

A model of molecular interactions on short oligonucleotide microarrays

Li Zhang¹, Michael F Miles² & Kenneth D Aldape³

High-density short oligonucleotide microarrays have become a widely used tool for measuring gene expression on a large scale^{1,2}. However, details of the mechanism of binding on microarrays remain unclear³. Short oligonucleotide probes currently synthesized on microarrays are often ineffective as a result of limited sequence specificity or low sensitivity. Here, we describe a model of binding interactions on microarrays that reveals how probe signals depend on probe sequences and why certain probes are ineffective. The model indicates that the amount of nonspecific binding can be estimated from a simple rule. Using this model, we have developed an improved measure of gene expression for use in data analysis.

A key issue in microarray technology using short oligonucleotide probes (such as those produced by Affymetrix, Inc.) is how to select probe sequences with high sensitivity and specificity. The current approach to this problem is to use multiple probe pairs, referred as a 'probe set'⁴, to target a single gene; one of each pair exactly matches a fragment of the gene (PM probe) and the other contains a single mismatching nucleotide in the center (MM probe). With 11–20 probe pairs in a probe set, the existence of some low-sensitivity probes is tolerable; the MM probes offer a measure of nonspecific binding to improve specificity⁴.

However, the potential for further improvement exists, as around 30% of the probe pairs consistently yield negative signals^{5,6}, indicating that the use of MM probes for assessment of nonspecific binding is unreliable. Furthermore, the observed probe signals within a probe set typically vary over 2 orders of magnitude⁷, suggesting that not all probes have optimal sensitivity.

Choosing optimal probes has been difficult because of the present poor understanding of how the sensitivity and specificity of a probe depend upon its sequence. Although duplex formation in solution has been extensively studied using a nearest-neighbor model^{8,9}, it has been difficult to apply those results to microarrays. Binding interactions on the microarrays seem to be complicated by many factors such as steric hindrance on the microarray surface¹⁰, probe-probe interaction¹¹ and RNA secondary structure formation³. Many recent studies have focused on developing statistical methods^{5,7,12–16} or experimental calibration^{17,18} to refine microarray technology. In our present study, we sought to develop a simple free energy model for the formation of RNA-DNA duplexes on short oligonucleotide microarrays. Our

model is based on the nearest-neighbor model⁸, with two modifications: (i) we assign a different weight factor at each nucleotide position on a probe to reflect the fact that different parts of the probe may contribute differently to the stability of binding; and (ii) we take into account two different modes of binding on the probes, gene-specific binding (GSB) and nonspecific binding (NSB). We call this model the positional-dependent-nearest-neighbor model (PDNN).

Here, GSB refers to the formation of DNA-RNA duplexes with exact complementary sequences, whereas NSB refers to the formation of duplexes with many mismatches between the probe and the attached RNA molecule. The number of duplexes with few mismatches should be negligible because the probes are preselected to avoid this type of binding⁴.

In addition, we make two assumptions: (i) at the completion of a hybridization experiment, a thermodynamic equilibrium state is reached between the RNA molecules bound to the microarray surface and the RNA molecules in solution; and (ii) binding of various RNA species is independent and noncompetitive.

The amount of signal expected to be observed for one probe is decomposed into GSB and NSB modes, which are determined from the level of gene expression and the binding affinities (see Methods for details). The model involves a modest number of unknown parameters: there are 16 stacking energy parameters and 24 weight factors for GSB. The same number of parameters is used for NSB. In addition, N_p , N^* , and B (see Methods for definitions) are also unknowns. However, there is little concern about over-parameterization because the number of probes on an array is on the order of 10^5 , far exceeding the number of parameters employed. In our experience, energy parameters derived from microarray data of different RNA samples are nearly identical (data not shown).

We fitted the observed microarray data with the model's expected signals to obtain the unknown energy parameters as well as the gene expression levels. The good fit (Fig. 1) shows that the model has largely captured the sequence dependence of probe binding affinity. This result is somewhat surprising given the simplicity of the model. The model may be further refined when other factors affecting hybridization³ are taken into account.

We tested the model with the same raw microarray data that were used to validate the current commercial software (MAS5.0, Affymetrix, Inc.). The model-estimated concentrations correlated well

¹Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd. 447, Houston, Texas 77030, USA. ²Departments of Pharmacology/Toxicology and Neurology and the Center for Study of Biological Complexity, Virginia Commonwealth University, Richmond, Virginia 23298-0599, USA. ³Department of Pathology & Brain Tumor Center, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd. 447, Houston, Texas 77030, USA. Correspondence should be addressed to L.Z. (lzhangli@mdanderson.org).

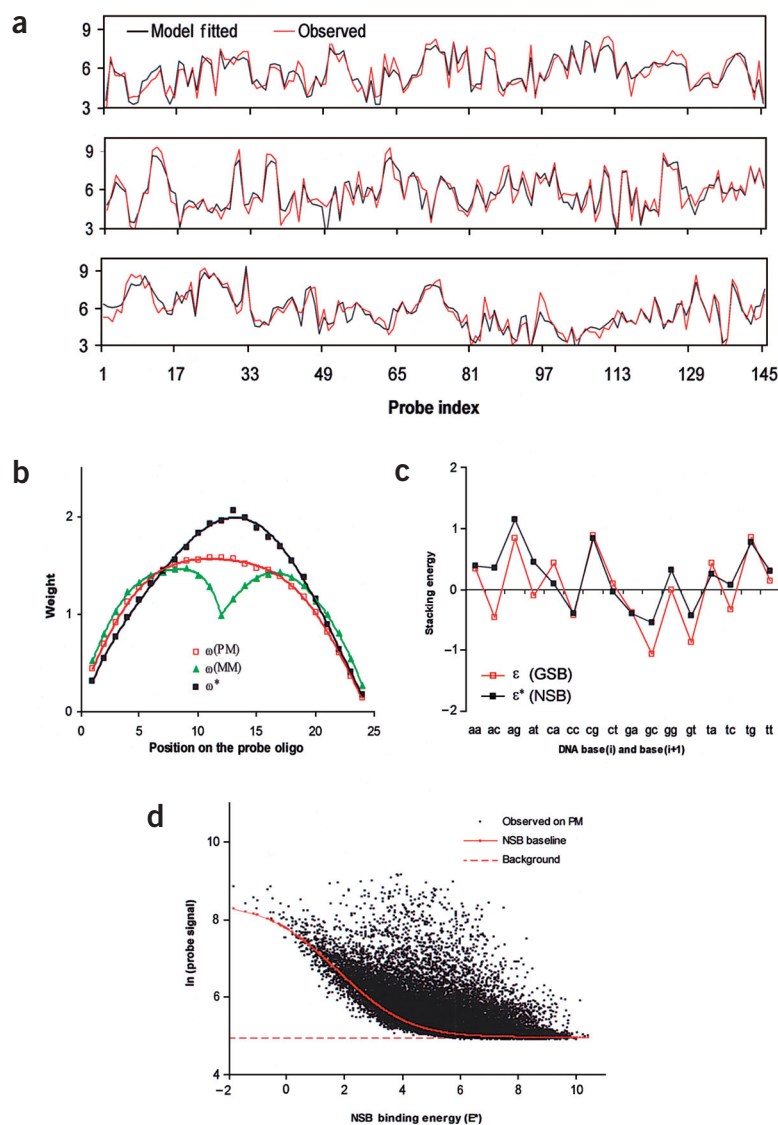


Figure 1 Model fitting and parameterization. (a) Model fitting. The y-axis represents probe signals on a log scale. Each stretch of 16 probes along the x-axis represents a probe set. The probe sets were arbitrarily chosen. Typically, the minimized fitness value, F (see Equation (4) for definition), ranges from 0.04 to 0.11; $\ln \hat{I}_{ij}$ and $\ln I_{ij}$ are well correlated ($r > 0.9$). Outliers are rare. Roughly 2% or fewer of the probes have $|\ln \hat{I}_{ij} - \ln I_{ij}| > 3\sigma$, where $\sigma^2 = F$. (b) Nearest-neighbor stacking energy. These stacking energies correlated weakly ($r = 0.6$) with those found in aqueous solution⁸ and are smaller in magnitude. (c) Weight factors. These parameters were obtained by modeling PM and MM probe signals separately. The parameters for PM and MM probes are the same except for the weight factors of GSB. (d) Background and NSB baseline. The model estimated baseline of NSB and observed PM probe signals (y-axis) are plotted on a natural logarithmic scale against the NSB energy E^* . The symbol * indicates parameters for NSB throughout this report.

(average $r = 0.988$) with known concentrations of 'spike-in' genes (Fig. 2a). We compared these results with those obtained with MAS5.0 and dChip⁷, a commonly used alternative to MAS5.0. All three methods yielded expression levels that correlated well with known concentrations (Fig. 2a–c). The contrast was in consistency: the variances were systematically lower with the PDNN model (Fig. 2d–f). We noticed that the apparent reduced variance with the PDNN model was partly due to a reduction of the range of gene expression levels, which does not represent a practical improvement by itself. On a log scale, changes in expression obtained from MAS5.0 appear to be 1.24 times

larger than those from the PDNN model (compare Fig. 2a and b). However, the effect of the reduced range of expression levels is considered to be minor because for most of the genes, the standard deviation of log-transformed PDNN expression levels across replicate samples is three times smaller than that obtained from MAS5.0 (compare Fig. 2d and e). Therefore, there seems to be a clear advantage to using PDNN model for error reduction.

The PDNN model appears to indicate that the two ends of probes contribute less to binding stability according to the weight factors (Fig. 1b)—perhaps because the ends of duplexes bound on the microarray surface are fraying. This behavior was not observed in duplexes in aqueous solution. Interestingly, there is a dip in the GSB weight factors of MM probes around the mismatch position (Fig. 1b). Presumably, the dip is caused by the mismatch, which destabilizes the duplex structure. Given that the mismatch is not prespecified in the model, that fact that it can be 'recovered' by the model is striking.

We also noted that stacking energies in the PDNN model (Fig. 1c) might explain the presence of negative probe pair signals (PM – MM < 0), which have remained a mystery because the mismatch has been thought to always lower the binding affinity. However, according to the PDNN model, the mismatch represents an energy cost for GSB but not for NSB. (The stacking energies and weight factors for NSB are the same for PM and MM probes.) Because of the mismatch in the center base, the NSB stacking energies relating to the middle three bases are different in a probe pair. This energy difference becomes a determining factor for $\ln(\text{PM}/\text{MM})$ values especially when gene expression level is low. This energy difference is highly correlated with the averages of $\ln(\text{PM}/\text{MM})$ values ($r = 0.95$) (Fig. 3).

To further validate our model, we analyzed samples in which either GSB or NSB was present, but not both. First, we observed that samples containing only a handful of genes, in which cases the source of NSB were minimal, produced virtually no negative PM – MM values (detailed data not shown). We then hybridized two *Drosophila melanogaster* RNA samples on two mouse arrays (MG_U74a), where little GSB was expected to be present. We observed that even though the overall signal intensity on these arrays was ~20-fold less than that observed with mouse RNA samples hybridized to the array, the relative values of probe signals in a probe set remained about the same. Therefore, $\ln(\text{PM}/\text{MM})$ values (Fig. 3, red line) were also similar. These results support the model's prediction that negative PM – MM values are caused by differences in NSB stacking energies.

Quantification of NSB is critical for interpretation of microarray data. For this task our model offers a simple rule: following Equation

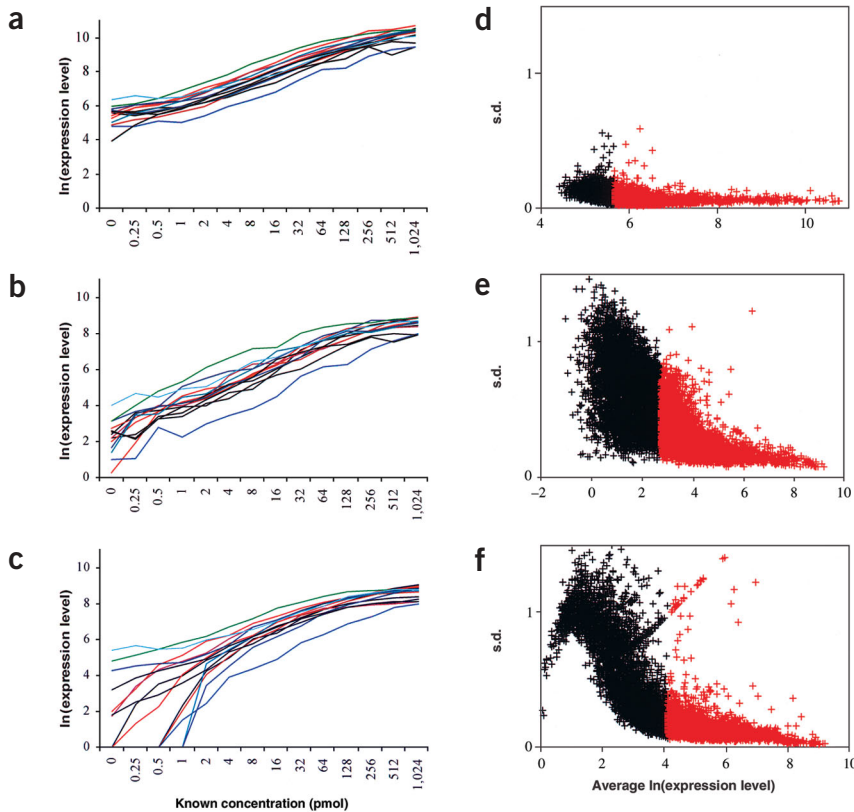


Figure 2 Accuracy test. Known concentrations of 14 'spike-in' genes are compared with those obtained from (a) PDNN, (c) MAS5.0 and (e) dChip. Each line represents a gene in 14 samples. The microarray raw data were obtained from the '1532 series' human data (see Methods for URL). For genes other than the 'spike-ins', standard deviations (s.d.) versus the averages of the log-transformed expression levels are shown in b, d and f as determined using PDNN, MAS5.0 and dChip⁷, respectively. Each of these figures contains 12,474 genes; top half shown in red.

Finally, the model provides a practical guide for microarray design in terms of probe selection. A previous study¹⁹ of probe selection warned that because hybridization behavior on microarrays is different from that in solution, simply using the parameters obtained from solution studies would be inappropriate. With our model, it is relatively straightforward to select probes according to the following criteria: (i) the chosen sequence should have a unique relationship with a targeted gene; and (ii) it should have a low E value and a high E^* value, so that the GSB signal is high while the NSB signal is low. Further improvement may also be achieved by taking into account alternative splicing sites and removing all MM probes. We have noticed that the newer array

(1), $N^*/(1 + \exp(E^*))$ is the expected amount of NSB for a probe. The only parameter that needs to be estimated for a given sample is N^* . The absence of probe signals substantially below the NSB baseline for any given level of E^* indicates that the estimate of NSB is reasonable (Fig. 1d).

A key advantage of the PDNN model is that it offers a means to check data quality and appropriateness of probe design. Problematic probe signals can be detected directly from the model fitting. Detection of outliers is also possible with dChip⁷, but the program requires multiple samples to determine statistical outliers and it cannot identify probe signals that are consistently in error owing to array design. The PDNN model does not have these limitations. Another feature of the PDNN model is that determination of NSB and GSB requires only PM probe signals. Therefore, the capacity of the array can be doubled through replacement of MM probes with additional PM probes.

designs are better than their predecessors in terms of reduced NSB and background noise (our unpublished observations). Given the short history of microarray technology, it is safe to predict that the technology will become much more powerful in the near future as more is learned about the physical mechanisms of microarrays.

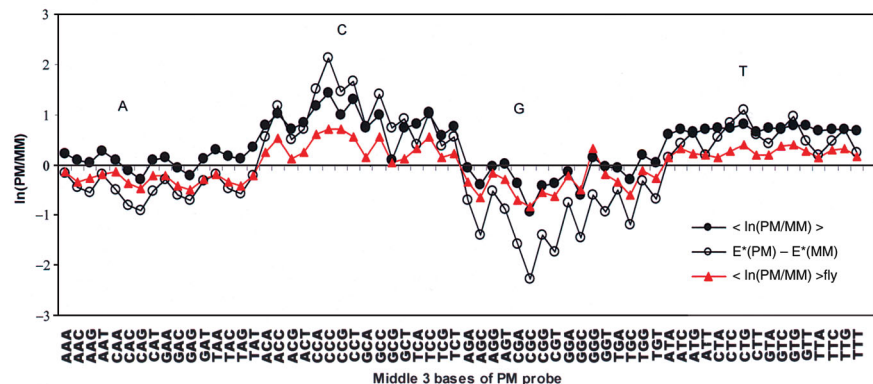
METHODS

PDDN model. In the PDNN model, a probe's signal is decomposed into three components as follows:

$$\hat{I}_{ij} = N_j / (1 + \exp(E_{ij})) + N^* / (1 + \exp(E_{ij}^*)) + B \quad (1)$$

where \hat{I}_{ij} is the expected signal of the i th probe in a probe set targeted to detect gene j ; the three terms on the right represent GSB, NSB and a uniform background (B), respectively. N_j is the number of expressed mRNA molecules from gene j and N^* is the population of RNA molecules that contributes to NSB. From this equation, E_{ij} is the free energy for formation of the specific

Figure 3 Evaluating probe pair signals with PDNN model. Average $\ln(\text{PM}/\text{MM})$ values were obtained from background-subtracted probe signals. Filled black circles represent data from a human RNA sample hybridized on HG-U94av2 array; Filled red circle, data from a *D. melanogaster* RNA sample hybridized on a mouse MG-U74a array. Standard deviations of $\ln(\text{PM}/\text{MM})$ values are -0.5 – 0.7 . Open circles represent $E^*(\text{PM}) - E^*(\text{MM})$ as determined using Equation (3) and the parameter values shown in **Figures 1b** and **c**. Note that the level of gene expression also affects $\ln(\text{PM}/\text{MM})$ values; $E^*(\text{PM}) - E^*(\text{MM})$ is a good predictor of $\ln(\text{PM}/\text{MM})$ value when expression level is low so that the effects of GSB are negligible.



RNA-DNA duplex with the targeted gene. By analogy, E_{ij}^* is the average free energy for NSB, that is, formation of duplexes with many different genes.

Note that N_j is assumed to have the same value within a probe set, while N^* and B are assumed to be constant throughout an array. Energy values in Equation (1) are in units of $k_B T$, where k_B is the Boltzmann constant. Given the sequence of a probe as $(b_1, b_2, \dots, b_{25})$, E_{ij} and E_{ij}^* are calculated as weighted sums of stacking energies:

$$E_{ij} = \sum \omega_k \varepsilon(b_k, b_{k+1}) \quad (2)$$

$$E_{ij}^* = \sum \omega_k^* \varepsilon^*(b_k, b_{k+1}) \quad (3)$$

where ω_k and ω_k^* are weight factors that depend on the position along the probe from the 5' end to the 3' end. The $\varepsilon(b_k, b_{k+1})$ term is the same as the stacking energy used in the nearest-neighbor model⁸.

Best values for all the parameters involved in the model can be obtained by minimizing the fitness function F to optimize the match between the expected signal intensity values (\hat{I}_{ij}) and the observed signal intensity values (I_{ij}) with

$$F = \sum (\ln \hat{I}_{ij} - \ln I_{ij})^2 / M \quad (4)$$

where M is the total number of probes on an array.

Minimization of F was performed through a Monte Carlo simulation procedure. Initially, the stacking energies were set to be random between -1 and 1 while the weight factors were held constant to be 1 . After F reached its minimum, the weight factors were allowed to change along with stacking energies to further minimize F .

Given the energy parameters, the gene expression level was calculated as follows:

$$N_j = \sum \{ [I_{ij} - B - N^* / (1 + \exp(E_{ij}^*))] / \lambda_{ij} \} / \sum [1 / (1 + \exp(E_{ij}))] / \lambda_{ij} \quad (5)$$

where $\lambda_{ij} = \sqrt{[I_{ij} (1 + \exp(E_{ij}))]}$. The summations include all probes in a probe set except for probes that have $I_{ij} - B - N^* / (1 + \exp(E_{ij}^*)) < 0$ (that is, there is no gene-specific signal) or have $|\ln \hat{I}_{ij} - \ln I_{ij}| > 3\sigma$ (that is, an outlier), where $\sigma^2 = F$. The expression levels were scaled so that the average is 500 on an array.

URLs: A computer program, PerfectMatch, designed for data analysis using the model is available online at <http://bioinformatics.mdanderson.org>. Details of the '1532 series' human microarray data are available at http://www.affymetrix.com/analysis/download_center2.affx.

ACKNOWLEDGMENTS

We thank Keith Baggerly, Kevin R. Coombes, Kenneth Hess, Jan Hermans, Roberto Carta, Jing Wang, David Gold, LeeAnn Chastain and Zoltan Szallasi for suggestions on the manuscript and Nobert Wilke and Mini Kapoor for technical support. This

work was supported by The University of Texas M.D. Anderson Cancer Center, a grant (DA14167) from the National Institute for Drug Abuse and funding from the State of California.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 3 December 2002; accepted 4 March 2003
Published online 8 June 2003; doi:10.1038/nbt836

1. Lockhart, D.J. & Winzler, E.A. Genomics, gene expression and DNA arrays. *Nature* **405**, 827–836 (2000).
2. van Berkum, N.L. & Holstege, F.C. DNA microarrays: raising the profile. *Curr. Opin. Biotechnol.* **12**, 48–52 (2001).
3. Southern, E., Mir, K. & Shchepinov, M. Molecular interactions on microarrays. *Nat. Genet.* **21**, 5–9 (1999).
4. Lockhart, D.J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14**, 1675–1680 (1996).
5. Zhou, Y. & Abagyan, R. Match-Only Integral Distribution (MOID) Algorithm for high-density oligonucleotide array analysis. *BMC Bioinformatics* **3**, 3 (2002).
6. Naef, F., Hacker, C.R., Patil, N. & Magnasco, M. Characterization of the expression ratio noise structure in high-density oligonucleotide arrays. *Genome Biol.* **3**, research0018 (2002).
7. Li, C. & Wong, W.H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **98**, 31–36 (2001).
8. Sugimoto, N. *et al.* Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry* **34**, 11211–11216 (1995).
9. Breslauer, K.J., Frank, R., Blocker, H. & Marky, L.A. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA* **83**, 3746–3750 (1986).
10. Shchepinov, M.S., Case-Green, S.C. & Southern, E.M. Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays. *Nucleic Acids Res.* **25**, 1155–1161 (1997).
11. Forman, J.E., Walton, I.D., Stern, D., Rava, R.P. & Trulson, M.O. in *Molecular Modeling of Nucleic Acids* vol. 682 206–228 (American Chemical Society, Washington, DC, USA, 1998).
12. Lazaridis, E.N., Sinibaldi, D., Bloom, G., Mane, S. & Jove, R. A simple method to improve probe set estimates from oligonucleotide arrays. *Math. Biosci.* **176**, 53–58 (2002).
13. Lemon, W.J., Palatini, J.J.T., Krahe, R. & Wright, F.A. Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. *Bioinformatics* **18**, 1470–1476 (2002).
14. Chu, T.M., Weir, B. & Wolfinger, R. A systematic statistical linear modeling approach to oligonucleotide array experiments. *Math. Biosci.* **176**, 35–51 (2002).
15. Chudin, E. *et al.* Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biol.* **3**, research0005 (2002).
16. Irizarry, R.A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).
17. Dudley, A.M., Aach, J., Steffen, M.A. & Church, G.M. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc. Natl. Acad. Sci. USA* **99**, 7554–7559 (2002).
18. Yuen, T., Wurmbach, E., Pfeffer, R.L., Ebersole, B.J. & Sealfon, S.C. Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res.* **30**, e48 (2002).
19. Li, F. & Stormo, G.D. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics* **17**, 1067–1076 (2001).

Erratum: Development of potent monoclonal antibody auristatin conjugates for cancer therapy

Svetlana O Doronina, Brian E Toki, Michael Y Torgov, Brian A Mendelsohn, Charles G Cerveny, Dana F Chace, Ron L DeBlanc, R Patrick Gearing, Tim D Bovee, Clay B Siegall, Joseph A Francisco, Alan F Wahl, Damon L Meyer & Peter D Senter
Nat. Biotechnol. 21, 778–784 (2003)

In the legend to **Figure 4d** on page 781, cIgG Ag⁻ should be cAC10 Ag⁻.

Corrigendum: A model of molecular interactions on short oligonucleotide microarrays

Li Zhang, Michael F Miles & Kenneth D Aldape
Nat. Biotechnol. 21, 818–821 (2003)

In the legend to **Figure 1** on page 819, text in parts **b** and **c** was transposed. The legend should have read as follows:

(**b**) Weight factors. (**c**) Nearest-neighbor stacking energy. These stacking energies weakly correlated ($r = 0.6$) with that found in aqueous solution⁸, and are smaller in magnitude.

In the legend to **Figure 2** on page 820, figure parts were referred to incorrectly. The legend should have read as follows:

Accuracy test. Known concentrations of 14 'spike-in' genes are compared with those obtained from (a) PDNN, (b) MAS5.0 and (c) dChip. Each line represents a gene in 14 samples. The microarray raw data were obtained from the '1532 series' human data (see Methods for URL). For genes other than the 'spike-ins', standard deviations (s.d.) versus the averages of the log-transformed expression levels are shown in **d**, **e** and **f** as determined using PDNN, MAS5.0 and dChip⁷, respectively. Each of these figures contains 12,474 genes; top half shown in red.