Chapter 8

Universal Gene Expression Analysis with Combinatorial Arrays

8.1 Introduction

The ability of DNA microarrays to simultaneously measure thousands of binding interactions has led to their rapid adoption in many applications: gene expression profiling [252, 183], DNA sequencing [65], genomic fingerprinting [157], and studies of DNA binding proteins [28], to name a few. Gene expression profiling, in particular, has exploded into an enormous field encompassing a wide variety of applications. For example, in the field of functional genomics, comparison of expression levels across many different experimental conditions [127, 24, 61], or between wildtype and knockout or overexpressed cells, helps to determine gene function and regulatory network structure. Differences in induced expression changes in closely related types of cancer have been used as a means for reliable diagnosis [4]. In the pharmaceutical industry, expression studies help to correlate drug response (positive or negative effects) with genetic profiles to predict the effects of the drug in new patients. Gene expression profiling is also used in the field of developmental biology to untangle the mysteries of development and aging, and in a variety of other fields to determine the biological

Text of this chapter copyright Cold Spring Harbor Laboratory Press. Originally published in [275]. Used with permission.

response to drugs, infections, and environmental toxins. A typical experimental setup is illustrated in Figure 8.1.

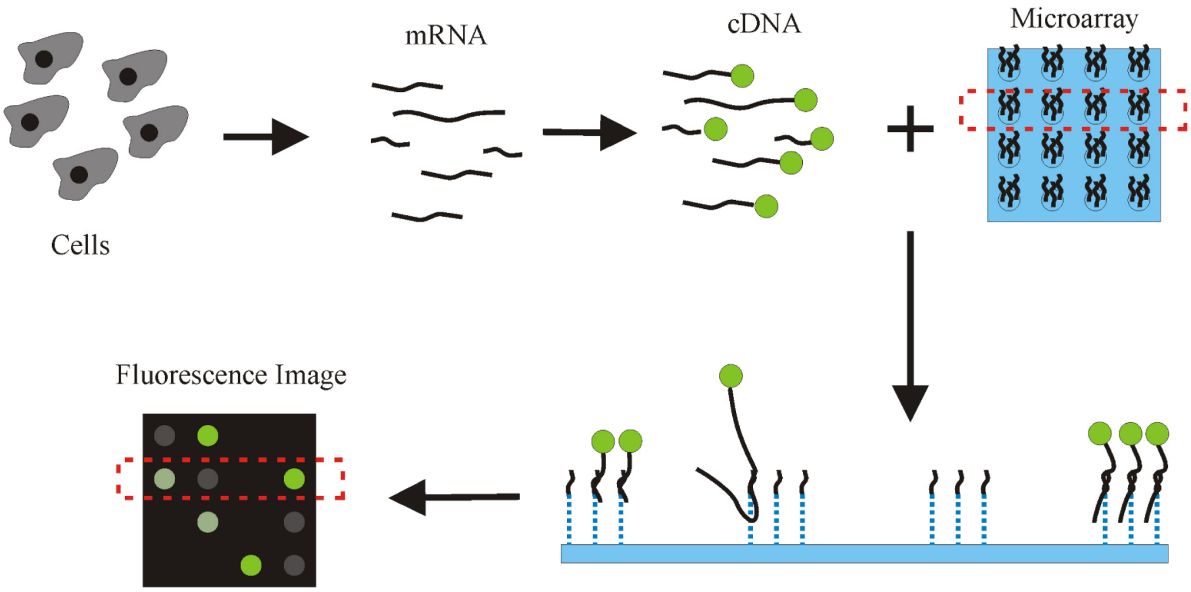

Figure 8.1: **Typical experimental setup for gene expression profiling**. Cells are harvested from the sample of interest and the messenger RNA (mRNA) is extracted and labeled. One common labeling technique involves reverse transcription of the mRNA into complementary DNA (cDNA) in the presence of fluorescently labeled nucleotides. The labeled sample is then hybridized to a microarray consisting of many spots, each with single-stranded DNA of a particular (known) sequence tethered to the array substrate. cDNA from the sample binds to complementary sequences on the microarray and can be detected quantitatively by fluorescence imaging. The brightness of each spot reflects the amount of the corresponding mRNA present in the original sample and is thus an estimate of the level of expression of the corresponding gene. Often experiments are performed in a differential fashion to cancel out many sources of errors. In this case, two labeled samples (with different fluorophores) are hybridized simultaneously to the microarray. One sample acts as a reference and is compared to a sample prepared under a different condition. For example, one can compare cells that have and have not been exposed to a drug, or one can compare cells from normal tissue with those from cancerous tissue. The pattern of binding of each sample to the microarray is observed in a different fluorescence "channel" and an analysis is performed to determine for each gene a ratio of the expression in the experimental sample to the reference sample. The ratios are assumed to represent expression changes induced by the differences in the experimental conditions between the two samples. Note that later in this chapter, the term "gene" is used loosely to refer to the cDNA being hybridized to the array, and the term "oligo" or "n-mer" is used to refer to the probe DNA tethered to a single spot on the array.

Diverse methods for fabricating expression arrays have been developed in the past several years, some based on the deposition of cDNA libraries and/or oligonucleotides by robots [238] or ink-jet printers [204], and others based on *in situ* DNA synthesis employing photolithographic [37], micromirror [203], or ink-jet technology [113].

A drawback of all of these methods is that one must carefully choose—in advance—which sequences to probe. As a result, revisions to the arrays to correct mistakes or incorporate new genomic information are costly, requiring arrays to be redesigned and manufactured. It is desirable to have a universal gene expression chip that is applicable to all organisms, ranging from bacteria to human, including those that lack complete cDNA libraries or whose genomes are not yet fully sequenced.

One way to realize universality is to synthesize a combinatorial n-mer array containing all 4^n possible oligos of length n, the key problem being to find a value of n that is large enough to afford sufficient specificity, yet is small enough for practical fabrication and readout. Combinatoric n-mer arrays can conveniently be fabricated in a small number of simple steps using conventional solid phase synthesis chemistry and arrays of parallel fluid channels in perpendicular orientations to mask the reagents. This microfluidic synthesis technique, described in detail in Chapter 7, has the potential to fabricate arrays with spot sizes as small as the tiniest microchannels that have been demonstrated—about 100 nm.

Until high-resolution, non-optical readout methods become practical, microarray densities will ultimately be constrained by the optical diffraction limit. With this lower bound of about 0.28 μm on pixel size, n-mer arrays are limited to 8×10^9 distinct spots per square inch, corresponding roughly to a 16-mer array on a 1-inch-square chip. While it is possible to fabricate arrays with larger surface areas we consider here arrays whose sizes are comparable to the current state-of-the-art in order to facilitate sensitivity comparisons. We therefore address the question of whether one can extract useful gene expression information from combinatorial arrays of short (i.e., $n \leq 16$) oligonucleotides.

We first develop an analytical model to predict, for a given value of n and a particular genome, the average "ambiguity" of the resulting hybridization pattern. With this model, we argue that for a certain minimum value of n, the ambiguity is sufficiently low that individual gene expression levels can be extracted from the hybridization data.

8.2 Results and discussion

8.2.1 Basic analytical model

Hybridization of a single labeled mRNA species to an n-mer array will cause numerous spots to fluoresce, yielding a characteristic "fingerprint" pattern. A diverse sample of mRNA transcripts yields an equilibrium hybridization pattern which is a linear superposition of numerous overlapping fingerprints, a pattern from which gene expression levels can be deduced by inverting a huge matrix of size 4^n—the number of distinct sequences on the array (see Section 8.3.1). This calculation is impractically large, but can be avoided by taking advantage of the vast redundancy inherent in a combinatorial array. One can ignore the ambiguous oligonucleotides that bind many different transcripts, instead concentrating on the information-rich oligonucleotides that bind few transcripts. We formalize this approach by defining the "degeneracy" of an n-mer as the number of different mRNA transcripts it can capture, which of course depends on the transcriptome being analyzed. In the best case one could find an oligonucleotide that binds each transcript uniquely; however, it is more realistic to expect to find small oligonucleotide groups, each oligo of which binds only to transcripts in a small independent group. In these cases the aforementioned matrix has vastly reduced dimension, is sparse, and is in block-diagonal form, greatly simplifying its inversion. The lower the average degeneracy, the easier is the construction of the block-diagonal matrix.

We now describe an analytic model that predicts the average degeneracy of the $N_o = 4^n$ distinct oligonucleotides on an n-mer array when analyzing a transcriptome of N_g "genes". An individual mRNA transcript of length ℓ has $b = \ell + 1 - n \approx \ell$ subsequences[1] of length n, any of which can serve as a site for binding the complementary n-mer affixed to the array. Assuming the transcript has a *random* nucleotide sequence, the probability that a *particular* n-mer "captures" the transcript is $p = b/N_o$. This is a simple Bernoulli trial. To compute the expected number of *different* transcripts to which the n-mer binds (i.e., its degeneracy, d), it is necessary to carry out N_g Bernoulli trials—one for each transcript. The result is a binomial distribution of degeneracies, which can be approximated

[1] Transcripts typically have lengths on the order of 1000 nucleotides, thus $\ell \gg n$.

by the Poisson distribution,

$$P_{\text{Binomial}}(d; N_g) = \binom{N_g}{d} p^d (1-p)^{(N_g-d)} \approx P_{\text{Poisson}}(d; \lambda) = \frac{\mathrm{e}^{-\lambda}\lambda^d}{d!}, \tag{8.1}$$

where $\lambda = N_g p$ is the average degeneracy.

Not all genes have exactly the same length. One can account for non-uniform transcript length by computing the degeneracy distribution as a weighted average of Poisson distributions:

$$P(d; \bar{d}) = \sum_{\ell=0}^{\infty} P_{\text{Poisson}}(d; \lambda(\ell)) f(\ell), \tag{8.2}$$

where $f(\ell)$ is the fraction of transcripts with length ℓ. The mean value of this new distribution is:

$$\bar{d} = N_g \bar{p} = \frac{N_g \bar{b}}{N_o} \approx \frac{N_g \bar{\ell}}{N_o}, \tag{8.3}$$

where $\bar{\ell}$ is the average transcript length.

The predictions of this model are compared with the true degeneracies calculated from yeast ORFs and mouse transcripts in Table 8.1 and Figure 8.2. It is well known that there are significant statistical biases in nucleotide and codon distributions [202]. Despite the fact that this model neglects these variations, its predictions agree surprisingly well with the genomic data. The reduced agreement for larger average degeneracy values can be attributed primarily to a "clipping" effect that occurs when the average degeneracy value is close to the maximum possible degeneracy value (i.e., the number of genes), a regime in which we are not interested.

8.2.2 Accounting for mismatches

In practice, hybridization is imperfect and stable duplexes can form between strands that are not perfect complements. As a first approximation, we suppose that the hybridization stringency can be tailored to prevent duplex formation when the number of mismatched positions exceeds some thresh-

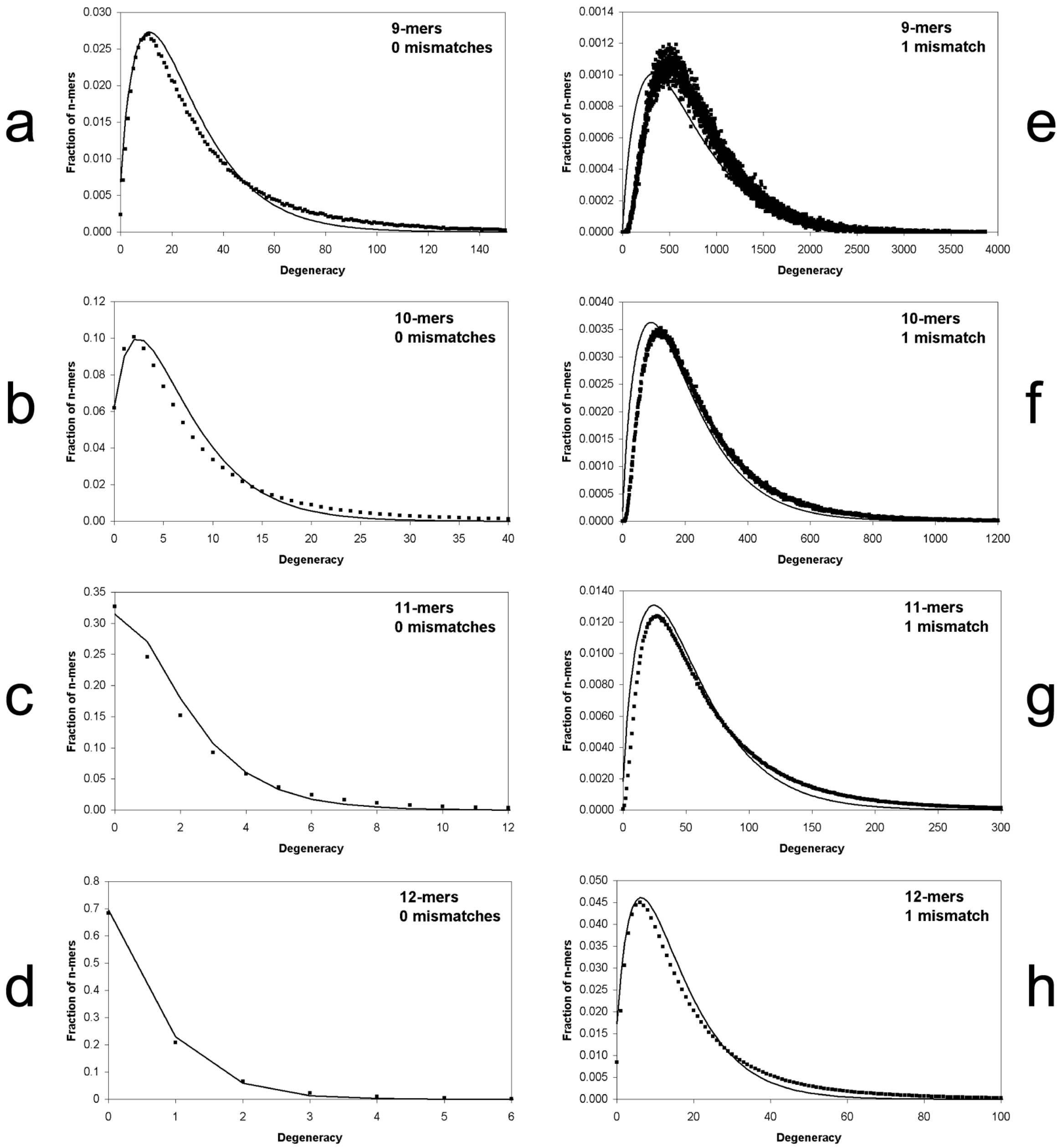

Figure 8.2: **Comparison of predicted and actual degeneracy histograms**. Degeneracy histograms determined from actual yeast genomic sequences (square markers) are compared with predictions of the analytical model (continuous line). Each histogram shows the fraction of n-mers having each degeneracy value. Predicted curves were obtained by taking a weighted average of Poisson distributions as in Equation 8.2, with the weights corresponding to the distribution of transcript lengths in yeast. There are no fitted parameters. Actual histograms were generated with custom computer software that counted the degeneracy of each n-mer in the yeast genome. (a–d) Histograms for the case of 0 mismatches, for $n = 9$, $n = 10$, $n = 11$, and $n = 12$, respectively. (e–h) Histograms for the case of 1 mismatch for the same range of n-values. Similar results were obtained for the mouse genome (not shown). (Reproduced from [275] with permission. Copyright Cold Spring Harbor Laboratory Press, 2002.)

		0 mismatches		1 mismatch	
Organism	n**-mer size**	$\bar{d}$ **(actual)**	$\bar{d}$ **(predicted)**	$\bar{d}$ **(actual)**	$\bar{d}$ **(predicted)**
yeast	7	479.3	544.2	4190	11970
yeast	8	130.2	135.9	2120	3399
yeast	9	33.42	33.96	790.0	950.9
yeast	10	8.420	8.485	245.8	263.0
yeast	11	2.110	2.120	70.29	72.07
yeast	12	0.5275	0.5295	19.39	19.59
mouse	9	130.2	134.1	3308	3754
mouse	10	32.66	33.44	976.2	1037
mouse	11	8.161	8.343	273.8	283.6
mouse	12	2.037	2.081	74.96	77.00
mouse	13	0.518	0.519	20.27	20.77
mouse	14	0.127	0.130	5.442	5.569

Table 8.1: **Comparison of average degeneracy predictions with actual data**. Average degeneracies are tabulated for both yeast and mouse, for various values of n and two different hybridization stringencies (0 mismatches and 1 mismatch). Predicted values were determined from the analytical model, Equation 8.3, while actual values were tabulated from actual yeast and mouse genomic sequence data.

old, m. Implementing this assumption requires one to establish hybridization and wash conditions that simultaneously provide adequate stringency for all spots on the array.

Comparing the melting curve for a perfectly matched duplex with that of a mismatched duplex indicates that a "window" of hybridization temperatures exists within which the perfect match is stable and the mismatch sufficiently unstable that the two can be distinguished. Fortunately, the width of this temperature window is largest for short oligonucleotides, due to single-nucleotide mismatches having an increasing destabilizing effect as oligo length is reduced. Numerous experiments have demonstrated that single-nucleotide mismatches can reliably be distinguished from perfect matches. This capability is exemplified by Wang *et al.* [284], who designed several huge microarrays (with 150,000–300,000 features) to detect single nucleotide polymorphisms (SNPs) in the human genome. They were able to resolve single-nucleotide central mismatches for all features on each chip simultaneously. The discrimination of end mismatches is somewhat more difficult due to the narrower range of suitable temperatures [64], but successful techniques have been demonstrated by several groups. Kutyavin *et al.* [156] employ minor-groove-binding molecules that stabilize properly formed double helices. Yershov *et al.* [301], Stomakhin *et al.* [255], and Maldonado-Rodriquez *et al.* [176] describe methods whereby duplexes with properly matched ends are stabilized by the phenomenon of "contiguous base stacking". It has also been reported that this level of discrimination can be achieved by hybridizing DNA to a PNA array, due to the higher mismatch sensitivity of DNA-PNA binding compared to DNA-DNA binding [117, 223, 291]. Of particular note for n-mer arrays where n is relatively short, it has also been observed that discrimination is simpler with shorter oligonucleotides due to the larger *relative* differences in binding of the perfect and non-perfect matches to the target [64].

To simultaneously achieve adequate discrimination across the whole n-mer array requires a means to reduce the intrinsic variation in melting temperatures (due to the variation in CG content from 0%–100%, among other factors). This is an active area of research, and already a number of groups have demonstrated successful techniques with small arrays. For example, Sosnowski *et al.* [245] report single-nucleotide mismatch discrimination under the same hybridization and wash conditions for two different sequences differing in intrinsic melting temperature by 20°C. More recently, chips with several thousand addressable spots have been produced based on this "electronic stringency control" method [105]. Other approaches include the addition of auxiliary molecules during hybridization [224, 128], the use of modified bases, or modification of the DNA backbone [110], to homogenize melting temperatures. Despite the fact that progress in array technology may yield nearly perfect hybridizations, for practical purposes we have relaxed this requirement in the conclusions that follow. We therefore assume that sequences can bind with up to one mismatch.

Mismatches increase the probability p that a gene binds to a particular immobilized n-mer. The increase is a simple multiplicative factor,

$$c = \sum_{k=0}^{m} \binom{n}{k} 3^k, \tag{8.4}$$

reflecting the increased number of subsequences that are *sufficiently* complementary (i.e., having $\leq m$ mismatches) for binding to the n-mer. The factor c enters the equation for average degeneracy (Equation 8.3) simply as a multiplier. An alternative viewpoint is that the number of distinct oligonucleotides on the array is reduced by this factor to $N'_o = 4^n/c$. Furthermore, because the decreased number of spots corresponds to a lower effective value for the n-mer length:

$n' = \log_4(4^n/c) = n - \log_4(c)$, one can quantify the effect of mismatches. When the analytical model is modified to include mismatches, we find excellent agreement between predictions and actual calculation (Table 8.1).

8.2.3 Truncation of transcripts

The size of the n-mer array is not the sole degree of freedom available to reduce the average degeneracy; one can also reduce $\bar{\ell}$, the average transcript length (see Equation 8.3). With appropriate nucleases and controlled reaction conditions it should be possible to truncate the length of all transcripts before hybridization according to one of two schemes: (1) reduction in transcript length by an average length $\overline{\Delta L}$ from one end, or (2) reduction of all transcripts to the same average length $\bar{L}$. For example, the duration of enzymatic digestion could be tailored to remove a desired average number of nucleotides from all transcripts (scheme 1). To implement scheme 2, one could protect the transcripts along a desired length (e.g., by polymerizing a second strand for a controlled time), subsequently digesting away the remaining unprotected portion. Since truncation would occur prior to hybridization, it can be incorporated into the analytic model simply by replacing $\bar{\ell}$ everywhere with $\bar{\ell} - \overline{\Delta L}$ or with $\bar{L}$, depending on the truncation scheme. Figures 8.3a and 8.3b demonstrate that the model continues to yield accurate predictions with truncated transcripts in addition to mismatches.

8.2.4 Estimating n

Having validated the model over a wide range of parameter values, we can estimate useful sizes for n-mer arrays. Figure 8.3c illustrates combinations of parameter values that are predicted to yield an average degeneracy of 1, (i.e., the "ideal" case), for which gene expression levels can be trivially solved. As shown for the case of 1 mismatch, to achieve this target in yeast requires a 14-mer array if transcripts are untruncated or a 12-mer array after transcript truncation to about 80 bp. In mouse, the target degeneracy is nearly realized with a 15-mer array without truncation or a 14-mer array after truncation to 90 bp.

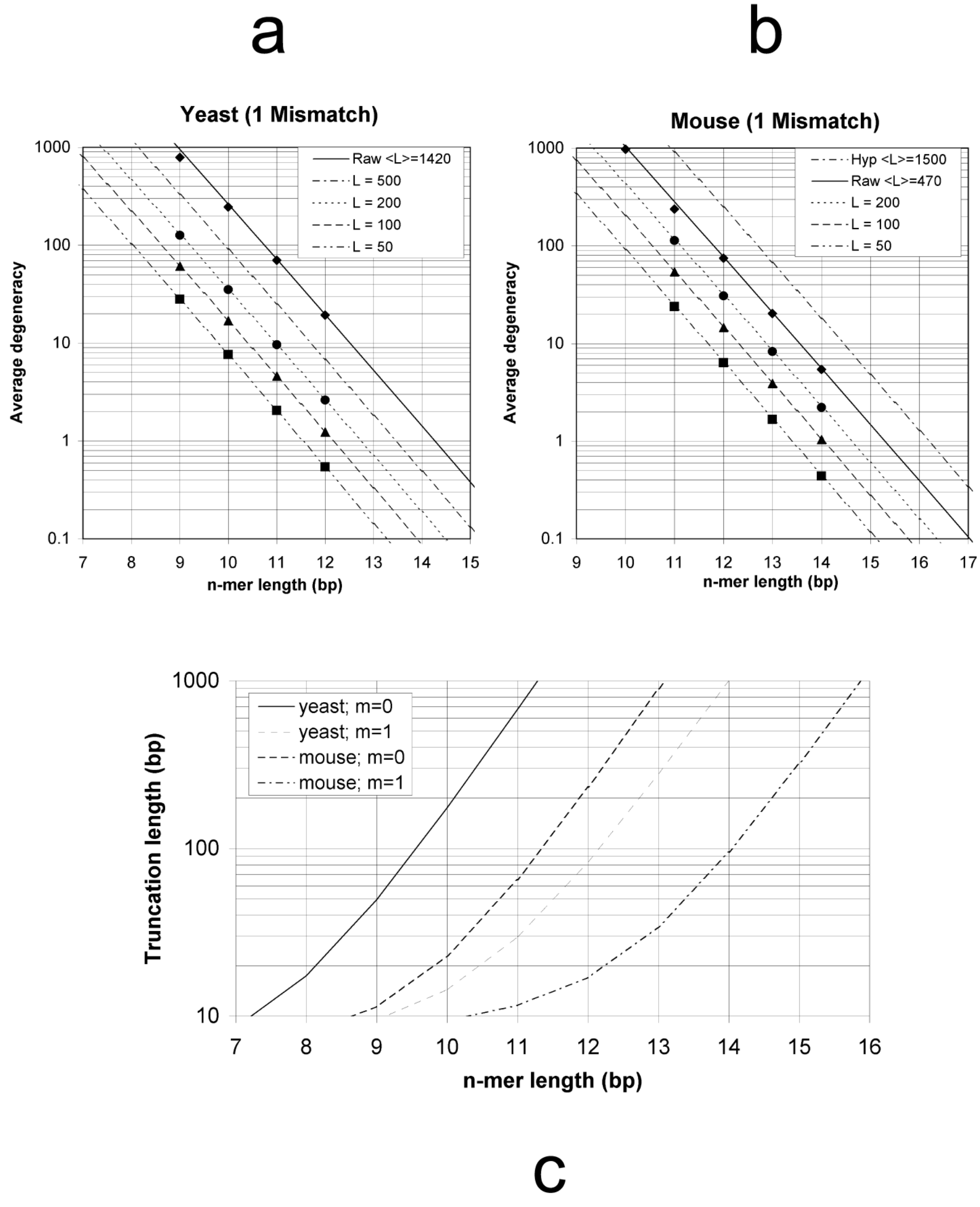

Figure 8.3: **Comparison of predicted and actual average degeneracy**. Predictions of average degeneracy are compared with calculations from actual sequence data, for the case of 1 mismatch: (a) yeast; and (b) mouse. Continuous lines represent predictions (with no fitted parameters) of average degeneracy as a function of the n-mer length, n, for varying degrees of transcript length truncation to a fixed length, L, computed from Equation 8.3 with modifications for mismatches and length truncation. ("Raw" designates the untruncated cases.) Discrete points represent the actual average degeneracy values tabulated from genomic sequence data. Due to the presence of many ESTs in the mouse UniGene database, the average transcript length for mouse is reported as much lower than yeast, so we have included a predicted curve for a hypothetical average gene length of 1500 bp. (c) Predicted relationship between parameter values to achieve an average degeneracy of 1 (the trivial case). (Adapted from [275] with permission. Copyright Cold Spring Harbor Laboratory Press, 2002.)

Our results so far have considered the average degeneracy of *all* n-mers on the array. However, when the degeneracy is sufficiently low, only a tiny subset of the oligos are needed for monitoring individual gene expression levels. A logical starting point is to consider, for each gene, the minimum degeneracy n-mer to which it can bind. Transcripts having "minimum degeneracy" equal to 1 are obvious trivial cases, as they can be monitored uniquely by a single array spot. Of the remaining transcripts, those that share their minimum degeneracy oligo with only trivial genes are also trivial by such an association. Statistically, a sufficiently large fraction of genes having a minimum degeneracy of 1 should render all genes trivial. Modifications to our purely analytic model fail to make accurate predictions for small subsets of oligonucleotides, presumably due to the underlying non-randomness of real genomes. However, beginning only with a histogram of the minimum degeneracy values for all genes in an organism (Figure 8.4), it is easy to estimate the likelihood of the above associations and predict the total fraction of genes whose expression levels can be trivially solved (see Section 8.3.4). To check these predictions, we wrote a computer program to determine exactly the fraction of trivially solvable genes based on the individual gene sequences.

A few results for the case of 1 mismatch are summarized in Table 8.2. In general, we found that nearly all genes turn out to be trivial if the fraction of genes having minimum degeneracy equal to 1 (Figures 8.5a and 8.5b) is at least about 80%. With a 10-mer array and transcript truncation to 50 bp, 98.8% of yeast transcripts are trivial. Most of the non-trivial genes are in fact unsolvable because they have identical sequences after truncation. Omitting the truncation would eliminate this problem and also simplify the experimental protocol. No truncation is needed with a 12-mer array, in which case 99.8% of transcripts are trivial. Upon close inspection, we found that most of the non-trivial genes may actually be unsolvable because they differ by only a few base pairs from one another. Similar results were obtained for mouse. With a 12-mer array and truncation to 100 bp, 97.9% of mouse transcripts are trivial; 99.6% of mouse transcripts are trivial with a 13-mer array and no truncation. Note that these required n values for both yeast and mouse are lower (by 1 or 2) than the previous predictions (Figure 8.3c), which were based on the *average* degeneracy taken

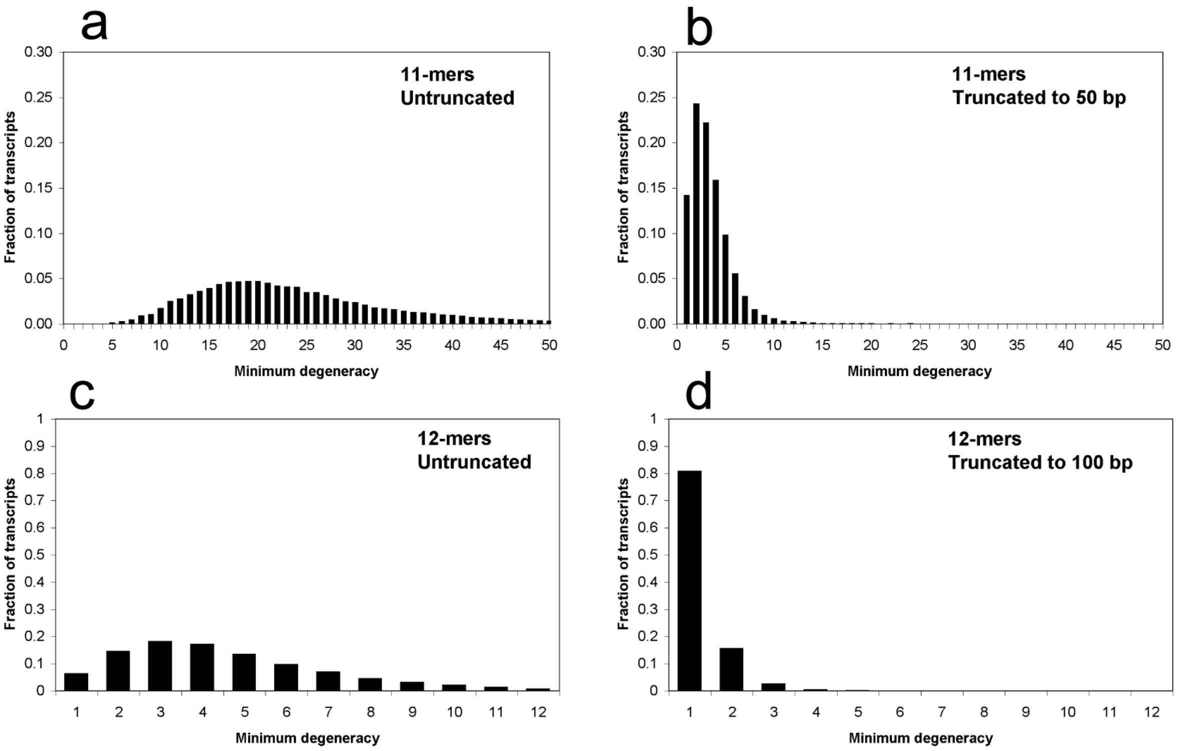

Figure 8.4: **Minimum degeneracy histograms for the mouse genome, assuming 1 mismatch**. Each histogram shows the fraction of transcripts having a given *minimum* degeneracy value. The minimum degeneracy of a transcript is determined by finding the degeneracy of all oligos to which it can bind and then selecting the lowest. Data for the histograms were generated by custom computer software that examined actual sequence data to find the n-mer with lowest degeneracy that binds to each transcript (allowing for up to 1 mismatch). As expected, increasing n and decreasing the transcript length both increase the proportion of genes having low minimum degeneracy. (a) 11-mers, no truncation; (b) 11-mers, truncation to 50 bp; (c) 12-mers, no truncation; (d) 12-mers, truncation to 100 bp. (Reproduced from [275] with permission. Copyright Cold Spring Harbor Laboratory Press, 2002.)

over *all* n-mers. It is likely that even smaller arrays can be used if one is willing to expend more computational effort and address also the non-trivial cases.

Organism	**n**	**Truncation**	**Fraction with** $d_{min} = 1$	**Fraction trivial (predicted)**	**Fraction trivial (actual)**	**Inherent redundancy**
yeast	10	50 bp	0.887	0.988	0.987	10.96
yeast	12	none	0.966	1.000	0.998	54.14
mouse	12	100 bp	0.809	0.996	0.979	6.17
mouse	13	none	0.906	1.000	0.996	20.28

Table 8.2: **Predicted and actual fraction of genes that can be trivially solved for several useful array sizes**. All data assume single mismatches. For each set of parameters, several quantities are listed. The fraction of transcripts with a minimum degeneracy of 1 ($d_{min} = 1$) was tabulated from the raw genome sequence data based on the n-mer size and truncation length. The predicted and actual fractions of transcripts that can be trivially solved were determined by the methods in Section 8.3.4. It is notable that in the cases shown here (and others not shown), nearly all genes could be trivially solved even when the fraction of genes with $d_{min} = 1$ was only 80%. Inherent redundancy (i.e., the average number of "unique oligos" per transcript) is also included for reference. In most cases where a high fraction of transcripts are trivially solvable, the intrinsic redundancy was observed to be on the order of 10.

8.2.5 Redundancy

Microarrays using oligonucleotides generally require more than one probe per gene to produce reliable results. With the decreased feature sizes and shorter probe lengths of combinatorial n-mer arrays, the importance of redundancy is likely to be even greater. Thus, while in principle only a single oligo is needed to monitor each gene, in practice one would use multiple oligos to allow averaging over independent measurements. Redundant measurements reduce the relative impact of experimental variations in binding and readout and increase the level of confidence in the measured values [162], particularly for genes expressed at low levels [136].

An approximate measure of the inherent level of redundancy in an array is the average number of "unique oligos" per gene. This quantity can be predicted by dividing the total number of unique oligos (i.e., oligos that bind to only one gene)—determined from either the Poisson model or the actual genomic data—by the number of genes. For the four array sizes discussed in the previous section, the average redundancy is on the order of 10 unique oligos per gene (see Table 8.2).

To ensure that a high fraction of genes have *at least* 10 unique oligos per gene, computing the *average* is not sufficient: the fraction must be calculated directly from the genomic sequence

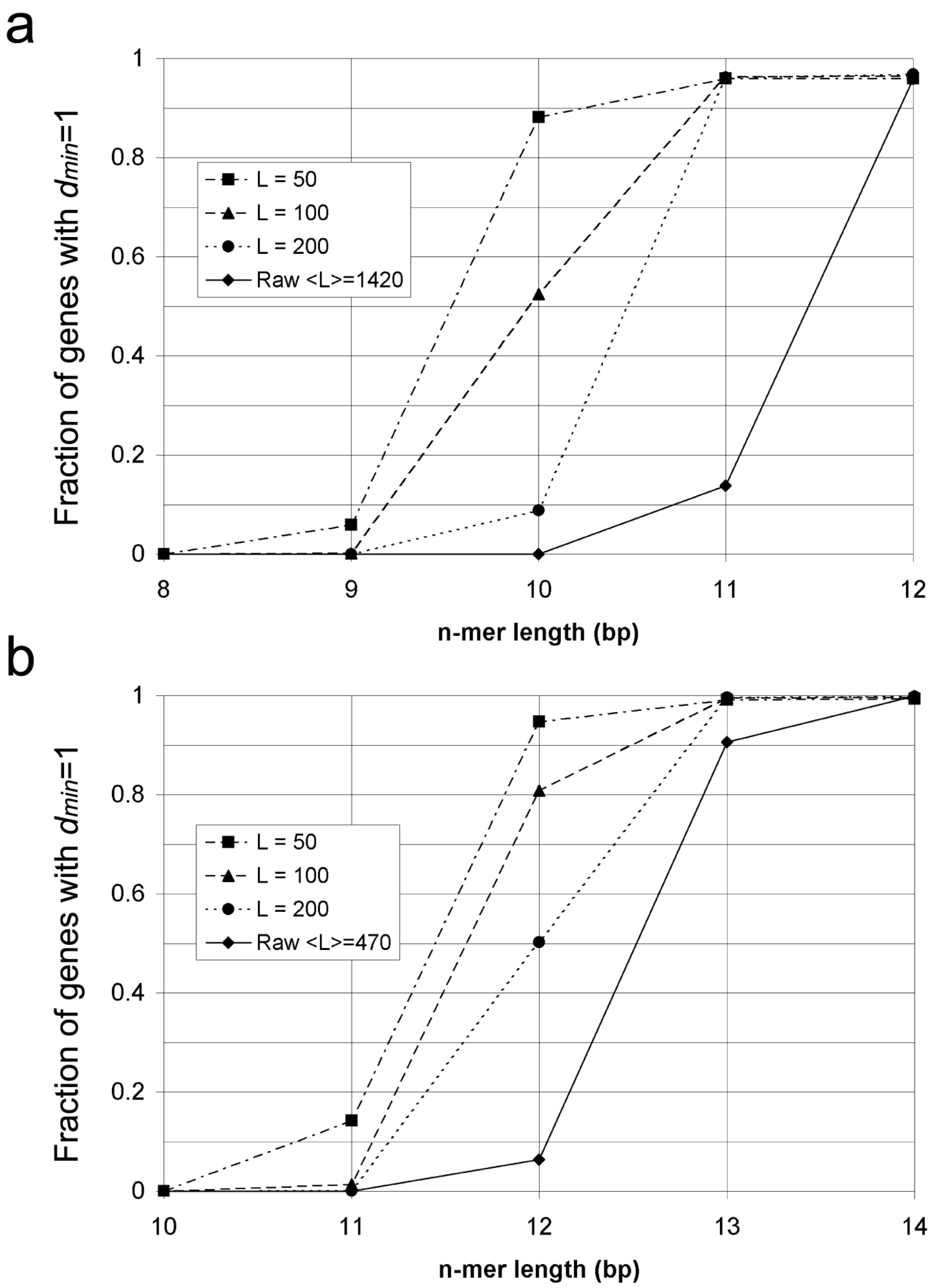

Figure 8.5: **Fraction of transcripts having minimum degeneracy equal to 1**. Plots show the fraction of transcripts having minimum degeneracy (d_{min}) equal to 1 (i.e., binding to an oligo that does not bind to any other transcripts) over a range of n-mer sizes and truncation lengths L, assuming 1 mismatch. ("Raw" designates untruncated cases.) (a) yeast; (b) mouse. It turns out that when at least a sufficient fraction ($\approx 80\%$) of transcripts have $d_{min} = 1$, nearly all gene expression levels can be trivially solved. (Reproduced from [275] with permission. Copyright Cold Spring Harbor Laboratory Press, 2002.)

data. We used customer computer software to do so. For yeast and a hybridization stringency of 1 mismatch, an 11-mer array with truncation to 100 bp ensures that 97.0% of genes bind to at least 10 unique oligos. A 13-mer array with truncation to 200 bp ensures that 99.6% of genes bind to at least 10 unique oligos in mouse with 1 mismatch. In Figure 8.6, actual redundancy is plotted against predicted redundancy for several sets of parameter values. These plots suggest that in order to have a large fraction of genes with the desired redundancy x, one should choose a set of parameters that *predicts* an average redundancy of about $10x$.

8.2.6 Conclusions

Since the mouse genome is only slightly smaller than the human genome, the results above provide an estimate of the required size for a *universal* array, namely $n \geq 12$ for truncated transcripts or $n \geq 13$ for untruncated transcripts. To ensure a redundancy of at least 10 unique oligos per gene, the required size is $n \geq 13$. Both figures are well within the limit of practical fabrication and readout ($n \leq 16$). While not universal, arrays as small as $n = 10$ would permit the study of microorganisms as complex as yeast.

In addition to universality, combinatorial n-mer arrays offer other significant advantages. For instance, since selection of n-mers with which to identify transcripts is performed in software, data can be reanalyzed (avoiding additional experiments) as genomic sequence data is updated. In addition, the selection criteria can easily be modified to incorporate additional constraints on parameters, such as spot quality and melting temperatures, to yield higher quality results. Besides gene expression analysis, combinatorial n-mer arrays have potential applications in such diverse areas as DNA sequencing by hybridization [65], the study of DNA binding proteins [28], and genomic fingerprinting [157].

As a final note, we point out that while combinatorial n-mer arrays can be *fabricated* without genomic knowledge, our analysis strategy does make use of known genomic sequence data as a prerequisite for *interpreting* the data. These data now exist in an essentially complete form for several bacteria, yeast, worm, fly, mouse, and human, among many other organisms. A comprehensive list

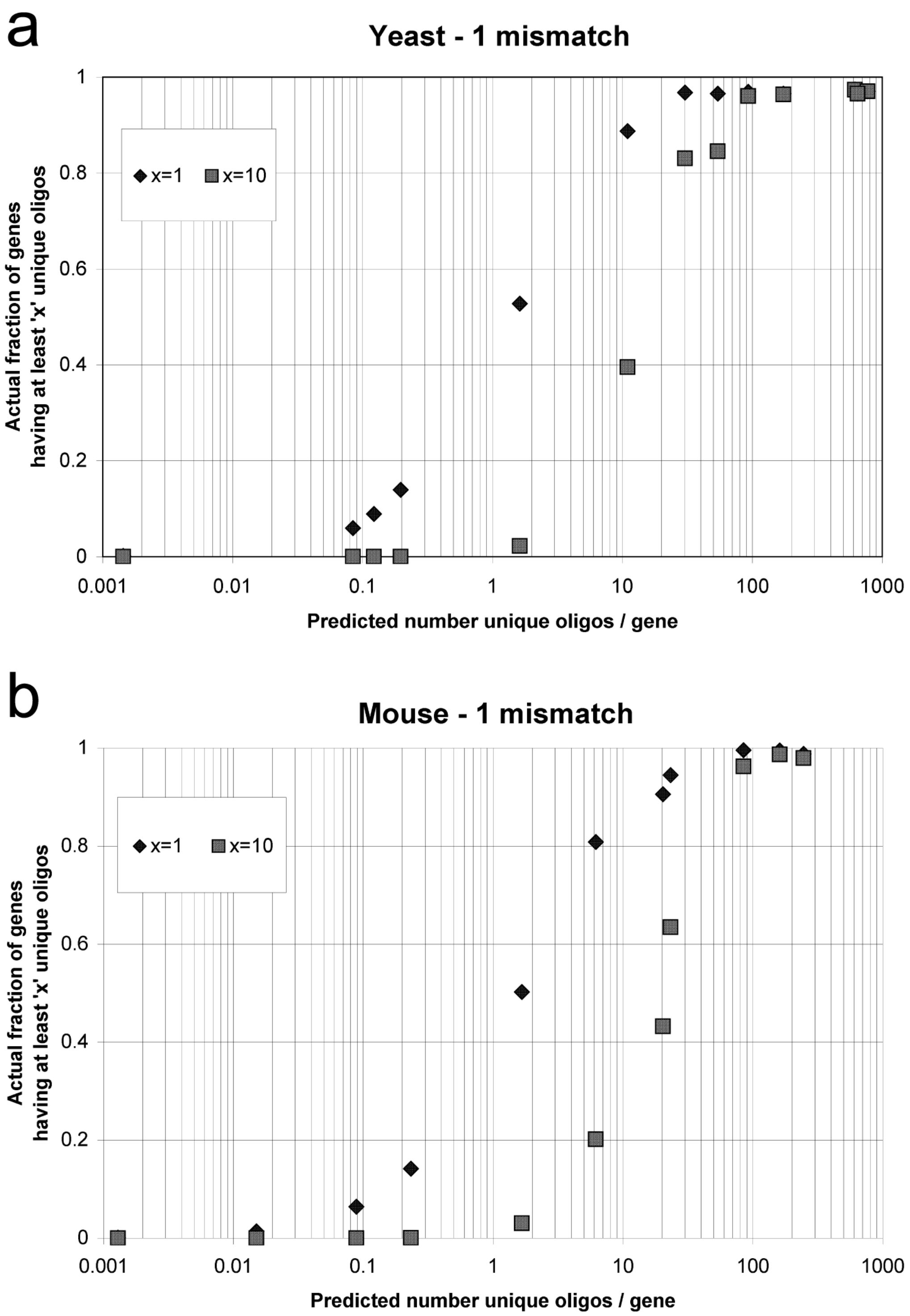

Figure 8.6: **Comparison of predicted and actual redundancy**. Redundancy is defined here as the number of unique oligos that bind to a gene. Since unique oligos bind to no other genes, their hybridization signal serves as an unambiguous measure of the expression of that gene. If there are several unique oligos for a gene, then several independent measurements are made—this is what is meant by “redundancy”. On the horizontal axis is the predicted redundancy, computed simply as the average number of unique oligos per gene (i.e., the total number of oligos with a degeneracy of 1 divided by the total number of genes). On the vertical axis is the actual fraction of genes having at least x unique oligos. Each plotted data point represents a particular n-mer size and truncation length. Some specific combinations are shown in Table 8.2. A general trend that can be observed in these plots is that in order to ensure that a large fraction of oligos have the desired redundancy x, one should choose a set of parameters that gives a *predicted* redundancy of about $10x$. (a) Data for yeast (1 mismatch); (b) Data for mouse (1 mismatch).

of completed and ongoing genome projects is available at `http://www.genomesonline.org/` [20]. For unsequenced organisms, by performing multiple hybridization experiments, we believe that it may be possible to deduce partial gene expression information without prior genomic knowledge.

8.3 Methods

8.3.1 Mathematical analysis of gene expression

A hybridization experiment can be expressed as the matrix equation $\mathbf{S} = \mathbf{H} \cdot \mathbf{E}$, where $\mathbf{S} = (S_1, S_2, ..., S_i, ..., S_{N_o})^T$ is the vector of measured signal intensities and $\mathbf{E} = (E_1, E_2, ..., E_j, ..., E_{N_g})^T$ is the vector of unknown transcript concentrations (i.e., expression levels). For a particular set of hybridization conditions, $\mathbf{H}$ is a constant matrix if the system is in chemical equilibrium and the array is not saturated. Each coefficient H_{ij} of $\mathbf{H}$ is closely related to the melting temperature (affinity) of the binding interaction between transcript j and oligo i, and can be estimated using semi-empirical formulae [27, 101] or measured by calibration experiments with known quantities of various mRNA species. Deducing transcript expression levels is reduced to the computational problem of solving the above system of equations for $\mathbf{E}$. Since it is impractical to directly invert $\mathbf{H}$, our approach is to find a projection $\mathbf{P}$, such that $\mathbf{H}' = \mathbf{P} \cdot \mathbf{H}$ is a square $N_g \times N_g$ matrix. The vast reduction in dimensionality allows one considerable freedom in choosing a projection, and choosing $\mathbf{P}$ such that $\mathbf{H}'$ is invertible and in block diagonal form permits trivial determination of expression levels: $\mathbf{E} = (\mathbf{P} \cdot \mathbf{H})^{-1} \cdot (\mathbf{P} \cdot \mathbf{S}) = \mathbf{H}'^{-1} \cdot \mathbf{S}'$. We simplify the problem by choosing a hybridization stringency m and setting all elements of $\mathbf{H}$ to zero for which the corresponding transcript and oligo require more than m mismatches to bind. We then search for a projection by beginning with the minimum degeneracy oligo for each gene and then selecting additional oligos until $\mathbf{H}'$ is invertible and the desired level of redundancy is achieved. The projection is simplest to construct when many rows have mostly zero entries—that is, when many oligos have low degeneracy.

8.3.2 Source of sequence data

Genomic sequence data for degeneracy calculations were drawn from public gene sequence databases for two organisms: yeast (*Saccharomyces cerevisiae*) and mouse (*Mus musculus*). These two organisms were selected because of their availability and because they are representative of the two ends of the eukaryotic genome size spectrum.

Yeast sequence data was obtained from the *Saccharomyces* Genome Database at Stanford University (`http://www.yeastgenome.org/`). We downloaded the complete set of coding sequences from `ftp://genome-ftp.stanford.edu/pub/yeast/yeast_ORFs/orf_coding.fasta.Z` on December 14, 1999. For this database, $N_g = 6306$ and $\bar{\ell} \approx 1420$. Since identical gene sequences cannot be distinguished by any microarray, duplicates were removed, leaving $N_g = 6276$ unique genes.

Sequences for mouse were downloaded from the UniGene system at the National Center for Biotechnology Information (`http://www.ncbi.nlm.nih.gov/UniGene/`). We downloaded the file `ftp://ftp.ncbi.nlm.nih.gov/repository/UniGene/Mm.seq.uniq.Z`, Build 74. Though this database does not contain the complete genome of mouse, it contains both genes and ESTs representing a substantial portion of the expressed genome. For this database, $N_g = 75963$ and $\bar{\ell} \approx 471$. Due to the many ESTs, the average transcript length is quite small. Thus we included some calculations with a longer hypothetical average gene length.

8.3.3 Degeneracy calculations

To calculate degeneracy values from actual sequence data, we wrote a computer program that scans through the sequences comprising a transcriptome, tallying the number of times each subsequence of length n (n-mer) is encountered in different transcripts. Accounting for length truncation to length $\bar{L}$ is accomplished by examining only the first $\bar{L}$ characters of each transcript. To deal with mismatches, each subsequence of length n within a transcript is expanded into a set of all sequences that differ by at most m nucleotides from the original subsequence. Sequences containing non-A,C,G,T characters were ignored (0% of sequence data in yeast; 1–2% in mouse). From the list of degeneracy values for each of the 4^n possible n-mers, the *average* degeneracy is easily calculated for comparison with the analytic model. In addition, the degeneracy list itself is used to generate a histogram showing the fraction of n-mers having each degeneracy value, for comparison with theoretical histograms calculated from Equation 8.2.

While counting n-mers is very simple in principle, the enormous size of the problem for large values of n (e.g., there are over a billion 15-mers) introduced several technical challenges. In early versions of our program, we observed that the memory requirements exceeded the physical RAM (random access memory) of the computer and thus caused a significant amount of swapping to disk, which slowed the program by several orders of magnitude. To avoid this problem, our counting program was coded in C++, a language that permits a high degree of control over memory usage. At the start of a "run" (for a particular organism, value of n, and number of mismatches m), the program loaded the genome sequence into memory and declared a large array with one entry to store the tally for each n-mer. As the genome was scanned, each encountered n-mer was converted to a number, determined by interpretting its DNA sequence as a base-4 number, with nucleotides representing the digits. This number served as an index into the array of tallies, allowing the proper tally to be incremented. Once all genes were scanned, the tally data was written to a file—one value per line—with implicit line numbers serving as the n-mer identities. For very large values of n, the entire array did not fit in memory so the program was run in several "passes". If, for example, two passes were required, the program would first count only n-mers occurring in the first half of the list of all possible n-mers, ignoring any n-mers from the second half that were encountered while scanning the genome. After writing the data to a file, the genome was scanned a second time, this time tallying only n-mers in the second half (ignoring those in the first). The resulting data was appended to the first file.

It should be noted that we settled on this strategy after trying several other options. Languages such as perl and PHP permit arrays to be created dynamically as the program runs; thus tallies only need to be stored for n-mers that have been encountered at least once while scanning the genome. Since most n-mers are not encountered at all (for large n), far fewer array elements need to be stored. However, this advantage is offset by the fact that accessing each element in a dynamic array is much slower than in a static array. Furthermore, it seemed that dynamic languages required about 100–1000× more memory to store the same amount of data, thus requiring a much larger number of passes through the genome. As a result, the C++ program ran considerably faster overall.

8.3.4 Predicting the fraction of solvable expression levels

The fraction of trivially solvable expression levels is estimated in a probabilistic fashion from a minimum-degeneracy histogram derived from actual sequence data (e.g., Figure 8.4). These histograms were generated by a computer program that makes use of the list of n-mer degeneracies to determine the lowest degeneracy oligo to which the transcript can bind. A minimum-degeneracy histogram indicates the fraction of genes (x_i) having each value of minimum degeneracy, i.

Genes having minimum degeneracy equal to 1 are clearly trivial because their expression level can be deduced unambiguously from the fluorescence of the minimum degeneracy oligo. A fraction x_1 of all transcripts fall into this category. Those genes having minimum degeneracy equal to 2 are trivial if the other gene that shares the degeneracy 2 oligo has minimum degeneracy equal to 1. Of all transcripts, a fraction $x_2 \cdot x_1$ are expected to fall into this category. Similarly, those genes having minimum degeneracy equal to 3 are trivial if both of the other (distinct) genes that share the degeneracy 3 oligo have minimum degeneracy equal to 1. Statistically, a fraction $x_3 \cdot x_1 \cdot (x_1 - 1/N_g)$ of genes should fall into this category. Continuing in this fashion, one obtains a summation that estimates the fraction of genes whose expression levels can be solved trivially.

A computer program was written to examine actual gene sequences in order to determine the exact total fraction that could be trivially solved. As above, all genes having minimum degeneracy equal to 1 are clearly trivial. Each of the remaining genes is handled in the following manner. First, all n-mers to which the gene binds are identified and sorted in increasing order of degeneracy. Then, for each n-mer in turn, the *other* genes that bind the n-mer are identified. If all of these other genes have minimum degeneracy equal to 1, then the original gene is trivial by its association. If this condition is not met for any of the n-mers to which the gene binds, then the gene is declared non-trivial.