Automated Zone Correction in Bitmapped Document Images

Susan E. Hauser, Daniel X. Le, George R. Thoma

Lister Hill National Center for Biomedical Communications
National Library of Medicine
Bethesda, Maryland
hauser@nlm.nih.gov

ABSTRACT

The optical character recognition system (OCR) selected by the National Library of Medicine (NLM) as part of its system for automating the production of MEDLINE® records frequently segments the scanned page images into zones which are inappropriate for NLM's application. Software has been created in-house to correct the zones using character coordinate and character attribute information provided as part of the OCR output data. The software correctly delineates over 97% of the zones of interest tested to date.

Keywords: Document Image Analysis, Page Segmentation.

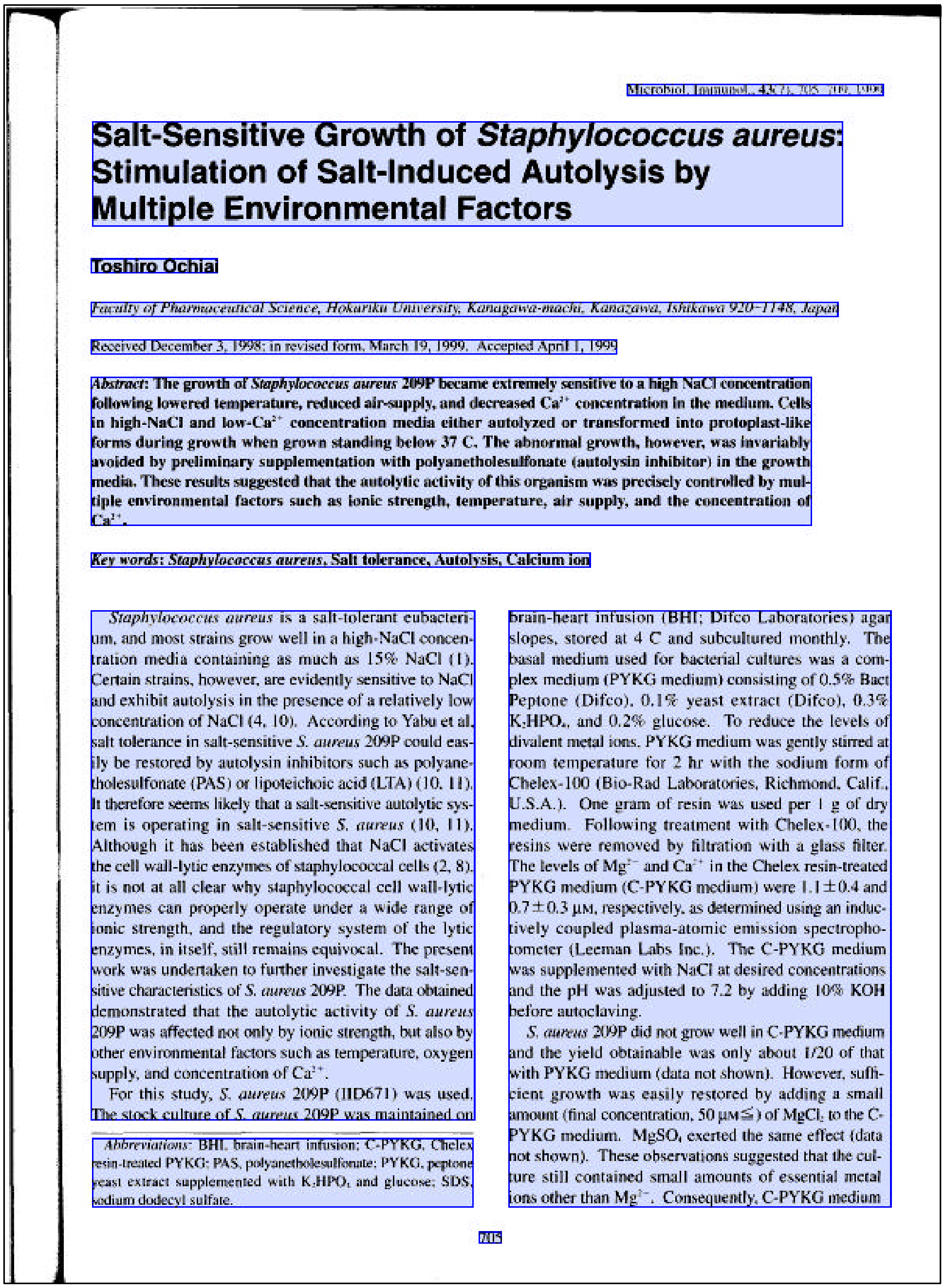
1. INTRODUCTION

The National Library of Medicine is automating the production of bibliographic records for its premier database, MEDLINE. As a first step, the Communications Engineering Branch of the Lister Hill National Center for Biomedical Communications has developed a system called MARS for Medical Article Record System. The first version, MARS-1, involves scanning and converting by optical character recognition the abstracts that appear in journal articles, while keyboarding the remaining fields (e.g., article title, authors, affiliations, etc). We are designing and developing the second generation system, MARS-2 [1], to automate the entry of these other fields also, and thereby increase the production rate. This new system employs scanning and OCR as well, in addition to modules that automatically zone the scanned pages, identify, or "label", the zones as particular fields, and reformat the field syntax to adhere to MEDLINE conventions.

In MARS-2 the entire image of the first page of each article in the biomedical journal is submitted for OCR conversion, and at least four pieces of textual information will be extracted: *title, authors, affiliation* and *abstract*. The first step after OCR conversion is to automatically identify these four areas of the image. In the context of the MARS system, a "zone" is a rectangular area of the scanned image, designated by two pairs of X,Y coordinates. A correct zone delineates the portion of the image containing characters that all belong to the same text group, such as Title, Author, Affiliation or Abstract. Figure 1 illustrates a page image with boxes drawn around typical, correct zones.

2. BACKGROUND

The commercial OCR system used in the MARS system includes a package to perform automatic zoning. However, this feature does not reliably segment the images of biomedical journal pages into zones that are appropriate for individual identification of title, author, affiliation and abstract text groups. The most common error made by the commercial automatic zone function is that zones are too large and include more than one significant text group. Figure 2 illustrates a typical case where the title, author and affiliation are all in one zone along with extraneous publication information. Figure 3 illustrates a similar case where the two columns of the abstract are also grouped into one zone. In this example, although both columns are part of the abstract, the result is that the text lines in the two columns are joined, disrupting the proper reading order. For example, the middle text of the first line of the zone is incorrectly read as "models have opment of."



Microbiol. Immunol., 43(7), 705–710, 1999

Salt-Sensitive Growth of *Staphylococcus aureus*: Stimulation of Salt-Induced Autolysis by Multiple Environmental Factors

Toshiro Ochiai

Faculty of Pharmaceutical Science, Hokuriku University, Kanagawa-machi, Kanazawa, Ishikawa 920–1148, Japan

Received December 3, 1998; in revised form, March 19, 1999. Accepted April 1, 1999

***Abstract*: The growth of *Staphylococcus aureus* 209P became extremely sensitive to a high NaCl concentration following lowered temperature, reduced air-supply, and decreased Ca^{2+} concentration in the medium. Cells in high-NaCl and low-Ca^{2+} concentration media either autolyzed or transformed into protoplast-like forms during growth when grown standing below 37 C. The abnormal growth, however, was invariably avoided by preliminary supplementation with polyanetholesulfonate (autolysin inhibitor) in the growth media. These results suggested that the autolytic activity of this organism was precisely controlled by multiple environmental factors such as ionic strength, temperature, air supply, and the concentration of Ca^{2+}.**

***Key words*: *Staphylococcus aureus*, Salt tolerance, Autolysis, Calcium ion**

Staphylococcus aureus is a salt-tolerant eubacterium, and most strains grow well in a high-NaCl concentration media containing as much as 15% NaCl (1). Certain strains, however, are evidently sensitive to NaCl and exhibit autolysis in the presence of a relatively low concentration of NaCl (4, 10). According to Yabu et al, salt tolerance in salt-sensitive *S. aureus* 209P could easily be restored by autolysin inhibitors such as polyanetholesulfonate (PAS) or lipoteichoic acid (LTA) (10, 11). It therefore seems likely that a salt-sensitive autolytic system is operating in salt-sensitive *S. aureus* (10, 11). Although it has been established that NaCl activates the cell wall-lytic enzymes of staphylococcal cells (2, 8), it is not at all clear why staphylococcal cell wall-lytic enzymes can properly operate under a wide range of ionic strength, and the regulatory system of the lytic enzymes, in itself, still remains equivocal. The present work was undertaken to further investigate the salt-sensitive characteristics of *S. aureus* 209P. The data obtained demonstrated that the autolytic activity of *S. aureus* 209P was affected not only by ionic strength, but also by other environmental factors such as temperature, oxygen supply, and concentration of Ca^{2+}.

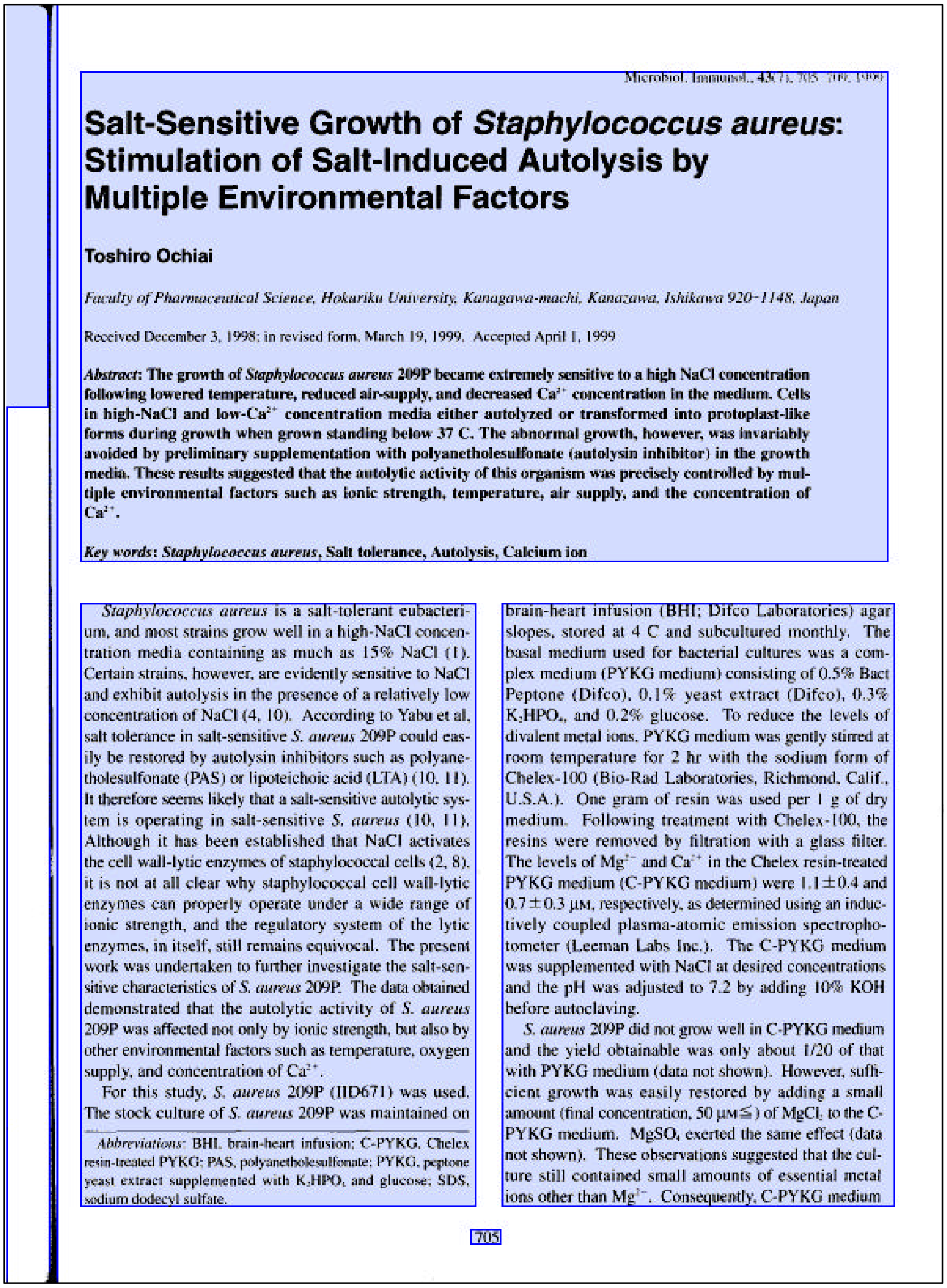
For this study, *S. aureus* 209P (IID671) was used. The stock culture of *S. aureus* 209P was maintained on brain-heart infusion (BHI; Difco Laboratories) agar slopes, stored at 4 C and subcultured monthly. The basal medium used for bacterial cultures was a complex medium (PYKG medium) consisting of 0.5% Bact Peptone (Difco), 0.1% yeast extract (Difco), 0.3% K_2HPO_4, and 0.2% glucose. To reduce the levels of divalent metal ions, PYKG medium was gently stirred at room temperature for 2 hr with the sodium form of Chelex-100 (Bio-Rad Laboratories, Richmond, Calif., U.S.A.). One gram of resin was used per 1 g of dry medium. Following treatment with Chelex-100, the resins were removed by filtration with a glass filter. The levels of Mg^{2+} and Ca^{2+} in the Chelex resin-treated PYKG medium (C-PYKG medium) were 1.1±0.4 and 0.7±0.3 μM, respectively, as determined using an inductively coupled plasma-atomic emission spectrophotometer (Leeman Labs Inc.). The C-PYKG medium was supplemented with NaCl at desired concentrations and the pH was adjusted to 7.2 by adding 10% KOH before autoclaving.

S. aureus 209P did not grow well in C-PYKG medium and the yield obtainable was only about 1/20 of that with PYKG medium (data not shown). However, sufficient growth was easily restored by adding a small amount (final concentration, 50 μM≦) of $MgCl_2$ to the C-PYKG medium. $MgSO_4$ exerted the same effect (data not shown). These observations suggested that the culture still contained small amounts of essential metal ions other than Mg^{2+}. Consequently, C-PYKG medium

Abbreviations: BHI, brain-heart infusion; C-PYKG, Chelex resin-treated PYKG; PAS, polyanetholesulfonate; PYKG, peptone yeast extract supplemented with K_2HPO_4 and glucose; SDS, sodium dodecyl sulfate.

705

Figure 1. Correct zones, generated by the zone correction algorithm.



Salt-Sensitive Growth of *Staphylococcus aureus*: Stimulation of Salt-Induced Autolysis by Multiple Environmental Factors

Toshiro Ochiai

Faculty of Pharmaceutical Science, Hokuriku University, Kanagawa-machi, Kanazawa, Ishikawa 920–1148, Japan

Received December 3, 1998; in revised form, March 19, 1999. Accepted April 1, 1999

***Abstract*: The growth of *Staphylococcus aureus* 209P became extremely sensitive to a high NaCl concentration following lowered temperature, reduced air-supply, and decreased Ca^{2+} concentration in the medium. Cells in high-NaCl and low-Ca^{2+} concentration media either autolyzed or transformed into protoplast-like forms during growth when grown standing below 37 C. The abnormal growth, however, was invariably avoided by preliminary supplementation with polyanetholesulfonate (autolysin inhibitor) in the growth media. These results suggested that the autolytic activity of this organism was precisely controlled by multiple environmental factors such as ionic strength, temperature, air supply, and the concentration of Ca^{2+}.**

***Key words*: *Staphylococcus aureus*, Salt tolerance, Autolysis, Calcium ion**

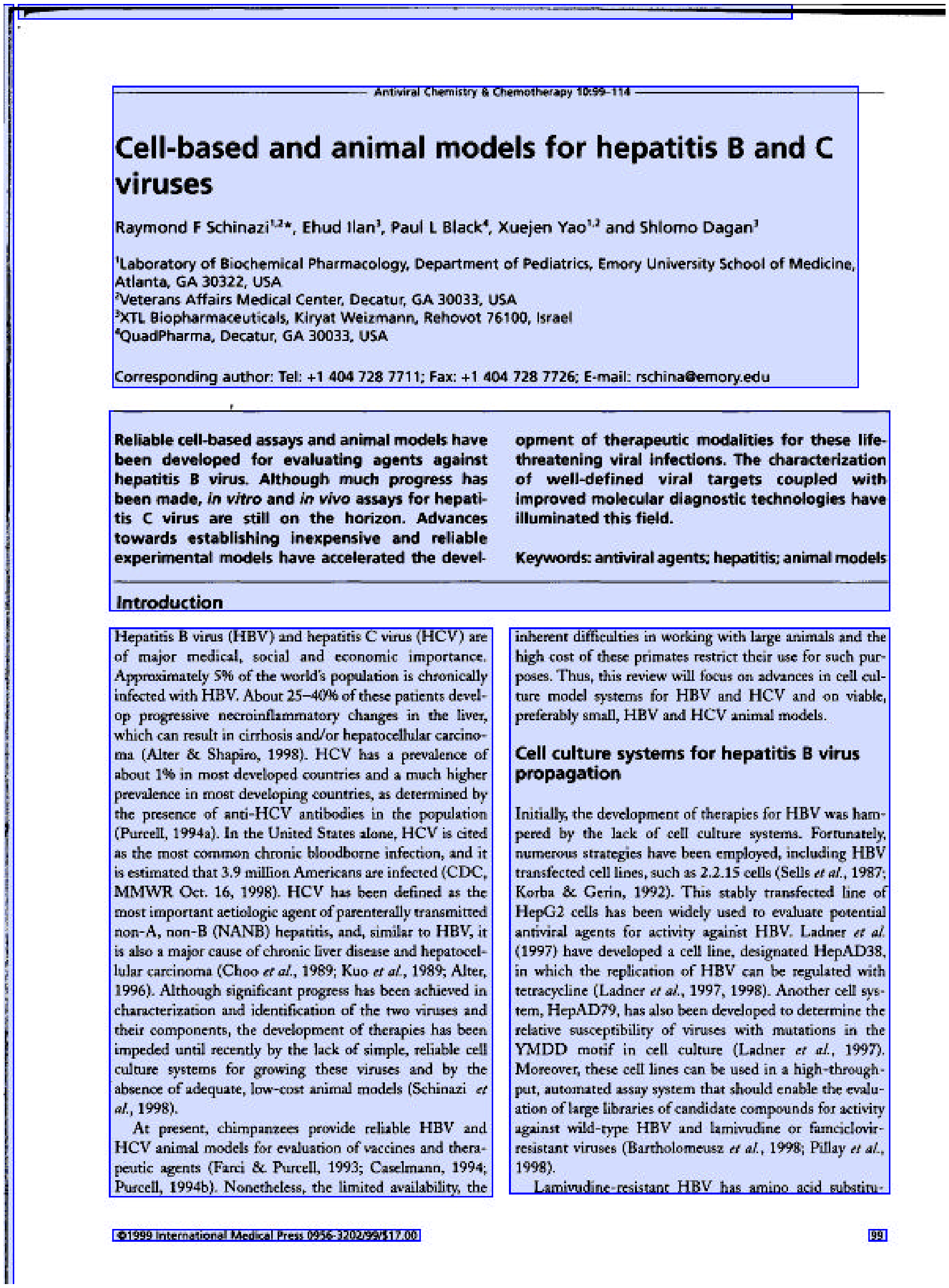
Staphylococcus aureus is a salt-tolerant eubacterium, and most strains grow well in a high-NaCl concentration media containing as much as 15% NaCl (1). Certain strains, however, are evidently sensitive to NaCl and exhibit autolysis in the presence of a relatively low concentration of NaCl (4, 10). According to Yabu et al, salt tolerance in salt-sensitive *S. aureus* 209P could easily be restored by autolysin inhibitors such as polyanetholesulfonate (PAS) or lipoteichoic acid (LTA) (10, 11). It therefore seems likely that a salt-sensitive autolytic system is operating in salt-sensitive *S. aureus* (10, 11). Although it has been established that NaCl activates the cell wall-lytic enzymes of staphylococcal cells (2, 8), it is not at all clear why staphylococcal cell wall-lytic enzymes can properly operate under a wide range of ionic strength, and the regulatory system of the lytic enzymes, in itself, still remains equivocal. The present work was undertaken to further investigate the salt-sensitive characteristics of *S. aureus* 209P. The data obtained demonstrated that the autolytic activity of *S. aureus* 209P was affected not only by ionic strength, but also by other environmental factors such as temperature, oxygen supply, and concentration of Ca^{2+}.

For this study, *S. aureus* 209P (IID671) was used. The stock culture of *S. aureus* 209P was maintained on brain-heart infusion (BHI; Difco Laboratories) agar slopes, stored at 4 C and subcultured monthly. The basal medium used for bacterial cultures was a complex medium (PYKG medium) consisting of 0.5% Bact Peptone (Difco), 0.1% yeast extract (Difco), 0.3% K_2HPO_4, and 0.2% glucose. To reduce the levels of divalent metal ions, PYKG medium was gently stirred at room temperature for 2 hr with the sodium form of Chelex-100 (Bio-Rad Laboratories, Richmond, Calif., U.S.A.). One gram of resin was used per 1 g of dry medium. Following treatment with Chelex-100, the resins were removed by filtration with a glass filter. The levels of Mg^{2+} and Ca^{2+} in the Chelex resin-treated PYKG medium (C-PYKG medium) were 1.1±0.4 and 0.7±0.3 μM, respectively, as determined using an inductively coupled plasma-atomic emission spectrophotometer (Leeman Labs Inc.). The C-PYKG medium was supplemented with NaCl at desired concentrations and the pH was adjusted to 7.2 by adding 10% KOH before autoclaving.

S. aureus 209P did not grow well in C-PYKG medium and the yield obtainable was only about 1/20 of that with PYKG medium (data not shown). However, sufficient growth was easily restored by adding a small amount (final concentration, 50 μM≦) of $MgCl_2$ to the C-PYKG medium. $MgSO_4$ exerted the same effect (data not shown). These observations suggested that the culture still contained small amounts of essential metal ions other than Mg^{2+}. Consequently, C-PYKG medium

Abbreviations: BHI, brain-heart infusion; C-PYKG, Chelex resin-treated PYKG; PAS, polyanetholesulfonate; PYKG, peptone yeast extract supplemented with K_2HPO_4 and glucose; SDS, sodium dodecyl sulfate.

Figure 2. An example of large zones generated by the commercial OCR system.



Antiviral Chemistry & Chemotherapy 10:99–114

Cell-based and animal models for hepatitis B and C viruses

Raymond F Schinazi[1,2]*, Ehud Ilan[3], Paul L Black[4], Xuejen Yao[1,2] and Shlomo Dagan[3]

[1]Laboratory of Biochemical Pharmacology, Department of Pediatrics, Emory University School of Medicine, Atlanta, GA 30322, USA
[2]Veterans Affairs Medical Center, Decatur, GA 30033, USA
[3]XTL Biopharmaceuticals, Kiryat Weizmann, Rehovot 76100, Israel
[4]QuadPharma, Decatur, GA 30033, USA

Corresponding author: Tel: +1 404 728 7711; Fax: +1 404 728 7726; E-mail: rschina@emory.edu

Reliable cell-based assays and animal models have been developed for evaluating agents against hepatitis B virus. Although much progress has been made, *in vitro* and *in vivo* assays for hepatitis C virus are still on the horizon. Advances towards establishing inexpensive and reliable experimental models have accelerated the development of therapeutic modalities for these life-threatening viral infections. The characterization of well-defined viral targets coupled with improved molecular diagnostic technologies have illuminated this field.

Keywords: antiviral agents; hepatitis; animal models

Introduction

Hepatitis B virus (HBV) and hepatitis C virus (HCV) are of major medical, social and economic importance. Approximately 5% of the world's population is chronically infected with HBV. About 25–40% of these patients develop progressive necroinflammatory changes in the liver, which can result in cirrhosis and/or hepatocellular carcinoma (Alter & Shapiro, 1998). HCV has a prevalence of about 1% in most developed countries and a much higher prevalence in most developing countries, as determined by the presence of anti-HCV antibodies in the population (Purcell, 1994a). In the United States alone, HCV is cited as the most common chronic bloodborne infection, and it is estimated that 3.9 million Americans are infected (CDC, MMWR Oct. 16, 1998). HCV has been defined as the most important aetiologic agent of parenterally transmitted non-A, non-B (NANB) hepatitis, and, similar to HBV, it is also a major cause of chronic liver disease and hepatocellular carcinoma (Choo *et al.*, 1989; Kuo *et al.*, 1989; Alter, 1996). Although significant progress has been achieved in characterization and identification of the two viruses and their components, the development of therapies has been impeded until recently by the lack of simple, reliable cell culture systems for growing these viruses and by the absence of adequate, low-cost animal models (Schinazi *et al.*, 1998).

At present, chimpanzees provide reliable HBV and HCV animal models for evaluation of vaccines and therapeutic agents (Farci & Purcell, 1993; Caselmann, 1994; Purcell, 1994b). Nonetheless, the limited availability, the inherent difficulties in working with large animals and the high cost of these primates restrict their use for such purposes. Thus, this review will focus on advances in cell culture model systems for HBV and HCV and on viable, preferably small, HBV and HCV animal models.

Cell culture systems for hepatitis B virus propagation

Initially, the development of therapies for HBV was hampered by the lack of cell culture systems. Fortunately, numerous strategies have been employed, including HBV transfected cell lines, such as 2.2.15 cells (Sells *et al.*, 1987; Korba & Gerin, 1992). This stably transfected line of HepG2 cells has been widely used to evaluate potential antiviral agents for activity against HBV. Ladner *et al.* (1997) have developed a cell line, designated HepAD38, in which the replication of HBV can be regulated with tetracycline (Ladner *et al.*, 1997, 1998). Another cell system, HepAD79, has also been developed to determine the relative susceptibility of viruses with mutations in the YMDD motif in cell culture (Ladner *et al.*, 1997). Moreover, these cell lines can be used in a high-throughput, automated assay system that should enable the evaluation of large libraries of candidate compounds for activity against wild-type HBV and lamivudine or famciclovir-resistant viruses (Bartholomeusz *et al.*, 1998; Pillay *et al.*, 1998).

Lamivudine-resistant HBV has amino acid substitu-

©1999 International Medical Press 0956-3202/99/$17.00

99

Figure 3. A second example of large zones generated by the commercial OCR system.

Correct zones are critical to downstream processes in MARS-2. For example, in the step following OCR and automatic zoning, those zones corresponding to title, authors, affiliation and abstract are automatically identified and labeled as such [2]. This complex labeling process uses several pieces of information in each zone to determine its identity. Information used to label a zone include absolute and relative location of the zone, and key words within the zone. The zone region must be correct if it is to provide useful information to the labeling program.

In another automatic process downstream from automatic zoning and automatic labeling, the title, author and affiliation fields are reformatted to comply with MEDLINE conventions [3]. This process also depends on correctly sized and labeled zones to be effective. Incorrect zones confound these other automated processes and ultimately require time-consuming manual intervention to be corrected, thereby offsetting the advantage expected from an automated system.

An alternative to automatic zoning is to have operators manually, using special software and the mouse, draw correct zones onto the bitmapped images prior to the OCR process. This is done in the MARS-1 system to identify the title and abstract zones. We find that it takes operators about 14 seconds per image to draw these two zones. For the four zones needed in MARS-2, we can estimate that it would require about 28 seconds per image of operator time to perform manual zoning. The purpose of the MARS-2 system is to remove these significant portions of manual labor. For example, for our target production rate of 1,000 records a day, this manual zoning would add over 7 person hours of labor.

Since we cannot depend on the commercial OCR system to correctly zone images, and seek to eliminate manual zoning, we are developing our own automatic zoning capability. With our own process, we free ourselves from depending on the commercial OCR system for automatic zoning, and can tailor the zone program design and operating parameters for images from the specific biomedical journals relevant to MEDLINE.

3. METHODS AND PROCEDURES

Recent improvements in OCR recognition accuracy have come coincidentally with improvements in automatic zoning. In many cases the same methods used to isolate and separate characters can be expanded to isolate and separate zones. A brief survey of activity in automatic zoning methods is given in Jain [4]. Approaches include "top-down" [5], which segment a page by X-cuts and Y-cuts into smaller regions, "bottom-up" [6,7], which recursively grow homogeneous regions from small components, and combinations of both [4,8]. Notable tradeoffs among the methods are granularity (finding small enough zones), computation time, and sensitivity to input parameters, noise, skew and page orientation. Top-down methods tend to be faster and less sensitive to input parameters and page orientation, but require pages to have a "Manhattan layout", which means the blocks can be separated by vertical and horizontal lines. Bottom-up and combination methods often result in greater accuracy at the expense of computational complexity and sensitivity to input parameters. All of these methods zone the page using image data alone, prior to OCR conversion. Our approach, in contrast, is to use the output data of the OCR system to improve automatic zoning.

In addition to ASCII text, the commercial OCR system also provides information about each of the converted characters in the output file. This information includes the level of confidence that the character was correctly recognized, character attributes such as italic or bold, character point size, and the x and y coordinates of the rectangle that bounds the character (bounding boxes) [9]. Thus we have both geometric and feature information available for each converted character. Our approach is to utilize all of this information to group text into correct zones. We use the bounding box coordinates to determine which characters are grouped in the same location on the page. Information on character size and attributes provide additional clues for keeping groups of adjacent characters together or putting them in separate zones.

Our zone correction program uses both top-down and bottom-up design strategies normally used on image data on our post-OCR non-image data. The overall method is outlined in Table I.

Table I. Zone Correction Program Steps.

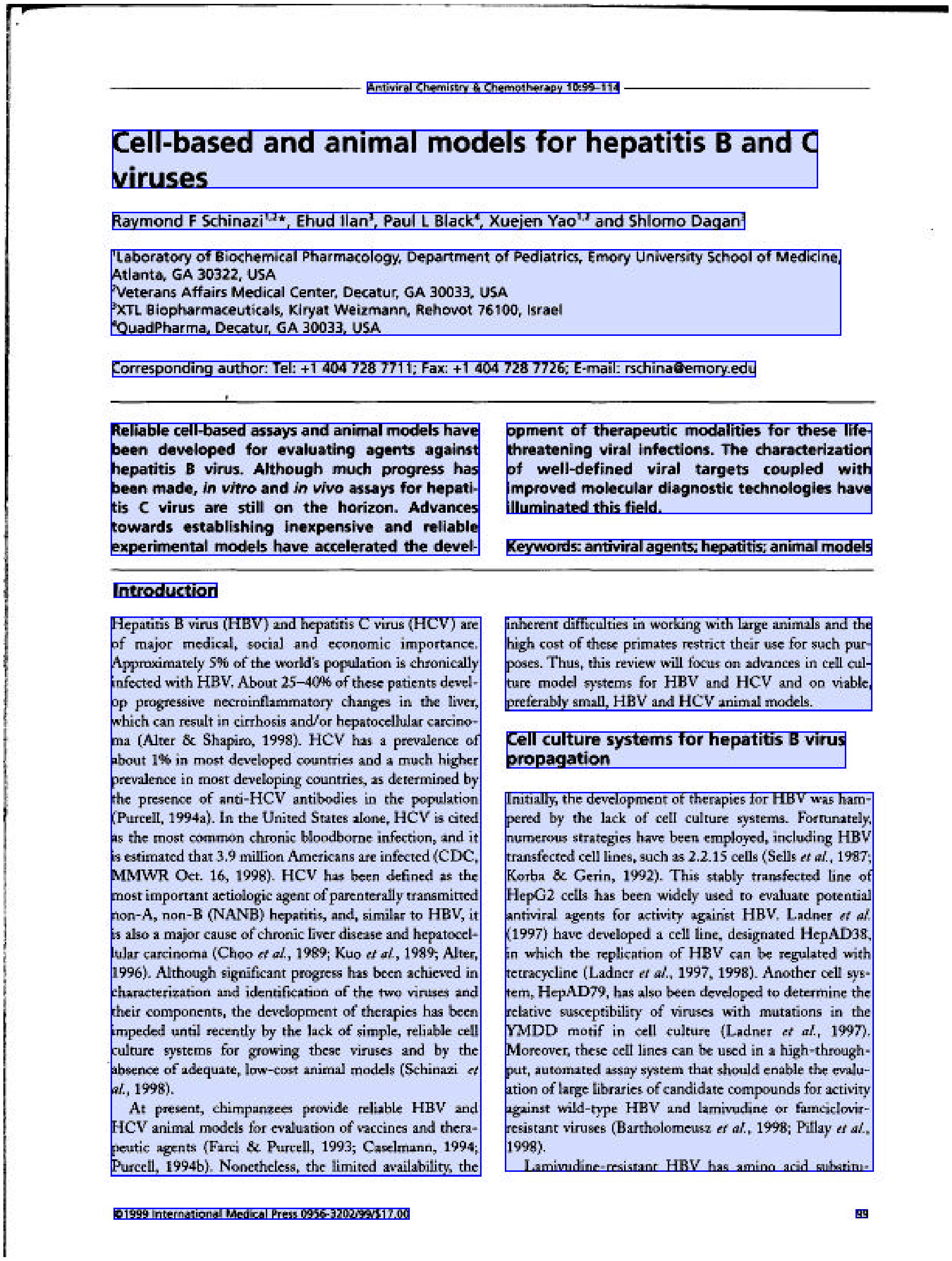
	Input	Function	Output
1.	Zones and data from OCR system	Separate zones into text lines	Text lines
2.	Text lines	Separate lines into fragments	Text lines
3.	Lines and line fragments	Combine lines vertically into zones	Initial zones
4.	Initial zones	Combine zones horizontally into zones	Final zones

The first step in creating new zones is to disassemble the original zones from the OCR system. Each zone is divided into individual text lines. In step 2 lines are further split horizontally into multiple lines when the space between words exceeds a distance threshold. This occasionally results in unnecessarily splitting lines into multiple parts, but is needed in order to split lines that originally span across two closely-spaced columns, as shown in Figure 3. Some of these lines will be rejoined in later steps. The bounding box enclosing each line is computed, as are several features such as percent italic characters and average character height. Some character features, such as bold or italic, are available directly from the OCR output data. Others, such as character height or case (upper or lower), are derived from the OCR output data.

Step 3 is to combine the lines vertically into initial zones. The criteria for combining are that (a) the vertical distance between lines must be less than a threshold; (b) either the left edge, right edge or midpoint must be horizontally aligned; and (c) the features computed in the previous step must be similar. When a line is added to a zone, the zone's rectangular boundary is expanded to include the new line. Then all remaining lines are checked to see if they fall within the new zone. If so, they are added to the zone. Many of the horizontally split lines are recombined in this way.

The fourth and last step is to combine zones horizontally using criteria similar to those in the previous step. In this case the zones are combined if (a) the horizontal distance between the zones are less than a threshold; (b) either the top or bottom edges of the zones are vertically aligned; and (c) the computed features of the two zones are similar. When zones are thus merged, a new zone boundary rectangle is created to include both zones. Any other smaller zones that fall within the rectangle are included with this zone.

Figures 1 and 4 shows the results of these steps applied to the two images used as examples in Figures 2 and 3. In both of these images, the title, author, affiliation and abstract are in separate zones. In addition, in Figure 4, the 2 columns of the abstract are in separate zones. The zone labeling process, which follows the zone correction process, will identify both zones as abstract, and join the enclosed text in reading order.



Antiviral Chemistry & Chemotherapy 10:99–114

Cell-based and animal models for hepatitis B and C viruses

Raymond F Schinazi[1,2]*, Ehud Ilan[3], Paul L Black[4], Xuejen Yao[1,2] and Shlomo Dagan[3]

[1]Laboratory of Biochemical Pharmacology, Department of Pediatrics, Emory University School of Medicine, Atlanta, GA 30322, USA
[2]Veterans Affairs Medical Center, Decatur, GA 30033, USA
[3]XTL Biopharmaceuticals, Kiryat Weizmann, Rehovot 76100, Israel
[4]QuadPharma, Decatur, GA 30033, USA

Corresponding author: Tel: +1 404 728 7711; Fax: +1 404 728 7726; E-mail: rschina@emory.edu

Reliable cell-based assays and animal models have been developed for evaluating agents against hepatitis B virus. Although much progress has been made, *in vitro* and *in vivo* assays for hepatitis C virus are still on the horizon. Advances towards establishing inexpensive and reliable experimental models have accelerated the development of therapeutic modalities for these life-threatening viral infections. The characterization of well-defined viral targets coupled with improved molecular diagnostic technologies have illuminated this field.

Keywords: antiviral agents; hepatitis; animal models

Introduction

Hepatitis B virus (HBV) and hepatitis C virus (HCV) are of major medical, social and economic importance. Approximately 5% of the world's population is chronically infected with HBV. About 25–40% of these patients develop progressive necroinflammatory changes in the liver, which can result in cirrhosis and/or hepatocellular carcinoma (Alter & Shapiro, 1998). HCV has a prevalence of about 1% in most developed countries and a much higher prevalence in most developing countries, as determined by the presence of anti-HCV antibodies in the population (Purcell, 1994a). In the United States alone, HCV is cited as the most common chronic bloodborne infection, and it is estimated that 3.9 million Americans are infected (CDC, MMWR Oct. 16, 1998). HCV has been defined as the most important aetiologic agent of parenterally transmitted non-A, non-B (NANB) hepatitis, and, similar to HBV, it is also a major cause of chronic liver disease and hepatocellular carcinoma (Choo *et al.*, 1989; Kuo *et al.*, 1989; Alter, 1996). Although significant progress has been achieved in characterization and identification of the two viruses and their components, the development of therapies has been impeded until recently by the lack of simple, reliable cell culture systems for growing these viruses and by the absence of adequate, low-cost animal models (Schinazi *et al.*, 1998).

At present, chimpanzees provide reliable HBV and HCV animal models for evaluation of vaccines and therapeutic agents (Farci & Purcell, 1993; Caselmann, 1994; Purcell, 1994b). Nonetheless, the limited availability, the inherent difficulties in working with large animals and the high cost of these primates restrict their use for such purposes. Thus, this review will focus on advances in cell culture model systems for HBV and HCV and on viable, preferably small, HBV and HCV animal models.

Cell culture systems for hepatitis B virus propagation

Initially, the development of therapies for HBV was hampered by the lack of cell culture systems. Fortunately, numerous strategies have been employed, including HBV transfected cell lines, such as 2.2.15 cells (Sells *et al.*, 1987; Korba & Gerin, 1992). This stably transfected line of HepG2 cells has been widely used to evaluate potential antiviral agents for activity against HBV. Ladner *et al.* (1997) have developed a cell line, designated HepAD38, in which the replication of HBV can be regulated with tetracycline (Ladner *et al.*, 1997, 1998). Another cell system, HepAD79, has also been developed to determine the relative susceptibility of viruses with mutations in the YMDD motif in cell culture (Ladner *et al.*, 1997). Moreover, these cell lines can be used in a high-throughput, automated assay system that should enable the evaluation of large libraries of candidate compounds for activity against wild-type HBV and lamivudine or famciclovir-resistant viruses (Bartholomeusz *et al.*, 1998; Pillay *et al.*, 1998).

Lamivudine-resistant HBV has amino acid substitu-

©1999 International Medical Press 0956-3202/99/$17.00

99

Figure 4. Another example of zones generated by the zone correction algorithm.

4. EVALUATION

The preliminary version of the zone correction program was tested on 4175 page images from 261 issues of 245 separate journal titles. The zones created on the 4175 images were manually viewed and scored [10]. The title, authors, affiliation and abstract fields were designated as correct, too small, too large, merged (with other specified zones), split into multiple zones, or miscellaneous errors. The error types are differentiated on the basis of "cost", a function of the computational expense and human labor needed to correct the error [11]. The least costly error is a split zone. There is a high probability that the zone labeling program that follows the zone correction program will be able to correctly label all of the zone fragments of a split zone, thus incurring a small computational expense and no human labor. The most costly error is a merged zone, i.e. a zone containing two or more of the title, author, affiliation or abstract regions. At best, the zone labeling program will label the zone as one of the regions included in the merged zone, which makes one zone too large and one or more other zones either empty or incorrectly associated with some other portion of text. Manual correction of merged zones requires some combination of cutting, pasting and typing, plus time spent to determine what needs to be done. The cost of correcting zones which are too big or too small fall in the middle, most likely requiring cutting and typing, respectively. Another cost factor is the extent to which a zoning error will affect the automatic labeling and automatic reformatting processes. We anticipate that split zones are easiest for those processes to handle, while merged zones are most difficult. Because merged zones are the most expensive error, the distance thresholds and similarity criteria used in zone correction deliberately favor split zones over merged zones.

Table II is a summary of the scores for these 4175 images. In this and subsequent tables the two or more zones that are merged are each counted as an error, which give merge errors at least twice as much weight in the final percentages as the other errors. The error rates were not so high as to be discouraging for a preliminary algorithm, but they are too high to be useful for the automated downstream processes that depend on good zones. Finding the reasons for these errors served as the basis for improving the algorithm.

Table II. Results of Zone Correction for 4175 Images.

Field	**Error Type**					
	split	**too big**	**too small**	**merged**	**totals**	**% images with an error in this field**
Title	225	2	8	24	**259**	6.2
Author	100	3	4	96	**203**	4.9
Affiliation	149	266	23	108	**546**	13.1
Abstract	773	57	20	24	**874**	20.9
totals	**1247**	**328**	**55**	**252**	**1882**	
% images with this error	30.9	7.9	1.3	6.0		

The next phase of the project was to refine the zoning criteria and parameters for an initial set of page images from 59 journals that would become the first set of journals to be processed by the new system. Journals selected have a page layout in which the title, authors, affiliations and abstract are all in only one column, and appear on the page in that order. This page layout format, which we designate as Type 1, is well-matched with the current capabilities of the zone labeling process that immediately follows the zone correction process. To improve the success of the zone correction process, journals were chosen in which the four zones of interest were distinct, either by being widely spaced or having different character attributes. Candidate journals were tested by both the zone correction process and the zone labeling process. If those processes were able to generate correctly bounded and labeled zones for the tested journal, the journal was added to the list of journals acceptable for processing by MARS-2.

For this test phase, images from the set of 59 journals were tested for both zone correction and zone labeling, and selected for MARS-2. During the testing period, the zone correction algorithm was continually updated and improved. The resulting zone correction program was re-tested and scored on the first five articles of one issue of each of the 59 journals. Table III summarizes the scores for these 295 images.

Table III. Results of an Improved Zone Correction for 59 Journals.

Field	**Error Type**					
	split	**too big**	**too small**	**merged**	**totals**	**% images with an error in this field**
Title	7				**7**	2.4
Author	1			4	**5**	1.7
Affiliation	4			5	**9**	3.1
Abstract	3			1	**4**	1.4
totals	**15**	**0**	**0**	**10**	**25**	
% images with this error	5.1	0	0	3.4		

Overall, of the 1,180 possible zones of interest, the zone correction program generated 1,155 correct zones, for 97.9% correct. The reasons for the split zones are understood and are expected to generate occasional zoning errors. These include incorrect character boundary coordinates or incorrect character attributes generated by the OCR process, and instances where the attributes of a single line of a title, affiliation or abstract are very different from the other lines of the same field. This occurs, for example, when a short line includes the italicized name of an organism, causing the average attributes of that line to be very different from the average attributes of its neighbors. In general, the criteria resulting in occasional split zones are, on balance, more useful than harmful.

5. PRODUCTION PERFORMANCE

The goal of the zone correction program is to be able to generate correct zones with an error rate less than 3% for a significant portion of the approximately 4000 journals indexed by NLM. We anticipate that we will eventually be able to handle more than 1000 journals. We are proceeding toward that goal on three fronts.

1. Finding and Testing Type1 Journals: Journals indexed by NLM are being categorized by visual inspection into groups according to page layout type. The current layout types are defined by the number of columns in which the title, author, affiliation and abstract appear, and the location of the affiliation, for example upper, lower left, lower right, or lower middle. The zone correction and the zone labeling process will use type data to adjust parameters or instructions for the journal issue being processed. We are currently concentrating on testing those journals that are Type 1, as described earlier. The first page of each article in one issue of the journal being tested is scanned, converted by OCR and processed by the zone correction and zone labeling. The resulting zones are viewed and scored. If the error rate is less than 3%, the journal is added to the list of those that can be processed by MARS-2. If the error rate is higher than 3%, the journal will be deferred for later consideration.

2. Testing Type 1 Journals from MARS-1: Journals that are currently being processed by MARS-1 are visually inspected and categorized by page layout type. Journals that are Type 1 are routed through the MARS-2 system, even though they may not have been previously tested. If the journal proceeds through MARS-2 with few zoning errors, it is added to the MARS-2 list. If the production operators decide that the zoning error rate is too high for efficient downstream processes, they have the option to return the journal to the MARS-1 system. To date, the zone correction process has performed well for these journals. For a detailed performance evaluation, zones from the first 30 journals processed by the MARS-2 system were visually examined and scored. The results for those 419 pages are shown in Table IV. The total number of 32 errors for the 1676 zones of interest is an overall error rate of 1.9%.

Table IV. Results of Zone Correction for 419 Pages from MARS-2 Production.

Field	**Error Type**					
	split	**too big**	**too small**	**merged**	**totals**	**% images with an error in this field**
Title	9			5	**14**	3.3
Author	0			5	**5**	1.2
Affiliation	4			0	**4**	1.0
Abstract	9			0	**9**	2.1
totals	**22**	**0**	**0**	**10**	**32**	
% images with this error	5.3	0	0	2.4		

At the time of writing, through these two methods, 104 journals have been tested . Of those journals, 98 have an acceptably low error rate in both the zone correction and zone labeling algorithms for the journal to be processed by MARS-2.

3. Processing Additional Types: The zone correction algorithm is being enhanced to process page layout formats other than Type 1. Journal specific information may also be necessary for certain journals, for example, journals in which each of multiple authors is followed by that author's affiliation, or journals with layouts that do not fall within one of our Types. Others have had moderate success in segmenting images by exploiting publication-specific layout information [12,13].

6. CONCLUSIONS

The algorithm developed for automatic zone correction accurately generates title, author, affiliation and abstract zone boundaries more than 97% of the time for selected journals with a Type 1 page layout format. The program also performs well in MARS-2 production on Type 1 journals that have not been pre-tested for inclusion. We anticipate comparable success with other page layout formats.

7. SUMMARY

Software based on our zone correction algorithm has been developed to automatically group text into zones in bitmapped document images using data available from the commercial OCR system. The program deconstructs the zones created by the commercial OCR system into individual lines and then, using geometric information and character features, combines lines into new zones. The program has an error rate less than 3% for selected journals with a page layout format that we call Type 1. Using page layout type and other publication-specific page layout information, the program will be enhanced to process other page layout formats as well. Current work involves testing Type 1 journals to add to the list of journals that can be processed with a low error rate by the current algorithm, and modifying the program to handle other types.

REFERENCES

1. Thoma GR, and Le DX. Medical database input using integrated OCR and document analysis and labeling technology. *Proceedings of the 1997 Symposium on Document Image Understanding Technology*, College Park, MD: University of Maryland Institute for Advances in Computer Studies; pp. 180-181, 1997.

2. Le DX, Kim J, Pearson GF, Thoma GR. Automated labeling of zones from scanned documents. *Proceedings of the 1999 Symposium on Document Image Understanding Technology*, College Park, MD: University of Maryland Institute for Advances in Computer Studies; pp. 219-226, 1999.

3. Ford GM, Hauser SE, Thoma GR. Automatic reformatting of OCR text from biomedical journal articles. *Proceedings of the 1999 Symposium on Document Image Understanding Technology*, College Park, MD: University of Maryland Institute for Advances in Computer Studies; pp. 321-325, 1999.

4. Jain AK, Yu B. Document representation and its application to page decomposition. . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, 1998, pp. 294-308.

5. Nagy G, Seth S, Viswanathan M. A prototype document image-analysis system for technical journals. *Computer*, Vol. 25, 1992, pp. 10-22.

6. O'Gorman L. The document spectrum for page layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, 1993, pp. 1162-1173.

7. Niyogi D., Srihari SN. Integrated approach to document decomposition and structual analysis. *International journal of Imaging Systems and Technology*, Vol. 7, 1996, pp. 330-342.

8. Farrow GSD, Xydeas CS, Oakley JP, Khorabi A, Prelcic NG. A comparison of system architecture for intelligent document understanding. *Signal Processing-Image Communication*, Vol. 9, 1996, pp. 1-19.

9. PrimeOCR Access Guide, Version 3.0. Prime Recognition, 1998.

10. Pearson GF, Thoma GR. Manual verification and correction of automatically labeled zones: user interface considerations. *Proceedings of the 1999 Symposium on Document Image Understanding Technology*, College Park, MD: University of Maryland Institute for Advances in Computer Studies, pp. 326-339, 1999.

11. Kanai J, Rice SV, Nartker TA, Nagy G. Automated evaluation of OCR zoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 17, 1995, pp. 86-90.

12. Krishnamoorthy M, Nagy M, Seth S. Viswanathen M. Syntactic segmentation and labeling of digitized pp. from technical journals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, 1993, pp. 737-747.

13. Baird HS. Model-directed document image analysis. *Proceedings of the 1999 Symposium on Document Image Understanding Technology*, College Park, MD: University of Maryland Institute for Advances in Computer Studies, pp. 42-49, 1999.