Cryptococcus neoformans Chemotyping by Quantitative Analysis of $^1$H Nuclear Magnetic Resonance Spectra of Glucuronoxylomannans with a Computer-Simulated Artificial Neural Network

ROBERT CHERNIAK,1* HOMAYOUN VALAFAR,2 LAURA C. MORRIS,1 AND FARAMARZ VALAFAR2

Department of Chemistry, Laboratory for Biological and Chemical Sciences, Georgia State University, Atlanta, 1 and Complex Carbohydrate Research Center, University of Georgia, Athens, 2 Georgia

Received 15 September 1997/Returned for modification 11 December 1997/Accepted 30 December 1997

The complete assignment of the proton chemical shifts obtained by nuclear magnetic resonance (NMR) spectroscopy of de-O-acetylated glucuronoxylomannans (GXMs) from Cryptococcus neoformans permitted the high-resolution determination of the total structure of any GXM. Six structural motifs based on an $\alpha$-(1→3)-mannotriose substituted with variable quantities of 2-0-$\beta$- and 4-0-$\beta$-xylopyranosyl and 2-0-$\beta$-glucopyranosyluronic acid were identified. The chemical shifts of only the anomeric protons of the mannosyl residues served as structure reporter groups (SRG) for the identification and quantitation of the six triads present in any GXM. The assigned protons for the mannosyl residues resonated at clearly distinguishable positions in the spectrum and supplied all the information essential for the assignment of the complete GXM structure. This technique for assigning structure is referred to as the SRG concept. The SRG concept was used to analyze the distribution of the six mannosyl triads of GXMs obtained from 106 isolates of C. neoformans. The six mannosyl triads occurred singularly or in combination with one or more of the other triads. The identification and quantitation of the SRG were simplified by using a computer-simulated artificial neural network (ANN) to automatically analyze the SRG region of the one-dimensional proton NMR spectra. The occurrence and relative distribution of the six mannosyl triads were used to chemotype C. neoformans on the basis of subtle variations in GXM structure determined by analysis of the SRG region of the proton NMR spectrum by the ANN. The data for the distribution of the six SRGs from GXMs of 106 isolates of C. neoformans yielded eight chemotypes, Chem1 through Chem8.

Cryptococcus neoformans is an opportunistic pathogenic yeast. In most healthy individuals the organism does not progress beyond the lungs, its normal portal of entry. Hematogenous dissemination occurs in patients who are immuno-suppressed due to immunotherapy or some form of immunodeficiency (38, 44, 52). AIDS patients are vulnerable to opportunistic infections due to the precipitous decline in the competency of the immune system (35). Consequently, the occurrence of cryptococcosis has increased dramatically, concomitantly with the emergence of AIDS (24, 41, 67). Despite antifungal therapy, AIDS patients with cryptococcosis are extremely susceptible to relapse. With the exception of Pneumocystis carinii, cryptococcosis is the most common invasive life-threatening fungal infection associated with AIDS (27); it is a leading cause of death in these patients (25). The yeast has a propensity to involve the central nervous system, where it causes meningocerebritis (26, 48). Other frequent sites of infection are the liver and spleen. Diagnosis of cryptococcosis in an AIDS patient may portend the progression to further immunosuppression and to a more severe phase of the disease (47). The insidious association of cryptococcosis with AIDS has resulted in increased attempts to define the virulence factors associated with the yeast and the host at the molecular level.

C. neoformans is atypical of pathogenic fungi in that it produces a well-developed polysaccharide capsule (18, 60). The major capsular polysaccharide is glucuronoxylomannan (GXM), which is an important contributor to the virulence of C. neoformans (60). GXM is antiphagocytic (20, 36) and poorly immunogenic (37, 45), and acapsular strains have diminished virulence (32). In vitro, GXM inhibits leukocyte migration (28), enhances human immunodeficiency virus (HIV) infection in human lymphocytes (46, 47), induces the release of tumor necrosis factor alpha by peripheral-blood mononuclear cells (21, 42), and promotes i-selectin shedding from neutrophils (28).

Two distinct varieties of C. neoformans have been described: C. neoformans var. neoformans and C. neoformans var. gattii (39). The two varieties were subdivided serologically into four serotypes based on the reactivities of whole yeast cells to polyclonal sera that were prepared by selective absorption with whole yeast cells (31, 33, 68). The variety neoformans consists of serotypes A, D, and A/D, and the variety gattii consists of serotypes B and C. Most cryptococcal infections in AIDS patients are due to C. neoformans var. neoformans, with serotype A comprising the majority of isolates (19, 40, 51, 57). Monoclonal antibodies and factor sera react specifically with GXMs. These observations substantiate the role GXMs play in conferring serotypes on C. neoformans.

The presence of at least eight antigenic factors, distributed among the serotypes of C. neoformans, has been proposed based on the reactivities of factor sera in yeast cell agglutination reactions (33). No structure corresponding to any one of the eight antigenic factors has been delineated.

The typical GXM consists of a linear (1→3)-$\alpha$-D-mannopyranan bearing $\beta$-D-xylopyranosyl (Xylp), $\beta$-D-glucopyranosyluronic acid (GlcA), and 6-O-acetyl substituents (18, 60, 63). The disposition of the O-acetyl substituents is the major determinant of the antigenic activity observed among GXMs...
obtained from all serotypes (A, B, C, D, and A/D) (15). Generally, the serological activity observed with type-specific antibody is lost after de-O-acetylation (15, 36). Despite considerable structural and antigenic diversity, a simple structural relationship exists between GXM of reference isolates for the four serotypes. They are comprised of a core repeating unit (Fig. 1) to which (1→2)-linked and (1→4)-linked \(\beta-D-Xylp\) units are added in increments of one to four residues. In this way, explicit molar ratios of Xyl/Man/GlcA in serotypes D, A, B, and C have been assigned as 1:3:1, 2:3:1, 3:3:1, and 4:3:1, respectively (Fig. 2, M1 through M4) (18). GXM from serotypes A and D are mainly substituted with Xylp at O-2, whereas GXM from serotypes B and C are substituted at O-2 and at O-4. Additional analytical data show that the precise molar ratios and typical substitution patterns, as proposed in the original models of GXM structure, are an oversimplification. In addition, substituent dispositions previously thought to be characteristic of one serotype have been identified in heterologous isolates (23, 24, 62, 63).

The primary sequences of de-O-acetylated GXM from serotypes A, B, C, and D were determined previously by two-dimensional (2D) nuclear magnetic resonance (NMR) spectroscopy (11, 56, 58, 59), and the assignments of the \(^1H\) and \(^{13}C\) chemical shifts were reported. These data, in conjunction with earlier chemical and NMR analyses of GXM, document the existence of antigenic multiplicity among the serotypes, particularly in serotypes A and C (11, 24, 62). GXM within a particular serotype cannot be subdivided serologically because factor sera and monoclonal antibodies have low discriminatory power for determining structural heterogeneity (15, 17). Consequently, the variation in GXM structure, as determined serologically, cannot be correlated to the range of virulence observed among \(C.\) neoformans isolates.

The complete assignment of the proton chemical shifts for the structural elements present in various GXM permits the high-resolution determination of the total structure of any GXM. In addition, the primary structural assignment of GXM can be made by using only the anomeric protons of the mannose residues. This is based on the fact that the assigned proton chemical shifts for the mannose residues resonate at clearly distinguishable positions in the spectrum and supply all the information essential for the assignment of the complete structure.

![FIG. 1. The core repeating structure of GXM.](image1)

![FIG. 2. The six mannosyl triads found in GXM of \(C.\) neoformans.](image2)
structure. This technique is referred to as the structural reporter group concept (SRG) (65, 66).

The identification and quantitation of the SRG can be simplified by using a computer-simulated artificial neural network (ANN) (53) to automatically analyze 1D proton NMR spectra as illustrated in Fig. 3. The 1D proton NMR spectrum serves as the input data for the ANN, and the ANN produces an output that represents the relative ratio of each SRG. The selection of an ANN to predict the chemotype of *C. neoformans* is not arbitrary. ANNs are extremely flexible in design, size, and method of training. The ANN is trained with data (1D proton NMR spectra) that define the structure of the polysaccharides obtained by conventional analytical methods.

ANN is a network that consists of simple processing elements (PEs) and connecting elements that provide communication between the PEs (53). The PEs are also referred to as “artificial neurons” or simply “neurons,” and the connecting elements are referred to as “synaptic pathways” or “synaptic connections” (53).

ANNs are ideal for extracting distinguishing features from complex data patterns. ANNs have been used for a variety of classification and identification tasks, including the identification of NMR spectra of polysaccharides and other spectral data (43, 49, 50, 54, 61), the diagnosis of diseases (1–8, 14, 15), and other classification problems (8–10, 13, 29, 55, 64). This article describes the development of a computer-simulated ANN for the quantitative analysis of GXM fine structure by using the proton NMR SRG concept. The data were used to develop a chemotyping system based on the quantitation of the subtle variations in GXM structure that occur in *C. neoformans* isolates. The ANN is potentially a powerful tool for investigating the impact that the variation in the fine structure of GXM from any particular isolate has on the effects observed with in vitro and in vivo biological systems.

**MATERIALS AND METHODS**

GXM. *C. neoformans* isolates used in this study are listed in Table 1. Many GXM samples were available from earlier studies (11, 15, 23, 24, 58, 59, 62). GXM from *C. neoformans* isolates not studied previously were prepared as described elsewhere (56). A streamlined method for the isolation of GXM from selected isolates used in this study is referenced in Table 1. The serotype of each GXM, whether determined previously or by dot enzyme assay (16) as part of this study, is given in Table 1.

**NMR spectroscopy.** De-O-acetylated GXM samples (-10 mg) were exchanged in 99.9% D₂O and lyophilized. Each sample was dissolved in 0.80 ml of 99.96% D₂O, then filtered through a Millipore MILLEX-GS 0.22-μm pore-size filter, and the filtrate was transferred into a 5-mm-diameter NMR tube (Wilmad 528-PP). All 1H NMR experiments were performed at 80°C on a Varian VXR-520 or on a Varian Unity+500 spectrometer equipped with a 5-mm 1H/19F probe. 1H chemical shifts were measured relative to internal sodium 4,4-dimethyl-2-silapentane-1-sulfonate taken as 0.00 ppm. The data were processed off-line by using a FELIX 2.30 software package (Biosym/Molecular Simulations, San Diego, Calif.) on a Silicon Graphics Indigo workstation. Each spectrum was analyzed with the Marquardt-Levenberg fitting algorithm. The chemical shifts of the resolved peaks and their areas were selected. Final data analysis was done with the Autofit Peaks I selection (by using Lorentzian functions). The peaks designated by PeakFit were processed manually based on the assignment of the chemical shifts of the SRG regions, determined previously from the resolution-enhanced spectra (see above). Rough fitting was done visually after all the peaks were selected. Final data analysis was done with the Marquardt-Levenberg fitting algorithm. The chemical shifts of the resolved peaks and their corresponding areas for GXM were saved as an ASCII file. The data were used to quantitate the presence of particular SRGs in GXM.

**ANN.** In this study we used a universal ANN design that consisted of a two-stage feed-forward neural-network topology and the standard, backpropagation training algorithm (53), as well as a modified version. Some parameters, such as hidden layer size, step size, etc. (53) were optimized experimentally. The final step was to train the ANN by using the optimum parameters. FELIX ASCII files representing the 1H NMR spectra for GXM obtained from 106 individual *C. neoformans* isolates were divided into a training set (69 spectra) and a testing set (37 spectra). The 50-50% split of the NMR data was used for the training set. The results of the PeakFit analysis (percent each SRG)
were used as the desired output data. Each file was submitted individually, via the Complex Carbohydrate Research Center’s web site (http://www.ccrc.uga.edu) by using a web browser (such as Netscape Navigator), for analysis by the trained neural network. A report page containing the percentage of each SRG in the sample data was generated by the neural network. The output of the trained neural network was compared to the data obtained independently by PeakFit. The initial experiments used the standard backpropagation training algorithm described by Rumelhart and McClelland (53). However, a modification to the definition of the error function improved the performance of the ANN. This modification included an additional mathematical term that required that the sum of the SRG ratios add up to 100%. The final ANN was trained with this modification. The trained ANN was cross-validated (53) with 1H NMR data which were not part of the training set. The purpose of cross-validation was to determine when to terminate training.

### TABLE 1. C. neoformans isolates used in this study

| Isolate | Serotype | Source |
|---------|----------|--------|
| 6       | A        | T. G. Mitchell (Duke University) |
| 18      | B/C      | K. J. Kwon-Chung (National Institutes of Health) |
| 34      | C        | K. J. Kwon-Chung |
| 98      | A        | T. G. Mitchell |
| 110     | A/D      | T. G. Mitchell |
| 118     | A        | T. G. Mitchell |
| 127     | D        | E. Jacobson (Medical College of Virginia) |
| 132     | A/D      | T. Shinoda (Meiji College of Pharmacy, Tokyo, Japan) |
| 371     | A        | K. J. Kwon-Chung |
| 401     | C        | K. J. Kwon-Chung |
| 409     | B        | E. Reiss (Centers for Disease Control and Prevention) |
| 430     | D        | K. J. Kwon-Chung |
| 1254    | D        | E. Reiss |
| 3168b   | D        | E. Reiss |
| 3183    | C        | E. Reiss |
| 3939    | B        | E. Reiss |
| 9375    | D        | H. J. Shadowy (Medical College of Virginia) |
| 9759    | A        | E. Reiss |
| 9759a   | A        | R. Cherniak; subculture of 9759 |
| 24067 a | D        | A. Casadevall (Albert Einstein College of Medicine) |
| 24067 b | D        | A. Casadevall |
| 24067 c | D        | A. Casadevall |
| 24067 d1| D        | A. Casadevall |
| 24067 d2| D        | A. Casadevall |
| 24067 e | D        | A. Casadevall |
| 24067 f1| D        | A. Casadevall |
| 24067 f2| D        | A. Casadevall |
| 24067 g | D        | A. Casadevall |
| A104    | A        | T. R. Kozel (University of Nevada, Reno) |
| A271    | A        | T. R. Kozel |
| A288    | A        | T. R. Kozel |
| A224064 | A        | T. R. Kozel |
| ATCC 6352 | A | F. I. Molina (American Type Culture Collection) |
| ATCC 24065 | B | F. I. Molina |
| ATCC 24066 | B/C | F. I. Molina |
| ATCC 28958 | D | F. I. Molina |
| ATCC 32269 | B | F. I. Molina |
| ATCC 32608 | C | F. I. Molina |
| ATCC 34875 | D | F. I. Molina |
| ATCC 34878 | D | F. I. Molina |
| ATCC 34880 | C | F. I. Molina |
| ATCC 48184 | A/D | F. I. Molina |
| ATCC 62066 | A | F. I. Molina |
| ATCC 90126 | A/D | F. I. Molina |
| AW      | A        | S. Levitz (Boston University Medical Center) |
| B3502   | D        | K. J. Kwon-Chung |
| B4538   | A        | K. J. Kwon-Chung |
| C917    | C        | T. R. Kozel |
| C24066  | C        | T. R. Kozel |
| C24066T | C        | R. Cherniak; derived from C24066 |
| C24066W | B/C      | R. Cherniak; derived from C24066 |

### TABLE 1—Continued

| Isolate | Serotype | Source |
|---------|----------|--------|
| CDC92-64 | A/B/D | M. E. Brandt (Centers for Disease Control and Prevention) |
| CDC92-159 | D | M. E. Brandt |
| CDC92-162 | A/B/D | M. E. Brandt |
| CDC92-187 | A/B/D | M. E. Brandt |
| CDC92-232 | D | M. E. Brandt |
| D52      | D        | T. R. Kozel |
| D529     | D        | T. R. Kozel |
| D24067   | D        | T. R. Kozel |
| F12      | D        | E. Jacobson |
| F27      | A        | J. Murphy (University of Oklahoma) |
| F113     | A        | J. Murphy |
| F117     | A        | J. Murphy |
| Houpt 24066 | B/C | T. R. Kozel |
| J9a      | D        | A. Casadevall |
| J9b      | D        | A. Casadevall |
| J9c      | D        | A. Casadevall |
| J9d      | D        | A. Casadevall |
| J11a     | A        | A. Casadevall |
| J11b     | A        | A. Casadevall |
| J22a     | A        | A. Casadevall |
| J22b     | A        | A. Casadevall |
| Kuhl 94-8069 | B | J. Murphy |
| M0010    | A        | T. G. Mitchell |
| M0013    | A        | T. G. Mitchell |
| M0024    | A/D      | T. G. Mitchell |
| M0025    | A        | T. G. Mitchell |
| M0027    | A        | T. G. Mitchell |
| M0029    | A        | T. G. Mitchell |
| M0046    | A        | T. G. Mitchell |
| M0048    | A        | T. G. Mitchell |
| M0050    | D        | T. G. Mitchell |
| M0051    | A        | T. G. Mitchell |
| M0053    | A        | T. G. Mitchell |
| M0062    | A        | T. G. Mitchell |
| M0095    | A        | T. G. Mitchell |
| MU-1     | A        | R. Cherniak; derived from 9759 |
| NCPF 3168 | A | A. Verheul (University Hospital, Utrecht, The Netherlands) |
| NIH 37   | A        | A. Verheul |
| NU-2     | A        | J. Murphy |
| P0       | NDc      | A. Casadevall |
| R1064    | A        | K. J. Kwon-Chung |
| SB4a     | A/D      | A. Casadevall |
| SB4b     | A/D      | A. Casadevall |
| SB4c     | D        | A. Casadevall |
| SB6a     | D        | A. Casadevall |
| SB6b     | A/D      | A. Casadevall |
| TYCC 38-602 | A | K. J. Kwon-Chung |

a B>C, activity with B greater than activity with C.
b Submitted to us as serotype A but chemically and serologically identical to serotype D.
c ND, not determined.

...
TABLE 2. Proton NMR chemical shifts of the SRG triads of
C. neoformans GXMs

| SRG | No. | Chemical shift (ppm) |
|-----|-----|----------------------|
|     | M1  | 5.241 ± 0.002         |
|     | M2  | 5.244 ± 0.002         |
|     | M3  | 5.208 ± 0.003         |
|     | M4  | 5.243 ± 0.003         |
|     | M5  | 5.217 ± 0.002         |
|     | M6  | 5.260 ± 0.002         |
| Ma  |     | 5.193 ± 0.003         |
| Mb  |     | 5.306 ± 0.004         |
| Mc  |     | 5.325 ± 0.005         |
|     |     | 5.301 ± 0.005         |
|     |     | 5.192 ± 0.003         |
|     |     | 5.192 ± 0.003         |
|     |     | 5.117 ± 0.002         |

- Number of GXMs analyzed to determine the chemical shifts of the SRG.
- Nomenclature is as presented in Fig. 1.
- In the M1 SRG, Ma and Mb are degenerate.
- In the M5 SRG, Mb and Mc are degenerate and overlap with Mb and Mc of the M1 SRG.
- In the M6 SRG, Mb is degenerate with Mb and Mc of the M1 and M5 SRGs.

RESULTS

SRGs. The complete assignment of the proton NMR chemical shifts for all the sugar residues present in GXMs selected from a broad range of C. neoformans isolates has been described previously (11, 12, 56, 58, 59). Based on the analysis of these data, we identified six recurring triad structures. The first mannosyl residue from the nonreducing end of a triad substituted at O-2 with β-d-glucuronic acid was always labeled Ma (see Fig. 1). For the development of the C. neoformans chemotyping system, the six structural triads were designated M1 through M6 (Fig. 2). The proton chemical shifts of the mannosyl residues characteristic of the six triads were obtained from our previous studies. The six triads, M1 through M6, and their characteristic chemical shifts (Table 2), constitute the SRG region of GXM. The development of the chemotyping system for C. neoformans was predicated on our ability to identify and quantify the SRGs for GXM by using high-field proton NMR.

Identification of SRGs by using proton NMR. The 1D proton spectrum of each GXM was recorded as illustrated for C. neoformans isolate 150 in Fig. 4a (1D full spectrum with SRG region marked). The SRG region of the proton spectrum was selected; resolution enhancement and expansion of the data were done to simplify the identification process (Fig. 4b). The presence of SRGs M1 through M6 in the GXM proton spectrum was determined by using the chemical-shift assignments listed in Table 2. Two SRGs, M2 and M3 were identified by using the data for the GXM of C. neoformans isolate 150 (Fig. 4b) and the data in Table 2. Similarly, 106 isolates were investigated and the SRG distributions in individual GXMs were determined (Table 3).

Quantitation of SRGs: PeakFit. The original proton NMR data for the SRG region were analyzed by using PeakFit as described in Materials and Methods. The PeakFit analysis using the FELIX ASCII file generated from the primary proton NMR data for the SRG region recorded for the GXM from C. neoformans isolate 150 is given in Fig. 5. The area for each identifiable resonance appearing in the proton NMR spectrum (Fig. 4) was assigned to a particular SRG based on its characteristic proton chemical shifts, identified as described above (Table 2). The area data depicted in Fig. 5 were used to calculate the percent occurrence of the SRGs present in the GXM from C. neoformans isolate 150 (Table 3). The analysis was repeated for all the GXMs available, and the results were tabulated (Table 3). The PeakFit data were normalized to 100%. Ordering preferences for entering SRG data in Table 3. The PeakFit data were used to determine the order of preference for entering the quantitative distribution of the SRGs in Table 3. The first order of preference was the identification of SRG M1. The data were entered in descending order of area assigned to SRG M1. When SRGs M5 and M6 were also present, they were listed without regard to their quantitative appearance in the GXM. M5 and M6 were not used in determining the order of preference here or in later entries, even in those few cases where one of them was present as the primary component, and this data set was defined as chemotype 1 (Chem1) (Table 3). The data for the ensuing set of entries were selected based on the concurrent appearance of M1 and M2, with the relative area of M1 determining the order of preference (Table 3). This data set was defined as Chem2 (Table 3). The third order of preference was based on the concurrent appearance of M1, M2, and M3, with the relative area of M1 determining the order of entry. This data set was defined as Chem3 (Table 3). The fourth order of preference was based on the sole appearance of M2, and this data set was defined as Chem4 (Table 3). The fifth order of preference was based on the appearance of M2 and M3, with the relative area of M2 determining the order of entry, and this data set was defined as Chem5 (Table 3). The sixth order of preference was based on the appearance of M2, M3, and M4, with the relative area of M2 determining the order of entry. Only two GXMs met these criteria, and this data set was defined as Chem6 (Table 3). The seventh order of preference was based on the appearance of M3, and this data set was defined as Chem7 (Table 3). The eighth, and final, order of preference was based on the appear-
| Isolate       | M1 | M2 | M3 | M4 | M5 | M6 | SRG | Total |
|--------------|----|----|----|----|----|----|-----|-------|
| Chemotype 1  |    |    |    |    |    |    | M1  |       |
| P0           | 100|     |    |    |    |    | M1  |       |
| 24067 d1     | 100|     |    |    |    |    | M1  |       |
| 24067 d2     | 100|     |    |    |    |    | M1  |       |
| 24067 e      | 100|     |    |    |    |    | M1  |       |
| 24067 f1     | 100|     |    |    |    |    | M1  |       |
| 24067 f2     | 100|     |    |    |    |    | M1  |       |
| 24067 c      | 100|     |    |    |    |    | M1  |       |
| 24067 a      | 100|     |    |    |    |    | M1  |       |
| 24067 b      | 100|     |    |    |    |    | M1  |       |
| D529         | 79 |     |    |    |    |    | M1  |       |
| 1254         | 79 |     |    |    |    |    | M1  |       |
| 9375         | 74 |     |    |    |    |    | M1  |       |
| M0050        | 70 |     |    |    |    |    | M1  |       |
| D52          | 65 |     |    |    |    |    | M1  |       |
| 3168         | 68 |     |    |    |    |    | M1  |       |
| 127          | 64 |     |    |    |    |    | M1  |       |
| J9a          | 59 |     |    |    |    |    | M1  |       |
| F12          | 57 |     |    |    |    |    | M1  |       |
| NIH 37       | 56 |     |    |    |    |    | M1  |       |
| J9b          | 53 |     |    |    |    |    | M1  |       |
| B3502        | 52 |     |    |    |    |    | M1  |       |

Continued on following page
| Isolate       | M1 | M2 | M3 | M4 | M5 | M6 | SRG | Total |
|--------------|----|----|----|----|----|----|-----|-------|
| J9c          | 50 | 29 | 21 |    |    |    | M1  | [87]  |
| J22b         | 49 | 22 | 29 |    |    |    | M1  | [107] |
| J22a         | 39 | 18 | 43 |    |    |    | M1  | [107] |
| 430          | 38 | 44 | 18 |    |    |    | M1  | [97]  |
| CDC92-159    | 38 | 28 | 34 |    |    |    | M1  | [77]  |
| ATCC 28958   | 34 | 19 | 47 |    |    |    | M1  | [107] |
| ATCC 48184   | 32 | 16 | 52 |    |    |    | M1  | [116] |
| CDC92-232    | 31 | 20 | 53 |    |    |    | M1  | [89]  |
| M0024        | 30 | 20 | 53 |    |    |    | M1  | [96]  |
| 98           | 27 | 20 | 53 |    |    |    | M1  | [89]  |
| ATCC 34875   | 27 | 25 | 48 |    |    |    | M1  | [115] |
| Chem2        |    |    |    |    |    |    |     |       |
| J9d          | 63 | 18 | 19 |    |    |    | M1>M2b | [89] |
| SB4a         | 47 | 28 | 25 |    |    |    | M1>M2 | [80]  |
| M0025        | 46 | 25 | 25 |    |    |    | M1>M2 | [70]  |
| SB6a         | 32 | 14 | 30 |    |    |    | M1>M2 | [90]  |
| 132          | 31 | 2 | 44 |    |    |    | M1>M2 | [95]  |
| 110          | 31 | 6 | 42 |    |    |    | M1>M2 | [112] |
| SB4c         | 28 | 12 | 36 |    |    |    | M1>M2 | [90]  |
| M0095        | 19 | 22 | 22 |    |    |    | M2>M1 | [88]  |
| Chem3        |    |    |    |    |    |    |     |       |
| M0051        | 32 | 28 | 18 |    |    |    | M1>M2>M3 | [90] |
| M0010        | 28 | 25 | 6  |    |    |    | M1>M2>M3 | [112] |

*Continued on following page*
| Isolate | M1 | M2 | M3 | M4 | M5 | M6 | SRG | Total |
|---------|----|----|----|----|----|----|-----|-------|
| M0062  | 25 | 17 | 8  | 20 | 30 |    | M1>M2>M3 | 100   |
|         | [18]| [39]| [2]| [3]| [38]|    |       |       |
| SB6b    | 23 | 34 | 15 |    | 28 |    | M2>M1>M3 | 96    |
|         | [18]| [33]| [3]|    | [42]|    |       |       |
| M0046   | 23 | 18 | 4  | 30 | 25 |    | M1>M2>M3 | 103   |
|         | [12]| [56]| [3]| [3]| [29]|    |       |       |
| M0027   | 23 | 46 | 31 |    |    |    | M2>M3>M1 | 68    |
|         | [2]| [49]| [17]|    |    |    |       |       |
| M0013   | 19 | 38 | 15 | 6  | 22 |    | M2>M1>M3 | 105   |
|         | [2]| [80]| [15]| [1]| [7]|    |       |       |
| AW      | 8  | 47 | 9  | 12 | 24 |    | M2>M1>M3 | 83    |
|         | [14]| [37]| [2]| [5]| [25]|    |       |       |
| Chem4   |    |    |    |    |    |    |       |       |
| MU-1    | 100|    |    |    |    |    | M2     | 103   |
|         | [89]|    |    |    |    |    |       |       |
| 201     | 100|    |    |    |    |    | M2     | 136   |
|         | [5]| [25]| [85]| [6]| [10]| [5]|       |       |
| A24064  | 100|    |    |    |    |    | M2     | 92    |
|         | [92]|    |    |    |    |    |       |       |
| 118     | 100|    |    |    |    |    | M2     | 87    |
|         | [87]|    |    |    |    |    |       |       |
| SB4b    | 77 |    |    |    | 23 |    | M2     | 80    |
|         | [72]|    |    |    | [8]|    |       |       |
| 9759a   | 76 |    |    |    | 24 |    | M2     | 79    |
|         | [63]|    |    |    | [16]|    |       |       |
| 9759    | 75 |    |    |    | 25 |    | M2     | 80    |
|         | [68]|    |    |    | [12]|    |       |       |
| M0029   | 74 |    |    |    | 26 |    | M2     | 84    |
|         | [76]|    |    |    | [8]|    |       |       |
| J11b    | 63 |    |    |    | 17 |    | M2     | 97    |
|         | [76]| [12]|    | [1]|    | [8]|       |       |
| A104    | 58 |    |    |    | 42 |    | M2     | 92    |
|         | [57]| [35]|    | [3]|    | [25]|       |       |
| Chem5   |    |    |    |    |    |    |       |       |
| R1064   | 91 |    |    |    | 9  |    | M2>M3  | 99    |
|         | [83]| [16]|    |    | [9]|    |       |       |
| M0053   | 85 |    |    |    | 15 |    | M2>M3  | 97    |
|         | [84]| [13]|    |    |    |    |       |       |
| ATCC 62066 | 82 |    |    |    | 18 |    | M2>M3  | 99    |
|         | [86]| [13]|    |    |    |    |       |       |
| F113    | 81 |    |    |    | 19 |    | M2>M3  | 103   |
|         | [86]| [17]|    |    |    |    |       |       |
| B4538   | 76 |    |    |    | 11 |    | M2>M3  | 102   |
|         | [81]| [15]|    |    | [6]|    |       |       |
| Isolate     | M1 | M2 | M3 | M4 | M5 | M6 | SRG   | Total |
|-------------|----|----|----|----|----|----|-------|-------|
| TYCC 38-602 | 75 | 25 |    |    |    |    | M2>M3 | [98]  |
| 194         | 75 | 25 |    |    |    |    | M2>M3 | [94]  |
| F27         | 75 | 25 |    |    |    |    | M2>M3 | [103] |
| F117        | 73 | 27 |    |    |    |    | M2>M3 | [94]  |
| ATCC 6352   | 71 | 29 |    |    |    |    | M2>M3 | [109] |
| 196         | 70 | 30 |    |    |    |    | M2>M3 | [88]  |
| NU-2        | 69 | 31 |    |    |    |    | M2>M3 | [98]  |
| A289        | 69 | 31 |    |    |    |    | M2>M3 | [95]  |
| CDC92-64    | 67 | 33 |    |    |    |    | M2>M3 | [91]  |
| ATCC 90126  | 65 | 35 |    |    |    |    | M2>M3 | [101] |
| 6           | 65 | 14 | 21 |    |    |    | M2>M3 | [98]  |
| J11a        | 64 | 12 | 24 |    |    |    | M2>M3 | [95]  |
| M0048       | 63 | 37 |    |    |    |    | M2>M3 | [103] |
| 371         | 60 | 40 |    |    |    |    | M2>M3 | [99]  |
| CDC92-162   | 59 | 41 |    |    |    |    | M2>M3 | [100] |
| A288        | 58 | 20 | 11 | 11 |    |    | M2>M3 | [53]  |
| NCPF 3168   | 52 | 48 |    |    |    |    | M2>M3 | [100] |
| CDC92-187   | 49 | 44 | 7  |    |    |    | M2>M3 | [100] |
| 150         | 45 | 55 |    |    |    |    | M3>M2 | [100] |
| A271        | 43 | 57 |    |    |    |    | M3>M2 | [96]  |
| ATCC 32608  | 39 | 46 | 15 |    |    |    | M3>M2 | [100] |
| Chem6       | 24 | 60 | 16 |    |    |    | M3>M2>M4 | [89] |

Continued on following page
| Isolate       | M1 | M2 | M3 | M4 | M5 | M6 | SRG | Total |
|--------------|----|----|----|----|----|----|-----|-------|
| C917         | 13 | 74 | 13 | 13 |    |    | M3>M2=M4 | [106] |
| Chem7        |    |    |    |    |    |    |      |       |
| C24066       | 100| [93]| 6  |    |    |    | M3 | [99]  |
| Houpt 24066  | 100| [93]| 5  |    |    |    | M3 | [98]  |
| C24066W      | 100| [16]| 86 | 5  |    |    | M3 | [107] |
| Kuhl 94-8069 | 100| [96]| 9  | 5  |    |    | M3 | [96]  |
| 3939         | 100| [93]| 9  |    |    |    | M3 | [93]  |
| 3183         | 100| [94]| 9  |    |    |    | M3 | [94]  |
| ATCC 24065   | 100| [96]| 5  |    |    |    | M3 | [96]  |
| ATCC 24066   | 100| [95]| 5  |    |    |    | M3 | [95]  |
| ATCC 32269   | 100| [92]| 9  |    |    |    | M3 | [92]  |
| ATCC 34878   | 100| [90]| 9  |    |    |    | M3 | [90]  |
| 409          | 100| [94]| 9  |    |    |    | M3 | [94]  |
| 184          | 100| [96]| 9  |    |    |    | M3 | [96]  |
| 18           | 100| [94]| 9  |    |    |    | M3 | [94]  |
| Chem8        |    |    |    |    |    |    |      |       |
| D24067       | 79 | 21 |    |    |    |    | M3>M4 | [96]  |
| 298Warty     | 72 | 28 |    |    |    |    | M3>M4 | [101] |
| ATCC 34880   | 65 | 35 |    |    |    |    | M3>M4 | [104] |
| 298          | 48 | 52 |    |    |    |    | M3>M4 | [104] |
| 34           | 33 | 67 |    |    |    |    | M3>M4 | [92]  |
| C24066T      | 100| [60]| 30 |    |    |    | M4 | [90]  |

* Values are percentages. For each GXM file, values on the first line are the PeakFit results normalized to 100%; values on the second line are obtained directly from the neural network analysis of the SRG region of the proton NMR spectrum and are enclosed in brackets.

b M1>M2, the area of M1 is greater than the area of M2. See Results, “Order of preference for entering SRG data in Table 3.”
ance of M3 and M4, and this set of data was defined as Chem8 (Table 3). The one archetypal isolate containing M4 only was included in this group.

Quantitation of SRGs: ANN model. Several specific ANN models of the feed-forward multistaged backpropagation type were considered (53). The architectures of these models differ in the number of hidden neurons, the complexity of input patterns, and the formulation of the output. The performances of several models were investigated to determine the one most suited for our study. Parameters such as hidden layer size, step size, etc. (53) were optimized experimentally. Fifty different networks were trained in order to determine the optimum parameters. The optimum network consisted of 1,000 input neurons, 5 hidden neurons, and 6 output neurons. A step size of 0.01 was used without a momentum term (53). The 1,000 input neurons covered a small portion of the 1D proton spectrum between 5.0 and 5.5 ppm, representing the mannose anomeric region (SRG). The results of the ANN analysis were tabulated and appear in Table 3 in brackets.

Chemotype distribution. The distribution of the eight suggested chemotypes (Chem1 through Chem8) in the 106 isolates of \textit{C. neoformans} investigated in this study is summarized in Table 4.

**DISCUSSION**

Latex agglutination tests and enzyme immunoassays for the detection of GXM in cerebral spinal fluid and serum, and polyclonal and monoclonal antibodies for serotyping, have been used for diagnostic and epidemiological studies (44). The detection of antigen is diagnostic for infection by \textit{C. neoformans}. The continued presence of high antigen levels indicates a poor prognosis for the patient, whereas antigen clearance is indicative of a successful course of treatment (34). The complete eradication of the organism by current treatment regimens is unlikely, and relapse usually occurs. Comparative investigations of the immunochemistry, immunology, and genetics of \textit{C. neoformans} are intimately linked to GXM. In turn, the results of these investigations are intrinsically related to the chemical structure of GXM at the residue level. The correlation between virulence, the course of the disease, the severity of the disease, and the chemical structure of GXM has not been studied. The recognition of the importance of GXM structure served as the driving force for this study.

Currently, \textit{C. neoformans} serotypes cannot be further subdivided because the available serological reagents do not have the discriminating power required to recognize structural variation in GXM (38). This phenomenon is due to the fact that although \textit{O}-acetyl is only a minor contributor to the epitope structure, it is a major determinant of the observed serological activity; de-\textit{O}-acylated GXM do not react with serotype-specific factor sera (15, 16, 18). Therefore, the principal variations in structure, which are based on the disposition of the sugar residues in GXM, are not detected. This lack of a strain-typing method has prevented the demonstration of point sources of infection that is required for epidemiological investigations of cryptococcosis (38). Bhattacharjee et al. (18) suggested the occurrence of four discrete model structures, one for each of the serologically defined types shown in Fig. 2 as M1 through M4, and implied that heterogeneity may exist between isolates. Heterogeneity, real or perceived, may be due to the usual pitfalls encountered in the study of complex polysaccharides by classical methods of structural analysis used in earlier studies (such as incomplete acid hydrolysis, differential loss of certain residues over others, low yields by partial acid hydrolysis, and incomplete methylation).

NMR spectroscopy, a nondestructive, highly discriminating analytical method, was used by us to probe the fine structure of de-\textit{O}-acylated polysaccharides. No secondary chemical reactions, hydrolysis, or derivatization was required. In a series of

![FIG. 5. PeakFit analysis of the SRG region of the 1D proton NMR spectrum of \textit{C. neoformans} isolate 150.](image_url)
in-depth studies, we demonstrated that heterogeneity in GXM structure is the norm. NMR analysis identified sequences originally relegated to one serotype in other types. Structures thought to occur in only one variety were commonly found in the other variety. In addition, two other triads, M5 and M6, were identified, and their structures were characterized by NMR spectroscopy. The six triads, M1 through M6 (Fig. 2), were used as a set of SRGs to develop the chemotyping system. The proton NMR spectra obtained for de-O-acetylated GXMs were resolution enhanced in order to determine precisely the chemical shifts present in a GXM. The chemical-shift data were used to identify which SRGs were present in a particular GXM by comparison to the assignments found in Table 2. Analysis of the SRG region of the original proton NMR spectrum was repeated by using PeakFit. The PeakFit analysis was used to quantify the distribution of SRGs present in a particular GXM. A large selection of GXMs was used to create a database for 106 GXMs (Table 3).

The SRG data formed a pattern of eight clusters based on the preferences used to generate Table 3. Chem1 (33 isolates [Table 4]) was determined by the presence of the SRG M1. M1 in Chem1 was almost always found associated with M6 or with M6 and M5. M5 never occurred alone or without M1 in any GXM. Chem2 (eight isolates [Table 4]) was determined by the concurrent presence of SRGs M1 and M2. Chem2 isolates were always associated with M6 or with M5 and M6. Chem3 (eight isolates [Table 4]) was determined by the concurrent presence of SRGs M1, M2, and M3. Chem3 isolates were associated with M6 in two cases and with M5 and M6 in one case; in one instance, M5 and M6 were absent. Chem4 (10 isolates [Table 4]) was determined solely by the presence of M2. In the Chem4 isolates, M6 was found associated with half the isolates and M5 was never detected. Chem5 (26 isolates [Table 4]) was determined by the presence of M2 and M3. Only four isolates in Chem5 were found associated with M6; M5 was never found. Chem6 (two isolates [Table 4]) was characterized by the concurrent presence of M2, M3, and M4; M5 and M6 were absent. Chem7 (13 isolates [Table 4]) was determined by the presence of M3; M5 and M6 were absent. This chemotype consisted of the largest structurally homogeneous set of isolates. Chem8 (6 isolates [Table 4]) was determined by the presence of M3 and M4; M5 and M6 were absent. One of the isolates was an M4 archetype isolate but was included in this chemotype. The data in Table 3 demonstrated that a continuum of SRGs characterized the GXMs of C. neoformans. A blurred demarcation between the structural elements previously assigned specifically to one or the other of the two varieties of C. neoformans (varieties neoformans and gattii) is apparent. As the degree of substitution of the core structure with Xyl increased, a concomitant decrease in the occurrence of SRGs M5 and M6 was observed. The identification of GXM archetype isolates was the exception rather than the rule. The 10 isolates in Chem7 that consisted of SRG M3 only were exceptional. However, M3 was frequently associated with Chem3, Chem5, Chem6, and Chem7.

Both the test set and the learning set of NMR data files, which were analyzed previously by PeakFit, were used to evaluate the trained neural network. The quantitative distribution of the SRGs obtained by the neural network (data in brackets) was compared to the data obtained by PeakFit (Table 3). At present there is not an absolute correspondence between the data obtained by PeakFit and those obtained by the neural network. In fact, a comparatively large percentage of M6 triad was retrieved when we performed a validation test of the ANN with a series of GXMs that essentially comprised only M1 triads (24067 isolates) (Tables 1 and 3). This phenomenon may be due to the overlap of M6b resonance with the degenerate M2b,c resonances (Table 2). Therefore, we know that training needs to be modified in order to remove any remaining ambiguity in the output data obtained by the present ANN.

The results of the standard backpropagation model for the ANN are not shown in Table 3 because of the large deviation of the sum of the ratios for the SRGs from the target of 100%. The sum of the ratios reported by the modified backpropagation model also diverged from the desired 100% value. But this divergence was much less than that observed for the standard backpropagation model. Although the ratios listed in Table 3 for the PeakFit analyses sum to 100%, this was accomplished by the manual scaling of the initial results. We believe that the small amount of residual chemical noise and resonance overlap present in the spectra affects the results of all mathematical analyses of the data.

An advantage of the ANN’s analysis of this type of data is that it can be shown mathematically that the further the sums of the ratios diverge from 100%, the less confidence one should place in the results (30). Therefore, a confidence index can be computed for each output. We did not scale the ANN output data algebraically to give a sum of the ratios equal to 100% because we believe that the observed divergence carries important information that could indicate the presence of contaminants and other possible artifacts.

In the future, we will refine the neural network by increasing the size of the learning set and by using NMR data from several archetypal isolates of C. neoformans that produce GXMs with a single mannosyl triad in the SRG region of the spectra. We will conduct experiments using larger segments of the spectrum as input data for the ANN in an attempt to increase the accuracy of the results. We plan to test the accuracy of the trained ANN by using archetype GXMs to prepare synthetic mixtures that contain predetermined ratios of the various SRGs.

The correlation of chemotype, based on the distribution of SRGs, with serotype was not possible because the O-acetyl substituent is instrumental in determining the structure of the epitope reactive with serological reagents. The GXM of 24066Tan is composed of M4 only and has no detectable acetyl substitution. Therefore, we can conclude that M3 acetylation fixed the conformation of M3 to produce the same epitope structure created by O-4 substitution in M4.

The chemotyping system and the identification of subtypes within each chemotype, based on the quantitative distribution of the SRGs, could lead to (i) a better understanding of how the polysaccharide antigens influence type 1 and type 2 T-cell responses, (ii) a better understanding of how the polysaccharides potentiate opportunistic infections in HIV-infected patients, (iii) the design of effective haptons for the preparation of neoglycoconjugate vaccines, (iv) a better understanding of how virulence is related to molecular structure, (v) the development of a computer-assisted automated system for the chemotyping of C. neoformans isolates and for the monitoring of phenotype switching, (vi) the development of a new epidemiological tool for the surveillance of cryptococcosis, and (vii) a possible explanation of the variation in antigen titer determined by common diagnostic tests.

ACKNOWLEDGMENTS

This research was supported in part by Public Health Service grant AI-31769 from the National Institutes of Health. The NMR spectrometers were purchased with the help of the National Science Foundation and the Georgia Research Alliance.
53. Rumelhart, D. E., and J. L. McClelland. 1987. Parallel distributed processing: explorations in the microstructure of cognition, vol. 1 and 2. MIT Press, Cambridge, Mass.

54. Sellers, J., W. York, P. Albersheim, A. Darvill, and B. Meyer. 1990. Identification of the mass spectra of partially methylated alditol acetates by artificial neural networks. Carbohydr. Res. 207:C1–C5.

55. Shadmehr, R., and D. Z. D’Argenio. 1990. A neural network for nonlinear Bayesian estimation in drug therapy. Neural Comput. 2:216–225.

56. Sheng, S., and R. Cherniak. 1997. Structure of the 13C-enriched O-deacetylated glucuronoxylomannan of Cryptococcus neoformans serotype A determined by NMR spectroscopy. Carbohydr. Res. 301:33–40.

57. Shimizu, R. Y., D. H. Howard, and M. N. Clancy. 1986. The variety of Cryptococcus neoformans in patients with AIDS. J. Infect. Dis. 154:1042.

58. Skelton, M. A., H. van Halbeek, and R. Cherniak. 1991. Complete assignment of the 1H and 13C-NMR spectra of the de-O-acetylated glucuronoxylomannan from Cryptococcus neoformans serotype B. Carbohydr. Res. 221:259–268.

59. Skelton, M. A., R. Cherniak, L. Poppe, and H. van Halbeek. 1991. Structure of the de-O-acetylated glucuronoxylomannan from Cryptococcus neoformans serotype D, as determined by 2D NMR spectroscopy. Magn. Reson. Chem. 29:766–793.

60. Sundstrom, J. B., and R. Cherniak. 1994. Polysaccharide antigens of the capsule of Cryptococcus neoformans. Infect. Immun. 62:1507–1512.

61. Thomsen, J., and B. Meyer. 1989. Pattern recognition of 1H NMR spectra of sugar alditols using a neural network. J. Magn. Reson. 84:212–217.

62. Turner, S. H., R. Cherniak, E. Reiss, and K. J. Kwon-Chung. 1992. Structural variability in the glucuronoxylomannan of Cryptococcus neoformans serotype A isolates determined by 13C NMR spectroscopy. Carbohydr. Res. 238:205–218.

63. Turner, S. H., and R. Cherniak. 1991. Multiplicity in the structure of the glucuronoxylomannan of Cryptococcus neoformans. NATO ASI (Adv. Sci. Inst.) Ser. Ser. H 53:123–142.

64. Valafar, F., H. Valafar, O. K. Ersoy, and R. J. Schwartz. 1995. Comparative studies of two neural network architectures for modeling of human speech production, p. 2056–2059. In Proceedings of the International Conference on Neural Networks, vol. 4.

65. van Halbeek, H. 1984. Structural analysis of the carbohydrate chains of mucin-type glycoproteins by high-resolution 1H-NMR spectroscopy. Biochem. Soc. Trans. 12:601–605.

66. Vliegenthart, J. F. G., L. Dorland, and H. van Halbeek. 1983. High-resolution 1H-nuclear magnetic resonance spectroscopy as a tool in the structural analysis of carbohydrates related to glycoproteins. Adv. Carbohydr. Chem. Biochem. 41:209–374.

67. Welisk, R. M., E. T. Starcer, and J. W. Curran. 1987. Opportunistic disease reported in AIDS patients: frequencies, associations and trends. AIDS 1:175–182.

68. Wilson, D. E., E. Bennett, and J. W. Bailey. 1968. Serologic grouping of Cryptococcus neoformans. Proc. Soc. Exp. Biol. Med. 127:820–823.