Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Three-dimensional visualization of human hemoglobin phenotypes with HPLC

L.M. Storti-Melo[1,2], P.H. Mangonaro[3], C.R. Valencio[3], C. Traina Junior[4] and C.R.B. Domingos[1]

[1]Departamento de Biologia,
Universidade Estadual Paulista Júlio de Mesquita Filho,
Universidade Estadual Paulista, Instituto de Biociências,
Letras e Ciências Exatas, São José do Rio Preto, SP, Brasil
[2]Centro de Investigação de Microorganismos,
Faculdade de Medicina de São José do Rio Preto,
São José do Rio Preto, SP, Brasil
[3]Departamento de Estatística e Ciências da Computação,
Universidade Estadual Paulista Júlio de Mesquita Filho,
Universidade Estadual Paulista, Instituto de Biociências,
Letras e Ciências Exatas, São José do Rio Preto, SP, Brasil
[4]Departamento de Ciência da Computação,
Universidade de São Paulo, São Paulo, SP, Brasil

Corresponding author: L.M. Storti-Melo
Email: stortilu@yahoo.com.br

Genet. Mol. Res. 8 (1): 354-363 (2009)
Received September 18, 2008
Accepted January 6, 2009
Published March 31, 2009

ABSTRACT. Hemoglobinopathies were included in the Brazilian Neonatal Screening Program on June 6, 2001. Automated high-performance liquid chromatography (HPLC) was indicated as one of the diagnostic methods. The amount of information generated by these systems is immense, and the behavior of groups cannot always be observed in individual analyses. Three-dimensional (3-D) visualization techniques can be applied to extract this information, for extracting patterns, trends or relations from the results stored in databases. We applied the 3-D visualization tool to analyze patterns in the results of hemoglobinopathy based on neonatal diagnosis by

HPLC. The laboratory results of 2520 newborn analyses carried out in 2001 and 2002 were used. The “Fast”, “F1”, “F” and “A” peaks, which were detected by the analytical system, were chosen as attributes for mapping. To establish a behavior pattern, the results were classified into groups according to hemoglobin phenotype: normal (N = 2169), variant (N = 73) and thalassemia (N = 279). 3-D visualization was made with the FastMap DB tool; there were two distribution patterns in the normal group, due to variation in the amplitude of the values obtained by HPLC for the F1 window. It allowed separation of the samples with normal Hb from those with alpha thalassemia, based on a significant difference ($P < 0.05$) between the mean values of the “Fast” and “A” peaks, demonstrating the need for better evaluation of chromatograms; this method could be used to help diagnose alpha thalassemia in newborns.

Key words: Neonatal screening; Hemoglobinopathies; 3-D visualization; FastMap DB

INTRODUCTION

The hemoglobinopathies are a group of genetic alterations that represent a public health problem in countries where its incidence is high (Galacteros, 1996). In Brazil, the colonization process had a great influence in the dispersal of globin mutant genes, and the hemoglobin (Hb) distribution is related to the ethnicities that comprise the population. The hemoglobinopathies were officially included in the Brazilian National Neonatal Screening Program in June 2001 (Ministério da Saúde, 2001). Studies on the identification of newborn hemoglobinopathies have been carried out by different groups, each one focusing on different collection and analysis methods (Januário, 1998). Diagnostic methods in this age group must be determined by their efficiency in the identification of fractions of small percentages and their application in population studies (Hayashi et al., 1987; Fucharoen et al., 1998; Ou and Rognerud, 2001). The Brazilian program recommends the use of isoelectric focusing or high-performance liquid chromatography (HPLC) as diagnostic methods, because they are highly sensitive and highly reproducible, where they are able to analyze a large number of samples (Ministério da Saúde, 2002). HPLC analyses supply a great amount of information, making individual analysis difficult, when the aim is to determine patterns of alterations in the population.

For optimal utilization of the information contained in a great volume of data, the use of information mapping techniques is necessary. These techniques rely on data graphical presentation as the primary tool (Chittaro, 2001). This presentation consists of the visualization of data to be analyzed. It is centered on mechanisms that make it possible for the user to comprehend rapidly the information presented. The visual exploration of data supplies a higher degree of reliability in the conclusions of the exploration of a numerical or literal representation (Keim, 2001). The automated techniques for the exploration of great amounts of data are defined as data mining, developed with the objective of finding new patterns, trends and relations. Through these techniques, the extraction of useful information is possible, which is better interpreted when presented in a graphical form (Traina et al., 2001).

FastMap DB is a tool for the visualization of data such as dates, numbers and text, where they are presented through the data multidimensional mapping for a three-dimensional (3-D) space, producing a reduction in dimensionality. Dimensionality reduction is aimed at presenting the dataset with the least number of attributes, but preserving the inherent characteristics of the stored information, and defining which attributes are important or that concentrate the information (Faloutsos and Lin, 1995). It makes it possible to determine points outside the pattern (outliers) and the formation of clusters (Tronco, 2003).

In view of the diversity of Hb phenotypes observed in the Brazilian population and the amount of information generated in newborn screening programs, the aim of this study was the application of the 3-D visualization data mining tool by FastMap DB for the identification of behavior patterns in the quantitative results in the diagnosis of hemoglobinopathies in newborns, obtained by HPLC using the sickle cell short program.

MATERIAL AND METHODS

The laboratory results of 2520 cord blood samples from the São Paulo State northwest region collected in 2001 and 2002, without distinction of gender and ethnic origin, were used. The present study was approved by CONEP and is in agreement with the norms established by Resolution CNS 196/96.

The final diagnosis for Hb phenotype was carried out by the combination of different laboratory methods, including: osmotic fragility in 0.36% NaCl solution (Silvestroni and Bianco, 1975); analysis of red blood cell morphology (Bonini-Domingos, 1993); electrophoresis in cellulose acetate, pH 8.6 and 7.0 (Marengo-Rowe, 1965; Dacie and Lewis, 1995); electrophoresis in agar-phosphate, pH 6.2 (Vella, 1968); visualization of Hb H and Heinz bodies (Papayannopoulos and Stamatonyannopoulos, 1974), and HPLC carried out with the VARIANT automated system (BIO-RAD Laboratories), using the sickle cell short program (Bio-Rad, 1994).

FastMap DB is a visual data mining tool that processes data stored in tabular format, so that it can be represented as a 3-D entity, even if the original data do not display an intrinsic spatial distribution. It is based on humans' high capability to interpret graphic information and allows graphic visualization of data stored in a relational base. FastMap DB allows users to play an interactive role in the process of knowledge discovery.

The tool allows its users to interactively create a vectorial distance function from any number of attributes of a table in a database. After the distance function is defined, the program chooses pivots among the tuple pairs (each object in the table or each sample) more distant between them. After the definition of the pivots, the system maps the data set for a 3-D representation, such that each object in the table is represented by a point in space. This system still allows the use of an attribute from the table to classify tuples and the visualization of the classes represented by different colors and shapes (Tronco, 2003).

The data mining tool was applied to the quantitative results obtained by HPLC. This system identifies the percentages of the Hb fractions in absorbance peaks determined as "Fast", "F1", "F", and "A". These peaks were used as attributes for spatial mapping, and the Hb phenotypes were used to classify the samples. To establish a spatial behavior pattern, the results were classified into groups, according to the delineation illustrated in Table 1.

Table 1. Classification of the FastMap DB analysis groups according to hemoglobin (Hb) phenotype observed.

N	Hemoglobin phenotype	FastMap DB analysis group
2169	Normal Hb profile for age	"Normal"
55	Hb S in heterozygosis	"Variant"
18	Hb C in heterozygosis	
21	Thalassemia beta heterozygote	"Thalassemia"
258	Thalassemia alpha heterozygote	

RESULTS

With the aim of establishing the normal phenotype profile for newborn blood samples, images using the four peaks discriminated as attributes for the data mining tool were generated. The mapping of these samples showed two clusters marked by the circles in Figure 1A. To determine which is the factor responsible for the sample division in spatial distribution, images eliminating some peaks were generated, in an attempt to join the data into only one distribution set.

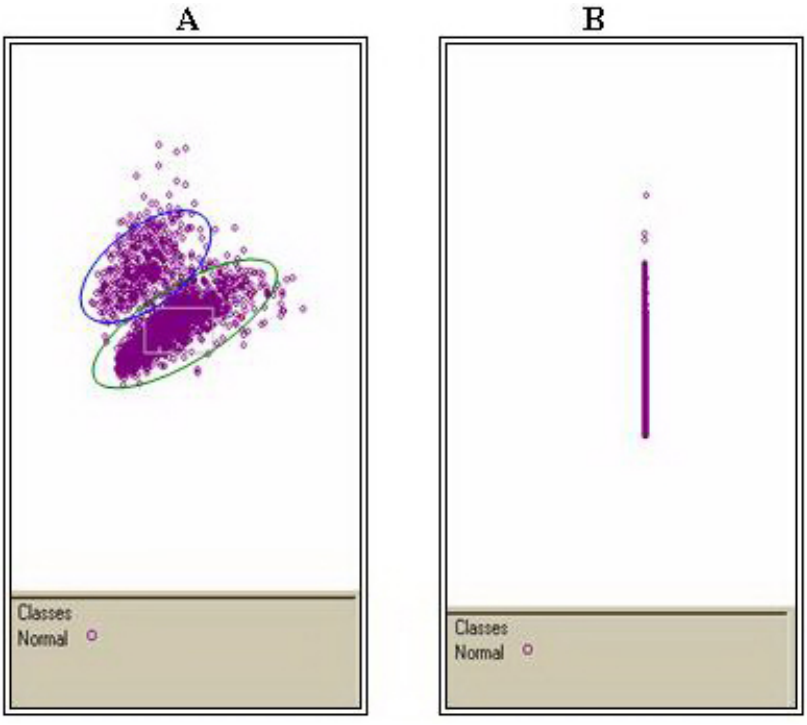

Figure 1. A. Visualization of "normal" group with the four attributes in 3-D mapping. **B.** Visualization of "normal" group with the "F" and "A" peaks as attributes in 3-D mapping.

The image generated without the "Fast" peak showed an approximation of the clouds, but the presence of two separate groups could still be observed. When the "F1" peak was left out, the data clouds merged, suggesting that these values determine the separation of the normal samples into two distinct patterns of normality. When the two observations were combined, an image only with the "F" and "A" peaks was generated. A linear data dispersion can be observed in Figure 1B.

To evaluate the "Fast" and "F1" values, images were generated using the capability of the FastMap DB tool that allows the division of the samples by the average value, constituting two classification groups for each analysis peak selected. The image illustrated in Figure 2 included the four attributes, where the "Fast" peak was used in the sample classification. The two groups obtained are represented in the image by the difference in point size. The points marked by the blue circle show "Fast" values between 0 and 5.5% and the points inside the green circle correspond to the "Fast" values from 5.5 to 10.9%. In Figure 2, it can be observed that there is no point overlapping, indicating that the two sub-groups correspond to two distinct clouds. By statistical analysis of these data, an average of 2.34% for the sub-group with a range of 0 to 5.5% was obtained, and it consisted of 2040 samples. The sub-group varying from 5.5 to 10.9% was composed of only 129 samples and an average of 7.35% was observed. The percent average obtained in the "Fast" peak, including all the normal samples was 2.64%, showing a statistically significant difference ($P < 0.05$) between the averages obtained for the two sub-groups described above.

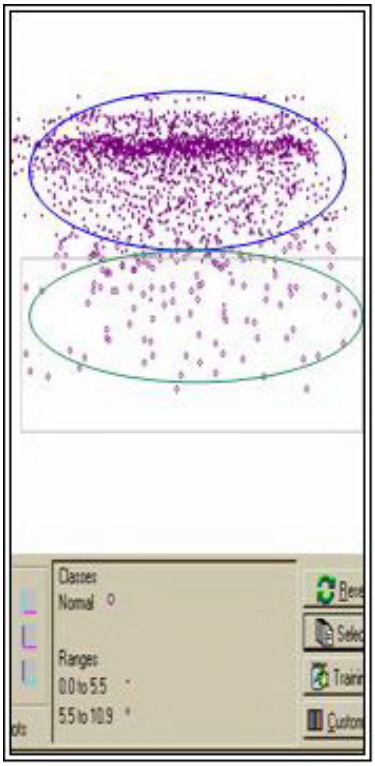

Figure 2. Visualization of "normal" group with the four attributes in 3-D mapping. "Fast" peaks were used to classify the samples, where they were divided into two sub-groups, one varying 0 to 5.5% and other 5.5 to 10.9%.

When the classification was made using the "F1" peak dividing the samples at the average value, a data mapping was generated very similar to the one in Figure 1A. Point overlapping does not exist, indicating that the two sub-groups correspond to two distinct clouds. To characterize the division in the two clouds, the average percentage for each sub-group obtained for the "F1" peak was calculated. For the sub-group varying from 0 to 10.3%, an average of 9.45% was obtained, grouping 1369 samples, and for the sub-group varying from 10.3 to 20.6%, an average of 12.25% was obtained, consisting of 800 samples. The percentage average obtained for all normal samples was 10.48%, which differed significantly ($P < 0.05$) from the averages of the two values obtained for the "F1" sub-groups.

The "normal", "thalassemia" and "variant" groups were analyzed with the four peaks as attributes, aiming to establish a pattern for the different hemoglobin phenotypes. In the mapping generated, the analysis groups were shown to be distributed into three distinct clouds, as illustrated in Figure 3, and they did not show overlapping among the points. The "normal" and "thalassemia" groups appeared to be diametrically opposed. Thus, the "normal" and "thalassemia" groups were analyzed first based on the four attributes. This mapping showed a distinction between the groups. Later, these groups were mapped using the "F" and "A" peaks as attributes because it represented the higher Hb concentrations at this age. In the image generated, a linear and parallel distribution of the groups was observed, indicating that the percentage values of the "F" and "A" fractions showed significant differences for the spatial mapping.

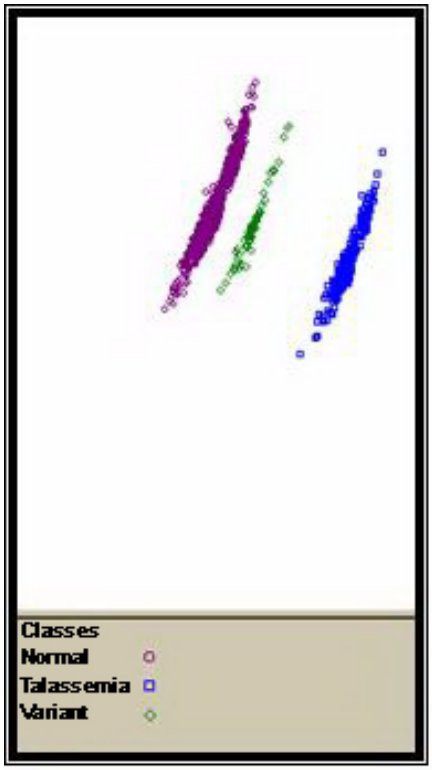

Figure 3. Visualization of "normal", "thalassemia" and "variant" groups with the four attributes in 3-D mapping.

For a better evaluation of the "thalassemia" group, images were generated with the alpha and beta thalassemia phenotypes separated and compared with the normal group. The mapping of these groups including the four attributes showed overlapping of sample points. Figure 4 illustrates the mapping with the "F" and "A" peaks, in which it can be observed that the samples are distributed linearly in the space, being well defined as three distinct sets. The image showed that the alpha and beta thalassemia groups are closer to each other and distant from the normal group.

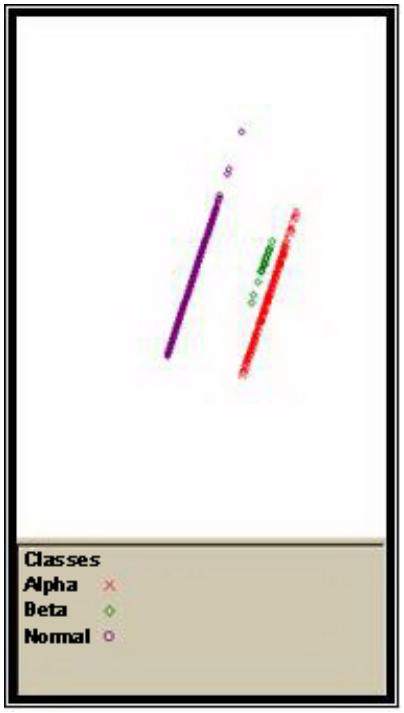

Figure 4. Visualization of "normal", "alpha" and "beta" groups with the "F" and "A" peaks as attributes in 3-D mapping.

In the spatial visualization for the "normal" and "variant" groups, there was an interface between the groups, as can be seen in Figure 5. The samples showed an overlapping area in the extremity, which persisted even with spatial rotation of the image. The "variant" group was composed of samples with Hb AS and Hb AC, in an attempt to establish a pattern between these variants and to differentiate them from samples with normal Hb. The pattern of spatial dispersion was maintained, where it was not possible to separate the samples into distinct clouds. Figure 5B represents the mapping encompassing the "normal" group, Hb AS and Hb AC with the four attributes. It should be noted that the variant fractions are represented in the HPLC for specific peaks called "S" and "C", and that these were not included as attributes for spatial mapping because they are absent in the samples with normal Hb and are specific characteristics for hemoglobinopathies.

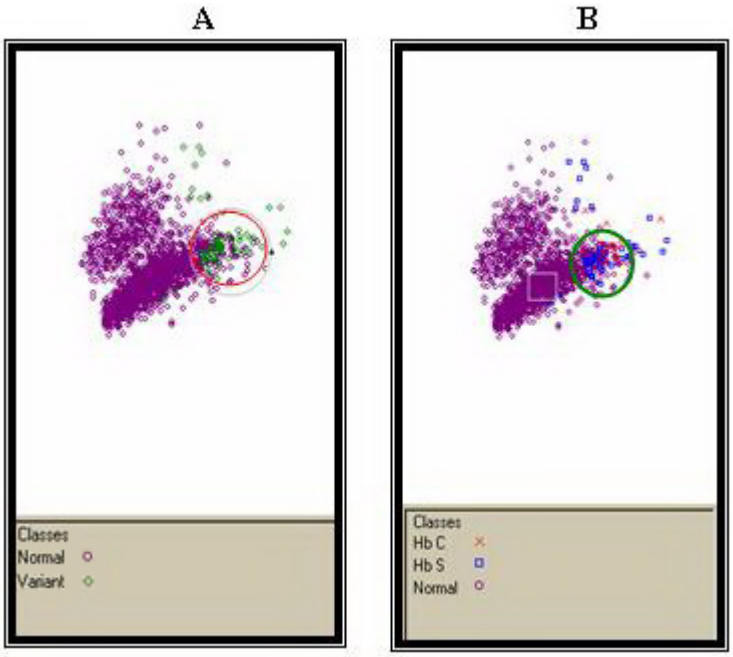

Figure 5. A. Visualization of "normal" and "variant" groups with the four attributes in 3-D mapping. Overlapping area indicated by the red circle. **B.** Visualization of "normal", "Hb AS" and "Hb AC" groups with the four attributes in 3-D mapping. Overlapping area indicated by the green circle.

DISCUSSION

The techniques and tools for data visualization are indispensable instruments in the process of knowledge discovery and have proven their value in data exploration and analysis, where these techniques have been used in visual data mining systems. With the use of visualization techniques, great volumes of data can be presented on the screen at the same time; different colors allow the user to immediately recognize similarities and differences in millions of items, and these can be arranged to determine the existence of any relationship.

The application of visual data mining techniques is a new approach in hemoglobinopathy research. For the Hb neonatal screening data, carried out by the diagnosis laboratory, the analyses using the FastMap DB tool showed very clear patterns. In the sample projection with normal Hb based on age, two distribution clouds were observed. This resource allowed us to divide the "Fast" peak values into two subgroups without overlapping. However, there was disequilibrium in the distribution of the samples, such that the first subgroup included 94% of the samples, where it was not representative of the distribution seen in Figure 1. These results indicate that the "Fast" peak was not the one responsible for the formation of two clouds observed in the samples in the mapping of samples with normal Hb for the age.

In the mapping using the "F1" peak division for the average value, two groups were found and equally distributed, where the first encompassed 63.1% of the samples. The averages

obtained for these two subgroups were shown to be significantly different, where the two also differed from the overall average for all normal samples. These results indicate the presence of two different distribution patterns for the "F1" values in those samples with normal Hb.

The global analysis of the results, visualized in spatial projection using the FastMap DB tool, evidenced the possible existence of a pattern for the thalassemic samples by HPLC, since there was no spatial overlapping of the sample points, suggesting differences in the profiles obtained by HPLC for the samples with normal values for the age studied and those with thalassemic forms, alpha or beta, even in heterozygosis. HPLC, with the sickle cell program, is not an accurate method for the diagnosis of thalassemia in newborns, but it is recommended for the national screening program. The isolated results suggested a characteristic profile for the samples that represented thalassemia, diagnosed by other methods. Mendes-Siqueira (2004) observed a statistically significant increase for the "Fast" and "F1" peaks in samples with alpha thalassemia. For samples with beta thalassemia, values for the "F" peak were increased and those for the "A" peak were diminished, differing significantly from values obtained for the normal group, suggesting the existence of a different profile for the chromatograms with thalassemic samples (Mendes-Siqueira, 2004). Although HPLC is not used as a diagnostic method for thalassemia, the different profiles observed in the spatial projection indicate the necessity of a better evaluation of these results in the screening test. Therefore, small differences in the chromatograms can indicate a probable thalassemia phenotype, making HPLC an additional method in the diagnosis of these alterations in newborns, in order to contribute to analyses associated with other specific methods to confirm this alteration.

The Hb S and Hb C are diagnosed with precision by HPLC, which identifies the small percentages of the variant fractions with high sensitivity. However, the spatial mapping of the heterozygote samples with the "normal" group showed overlapping of points, making it impossible to differentiate the groups. The Hb fetal peak represents the major hemoglobin component in this age group and does not show alterations in the percentage values between normal samples and variant samples. This finding, along with the representative difference in the number of samples, can be responsible for the overlapping of sample points. It should also be pointed out that the Hb variants elute as specific chromatographic peaks and that these were not mapped because they were absent in the normal samples, and thus, comparison was not possible.

CONCLUSIONS

The spatial projection results allowed the differentiation of behaviors between normal and thalassemic samples through their separation into distinct clouds, and evidenced the necessity to evaluate the chromatographic results together with other methods for the diagnosis of thalassemia. FastMap DB allowed information grouping and the data to be projected in 3-D space, facilitating the visualization of profiles of the different phenotypes and the interface among theoretically distinct groups.

REFERENCES

Bio-Rad (1994). Instruction Manual Variant Sickle Cell Short Program. Bio-Rad Laboratories, Hercules.

Bonini-Domingos CR (1993). Hemoglobinopatias no Brasil - Variabilidade genética e metodologia laboratorial. Doctoral thesis, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São José do Rio Preto.

Chittaro L (2001). Information visualization and its application to medicine. *Artif. Intell. Med.* 22: 81-88.

Dacie JV and Lewis SM (1995). Practical Haematology. 8th edn. Churchill Livingstone, Edinburgh.

Faloutsos C and Lin K-I (1995). FastMap: A Fast Algorithm for Indexing, Data Mining and Visualization of Traditional and Multimedia Datasets. In: Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, San Jose, California ACM Sigmod, Zurich, 163-174.

Fucharoen S, Winichagoon P, Wisedpanichkij R, Sae-Ngow B, et al. (1998). Prenatal and postnatal diagnoses of thalassemias and hemoglobinopathies by HPLC. *Clin. Chem.* 44: 740-748.

Galacteros F (1996). Neonatal detection of sickle cell disease in metropolitan France. Association francaise pour le depistage et la prevention des handicaps de l'enfant (AFDPHE). *Arch. Pediatr.* 3: 1026-1031.

Hayashi A, Wada Y, Matsuo T, Katakuse I, et al. (1987). Neonatal screening and mass-spectrometric analysis of hemoglobin variants in Japan. *Acta Haematol.* 78: 114-118.

Januário JN (1998). Programa de triagem neonatal apresenta primeiros resultados. *J. Hemominas* 8: 5-6.

Keim DA (2001). Visual exploration of large data sets. In: Communications of the ACM. ACM (Association for Computing Machinery), New York, 38-44.

Marengo-Rowe AJ (1965). Rapid electrophoresis and quantitation of haemoglobins on cellulose acetate. *J. Clin. Pathol.* 18: 790-792.

Mendes-Siqueira FA (2004). Contribuição para o Estudo das Alterações Moleculares e Interferentes na Expressão Fenotípica das Hemoglobinopatias a Partir de um Programa de Diagnóstico Neonatal. Doctoral thesis, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São José do Rio Preto.

Ministério da Saúde (2001). Portaria da Saúde, No. 822, 6 de Junho de 2001. Diário Oficial da União, Brasília.

Ministério da Saúde (2002). Manual de Normas Técnicas e Rotinas Operacionais do Programa de Triagem Neonatal. Anvisa, Brasília.

Ou CN and Rognerud CL (2001). Diagnosis of hemoglobinopathies: electrophoresis *vs* HPLC. *Clin. Chim. Acta* 313: 187-194.

Papayannopoulou T and Stamatoyannopoulos G (Editors) (1974). Stains for inclusion bodies. In: Standardization laboratory reagents and methods for detection of haemoglobinopathies. Hew Publications, Atlanta. Available at [http://www.informapharmascience.com/doi/abs/10.3109/10408367409107630].

Silvestroni E and Bianco I (1975). Screening for microcytemia in Italy: analysis of data collected in the past 30 years. *Am. J. Hum. Genet.* 27: 198-212.

Traina AJM, Traina C Jr, Botelho E, Barioni MCN, et al. (2001). Visualização de Dados em Sistemas de Banco de Dados Relacionais. In: Proceeding XVI Simpósio Brasileiro de Banco de Dados, Rio de Janeiro, 95-109.

Tronco MN (2003). Implementação de Recursos de Visualização e Interação em Ferramenta Data Mining. UNESP, São José do Rio Preto.

Vella F (1968). Acid-agar gel electrophoresis of human hemoglobins. *Am. J. Clin. Pathol.* 49: 440-442.