REVIEW

Automation, parallelism, and robotics for proteomics

Gil Alterovitz[1, 2, 3], *Jonathan Liu*[4], *Jijun Chow*[4] *and Marco F. Ramoni*[1, 2, 3]

[1] Division of Health Sciences and Technology (HST), Harvard Medical School and Massachusetts Institute of Technology, Boston, MA, USA
[2] Children's Hospital Informatics Program at HST, Boston, MA, USA
[3] Harvard Partners Center for Genetics and Genomics, Harvard Medical School, Boston, MA, USA
[4] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA

The speed of the human genome project (Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C. *et al.*, *Nature* 2001, *409*, 860–921) was made possible, in part, by developments in automation of sequencing technologies. Before these technologies, sequencing was a laborious, expensive, and personnel-intensive task. Similarly, automation and robotics are changing the field of proteomics today. Proteomics is defined as the effort to understand and characterize proteins in the categories of structure, function and interaction (Englbrecht, C. C., Facius, A., *Comb. Chem. High Throughput Screen*. 2005, *8*, 705–715). As such, this field nicely lends itself to automation technologies since these methods often require large economies of scale in order to achieve cost and time-saving benefits. This article describes some of the technologies and methods being applied in proteomics in order to facilitate automation within the field as well as in linking proteomics-based information with other related research areas.

Received: January 24, 2006
Revised: April 6, 2006
Accepted: April 7, 2006

Keywords:
Multiplexing / Proteomics methods / Protein chips

1 Proteomics and automation

Robotics and intelligent systems technologies can help overcome technological hurdles such as speed, cost, and precision in large-scale biological endeavors. High-throughput automation is also useful to capture higher quality snapshots of cellular activity. Researchers are becoming more reliant on modern automated systems and robotics to find therapeutic targets in the proteome. Robotics increases laboratory efficiency by reducing contamination and human error (see Fig. 1). It also can help in making sense of raw data, allowing researchers to concentrate on other tasks [1, 2]. The need for applying automation in proteomics is increasing daily as larger and more complex proteome datasets are being generated. For example, innovation among drug companies has progressed to the point where automated methods of targeting

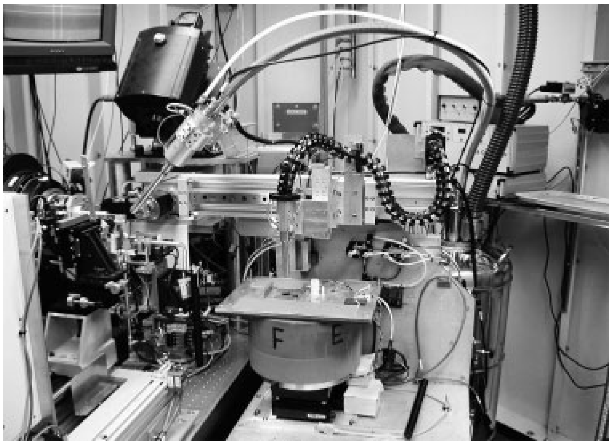

Figure 1. Automated robot used to mount and align protein crystals at Berkeley Lab Advanced Light Source. Reproduced, with permission, from Thomas Earnest (LBL) [40].

Correspondence: Dr. Gil Alterovitz, Bioinformatics Core, Harvard Medical School, New Research Building, Room 250, 77 Ave Louis Pasteur, Boston, MA 02115, USA
E-mail: ga@alum.mit.edu
Fax: +1-617-525-4488

Abbreviations: CD, compact disk; **LIMS**, Laboratory Information Management Systems; **SPR**, surface plasmon resonance; **Y2H**, yeast two-hybrid and screening have replaced traditional, manual techniques [1]. Moreover, in order to make whole-organism proteome-based experiments a reality, machine learning techniques, such as predicting the functional implications of certain post-translational alteration, are becoming an indispensable part of the research [1]. Thus, automation is permeating into many areas of laboratory research: from assays to statistical analysis.

This article focuses on automated technologies in the three areas of proteomics research: (i) 2-DE and MS, uses traditional tools in new applications, providing comprehensive analysis of proteins; (ii) array-based proteomics builds upon the same concept behind the widely used cDNA microarray technology in genomic research [3]; (iii) protein structure and imaging, uses information about 3-D structure and interaction to provide a complete picture of protein character [4].

These technologies are currently being applied to study cellular processes and regulations in high-throughput formats. Unfortunately, many current automation schemes generate large quantities of unwieldy data. Researchers are therefore relying on Laboratory Information Management Systems (LIMS) to assist them with the extraction of useful information [5–7].

Improvements have been made since the first generations of LIMS used in genomics research. Newer incarnations of LIMS have attracted considerable commercial interest such as Decodon with its Protecs and Bruker Daltronics with its ProteinScape database. Genologics, with its ProteusLIMS system is comprised of four capabilities: lab management, instrument and data integration, bioinformatics and data management, and analytics and reporting (see Fig. 2). A comprehensive list of available current LIMS packages may be found at The LIMSource (www.limsource.com). LIMS optimization is ongoing, as evidenced by Modas, a current project under development by nonlinear dynamics. It supports 2-DE and MS, integrating analysis with LIMS into one package. The LIMS software will be key to being able to take full advantage of automation advancements.

2 Automated gel electrophoresis and MS

2-DE is used for the characterization and separation of protein samples on the basis of charge and molecular weight. Its applications have included investigating protein quantity

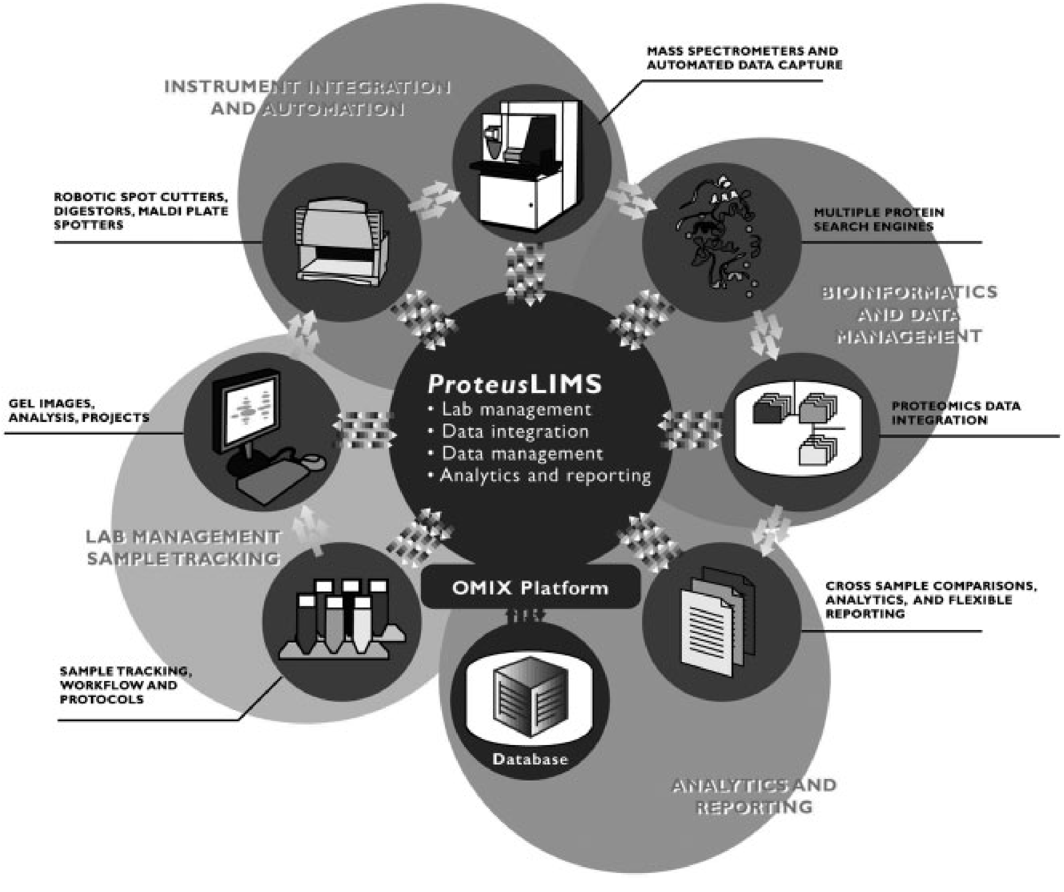
INSTRUMENT INTEGRATION AND AUTOMATION
MASS SPECTROMETERS AND AUTOMATED DATA CAPTURE
ROBOTIC SPOT CUTTERS, DIGESTORS, MALDI PLATE SPOTTERS
MULTIPLE PROTEIN SEARCH ENGINES
BIOINFORMATICS AND DATA MANAGEMENT
GEL IMAGES, ANALYSIS, PROJECTS
ProteusLIMS
• Lab management
• Data integration
• Data management
• Analytics and reporting
PROTEOMICS DATA INTEGRATION
LAB MANAGEMENT SAMPLE TRACKING
OMIX Platform
CROSS SAMPLE COMPARISONS, ANALYTICS, AND FLEXIBLE REPORTING
SAMPLE TRACKING, WORKFLOW AND PROTOCOLS
Database
ANALYTICS AND REPORTING

Figure 2. Four functional areas of the ProteusLIMS system illustrating the typical proteomics workflow. Reproduced, with permission from GenoLogics.

and character, identifying PTMs and generating proteome maps [8]. While 2-DE is extremely useful, it suffers from several technical limitations. These include the manual handling of gels. In addition, 2-DE is time consuming. It can take days to run and analyze a single gel [9].

The quest for high-throughput formats has become important for protein characterization. One of the developments in automated production of 2-DE has been marketed by commercial entities such as NextGen Sciences and Large Scale Biology. The a2DEoptimizer by NextGen Sciences features automated gel casting, saving money for core facilities that previously bought precast gels. It can run IEF at up to 10 000 V, cast multiple gels simultaneously, and is completely controlled and monitored by computer. NextGen Sciences claims that this improves 2-DE by increasing reproducibility and enhancing resolution of protein separations (see Fig. 3). It also has the ability to create user-defined gradient gels. Such gradients can be difficult to create manually.

Large Scale Biology, under their subsidiary, Predictive Diagnostics, has released BAMF (Biomarker Amplification Filter), a computer platform combining 2-DE, NMR, MS, and biomarkers to identify individual proteins. According to the company, BAMF successfully diagnosed 100% of ovarian cancer patients using a dataset from the National Cancer Institute.

There are several features that are commonly offered by many of the newer automated gel processing systems including the ability to: (i) import and export gels into standard bit-mapped graphics formats; (ii) manipulate, preprocess, filter, and organize gel bitmaps; (iii) visualize and compare gels; (iv) create, queue, and monitor computational analysis tasks; and (v) present results (*e.g.*, peptide matches in an excised, digested protein spot) [10].

A new generation of 2-DE image analysis packages is featuring new flexibility and automation. These systems include Progenesis by Nonlinear Dynamics, Decyder from GE Healthcare, Delta2D from Decodon among many others.

To overcome the calculation-intensive process of image analysis of 2-DE gels, Dowsey *et al.* [10] introduced Grid-enabled cluster computing and the proTurbo framework in the newly updated ProteomeGRID bioinformatics pipeline (see Fig. 4). ProteomeGRID, a high-throughput 2-DE image-analysis computing platform now utilizes a gel-matching

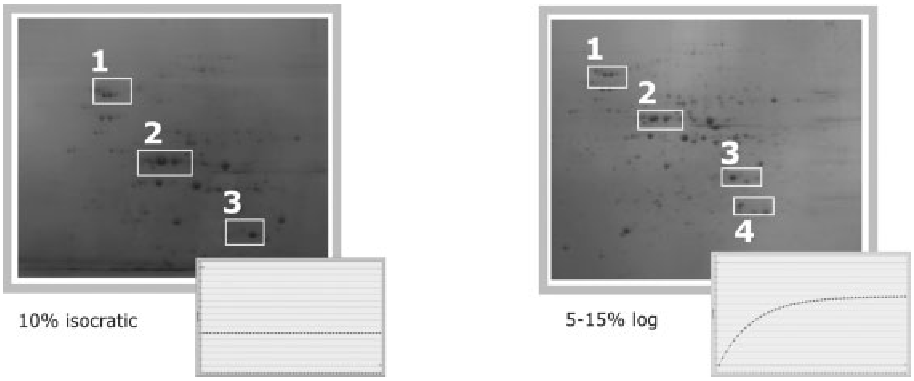

Figure 3. Using the a2DEoptimizer, these images show the effect of a suitable acrylamide gradient on the resolution of proteins in the second dimension of separation. Note that with this particular gradient, the small molecular weight proteins are better resolved than in homogenous gel. This enables enhanced spot detection and identification when processed by MS. Reproduced, with permission, from NextGen Sciences.

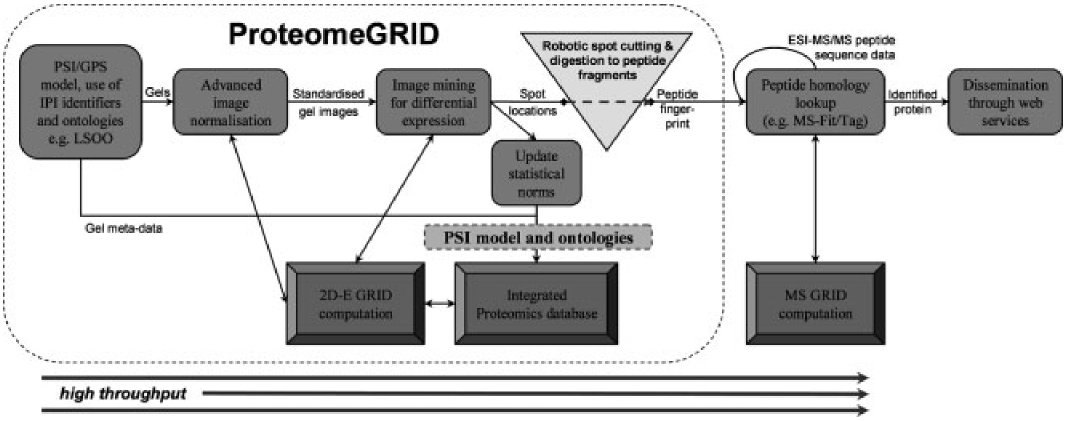
ProteomeGRID
PSI/GPS model, use of IPI identifiers and ontologies e.g. LSOO
Gels
Advanced image normalisation
Standardised gel images
Image mining for differential expression
Spot locations
Robotic spot cutting & digestion to peptide fragments
Peptide finger-print
ESI-MS/MS peptide sequence data
Peptide homology lookup (e.g. MS-Fit/Tag)
Identified protein
Dissemination through web services
Update statistical norms
Gel meta-data
PSI model and ontologies
2D-E GRID computation
Integrated Proteomics database
MS GRID computation
high throughput

Figure 4. A schematic illustration of the ProteomeGRID pipeline. ProteomeGRID is a high-throughput computing platform designed for image analysis of 2-DE. Reproduced, with permission from [5].

algorithm to overcome the bottleneck of spot matching. Researchers are currently working on implementing expression quantification as well.

An important feature common to automated approaches is the use of standards for interoperability. For example, ProteomeGRID was designed to follow the HUPO Proteomics Standards Initiative General Proteomics Standards (HUPO PSI GPS) ontology for image mining [10].

MS and sample preparation are undergoing automation as well. SELDI MS allows parallel processing by using several spots at once on arrays [11]. Ciphergen has manufactured a platform using SELDI technology to identify protein biomarkers. With integrated hardware, software and arrays, the ProteinChip® System Series 4000 boasts four modes including biomarker discovery, purification, identification, and biomarker assays. Research in this area has shown that integrating software and robotics-based hardware into a pipeline can yield benefits in terms of increased signal quality, reduced labor time, and lower costs [12].

By using sample volumes on the order of micro- or nanoliters and capillary force, technologies such as MALDI and peptide sequencing are shrinking to the size of a compact disk (CD). These CDs can processes 96 protein samples in parallel [13]. An example of this technology is the MALDI SP1 CD Microlaboratory made by Gyrolab. The SP1 processes up to 96 protein digests at once, detecting peptides with as low abundance as a femtomole. It offers parallel processing and automation at the nanoliter scale, improving reproducibility, and reducing sample loss.

Researchers at Cranfield University have developed the Genome Annotating Proteomic Pipeline (GAPP), currently in its beta release [14]. It is a scalable, grid-based pipeline designed to analyze MS data automatically. Linking users to a virtual database allows them to access GAPP's supercomputers to do time-consuming computations over a secure connection.

As automation and miniaturization develops, fully automated protein analysis may be available on a single chip [1].

3 Array-based proteomics

The term protein microarray [15–18] can have many types of instantiations. The core, common idea is that operations occur in parallel and are often miniaturized. Protein arrays typically operate by using immobilized proteins on surfaces such as glass, membranes, and beads (or other particles). The surface chemistry is designed to fix the proteins in place (see Fig. 5). Binding molecules are exposed to the array and seek out their unique target proteins. The binding of a target protein with a binding molecule is signified by a detection system. Their advantages include speed, high specificity, and the great amount of information derived from one test.

Protein microarray technology can be broken into several categories including functional arrays, detection arrays, and RP arrays [19]. Functional arrays can be used to study protein properties and interactions [20]. This tool can be used to predict functions based on coexpression after analyzing small molecule and drug-binding properties. In creating chips, individual ligands (peptides, antibodies, *etc.*) are spotted onto a surface. Protein expression is then differentially analyzed *via* biochemical interactions, providing potential, functional, and binding-based information [4, 19].

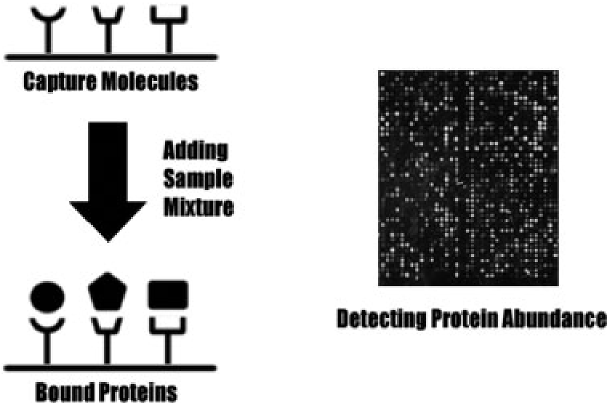

Figure 5. Protein array detection system – protein abundance is measured using labeled markers binding specifically to their substrate. This technique is the backbone of microarray technology.

Detection arrays operate in an almost opposite manner. In these, antigens or antibodies are immobilized to a surface and proteins are exposed to them. This technique is useful when assaying antibodies and monitoring protein expression [19]. Commercial applications of antibody arrays include the Whatman FAST® Quant system. Each kit contains 64 arrays of eight to ten mAb with affinities to human or mouse cytokines. It features parallel processing, performing over 500 measurements from 56 samples. Furthermore, quantitative analysis, using ArrayVision™ FAST can be performed minutes after scanning.

The Panorama™ Ab Microarray Cell Signaling Kit by Sigma-Aldrich features 224 different antibodies spotted onto 4 × 8 glass slides (see Fig. 6). By minimizing the background noise, the kit can analyze and detect minute quantities of protein – as low as a few nanograms *per* milliliter. In less than 5 hr, the relative abundance of several hundred proteins may be determined with less than 10% variation in spot morphology.

BD Biosciences BD Lyoplate™ Technology allows plate design flexibility. This is a custom service that takes user-defined specifications and prepares specialized plates that contain antibody and other reagents in a GMP environment.

In RP arrays [21], cell lysates are immobilized to a surface and probed with antibodies, resulting in a multiplexed output [19]. This also allows the analysis of modified proteins and is often used in profiling cancer [22].

New developments in protein chip format and methodology has resulted in the commercially available ProteinChip technology from Ciphergen (it is also considered an MS

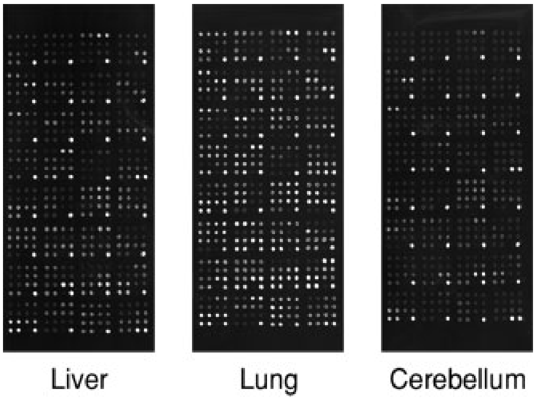

Figure 6. Protein expression differences within mouse tissues using Panorama™ Ab Microarray Cell Signaling Kit. One microgram of each protein extract from different tissues was labeled with Cy3 and incubated on separate arrays. Reproduced, with permission, from Sigma-Aldrich.

automation technology and was noted in Section 2 of this review). ProteinChip is a parallelized approach that can provide information on protein structure, character, and PTMs [23]. It has been widely used for clinical biomarker discovery [24].

Using microfluidics technology, chips can be etched with microscopic channels in which miniature assays are performed [2, 25]. Caliper Technologies has developed one such device called the LabChip® 3000 Drug Discovery System. The LabChip 3000 includes many functions and has been most commonly used for kinase profiling [26]. For channels 10–50 μm in diameter used in the present study, less reagent was required and fewer sample cells were necessary.

The effects of protein microarrays have been seen across several fields such as medicine and pharmacology. In future, blood tests may be performed by providing fewer drops of blood onto a chip with specific protein markers, providing valuable diagnostic and real-time prognostic information [20]. Advantages of the protein microarray approaches include the ability to use small samples and built-in experimental controls for calibration.

Another avenue of protein identification and screening is yeast two-hybrid (Y2H) [27]. Y2H involves the measurement of physical interactions. Using an easily differentiable reporter gene, hybrid transcription factors are made. If the two hybrid transcription factors react, the reporter gene is transcribed and the cell will thus phenotypically indicate the interaction. Advantages of Y2H include its ability to be done *in vivo* and ease with which parallelization can be implemented to test many interaction combinations quickly.

The tandem affinity purification (TAP) [28] method uses a TAP-tag to target specific proteins. The proteins are then purified and further analyzed *via* SDS-PAGE, MS, and functional assays. These interactions must be completed, for the most part, *in vitro* and are time consuming, requiring the addition of many reagents. There is, however, much room for optimization.

4 Structure and imaging

The determination of protein 3-D quaternary structures has greatly increased in the past few years. Analysis of the large number of proteins requires new high-throughput and multiplexed techniques [29]. Automated systems can help increase reliability, objectivity, reproducibility, and repeatability.

The most common techniques for structural analysis include NMR, X-ray crystallography, structure prediction methodology, and even MS to a certain degree [30]. These techniques allow researchers to determine or constrain the potential 3-D structure of proteins and protein subunits.

Imaging proteomics can also detect protein–protein interactions and protein localization. Techniques include transfected cell arrays, green fluorescent protein-based (GFP) labeling, and fluorescence resonance energy transfer (FRET) [31, 32]. FRET is used to detect distance-dependent interactions by using an energized fluorophore that can be transferred to an acceptor less than 100 Å away [33].

Multiplexed surface plasmon resonance (SPR) [34, 35] is another automated approach to the quantitative analysis of protein interactions. SPR is a technique that studies bioaffinity on gold and noble metal thin films. It can provide information on concentrations in a solution several hundred nanometers above the film by detecting changes in optical properties that result when proteins in the solution interact [36]. Advantages of SPR include its low target consumption and freedom from radioactive labeling [37].

VEGA ZZ, a molecular modeling software package, is designed for use by researchers analyzing protein structures. It has an extensive list of features including multiple file format support, atomic potential attribution, 3-D molecular editor, and a protein–protein docking system. Figure 7 shows the M1 muscarinic receptor–acetylcholine complex (MEP surface and tube) as explored with VEGA [38].

Despite great improvements in the automation of structural proteomics, there remain numerous challenges in the field. For example, when analyzing proteins *via* X-ray diffraction, proteins must be able to crystallize. This process is largely based on trial and error – as many factors such as pH and salt concentration are involved [39]. Commercial solutions have been developed, including the Index™ product from Hampton Research, which can screen for crystallization based on factors such as pH, salt, and ionic strength. New techniques in automation and related technologies are thus not only facilitating experiments in terms of speed, but also making a new type of large-scale proteomics research possible.

5 Conclusion

Proteomics is leveraging some of the automated methodologies developed for other fields such as functional genomics. However, some challenges that are new and proteomics-

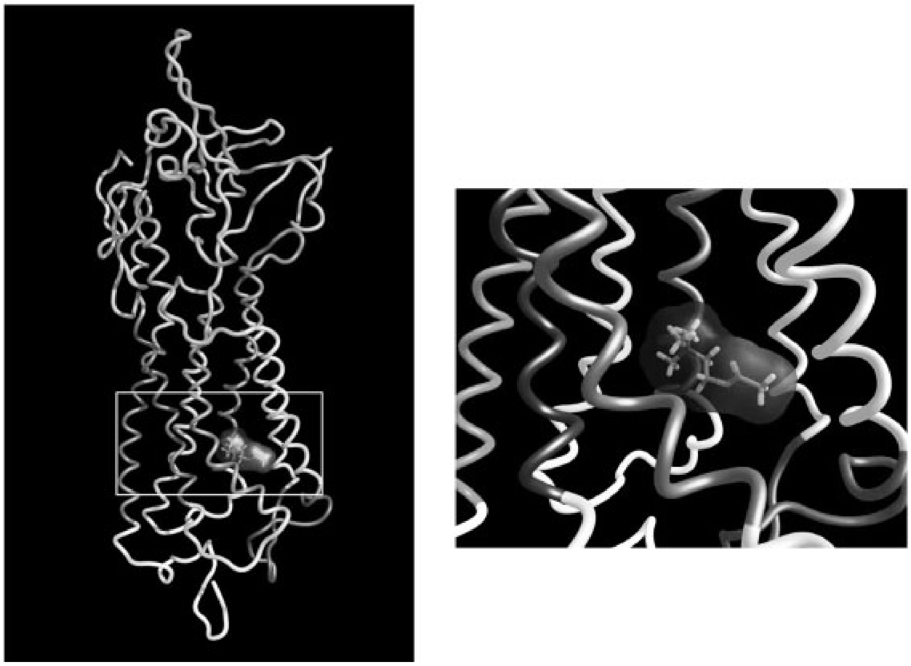

Figure 7. M1 muscarinic receptor–acetylcholine complex as explored with VEGA. An enlarged view of the boxed interaction to the right. Reproduced, with permission, from Alessandro Pedretti.

specific (such as post-transitional modifications) may benefit from custom-designed solutions. In addition, automation in future may be able to take advantage of the synergistic developments in other fields – by linking to that data. Thus, one area for future work involves the design of interfaces for integration of data from many heterogeneous sources in an automated manner. In future, automated and parallelized pipeline solutions that integrate genomic, proteomic, and relational information (*e.g.*, networks) may be used in biology for discoveries that would not have been possible through isolated, manual analysis.

6 References

[1] Alterovitz, G., Afkhami, E., Ramoni, M., in: Liu, J. X. (Ed.), *New Developments in Robotics Research*, Nova Science Publishers, New York, NY 2005, pp. 217–252.

[2] Chapman, T., *Nature* 2003, *422*, 665–666.

[3] Kohane, I. S., Kho, A. T., Butte, A. J., *Microarrays for an Integrative Genomics*, MIT Press, Cambridge 2002.

[4] de Hoog, C. L., Mann, M., *Annu. Rev. Genomics Hum. Genet.* 2004, *5*, 267–293.

[5] Amin, A. A., Faux, N. G., Fenalti, G., Williams, G. *et al.*, *Proteins* 2006, *62*, 4–7.

[6] Goh, C. S., Lan, N., Echols, N., Douglas, S. M. *et al.*, *Nucleic Acids Res.* 2003, *31*, 2833–2838.

[7] Turner, E., Bolton, J., *Qual. Assur.* 2001, *9*, 217–224.

[8] Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balazsi, G. *et al.*, *J. Bacteriol.* 2003, *185*, 5673–5684.

[9] Hille, J. M., Freed, A. L., Watzig, H., *Electrophoresis* 2001, *22*, 4035–4052.

[10] Dowsey, A. W., Dunn, M. J., Yang, G. Z., *Proteomics* 2004, *4*, 3800–3812.

[11] Simpkins, F., Czechowicz, J. A., Liotta, L., Kohn, E. C., *Pharmacogenomics* 2005, *6*, 647–653.

[12] Alterovitz, G., Aivado, M., Spentzos, D., Libermann, T. A. *et al.*, *Proceedings of the International Conference of IEEE Engineering in Medicine and Biology*, San Francisco, CA, USA 2004.

[13] Gustafsson, M., Hirschberg, D., Palmberg, C., Jornvall, H., Bergman, T., *Anal. Chem.* 2004, *76*, 345–350.

[14] Shadforth, I., Crowther, D., Bessant, C., *Proteomics* 2005, *5*, 4082–4095.

[15] Dupuy, A. M., Lehmann, S., Cristol, J. P., *Clin. Chem. Lab. Med.* 2005, *43*, 1291–1302.

[16] Clarke, W., Chan, D. W., *Clin. Chem. Lab. Med.* 2005, *43*, 1279–1280.

[17] Bertone, P., Snyder, M., *Febs. J.* 2005, *272*, 5400–5411.

[18] MacBeath, G., Schreiber, S. L., *Science* 2000, *289*, 1760–1763.

[19] Cretich, M., Damin, F., Pirri, G., Chiari, M., *Biomol. Eng.* 2005, *23*, 77–88.

[20] Xu, Q., Lam, K. S., *J. Biomed. Biotechnol.* 2003, *2003*, 257–266.

[21] Speer, R., Wulfkuhle, J. D., Liotta, L. A., Petricoin, E. F., IIIrd, *Curr. Opin. Mol. Ther.* 2005, *7*, 240–245.

[22] Kreutzberger, J., *Appl. Microbiol. Biotechnol.* 2006, *70*, 383–390.

[23] Merchant, M., Weinberger, S. R., *Electrophoresis* 2000, *21*, 1164–1177.

[24] Issaq, H. J., Conrads, T. P., Prieto, D. A., Tirumalai, R., Veenstra, T. D., *Anal. Chem.* 2003, *75*, 148–155.

[25] Dittrich, P. S., Manz, A., *Nat. Rev. Drug Discov.* 2006, *5*, 210–218.

[26] Schutkowski, M., Reineke, U., Reimer, U., *Chembiochem.* 2005, *6*, 513–521.

[27] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A. *et al.*, *Nature* 2000, *403*, 623–627.

[28] Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M. *et al.*, *Nat. Biotechnol.* 1999, *17*, 1030–1032.

[29] Lee, H. J., Yan, Y., Marriott, G., Corn, R. M., *J. Physiol.* 2005, *563*, 61–71.

[30] Stults, J. T., Arnott, D., *Methods Enzymolgy* 2005, *402*, 245–289.

[31] Nakanishi, J., Takarada, T., Yunoki, S., Kikuchi, Y., Maeda, M., *Biochem. Biophys. Res. Commun.* 2006, *343*, 1191–1196.

[32] Meyer, B. H., Martinez, K. L., Segura, J. M., Pascoal, P. *et al.*, *FEBS Lett.* 2006, *580*, 1654–1658.

[33] Phizicky, E., Bastiaens, P. I., Zhu, H., Snyder, M., Fields, S., *Nature* 2003, *422*, 208–215.

[34] Unfricht, D. W., Colpitts, S. L., Fernandez, S. M., Lynes, M. A., *Proteomics* 2005, *5*, 4432–4442.

[35] Besenicar, M., Macek, P., Lakey, J. H., Anderluh, G., *Chem. Phys. Lipids* 2006, DOI: 10.1016/j.chemphyslip. 2006.02.010.

[36] Alves, I. D., Park, C. K., Hruby, V. J., *Curr. Protein Pept. Sci.* 2005, *6*, 293–320.

[37] Besenicar, M., Macek, P., Lakey, J. H., Anderluh, G., *Chem. Phys. Lipids* 2006, DOI: 10.1016/j.chemphyslip. 2006.02.010.

[38] Pedretti, A., Villa, L., Vistoli, G., *J. Mol. Graph Model* 2002, *21*, 47–49.

[39] Rayment, I., *Structure* 2002, *10*, 147–151.

[40] Snell, G., Cork, C., Nordmeyer, R., Cornell, E. *et al.*, *Structure* 2004, *12*, 537–545.