Original article

Biological Database of Images and Genomes: tools for community annotations linking image and genomic information

Andrew T Oberlin[1], Dominika A Jurkovic[2], Mitchell F Balish[2] and Iddo Friedberg[1,2,*]

[1]Department of Computer Science and Software Engineering, Miami University, 700 E. High St., Oxford, OH 45056, USA and
[2]Department of Microbiology, Miami University, 700 E. High St., Oxford, OH 45056, USA

*Corresponding author: Tel: +1(513) 529 3182; Fax: +1 (513) 529-2431; Email: i.friedberg@miamioh.edu

---

Genomic data and biomedical imaging data are undergoing exponential growth. However, our understanding of the phenotype–genotype connection linking the two types of data is lagging behind. While there are many types of software that enable the manipulation and analysis of image data and genomic data as separate entities, there is no framework established for linking the two. We present a generic set of software tools, BioDIG, that allows linking of image data to genomic data. BioDIG tools can be applied to a wide range of research problems that require linking images to genomes. BioDIG features the following: rapid construction of web-based workbenches, community-based annotation, user management and web services. By using BioDIG to create websites, researchers and curators can rapidly annotate a large number of images with genomic information. Here we present the BioDIG software tools that include an image module, a genome module and a user management module. We also introduce a BioDIG-based website, MyDIG, which is being used to annotate images of mycoplasmas.

Database URL: BioDIG website: http://biodig.org
BioDIG source code repository: http://github.com/FriedbergLab/BioDIG
The MyDIG database: http://mydig.biodig.org/

Introduction

The large amount of genomic data has led to a concurrent deluge of new and creative ways of genomic data processing and visualization. These are manifest in the many varieties of genome browsers (1–11) and in the large number of genome annotation tools both manual and computational [See reviews: (12, 13)].

However, genomic data are not the only biological data that require high-throughput analysis and annotation. Since long before the genomic revolution, biological image data have been produced steadily from sources as diverse as drawings, medical radiology (magnetic resonance imaging, radiography), microscopy and regular photography. Like genomic data, biological images also require analysis and annotation that include feature detection, segmentation and association with biological function. Computer-aided and computational image analysis has been performed since computers emerged in the 1940s, with major advances made as the acquisition, storage, analysis and manipulation of images have become cheaper and more accessible.

With the deluge of both genomic and biological image data, there is a need for infrastructure capable of handling both types of data and at the same time associating them

---

in a meaningful fashion. There are many projects that catalog biological images in fields as diverse as zoology, microbiology, neurobiology, developmental biology and cancer research, to name a few. However, to date there has been no freely available software to enable the association of image features with genomic features. With the large number of both image and genomic data, and with the understanding that phenotypic data manifest in images is ultimately a result of genomic features, there is a need to easily associate image data and genomic data. While databases providing such information do exist (14, 15), there are no general-purpose software tools that can be used to construct such an association.

Here we present the Biological Database of Images and Genomes or BioDIG. BioDIG is a set of tools to construct databases with genomic and image data and place them in semantic association. BioDIG provides the following features: a genome uploader and browser, an image uploader, browser and tagging workbench, association of image features with genomic features, user management that features circles of trust for image feature tagging and web services. We demonstrate an example website, MyDIG, which we have constructed using BioDIG. BioDIG is released under an open source license, and is freely available. We call on the community of bioinformatics programmers to join in its development.

Materials and methods

General

BioDIG is a web-based workbench that enables users to upload images, annotate parts of them with semantic data and associate these tags with genomic features. BioDIG includes user management with increasing circles of privileges. BioDIG is implemented using three modules: the genome module, the image module and the users module. The genome module is used to upload, store and visualize genomes. The image module is used to upload, store, annotate and visualize images. The user module is used to manage the users and their privileges in the system.

Genome module

The genome module implements the storage of genomic information that should be readily accessible for both queries and visualization. As a foundation, we use the Generic Model Organism Database, GMOD, database schema, Chado, which is implemented in PostgreSQL. This database was designed to represent genomic information provided as Generic Feature Format or GFF files. BioDIG is implemented as an extension of Chado, adding links from the genes in the feature table of Chado to the images in the image database extension. The features table of Chado is used to store data about genes, gene components, mRNA

and other available data in the GFF format. BioDIG is implemented using PostgreSQL; however, it uses a Django framework (djangoproject.com) for web service and web client access to support community annotation. For visualization of the genomic data, GMOD also offers a genome visualization tool, GBrowse (10), which is a Perl-based system that can link to a multitude of genomic databases including Chado. However, when connecting to Chado, GBrowse runs much more slowly than when connecting to other databases. Thus, we have chosen to duplicate the data in SQLite3, which will be used for the rendering of GBrowse. These files break up the data into separate databases so that each genome can be separated, decreasing search time. The SQLite3 database that is used in conjunction with GBrowse visualization is Bio::DB::Seq Feature::Store, which stores the features as flat Perl objects, which eliminate all of the table joins that Chado must go through to prepare the information for visualization. Normally, one would be concerned with the maintainability of a system that duplicates information. However, the genomic information in these databases is expected to change only rarely, leaving little risk that these sources will become unsynchronized. We expect that in this database schema, one would perform these information updates. For instance, when a gene's function becomes known for an organism that is already in the database, a batch procedure will be performed, making minimal changes to the database without breaking any foreign keys. Using this genome module, we are able to visualize and store information about genes so that queries are optimized and the visualization workload is not burdening the search system.

Image module

The image module's goal is to handle the creation and storage of images and image metadata. Here 'image metadata' is taken to mean identifiable components of biological interest within the images. Organelles, cells, tissues, organisms and other biological objects in the images are all considered identifiable components. The image module supports the classification of these components. We have implemented a web user interface that allows the tagging of images by drawing on the image and describing the substructure using free text and/or an ontology. Finally, once components have been identified, they can be linked to the associated genes via 'gene links' mapping the phenotypic data in the images to the genomic data (Figure 1). The implementation of gene links is explained below.

The image module's most important task deals with storing the image files in a fashion that makes them highly accessible. First, the picture table stores the path to the image on the file system. There are two options often used for image storage. One is to store the image in the database as a Binary Large Object (BLOB), another is to store a path to an image file, either in a media directory or on an external

media server. On the one hand, storing an image as a BLOB in a database has certain advantages over storing it as a file, such as lower read throughput for objects <1 MB in size for short intervals and atomicity for overwriting. On the other hand, storing an image as a file provides other advantages such as high throughput for files >1 MB and low fragmentation in the long term (16). We have decided to use files for storing images in the BioDIG system to minimize the load on the database. Because the system was designed with a web client through web services in mind, we expect that images shall often be downloaded straight from the server by scripts to manipulate the image data. The services and server structure also needed to support the addition of future automated systems such as image analysis and tag suggestions from users. These scripts would create an extra strain on an SQL database, which is supposed to be focused on the web client. It is often important to consider where the possible stress on a system could occur, and in the BioDIG system it could be from the web users, programmers and even the server itself.

Because we use the Django framework, Python objects are linked to tables in the database. When one retrieves one of these objects, all the data are made available to the user. Thus, storing the path can decrease search times that may only use the picture table as a foreign key to link information. To increase throughput even more, the images are kept in consistent format, meaning that all pictures are encoded in PNG format to remain web compliant. The names of these images are also changed to a unique key, which can be used for hashing. Once the database grows in size, these hashed names will be used to limit the number of images per directory by taking two-letter prefixes as the names of the directory in which the image resides, a method suggested by (16) for increasing server performance.

The rest of the schema for the images represents other image metadata such as the image description, upload date, the user who owns the image and the image tags. The tags themselves are represented by a combination of three tables, tag points, tags and tag groups, as an extension to Chado. The tag points are an ordered list of points that when drawn in an Application Programming Interface (API), form a polygon. These points are held together by the tags, which hold the metadata such as the structure related to the tag in the image. We also store information about tag groups for users to make their own tag groups. These groups will be unique to a user until that user wishes to merge with another user's tag group on the same image. When image tags are linked to the genes in the feature table of Chado via gene links, data querying is enabled (Figure 2).
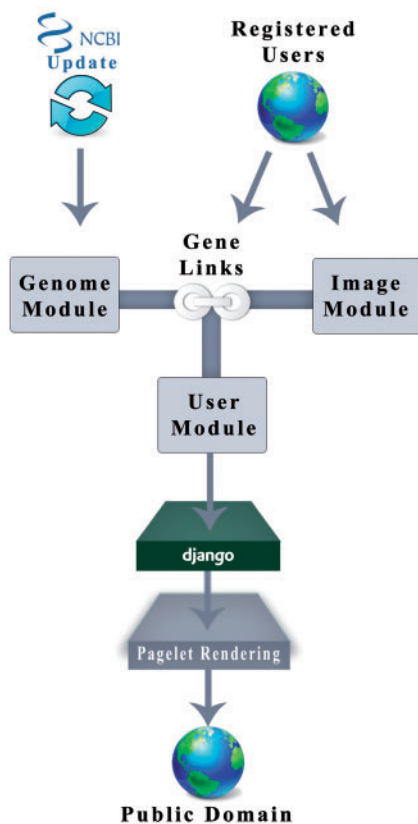
Figure 1. BioDIG Schema. BioDIG comprises three modules: image, genome and user management. Data in the genome module are deposited initially by registered users, but gets updated from NCBI. Level 2 and 3 users can upload images, and update image tags and image links to genomic data.
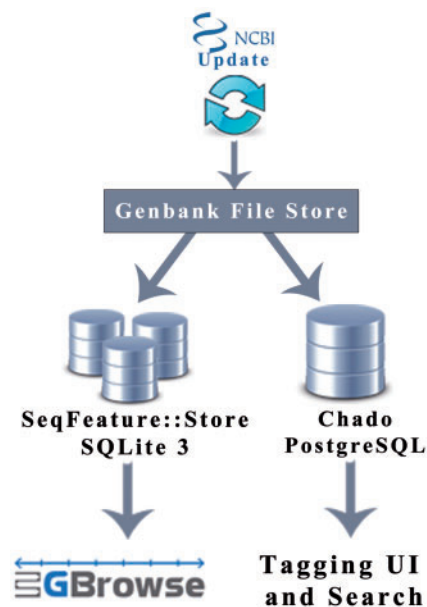

Figure 2. Genome module. The genome module consists of the GMOD PostGreSQL database, Chado, which is used to house genome information. SQLite3 is used to interface with the GBrowse genome database, owing to its speed. The duplication and maintenance of data integrity are not costly. See text for details.

User management module

BioDIG enables collaborative editing of image metadata, with the aim of improving image annotations through rapid collaborative work by a large number of users. At the same time, there is a need to maintain high annotation quality and prevent vandalism. To achieve these goals, BioDIG features a user module that provides three user permission levels, with each level including the permissions of the preceding levels.

Level 1 users have the most basic set of permissions, being able to search published information on the website and download the results of their search, including genomic, image, image tag and gene link information, for further analysis. The majority of users are intended to have Level 1 read-only access to a BioDIG-based site. Level 2 users, or annotators, have site-wide read privileges, and in addition, can employ tagging user interfaces and pipelines for publishing. Level 2 users constitute a much smaller number, and assigning a user to Level 2 is supposed to be based on a circle of trust. Level 3 users, or administrators, have permissions to access the databases and edit information directly. These users will also be in charge of monitoring requests for publication and merge requests for tag groups as the result of the pipeline. Figure 3 illustrates the user permissions scheme.

In this community-annotation effort, crediting users with the publication of images can connect labs performing similar research and allow the user to edit these photos in their own workbench. Many websites attempt to have private and public layers; however, this implementation of the process is aimed to allow for approved and unapproved information to maximize community engagement. Once a tag is sent through the pipeline, it
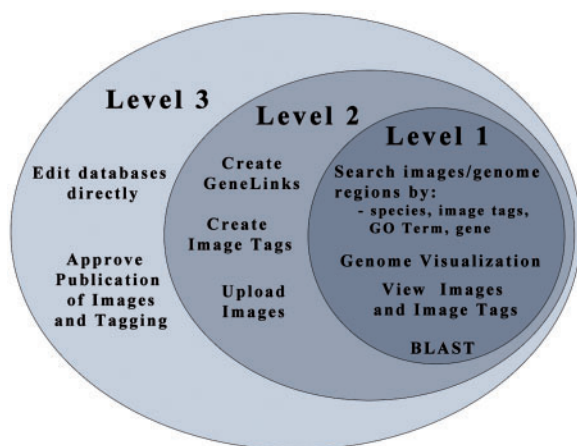
Figure 3. User management module. User management is done by a circle of increasing privileges. Most users are Read Only, Level 1. Image and genome curators are Level 2 trusted users. Level 3 users can add and revoke user privileges, in additions to all Level 1 and 2 permissions.

is available for all users in to see publicly. However, if the administrators of BioDIG approve the tag, it is said to be of a trustworthy source having no obvious errors. This idea is implemented to increase the amount of quality information in the database, while not limiting the capabilities of the system. Inclusion of the user who owns the image also allows the person to be involved in the tagging of the picture by other people; the approval process of the tags and associated gene links is done by trusted users only, helping maintain quality. With the administrators monitoring this submission process, we will be able to avoid deadlocks and make the data more accessible.

Web services and data analysis

Web services provide an opportunity for researchers to mine data from BioDIG-enabled websites. Most of the search features of the website itself are, or soon will be, available via web services. This will allow for a command line interface to be implemented along with a Representational State Transfer (REST) API in common programming languages for interacting with these services programmatically. For instance, if someone wishes to retrieve a phenotype caused by the expression of several orthologs in different species, a client API could be used to query for tags related to a set of genes across a list of species, and retrieve the relevant images and image tags. The client could then access the information for each of these tags through a predetermined object-oriented interface, allowing for more dynamic searches than one would naturally make by visiting the website. The user could then run metrics to analyze the image and the tag data in a program rather than just by hand to increase the speed at which this data can be analyzed.

For example, to return information about image 52, a caller program can use the URL: http://wan1.mbi.muohio.edu/dome/images/getImageMetadata?imageKey=52 which will return a JSON object:

```
{
    "description": ''Mycoplasma pneumoniae",
    "organisms":   [{"common_name":   ''Mycoplasma
  pneumoniae M129",
    "organism_id": 7}],
    "url":   ''/dome/media/pictures/85c4bc19fa0d481460
  bcb522349fba27.png",
    "errorMessage": ''Success",
    "uploadDate": ''2012-11-24 19:14:08",
    "error": false,
    "uploadedBy": ''oberliat"
}
```

Documentation for the REST API is provided automatically via Swagger (http://swagger.wordnik.com/).

Example site: the MyDIG website

We describe MyDIG, a BioDIG-derived website implemented to show mycoplasma images and genomes. Mycoplasmas are a clade of bacteria whose main characteristics are parasitism, the lack of a cell wall and a small rapidly evolving genome (17). The cellular morphology of mycoplasmas is especially interesting, as the limited complexity of mycoplasmal cellular structure, with small, wall-less, single membrane-bound cells, enables researchers to focus on universal principles of cellular organization (18), providing a platform for understanding more derived types of cells.

For these reasons, we created the Mycoplasma Database of Image and Genomes or MyDIG using the BioDIG infrastructure. Phylogenetically associated cellular morphology is cardinal in the description of mycoplasmas. Morphological features in mycoplasmas are often associated with known genomic features. We present a series of screenshots showing a typical use-case scenario for the image-genomic association for MyDIG. *Mycoplasma pneumoniae* is a significant human pathogen whose cell shape is distinctly polarized (19). From the cell body, which contains most of the cell volume, emanate both the attachment organelle and the trailing filament (18). They are both prosthecal structures, continuous with the cytoplasm of the cell body. Although the trailing filament is not well understood, the attachment organelle of *M. pneumoniae* and its close relatives is essential for adherence to host cells (20), which is necessary for all growth in a non-laboratory environment, and also contains a poorly characterized motor that contributes to cell motility and cell division (21–23). The composition and assembly of *M. pneumoniae*'s attachment organelle are well studied, having been established through analysis of non-adherent mutants (24). Essential to the formation of the attachment organelle is a cytoskeletal structure called the electron-dense core (25), which has been imaged extensively and carefully (26, 27); despite this, the specific locations of its protein components are only vaguely defined. In Figure 4, the general locations of these proteins as well as the transmembrane proteins involved in adherence are indicated (28–30) on a scanning electron micrograph of an *M. pneumoniae* cell grown attached to a glass cover slip, as previously described in (31) and imaged at high magnification.

The need for an image-genomic database is illustrated by the case of *Mycoplasma microti*, a mycoplasma associated with prairie voles. The original publication describing this species includes a transmission electron micrograph of sections through several *M. microti* cells, which reveal the variety of polarization idiosyncratic to the *Mycoplasma muris* phylogenetic cluster, of which *M. microti* is a member. At the time of its publication (1995), images showing similar structures of other members of this cluster had

been published (32, 33). However, the authors of the *M. microti* study, perhaps unaware of these images, described the morphology in terms that emphasize pleomorphy and imply the absence of polarization (34, 35). This flawed description led to *M. microti* being described subsequently as non-motile, as absence of polarity is associated with absence of motility (36). However, we recently determined that like other species of the *M. muris* cluster that exhibit polarization, *M. microti* exhibits robust gliding motility (37) (S.L. Distelhorst and M.F. Balish, unpublished data). Had a database of mycoplasma morphological features been available to the authors, this error could have been prevented.

Conclusions

We have described a new software package for constructing websites linking biological image and genomic data. When designing BioDIG, we did so with generic implementation in mind, so that anyone who associates genomic data with image data may benefit from its adoption. BioDIG is suitable for uses as diverse as cancer pathology, zoology, plant biology, cell biology and microbiology, to name a few. The web services feature is especially useful for linking to other web resources that can capitalize on BioDIG-enabled sites.

Being a collection of modules, rather than a single application, BioDIG can be applied to a wide range of uses. Those can range from microbiologists researching certain taxa (as the mollicute-oriented MyDIG is intended, see example above) to zoologists or plant scientists interested in whole-organism morphologic features and their genomic associations. The biomedical research community, both applied and basic, can also benefit from BioDIG. Image databases of cellular and tissue morphology, radiological images and pathology can all benefit from incorporating BioDIG if links to genomic information are useful. BioDIG-based websites can also be used as a teaching resource, in application ranging from primary secondary to post-secondary and higher education. We call on the community of biology software developers to join us in the development of BioDIG, which is distributed under an open source license. We encourage life science researchers to contact us regarding applications of BioDIG suitable for their needs.

Image feature tagging currently uses freeform drawing, and can be tagged using free text. However, for certain types of applications, it may be better to provide the annotator with a controlled vocabulary or an ontology to make annotations standard and computationally tractable. In the future, we intend to provide a mechanism for doing so. Other features that require implementation are searches by taxon, image tags, gene names, gene annotations and boolean combinations of all of the above.

Figure 4. Case study: *M. pneumoniae*. *M. pneumoniae* image tagged with locations of attachment organelle gene products. *M. pneumoniae* strain 19294 was passaged 10 times and grown and processed for imaging by scanning electron microscopy essentially as previously described (31), but at higher magnification that is well suited for morphological comparison across species. (**A**) (Blue) image highlighting components of the proximal portion of the terminal organelle, consisting of the base of the electron-dense core. (**B**) (Green) image highlighting components of the long portion of the terminal organelle, consisting of the shaft, within which is the rod portion of the electron-dense core. Gene designations are keyed to the GenBank annotation of *M. pneumoniae* strain M129.

References

1. Fiume,M., Smith,E.J., Brook,A. *et al*. (2012) Savant Genome Browser 2: visualization and analysis for population-scale genomics. *Nucleic Acids Res.*, **40**, W615–W6621.

2. Kong,L., Wang,J., Zhao,S. *et al*. (2012) ABrowse–a customizable next-generation genome browser framework. *BMC Bioinformatics*, **13**, 2.

3. Abeel,T., Van Parys,T., Saeys,Y. *et al*. (2012) GenomeView: a next-generation genome browser. *Nucleic Acids Res.*, **40**, e12.

4. Wang,J., Kong,L., Zhao,S. *et al*. (2011) Rice-Map: a new-generation rice genome browser. *BMC Genomics*, **12**, 165.

5. Kawahara,Y., Sakate,R., Matsuya,A. *et al*. (2009) G-compass: a web-based comparative genome browser between human and other vertebrate genomes. *Bioinformatics*, **25**, 3321–3322.

6. Friedel,M., Nikolajewa,S., Suhnel,J. and Wilhelm,T. (2009) DiProGB: the dinucleotide properties genome browser. *Bioinformatics*, **25**, 2603–2604.

7. Skinner,M.E., Uzilov,A.V., Stein,L.D. *et al*. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.

8. Loveland,J. (2005) VEGA, the genome browser with a difference. *Brief. Bioinform.*, **6**, 189–193.

9. Stalker,J., Gibbins,B., Meidl,P. *et al*. (2004) The Ensembl Web site: mechanics of a genome browser. *Genome Res.*, **14**, 951–955.

10. Stein,L.D., Mungall,C., Shu,S. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.

11. Kent,W.J., Sugnet,C.W., Furey,T.S. *et al*. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

12. Friedberg,I. (2006) Automated protein function prediction–the genomic challenge. *Brief. Bioinform.*, **7**, 225–242.

13. Rentzsch,R. and Orengo,C.A. (2009) Protein function prediction the power of multiplicity. *Trends Biotechnol.*, **27**, 210–219.

14. Blake,J.A., Richardson,J.E., Davisson,M.T. and Eppig,J.T. (1997) The Mouse Genome Database (MGD). A comprehensive public resource of genetic, phenotypic and genomic data. The Mouse Genome Informatics Group. *Nucleic Acids Res.*, **25**, 85–91.

15. Shen,E.H., Overly,C.C. and Jones,A.R. (2012) The Allen Human Brain Atlas: comprehensive gene expression mapping of the human brain. *Trends Neurosci.*, **35**, 711–714.

16. Sears,R., Van Ingen,C. and Gray,J. (2007) To BLOB or Not To BLOB: Large Object Storage in a Database or a Filesystem? http://arxiv.org/abs/cs/0701168.

17. Rottem,S. (2003) Interaction of mycoplasmas with host cells. *Physiol. Rev.*, **83**, 417–432.

18. Balish,M.F. (2006) Subcellular structures of mycoplasmas. *Front. Biosci.*, **11**, 2017–2027.

19. Atkinson,T.P., Balish,M.F. and Waites,K.B. (2008) Epidemiology, clinical manifestations, pathogenesis and laboratory detection of Mycoplasma pneumoniae infections. *FEMS Microbiol. Rev.*, **32**, 956–973.

20. Krause,D.C. and Balish,M.F. (2001) Structure, function, and assembly of the terminal organelle of Mycoplasma pneumoniae. *FEMS Microbiol. Lett.*, **198**, 1–7.

21. Radestock,U. and Bredt,W. (1977) Motility of Mycoplasma pneumoniae. *J. Bacteriol.*, **129**, 1495–1501.

22. Hasselbring,B.M. and Krause,D.C. (2007) Cytoskeletal protein P41 is required to anchor the terminal organelle of the wall-less prokaryote Mycoplasma pneumoniae. *Mol. Microbiol.*, **63**, 44–53.

23. Lluch-Senar,M., Querol,E. and Pinol,J. (2010) Cell division in a minimal bacterium in the absence of ftsZ. *Mol. Microbiol.*, **78**, 278–289.

24. Krause,D.C. and Balish,M.F. (2004) Cellular engineering in a minimal microbe: structure and assembly of the terminal organelle of Mycoplasma pneumoniae. *Mol. Microbiol.*, **51**, 917–924.

25. Biberfeld,G. and Biberfeld,P. (1970) Ultrastructural features of Mycoplasma pneumoniae. *J. Bacteriol.*, **102**, 855–861.

26. Henderson,G.P. and Jensen,G.J. (2006) Three-dimensional structure of Mycoplasma pneumoniae's attachment organelle and a model for its role in gliding motility. *Mol. Microbiol.*, **60**, 376–385.

27. Seybert,A., Herrmann,R. and Frangakis,A.S. (2006) Structural analysis of Mycoplasma pneumoniae by cryo-electron tomography. *J. Struct. Biol.*, **156**, 342–354.

28. Seto,S. and Miyata,M. (2003) Attachment organelle formation represented by localization of cytadherence proteins and formation of the electron-dense core in wild-type and mutant strains of Mycoplasma pneumoniae. *J. Bacteriol.*, **185**, 1082–1091.

29. Kenri,T., Seto,S., Horino,A. *et al*. (2004) Use of fluorescent-protein tagging to determine the subcellular localization of mycoplasma pneumoniae proteins encoded by the cytadherence regulatory locus. *J. Bacteriol.*, **186**, 6944–6955.

30. Cloward,J.M. and Krause,D.C. (2009) Mycoplasma pneumoniae J-domain protein required for terminal organelle function. *Mol. Microbiol.*, **71**, 1296–1307.

31. Hatchel,J.M., Balish,R.S., Duley,M.L. and Balish,M.F. (2006) Ultrastructure and gliding motility of Mycoplasma amphoriforme, a possible human respiratory pathogen. *Microbiology*, **152**(Pt 7), 2181–2189.

32. Lo,S.C., Hayes,M.M., Wang,R.Y. *et al*. (1991) Newly discovered mycoplasma isolated from patients infected with HIV. *Lancet*, **338**, 1415–1418.

33. Mirsalimi,S.M., Rosendal,S. and Julian,R.J. (1989) Colonization of the intestine of turkey embryos exposed to Mycoplasma iowae. *Avian Dis.*, **33**, 310–315.

34. Dillehay,D.L., Sander,M., Talkington,D.F. *et al*. (1995) Isolation of mycoplasmas from prairie voles (Microtus ochrogaster). *Lab. Anim. Sci.*, **45**, 631–634.

35. Brown,D.R., Talkington,D.F., Thacker,W.L. *et al*. (2001) Mycoplasma microti sp. nov., isolated from the respiratory tract of prairie voles (Microtus ochrogaster). *Int. J. Syst. Evol. Microbiol.*, **51**(Pt 2), 409–412.

36. Brown,D.R., May,M., Bradbury,J.M. *et al*. (2011) Genus I. Mycoplasma Nowak 1929, 1349 nom. cons. Jud. Comm. Opin. 22, 1958, 166. In: *Bergey's Manual of Systematic Bacteriology,* 2nd edn. Springer (New York).

37. Jurkovic,D.A., Newman,J.T. and Balish,M.F. (2012) Conserved terminal organelle morphology and function in Mycoplasma penetrans and Mycoplasma iowae. *J. Bacteriol.*, **194**, 2877–2883.