Engineering in Genomics

Harold "Skip" Garner

After Sequencing: Quantitative Analysis of Protein Localization

Michael V. Boland and Robert F. Murphy, *Carnegie Mellon University*

The completion of the Human Genome Project will be no more than an important milestone in our effort to understand how billions of DNA nucleotides give rise to a human being. After accumulating raw DNA sequence information, it will be necessary for the scientific community to dedicate many more years to characterizing each gene and its protein product. Structural characterization will require the use of NMR [1], x-ray crystallography [2], and other techniques. Biochemical assays will be necessary to discover the function of each protein and to place each protein in the appropriate cellular pathway(s). Above all, these data will have to be organized and combined in ways that make them useful in further research (see [3], for example). For a variety of reasons, including the sheer size of the human genome, this information will need to be quantitative in nature so that analysis can be both automated and systematic.

One area of protein characterization that is not yet developed but which can be anticipated to be extremely useful in the post-genomic era is the study of protein localization (i.e., where within a particular cell type does one find a given protein?). As with analysis of protein structure and function, two complementary approaches to discovering the subcellular location of newly identified proteins can be envisaged: prediction and experimental determination. References addressing both predictive and experimental approaches to protein characterization are included in Table 1.

Given that the subcellular location is known even in the most general terms for only a small fraction of all proteins, the accuracy and completeness of predictive systems is currently limited. This leads to a compelling need for automated approaches to experimentally determine subcellular localizations. The starting point is a systematic, quantitative method for describing protein localization patterns.

A numerical description of protein localization is fundamentally useful because it provides information about each protein that is complementary to the sequence and structure data that are currently available, and because it obviates the need to describe protein localization subjectively, as is currently the practice. Furthermore, quantitative analysis facilitates repetition of an experiment by others and can provide a systematic approach to answering a question where otherwise there would only be subjective judgements. With numerical output from an experiment, it is possible to quantitate the confidence one has in the result, or to assess the statistical significance of results obtained under different conditions. Lastly, the possibility that automated methods may provide biological insights not readily available from visual examination of localization patterns must be considered.

Table 1. Predictive and Experimental Analysis of Protein Structure, Function, and Localization

Predictive Analysis
Sequence homology searches [4] Protein structure prediction [5, 6] Functional class prediction [7, 8] Protein localization prediction [9, 10, 11]
Experimental Analysis
Protein structure determination [12, 13] Determination of protein function - biochemical analysis Tissue/developmental expression analysis [14, 15] Determination of protein localization [16, 17, 18]

Automating the analysis of protein localization will bring some of the same benefits that have come with the automated analysis of protein and DNA sequences. In the time before quantitative sequence analysis, comparisons could only be done subjectively (i.e., the assessment of similarity between two sequences was done by looking at them). Today, on the other hand, it is possible to send a new sequence to a variety of database servers (e.g., http://www.ncbi.nlm.nih.gov/BLAST/ and http://www.ncgr.org/) and, in short order, receive a list of existing sequences that are similar to it.

Although the automated analysis of protein localization may be more complex than the automated analysis of sequences, the benefits to be derived are the same. For example, it is not far-fetched to imagine a time in which it will be possible to send images of a new protein's localization to a database and obtain a list of proteins that localize in a similar manner. The ability to identify known proteins with similar sequence *and* similar localization is a significant advance over the current state of the art. Furthermore, as we move beyond the initial stages of the Human Genome Project (i.e., mapping and sequencing), it is becoming increasingly clear that we need structural, functional, and localization information to accompany the raw sequences.

Some recent work with the goal of introducing quantitative analysis to the description of protein localization in a way that is biologically useful is described below. The usefulness of the quantitative descriptions is assessed by using those descriptions to address biologically relevant questions. Two applications developed thus far include the classification of protein localization patterns into known categories and the selection of a typical image from a set in which the images depict protein localization patterns.

The techniques used to develop these applications are drawn from the fields of fluorescence microscopy, pattern recognition, and machine learning. Images depicting the localization of a particular protein are generated by specifically labeling that protein with a fluorescent marker and then collecting images with a microscope. The availability of antibodies directed against many cellular proteins makes it possible to generate images for a wide variety of protein localization patterns. Sample microscope images are shown in Fig. 1.

Pattern Analysis

The automated analysis of images most frequently involves four common steps, regardless of the desired final results (see Fig. 2). These steps include image collection, image restoration (if necessary), image processing, and numerical feature extraction. The final step in this process can then be one of a variety of pattern-analysis techniques.

Fluorescence images to be analyzed can be acquired using any of a variety of microscope modalities followed by appropriate image restoration methods. Image restoration consists of those steps intended to turn the output from the microscope system into an image that better represents the original sample. Once the images are acquired and reconstructed, they usually need to be processed further. In this case, outcomes desired from image processing include identification and isolation of single cells in an image (segmentation) and selection of pixels of interest via thresholding.

Once the images have been processed, it is necessary to "describe" the pattern in the image numerically. This step is commonly referred to as feature extraction and involves the calculation of a small number of values (frequently 5-50) that are intended to summarize the information in the image more concisely than the individual pixels. Examples of features that might be useful for describing protein localization patterns are the number of fluorescent objects in a cell (where an object is defined to be a group of contiguous pixels that all have an intensity above a threshold value) and the average area of those objects. While such a description is clearly biased by the choice of features (e.g., the two example features described above do not take into account where the objects are in the cell), the bias is explicit and can be incorporated in subsequent analysis and discussion. Once the features have been calculated, it is possible to use them to perform a variety of pattern-analysis tasks.

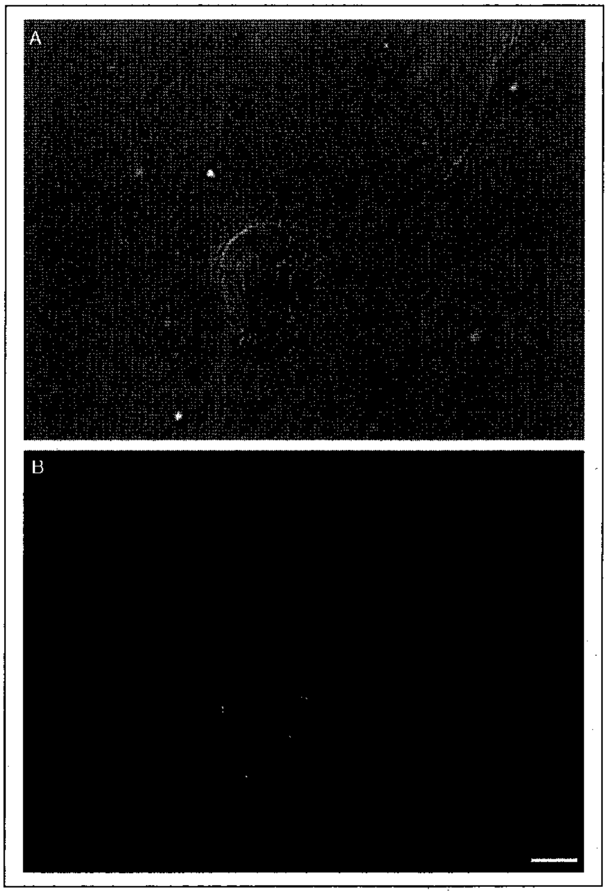
Classification of Protein Localization Patterns

Fluorescence images such as the one in Fig. 1(B) represent the localization of a single protein within a cell. By calculating numeric features to summarize these images, one is in fact describing the localization pattern of the protein. One method for assessing the biological utility of these descriptions is to develop a system that is able to recognize new images containing the same pattern.

An obvious use for this approach can be found in the field of high-throughput screening. To determine which of 1,000 potential drugs serve to prevent the translocation of a protein from one cellular compartment (the endoplasmic reticulum - ER) to another (the Golgi), it would be advantageous to have a system that can automatically recognize, and therefore classify, images depicting the fluorescently tagged protein's ER localization pattern and its Golgi localization pattern. After applying the potential drugs to cultured cells and labeling the protein of interest, only those samples that are classified as "ER" (i.e., those in which the drug has blocked trafficking of the protein) need be studied further.

Classification of protein localization patterns can also facilitate the automation of microscope function. As an example, an investigator studying the disassembly and reassembly of the Golgi apparatus during cell division [19] might wish to study drugs that inhibit the process. An automated microscope might be programmed to find all cells displaying a fluorescence pattern characteristic of a dispersed mitotic Golgi apparatus and then, for example, image those cells over time after photoactivating a caged drug in half of them (to explore the role of the drug's target in mitotic reassembly).

The feasibility of automated classification has been demonstrated by the development of a system that is able to classify the localization patterns characteristic of five cellular molecules (proteins and DNA) in Chinese Hamster Ovary (CHO) cells [16]. Images were acquired on an epifluorescence microscope after the cells had been fixed, permeabilized, and labeled with appropriate fluorescent reagents (antibodies conjugated to fluorescent dyes, and the DNA intercalating agent Hoechst 33258). The labels used were directed against a Golgi protein (giantin), a lysosomal protein (LAMP2), a nuclear protein (NOP4), a cytoskeletal protein (tubulin), and DNA. Note that the same analysis described here can be done with live cells if appropriate labels (e.g., green fluorescence protein - GFP) are available for the molecule of choice. The images were processed to remove out-of-focus and background fluorescence and cropped to allow analysis of single cells. Numerical features (Zernike moments [20] and Haralick's texture features [21]) were then extracted from the processed images. These particular features were chosen based on their invariance to image rotation and for their ability to describe a variety of patterns efficiently. They were also selected without consideration of the particular protein localiza-



1. Sample images collected from a multimode (i.e., capable of automated switching between transmitted and fluorescence illumination of the specimen) microscope. A transmitted light (differential interference contrast) image (A) is included to show the full extent of the cell. A fluorescence image (B) of the same cell depicted in A shows only those parts of the cell that were labeled with a fluorescently conjugated antibody against a mitochondrial protein. Scale bar = 10 µm.

tion patterns available at the time. While it would be easy to design features to discriminate the initial five classes, it is unlikely that a feature set tailored to those classes could be used to adequately describe new localization patterns added later.

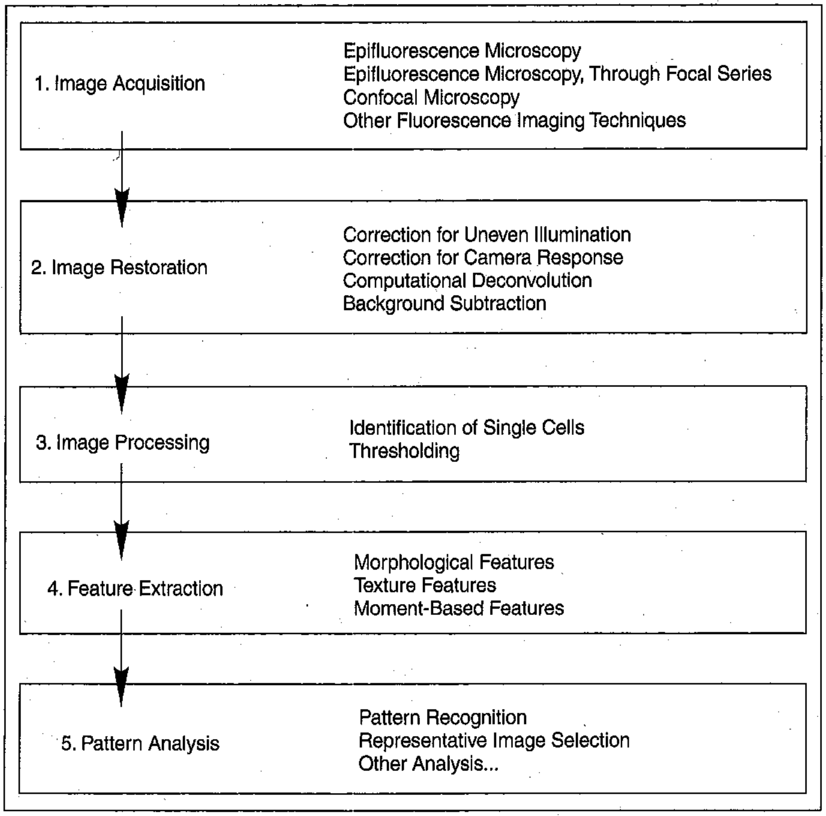
Once the images are described by the numerical features, pattern-recognition techniques are brought to bear on the problem. Briefly, once a protein localization image has been described by its features, it constitutes a single point in an n-dimensional feature space, where n is the number of features describing each image. The basic goal of pattern recognition is to generate boundaries in the feature space that separate the classes from one another. Classification schemes include statistical methods [22], decision trees [23], and neural networks [24]. A large number of such classifiers have been designed and tested with an almost equally large set of problems, including target recognition for the military, parts inspection for manufacturers, and handwriting recognition for the U.S. Post Office.

Of the classifiers we investigated for application to this problem (linear discriminant analysis, classification trees, and a backpropagation neural network), the backpropagation neural network proved to be the best at recognizing images from the five classes described above. After being trained on one subset of image feature data, a neural network was able to correctly identify an average of 88% of previously unseen images. This rate is impressive given the heterogeneity between individual cells, even within a particular class (i.e., not all giantin localization patterns are identical—there is considerable variation from one cell to another). This work is currently being extended to a set of 10 localization patterns and a new numerical feature set that is designed to utilize biological knowledge in describing the localization of the proteins. Early results indicate that this new system can be used to correctly classify a *set* of 10 previously unseen images at a rate of 99% (Boland MV, Murphy RF - manuscript in preparation). Classification of sets depends on the members of the set being drawn from a homogeneous population (e.g., cells that were labeled at the same time with the same fluorescent markers), and it is accomplished by first classifying each member of the set individually. The classification for the set is then defined to be the class to which a plurality of its members belongs. An unknown class is included to handle cases in which two classes are equally common. The same classification system is able to correctly recognize 83% of individual images.

Selection of a Representative Localization Pattern

A second application of pattern analysis to protein localization patterns is the selection of a representative from a set of patterns. This is believed to be a novel application of pattern analysis that is not limited to biological images. Investigators are regularly required to choose a single image from a collection for presentation or publication. The often unspoken implication of this choice is that the selected image is representative of the set. The problem with this approach is that there is no way to describe the biases (both conscious and unconscious) of the investigator in making the selection. Furthermore, without the benefit of quantitative analysis, this choice is not repeatable by other investigators, and subsequent results and discussion are not rigorously comparable to each other. One might argue that any numeric description of a pattern is biased by the formulas used to generate the numbers. Fortunately, however, this bias is almost always explicit, whereas the subjective biases of individual investigators are implied, at best, and unstated, at worst. Automated selection of representative images, particularly those depicting protein localization, will be useful for selecting images for publication, for summarizing the contents of a protein localization database with a handful of images, and for facilitating the analysis of experiments by providing the investigator with new insight into the data.

With these goals in mind, a system has been developed that is able to systematically choose a representative image from a set in which each image depicts the localization of a protein in a single cell [17]. Using the images and features generated for the classification study (above), methods were investigated for assigning a measure of typicality to the images in a set. The first four steps in this process (i.e., image collection, image restoration, image processing, and feature extraction—see Fig. 2) are identical to those used for classification. As mentioned above, after feature extraction, each image can be thought of as a point in n-dimensional space, where n is the number of features used to describe the image. Intuitively, the most typical (or representative) image of a set has been defined to be that image that falls closest to the middle of the distribution of im-


1. Image Acquisition
Epifluorescence Microscopy
Epifluorescence Microscopy, Through Focal Series
Confocal Microscopy
Other Fluorescence Imaging Techniques
2. Image Restoration
Correction for Uneven Illumination
Correction for Camera Response
Computational Deconvolution
Background Subtraction
3. Image Processing
Identification of Single Cells
Thresholding
4. Feature Extraction
Morphological Features
Texture Features
Moment-Based Features
5. Pattern Analysis
Pattern Recognition
Representative Image Selection
Other Analysis...

2. Overview of pattern-analysis techniques applied to cellular protein localization. General steps in the process are on the left and examples specific to the analysis described here are on the right.

ages in feature space (see Fig. 3). The distance metric used to find the image closest to the center of the distribution must be chosen wisely, however, and it turns out that statistically robust methods of calculating distance are crucial to the success of typical image selection. These robust methods [25] are able to recognize and then ignore spurious outlier points that tend to bias the estimates of the population mean and covariance (see Fig. 3). For the images used in this work, outliers consisted of localization patterns from sick or damaged cells, cells in mitosis, cells that labeled poorly, or cells not in proper focus. Since the multivariate mean is defined to be the center of the population for the purposes of distance calculation, it is important that the estimate of the mean not include features from unusual patterns. With such distance metrics in place, reasonable results were obtained for four different classes of fluorescence images [17].

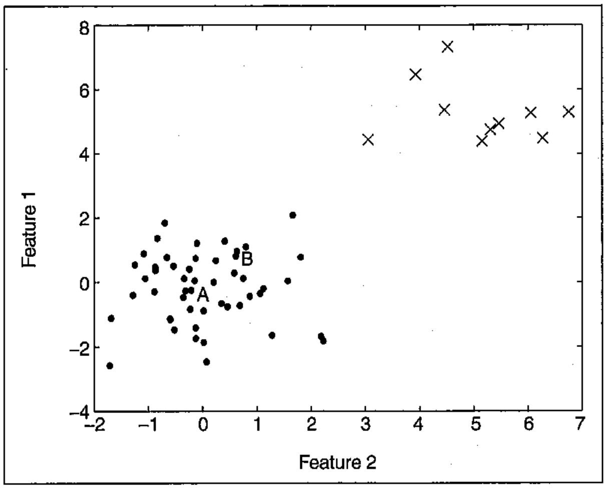
Satisfactory evaluation of the methods developed for choosing a representative pattern requires two approaches. First, a statistical test was implemented to confirm that the method performed as expected. Briefly, a chi-squared test was used to determine whether, in the case of an intentionally contaminated data set, the images of the majority class were assigned typicality scores higher than those assigned to the contaminant class (see [17] for details). Second, we tested the methods by subjectively evaluating the localization patterns ranked as most typical, asking whether (without the benefit of the automated system) we might have picked them as most typical as well. Figure 4 includes the three most typical (A, B, C) and three least typical (D, E, F) localization patterns from the set of lysosomal protein patterns used in our classification study (the images are available at http://murphylab.web.cmu.edu/data/).

In the case of these data, the most typical patterns all depict the kind of localization that one would expect from a lysosomal protein: many vesicles distributed throughout the cytoplasm. The least typical images, on the other hand, are all less than ideal for one reason or another. Images D and E both contain a single very bright object that artificially dominates the image. Image F had a weak signal to begin with and, after processing, depicts only a few objects. By using both objective and subjective approaches to testing, it was possible to take advantage of the quantitation,while being reassured by results that met biological expectations.

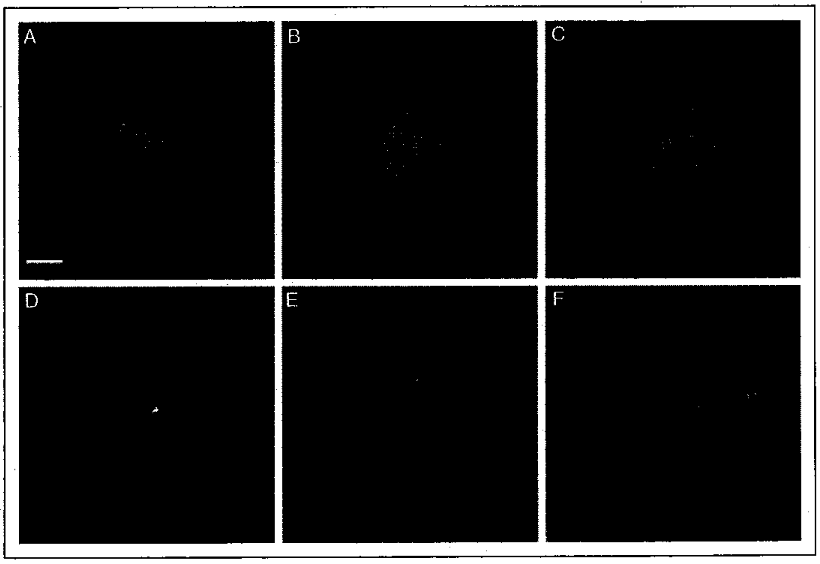
A service to demonstrate these typical image selection methods has been developed (http://murphylab.web.cmu.edu/services/TypIC/). It allows an investigator to upload a collection of images and, after processing, returns a list of the images ranked by their typicality.

Concluding Remarks

This work addresses two principles that will be integral to the post-genomic or proteomic era (i.e., after sequencing). The first is that any analysis of data from or related to the Human Genome Project will need to be designed with high-throughput in mind. Just the sequence information will encompass some 3 billion nucleotides, and that does not include information about introns, exons, promoters, and many other features of interest. The volume of information that must be synthesized is even larger than the genome itself, and it is diverse in nature. It includes sequence, structural, functional, and localization information for each gene, and each of those constituents has its own levels of organization as well (e.g., functional information for a protein can be obtained at the molecular, cellular, and organismal levels). Computational analysis must be able to handle all these data in a reasonable amount of time. The second principle, which has been alluded to here, is that analysis techniques must incorporate data from a variety of sources. Archiving and indexing of sequence data, for example, must include sequences from multiple organisms and from diseased and healthy states to be maximally useful. The other levels of information, including structure, function, and localization will need to be similarly organized.



3. The effect of outliers on the calculation of a population mean. In this simplified example, in which each image is described by two features, the presence of outliers (points marked with 'x') causes the overall mean (B) to be shifted towards the outliers. Removal of the outliers allows proper estimation of the mean of the majority population (A).



4. The most (A-C) and least (D-F) typical images from a collection of 97 fluorescence images of the distribution of LAMP2, a membrane protein found predominantly in lysosomes. Scale bar = 10 μm.

We have demonstrated two different applications of pattern analysis to protein localization patterns from cultured mammalian cells. We have further shown that those applications are biologically useful and that they are consistent with the principles just described. The classification system is able to recognize previously unseen patterns as belonging to a known class, and this will clearly be useful in the area of high-throughput screening based on protein localization. The typicality work demonstrates that it is possible to objectively select a pattern to represent a set. Objective selection of one or more representatives will be an important method for summarizing potentially large data sets containing a variety of types of information.

The results we have obtained are promising, but also very early in the development of this field. Precisely because of this immaturity, however, there are many interesting and novel questions to ask and many challenging problems to solve.

Acknowledgments

The research in this article was supported in part by research grant RPG-95-099-03-MGO from the American Cancer Society (R.F.M.), by NSF grant BIR-9217091, and by NSF Science and Technology Center grant MCB-8920118. M.V.B. was supported by NIH training grant T32 GM08208 and by NSF training grant BIR-9256343.

Michael Boland received his B.S. (1992) in electrical and computer engineering from the University of Colorado at Boulder. Since 1993, he has been enrolled in the M.D./Ph.D. program run jointly by the University of Pittsburgh School of Medicine and Carnegie Mellon University. He is currently working toward his Ph.D. in biomedical engineering at CMU, where his research interests are in the automated analysis of fluorescence microscopy images.

Robert F. Murphy earned an A.B. in biochemistry from Columbia College in 1974 and a Ph.D. in biochemistry from the California Institute of Technology in 1980. He was a Damon Runyon-Walter Winchell Cancer Foundation postdoctoral fellow with Dr. Charles R. Cantor at Columbia University from 1979 through 1983, after which he became an assistant professor of biological sciences at Carnegie Mellon University in Pittsburgh, Pennsylvania. He received a Presidential Young Investigator Award from the National Science Foundation in 1984 and has received research grants from the National Institutes of Health, the National Science Foundation, the American Cancer Society, the American Heart Association, and the Arthritis Foundation. His laboratory at Carnegie Mellon focuses primarily on the application of fluorescence methods to problems in cell biology, with particular emphasis on pH regulation in endosomes and lysosomes. He also has a long-standing interest in computer applications in biology, and he developed the Computational Biology curriculum taught at Carnegie Mellon since 1989. In 1984, he co-developed the Flow Cytometry Standard data file format used throughout industry, and he chairs the Data File Standards Committee of the International Society for Analytical Cytology. Dr. Murphy is also Chair of the Cytometry Development Workshop held each year in Asilomar, California. He is currently an associate professor in the Department of Biological Sciences and the Faculty of Biomedical Engineering at Carnegie Mellon.

Address for Correspondence: Dr. Robert F. Murphy, Center for Light Microscope Imaging and Biotechnology, Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA. Phone: +1.412.268.3480. Fax: +1.412.268.6571. E-mail: murphy@cmu.edu

References

1. **Wagner G:** An account of NMR in structural biology. *Nat Struct Biol*, 4 Suppl:841-844, 1997.
2. **Brünger AT:** X-ray crystallography and NMR reveal complementary views of structure and dynamics. *Nat Struct Biol*, 4 Suppl:862-865, 1997.
3. **Kanehisa M:** A database for post-genome analysis. *Trends Genet*, 13:375-376, 1997.
4. **Altschul SF, Boguski MS, Gish W, Wootton JC:** Issues in searching molecular sequence databases. *Nat Genet*, 6:119-129, 1994.
5. **Jones DT:** Progress in protein structure prediction. *Curr Opin Struct Biol*, 7:377-387, 1997.
6. **Finkelstein AV:** Protein structure: What is it possible to predict now? *Curr Opin Struct Biol*, 7:60-71, 1997.
7. **Bork P and Koonin EV:** Predicting functions from protein sequences-where are the bottlenecks? *Nat Genet*, 18:313-318, 1998.
8. **Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, et al.:** Predicting function: From genes to genomes and back. *J Mol Biol*, 283:707-725, 1998.
9. **Eisenhaber F and Bork P:** Wanted: Subcellular localization of proteins based on sequence. *Trends Cell Biol*, 8:169-170, 1998.
10. **Horton P and Nakai K:** Better prediction of protein cellular localization sites with the k nearest neighbor classifier. *Intell Syst Mol Biol*, 5:147-152, 1997.
11. **Nakai K and Kanehisa M:** A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, 14:897-911, 1992.
12. **Wüthrich K:** Protein structure determination in solution by NMR spectroscopy. *J Biol Chem*, 265:22059-22062, 1990.
13. **MacArthur MW, Driscoll PC, Thornton JM:** NMR and crystallography—Complementary approaches to structure determination. *Trends Biotechnol*, 12:149-153, 1994.
14. **Bassett, DEJ, Eisen MB, Boguski MS:** Gene expression informatics-It's all in your mine. *Nat Genet*, 21:51-55, 1999.
15. **Bowtell DD:** Options available—from start to finish—for obtaining expression data by microarray. *Nat Genet*, 21:25-32, 1999.
16. **Boland MV, Markey MK, and Murphy RF:** Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry*, 33:366-375, 1998.
17. **Markey MK, Boland MV, and Murphy RF:** Towards objective selection of representative microscope images. *Biophys J*, 76:2230-2237, 1999.
18. **Spector DL, Goldman RD, Leinwand LA:** *Cells: A Laboratory Manual: Subcellular Localization of Genes and Their Products*, Vol. 3. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1998.
19. **Jesch SA and Linstedt AD:** The golgi and endoplasmic reticulum remain independent during mitosis in hela cells. *Mol Biol Cell*, 9:623-635, 1998.
20. **Khotanzad A and Hong YH:** Rotation invariant image recognition using features selected via a systematic method. *Pattern Recognit*, 23:1089-1101, 1990.
21. **Haralick RM:** Statistical and structural approaches to texture. *Proc IEEE*, 67:786-804, 1979.
22. **Duda RO and Hart PE:** *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.
23. **Breiman L, Friedman J, Olshen R, and Stone C:** *Classification and Regression Trees*, Wadsworth and Brooks/Cole, Monterey, CA, 1984.
24. **McClelland JL and Rumelhart DE:** *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 2. The MIT Press, Cambridge, MA, USA, 1986.
25. **Rocke DM and Woodruff DL:** Identification of outliers in multivariate data. *J Am Stat Assn*, 91:1047-1061, 1996.