A Statistical Analysis of Spot Variation Using the Two-Dimensional Polyacrylamide Gel Electrophoresis

RORK KUICK,* ERIC BOERWINKLE,* SAMIR M. HANASH,†
AND CHARLES F. SING*

**Department of Human Genetics, and †Department of Pediatrics, University of Michigan Medical School, Ann Arbor, Michigan 48109*

Received February 18, 1985

For the development of valid algorithms for matching protein spots between two-dimensional gels, it is essential that one has an understanding of the relative roles of the many sources of variability affecting the location of spots. We first consider the contribution of observers to the measurement variability of spot location, arriving at a simple model for these effects. Next we present an analysis of the variability in spot locations for a sample of gels containing duplicate gels for each sample. Our data indicate that both differences between duplicates and between samples are considerable, and that the size of the effects depends on the region of the gel being considered. In the discussion we examine several matching strategies that match large groups of gels based on algorithms which match two gels at a time. © 1986 Academic Press, Inc.

INTRODUCTION

A number of technologies have emerged recently for obtaining precise measurements of gene structure and gene action. One of these, two-dimensional polyacrylamide gel electrophoresis (2-D PAGE), holds the promise of conveniently examining a large number of the gene products that are expressed in a specific tissue at a given point in time. Because it is possible to simultaneously identify the products of a very large number of loci in a tissue sample from an individual, this method is far more efficient for population studies than multiple one-dimensional gels. The use of the 2-D gel is also thought to provide a more representative sample of the products of all genetic loci than multiple 1-D gels (*1*).

Application of the 2-D PAGE method to distinguish those gene loci that have allelic variations from those that do not is one example of the utility of the 2-D gel method that is of great interest to all geneticists. Variation in the location and/or staining intensity of a protein can be the basis for studies of genetic polymorphisms and the fundamental mutational processes that alter gene expression (*2*). In the study of the common diseases of man, allelic variation detected by the 2-D gel can be used as a direct measurement of the contribution

0010-4809/86 $3.00

of genotypic variability at multiple loci to phenotypes at one or more of the levels in the hierarchy of biological organization. Such studies seek to understand the interaction between genetic and environmental factors that are a part of the etiology of the phenotypes that determine health and disease (*3*).

In order to perform genetic studies, one must first reliably "match" the spots that define gene products across the sample of gels that represent the individuals being studied. At present, because of the variability of the spot patterns from gel to gel, matching is not a trivial task. For the development of valid algorithms for matching spots, it is essential that one has an understanding of the relative roles of the many sources of biological and nonbiological variability that influence the variability among gels for the location (and staining intensity) of a spot. Genetic and/or environmental factors may cause biological variability of spot phenotypes among tissue types and developmental stages of the same individual, and among individuals measured on the same tissue and developmental stage. Nonbiological variation may be a consequence of variation in laboratory technique and errors in measuring spot location and intensity. Our ability to accurately evaluate the contributions of the biological sources of variability is determined by the relative magnitude of the technical and measurement variations inherent in using the 2-D gel method.

For a given tissue and stage of development, the impact of genetic variation among individuals on spot location may be of two sorts. A genetic variant may have a major impact on the location of a spot that sets the variant phenotype apart from its wild-type equivalent. This sort includes variants which differ by charge changes. Or, there may be minor deviations of spot location that are attributable to genetic differences among individuals that determine the normal variability within each class of major locational variants. This sort may include variation due to slight conformational changes. We have chosen to begin our studies of locational variability of the 2-D spots by estimating the sources of laboratory and measurement variability. Our goal is to identify sources of experimental variation that will direct our attention to those laboratory and image-processing methods that might be improved.

METHODS

Laboratory Methods

Each sample of whole blood is first subdivided into six subfractions: plasma, platelets, nonpolymorphonuclear cells, polymorphonuclear cells, red cell cytosol, and red cell membranes. First, a platelet-rich plasma is isolated using low-speed centrifugation. The platelets are then separated from the plasma by high-speed centrifugation. The platelet pellet is washed with phosphate-buffered saline (PBS). The white and red cells remaining after removal of the platelet-rich plasma are diluted with PBS to the original volume and the nonpolymorphonuclear cells separated by centrifugation on Ficoll–Hypaque. After removal of this fraction the remaining mixture of polymorphonuclear cells and erythrocytes is separated by sedimentation on 3% dextran sulfate. The red cell fraction is lysed to obtain the cytosol and membrane fractions. Each fraction is processed according to procedures reported in (*2*). Samples not isofocused on the day of preparation are stored as pellets at −80°C until processing can be completed.

We chose to study the variability of spot locations on 2-D gel patterns obtained from the white cell (lymphocyte) fraction. A pellet consisting of 4 to 6 million cells is solubilized in 60 to 100 μl of a mixture consisting of, per liter, 9 mole of urea, 2% Nonidet P-40 surfactant, 2% ampholines (pH 3.5–10; LKB Instruments, Inc., Rockville, Md.), and 2% (v/v) 2-mercaptoethanol in distilled deionized water. A 10% α-toluenesulfonyl fluoride (PMSF) solution (15.5 mg/ml ethanol) is added to the sample to inhibit proteolysis. Solubilized samples are centrifuged for 3 min in a microfuge and the resulting supernatant applied onto isoelectric focusing gels. First-dimension gels contained 0.75 ml of pH 3.5–10 ampholines. Isoelectric focusing was done at 700 V for 16 hr and 1200 V for an additional 2 hr. Second-dimension sodium dodecyl sulfate (SDS) gels had an 11.25–13.75% acrylamide gradient (*4*). Twenty gels are processed simultaneously using the DALT apparatus (*5*). Two-dimensional gels were fixed and stained by the silver technique of Merril *et al.* (*6*).

Definitions of Data

Data were collected on the X–Y coordinate locations of study spots using a Sumagraphics bit pad interfaced with a Tecktronics microcomputer. The spots selected for the studies reported here represent a subset of those that could be unambiguously identified on each gel. The location of each selected spot was first determined by placing a photograph of the gel onto the digitizing pad, placing the stylus at the position of each spot, and then recording the displayed X–Y location in a data base of raw measurements. All measurements were recorded in millimeters. Our transformation involved aligning across all gels three reference spots that bound a local region containing the data spot of interest. The three spots bounding the region of interest form a reference trio (RT). The location of the data spots on the jth gel within such a region were transformed according to information on the RT by the linear transformation

$$\mathbf{X}' = A_j\mathbf{X} + \mathbf{b}_j$$

where $\mathbf{X}'$ are the transformed and $\mathbf{X}$ are the untransformed 2 × 1 vectors of X, Y coordinates. The A_j and $\mathbf{b}_j$ summarize the information about the deviations of the three reference spots on the jth gel from the average of the three reference spot locations for all gels considered. The A_j is a 2 × 2 matrix and $\mathbf{b}_j$ is a 2 × 1 vector uniquely determined by solving

$$\mathbf{X}_{i.} = A_j\mathbf{X}_{ij} + \mathbf{b}_j \qquad i = 1,2,3$$

where $\mathbf{X}_{ij}$; $i = 1,2,3$, are the initial positions of the three reference spots on the jth gel and $\mathbf{X}_{i.}$ are the average (X and Y coordinates) of the $\mathbf{X}_{ij}$ over all gels ($j = 1, \ldots$). The RT spots will have identical coordinates on each gel after the transformation. These "RT transformations" have the properties of being simple, linear, and local. They are similar but not identical to the transformations used by Vo *et al.* (*7*). We performed a series of studies of the local variation of spot location attributable to differences among observer, repeated measurements by the same observers, differences among replicates of the same sample of tissue, and differences between samples.

Sampling Design

A first study was undertaken to estimate observer bias and the effect of repeated measurements on the estimate of the bivariate distribution of spot location. One gel, considered to be good by our lab, was selected for scoring. Twenty spots were chosen to be representative of the spectrum of spot sizes, shapes, proximity to streaks, and overlap with other spots that one may encounter on a gel. Each spot was scored on each of 5 days by four observers, two of whom could be classified as being more experienced than the others at scoring gels.

Our second study dealt with the contribution to spot variability of differences between gels run on the same sample of tissue and differences between gels run on samples from different individuals. Two gels were run on each sample of lymphocytes taken from 11 individuals. For this study the gels were scored by one experienced person on each of 3 days. Five regions of the gel, each consisting of three data spots bounded by three reference spots, were selected for scoring based on the criteria that the regions be of approximately the same size and that the location of the regions be distributed over as much of the gel as possible. Prior to statistical analysis each region was transformed to align the reference trio of spots for the region as described above.

Statistical Methods and Results

Contribution of Observers and Repeated Measurements

In the first study the major interest was to estimate observer biases and sources of error variance in the measurement of spot location. Spots chosen varied in size, shape, proximity to streaks, and the degree of overlap with other spots, all of which may increase undesired variability and therefore decrease our ability to detect important biological variation.

Letting $\mathbf{X}_{ijk}$ be the ith observer's vector of X,Y coordinates for spot j and repeated measurement (day) k, the full model chosen was

$$\mathbf{X}_{ijk} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \boldsymbol{\gamma}_{ij} + \boldsymbol{\varepsilon}_{ijk},$$

$$\boldsymbol{\varepsilon}_{ijk} \sim \text{independent } N_2(\mathbf{0},\Sigma_{ij}),$$

$$i = 1,2,3,4; j = 1,2, \ldots, 20; k = 1,2, \ldots, 5.$$

The main spot effects, $\boldsymbol{\beta}_j$, will be very large since the spots chosen ranged over the entire gel. The $\boldsymbol{\alpha}_i$ measure the ith observers' deviation from the spot means averaged over all the spots in the sample and are expected to be very small, since it is unlikely that an observer will be biased in the same direction for all the spots scored. The γ_{ij} measure the ith observers bias for spot j. The model assumes that the ε_{ijk} are independent bivariate normal (denoted N_2) with mean zero and a variance–covariance matrix (Σ_{ij}) that depends on both the observer and the particular spot.

We began by testing the null hypothesis, H_0, that the error variances do not depend on the spot being scored,

$$\mathrm{H}_0\colon\ \varepsilon_{ijk} \sim \text{independent } N_2(\mathbf{0},\Sigma_i) \qquad i = 1,2,3,4.$$

The modified likelihood ratio test uses

$$-80 \ln \lambda^* \sim \text{approx. } \chi^2_{228}$$

(i.e., is approximately distributed as a χ^2 with 228 degrees of freedom)

where $\ln \lambda^* = \sum_{i=1}^{4} \ln \lambda_i^*$,

$$\lambda_i^* = \frac{1}{20} \sum_{j=1}^{20} \{\ln|A_{ij}| - \ln|A_i| + 2 \ln 20\},$$

$$A_{ij} = \sum_{k=1}^{5} (\mathbf{X}_{ijk} - \mathbf{X}_{ij.})(\mathbf{X}_{ijk} - \mathbf{X}_{ij.})', \text{ and}$$

$$A_i = \sum_{j=1}^{20} A_{ij}.$$

However, using an asymptotic expansion of $-m \ln \lambda_i^*$ shows that $m = 64.8$ (instead of $m = 80$) makes terms of order $\mathrm{O}(m^{-1})$ vanish so that

$$-64.8 \ln \lambda^* \sim \text{approx. } \chi^2_{228}$$

is a better approximation (*8*). An approximate p value of 0.58 was obtained, leading us to accept H_0, that the error terms do not depend on which spot was scored.

One may also consider the statistics

$$-64.8 \ln \lambda_i^* \sim \text{approx. } \chi^2_{57}$$

which test H_0 for each observer separately. The smallest of the four p values thus obtained was 0.21, which further reassures us of the homogeneity of error variances across spots.

We next investigated the effect of observers on the error variance by testing the hypothesis

$$\mathrm{H}_\Sigma\colon\ \varepsilon_{ijk} \sim \text{independent } N_2(\mathbf{0},\Sigma)$$

which states that the Σ_i are the same for all the observers. This model was rejected ($p \leqslant 0.001$) by a test analogous to that given above, indicating, that for the data considered, the variance–covariance matrix of measurement errors varies among observers.

We next turned to investigating the nature of the heterogeneity of observer variances. First, measurement errors in the X and Y coordinates were tested for

TABLE 1

ESTIMATES OF X AND Y COORDINATE ERROR VARIANCES FOR EACH OBSERVER AND ANOVAS TO DETERMINE IF OBSERVERS HAVE THE SAME ERROR VARIANCES

Observer i	$\hat{\sigma}^2_{X,i}$	$\hat{\sigma}^2_{Y,i}$	Experienced
1	3.60	4.97	No
2	1.15	2.35	Yes
3	1.69	1.50	Yes
4	2.63	2.79	No

Source	df	MS	F	p value
	ANOVA of log $s^2_{X,ij}$			
Between	3	5.12	8.54	~0.0001
Within	76	0.599		
Total	79			
	ANOVA of log $s^2_{Y,ij}$			
Between	3	2.75	3.39	0.022
Within	76	0.811		
Total	79			

independence. The error model $H_{X,Y}$: "the Σ_i are diagonal" is easily accepted by the usual parametric test. Only 2 of the 80 Σ_{ij} had significant covariance terms at the 0.05 level, where 4 would be expected by chance alone.

Letting $\sigma^2_{X,i}$ and $\sigma^2_{Y,i}$ be the X and Y coordinate variances for the ith observer, we next tested the one-dimensional hypotheses of no differences in the $\sigma^2_{X,i}$ and similarly $\sigma^2_{Y,i}$ among observers,

$$H_{0,X}\colon \sigma^2_{X,1} = \sigma^2_{X,2} = \sigma^2_{X,3} = \sigma^2_{X,4}.$$

Since H_Σ was rejected above, it was expected that $H_{0,X}$ (or $H_{0,Y}$) would also be rejected. Following the suggestion of Scheffé (9), an ANOVA was performed on log $s^2_{X,ij}$ to test $H_{0,X}$, where $s^2_{X,ij}$ is the unbiased estimator of $\sigma^2_{X,ij}$, the ith observer's X coordinate variance for spot j. Table 1 summarizes the results of the ANOVAs: p values of 0.001 and 0.022 were obtained for $H_{0,X}$ and $H_{0,Y}$, respectively. Scheffé-type confidence intervals reveal that differences between the two "experienced" and the two "inexperienced" observers contributed to the rejection of these hypotheses, the experienced observers having smaller error variances. There were no significant differences between the two observers within each "experience group."

Using the data from the two experienced observers, we estimated the effects of observers and observer by spot interaction for each dimension. Specifically, the X-coordinate model

TABLE 2

THE RELATIONSHIP BETWEEN RUNS, FAMILIES, AND INDIVIDUALS IN THE GEL VARIABILITY STUDY

Family No.	Individual No.	Run No.
1	1	1
2	2	2
3	3	2
4	4,5,6,7	3
5	8	4
6	9,10,11	5

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$
$$\varepsilon_{ijk} \sim \text{independent } N(\mathbf{0},\sigma^2),$$
$$i = 1,2; j = 1,2, \ldots, 20; k = 1,2, \ldots, 5,$$

was used, with observers as a random effect. As expected, the spot effects, β_j, were very large and the main observer bias effects, α_i, were very small. The usual unbiased estimator for the interaction variance component was negative for the X coordinate and 4×10^{-5} for the analogous Y-coordinate model.

Several points emerge from this analysis concerning the design of the next experiment. First, the assumption that the measurement errors are independent of the spot selected for study seems reasonable. Second, the X- and Y-coordinate measurement errors are independent of each other. Finally, observers agreed on the location of each spot, so that differences observed in measurements of spot locations on different gels would be due only to "treatment" effects (plus a random error) and not observer bias. This last point justifies the use of only a single observer in the second study reported below.

Gel-to-Gel Variability

Next we studied 11 individuals from six families each of whom had duplicate gels run from the same sample. The gels are made in "runs" of 20 gels and variability between runs is expected. Duplicate gels were made by dividing a sample in half and then randomly assigning each subsample to a gel, always within the same run of 20 gels. Most of the gels in each run had no duplicates and were not included in the study. Table 2 shows the relationship between runs, families, and individuals.

The main purpose of this second study was to estimate the variability in spot location among gels attributable to samples and duplicates. Because of the small number of gels on each run, a balanced design in which families are nested within runs and with duplicate gels on each individual would leave few degrees of freedom for estimating the relevant variance components. Thus, partitioning out biologically caused "family" variability and technically caused

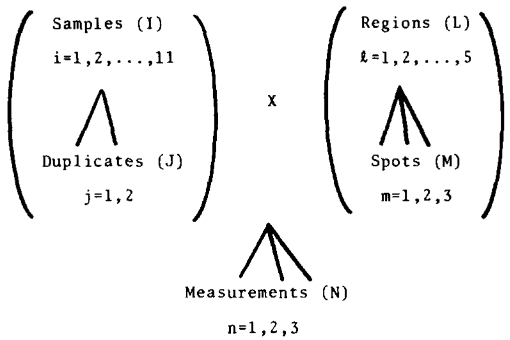

FIG. 1. Design and subscripts used in the gel variability study.

"run" variability will be left for a future study (the design of which will be aided by the results of the current study). We will use the term "sample" instead of "individual" to aid in recalling that sample differences will be caused by a mixture of individual, family, and run effects. Figure 1 depicts the nesting and crossing in the design, as well as the associated subscripts used in the model.

The (X-coordinate) model is

$$X_{ijlmn} = \mu + \alpha_l^{\mathrm{L}} + \alpha_{lm}^{\mathrm{M}} + a_i^{\mathrm{I}} + a_{ij}^{\mathrm{J}} + a_{il}^{\mathrm{IL}} + a_{ilm}^{\mathrm{IM}} + a_{ijl}^{\mathrm{JL}} + a_{ijlm}^{\mathrm{JM}} + \varepsilon_{ijlmn}.$$

Here μ, α_l^{L} (regions), and α_{lm}^{M} (spots within regions) are fixed effects. These are also large and uninteresting effects, indicating only that the different spots selected for scoring do not occupy the same position on the gel. Table 3 presents the ANOVA for the remaining effects and the (unbiased) variance components estimates for both coordinates. Except for the main "sample" effect in the X coordinate, the F tests for the variance components are fairly significant ($p < 0.05$).

Only a relatively small amount ($\simeq$10% here) of the total variability is due to the main effects of samples (σ_{I}^2) and duplicates (σ_{J}^2). With our RT transforma-

TABLE 3

RESULTS FROM THE ANOVAS OF OUR OBSERVATIONS ON X AND Y COORDINATES OF SPOTS, COMPARING SAMPLES, DUPLICATES, REGIONS, AND SPOTS AND THEIR INTERACTIONS

Source	df	MS		F test expression	F		$\hat{\sigma}^2$		% total $\hat{\sigma}^2$	
		X	Y		X	Y	X	Y	X	Y
(I) Samples	10	0.656	0.753	$\mathrm{MS_I/MS_J}$	2.13[a]	7.82[c]	0.0039	0.0073	2.9	7.6
(J) Duplicates	11	0.307	0.096	$\mathrm{MS_J/MS_E}$	16.8	5.55	0.0064	0.0017	4.8	1.8
(IL) Sample × Region	40	0.544	0.399	$\mathrm{MS_{IL}/MS_{JL}}$	1.98[b]	2.96	0.0149	0.0146	11.2	15.2
(JL) Duplicates × Region	44	0.275	0.135	$\mathrm{MS_{JL}/MS_E}$	15.0	7.77	0.0285	0.0131	21.4	13.6
(IM) Sample × Spot	100	0.306	0.230	$\mathrm{MS_{IM}/MS_{JM}}$	3.10	4.14	0.0345	0.0291	25.9	30.3
(JM) Duplicates × Spot	110	0.099	0.056	$\mathrm{MS_{JM}/MS_E}$	5.38	3.21	0.0268	0.0128	20.1	13.3
(E) Error	660	0.0183	0.0173				0.0183	0.0173	13.8	18.2
							0.133	0.096		

[a] $p = 0.12$; [b] $p = 0.014$; [c] $p = 0.0011$; all other $p < 0.0005$.

TABLE 4

UNBIASED VARIANCE COMPONENT ESTIMATES FOR X AND Y COORDINATES ESTIMATED SEPARATELY FOR EACH REGION

	Region				
Source	1	2	3	4	5
	X-Coordinates				
(I) Samples	0.010	0.006	0.006	0.009	0.048
(J) Duplicates	0.030	0.010	0.056	0.013	0.038
(IM) Sample × Spot	−0.002	0.026	0.032	0.033	0.083
(JM) Duplicates × Spot	0.014	0.014	0.008	0.025	0.073
(E) Error	0.013	0.023	0.019	0.018	0.019
"Total"	0.064	0.079	0.121	0.098	0.261
	Y-Coordinates				
(I) Samples	0.000	0.003	0.074	0.005	0.010
(J) Duplicates	0.023	0.004	0.012	0.011	0.011
(IM) Sample × Spot	0.013	0.076	0.023	0.036	−0.003
(JM) Duplicates × Spot	0.004	0.015	0.004	0.011	0.030
(E) Error	0.024	0.013	0.011	0.015	0.024
"Total"	0.064	0.111	0.124	0.079	0.072

tions this is not surprising, for these effects measure the extent to which all of the spots are shifted (together) in the same direction on a particular gel (σ_J^2) or sample (σ_I^2). Lumping the three sample-related effects (I,IL,IM) and duplicate related effects (J,JL,JM) effects together, we may summarize as follows. For the X coordinates (respectively, Y coordinates) 40% (53%) of the total spot variability was due to sample-type effects, 46% (29%) was due to duplicate type effects, and 14% (18%) was due to measurement error. It should be kept in mind that different matching schemes may use different types of transformations, so that the size of various effects may differ among laboratories.

Table 4 gives variance component estimates for each region separately. Generally, the size of the variance component estimates vary considerably over the regions. Note that the effects for the X coordinates in region 5 are quite large. This region was chosen at the basic end of our gels where experience had already indicated that X coordinate distortions were large. These distortions are apparently not linear as our transformations did not remove their effects. Another striking aspect is the large σ_I^2 in region 3 for the Y coordinates (and also the σ_{IM}^2 in region 2). Closer inspection of the sum of squares in these cases indicates that much of what has appeared as sample variability is in fact caused by differences in runs 3 and 5 (or equivalently, families 4 and 6). Whether the between-samples variability is mostly due to biological "individual" or "family" factors rather than a technical "run" effect remains a basic question that only a larger sample would answer.

Discussion

A number of recommendations for designing spot-matching algorithms are suggested from the analyses presented here. Since most gel-matching strategies are based on algorithms that match two gel images at a time, we will consider the implications in this context. Here a gel image consists of a list of spot locations along with some measure of spot size. In order to simplify notation in what follows, we will write models only for a single spot and assume that each gel has the spot under consideration.

Given a sample of gels, A,B,C . . . , to be matched, present matchers proceed basically in one of two ways. One may match gels A and B, B and C, etc., and assume that the matching is transitive. That is, if spot a on gel A matches spot b on gel B, and b is found to match spot c on gel C, then a and c will be considered to match (*7*). Alternatively one could match each gel to a "standard" or "reference" gel S, matching the pairs S–A, S–B, S–C . . . , and again assume transitivity (*10, 11*). The standard gel image, S, can be updated in the light of matches made to other gels (*11*). We will investigate variants of a procedure in which the standard gel image is updated at the conclusion of every gel-matching routine. Calling such an image a "composite" gel image, one could match gels A and B, form a composite of A and B which is then matched to C, and so forth. One could also match two groups of gels by matching within each group and then match the composites for each of the groups with each other.

Using our RT transformations, a composite of a sample of gels can be precisely defined. Each spot on the composite will be located at the average location of that spot for the gels in the sample that possess it, the average being calculated after the transformation has been performed. Using this definition along with knowledge of the partitionable nature of the variability in spot location will allow us to compare the various possible schemes for using composite gel images.

Effects of Duplicates

Consider the observations for a spot on two duplicate gels for each sample of tissue defined by the model

$$Y_{ij} = \mu + \alpha_i^{\mathrm{I}} + \beta_{ij}^{\mathrm{J}}$$

where Y_{ij} is a measurement of the location on sample i, duplicate j. Here the α_i^{I} term absorbs the I, IL, and IM terms of the previous model while β_{ij}^{J} absorbs the J,JL,JM, and error terms. If repeated measurements have been taken, average them and divide the measurement error term by the appropriate factor; the model remains as above though the β_{ij}^{J} terms will be smaller. At present, assume no family or run effects.

Selecting one gel from each pair of duplicate gels (or not running duplicates to begin with) and matching across samples, one faces the variance

$$V(Y_{ij} - Y_{i'j'}) = 2\sigma_{\mathrm{I}}^2 + 2\sigma_{\mathrm{J}}^2 \qquad [1]$$

TABLE 2

COMPARISON BETWEEN DIAGNOSIS AND PROGRAM SUGGESTION (REVISED VERSION)

Disease entity	A[a]	B	C	D
Hypovolemia	5	0	0	58
Prerenal acute renal failure	4	0	0	59
Acute nephritic syndrome	5	0	0	58
Hepatic nephropathy	0	0	0	63
Wegener's granulomatosis	2	0	0	61
Systemic lupus erythematosus	0	0	0	63
Polyarteritis nodosa	0	0	0	63
Henoch–Schönlein purpura	1	0	0	62
Goodpasture's syndrome	2	0	0	61
Progressive systemic sclerosis	0	0	0	63
Essential mixed cryoglobulinemia	0	0	0	63
Upper urinary tract infection	2	0	0	61
Disseminated intravascular coagulation	0	0	0	63
Preeclampsia	0	0	0	63
Eclampsia	0	0	0	63
Total bilateral renal artery occlusion	1	0	0	62
Renal infarction	0	0	0	63
Acute noninfectious interstitial nephropathy	4	0	0	59
Acute uric acid nephropathy	0	0	1	62
Hemolytic–uremic syndrome	0	0	0	63
Thrombotic thrombocytopenic purpura	0	0	0	63
Dysproteinemia	3	0	0	60
Hypercalcemic nephropathy	0	0	1	62
Obstructive uropathy	3	0	0	60
Septicemia	18	0	1	44
Acute tubular necrosis[b]	42	0	1	20
Malignant hypertension	1	0	0	62
Chronic azotemia	1	0	0	62
Urinary retention	0	0	0	63
	94	0	4	

[a] A, true positive suggestions; B, false negative suggestions; C, false positive suggestions; D, true negative suggestions.

[b] Including toxic and pigment-induced nephropathy.

DISCUSSION

The present study demonstrates that a computer program based on the binary-choice task formulation strategy may be designed to imitate the diagnostic reasoning process of humans. Diagnostic definitions are arbitrary and therefore the task of the programmer is to copy the diagnostic criteria of the user of the program, and not to make "correct" diagnoses. Diagnostic criteria are changed from time to time and program design should be adjustable to allow the diagnostic criteria of the program to be changed.

being matched. Sequential matching is at a disadvantage because it ignores the grouping of gels into runs.

Using the Nesting/"Nested Sequential" Matching

Instead of matching in a strictly sequential manner one could first match sequentially in each run (or family) and then match composite images of the runs. If $\bar{Y}$ is the average of the first m individual's spots, where the individuals are nested within a run, then in matching to the m + 1st individual in that run we have

$$V(\bar{Y} - Y_{v,m}) = \left(\frac{1}{m} + 1\right) \sigma_{\mathrm{I}}^2, \qquad m = 1,2, \ldots, R \qquad [7]$$

which is the same as [6] when $k = 1$. Now letting

$$Y_{v,.} = \frac{1}{R}\sum_{i=1}^{R} Y_{v,i} \quad \text{and} \quad \bar{Y} = \frac{1}{k}\sum_{v=1}^{k} Y_{v,.}$$

then in the kth stage of matching across runs we have

$$V(\bar{Y} - Y_{k+1,.}) = \left(\frac{1}{k} + 1\right)\left(\sigma_{\mathrm{V}}^2 + \frac{1}{R}\sigma_{\mathrm{I}}^2\right), \qquad [8]$$

which is asymptotically $\sigma_{\mathrm{V}}^2 + (1/R)\sigma_{\mathrm{I}}^2$.

If $\sigma_{\mathrm{V}}^2 \ll \sigma_{\mathrm{I}}^2$ then [6] is smaller than [7] in the long run, but as σ_{V}^2 increases, using the nesting information becomes increasingly more important. For $\sigma_{\mathrm{V}}^2 \geq \sigma_{\mathrm{I}}^2$ using "nested sequential" matching is better at every stage.

The analysis of the second experiment indicates that both sample and gel-to-gel (duplicate) effects are large. Some of the sample effects are presumed to be due to biological differences that cannot be reduced by technical methods. Reducing the differences between gels, on the other hand, can be accomplished by more accurately standardizing the pouring, loading, and running of gels within a run. However, matching between duplicate gels to form composites is a possible image-processing strategy that theoretically reduces these nonbiological effects by one-half.

We have also presented, in the discussion of Table 4, that the size of the variance components differs for different regions of the gel. A model regressing the expected variance on the X,Y location might be expected to capture much of this information. Alternatively, the metric used to measure the distance between spots from different gels could be weighted in the X and Y coordinates in a region-dependent manner, for example, taking into account greater X coordinate variability at the basic end of the gel.

One of the goals of most 2-D PAGE laboratories is to minimize the effect of runs, and innovations such as computerized gel pouring and the commercial pouring of large batches of gels will help in this endeavor. Still, gels will need to be isofocused and electrophoresed in batches, which will continue to contribute to a nonzero run effect. Although at present we have no rigorous estimates of the relative sizes of σ_{V}^2 and σ_{I}^2 in the above model, our experience indicates that

σ_V^2 is large enough in our laboratory to warrant using the information about which run a gel belongs to.

Acknowledgments

We thank Leslie Baier and Dave Sing for technical assistance. This work was supported by grants from the Department of Energy (AC02-82ER60089) and the National Institutes of Health (PO1-CA26803 and RO1-CA32146).

References

1. Neel, J. V. A revised estimate of the amount of genetic variation in human proteins: Implications for the distribution of DNA polymorphisms. *Amer. J. Hum. Genet.* **36,** 1135 (1984).
2. Neel, J. V., Rosenblum, B. B., Sing, C. F., Skolnick, M. M., Hanash, S. M., and Sternberg, S. Adapting two-dimensional gel electrophoresis to the study of human germline mutation rates. *In* "Two-Dimensional Gel Electrophoresis of Proteins" (J. E. Celis and R. Bravo, Eds.), pp. 259–306. Academic Press, New York, 1984.
3. Sing, C. F., Boerwinkle, E., and Moll, P. P. Strategies for elucidating the phenotypic and genetic heterogeneity of a chronic disease with a complex etiology. *In* "Diseases of Complex Etiology in Small Populations. Ethnic Differences and Research Approaches" (R. Chakraborty and E. Szathmary, Eds.), pp. 39–66. Alan R. Liss, New York, 1985.
4. Rosenblum, B. B., Hanash, S. M., Yew, N., and Neel, J. V. Two-dimensional electrophoretic analysis of erythrocyte membranes. *Clin. Chem.* **28,** 925 (1982).
5. Anderson, N. L., and Anderson, N. G. Analytical techniques for cell fractions. XXII. Two-dimensional analysis of serum and tissue proteins. Multiple gradient–slab gel electrophoresis. *Anal. Biochem.* **85,** 341 (1978).
6. Merril, C. R., Switzer, R. C., and van Keuren, M. L. Trace polypeptides in cellular extracts and human body fluids detected by two-dimensional electrophoresis and a highly sensitive silver stain. *Proc. Natl. Acad. Sci. USA* **76,** 4335 (1979).
7. Vo, K. -P., Miller, M. J., Guiduschek, E. P., Nielsen, C., Olson, A., and Nguyen, H. X. Computer analysis of two-dimensional gels. *Anal. Biochem.* **112,** 258 (1981).
8. Muirhead, R. J. "Aspects of Multivariate Statistical Theory." Wiley, New York, 1982.
9. Scheffé, H. "The Analysis of Variance." John Wiley and Sons, New York, 1959.
10. Lemkin, P. F., Lipkin, L. E. A computer system for two-dimensional gel electrophoresis analysis. III. Multiple two-dimensional gel analysis. *Comput. Biomed. Res.* **14,** 407 (1981).
11. Garrels, J. I., Farrar, J. T., and Burwell, C. B. IV. The QUEST system for computer-analyzed two-dimensional electrophoresis of proteins. *In* "Two-Dimensional Gel Electrophoresis of Proteins" (J. E. Celis and R. Bravo, Eds.), pp. 37–91. Academic Press, New York, 1984.