS.

Search system description

Allie has four main pages, described below.

Top page

On the top page, Allie accepts a user query with advanced search options. A query can be an abbreviation, a long form, or a substring, and it must contain at least two ASCII characters. If a query matches either an abbreviation or a long form, Allie returns the corresponding long form or abbreviation clusters, respectively. A cluster is a group of conceptually identical abbreviations or long forms. In other words, if a query matches an abbreviation, the abbreviation becomes the search key and the list of its corresponding long form clusters is displayed. When a query matches both, the user may choose either one. We use the term 'item' to denote either abbreviation or long form, depending on the search key.

The user is provided options for the search method (i.e. exact match or partial match), the sorting order of the results (by hit clusters in a result page and by PubMed/MEDLINE information in each cluster), and the number of hit clusters shown per result page. The hit clusters can be sorted in ascending or descending order for each of the following:

- alphabetical order of cluster-representative items;
- appearance frequency (the number of the pairs appearing in MEDLINE);
- publication year of papers that contain the pairs in their titles or abstracts.

The sort order of PubMed/MEDLINE information for each cluster can be in ascending or descending order by publication year. The default values for the user options are exact match, descending frequency of appearance for the order of hit clusters, ascending publication year for PubMed/MEDLINE information, and 30 clusters per result page. Using these default values, a user can quickly find the most frequently used long form for an abbreviation and when the pair was first used in the literature along with its research field and co-occurring abbreviations.

Hit cluster-list page

When a query hits either abbreviations or long forms, Allie shows a hit cluster-list page (Figure 2A). Allie displays buttons for the user to choose if there are both or a message if it hits neither.

A hit cluster-list page is vertically divided into three parts. The top section shows the search conditions, where users can rearrange the order or change the cluster-list displayed if partial match is chosen for the search method and if the query hits multiple items. In addition, this section shows a menu by which the user can filter out those clusters that do not frequently appear in the articles of the chosen research field. The middle section shows the meta-data of the list shown, including the item that the query matches and the numbers of clusters and pairs. The lower section is a table with each cluster's information on a separate row. It contains the representative item (long form or abbreviation), research field, co-occurring abbreviations, and PubMed/MEDLINE information (publication years and titles of articles in which the pair appears in the titles or the abstracts). In cells of co-occurring abbreviations and PubMed/MEDLINE information, there are links to pages where users can find detailed information.

Co-occurring abbreviation page

On this page, there is a table where each row shows a co-occurring abbreviation, the frequency of co-occurrence with the hit pair, and its total appearance frequency (Figure 2B). Each abbreviation is anchor text that links to the hit cluster-list page for the abbreviation by exact match.

PubMed/MEDLINE information page

On this page, there is a table which lists the publication year, title, and co-occurring abbreviations that appear in the title or the abstract with the pair (Figure 2C). Each title is anchor text that links to the corresponding PubMed page. In addition, each co-occurring abbreviation is anchor text, similar to the co-occurring abbreviation pages.

Database download

While the Allie database for the search service is updated monthly, raw data extracted by ALICE are updated daily and are published weekly, which are freely downloadable from our FTP site (ftp://ftp.dbcls.jp/allie/). Since these are data that ALICE extracts without any post-processing such as clustering, these data may not reflect the same results as would be obtained from the Allie search system. These data are provided such that users can develop their own applications using them. These are tab-delimited text, where each line consists of a pair of an abbreviation and its corresponding long form with their unique IDs (i.e. an abbreviation ID and a long form ID), the PubMed ID of the title or the abstract where the pair appears, and its publication year.

Implementation

Allie consists of two main parts: an updating part (updater) and a search system part (searcher). The updater is a set of scripts that process MEDLINE data to generate a list of pairs and to update the database. The searcher was designed using Ruby on Rails and MySQL. Since some of the datasets needed for Allie are large (about one gigabyte) and since data retrieval takes time, the datasets are cached in the main memory of the Allie server.

As for the SOAP/REST interfaces, there are four types of searches, consisting of the combination of search methods (exact or partial) and search keys (abbreviation or long form). In addition, by using a pair ID obtained by a search, the co-occurring abbreviations and the PubMed/MEDLINE information can be obtained, respectively.

Results and an example usage

Table 1 shows examples of Allie's outputs. There are three abbreviations and their corresponding long forms with their research fields and the co-occurring abbreviations. It also shows the year of each long form's first appearance. At the time of writing of this manuscript, the total number of non-redundant pairs is 1 564 399, and the total numbers of the abbreviation and long form clusters are 406 372 and 1 341 981, respectively. The history of the updates shows that around 9000 new pairs are added monthly.

The following is an example usage of Allie. A researcher wants to know about the abbreviation 'SPF' that appears in a document without its long form nearby. The document describes a vaccine and that enzyme-linked immunosorbent assay (ELISA) was used in the experiment. Using Allie, he can find out that 'SPF' is a polysemous abbreviation, and that many articles are published in the journals pertaining to the research field of 'Veterinary Medicine' where it was used as an abbreviation of 'specific pathogen-free'. In addition, 'GF', 'IBDV' and 'ELISA' often co-occur with it as an abbreviation of 'specific pathogen-free' in MEDLINE. Since 'ELISA' appears in the document, he can speculate that the 'SPF' stands for 'specific pathogen-free'. By clicking the details link below the 'ELISA', and then clicking some of the listed abbreviations, he can find out that the pairs GF—'germ-free' and 'IBDV'—'infectious bursal disease virus' also co-occur with 'SPF'. Consequently, he can verify that the pair is correct.

We assume that the document in which an abbreviation in question appears contains several domain specific terms, and in many cases, he/she can find some clues to identify the proper long form by checking the research field, the co-occurring abbreviations, and the PubMed/MEDLINE

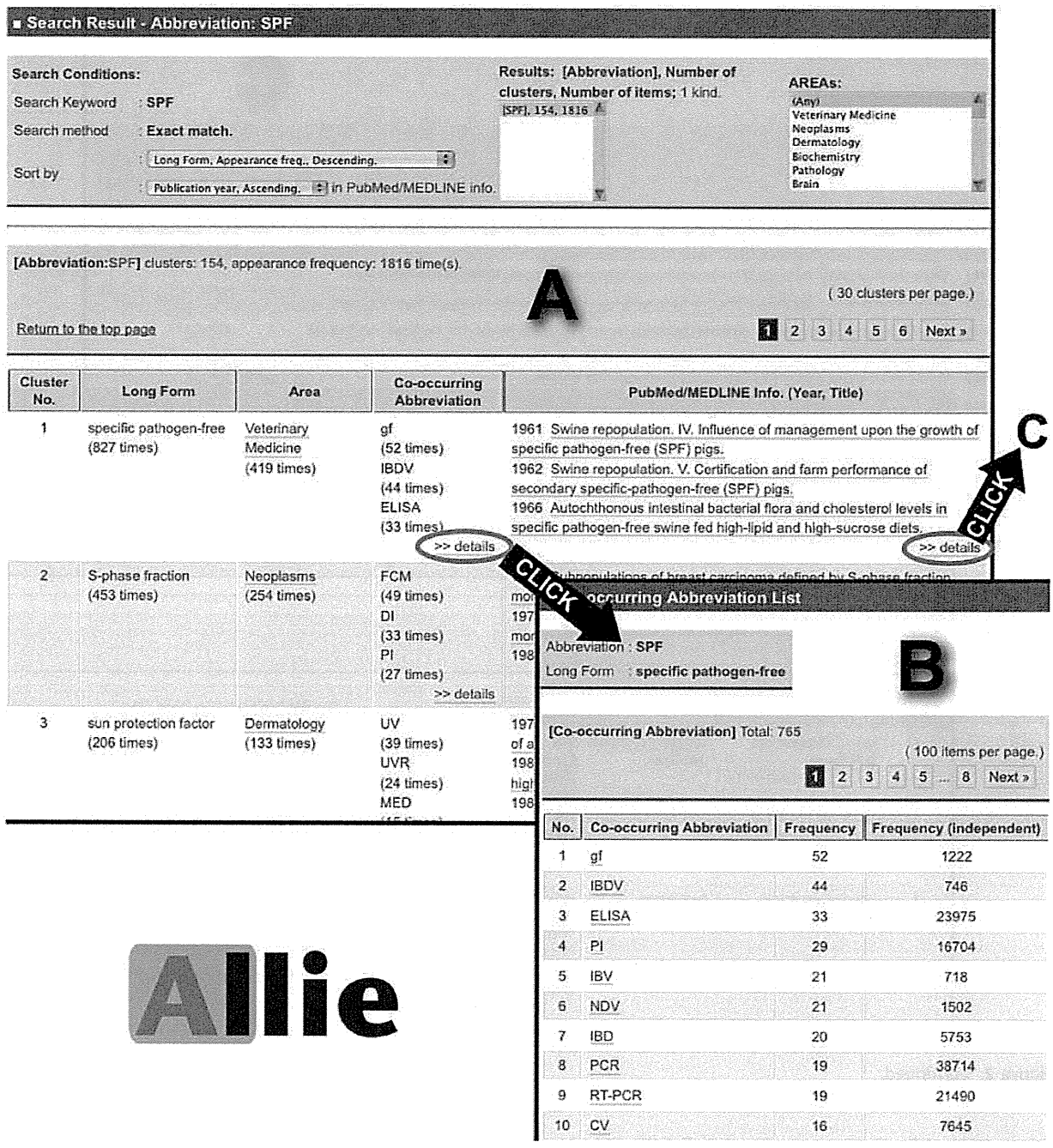

Figure 2. Images of Allie's outputs. (**A**) Hit cluster-list page for the abbreviation 'SPF'. By clicking links in the 'Co-occurring Abbreviation' or the 'PubMed/MEDLINE Info.' cells, the user can access these corresponding pages (A to B or A to C, respectively). (**B**) Co-occurring abbreviation page. Here, the user is provided with all the co-occurring abbreviations, and by clicking one of the listed abbreviations, one can access the hit cluster-list page. (**C**) PubMed/MEDLINE Information page. Here, the user is provided with all publication years, titles, and co-occurring abbreviations that appear in the titles or abstracts with the pair. Each title is anchor text that links to the corresponding PubMed page. By clicking one of the co-occurring abbreviations, the user can access the hit cluster-list page (C to D). (**D**) Hit cluster-list page for the abbreviation 'BVD'.

Database, Vol. 2011, Article ID bar013, doi:10.1093/database/bar013

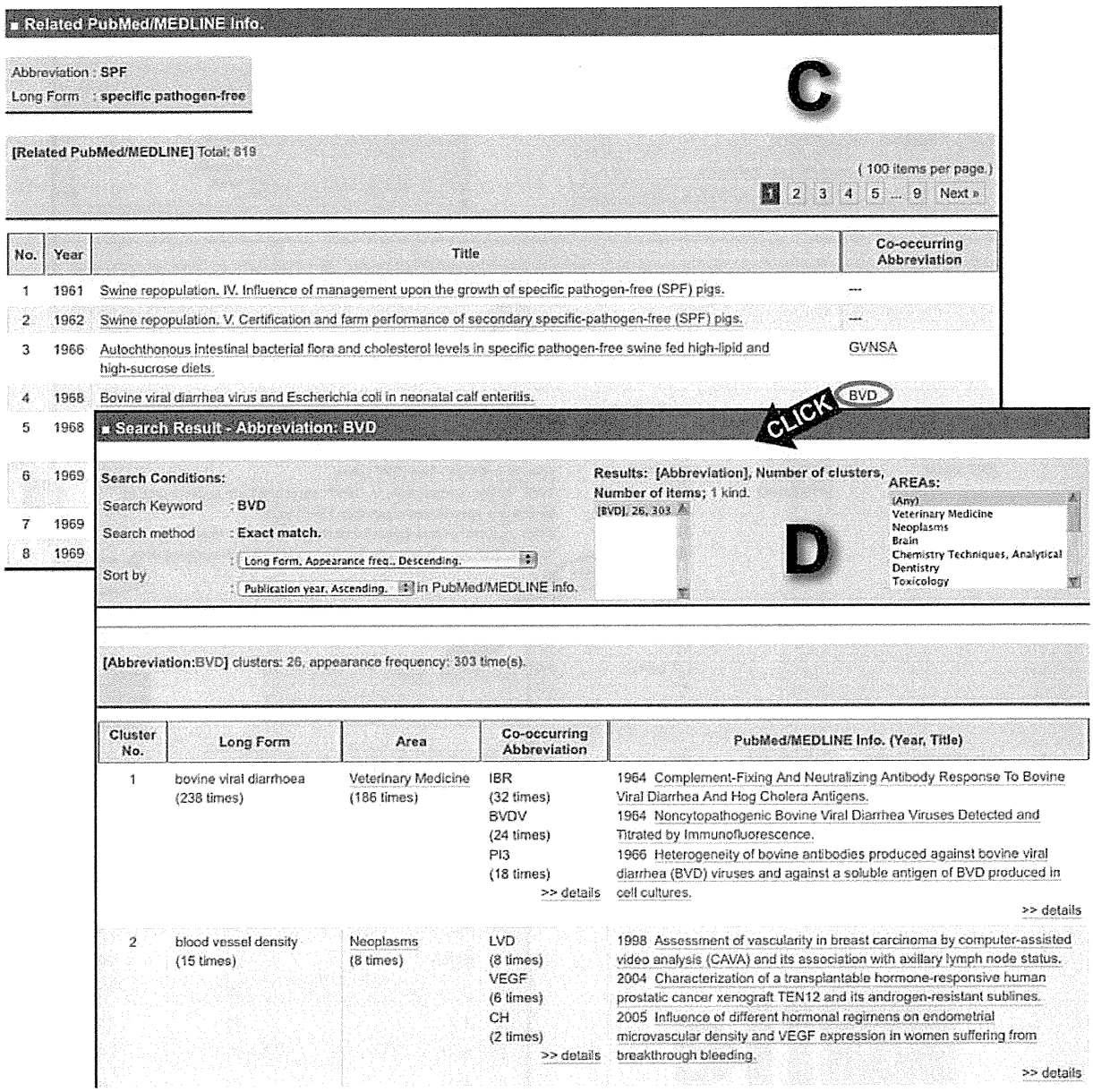

Figure 2. Continued.

information. Allie does not guess or predict the right long form for a given abbreviation, but instead it provides various related evidence for the user to quickly identify it.

Conclusion

Allie is a search system that returns not only pairs of abbreviations and their long forms appearing in the MEDLINE data, but also their relevant research fields and abbreviations. Providing the relevant information is a unique feature and makes it much easier for researchers in the life sciences to find the pair that they are looking for as demonstrated in the example usage. In addition, it is useful for users to utilize the database used in Allie in their environment since the entire database updated periodically is freely downloadable.

Discussions and future plans

The granularity of relevant research fields used to disambiguate polysemous abbreviations may be considered to be too coarse. To help users disambiguate easily, we

Table 1. Examples of Allie's outputs The long forms are sorted by descending frequency of appearance

Abbreviation	Long form	Research field	Co-occurring abbreviation	Year
SPF	Specific pathogen-free	Veterinary medicine	GF/IBDV/ELISA...	1961
	S-phase fraction	Neoplasms	FCM/DI/PI...	1978
	Sun protection factor	Dermatology	UV/UVR/MED...	1978
MAP	Mean arterial pressure	Physiology	HR/CO/CI...	1974
	Mitogen-activated protein	Biochemistry	ERK/JNK/PKC...	1991
	Mean arterial blood pressure	Physiology	HR/NO/CO...	1975
	Microtubule-associated protein	Neurology	AD/GFAP/NGF...	1979
BAC	Bacterial artificial chromosome	Genetics	FISH/YAC/PCR...	1994
	Blood alcohol concentration	Substance-Related Disorders	DUI/DWI/BrAC...	1994
	Bronchioloalveolar carcinoma	Neoplasms	AAH/NSCLC/EGFR...	1983
	Benzalkonium chloride	Ophthalmology	EDTA/CPC/CMC...	1979

considered that several granularity levels of the contexts where each pair appears should be provided. For example, if a user wants to know the correct long form for 'SPF' appearing in a document, as described above, providing several contexts assumed to be helpful. Enumerating from the finest to coarser levels, articles in which each pair appears are at the finest level of the granularity. At a coarser level there are co-occurring abbreviations. MeSH terms frequently annotated to articles in which each pair appears could be at a next coarser level. However, we considered MeSH terms at even coarser granularity would be better to be provided because no single user grasps the whole MeSH vocabulary. As for our pair database, a frequently appearing MeSH term would be chosen from about the 20 000 terms for each pair if we take the most frequently annotated MeSH term as a research field of each pair. The set of Journal Subject Terms is a subset of the MeSH terms, and the total number of these is 123. Therefore, these would be familiar enough to users to determine which research field is the right one even if they search for a pair outside of their areas of expertise. To our knowledge, there is no other vocabulary of its kind available now. Moreover, there is a delay (about three months) in the times of registration into MEDLINE and annotation of MeSH terms per each article since MeSH terms are manually annotated. Consequently, some newly emerged pairs do not have any MeSH terms. We also assume that abbreviation has one meaning within a research field, and we believe that the granularity is suitable. After providing relevant research fields, we have received positive feedbacks from users. Nevertheless, there may be a better way for users to disambiguate polysemous abbreviations. We will continue to survey user experiences and reflect its outcomes in Allie promptly.

Since the process of generating the data needed for Allie is fully automatic, the displayed results may contain incorrect pairs. These are caused by errors in extracting pairs, tagging terms, or grouping pairs. While ALICE exhibits good performance, a few new tools to extract pairs of abbreviations and long forms from text have been proposed such as BIOADI (8), Ab3P (14) and NatLab (15). To compare the extraction accuracy, we are evaluating ALICE on BIOADI and Ab3P corpora used in these studies and considering how we can increase the extraction performance. In addition, neither GENA nor SPECIALIST Lexicon have a complete list of conceptually identical terms, and we will research another way to complement it. Concerning handling of long forms that are lexically similar to each other but that are conceptually different, we have emphasized manually crafted dictionaries unless there is a special method. For example, chemical compound names are hard to determine whether two lexically similar names are conceptually identical or not ('albendazole sulphoxide' and 'albendazole sulfoxide' are conceptually identical to each other, but 'glutamic acid' and 'glutaric acid' are not). Therefore, we are developing an improved clustering method for grouping conceptually identical pairs based on graph algorithms and dynamic programming techniques. In this way, we are improving the performance as much as possible.

In addition, the downloadable database currently contains raw data generated by ALICE, but we are now planning to release a database used by the search system in which clustering results and the main research field and the co-occurrence abbreviations of each pair are included.

Acknowledgements

We thank Dr Shin Kawano for checking Allie and reporting an issue, Mr Toyofumi Fujiwara for helping with the development of Allie, and Mr Sebastian R. Riedel and Dr Kiyoko F. Aoki-Kinoshita for their comments on English writing.

Funding

Integrated Database Project, Ministry of Education, Culture, Sports, Science and Technology of Japan.

Conflict of interest. None declared.

References

1. Chang,J.T., Schutze,H. and Altman,R.B. (2002) Creating an online dictionary of abbreviations from MEDLINE. *J. Am. Med. Inform. Assoc.*, **9**, 612–620.
2. Bloom,D.A. (2000) Acronyms, abbreviations and initialisms. *BJU Int.*, **86**, 1–6.
3. Liu,H., Lussier,Y.A. and Friedman,C. (2001) A study of abbreviations in the UMLS. In: *Proceedings of AMIA Symposium*. Hanley & Belfus, Inc., Philadelphia, PA, USA, pp. 393–397.
4. Okazaki,N. and Ananiadou,S. (2006) Building an abbreviation dictionary using a term recognition approach. *Bioinformatics*, **22**, 3089–3095.
5. Rimer,M. and O'Connell,M. (1998) BioABACUS: a database of abbreviations and acronyms in biotechnology and computer science. *Bioinformatics*, **14**, 888–889.
6. Wren,J.D. and Garner,H.R. (2002) Heuristics for identification of acronym definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods Inform. Med.*, **41**, 426–434.
7. Zhou,W., Torvik,V.I. and Smalheiser,N.R. (2006) ADAM: another database of abbreviations in MEDLINE. *Bioinformatics*, **22**, 2813–2818.
8. Kuo,C.-J.J., Ling,M.H., Lin,K.-T.T. and Hsu,C.-N.N. (2009) BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature. *BMC Bioinformatics*, **10** (Suppl. 15), S7.
9. Okazaki,N., Ananiadou,S. and Tsujii,J. (2010) Building a high-quality sense inventory for improved abbreviation disambiguation. *Bioinformatics*, **26**, 1246–1253.
10. Ao,H. and Takagi,T. (2005) ALICE: An algorithm to extract abbreviations from medline. *J. Am. Med. Inform. Assoc.*, **12**, 576–586.
11. National Library of Medicine (US) (2009), UMLS® Reference Manual [Internet] http://www.ncbi.nlm.nih.gov/books/NBK9676/ (9 March 2011, date last accessed).
12. Koike,A. and Takagi,T. (2004) Gene/protein/family name recognition in biomedical literature *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*. Association for Computational Linguistics, Boston, Massachusetts, USA, pp. 9–16.
13. Schuyler,P.L., Hole,W.T., Tuttle,M.S. and Sherertz,D.D. (1983) The UMLS Metathesaurus: representing different views of biomedical concepts. *Bull. Med. Libr. Assoc.*, **81**, 217–222.
14. Sohn,S., Comeau,D.C., Kim,W. and Wilbur,W.J. (2008) Abbreviation definition identification based on automatic precision estimates. *BMC Bioinformatics*, **9**, 402.
15. Yeganova,L., Comeau,D.C. and Wilbur,W.J. (2010) Identifying Abbreviation Definitions—Machine Learning with Naturally Labeled Data. In: *2010 Ninth International Conference on Machine Learning and Applications*. IEEE Computer Society, Los Alamitos, CA, USA, pp. 499–505.

Biochemical and Biophysical Research Communications 407 (2011) 562–567

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

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

Gene expression profiling in multipotent DFAT cells derived from mature adipocytes

Hiromasa Ono [a,b], Yoshinao Oki [a], Hidemasa Bono [b], Koichiro Kano [a,*]

[a] Laboratory of Cell and Tissue Biology, College of Bioresource Sciences, Nihon University, 1866 Kameino, Fujisawa, Kanagawa 252-8510, Japan
[b] Database Center for Life Science (DBCLS), Research Organization of Information and Systems (ROIS), Faculty of Engineering Bldg.12 2-11-16 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan

ARTICLE INFO

Article history:
Received 10 March 2011
Available online 16 March 2011

Keywords:
Dedifferentiation
Microarray
Gene expression profiling
DFAT
Adipocytes

ABSTRACT

Cellular dedifferentiation signifies the withdrawal of cells from a specific differentiated state to a stem cell-like undifferentiated state. However, the mechanism of dedifferentiation remains obscure. Here we performed comparative transcriptome analyses during dedifferentiation in mature adipocytes (MAs) to identify the transcriptional signatures of multipotent dedifferentiated fat (DFAT) cells derived from MAs. Using microarray systems, we explored similarly expressed as well as significantly differentially expressed genes in MAs during dedifferentiation. This analysis revealed significant changes in gene expression during this process, including a significant reduction in expression of genes for lipid metabolism concomitantly with a significant increase in expression of genes for cell movement, cell migration, tissue developmental processes, cell growth, cell proliferation, cell morphogenesis, altered cell shape, and cell differentiation. Our observations indicate that the transcriptional signatures of DFAT cells derived from MAs are summarized in terms of a significant decrease in functional phenotype-related genes and a parallel increase in cell proliferation, altered cell morphology, and regulation of the differentiation of related genes. A better understanding of the mechanisms involved in dedifferentiation may enable scientists to control and possibly alter the plasticity of the differentiated state, which may lead to benefits not only in stem cell research but also in regenerative medicine.

1. Introduction

Mature adipocytes (MAs) possessing a single, large lipid droplet are generally considered to be in the terminal stage of differentiation, and having lost their proliferative ability, are stationary [1,2]. We previously established the preadipocyte cell line DFAT cells derived from dedifferentiated MAs from subcutaneous fat tissue of several animals using the ceiling culture method in the absence of any specific factors [3–5]. Interestingly, although DFAT cells represent a unique preadipocyte cell line that previously underwent terminal differentiation into MAs, these cells can not only redifferentiate into adipocytes but also differentiate into osteoblasts, chondrocytes, skeletal myocytes, smooth muscle cells, cardiomyocytes, vascular endothelial cells, and neural cells under appropriate culture conditions *in vitro* or *in vivo* [6–11]. These properties shown by DFAT cells indicate that differentiated cells having distinct functions *in vivo* can dedifferentiate *in vitro*. However, the mechanism remains obscure. In this study, to identify the transcriptional signatures of dedifferentiated and multipotent cells derived from functional cells, we performed transcriptome analyses of those cells using *in vitro* dedifferentiation culture system of porcine MAs.

2. Materials and methods

2.1. Cell isolation and dedifferentiation culture

Primary porcine MAs were isolated using the methods described by Nobusue and Kano [5]. In brief, fat tissue was minced and digested in 0.1% (w/v) collagenase solution (Collagenase type II; Sigma–Aldrich) at 37 °C for 1 h with gentle agitation [12]. The digested cell suspension was then filtered through 150 and 250 μm nylon meshes (Kyoshin Rikoh), allowing the cells to pass through but retaining unwanted stromal cells and tissue. The floating MAs in the top layer were collected and washed thrice by centrifugation. Isolated MAs were placed in 12.5 cm^2 culture flasks (BD Falcon) filled with Dulbecco's modified Eagle's medium (Nissui Pharmaceutical Co.) with 20% (v/v) fetal bovine serum (Moregate BioTech). The flask was filled with the medium, turned upside down, and incubated in a humidified 5% CO_2 atmosphere. Approximately 1 week later, the cells had firmly attached themselves to the ceiling and developed fibroblast-like shape without any visible fat droplets. The medium was changed every 4 days until the cells had grown to semiconfluence at 14 days.

* Corresponding author. Fax: +81 466 84 3657.
E-mail address: kkano@brs.nihon-u.ac.jp (K. Kano).

0006-291X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2011.03.063

2.2. RNA extraction and genechip microarray hybridization

Total RNA was extracted from differentiated and dedifferentiated porcine MAs using the Trizol reagent (Invitrogen) and RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions. The quality of the extracted RNA was assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies). Three biological replicates were prepared for data reproducibility. GeneChip One-Cycle Target Labeling and Control Reagents (Affymetrix) were used to convert 5 μg of total RNA to biotinylated labeled cRNA. This was then hybridized to the Affymetrix GeneChip Porcine Genome Array (Affymetrix). Fluorescent images were detected using the GeneChip Scanner 3000 (Affymetrix). Expression data and raw expression data (CEL files) were generated using the GeneChip Operating System software (Affymetrix).

2.3. Reannotation of porcine Affymetrix probe sets

Sequence data were obtained from the Affymetrix (Affymetrix Porcine Consensus Sequences in FASTA format) and Ensembl (human cDNA sequences) websites. All sequence analyses were performed using Ensembl (release 61; Jan. 2011) based on the *Homo sapiens* high-coverage assembly (hg19) (Feb. 2009). Each probe set was reannotated using the following method. The porcine sequences as query were compared at an amino acid sequence level with the Ensembl human cDNA sequences as database. We used the tblastx program with a cut-off e-value of significant homology with human sequence set at 1e-10.

2.4. Microarray data analysis

Microarray analyses were initially performed in R (http://www.r-project.org/), using packages from the Bioconductor project (http://www.bioconductor.org/). Raw intensity values were subjected to a preprocessing step using the robust multiarray average algorithm that summarizes and normalizes data into gene expression level [13]. Probe sets were defined as "differentially expressed" during dedifferentiation when the false discovery rate (FDR) was lower than 0.05 using the RankProd package of Bioconductor [14]. TIBCO Spotfire (TIBCO Software) was then used for the identification of differentially expressed probe sets, calculation of correlation coefficients, and hierarchical clustering by Ward's method. Functional analyses were performed using Ingenuity Pathways Analysis (IPA, version 8.8; Ingenuity Systems). Functional analysis identified the biological functions that were most significant to the dataset. IPA provides minimal support for porcine genes; therefore, equivalent human Affymetrix probe set (HG U133-PLUS-2) IDs (see reannotation, above) were imported. Genes from the dataset that met the FDR cut-off value of 0.05 and were associated with biological functions in the Ingenuity Pathways Knowledge Base were considered for analysis. Benjamini–Hochberg multiple testing correction was used to calculate the *p*-value, determining the probability that each biological function assigned to the dataset was by chance alone. To compare the genes regulated in DFAT cells with those in porcine iPS cells, publicly available datasets were obtained from the Gene Expression Omnibus (GEO) database. The microarray data "Induced Pluripotent Stem Cells from the Pig Somatic Cells" are accessible through GEO series accession number GSE15472 [15]. The datasets of iPS cells were used for comparative gene expression analysis. Genes significantly upregulated (FDR < 0.05) in both DFAT and iPS cells compared with MAs were selected and then subjected to functional analysis using IPA.

3. Results

3.1. Comprehensive assessment of differential gene expression during MA dedifferentiation

To identify the transcriptional signatures of dedifferentiated cells, we analyzed gene expression profiles of MA dedifferentiation using the Affymetrix GeneChip Porcine Genome Array, which can detect regulation of 23,256 porcine transcripts. We examined the differential mRNA levels of the genes in each experiment by comparing RNA samples from DFAT cells with those from MAs. Furthermore, three biological replicates were assessed for each dedifferentiation sample. The microarray data discussed in this publication have been deposited in NCBI's GEO database [16] and are accessible through GEO accession numbers GSM432404, GSM432405, GSM432406, GSM432407, GSM432408, and GSM432409 (series GSE17264).

The Affymetrix GeneChip Porcine Genome Array was poorly annotated, with less than approximately 10% of the probe sets in this array being described with gene symbols, thus posing a challenge to biological interpretation of data. Lack of annotation is possibly because of the limited availability of a full-length porcine cDNA sequence. First, we attempted to improve the annotation of this microarray. Each probe set was reannotated using the following method. Using BLAST, the Affymetrix Porcine consensus sequences were compared at an amino acid sequence level with the Ensembl human cDNA sequences as database. Using this method, we putatively identified 17,308 (72.31%) of 23,935 probe sets in the Affymetrix Porcine microarray. To analyze these data in a non-biased manner, correlation coefficients for all genes among each sample were calculated. Correlation values close to 1 were common when comparing the same samples within MAs or DFAT cells, whereas comparing MAs with DFAT cells yielded a correlation of 0.798 (Fig. 1A and Supplementary Table S1). Cluster analysis also demonstrated high similarity within MAs and DFAT cells but not between the two groups. Complete array expression values produced a dendrogram (Fig. 1B) reflecting global expression differences across the samples. Hierarchical clustering of all samples demonstrated a clear distinction on the basis of MA and DFAT cell samples. These results indicate that all microarray data were of high quality and extremely consistent within each cell type and that the transcriptional profiles of DFAT cells and MAs were distinguished. Therefore, the gene expression profiles of DFAT cells were compared with those of MAs. In the following analyses, genes were defined as differentially expressed in dedifferentiation when FDR was lower than 0.05. The analyses yielded 602 and 628 genes reproducibly downregulated and upregulated, respectively, during MA dedifferentiation. A complete list of genes discussed in this paper is provided in Supplementary Table S2.

3.2. Significant decrease in expression levels of typical MA genes during dedifferentiation

To provide further details of the functional relevance in differentially expressed genes, we performed functional analysis using IPA. On the basis of significant statistics, IPA was used to evaluate the biological functions related to up- or downregulated genes involved in MA dedifferentiation. As shown in Fig. 2A, 308 genes significantly downregulated during MA dedifferentiation showed significant enrichment for several biological functions in IPA. Functional classifications and the corresponding number of identified genes (in brackets) included those for "metabolism of lipid" (55), "modification of fatty acid" (29) and "oxidation of lipid" (25). Many of these genes play a major role in metabolism of lipid such as *ADIPOQ* (adiponectin), *LIPE* (lipase, hormone sensitive), *PDK4*

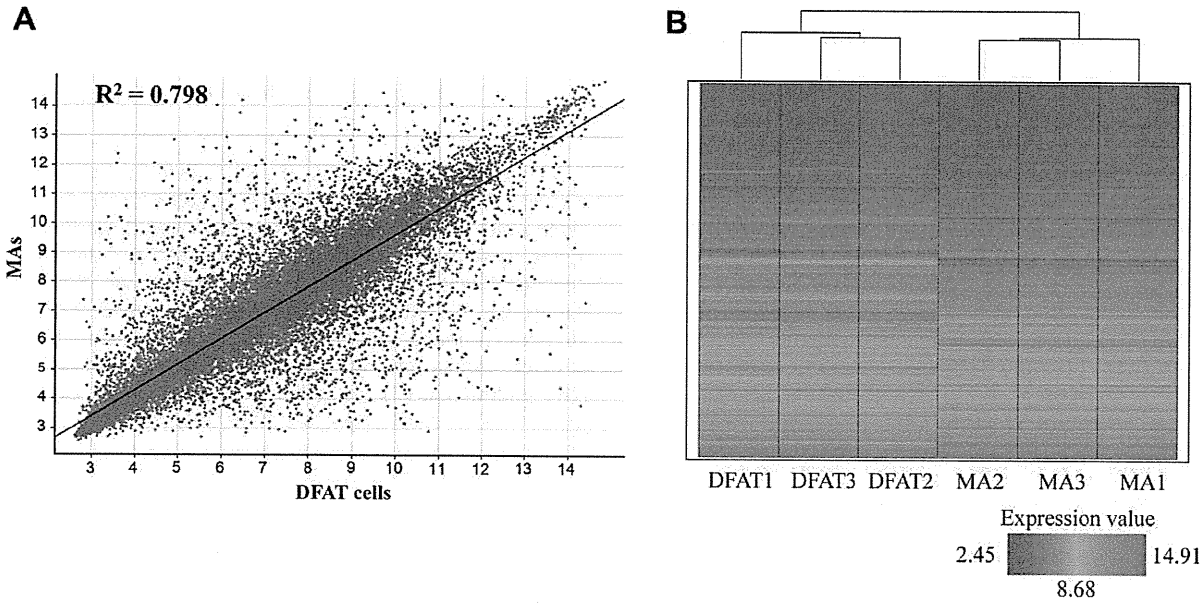

Fig. 1. Correlation coefficients and hierarchical clustering of microarray data during dedifferentiation. (A) Correlation coefficients were calculated among MAs and DFAT cells. The averaged expression signals of three biological replicates are plotted on the *x* and *y* axes. (B) Expression of profiled samples of MAs and DFAT cells was compared using hierarchical clustering by Ward's method.

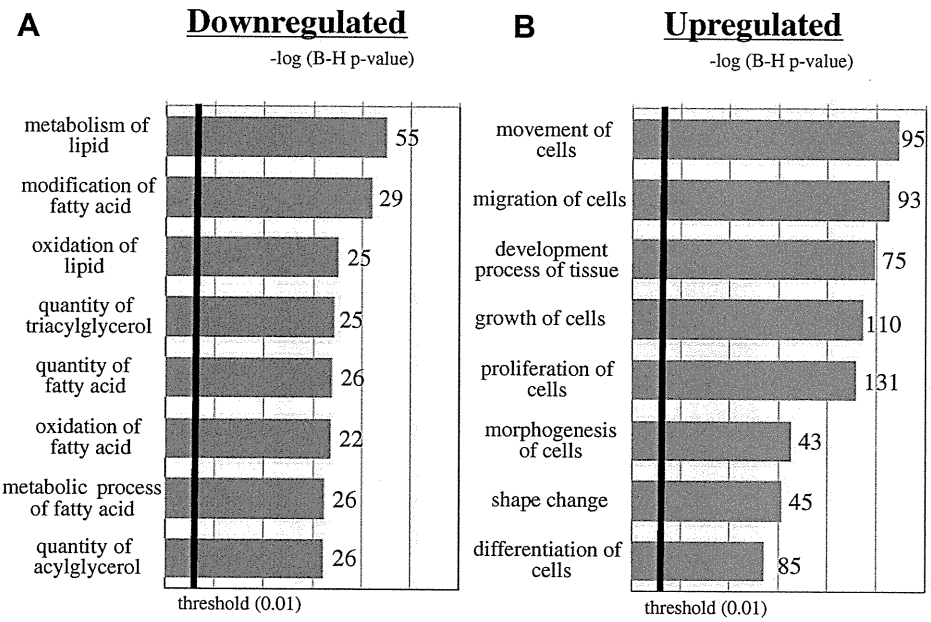

Fig. 2. Functional classification of differentially expressed genes in MAs during dedifferentiation. Using IPA, analyses were performed individually on 308 and 368 genes significantly (FDR < 0.05) down- and upregulated, respectively, during dedifferentiation. All functional categories demonstrated enhanced statistical representation. Bars represent the proportion of genes involved in each category for which statistical significance and number of genes are indicated.

(pyruvate dehydrogenase kinase, isozyme 4), *LPL* (lipoprotein lipase), *FASN* (fatty acid synthase), *PPARG* (peroxisome proliferator-activated receptor γ), and *FABP4* (fatty acid-binding protein 4). Supplementary Table S3A lists all genes involved in lipid metabolism. These findings demonstrate that DFAT cells lose the functional MA phenotype during dedifferentiation.

3.3. Significant upreguration of cell proliferation, cell morphology changes, and tissue specific-related genes during MA dedifferentiation

To gain insight into the dedifferentiation process, we next examined genes upregulated during MA dedifferentiation. As shown in Fig. 2B, 368 genes that were clearly upregulated during dedifferentiation showed significant enrichment for several biological functions, such as "movement of cell" (95), "migration of cells" (93), "developmental process of tissue" (75), "growth of cells" (110), "proliferation of cells" (131), "morphogenesis of cells" (43), "shape change" (41) and "differentiation of cells" (85). A complete list of genes involved in these functions can be found in Supplementary Table S3B–I. Two hundred and twenty genes were associated with at least one of these eight biological functions, of which "movement of cells" (Supplementary Table S3B), "migration of cells" (Supplementary Table S3C), and "proliferation of cells" (Supplementary Table S3F) shared many common genes, indicating

that these three functions are closely related. The representative genes involved in these functions were *SERPINE1* (serpin peptidase inhibitor, clade E, member 1), *VEGFC* (vascular endothelial growth factor-C), *CDH2* (*N*-cadherin), *MDK* (midkine), *TIMP1* (TIMP metallopeptidase inhibitor 1), *IL33* (interleukin-33), *PLAU* (plasminogen activator, urokinase), *ARNT2* (aryl-hydrocarbon receptor nuclear translocator 2), *TNFRSF12A* (tumor necrosis factor receptor superfamily, member 12A), and *IL18* (interleukin-18) (Fig. 3A). Similarly, many genes involved in "growth of cells" (Supplementary Table S3E) were shared with those involved in "morphogenesis of cells" (Supplementary Table S3G) and "shape change" (Supplementary Table S3H). The representative genes involved in these three functions were *BASP1* (brain abundant, membrane-attached signal protein 1), *SPP1* (secreted phosphoprotein 1), *ITGA5* (integrin alpha 5), *FN1* (fibronectin 1), *TGFB1I1* (transforming growth factor beta 1-induced transcript 1), *CD44*, *CAPG* [capping protein (actin filament), gelsolin-like], *SDC1* (syndecan-1), *PALLD* (palladin, cytoskeletal-associated protein), *FHL3* (four and a half LIM domains 3), and *IL1R1* (interleukin-1 receptor, type I) (Fig. 3B). This result suggests that MAs re-enter the cell cycle and gain a fibroblast-like appearance during dedifferentiation. The categories "developmental process of tissue" (Supplementary Table S3D) and "differentiation of cells" (Supplementary Table S3I) included genes involved in several tissue-specific functions and regulation of differentiation, such as *SFRP2* (secreted frizzled-related protein 2), *PRRX1* (paired related homeobox 1), *HEY2* (hairy/enhancer-of-split related with YRPW motif 2), *DLX2* (distal-less homeobox 2), *AEBP1* (AE-binding protein 1), *PEG10* (paternally expressed 10), *PRRX2* (paired related homeobox 2), *RUNX1* (runt-related transcription factor 1), *FZD7* [frizzled homolog 1 (Drosophila)], *IGFBP5* (insulin-like growth factor-binding protein 5), *ID4* (inhibitor of DNA binding 4), and *ID2* (inhibitor of DNA binding 2) (Fig. 3C). This result shows that DFAT cells express several lineage-specific genes and have a multipotent capacity for lineage differentiation.

3.4. Characteristics of DFAT cells determined by expression patterns of CD markers

The expression patterns of cell-surface antigens are used to characterization of cell types [17]. To better understand the characterization of DFAT cells, we identified the CD marker genes from genes significantly upregulated during MA dedifferentiation. We generated heat maps for these genes using normalized expression values (Fig. 3D). Heat maps demonstrated regulation of 23 CD marker genes uniquely induced during dedifferentiation. The CD markers were categorized by cell type, such as myeloid (12), lymphoid (14), endothelial (7), epithelial (7), and stromal cells (9) [17]. A number of CD marker genes significantly upregulated during MA dedifferentiation are associated with hematopoiesis, such as those for lymphoid cells (*CD10*, *CD24*, *CD39*, *CD40*, *CD44*, *CD49e*, *CD56*, *CD71*, *CD86*, *CD87*, *CD109*, *CD138*, *CD225*, and *CD232*) and myeloid cells (*CD10*, *CD13*, *CD40*, *CD44*, *CD49e*, *CD54*, *CD71*, *CD86*, *CD87*, *CD109*, *CD140a*, and *CD232*). This result shows that MAs upregulated several CD marker genes of other cell types, including hematopoietic cells, during dedifferentiation.

3.5. Comparative analysis of DFAT and iPS cells

Comparing the gene expression profiles of DFAT cells with those of iPS cells would be a useful strategy for identifying gene expression signatures of dedifferentiation. To identify common gene expression profiles in dedifferentiation, we compared the highly expressed genes of DFAT cells with those of iPS cells. The data showed that 194 genes were highly expressed in both cell types

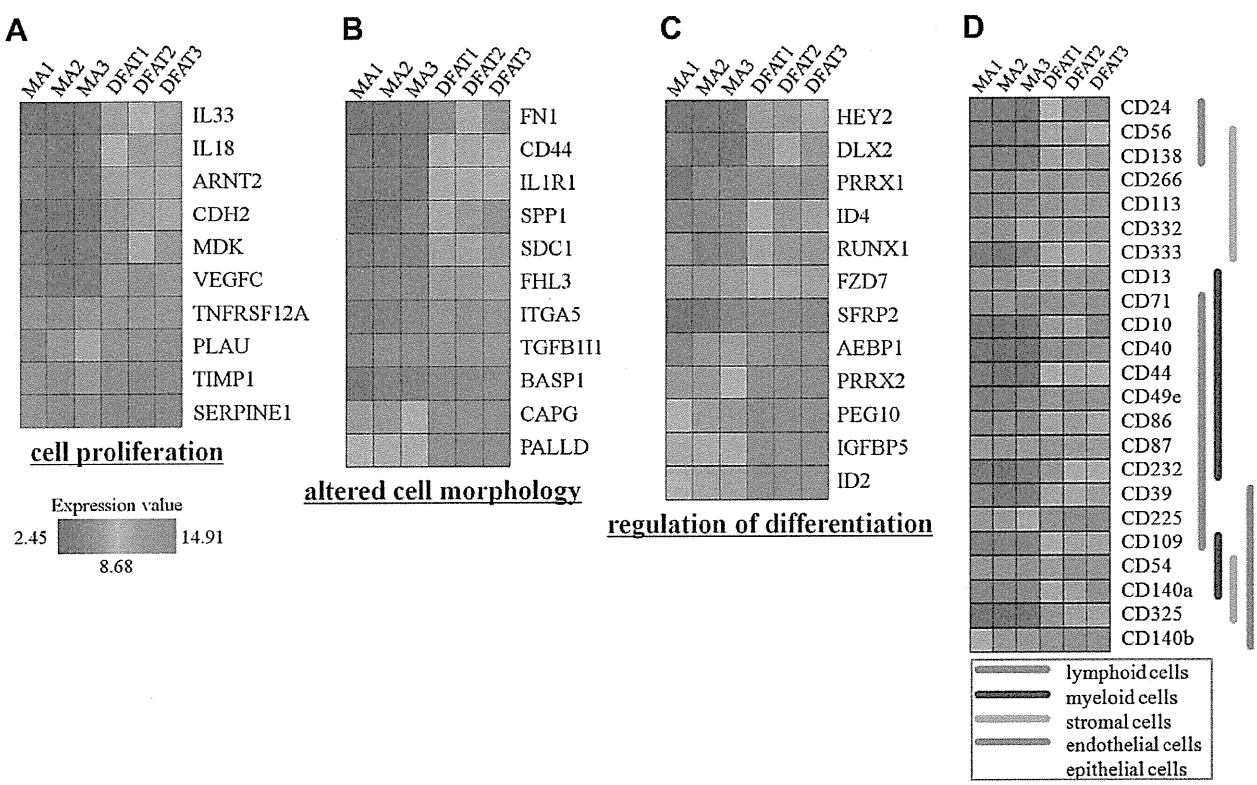

Fig. 3. Heat maps were generated for all genes involved in cell proliferation (A), altered cell morphology (B), regulation of differentiation (C), and CD markers (D) significantly induced during MA dedifferentiation. Columns represent samples and rows represent genes. The color scales were generated from the expression values. Red and blue denote high and low expression, respectively. Orange, purple, blue, green, and yellow bars indicate CD marker genes for lymphoid, myeloid, stromal, endothelial, and epithelial cells, respectively.