Applications of Coding Theory to the Design of Somatic Cell Hybrid Panels

TUSHAR MADHU GORADIA
Division of Health Sciences and Technology,
Harvard University–Massachusetts Institute of Technology, Boston, Massachusetts 02115
and Department of Biomathematics, UCLA School of Medicine, Los Angeles, California 90024

AND

KENNETH LANGE*
Department of Biomathematics, UCLA School of Medicine, Los Angeles, California 90024

Received 28 August 1987; revised 11 May 1988

ABSTRACT

The somatic cell hybridization technique for gene mapping depends on assembling panels of rodent-human hybrid clones containing random subsets of the human chromosomes. Such panels should be as informative as possible and permit error detection and error correction for assays of the human gene in the various clones. We derive estimates of the number of randomly generated clones required to be reasonably confident of accurately and unambiguously assigning a gene to a particular human chromosome. The collection of clones in such a random panel is contrasted with minimal panels suggested by algebraic coding theory. To approximate minimal panels we suggest the method of simulated annealing for selecting small, informative panels from larger existing collections of clones. These theoretical insights emphasize the need for more collaboration and coordination among gene mapping groups so that optimal clone panels can be assembled, stored, and distributed.

1. INTRODUCTION

Human gene localization by somatic cell hybridization has been actively applied for two decades since the pioneering work of Weiss and Green [43]. In brief outline, the method involves the formation of interspecies hybrid cells by fusing normal diploid human somatic cells with permanently transformed rodent cells—usually mouse or Chinese hamster [6, 18]. The result-

*To whom correspondence should be addressed.

655 Avenue of the Americas, New York, NY 10010 0025-5564/88/$03.50

ing hybrid somatic cells retain all of the rodent chromosomes while losing a random subset of the human chromosomes. A few generations after fusion, clones can be identified with stable subsets of the human chromosomes. All chromosomes, human and rodent, normally remain functional. With a broad enough collection of different hybrid clones, it is possible to establish a correspondence between the presence or absence of a given human gene and the presence or absence of each of the 24 distinct human chromosomes. From this pattern one can infer the particular chromosome on which the gene resides. With the above outline in mind, it is of interest to determine the minimal number of distinct hybrid clones required to accurately and unambiguously assign a gene to a particular human chromosome. Such collections or panels of hybrid clones ideally should be designed to detect and correct a small number of errors in assays for the human gene or its protein product. By using some ideas from algebraic coding theory and probability, we contrast such minimal panels with the random panels typically generated by molecular geneticists.

In developing some of the logical and mathematical consequences of the somatic cell hybrid technique, it seems prudent to state explicitly its underlying assumptions. Some violations of these assumptions will be noted below. The major assumptions are:

(a) The human gene G to be mapped is present on exactly one chromosome.

(b) Any rodent analogue of G is distinguishable from G at either the protein or the DNA level.

(c) Each of the 24 distinct human chromosomes (22 autosomes and the X and Y sex chromosomes) is either absent from a clone or is cytologically or biochemically detectable in the clone.

(d) All cells within a clone share the same chromosome constitution.

(e) The presence or absence of G can be accurately detected in each clone.

Although assumption (a) is generally satisfied, a few human genes are scattered at multiple dispersed sites within the human genome; the gene encoding ribosomal RNA is a case in point [9]. Assumption (b) can be fulfilled if the human gene G and its rodent analogue are distinguishable by either electrophoresis of their gene products or by hybridization of an appropriate human DNA probe. Assumption (c) is easily met because human chromosomes can be accurately identified by a combination of human specific isozyme marker assays and karyotypic analysis using standard banding procedures. Violations of this condition usually result from the presence of unrecognized chromosomal aberrations such as insertions, deletions, translocations, and fragmentations. However, such chromosome aberrations are often intentionally employed for the regional mapping of genes on a specific chromosome. Assumption (d) can be violated if some of the cells in a clone continue to undergo human chromosome loss. For this reason several cells from a clone should be karyotyped. It is the maximal subset of human chromosomes in the clone which is relevant to gene localization. As a safeguard, ambiguous clones should be disregarded. The last assumption, (e), causes the most trouble. For instance, ambiguities can arise when phenotypic polymorphism in structural genes cannot be distinguished from phenotypic polymorphism in associated regulatory genes [6]. In addition, not all genes are constitutive in the sense that they are expressed at all times in all cell types [32]. Use of DNA probes to detect the human gene neatly circumvents both of these problems. Finally, laboratory error can enter into both enzyme and probe hybridization assays.

The method of *in situ* hybridization carries the technique of human probe hybridization one step further. If a human cell is karyotyped and radioactive grains corresponding to the hybridized probe cluster predominantly on a given chromosome, then the gene is declared to reside on that chromosome. In practice, the independent results of somatic cell hybridization and *in situ* hybridization tend to reinforce each other [22].

In the current paper we will be concerned solely with issues of redundancy and efficiency in the somatic cell hybrid method. It has long been appreciated that certain redundancies in panels of somatic cell hybrid clones can self-detect and self-correct phenotyping errors representing violations of assumption (e) [17]. We will attempt to explain the utility of these error detection and correction capabilities as well as the amount of effort necessary to generate at random panels of clones for such purposes. These randomly generated panels we contrast with minimal panels suggested by algebraic coding theory. We also provide a practical solution to the combinatorial problem of selecting small, informative panels from larger existing collections of clones. As examples, we select good panels of sizes 5 through 20 clones from 189 published clones.

2. MATHEMATICAL MODELS FOR THE DESIGN AND GENERATION OF HYBRID CLONE PANELS

To model mathematically the design of hybrid clone panels, we borrow and reinterpret some concepts from communications engineering. A modicum of notation is required. Let n denote the number of distinct hybrid clones in a panel. Since in females the 22 autosomes and the X chromosome occur in homologous pairs, and since the Y chromosome bears few genes of interest, we focus on clones derived from human female cells. We may construct a karyotype matrix K consisting of n rows and 23 columns. The entry in row i and column j of K is 1 if clone i contains chromosome j; otherwise it is 0. See Figure 1 for an example. Note that the X chromosome

$$
\begin{matrix}
0&1&0&1&0&0&0&1&0&0&0&0&0&0&1&0&1&1&0&1&1&1&1\\
1&0&1&0&1&1&0&0&1&0&0&0&0&1&0&0&1&0&1&0&1&1&1\\
0&1&1&1&1&0&1&0&0&0&0&0&1&0&0&1&1&0&1&1&0&1&1\\
1&1&1&0&0&1&1&0&0&1&0&1&0&0&0&1&1&1&0&0&1&0&1\\
0&0&0&1&1&1&1&0&0&0&1&1&1&1&1&0&1&0&0&0&1&1&0\\
0&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&1&0&0&0&0&0&0\\
0&0&1&0&1&0&1&1&0&1&1&1&0&0&0&0&1&1&1&1&1&0&0\\
0&0&0&1&0&1&1&1&0&0&0&1&0&1&1&1&1&0&1&0&1&0&1\\
1&0&0&0&1&1&0&0&0&1&0&1&1&0&1&0&1&0&1&1&0&0&1
\end{matrix}
$$

FIG. 1. An example of an optimal karyotype matrix K with every column at least a distance 3 from every other column. The rows correspond to clones, and the columns correspond to chromosomes.

is counted as chromosome number 23. A convenient means of comparing two columns c_s and c_t of K is provided by the Hamming distance $\rho(c_s, c_t)$ [14, 20]. This is defined as the number of entries in which c_s and c_t differ. For instance in Figure 1, $\rho(c_1, c_2) = 5$ and $\rho(c_1, c_3) = 4$. It is easy to check that ρ satisfies the mathematical conditions for a metric, namely

$$
\begin{gathered}
\rho(c_s, c_t) = \rho(c_t, c_s), \\
\rho(c_s, c_t) = 0 \quad \text{if and only if} \quad c_s = c_t, \\
\rho(c_r, c_t) \leqslant \rho(c_r, c_s) + \rho(c_s, c_t) \qquad \text{(triangle inequality)}.
\end{gathered}
$$

Besides comparing columns of K, it is appropriate to compare the columns of K with the results of testing for a given gene G in the different hybrid clones. We can construct a phenotype column vector p whose ith entry is 1 when the ith hybrid clone contains G; otherwise it is 0. From the assumptions (a) through (e) of the introduction, G can reside on chromosome r only if $p = c_r$. If c_r is distinct from all other columns of K, then G can be assigned to chromosome r. In terms of the Hamming distance, this is equivalent to the two conditions $\rho(p, c_r) = 0$ and $\rho(c_s, c_r) > 0$ for all $s \neq r$. As the number n of randomly generated clones increases, satisfying these two conditions for unique chromosome assignment becomes more likely.

In practice, errors can occur in detecting G in the various hybrid clones. These errors affect p and are probably more common than errors affecting the definition of the karyotype matrix K. (Karyotype errors are considered in the discussion.) Let p_{obs} represent the observed phenotype test results for the different hybrid clones. The number of phenotyping errors is $\rho(p, p_{\text{obs}})$. Some of these errors will be false positives in detecting G, and some will be false negatives. If p and p_{obs} are identical, then there are no phenotyping errors.

Sufficient redundancy in K can compensate for a limited number of errors in p_{obs}, as two well-known propositions from coding theory demonstrate [14]. The first deals with the ability to detect errors. Suppose G lies on chromosome r and we know *a priori* that the number of errors $\rho(p, p_{\text{obs}}) \leqslant m$ for some positive integer m. If the minimal Hamming distance to column c_r satisfies

$$\min_{s \neq r} \rho(c_r, c_s) > m, \tag{1}$$

then the fact there are errors in p_{obs} is detectable. The idea of the proof consists in showing that p_{obs} is incompatible with G residing on any chromosome if $\rho(c_r, p_{\text{obs}}) > 0$. Indeed, consider any chromosome s different from r. Then by the triangle inequality and the condition (1),

$$\begin{aligned} \rho(p_{\text{obs}}, c_s) &\geqslant \rho(c_r, c_s) - \rho(c_r, p_{\text{obs}}) \\ &= \rho(c_r, c_s) - \rho(p, p_{\text{obs}}) \\ &> m - m \\ &= 0. \end{aligned}$$

Hence, p_{obs} cannot coincide with any column, and there must be at least one error.

Although error detection can alert us to inconsistencies, it will not remedy them. For error correction, suppose we know *a priori* that the number of errors $\rho(p, p_{\text{obs}}) \leqslant m$ for some positive integer m. Assuming again that G resides on chromosome r, the more stringent condition

$$\min_{s \neq r} \rho(c_r, c_s) > 2m, \tag{2}$$

permits error correction. To prove this proposition one needs to argue that $\rho(p, c_s) > 0$ for any chromosome s different from r. By the triangle inequality and the condition (2),

$$\begin{aligned} \rho(p, c_s) &> \rho(c_r, c_s) - \rho(c_r, p_{\text{obs}}) - \rho(p_{\text{obs}}, p) \\ &> 2m - m - m \\ &= 0. \end{aligned}$$

As a consequence one can infer that G resides on r.

Because we do not know in advance what chromosome G lies on, it is useful to construct karyotype matrices for which the condition (1) or (2) holds for all possible columns r. The matrix K in Figure 1 furnishes some examples. The first five rows alone permit correct gene placement in the absence of errors, since all columns are distinct. In fact, these columns just represent the binary expansions of various numbers between 0 and 31. The first six rows of Figure 1 have all column pairs a distance 2 or more apart. Hence, these six rows permit detection of one error regardless of which chromosome G lies on. The sixth row was constructed by forcing the column sums of the first six rows to be even. The pairwise column distances for the whole matrix in Figure 1 are always at least 3. Hence, all nine rows permit the correction of a single phenotyping error. There are no simple algorithms to construct this and more complicated examples, but many such matrices have been published in the coding theory literature [27, 36, 42].

The karyotype matrix in Figure 1 has much better error detecting and correcting properties than most random karyotype matrices of the same size. Our next aim is to investigate the number of random clones a laboratory geneticist would have to generate to achieve comparable results. To facilitate our analysis we require some more definitions and the introduction of simplifying assumptions. To begin with, we now view the intercolumn distances $\rho(c_s, c_t)$ as random variables X_{st}; the number of clones is still fixed at n. The joint distribution of the X_{st} over all column pairs $\{s, t\}$ can be well approximated under the following assumptions:

(f) Human chromsomes are lost independently of one another during the formation of a stable clone.

(g) The probability that at least one member of a homologous pair of human chromosomes is retained by a clone is $\frac{1}{2}$.

(h) The chromosome complements of different clones are independently determined.

These assumptions are almost certainly false in any strict accounting [33]. However, they are conservative assumptions in the sense that departures from them will result in panels with less information content on the average. In other words, generating good random panels is more difficult if they are violated. Assumption (g) represents a compromise motivated by the range of chromosome retention probabilities of .07 to .75 published by Rushton [33]. Assumption (h) can be almost guaranteed if different clones are generated by fusing different parental cell lines.

As a consequence of assumptions (f) through (h), the entries of the karyotype matrix K are independent random variables equally likely to take the values 0 or 1. From this, it is evident that each random distance X_{st} follows a binomial probability distribution

$$\begin{aligned} P(X_{st} = m) &= \binom{n}{m}\left(\tfrac{1}{2}\right)^m \left(\tfrac{1}{2}\right)^{n-m} \\ &= \binom{n}{m}\left(\tfrac{1}{2}\right)^n. \end{aligned}$$

Due to the central limit theorem, X_{st} will be approximately normally distributed even for n as small as 10 [10]. Less obvious is the fact that the X_{st} are uncorrelated. If collectively the X_{st} actually followed a multivariate normal distribution, lack of correlation among the X_{st} would imply they were independent [29]. We will exploit this near-independence momentarily. Returning to the problem of showing that they are uncorrelated, we note that when two pairs $\{s,t\}$ and $\{u,v\}$ do not overlap, it is intuitively obvious that X_{st} and X_{uv} are independent. Independence is a stronger property than lack of correlation. If the pairs $\{s,t\}$ and $\{u,v\}$ share one column in common, say $t=u$, then X_{st} and X_{tv} are still independent. This becomes clear when one conditions on the outcome of column c_t. It must be emphasized here that assumption (g) is critical. Only a retention probability of $\frac{1}{2}$ is consistent with independence. Even larger subsets of the X_{st} are independent. For instance, the collection of random distances from column r, $\{X_{rs} : s \neq r \text{ and } r \text{ fixed}\}$, is independent. Again this follows by conditioning on column c_r. However, it is false that the whole collection of X_{st} is independent. This subtlety enormously complicates the exact calculation of probabilities.

With these preliminaries, it is possible to approximate the probability distributions of two important random variables. N_d' denotes the random number of clones required for a fixed column r of K to be a distance d or greater from all other columns of K. N_d is the random number of clones required for all pairs of columns to be a distance d or greater apart. N_d is more relevant than N_d' when a laboratory group intends to map a large number of different genes using the same panel. N_d' is appropriate for mapping a single gene.

The distributions of N_d' and N_d can be derived using the random variables X_{st}. Thus,

$$\begin{aligned} P(N_d' \leqslant n) &= P\Big(\min_{s \neq r} X_{rs} \geqslant d\Big) \\ &= \prod_{s \neq r} P(X_{rs} \geqslant d) \\ &= \left[1 - \sum_{m=0}^{d-1} \binom{n}{m} \frac{1}{2^n}\right]^{c-1}, \end{aligned} \tag{3}$$

where $c = 23$ is the number of columns. The formula (3) is exact. The approximate distribution of N_d is

$$\begin{aligned} P(N_d \leqslant n) &= P\Big(\min_{\{s,t\}} X_{st} \geqslant d\Big) \\ &\simeq \prod_{\{s,t\}} P(X_{st} \geqslant d) \\ &= \left(1 - \sum_{m=0}^{d-1} \binom{n}{m} \frac{1}{2^n}\right)^{\binom{c}{2}}, \end{aligned} \tag{4}$$

TABLE 1

Summary Statistics for N_d' [a]

			Percentiles			
d	Mean	S.D.	50th	95th	99th	99.9th
1	5.8	1.8	6	9	12	15
2	9.0	2.1	9	13	16	19
3	11.9	2.4	12	16	19	23
4	14.7	2.6	14	19	22	26
5	17.4	2.7	17	22	25	30
6	19.9	2.9	20	25	28	33
7	22.5	3.0	22	28	31	36

[a] N_d' denotes the random number of distinct hybrid clones required to achieve a karyotype matrix with column r at least a distance d from all other columns. Because N_d' is discrete, we define the αth percentile as the first integer n such that $P(N_d' \leqslant n) \geqslant \alpha/100$.

where

$$\binom{c}{2} = \frac{c(c-1)}{2}$$

is the number of pairs of columns. We have derived a number of upper and lower bounds for $P(N_d \leqslant n)$ which lend support to the approximation (4), but the derivations of these bounds are too involved to present here.

The moments of N_d' and N_d can be computed from the formula

$$E(Z^m) = \sum_{n=0}^{\infty} \left[(n+1)^m - n^m\right] P(Z > n)$$

for the mth moment of a random variable in terms of its right tail probabilities [10]. Tables 1 and 2 display the mean, standard deviation, and representative percentiles for N_d' and N_d. Also recorded are some theoretical lower bounds for N_d given in the coding theory literature [2]. d is the obvious lower bound for N_d'.

One of the side effects of employing panels with large numbers of clones is that we increase the expected number of gene detection errors. A rigorous analysis of the chances for correct gene placement should take this fact into account. To model errors in the phenotype column vector p_{obs}, suppose they occur independently in the various clones and have common rate q. The total number of errors will then be binomially distributed. If there are m such errors, and G resides on chromosome r, then we can correct the errors

TABLE 2

Summary Statistics for N_d [a]

				Percentiles			
d	Mean	S.D.	Min.	50th	95th	99th	99.9th
1	9.3	1.9	5	9	13	15	18
2	13.1	2.1	6	13	17	19	23
3	16.4	2.2	9	16	20	23	27
4	19.5	2.4	10	19	24	27	31
5	22.4	2.5	11	22	27	30	34
6	25.3	2.6	12	25	30	33	37
7	28.1	2.7	15	28	33	36	4

[a] N_d denotes the random number of distinct hybrid clones required to achieve a karyotype matrix with every column at least a distance d from every other column. Because N_d is discrete, we define the αth percentile as the first integer n such that $P(N_d \leqslant n) \geqslant \alpha/100$.

provided

$$\min_{s \neq r} X_{rs} > 2m.$$

This is just the condition (2). Thus the probability of correct gene placement given n clones reduces to

$$\begin{aligned} &P(\text{correct gene placement} \mid n \text{ clones}) \\ &= \sum_{m=0}^{n} \binom{n}{m} q^m (1-q)^{n-m} P\Big(\min_{s \neq r} X_{rs} > 2m\Big) \\ &= \sum_{m=0}^{n} \binom{n}{m} q^m (1-q)^{n-m} \left[1 - \sum_{j=0}^{2m} \binom{n}{j} \frac{1}{2^n}\right]^{c-1}, \end{aligned}$$

with $c = 23$. Figure 2 plots this probability versus n for various values of q. For instance, with $q = 0.01$, about 12 randomly generated clones suffice to place a given gene with 95% certainty. Also, for absurdly large q, e.g., $q = 0.5$, the probability of correct gene placement diminishes as more randomly generated clones are added.

As a final comment, we note that all the above mathematical results continue to hold when hybrid clones are cultured in a selective medium which promotes the retention of a particular human chromosome. For example, if the parental rodent cells are deficient in the enzyme thymidine kinase and if hybrid cells are cultured in HAT medium, then only those

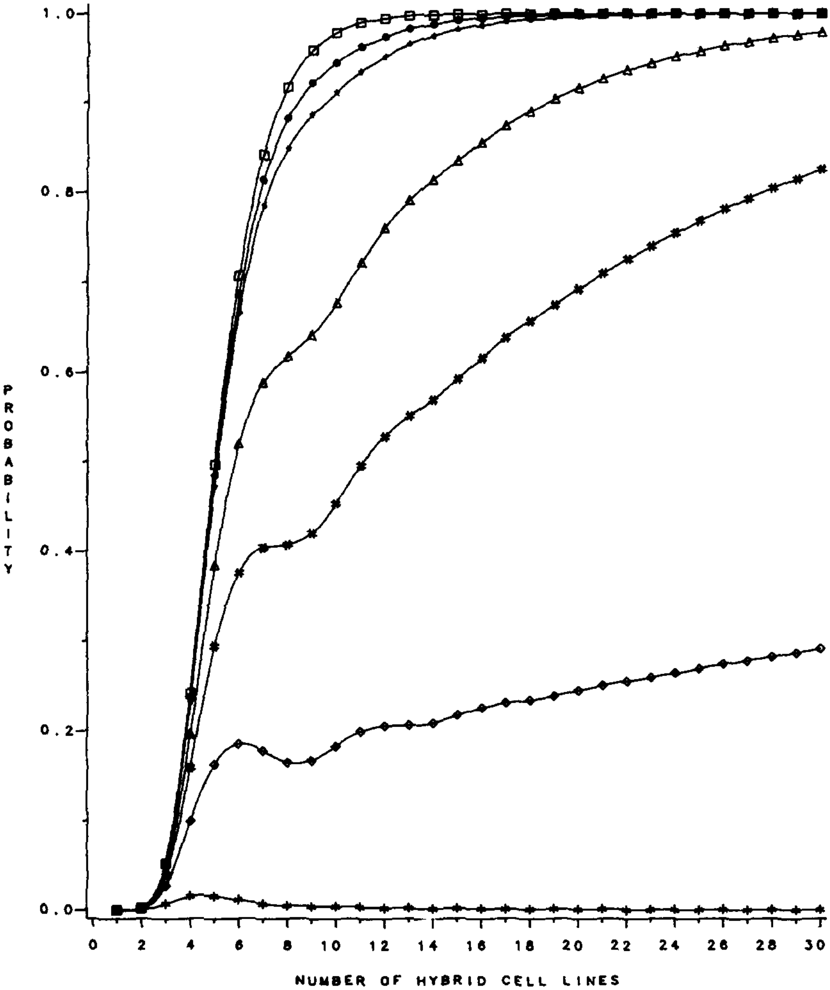

FIG. 2. Family of curves for the probability of correct gene placement with gene-detection error rates of $q = 0.000$ (□), $q = 0.005$ (○), $q = 0.010$ (☆), $q = 0.050$ (△), $q = 0.100$ (#), $q = 0.200$ (◇), and $q = 0.500$ (∗).

hybrid cells which carry the gene encoding thymidine kinase on chromosome 17 will survive. In this case, column 17 of the karyotype matrix is predetermined to be a column of 1's. Thus, the distribution of N_d' and N_d must be calculated conditional on a column of 1's. However, it is intuitively clear from symmetry considerations that the two events $A = \{$a given column contains all 1's$\}$ and $B = \{N_d' \leqslant n\}$ or $B = \{N_d \leqslant n\}$ are independent. This translates into

$$P(N_d' \leqslant n|\text{column of 1's}) = P(N_d' \leqslant n),$$
$$P(N_d \leqslant n|\text{column of 1's}) = P(N_d \leqslant n).$$

3. SELECTION OF GOOD HYBRID PANELS

Given an existing collection of clones, there are two prerequisites for choosing a subset of them to form a small, informative panel. First, some criterion of merit must be established for measuring the information content of each panel. As we have attempted to demonstrate, one reasonable criterion is the minimum Hamming distance between the column pairs of a panel. A refinement of this criterion is to take into account the number of column pairs which attain this minimum distance. Thus, we will adopt the criterion

$$E(\sigma) = -d + \frac{k-1}{\binom{23}{2}}, \tag{5}$$

where σ is a given panel of clones, d is the minimum Hamming distance for σ, k is the number of column pairs attaining this minimum distance, and

$$\binom{23}{2} = 253$$

is the total number of pairs. The best panels have a low value for E. Because of the scaling of the second term in (5), a panel with a higher d will always be preferred to a panel with a lower d. Note that a fixed panel size n is implicit in the definition (5).

Having decided on the criterion E, the next problem is to find a panel which furnishes a minimum or near-minimum of E. Exhaustive enumeration of all possible panels is infeasible. For instance, with 50 clones and a panel size of 12, there are

$$\binom{50}{12} = 1.2 \times 10^{22}$$

possible panels. Since no good deterministic algorithms exist for finding the minimum of E, we will describe three random sampling techniques. All three are implemented by a random exchange mechanism. Given an existing panel, a random clone currently in the panel is selected for exchange with a random clone outside the panel, but in the existing collection of clones. The first and most naive algorithm is to always exchange the two clones,

TABLE 3

List of 189 Hybrid Clones

No.[a]	Ref.	Clone name	No.[a]	Ref.	Clone name	No.[a]	Ref.	Clone name	No.[a]	Ref.	Clone name
1	37	84-2	49	23	WIL-6	97	15	CP72532X-6	145	21	CTP41E3
2	37	84-3	50	23	WIL-7	98	15	TLUC2LE5	146	1	L53C
3	37	84-7	51	23	WIL-8X	99	15	TLUC2LE5C9	147	1	L53F
4	37	84-13	52	23	WIL-8Y	100	39	MOG-2A2	148	1	L53G
5	37	84-26	53	23	WIL-14	101	39	HORL411B6P	149	38	HORP9.5
6	37	84-27	54	23	WIL-15	102	39	HORL411B6N	150	38	CTP41E3
7	37	84-35	55	23	XTR-3CSAZK	103	39	SIF-4D-B1	151	38	CTP34B4
8	37	84-21	56	25	I-MA9-4.11	104	39	SIF-4D-D1	152	38	SIF4A31
9	37	84-39	57	25	III-MA9-3	105	45	3W4-CL-5	153	38	DUR4.3
10	11	CP3-1	58	25	III-MA9-4	106	45	PLTI-1S	154	5	DUR-4.3
11	11	CP4-1	59	25	III-MA9-13	107	45	TRI-C4	155	5	IA9498
12	11	CP5-1	60	25	III-MA9-14	108	45	MOG-13.22	156	5	HORP-9.5
13	11	CP6-1	61	25	VI-MA9-3	109	45	HORL-9D2	157	5	CLONE-21
14	11	CP12-1	62	25	VI-MA9-14B	110	45	HORL-9-15	158	5	SIF4A24E3
15	11	CP14-1	63	31	41.06	111	44	C4A12	159	5	FG10
16	11	CP15-1	64	31	45.01	112	44	GAL-13	160	5	SIF15P5
17	11	CP16-1	65	31	45.43	113	44	TWIN-12.2	161	5	FIR5
18	11	CP17-1	66	31	76.14	114	44	C10B	162	5	MOG2C2
19	11	CP18-1	67	31	76.33	115	44	SHAG-1-2	163	34	DUA-3-BSAG
20	11	CP20-1	68	31	79.05B	116	44	SHAG-1-3	164	35	TSL-1
21	11	CP26-1	69	31	80.05D	117	44	SHAG-2-1	165	35	XTR-2
22	11	CP28-1	70	31	80.14C	118	44	SHAG-5-1	166	35	ALR-2

23	11	CP29-1	71	32	80.17A	119	44	SHAF-3-1	167	35	WIL-6
24	11	640-77A	72	31	82.82A	120	44	SHAH-2-2	168	35	RAS-8
25	11	640-63	73	31	85.16A	121	44	SHAH-3-2	169	35	RAS-9DT
26	11	879-16	74	31	89.27	122	44	SHAH-4-1	170	35	DUM-13
27	11	640-34	75	31	100.02B	123	44	SHAH-5-1	171	35	DUA-1
28	11	822-19-C15	76	31	102.05B	124	44	SHAH-5-1	172	35	DUA-5BSAGA
29	11	68-201-A	77	31	103.04	125	44	SHAM-2-9	173	35	DUM-6
30	11	706-D-1	78	31	111.02A	126	7	MOG2E5	174	35	XTR-22
31	11	826-26A	79	31	112.10A	127	7	MOG2C2	175	35	TSL-2
32	11	762-8A	80	31	120.33	128	8	MOG7	176	35	REX-12
33	23	DUA-3BSAGA	81	31	120.35	129	26	MOG-2C2	177	35	ATR-13
34	23	DUA-5BSAGA	82	31	133.05	130	16	C56L	178	35	TSL-4
35	23	ICL-15	83	31	134.02A	131	16	C35H	179	35	MAR-2
36	23	JSR-14	84	24	REW-7	132	4	DT-2-12	180	41	XV-D
37	23	REW-5	85	15	CP3-2	133	4	DT-2-1R3	181	41	XV-K
38	23	REW-8D	86	15	CP11-2	134	30	HORL4111B6	182	41	XIII-N
39	23	REW-11	87	15	CP14-2	135	30	MOG34A4	183	41	XXI-O
40	23	REX11BSAGB	88	15	CP15-2	136	30	HORL1	184	41	XXI-R
41	23	SIR-8	89	15	CP22	137	30	FIR5R3	185	41	XXI-R1
42	23	TSL-1	90	15	CP26-2	138	21	LSR34 + S49	186	41	XXI-R2
43	23	VTL-6	91	15	CP27	139	21	LSR8	187	41	XXI-R3
44	23	VTL-8	92	15	CP28	140	21	PLTI.S	188	41	XXI-R4
45	23	VTL-17	93	15	CP38-1	141	21	LSR34-S49	189	41	XXI-D1
46	23	WIL-1	94	15	CP23	142	21	3W4CL5			
47	23	WIL-2	95	15	CP23-4-1	143	21	DUR4R3			
48	23	WIL-5	96	15	CP43	144	21	DT1.2			

[a] Clone number used in Table 4.

producing a new panel with exactly one new member. As the exchange progress, a record is kept of the best panel encountered. This simple algorithm basically amounts to random sampling from the collection of all possible panels.

A second and more directed algorithm is to make an exchange only if the value of E_2 for the new panel is at least as low as the value of E_1 for the current panel. We will call this the random downhill algorithm. It wastes no time taking poor steps, but it potentially can get trapped at a local minimum.

Our third algorithm, the method of simulated annealing, represents a compromise between the first two algorithms [28, 19]. The early stages of simulated annealing resemble random sampling; later stages resemble the random downhill algorithm. Simulated annealing is motivated by the observation that a liquid cooled very slowly from a high temperature to a low temperature will crystallize in a state of minimum energy. To implement simulated annealing a parameter T analogous to temperature is gradually reduced to 0. The objective function E to be minimized is termed energy. In the present context simulated annealing can be realized as follows: Suppose a current panel with energy E_1 exists. Generate a new random panel with energy E_2 by the exchange step. Move to this new panel if $E_2 \leqslant E_1$. If $E_2 > E_1$, then move to the new panel with the Boltzmann probability $\exp[-(E_2 - E_1)/T]$. As T tends to 0, this probability tends to 0 also. Thus, fewer and fewer unfavorable steps are taken as T approaches 0. As with the other two algorithms, a record is kept of the best panel encountered.

To illustrate the above three algorithms for panel design we have amassed 189 different hybrid clones from several published articles [37, 11, 23, 25, 31, 24, 15, 39, 45, 44, 7, 8, 26, 16, 4, 30, 21, 1, 38, 5, 34, 35, 41]. Table 3 lists an appropriate identification code for each clone. We have omitted duplicate clones found in more than one reference and clones having ambiguous human chromosome complements. Table 4 presents the best panels resulting from the application of the three algorithms. The total number of iterations for each panel size ranged from 55,000 to 75,000 and was determined by the convergence criterion for simulated annealing. See [28] for a detailed description of how simulated annealing is implemented. Listed in Table 4 are the panel composition, the minimum Hamming distance, and the number of pairs of columns attaining this distance.

Table 4 makes it clear that random sampling is not a contender for panel design. The best panels are relatively rare and can only be identified by some type of directed search. Both the random downhill algorithm and simulated annealing produce excellent panels. Simulated annealing performs better, particularly for panel sizes $n = 12$, 17, and 19. In these three cases the best simulated annealing panels have a higher minimum Hamming distance than the best random downhill panels. As expected, random downhill typically

TABLE 4

Best Panels Achieved by the Three Algorithms

Panel Size	Random Sampling Energy[a]	Random Downhill Energy[a]	Simulated Annealing Energy[a]	Simulated Annealing Panel Composition[b]
5	(0,2)	(1,43)	(1,38)	{6,78,79,92,151}
6	(1,23)	(1,13)	(1,12)	{29,56,73,78,90,167}
7	(1,8)	(1,3)	(1,1)	{19,46,52,56,67,80,151}
8	(1,3)	(2,40)	(2,27)	{4,43,50,67,90,113,177,186}
9	(2,30)	(2,5)	(2,6)	{3,10,14,43,50,58,80,105,131}
10	(2,7)	(3,37)	(3,24)	{12,28,29,73,79,80,90,92,140,179}
11	(2,3)	(3,7)	(3,3)	{3,4,29,38,43,53,73,74,79,90,181}
12	(2,2)	(3,1)	(4,43)	{14,36,58,63,71,78,79,105,131,144,179,189}
13	(3,13)	(4,27)	(4,12)	{6,12,23,27,28,36,42,51,79,81,90,105,141}
14	(3,3)	(4,3)	(4,1)	{7,31,36,42,59,74,79,81,90,92,104,106,123,149}
15	(3,1)	(5,30)	(5,9)	{3,6,29,36,39,49,53,59,61,73,78,79,90,146,154}
16	(4,5)	(5,7)	(5,2)	{21,23,27,28,49,56,73,78,79, 106,124,130,144,151,179,183}
17	(4,6)	(5,3)	(6,38)	{7,21,38,42,52,56,59,63,73, 78,79,80,88,104,111,148,152}
18	(4,2)	(6,4)	(6,13)	{14,19,31,38,49,50,53,59,78, 80,90,91,104,106,116,154,160,165}
19	(5,10)	(6,2)	(7,32)	{7,10,21,28,56,58,63,74,78,79, 104,105,106,131,144,167,174,179,181}
20	(5,5)	(7,7)	(7,17)	{4,21,22,28,29,31,36,42,59,61, 78,101,113,119,128,131,140,146,159,179}

[a] The first number listed is the minimum Hamming distance for the column pairs, and the second number is the number of pairs attaining this distance.
[b] The numbers in braces correspond to the clone numbers in Table 3.

achieved its best panels in relatively few iterations, whereas simulated annealing often attained its best panels in the final stages of simulation.

It is interesting to contrast Tables 2 and 4. For instance, under simulated annealing $d = 6$ is first reached for the panel size $n = 17$. The minimum n possible for this d is 13. The average n when panels are randomly generated is 25.3, with a standard deviation of 2.6. In other words, by pooling clones one is able to reach the level $d = 6$ much sooner than by assembling a sequence of panels from clones which are randomly generated one after another. Note that when $d = 6$, up to five assay errors can be detected and up to two can be corrected.

4. DISCUSSION

We have attempted to formalize some notions of redundancy, efficiency, error detection, and error correction for the somatic cell hybrid method. For all practical purposes, it is clear that as the number of clones in a panel increases, the chance of correctly mapping a given gene also increases. Yet it is hardly economical to use large randomly constructed panels when small purposely designed ones will suffice. Even in the context of purely randomly generated panels, Figure 2 demonstrates a phenomenon of diminishing returns in adding further clones to an existing panel of hybrids. It is our contention that current laboratory practice encourages the use of random panels with two many hybrid clones.

Little thought has been devoted to the engineering rather than intuitive construction of panels. (See [17] for a partial exception to this observation.) By applying some simple concepts from algebraic coding theory, rational construction of panels is feasible. We have focused on the minimum Hamming distance for a panel as a measure of its discriminatory power. Selection of nearly optimal panels by this criterion from existing collections of clones is practical using random sampling techniques and results in panels which are uniformly good for all chromosomes. The alternative of choosing good panels by visual inspection is not practical. Once again the method of simulated annealing has proved its versatility. Two other applications of simulated annealing in genetics appear in [13] and [40]. We conjecture that the combinatorial optimization problem of panel design is intrinsically hard in the precise technical sense of being NP-complete [12]. For problems of this category, like the traveling salesman problem, simulated annealing offers a practical, easy to implement approximate solution [3].

Partially on the basis of this study, we recommend more collaboration and coordination among gene mapping groups so that good panels can be assembled, stored, and distributed. Besides being more efficient, small panels with the same information content as large panels can actually reduce the number of assay errors. More systematic design and distribution of panels will also enhance the proper cytogenetic characterization of clones within the best panels. In fact, the error detection and correction capabilities of a good panel permit careful monitoring of it for corrupted clones and clones experiencing continued chromosome loss. The panels in Table 4 are not meant to be definitive. Some of the clones represented may no longer be available or have stable chromosome complements. However, our techniques for achieving maximal panel redundancy with minimal panel size offer the opportunity to design and disseminate good panels regardless of exactly what clones are currently available.

Research supported in part by University of California, Los Angeles; USPHS Grant HL 07505; and NIH Grant AM 33329.

REFERENCES

1 M. Azoulay, I. Henry, F. Tata, D. Weil, K. H. Grzeschik, M. E. Chaves, N. McIntyre, R. Williamson, S. E. Humphries, and C. Junien, The structural gene for

lecithin:cholesterol acyl transferase (LCAT) maps to 16q22, *Ann. Hum. Genet.* 51:129–136 (1987).

2 M. R. Best, A. E. Brouwer, F. J. MacWilliams, A. M. Odlyzko, and N. J. A. Sloane, Bounds for binary codes of length less than 25, *IEEE Trans. Inform. Theory* IT-24:81–93 (1978).

3 E. Bonomi and J.-L. Lutton, The N-city travelling salesman problem: Statistical mechanics and the Metropolis algorithm, *SIAM Rev.* 26:551–568 (1984).

4 J. Burg, E. Conzelmann, K. Sandhoff, E. Solomon, and D. M. Swallow, Mapping of the gene coding for the human GM2 activator protein to chromosome 5, *Ann. Hum. Genet.* 49:41–45 (1985).

5 M. S. Davies, L. F. West, M. B. Davis, S. Povey, and R. K. Craig, The gene for human alpha-lactalbumin is assigned to chromosome 12q13, *Ann. Hum. Genet.* 51:183–188 (1987).

6 P. D'Eustachio and F. H. Ruddle, Somatic cell genetics and gene families, *Science* 220:919–924 (1983).

7 Y. H. Edwards, J. H. Barlow, C. P. Konialis, S. Povey, and P. H. W. Butterworth, Assignment of the gene determining human carbonic anhydrase, CAI, to chromosome 8, *Ann. Hum. Genet.* 50:123–129 (1986).

8 Y. H. Edwards, J. Lloyd, M. Parkar, and S. Povey, The gene for human muscle specific carbonic anhydrase (CA III) is assigned to chromosome 8, *Ann. Hum. Genet.* 50:41–47 (1986).

9 H. J. Evans, R. A. Buckland, and M. L. Pardue, Location of the genes coding for 18S and 28S ribosomal RNA in the human genome, *Chromosoma* 48:405–426 (1974).

10 W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, Wiley, New York, 1968, Chapters VII, XI.

11 J. H. Fisher, F. T. Kao, C. Jones, R. T. White, B. J. Benson, and R. J. Mason, The coding sequence for the 32,000-dalton pulmonary surfactant-associated protein A is located on chromosome 10 and identifies two separate restriction-fragment-length polymorphisms, *Amer. J. Hum. Genet.* 40:503–511 (1987).

12 M. R. Garey and D. S. Johnson, *Computers and intractability: A Guide to the theory of NP-completeness*, Freeman, San Francisco, 1979.

13 L. Goldstein and M. S. Waterman, Mapping DNA by stochastic relaxation, *Adv. in Appl. Math.* 8:194–207 (1987).

14 R. W. Hamming, Error-correcting codes, in *Coding and Information Theory*, Prentice-Hall, Englewood Cliffs, N.J., 1986, pp. 44–47.

15 G. S. Harrison, H. A. Drabkin, F. T. Kao, J. Hartz, I. M. Hart, E. H. Y. Chu, B. J. Wu, and R. I. Morimoto, Chromosomal location of human genes encoding major heat-shock protein HSP70, *Som. Cell Mol. Genet.* 13:119–130 (1987).

16 I. Henry, P. Gallano, C. Besmond, D. Weil, M. G. Mattei, C. Turleau, J. Boue, A. Kahn, and C. Junien, The structural gene for aldolase B (ALDB) maps to 9q13 → 32, *Ann. Hum. Genet.* 49:173–180 (1985).

17 M. E. Kamarck, P. E. Barker, R. L. Miller, and F. H. Ruddle, Somatic cell hybrid mapping panels, *Exp. Cell Res.* 152:1–14 (1984).

18 F. T. Kao, Somatic cell genetics and gene mapping, *Internat. Rev. Cytol.* 85:109–146 (1983).

19 S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, Optimization by simulated annealing, *Science* 220:671–680 (1983).

20 F. J. MacWilliams and N. J. A. Sloane, Linear codes, in *The Theory of Error-correcting Codes*, Elsevier, New York, 1977, pp. 7–13.

21 D. Martin, D. F. Tucker, P. Gorman, D. Sheer, N. K. Spurr, and J. Trowsdale, The

human placental alkaline phosphatase gene and related sequences map to chromosome 2 band q37, *Ann. Hum. Genet.* 51:145–152 (1987).

22 V. A. McKusick, The morbid anatomy of the human genome: A review of gene mapping in clinical medicine, *Medicine* 65:1–33 (1986).

23 J. C. Murray, K. H. Buetow, M. Donovan, S. Hornung, A. G. Motulsky, C. Disteche, K. Dyer, K. Swisshelm, J. Anderson, E. Giblett, E. Sadler, R. Eddy, and T. B. Shows, Linkage disequilibrium of plasminogen polymorphisms and assignment of the gene to human chromosome 6q26–6q27, *Amer. J. Hum. Genet.* 40:338–350 (1987).

24 S. L. Naylor, A. Y. Sakaguchi, E. Spindel, and W. W. Chin, Human gastrin-releasing peptide gene is located on chromosome 18, *Som. Cell. Mol. Genet.* 13:87–91 (1987).

25 J. M. Nigro, C. W. Schweinfest, A. Rajkovic, J. Pavlovic, S. Jamal, R. P. Dottin, J. T. Hart, M. E. Kamarck, P. M. M. Rae, M. D. Carty, and P. Martin-DeLeon, cDNA cloning and mapping of the human creatine kinase *M* gene to 19q13, *Amer. J. Hum. Genet.* 40:115–125 (1987).

26 I. R. Phillips, E. A. Shephard, S. Povey, M. B. Davis, G. Kelsey, M. Monteiro, L. F. West, and J. Cowell, A cyctochrome P-450 gene family mapped to human chromosome 19, *Ann. Hum. Genet.* 49:267–274 (1985).

27 M. Plotkin, Binary codes with specified minimum distance, *IRE Trans. Inform. Theory* IT-6:445–460 (1960).

28 W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, 1986. Minimization or maximization of functions, in *Numerical Recipes*, Cambridge U.P., Cambridge, 1986, pp. 274–334.

29 C. R. Rao, Multivariate analysis, in *Linear Statistical Inference and Its Applications*, Wiley, New York, 1973, pp. 516–525.

30 S. H. Rider, P. A. Gorman, J. M. Shipley, G. Moore, B. Vennstrom, E. Solomon, and D. Sheer, Localization of the oncogene c-erbA2 to human chromosome 3, *Ann. Hum. Genet.* 51:153–160 (1987).

31 N. J. Royle, D. M. Irwin, M. L. Koschinsky, R. T. A. MacGillivray, and J. L. Hamerton, Human genes encoding prothrombin and ceruloplasmin map to 11p11–q12 and 3q21–24, respectively, *Som. Cell Mol. Genet.* 13:285–292 (1987).

32 F. H. Ruddle and R. S. Kucherlapati, Hybrid cells and human genetics, *Sci. Amer.* 231:36–44 (July 1974).

33 A. R. Rushton, Quantitative analysis of human chromosome segregation in man-mouse somatic cell hybrids, *Cytogenet. Cell Genet.* 17:243–253 (1976).

34 A. Y. Sakaguchi and T. B. Shows, Coronavirus 229E susceptibility in man-mouse hybrids is located on human chromosome 15, *Som. Cell. Genet.* 8:83–94 (1982).

35 D. L. Silberstein and T. B. Shows, Gene for glutathione S-transferase-1 (GST1) is on human chromosome 11, *Som. Cell. Genet.* 8:667–675 (1982).

36 N. J. A. Sloane and D. S. Whitehead, New family of single-error correcting codes, *IEEE Trans. Inform. Theory* IT-16:717–719 (1970).

37 R. S. Sparkes, G. J. Dizikes, I. Klisak, W. W. Grody, T. Mohandas, C. Heinzmann, S. Zollman, A. J. Lusis, and S. D. Cederbaum, The gene for human liver arginase (ARG1) is assigned to chromosome band 6q23, *Amer. J. Hum. Genet.* 39:186–193 (1986).

38 N. K. Spurr, P. N. Goodfellow, and A. J. P. Docherty, Chromosomal assignment of the gene encoding the human tissue inhibitor of metalloproteinases to xp11.1–p11.4, *Ann. Hum. Genet.* 51:189–194 (1987).

39 D. M. Swallow, S. Povey, M. Parkar, P. W. Andrews, H. Harris, B. Pym, and P. Goodfellow, Mapping of the gene coding for the human liver/bone/kidney isozyme of alkaline phosphatase to chromosome 1, *Ann. Hum. Genet.* 50:229–235 (1986).

40 A. Thomas, Optimal computation of probability functions for pedigree analysis, *IMA J. Math. Appl. Med. Biol.* 3:167–178 (1986).

41 S. Vora, S. Durham, B. de Martinville, D. L. George, and U. Francke, Assignment of the human gene for muscle-type phosphofructokinase (PFKM) to chromosome 1 (region cen → q32) using somatic cell hybrids and monoclonal anti-M antibody, *Som. Cell Genet.* 8:95–104 (1982).

42 T. J. Wagner, A search technique for quasi-perfect codes, *Inform. and Control* 9:94–99 (1966).

43 M. Weiss and H. Green, Human-mouse hybrid cell lines containing partial complements of human chromosomes and functioning human genes, *Proc. Nat. Acad. Sci. U.S.A.* 58:1104–1111 (1967).

44 D. E. Wilson, R. Del Pizzo, B. Carritt, and S. Povey, Assignment of the human gene for beta-glycerol phosphatase to chromosome 8, *Ann. Hum. Genet.* 50:217–221 (1986).

45 D. E. Wilson, D. M. Swallow, and S. Povey, Assignment of the human gene for uridine 5′-monophosphate phosphohydrolase (UMPH2) to the long arm of chromosome 17, *Ann. Hum. Genet.* 50:223–227 (1986).