Peter F. Lemkin

Image Processing Section, NCI-FCRDC/NIH, Frederick, MD, USA

The 2DWG meta-database of two-dimensional electrophoretic gel images on the Internet

The 2DWG meta-database is a searchable database of two-dimensional (2-D) electrophoretic gel images found on the Internet. A meta-database contains information about locating data in other databases – but not that data itself. This database was constructed because of a need for an enriched set of World Wide Web (WWW) locations (URLs) of 2-D gel images on the Internet. These gel images are used in conjunction with the National Cancer Institute (NCI) Flicker Server to manipulate and visually compare 2-D gel images across the Internet. User's gels may also be compared with those in the database. The 2DWG is organized as a spreadsheet table with each gel image being represented by a row sorted by tissue type. Data for each gel includes tissue type, species, cell-line, image URL, database URL, gel protocol, organization URL, image properties, map URL if it exists, *etc.* The 2DWG may be searched to find relevant subsets of gels. Searching is done using the dbEngine – a WWW database search engine which accesses selected rows of gels from the full 2DWG table. The 2DWG meta-database is accessible on the WWW at http://www-lecb.ncifcrf.gov/2dwgDB/ and the NCI Flicker server at http://www-lecb.ncifcrf. gov/flicker/

1 Introduction

A variety of images play key supporting roles in biomedicine both in clinical and basic research (*cf.* Table 1). In this paper we concentrate on 2-D PAGE protein gel images and, in particular, the role of a meta-database for keeping track of these images. However, as we describe this 2DWG meta-database, keep in mind the utility of the meta-database concept for other biomedical domains. We will revisit this generalization of meta-databases for the Internet in Section 4. Scientists around the world often work on 2-D gel data with gels run on similar samples and sometimes with similar apparatus. Traditionally, spot maps (labeled images) that identify proteins and, sometimes, their post-translational modifications were published in the journal literature [1–6]. In the last few years many of these databases were put onto the World Wide Web (WWW), thereby providing wider access, with the first WWW database being SWISS-2DPAGE [7, 8] and many others following. A partial list of samples includes plasma [9–11], cerebrospinal fluid (CSF) [11, 12], red blood cells (RBCs) [9, 13], platelets [12], liver [14, 15], yeast [16–20], *E.coli* [21, 22], breast [22], heart [23, 24], *Drosophila melanogaster* [25], mouse [26], rat [18], and keratinocytes [27]. In addition to spot identification, some of these databases also have quantitative data of the identified proteins as a function of disease state, stimulation, and inactivation, and should thus help increase our understanding of normal and abnormal disease states [22].

Many of these 2-D gel WWW databases (see Table 2) provide protein spot maps as well as related data. The 2DWG is a catalog of 2-D gel images that exist on the Internet and has the goal of providing a convenient high-quality list of available data. Where feasible, visually comparing sample 2-D gel images against these database gel images and looking up spots in protein maps may suggest putative protein spot identifications. This may then suggest experiments to be run to verify these identifications for a researcher's own gels. The experiments may be as simple as running the gel with monoclonal antibodies for the putative proteins rather than having to resort to more expensive and time-consuming experiments such as sequencing or using mass-spectrometry methods.

Figure 1 shows corresponding regions of plasma protein gels from different laboratories – one run with IPGs and the other with carrier ampholytes. Even though the gels were run under different conditions, it is still useful to compare them since many proteins can be identified

Table 1. Images used in support of biomedical research

1	Gel electrophoresis 1-D and 2-D protein, RNA and DNA materials
2	HPLC spectra, mass-spectrometry
3	Capillary electrophoresis, chromatography
4	Serial section images produced by various microtome methods
5	Projections of reconstructed 3-D views: Visual Human anatomy
6	2-D projections of 3-D molecular models
7	Medical X-rays for comparing bone growth or tumor progression
8	MRI or PET imaging
9	MRM (MRI microscopy)
10	Protein and nucleic acid sequence logos
11	RNA structure comparison dot matrices
12	2-D DNA gels using restriction enzymes
13	Spectra, time series or graphs (of anything)
14	Problem domains which produce misaligned or distorted images
15	Problem domains which lose alignment during data acquisition
16	Groupware consultations on the same set of images

Correspondence: Dr. P. F. Lemkin, Image Processing Section/LECB, Bld 469 Room 150, NCI-FCRDC/NIH, Frederick, MD 21702, USA (E-mail: lemkin@ncifcrf.gov)

Nonstandard abbreviations: CGI, common gateway interface; **CSF,** cerebrospinal fluid; **2DWG,** 2-D world gel (database); **FTP,** file transfer protocol; **GIF,** graphics interchange format; **HTML,** hypertext markup language; **JPEG,** joint photographic experts group; **NCI,** National Cancer Institute; **RBC,** red blood cell; **TIFF,** tagged image file format; **URL,** universal resource locator; **WWW,** World Wide Web

Keywords: World Wide Web / Internet / Two-dimensional polyacrylamide gel electrophoresis / Databases / Meta-database / Proteins / Genetics / Image analysis

0173-0835/97/1515-2759 $17.50+.50/0

Table 2. Partial list of WWW 2-D electrophoretic gel databases[a)]

Material	WWW location (URL)	Organization
Liver, plasma, HepG2, RBC, lymphoma, CSF, macrophage-CL, erythroleukemia-CL, platelet, yeast, *E.coli*	http://www.expasy.ch/	ExPASy SWISS-2DPAGE
Mouse liver, human breast cell lines, pyrococcus	http://www.anl.gov/CMB/PMG/	Argonne Protein Mapping Group HSC-2DPAGE,
Human, rat and mouse heart	http://www.harefield.nthames.nhs.uk/	Heart Science Centre, Harefield
Human heart	http://www.chemie.fu-berlin.de/user/pleiss/	HEART-2DPAGE, German Heart Institute Berlin
Human heart	http://www.mdc-berlin.de/~emu/heart/	HP-2DPAGE, MDC, Berlin
Rat neuronal	http://sunspot.bioc.cam.ac.uk/NEURON.html	Cambridge 2D PAGE
Embryonal stem cells	http://www.ed.ac.uk/~nh/2DPAGE.html	Immunobiology, University of Edinburgh
Human colon carcinoma	http://www.ludwig.edu.au/www/jpsl/jpslhome.html	Joint Protein Structure Lab
Human: primary keratinocytes, epithelial, hematopoietic, mesenchymal, hematopoietic, tumors, urothelium, amniotic fluid, serum, urine, proteasomes, ribosomes, phosphorylations Mouse: epithelial, new born (ear, heart, liver, lung)	http://biobase.dk/cgi-bin/celis/	Danish Centre for Human Genome Research
E.coli	ftp://ncbi.nlm.nih.gov/repository/ECO2DBASE/ (new) http://pcsf.brcf.med.umich.edu/eco2dbase/	ECO2DBASE (in NCBI repository)
Rat, mouse, human liver, corn, wheat	http://www.lsbc.com/	Large Scale Biology Corp
Maize	http://moulon.moulon.inra.fr/imgd/	Maize Genome Database, INRA
Drosophila melanogaster	http://tyr.cmb.ki.se/	Karolinska Institute
A375 melanoma cell line	http://rafael.ucsf.edu/2DPAGEhome.html	UCSF 2D PAGE
Bacillus subtilis	http://pc13mi.biologie.uni-greifswald.de/	University of Greifswald Protein Disease
Plasma, CSF, urine	http://www-lecb.ncifcrf.gov/PDD	Database (PDD)
Prostate, phosphoprotein, breast cancer drug screen, FAS (plasma), Cd toxicity (urine), leukemia	http://www-lecb.ncifcrf.gov/ips-databases.html	IPS/LECB, NCI/FCRDC
Yeast	http://www.ibgc.u-bordeaux2.fr/YPM	Yeast 2D gel DB, Bordeaux
Yeast	http://yeast-2dpage.gmm.gu.se/	Yeast 2D-PAGE Göteborg
Yeast	http:/www.proteome.com/YPDhome.html	Proteome Inc (YPD – Yeast Protein DB)
Yeast, REF52, mouse embryo	http://siva.cshl.org/index.html	Quest Protein Database Center

a) These databases contain 2-D gel images for a variety of tissues as well as 2-D protein gel maps identifying some of their proteins. The user should investigate them individually since the relative URL paths for 2-D gel image files and maps differ.

visually in most regions of the gels. Using the Flicker Comparison WWW program [28] (*cf.* Section 6.1), further confidence in this visual identification can be achieved.

1.1 Image data on the Internet

The current generation of remote image communication and collaboration methods included postal mail, fax, E-mail, File Transfer Protocol (FTP), and WWW retrieval of static images. These are all passive methods with no opportunity for an investigator to manipulate materials dynamically. The new generation of Internet tools is beginning to provide active methods of collaborative computing with images. Images may now be manipulated to facilitate the viewing and comparison of image data. More research organizations are publishing directly on the Internet and providing data on the Internet through on-line databases. Some of these data are suitable for comparison between laboratories. Other data are not, but do serve as representative instances of a particular methodology, which may be useful for improving preparative methods with similar types of samples. Peer-reviewed WWW journals are also beginning to appear on the Internet [29] as a quick way to publish. This type of network-based review process helps address the data quality issue and will be discussed in Section 2. These Internet databases may also be used to provide standard samples for a variety of samples (*e.g.* 2-D gels with spot maps of identified proteins, protein or DNA sequences, structural motifs, *etc.*).

1.2 Finding relevant data on the Internet

In the early days of the WWW, one of the problems was an insufficient amount of data. With the explosive growth of the WWW and the Internet, this became a problem of finding too much data – most of it of low quality or not relevant. This occurs because WWW search engines are often not very discriminating. Their WWW-crawler indexing algorithms do not index all WWW sites equally well. This results in a great deal of marginal data, often unrelated to the problem at hand, which sometimes makes the search results worthless. We therefore need to find the relevant data to take better advantage of this potentially powerful collaborative medium.

WWW sites are indexed by "WWW crawlers" associated with search engines. They work by finding a popular site, then reading its top level WWW page, analyzing it for other links (both on that site and outside), and then visiting each of those links in turn. This lets it visit all sites around the world which are connected in some way to that initial site. They then use the content returned from these sites to create an index. When queried by WWW browser users, the associated search engines look through these indexes for keywords and return pages which have these keywords. However, because of the strictly syntactic analysis of the initial content indexing, this type of indiscriminant indexing often returns misleading links. Some search index WWW sites are now hiring people to "cruise" the WWW and define more intelligent indexing to improve indexing quality. There is a specialized WWW crawler for the 2-D gel community called SWISS 2DHunt (http://www.expasy.ch/ch2d/2DHunt/) that focuses on finding only 2-D gel electrophoresis WWW sites and so should provide an enriched data source for these types of WWW sites.

1.3 Meta-databases

We help address this problem for finding 2-D PAGE gels by creating a specialized Internet meta-database which provides a set of enriched links to specific data. In general, a meta-database contains information about locating data in other databases but does not contain that data itself. For example, SWISS-PROT provides an enriched set of data which also includes organized links to specific data in other WWW databases such as Medline, OMIM, *etc.* Similarly, the 2DWG meta-database provides an enriched set of links to 2-D gel images, associated databases and protein maps. Unlike a WWW crawler which picks up everything (including much irrelevant material) and is probably complete, a manually edited meta-database depends on submissions by others, including the editorial board. Therefore, it will only be

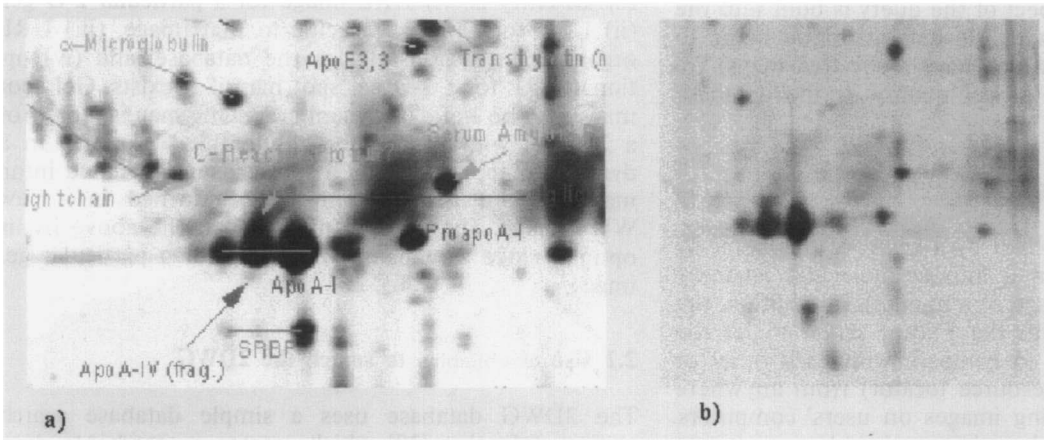


Figure 1. Corresponding regions of plasma protein gels from different labs. (a) Plasma gel run with immobilized pH gradient, nonlinear gradient (SWISS-2DPAGE); (b) plasma gel run with carrier ampholytes and a linear gradient (C. Merril, NIMH).

as good as the effort the editorial board and the 2-D gel communitity put into it.

1.4 Distributed WWW databases

A major advantage of distributed databases is their lower costs, which are amortized across institutions. The total cost is often more than could be supported by any single group. Sharing the costs then makes free access more likely. On the downside, distributed databases may be less reliable. Some of their risks include (i) nonuniformity of data-encoding conventions, robustness, consistency, and commitment of the group to maintain the database; (ii) data quality may be variable; and (iii) Internet access may be slow or unavailable. These quality and uniformity issues are beginning to be addressed by the scientific and Internet community.

1.5 Enriched collection of WWW images – the 2DWG meta-database

Because of the problem of finding enriched sources of 2-D gel image data on the WWW, we constructed the 2DWG meta-database of 2-D protein gel images and maps found on the WWW. This was done initially by manually data-mining the 2-D gel WWW databases found in Table 2. Data mining consists of going into a database and finding as much relevant content as possible – often using that content in new ways. The 2DWG is a spreadsheet organized by tissue with hypertext links to WWW 2-D gel databases for images, associated data, and protein spot maps. Because the data found on the WWW is of variable quality with varying amounts of documentation, we needed a way to help indicate that some data is of a higher quality than others. The definition of quality can take on several aspects: (i) is the sample and its preparation characteristic of the material? (ii) is the 2-D gel technique of a quality recognized by the 2-D gel community? This can only really be answered by a peer-review of the site and was one of the reasons we added the editorial board. Although not necessarily the optimal solution, we currently indicate that a gel may be higher quality if it has an associated protein "map image". Otherwise we indicate that it is a "raw gel image". This aspect of the query is built into the search query interface with the default option being to return only those gels which have associated maps. We may add other criteria of gel quality as the database develops.

1.6 Searching the 2DWG

The 2DWG can be searched by a keyword phrase or by clicking on an organ in a human molecular anatomy icon (a coronal view image of a human). In addition, the 2DWG is integrated with the Flicker server to let the users select gel images to compare with each other or with URLs (universal resource locator) from anywhere on the Internet, including images on users' computers. Because the 2DWG database is organized by row, it may be easily searched to find rows matching a combination of search terms present in that row. A set of terms is grouped into a search expression which is used in the search where a term is a word or word fragment without spaces and is case-independent. Each term may be delimited by either AND or OR connectives. For example, for terms "ventricle", "human" and "map", the search expression might be "ventricle AND human AND map".

1.7 2DWG database maintenance – submission of data from the WWW

After constructing the initial 2DWG database by data-mining known databases (*cf.* Table 2), our aim was to automate the growth of the database since we do not envision ourselves continuously searching the WWW for new data. That process is too labor-intensive. We therefore established a WWW-based data submission process for the 2DWG using data entry forms (entered from the 2DWG home page). We are also evaluating an (E-mail/WWW based) editorial-board peer-review of submitted data. Unfortunately, we can not guarantee completeness of the database if someone does not submit a new gel database or map. For this reason, we foresee the editorial board and the editor holding active roles in suggesting gels and databases to be added as well. We now discuss some of the specific issues for the 2DWG for its organization, searching, integration with Flicker image comparison and data submission.

2 Materials and methods

2.1 Organization of the 2DWG as a spreadsheet

Because a major goal was to organize the 2-D gel image data by tissue, we decided to allocate one gel image and its associated data per row. Section 6.2 is a glossary of 2DWG table column headings. Since we were pointing to data residing on the WWW server from which the gel came, we did not have to copy much data into the 2DWG – only the WWW location of that data. However, we did want to be able to delve quickly into the associated parts of that 2-D gel database WWW server. Therefore, we decided to provide a minimum set of hypertext links back to each 2-D gel image-associated database WWW server. The final set of hypertext links for an entry is: (i) URL image for a particular 2-D gel, (ii) URL for database specific to that image, (iii) URL organization responsible for the database, and (iv) optional URL for a 2-D gel spot map if it exists. Gel spot maps may be static or dynamic. A static map is a copy of a 2-D gel with the names of identified proteins. In a dynamic map, clicking on a spot causes associated information about that protein to be returned in a new WWW page. The 2-D gel map link in the above list is optimal since a map may not exist for a particular gel image.

2.2 Use of dbEngine to search the 2DWG

The 2DWG database uses a simple database search engine dbEngine [30] which creates a searchable database on a World Wide Web (WWW or Web) server. Data for the dbEngine is prepared from spreadsheet programs (such as Excel, dBase-IV, *etc.*) or from tables exported

from relational database systems. The table consists of records (rows) and each row has related fields (columns). As a common gateway interface (CGI) program [31], dbEngine is used with a WWW server such as is available commercially or in the public domain from NCSA, CERN and others. Capabilities of the dbEngine include: (i) searching records by matching them with an expression. The expression is a list of terms (without spaces) separated with ANDs or ORs. The search result is then presented as a hypertext list or table; (ii) mapping some fields returned in the search results to hypertext links to other WWW database servers; (iii) creating bidirectional hypertext links between pictures and the database entries; (iv) drawing an overlay region around objects in an image by clicking on a result in the search results (*e.g.*, draw a spot's location in a 2-D gel image). We have also used the dbEngine to (v) support federated 2-D protein gel databases with associated clickable 2-D maps (see http://www-lecb.ncifcrf.gov/dbEngineDatabases.html). Items (iv) and (v) are enhancements in the new version 2.0 of dbEngine.

2.3 Converting 2DWG HTML data to a searchable database format

The full database is maintained in hypertext markup language (HTML) format in a file called 2DWG.html.* We used the HTML format as the primary data format since it makes it easier to integrate new data from the data entry submission process to be discussed. The format required by dbEngine is a simple tab-delimited single record-per-line file, such as data available as interchange files from spreadsheets or relational databases. We constructed a conversion program cvhtml2db which extracts the HTML table row data from the full 2DWG.html database file and then creates a dbEngine compatible data file. The latter is saved in the WWW server and is used in subsequent searches.

2.4 Searching the 2DWG

Individual table entries or sets of entries may be accessed using the 2-D federated database paradigm. Since each table entry has a unique identifier, WG*nnnnn*, it can uniquely access a single row. The key point is to specify a unique search string. If it is not unique, multiple rows will be returned. This search string is passed to the dbEngine as if it had been typed in the search interface. For example "Human AND Ventricle AND map" is specified as "Human+Ventricle+map". To make it easier to type in search strings manually, we provide a standard search interface. This has buttons to define whether or not the image is an image map. The user interface is shown later in Fig. 4. The syntax for accessing this data as a federated database is shown for these two examples:

```
http://www-lecb.ncifcrf.gov/cgi-bin/dbEngine/2dwgDB,
getTableDataByID,WG00123
or
http://www-lecb.ncifcrf.gov/cgi-bin/dbEngine/2dwgDB,
getTableDataByID,Human+Ventricle+map
```

2.5 Searching the 2DWG with a molecular anatomy icon

In addition to searching by keyword phrase, the 2DWG can be searched by clicking on an organ in a cross section (coronal slice) of a human molecular anatomy image map. This WWW page is accessible from the 2DWG home page. This map is meant for demonstration purposes to indicate future possibilities and not to represent a complete human anatomic database. Clicking on the "+" sign in an organ causes the corresponding subset of the 2DWG entries to be returned. Only 2-D gels with associated human protein maps are reported.

Note that there are other gels of species and tissues with and without maps in the complete 2DWG, so this iconic search does not cover the whole database. The concept of molecular anatomy was proposed in the 1970s and early 1980s by Anderson *et al.* [32, 33] and today increasing numbers of identified proteins and their post-translational modifications are being cataloged by many groups. These human anatomical icons for male and female were derived from the Visible Human Simulation server (located at http://www.uchsc.edu/sm/chs/). Because not all organs are visible in these sections, we label the approximate location of where they would be.

2.6 Invoking flicker comparison on the 2DWG search results list of gels

The National Cancer Institute (NCI) Flicker comparison program [28] (*cf.* Section 6.1 for a short discussion on Flicker and image comparison) is a Java application running on your WWW browser for visually comparing images over the Internet. You can flicker-compare a gel image residing on your computer with one of the images in the 2DWG search results table. Alternately, you can flicker data among gels from just the 2DWG. There is a checkbox, *i.e.* ☐, adjacent to each **Image URL** entry for each gel in the table. The user can select one or two of these gels to Flicker and then press the "Go Flicker selected gels" button. To compare another gel not found in the 2DWG, the user would type in the URL, select only one gel and then press the button. This is clarified in Fig. 7.

In 2-D gel image databases, some images are reversed in the horizontal p*I* direction and others may be reversed in M_r in the vertical direction. We indicate this in the 2DWG p*I* range by the direction of the range (*e.g.* 8–4 for basic to acid, rather than 4–8 for acid to basic). Flicker will let you transform an image with the **Flip Horiz** and **Flip Vert** operations to flip an image in the horizontal and vertical dimensions. Because some images are on a different scale (either through the way they were run or because of the scanner resolution), you might consider using the **Affine Transform** option to make the regions being compared have a similar scale. The Flicker Reference Manual (accessible from the Flicker Server) contains additional information on running Flicker. Figure 2 illustrates the client-server paradigm used by Flicker and the distributed 2-D gel databases.

* This file can be accessed in your WWW browser. However, because of its size we recommend instead using the 2DWG search facility to load a relevant subset of the data.

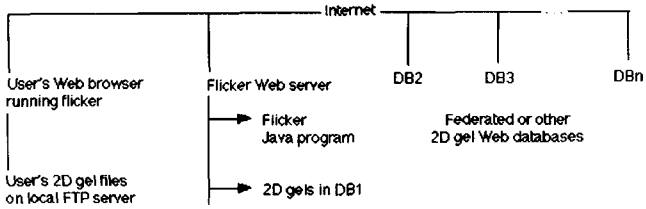


Figure 2. Distributed data client-server paradigm. This schematic presentation illustrates the relationship between the users' WWW browser with their own 2-D gel images residing on their local FTP server, the NCI Flicker server which contains the Flicker program, and the federated or other 2-D gel databases (DB) on other WWW servers. Two gel images to be compared may come from the Internet WWW databases or from the users' local FTP server. The images may be from either the Flicker 2-D gel image DB WWW server or from other federated 2-D gel image WWW databases DB2, DB3, ... DBn listed in Table 2 or elsewhere.

2.7 Internet access restrictions with current Java and WWW browsers

Current Java security restrictions prevent Flicker when it runs in your WWW browser from reading local files or URLs from WWW sites other than the NCI Flicker Server. This of course would prevent us from meeting a major goal of this project – to compare gel images between labs. There are several solutions. The first is to run the Flicker Java program as a stand-alone "Java" application. This is difficult since users would have to install the Flicker program and other related Java support software. The second, simpler, solution is to use indirect image fetching by the NCI Flicker proxy server to get an image from the WWW for Flicker rather than having Flicker do it (which is restricted from doing this).

The current versions of Netscape (3.x) and Microsoft Internet Explorer (3.x) enforce a highly restrictive security when used with Java applets. Applets running on these WWW browsers cannot read or write local files or download data from WWW location URLs other than the WWW server from which the applet came. We have implemented an interim second solution to allow access to images on any computer on the Internet without violating the browser security. When new versions of WWW browsers are released which allow direct access, we will shut down this service.

The interim solution (*cf.* Fig. 3) uses the NCI Flicker Server, which provides a URL proxy service integrated with Flicker. When a URL is requested in Flicker for a host other than the NCI Flicker server, Flicker passes this request to the NCI proxy server. It in turn gets the image data for the specified URL using the public domain wget program and saves it as a temporary disk file. If needed, it then converts the image file to graphics interchange format (GIF) using the ImageMagick convert program which greatly increases the variety of WWW image formats that Flicker can handle. Finally, it sends the GIF image back to Flicker.** This proxy service can handle images up to 1.5 Mbytes in size (to limit the load on our server) and handles http:// and ftp:// protocols. It may have problems with some CGI image access methods.

2.8 Extending the 2DWG using data submitted from the Internet

As was mentioned, the initial data for the 2DWG was obtained by data-mining 2-D gel Internet databases. We are now soliciting contributions of high-quality gels from the 2-D gel WWW server community to expand the database further. A WWW data entry form is used for submitting a new entry to the editor of the 2DWG database for review and eventual inclusion in the database. We are still in the process of setting up the network-based peer-review mechanism, but will outline it here.

2.8.1 Criteria for data to be submitted to the 2DWG

When you submit data describing one of your gels residing on the WWW server, the data should adequately describe that gel. Your WWW server should provide supporting information regarding that gel, including: (i) 2-D protein gel images. If you are using pseudo-images, we suggest also supplying raw (*i.e.* stained) gels, autoradiographs, blots, *etc.*, which are easier to compare with gel images from other labs. The file names should reflect the material and not simply be called "gel", "image", "slide", "figure", *etc.* (ii) The protocol for running the gel, including apparatus, detection method, carrier ampholytes (CA) or NEPHGE or other, IEF or IPG or other, pH range, M_r range, *etc.* (iii) The protocol for sample preparation. (iv) An image map identifying proteins. This can be a static labeled map or a dynamic map where one could click on a spot to cause that WWW server to report the information on the corresponding protein identification. There may also be additional information provided in tabular form. (v) References to the refereed literature where the biomedical data and protocols are described. (vi) Other information useful for searching the 2DWG or interpreting the data should also be included and may be entered in the Miscellaneous field.

It is much more convenient for readers to have access to this information on-line in a WWW server through hypertext links since it may be difficult for many readers to get journal papers, tech-reports or books which are

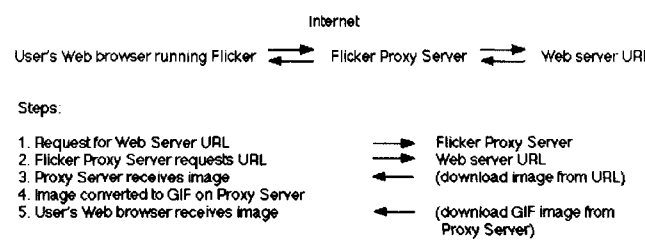


Figure 3. The NCI Flicker URL proxy server. When a URL is requested in Flicker for a host other than the NCI Flicker server, Flicker passes this request to the NCI Flicker server. It in turn acts as a proxy and gets the image data for the specified URL. If needed, it converts the image to GIF format – increasing the variety of image formats which Flicker can handle. The proxy service can handle images up to 1.5 Mbytes in size and with http:// and ftp:// protocols.

** The wget program is available at ftp://prep.ai.mit.edu/pub/gnu/. ImageMagick is available from ftp://ftp.wizards.dupont.com/pub/ImageMagick/. GIF is a standard graphics image format available on most computers.

not widely available. Few libraries can afford to subscribe to all specialized journals. Gel images themselves are not submitted to 2DWG, only information about them. This is because the 2DWG is a meta-database that describes data residing on other WWW or file transfer protocol (FTP) servers. The image files are generally in GIF, joint photographic experts group (JPEG) or tagged image file format (TIFF), which are suitable for WWW browsers. WWW addresses or URLs are used to define links to these other WWW databases. The 2DWG only accepts the "http://" or "ftp://" protocols. If the images are to be used with NCI Flicker, image size should be 8-bit to 24-bit color or gray scale and approximately 500–1000 pixels (in both dimensions). Since the data is only used for viewing – not quantitation, very high spatial and color resolution is not needed and would only slow down access for users who lack high-speed Internet connections.

2.8.2 Experimental Internet peer-review of submitted data

We are in the process of implementing an electronic peer-review for submitted materials by 2DWG Editorial Board members. When data is submitted to the 2DWG using WWW data entry forms, the compiled data document will be E-mailed to these reviewers who should respond within a short time. The reviewers will interact with a 2DWG staging server (different from the 2DWG) where they can judge the material. We plan to implement an "Accept/Reject/Revise with comments" system for the reviewers, with the submitter being notified by return E-mail. As this submission and review process is experimental, we foresee changes being made as we gain experience. We encourage the research community to submit high quality 2-D gel materials to the 2DWG.

We will encourage reviewers to use the following criteria (some of which was suggested by the ExPASy group) in evaluating this data: (i) database content is relevant and of academic value, (ii) correctness and completeness of the content, (iii) quality and clarity of the content, and (iv) contains valid and relevant links – especially maps and related information. The 2DWG Editor will automatically review links to the data periodically and mark entries which are truly no longer available (*i.e.*, not simply because their server is down one day), so that 2DWG accesion numbers will never be reused.

2.8.3 Submission process

When submitters complete the submission form (http://www-lecb.ncifcrf.gov/2dwgDB/2dwgSubmitData.html), they press the "Submit data" button that sends it to the 2DWG editor for review. However, they may get an error message back that some fields are incomplete or incorrect. The data entry program will return a list of exactly which fields are incorrect and suggest what they may need to do to correct them. Corrections can be easily made. First, they should click on the Back button on the WWW browser. This will bring them back one level to the submitted browser in the data form with the entries intact. Then the submitter would change just those fields which were incorrect and resubmit the form. By making the corrections this way, all of the fields do not have to be retyped just to correct a few errors. If the submission process to the Editor is successful, an accession number will be assigned and reported back to the submitter in the confirmation. The accession numbers are of the form "WG*nnnnn*" where *nnnnn* is a sequential number. If corrections regarding this entry are needed in the future, they should be sent by E-mail to the 2DWG editor describing the changes and indicating the "WG*nnnnn*" identifier. When entering multiple gels which are similar, this can entail repeatedly typing a great deal of data. However, there is a simple trick to avoid having to retype all of the fields which are the same. Simply click on the Back button on the WWW browser until the filled-in form is visible again. Then scroll to the specific fields that need changing and change only those fields. Verify that all of the fields are correct and then submit the new entry.

2.8.4 Publishing 2-D gel data on the World Wide Web

As noted in Section 1, there are many groups running 2-D gels on a variety of materials. Many of these gels have protein spots. However, not that many groups have created WWW server 2-D gel databases for their data. This is the case even though many of these groups have access to the Internet and Internet servers where they could put their data. In the past, the mechanics of creating a 2-D gel WWW site was daunting. Laboratories that want to publish 2-D gel databases on their WWW server now do so in a number of ways. Section 6.3 goes into more detail on the mechanics of how to do this with new, easier methods.

3 Results

The 2DWG can be searched in a variety of ways. We will present two examples. Because it is easy to experiment with the Internet from a WWW browser, we suggest the user investigate some of the other materials and search methods available in the 2DWG server. The user interface to the search engine is shown in Fig. 4. Figure 5 shows the five gels returned from a search of "Ventricle AND map". To return only human ventricle gels, the query would be "Ventricle AND human AND map". Figure 6 shows the Flicker selection and URL entry interface that is part of the search results. This then lets users select and compare gel images in the 2DWG with gels found elsewhere on the Internet. Users would select the desired gels from the table or URL type-in field and then press the "Go Flicker selected gels" button. In another example, Fig.7 shows part of the search results table of a search for "breast AND human AND map". This catalogs a number of the breast cancer cell-line gels generated in the large NCI drug screening project. Having this targeted list of gel images makes it easier to select and flicker-compare gels between different cell lines and between different laboratories.

4 Discussion

Meta-databases such as 2DWG are useful Internet resources of enriched data which simplify finding rele-

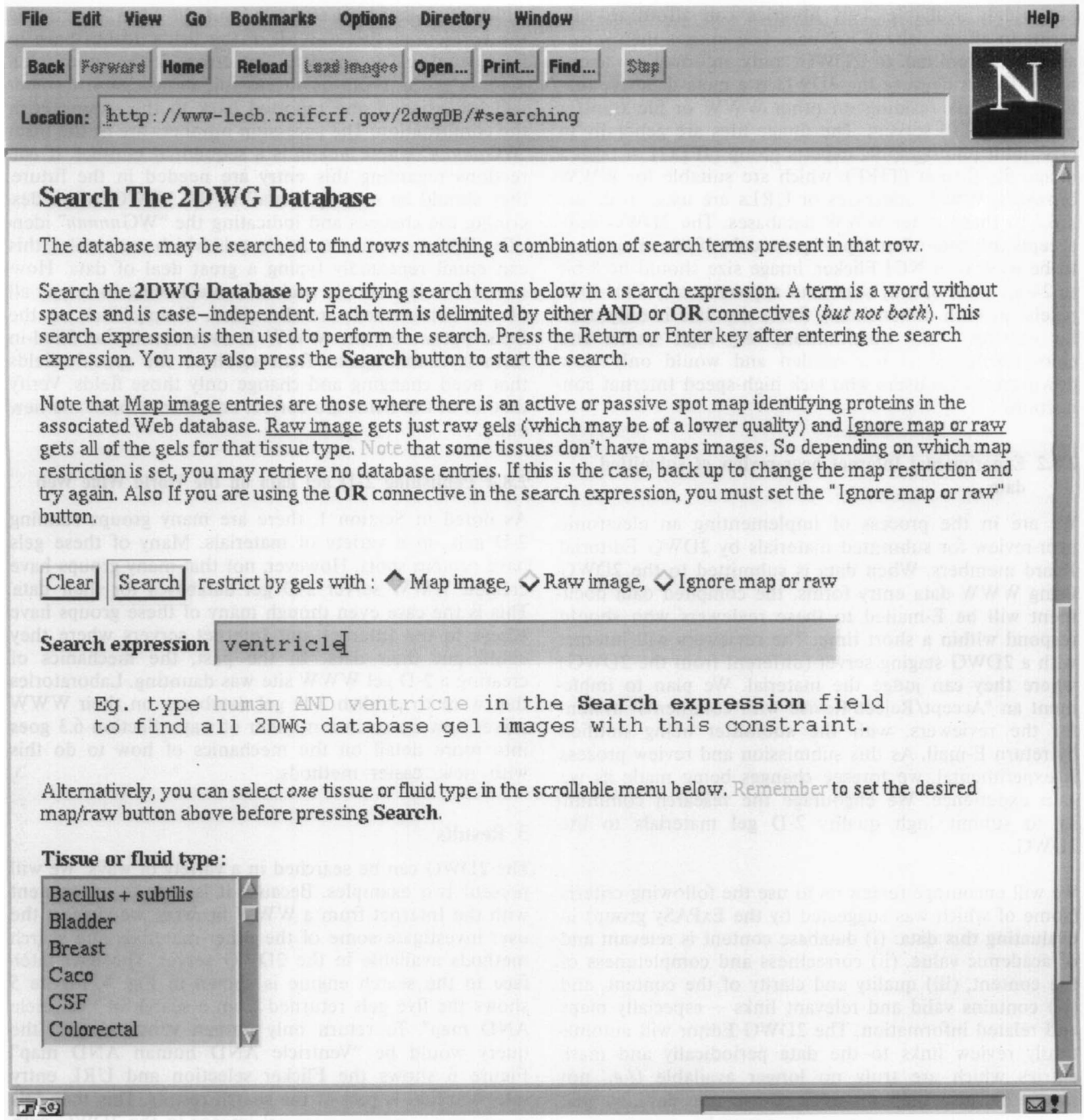
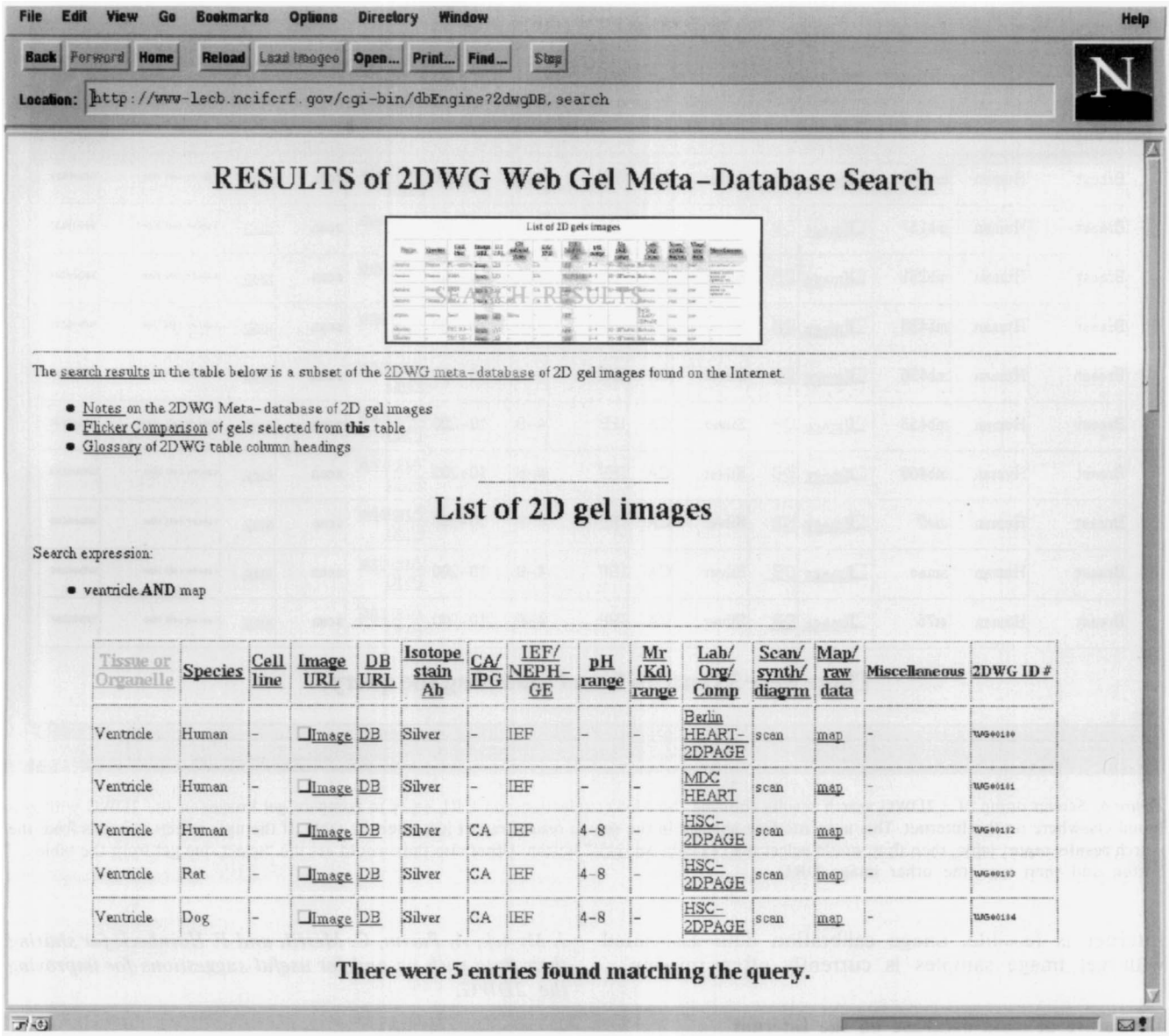


Figure 4. Screen dump of a 2DWG search interface. The query may be entered by typing a search expression or selecting a tissue or fluid type. In both cases, the search may be restricted by requiring that only gels with maps, raw gels, or neither be returned.

vant data for a specialized problem domain. There is a current need for such meta-databases since there are no general-purpose WWW indexing search engines available that provide uniformly high quality lists of URLs for arbitrary problem domains. Using the 2DWG provides easy access to tissue-specific standard 2-D gel images on the WWW. This data then greatly simplifies the process of researchers flicker-comparing their own local 2-D gel image against this standard data. In addition to their utility for 2-D gels, we foresee using meta-databases for other biomedical image domains – especially those where existence of a standard sample makes sense (*e.g.* catalogs of RFLP patterns, mass-spectra, or HPLC spectra, *etc.*). Other domains, such as a set of tumor progression images for a particular patient, of course, do not carry over between patients. However, these images may still be interesting from the point of view of defining typical changes as a function of time or progression of disease stages.

Integrating WWW data analysis tools with meta-databases improves access to that data for subsequent anal-



Tissue or Organelle	Species	Cell line	Image URL	DB URL	Isotope stain Ab	CA/ IPG	IEF/ NEPH-GE	pH range	Mr (Kd) range	Lab/ Org/ Comp	Scan/ synth/ diagrm	Map/ raw data	Miscellaneous	2DWG ID #
Ventricle	Human	-	Image	DB	Silver	-	IEF	-	-	Berlin HEART-2DPAGE	scan	map	-	
Ventricle	Human	-	Image	DB	Silver	-	IEF	-	-	MDC HEART	scan	map	-	
Ventricle	Human	-	Image	DB	Silver	CA	IEF	4-8	-	HSC-2DPAGE	scan	map	-	
Ventricle	Rat	-	Image	DB	Silver	CA	IEF	4-8	-	HSC-2DPAGE	scan	map	-	
Ventricle	Dog	-	Image	DB	Silver	CA	IEF	4-8	-	HSC-2DPAGE	scan	map	-	

Figure 5. Screen dump of a 2DWG search showing the resulting table from a search for "Ventricle AND map".

ysis. By providing a direct route for obtaining data from the distributed database, the meta-database makes it easier for occasional computer users to obtain data. Without integrated WWW tools, obtaining the data becomes increasingly complex and providing this data to the analysis tool may be overwhelming for the average user. An integrated approach automates both steps. Using a WWW-based data submission process, with peer review by E-mail and the WWW for 2DWG, will help the database expand in a controlled way that helps preserve data quality while ensuring rapid publication. The distributed data analysis paradigm helps enhance scientific collaboration within the research community by making large amounts of data available, often earlier than if provided by commercial database vendors. The latter tends to enter a market only when a critical mass of users is reached. On the downside, distributed databases may be less reliable but that should be ameliorated by the increased use of peer review on the Internet.

4.1 Flicker Comparison with the 2DWG

Using Flicker with the 2DWG, users can visually compare their own data with data from a wide range of Internet image database. With the WWW and Java, it is now possible to provide real-time interactive software for data analysis on a user's computer WWW browser using software distribution over the Internet which is transparent to the user. These tools open up possibilities for increased collaboration because collaborators separated by distance but having access to the WWW can visualize and discuss the same data at the same time. To achieve optimal results in a comparison, it is necessary to use samples and methods that are as similar as possible. The analysis is only potentially as good as the data being compared. Putative 2-D gel spot identification may suggest which antibodies to try without having to sequence spots or use other expensive methods. Although quantitation of arbitrary 2-D gel images from the

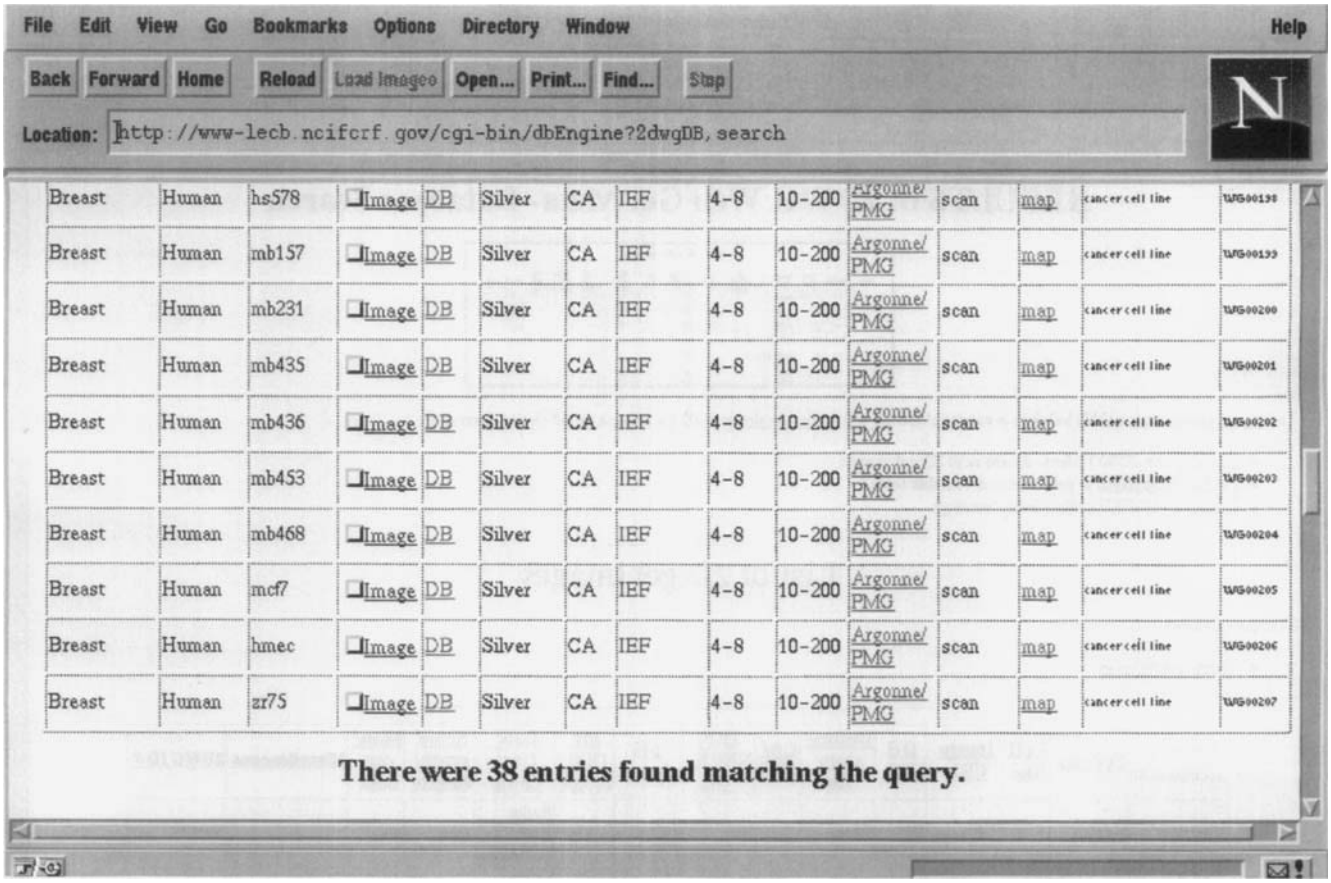


Figure 6. Screen dump of a 2DWG search results showing the Flicker selection and URL entry to compare gel images in the 2DWG with gels found elsewhere on the Internet. This user interface appears in the search results report just after the table. If the user selects two gels from the search results report table, then they would select the "exactly two gels" button. Otherwise they would set the "select one gel from the table ..." button and then type the other image URL.

Internet is feasible, image calibration data associated with gel image samples is currently often missing.

4.2 Future of meta-database on the Internet

The technology for searching the Internet is improving for high quality specialized data and at some point it may be much easier to find specific data. What may be more difficult is to present that data in a format that makes it easily accessible for these specialized users. The speed and power of the WWW has surprised everyone. Perhaps the next generation of WWW technology will surprise us as well with the ability to configure integrated work environments combining powerful WWW data analysis tools with ways of quickly finding targeted quality problem-domain specific data. We feel that integrated support environments will be a driving force for growth in these specialized research communities and other groups which might not normally have access to this data.

Thanks are due to E. Burchill, T. Schneider, and G. Thornwall for useful suggestions for improving this manuscript. Thanks also to the 2-D protein gel electrophoresis groups which have made their 2-D gel image and protein map data available on the Internet. In particular I want to thank D. Hochstrasser, J. Celis, M. Dunn, J. Weinstein, C. Giometti, J. Myrick, A. Partin, C. Merril, and P. Hornbeck for sharing their data with us and for useful suggestions for improving the 2DWG.

Received June 2, 1997

5 References

[1] Celis, J. E. (Ed.), Special issue: Two-dimensional Electrophoresis protein databases, *Electrophoresis* 1992, *13*, 891–1062.
[2] Celis, J. E. (Ed.), Special issue: Two-dimensional Electrophoresis protein databases, *Electrophoresis* 1993, *14*, 1089–1240.
[3] Celis, J. E. (Ed.), Special issue: Electrophoresis in Cancer Research, *Electroporesis* 1994, *15*, 305–556.
[4] Celis, J. E. (Ed.), Special issue: Two-dimensional Electrophoresis protein databases, *Electrophoresis* 1995, *16*, 2175–2264.
[5] Dunn, M. J. (Ed.), 2-D Electrophoresis: From Genome to Proteome, Proceedings of the International Meeting, Siena, Sept. 5–7, 1994, *Electrophoresis* 1995, *16*, 1077–1326.
[6] Dunn, M. J. (Ed.), From Genome to Proteome, Proceedings: 2nd Siena 2-D Electrophoresis Meeting, Siena, Italy, Sept. 16–18, 1996, *Electrophoresis* 1997, *18*, 307–661.
[7] Appel, R. D., Sanchez, J.-C., Bairoch, A., Golaz, O., Miu, M., Vargas, J. R., Hochstrasser, D. F., *Electrophoresis* 1993, *14*, 1232–1238.
[8] Sanchez, J.-C., Appel, R. D., Golaz, O., Pasquali, C., Ravier, F., Bairoch, A., Hochstrasser, D. F., *Electrophoresis* 1995, *16*, 1131–1151.
[9] Hughes, G. J., Frutiger, S., Paquet, N., Ravier, F., Pasquali, C., Sanchez, J.-C., James, R., Tissot, J.-D., Bjellqvist, B., Hochstrasser, D. F., *Electrophoresis* 1992, *13*, 707–714.

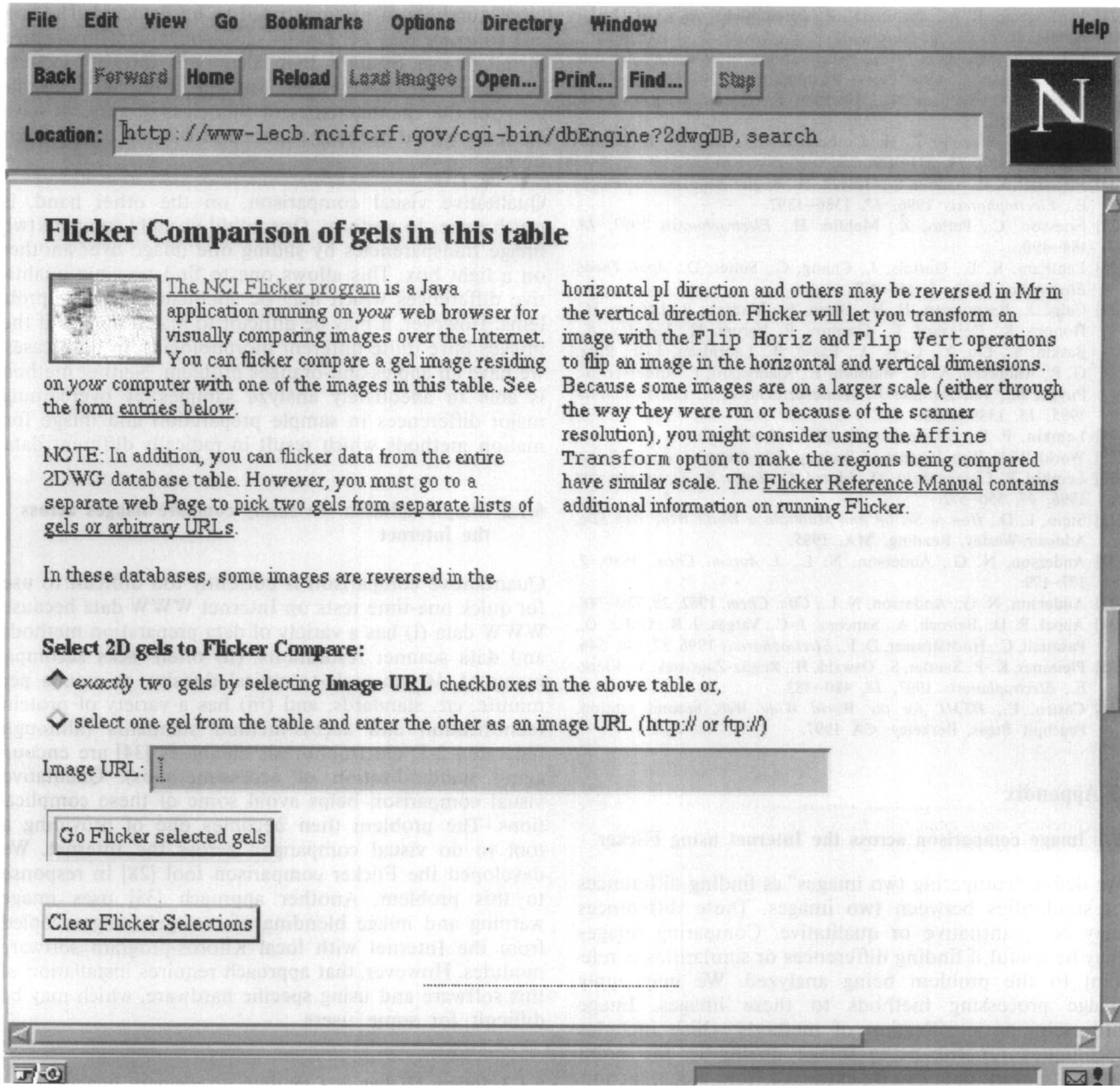


Figure 7. Screen dump of a 2DWG search results report showing part of the resulting table for the query "Breast AND map". Even for a tissue type with a large number of gel images (several breast cancer cell-lines from different labs), it is easier to use the search facility to retrieve data than to display the entire 2DWG database with its hundreds of gels. Having this target list makes it easier to select and compare gels from these different cell-lines laboratories.

[10] Golaz, O., Hughes, G. J., Frutiger, S., Paquet, N., Bairoch, A., Pasquali, C., Sanchez, J.-C., Tissot, J.-D., Appel, R. D., Walzer, C., Hochstrasser, D. F., *Electrophoresis* 1993, *14*, 1223–1231.

[11] Lemkin, P. F., Orr, G. A., Goldstein, M. P., Creed, J., Whitley, E., Myrick, J., Merril, C. R., *Electrophoresis* 1995, *16*, 1175.

[12] Gravel, P., Sanchez, J.-C., Walzer, C., Golaz, O., Hochstrasser, D. F., Balant, L., Hughes, G. J., Garcia-Sevilla, J., Guimon, J., *Electrophoresis* 1995, *16*, 1152–1159.

[13] Golaz, O., Walzer, C., Hochstrasser, D. F., Bjellqvist, B., Turler, H., Balant, L., *Appl. Theor. Electrophor.* 1992, *3*, 77–82.

[14] Hochstrasser, D. F., Frutiger, S., Paquet, N., Bairoch, A., Ravier, F., Pasquali, C., Sanchez, J.-C., Tissot, J.-D., Bjellqvist, B., Vargas, R., Appel, R. D., Hughes, G. J., *Electrophoresis* 1992, *13*, 992–1001.

[15] Hughes, G. J., Frutiger, S., Paquet, N., Pasquali, C., Sanchez, J.- C., Tissot, J.-D., Bairoch, A., Appel, R. D., Hochstrasser, D. F., *Electrophoresis* 1993, *14*, 1216–1222.

[16] Sanchez, J.-C., Golaz, O., Frutiger, S., Schaller, D., Appel, R. D., Bairoch, A., Hughes, G. J., Hochstrasser, D. F., *Electrophoresis* 1996, *17*, 556–565.

[17] Latter, G. I., Boutell, T., Monardo, P. J., Kobayashi, R., Futcher, B., McLaughlin, C. S., Garrels, J. I., *Electrophoresis* 1995, *16*, 1170–1174.

[18] Boutell, T., Garrels, J. I., Franza, B. R., Monardo, P. J., Latter, G. I., *Electrophoresis* 1994, *15*, 1487–1490.

[19] Pasquali, C, Frutiger, S., Wilkins, M. R., Hughes, G. J., Appel, R. D., Bairoch, A., Schaller, D., Sanchez, J.-C., Hochstrasser, D. F., *Electrophoresis* 1996, *17*, 547–555.

[20] Garrels, J. I., *Nucleic Acids Res.* 1995, *24*, 46–49.

[21] VanBogelen, R. A., Abshire, K. Z., Pertsemlidis, A., Clark, R. L., Neidhardt, F. C., in: Neidhardt, F. C., Gross, C. A., Ingraham, J. L., Riley, M. (Eds.), *Gene-Protein Database of Escherichia coli K-12*, Edition 6, ASM Press, Washington, DC 1996.
[22] Giometti, C., Williams, K., Tollaksen, S. L., *Electrophoresis* 1997, *18*, 573–581.
[23] Evans, G., Wheeler, C. H., Corbett, J. M., Dunn, M. J., *Electrophoresis* 1997, *18*, 471–479.
[24] Pleissner, K-P., Sander, S., Oswald, H., Regitz-Zagrosek, V., Fleck, E., *Electrophoresis* 1996, *17*, 1386–1392.
[25] Ericsson, C., Petho, Z., Mehlin, H., *Electrophoresis* 1997, *18*, 484–490.
[26] Lantham, K. E., Garrels, J., Chang, C., Solter, D., *Appl. Theor. Electrophor.* 1992, *2*, 163–170.
[27] Celis, J., Rasmussen, H. H., Olsen, E., Madsen, P., Leffers, H., Honoré, B., Dejgaard, K., Gromov, P., Vorum, H., Vassilev, A., Baskin, Y., Liu, X., Celis, A., Basse, B., Lauridsen, J.-B., Ratz, G. P., Andersen, A. H., Walbum, E., Kjaergaard, I., Andersen, I., Puype, M., Van Damme J., Vanderkerckhove, J., *Electrophoresis* 1995, *15*, 1349–1458.
[28] Lemkin, P. F., *Electrophoresis* 1997, *18*, 461–470.
[29] World Wide Web Journal of Biology, http://epress.com/w3jbio/.
[30] Lemkin, P., Chipperfield, M., Merril, C., Zullo, S., *Electrophoresis* 1996, *17*, 556–572.
[31] Stein, L. D., *How to Set up and Maintain a World Wide Web Site*, Addison-Wesley, Reading, MA, 1995.
[32] Anderson, N. G., Anderson, N. L., *J. Autom. Chem.* 1980, *2*, 177–179.
[33] Anderson, N. G., Anderson, N. L., *Clin. Chem.* 1982, *28*, 739–748.
[34] Appel, R. D., Bairoch, A., Sanchez, J.-C., Vargas, J. R., Golaz, O., Pasquali, C., Hochstrasser, D. F., *Electrophoresis* 1996, *17*, 540–546.
[35] Pleissner, K.-P., Sander, S., Oswald, H., Regitz-Zagrosek, V., Fleck, E., *Electrophoresis* 1997, *18*, 480–483.
[36] Castro, E., *HTML for the World Wide Web*, second edition, Peachpit Press, Berkeley, CA 1997.

6 Appendix

6.1 Image comparison across the Internet using Flicker

We define "comparing two images" as finding differences or similarities between two images. These differences may be quantitative or qualitative. Comparing images may be useful if finding differences or similarities is relevant to the problem being analyzed. We may apply image processing methods to these images. Image processing is a collection of methods which includes techniques for enhancing image quality to the point where relevant data can be extracted from a comparison.

6.1.1 Comparing images in a collaborative setting

How can scientists around the world compare similar image data? We can do this primarily the two ways we mentioned. However, both methods have difficulties. Quantitative numeric comparison requires building a computer database from quantitative data extracted from the images using special local software. This software measures objects in the images and is able to build composite databases where we can do various statistical tests. However, this is difficult or impossible to perform if calibrated gel image data is not available locally. Building composite gel databases is time-consuming if we only want to compare one data point between the two images. This method is very good for long-term multiple-image databases where many proteins are being analyzed under a variety of conditions. It is also required when subtle quantitative changes or changes involving large numbers of proteins need to be detected. The special software may be expensive, inconvenient to acquire, install or use. Systems that fall in this category include BioImage, GELLAB, LSB, Melanie, PDI, *etc.* Descriptions of the characteristics of this class of systems have been published in many papers and will not be reviewed here.

Qualitative visual comparison, on the other hand, is much easier to perform. One could visually compare two image transparencies by sliding one image over another on a light box. This allows one to find possible qualitative differences which may be adequate for some problems. However, it may be difficult to match objects if the images have quite different morphologies. In both cases, we have an apples-and-oranges problem. Neither method is able to adequately analyze samples by overcoming major differences in sample preparation and image formation methods which result in radically different data.

6.1.2 Simple solution – visually compare images across the Internet

Quantitative comparison is currently too difficult to use for quick one-time tests on Internet WWW data because WWW data (i) has a variety of data preparation methods and data scanner resolutions, (ii) often lacks accompanying p*I*, M_r grayscale-to-optical density, or counts per minute, *etc.* standards, and (iii) has a variety of protein identification and access-method standards (although federated 2-D electrophoresis databases [34] are encouraging standardization of access-methods). Qualitative visual comparison helps avoid some of these complications. The problem then becomes one of providing a tool to do visual comparison across the Internet. We developed the Flicker comparison tool [28] in response to this problem. Another approach [35] uses image warping and image blending of two gel images copied from the Internet with local Khoros-program software modules. However, that approach requires installation of this software and using specific hardware, which may be difficult for some users.

6.1.3 Image Flicker – a method to compare images visually

Image flickering is the rapid alternate display of two images which overlay the same visual space. The history of flickering includes various implementations including optical-mechanical and computer flickering methods. It was probably first used in astronomy, with later uses in military intelligence analysis, 2-D gel analysis comparisons, and other domains. Images are locally aligned by moving one image with respect to the other while flickering. Images appear to fuse together, enhancing differences, if (i) the Flicker rate is adjusted (0.1–1 s) for the type of material being used and the individual user, (ii) the user is reasonably close to the display and (iii) local regions are well aligned, with most features aligned. Another key variation on this method is to use differential flicker, which displays one image longer on screen than the other. This is useful for comparing light and dark samples.

6.1.4 Sources of data for flicker comparison

Because of the dynamic capabilities of Java and the Internet it is possible to read data from multiple databases in the same comparison. These images can reside on the investigator's own local computer (through its FTP or WWW server). Data can reside on different WWW sites. Figure 2 illustrates this distribution of resources. The 2DWG can serves as an enriched reference source of 2-D gel image data for Flicker, providing the direct connection between Flicker and this distributed data.

6.1.5 Robustness of Java and WWW browsers

Early versions of Java and WWW browsers were not as robust as we would like, but they are improving. The speed of the Java interpreter currently running in WWW browsers is slow, but new releases are faster. Image processing uses a large amount of memory and compute power. However, larger and faster computers are becoming available at lower cost.

6.1.6 Enhancements of the Flicker comparison technique

Flicker is being extended in a number of ways. These include (i) the quantitation (manual and automatic) of objects such as spots or bands, (ii) automatic alignment of spots, bands or objects between gels or other images, *etc.*, (iii) the integration of Internet databases so users can interact directly with them, (iv) adding more image processing transforms for enhancement prior to comparison, (v) better integration of Flicker with meta-databases. Recently, we added a tool to the Flicker server for WWW users to create their own Flicker WWW page. The created WWW page is returned to the users' WWW browsers where they may save it locally and install it on their own WWW server using images from that computer as data. When invoked, it runs the NCI-Flicker server, on their data. More information on running Flicker may be found in the on-line "Flicker Reference Manual" at http://www-lecb.ncifcrf.gov/flicker/flkInfo.html.

6.2 Glossary of 2DWG table headings

Tissue or Organelle – tissue or organelle of origin of sample. If there are subcategories, they are appended to the right (*e.g.* lymphocyte-T, *etc.*).

Species – species of the sample.

"Cell line" – cell line or strain of sample if applicable.

Image URL – URL to the gel image. Currently ftp:// and http:// protocols and GIF, JPEG or TIFF images are accepted.

DB URL – URL to the specific database where this image resides and which may discuss this data in more detail. This database may also include spot maps of identified proteins.

Isotope / stain / AB – detection method. Typically, isotope implies autoradiographs or phosphor-imaging (in which case the radioisotope is mentioned, *e.g.* [^{35}S]Met); stain (*e.g.* Coomassie blue or silver for silver stain); and Ab is antibody (*e.g.* anti-PY, anti-p53, *etc.*).

CA/IPG – first-dimension method: carrier ampholytes or immobilized pH gradients.

IEF/NEPHGE – type of gel: isoelectric focusing or nonequilibrium pH gradient electrophoresis.

pH range – the isoelectric pH range if known. If the range is specified as 8–4, then this means that acid is on the right, the default is acid on the left (*e.g.* 4–8).

M_r range – the molecular mass (kDa) range if known. If the range is specified with "mwm", this indicates that molecular weight markers are used to specify the range.

Lab / Org / Comp – the laboratory, organization or company where the WWW database resides. This entry is linked to the top level WWW page for that organization.

Scan/synthesized/diagram– whether the image is a scan of a single gel, a composite of a number of gels, a synthetic gel, or a diagram of spot positions.

Map/raw data– whether the image is a raw gel image or a spot map (with proteins identified by name or number indices used by the particular DB scheme). If a URL is available for the spot map, then it is provided. Note that for some databases, although a map exists, you must track down the spot mapping of numbers to protein identifications in associated papers. In other databases, there are active gel images where the user can click on an image to look up the database contents (if any) for that spot.

Miscellaneous – additional information about the gel, sample, *etc.* This should complement the other fields so the entry is sufficient to identify the material.

2DWG ID # – unique identification number WG00001, WG00002, ... for the 2DWG database assigned by the 2DWG during the submission process.

6.3 Setting up a 2-D gel database on the WWW

In the past, setting up a 2-D gel database on the WWW was difficult and required special expertise. However, a number of methods and tools have become available to help make publishing this data easier. We will explore two methods for creating dynamic gel maps: client-side image maps [26] and server-side maps using the dbEngine [30] to publish spreadsheet tables of protein data. Others, including the ExPASy group, will be offering 2-D gel WWW publishing software as well. In both cases the database should have a clickable map. However, client-side image maps have no easy mechanism to meet all of the requirements of the ExPASy federated 2-D gel database criteria. This mechanism is not as flexible or powerful as the server-side mechanism. There are many commercial publishing tools available for generating WWW HTML. Some of these also have the capability of generating client-side image maps. We present this mate-

rial on the general methodology to encourage groups to publish their data on the WWW.

In all cases, the database publisher needs to write HTML documents. This procedure is described in any good documentation on HTML and [36] is a typical reference describing HTML 3.2. You will need a tool to edit HTML. If you learn the underlying HTML, you can do this using any text editor. However, you do not need to learn HTML since commercial publishing tools such as Netscape 3.0 Gold are WYSIWYG (what you see is what you get) editors. However, we find that learning HTML is relatively straightforward and that editing some of the hypertext with a simple text editor gives you somewhat more control in creating the WWW site.

The general structure of information in a good 2-D protein database WWW site is flexible. There are no rigid design rules. However, there is some information which should always be provided in the site. We will not go into the specific organization of a site since that depends on many things, including the types of material. Without providing a specific checklist here, a site's author should attempt to include information related to each gel so users of the database can understand the conditions under which the gel was run. Many of the fields specified in Section 6.2 should be included. You might want to visit some of the 2-D gel WWW databases listed in Table 2 to get additional ideas.

We now present some of the HTML details specific to publishing image dynamic maps. An image map (such as a 2-D gel image map) is an active image viewable in the WWW browser with specific areas which respond to clicking with the mouse. In terms of 2-D gel spot maps, clicking on a spot should cause information on that spot to be returned from the WWW server. We now present examples of the two mechanisms for doing this.

6.3.1 Client-side image maps

Client-side image maps are a mechanism of HTML version 3. Earlier server-side versions of clickable WWW image maps required the image map be maintained in the server system file areas. When a user clicked on an image, a request was sent to the WWW server to service the request and return data. The problem here is that the database author may not have ready and continued access to a WWW server's system disk area. The client-side map mechanism on the other hand allows the mapping to be specified in the HTML and executed from the WWW browser rather than from the WWW server. We will now illustrate some of the details of a typical client-slide WWW page used for a 2-D gel image map.

The image map WWW page requires a named image map denoted by the <MAP> tag (a tag is a special HTML identifier). This tag contains an attribute NAME which is set equal to some map name you decide (in this case `2DgelMap`). The SHAPE attribute is the designation of how coordinates in the COORDS attribute are to be interpreted. Shapes include "circle", "rect" (rectangle) and "poly" (polygon). We suggest, the circle which is the simplest with COORDS=x,y,radius. It is easy to estimate the spot centroid coordinates using many of the common PC desktop publishing programs. The author specifies a <IMG SRC> tag with the name of the 2-D gel map image (GIF format is generally used). Next, the USEMAP attribute is used to specify the name of the map just defined. Finally, for each map entry corresponding to a protein spot, you need to create an HTML file. For the example below, these would be protein-1.html,protein-2.html, *etc.* (or whatever names you want to use). The information you want to publish for that spot would be in each of these latter HTML documents. The advantage of this method is its simplicity. The disadvantage is that you have to create a separate **protein-n.html** for each protein *n*, which can be a great deal of work for a large number of proteins. In addition, there is no simple way to search for specific proteins or to meet all of the federated 2-D gel database criteria. One way to partially get around this search problem is to make a list of proteins by name in a WWW page in your server and to use the WWW browser Find String capability to find entries in the current browser page.

```
<MAP NAME="2DgelMap">
<AREA SHAPE="circle" COORDS="93,193,10" HREF="protein_1.html">
<AREA SHAPE="circle" COORDS="61,182,10" HREF="protein_2.html">
<AREA SHAPE="circle" COORDS="97,35,10"  HREF="protein_3.html">
<AREA SHAPE="circle" COORDS="94,130,10" HREF="protein_4.html">
    etc...
</MAP>

<IMG SRC="2dGelImage.gif" USEMAP="#2DgelMap">
```

6.3.2 Use of dbEngine on spreadsheet tables

The dbEngine [30] used in 2DWG has been used to publish other types of databases including several table and 2-D gel image-oriented protein databases. As a CGI-BIN program, dbEngine supports server-side image maps. The dbEngine is a simple database search engine, used to publish spreadsheet-type data on the WWW. We will not go into the details here since they are described in detail in the paper and on our WWW server (http://www-lecb.ncifcrf.gov/Software/dbEngine.html). However, we will describe some of the files which need to be defined. The advantage of using a database engine with a few general files over the client-side method is that a database with many proteins will require fewer file.

Typically, we install the protein database in a WWW server directory dedicated to that database. For example, the phosphoprotein 2-D gel database http://www-lecb.ncifcrf.gov/phosphoDB/ of protein changes related to the cell cycle. The database consists of a table of proteins and links to other data, gel images, clickable gel maps linking the image to the table, derived images showing a spot's region on the gel for any spot in the database.

As an example, let the name of the database be **xxx**. We need to create a set of database files with the **xxxDB** prefix, which have several specific file extensions. This is discussed in detail in the dbEngine paper and reference manual. The database will also have a home page called **index.html** which includes a description of the database, links to a clickable 2-D gel map page, a type-in form for

keywords to search the database using dbEngine, references, and any other relevant material and hypertext links. Some typical HTML for the dbEngine search query for this home page would be:

```
<H2>Search The XXX Protein Database</H2>
The database may be searched to find entries matching a key phrases in
any of the data for that entry.  Search the database by specifying
search terms below (one per entry). Each entry is searched for the
<I>conjunction</I> (i.e. AND) or <I>disjunction</I> (i.e. OR) of the
terms. [Note <I>terms</I> may be any part of a database entry but may
not include any spaces.]
<P>
<FORM METHOD="POST" ACTION="/cgi-bin/dbEngine?xxxDB,search">
<INPUT TYPE="reset" VALUE="Reset form">
<INPUT TYPE="submit" VALUE="Search Database"> <BR>
<INPUT TYPE="checkbox" NAME="DB.useHTML3Table" VALUE="on" CHECKED>
Present results as a table
<BR>
Enter search terms (you may use either <B>AND</B> or <B>OR</B> term
connectives):
<BR>
<INPUT NAME="DB.grep" SIZE=55>
```

Although specific to the 2DWG, Fig. 4 shows approximately how the WWW browser would render this HTML. As you can see from the HTML it requires that dbEngine be installed in the WWW server's cgi-bin directory. Specifying the HTML to create a clickable 2-D gel map page is also relatively straightforward and is shown below. In this example, the name of the database is xxxDB and the name of the gel image used to map the x,y coordinates is gelImage.gif. Then the resulting HTML to define an image map is:

```
<A HREF="/cgi-bin/dbEngine/xxxDB,ismapTable,gelImage">
<IMG SRC="/xxxDB/gelImage.gif" ISMAP></A>
```

The other requirement for setting up the map is a little more difficult. It involves setting up a file called xxxDB.map, described in the paper and reference manual. Spreadsheet table data for the database is stored in a file called xxxDB.txt. This can be prepared on any spreadsheet program such as Excel or database program such as dBase-IV, *etc.* Typical columns include: xxxDB ID # – Unique protein database identifier XX0001, XX0002, ...; Protein Name – common or EC name; MW – apparent molecular weight (kDa); p*I* – apparent isoelectric point; Draw Spot – draws a region around the indicated spot in gelImage map image; Other ID # – found in another WWW database; Swiss-Prot AC – Swiss Protein database accession number; and Response – percent change of protein relative to some standard.

Some of these field entries ("Other ID #" and "Swiss-Prot AC") can be translated by dbEngine to hypertext links to external databases through a dbEngine mapping file called xxxDB.f2u (maps fields to URLs). It adds a field identifier in the database to the end of the corresponding base URL address. Of course, the WWW servers must exist which support this convention. Federated 2-D gel servers have this capability. Simplified examples of dbEngine database files can be found in the on-line documentation on dbEngine or by contacting the author for more information.