BIOINFORMATICS

Vol. 19 Suppl. 1 2003, pages i315–i322
DOI: 10.1093/bioinformatics/btg1044

Gene structure-based splice variant deconvolution using a microarry platform

Hui Wang[1, 2,*], Earl Hubbell[1], Jing-shan Hu[1], Gangwu Mei[1], Melissa Cline[1], Gang Lu[1], Tyson Clark[3], Michael A. Siani-Rose[1], Manuel Ares[3], David C. Kulp[1] and David Haussler[2]

[1]Affymetrix Inc. 3450 Central Expressway, Santa Clara, CA 95051, USA, [2]Department of Computer Science and Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA and [3]Department of Biology, University of California, Santa Cruz, 1156 High Street, CA 95064, USA

Received on January 6, 2003; accepted on February 20, 2003

ABSTRACT

Motivation: Alternative splicing allows a single gene to generate multiple mRNAs, which can be translated into functionally and structurally diverse proteins. One gene can have multiple variants coexisting at different concentrations. Estimating the relative abundance of each variant is important for the study of underlying biological function. Microarrays are standard tools that measure gene expression. But most design and analysis has not accounted for splice variants. Thus splice variant-specific chip designs and analysis algorithms are needed for accurate gene expression profiling.

Results: Inspired by Li and Wong (2001), we developed a gene structure-based algorithm to determine the relative abundance of known splice variants. Probe intensities are modeled across multiple experiments using gene structures as constraints. Model parameters are obtained through a maximum likelihood estimation (MLE) process/framework. The algorithm produces the relative concentration of each variant, as well as an affinity term associated with each probe. Validation of the algorithm is performed by a set of controlled spike experiments as well as endogenous tissue samples using a human splice variant array.

Contact: hui_wang@affymetrix.com

INTRODUCTION

Alternative splicing is an important regulatory mechanism, often controlled by developmental or tissue-specific factors. (Smith *et al.*, 1989; Hodges and Bernstein, 1994). Many alternatively spliced mRNAs may be expressed simultaneously in the same tissue, yielding an extensive set of proteins with distinct functions (Smith *et al.*, 1989; Kochiwa *et al.*, 2002). In human, approximately 30–60% of genes undergo alternative splicing (Sutcliffe and Milner, 1988; Croft *et al.*, 2000; Lander *et al.*, 2001; Venter *et al.*, 2001; Kochiwa *et al.*, 2002). In some cases, splice variants are associated with human diseases (Stallings-Mann *et al.*, 1996; Liu *et al.*, 1997; Siffert *et al.*, 1998).

Microarray technology has become a standard method for gene expression profiling. However, most microarray design and analysis is limited to detecting and measuring changes of expression for each gene. The current methods ignore, implicitly or explicitly, the presence of multiple splice variants in the same target mRNA pool. The reasons are many, but include the complexity of microarray designs to measure the multitude of splicing products and limitations of target labeling techniques. Being able to measure variant-level concentrations is important for accurate expression profiling, and consequently for obtaining a better understanding of the biological processes. Recently, several studies have applied microarray technology to this issue (Hu *et al.*, 2001; Miki *et al.*, 2001; Shoemaker *et al.*, 2001; Clark *et al.*, 2002; Kapranov *et al.*, 2002; Yeakley *et al.*, 2002). Genomic tiling arrays and exon arrays can be used to identify co-regulated exons, which allows the inference of variant mixtures (Shoemaker *et al.*, 2001; Kapranov *et al.*, 2002). Expression arrays with multiple probes have been retrospectively analyzed to identify exons that are differentially included or skipped in a tissue-specific manner (Hu *et al.*, 2001). RNA-mediated ligation combined with arrays presents a novel method for detecting exon-exon junction information of known splice variants (Yeakley *et al.*, 2002). Most recently splice junction spanning oligonucleotides representing nearly all yeast splicing events have been used to monitor the genome-wide effects of splicing factor mutations in yeast (Clark *et al.*, 2002), suggesting exon joining information can be accessed using oligonucleotide arrays. To date, there is no analysis

*To whom correspondence should be addressed.

that provides quantitative measure of different variants' expression levels.

Li and Wong adopted a model-based approach to estimate gene expression by fitting expression data at the probe level across multiple experiments (Li and Wong, 2001). Their model applies a simple formula linking intensity to concentration using the fact that all probes from the same probe set hybridize to the same target. This approach can reasonably address probe specific behavior and detect and eliminate outlier probes to give better expression estimates.

Inspired by their method, we developed an algorithm to estimate splice variant expression level by incorporating gene structure information. The gene structure specifies the features of each variant, where features can be both exons and exon-exon junctions. Probes are tiled selectively along certain features. Thus the probe intensity reflects the total concentration of the feature to which the probe belongs. Since a combination of features defines a variant, probe intensity then reflects the total concentration of one or more transcripts. By capturing these relationships, we are able to deconvolute the relative abundance of each variant in a set of samples. Data from these probes is fit across multiple experiments with a squared error loss function used to minimize the differences between predicted and observed values. Parameters are estimated iteratively using a maximum likelihood estimation (MLE) framework. The algorithm outputs the relative concentration of each variant, as well as an affinity term associated with each probe. Its efficiency is demonstrated through experiments on spiked clones and endogenous tissue samples.

METHODS AND MATERIALS

A special splice variant chip was designed using 21 well-characterized genes. The chip design process includes sequence selection, gene selection and probe selection.

Sequence selection

All mRNA and cDNA sequences were mapped to the Golden Path Genomic sequences (April 2001 release) using psLayout (Kent and Haussler, 2001). Based on the alignments, we characterized genes and generated unique splice variants. Then gene features including exons, introns and junctions were extracted and loaded into a relational database.

Gene selection

21 genes were selected from the literature (including ACHE, APAF1, BCL2L1, BCLG, BRCA1, CALCA, CALCR, CASP2, CD44, DNMT3B, FGF1, IL15, ITGA3, ITGA6, MAPT, MYLK, PSEN1, THRA, TNNT2, TPM2 and WNT2B). These genes were selected because their alternative splicing has been well studied using standard techniques of RNA analysis. The regulation of these genes at the level of splicing plays an important role in biological processes such as cancer and muscle development. Sequence information of each gene is extracted from the sequence selection database described above. A splice variant chip is then designed based on the information of these 21 genes.

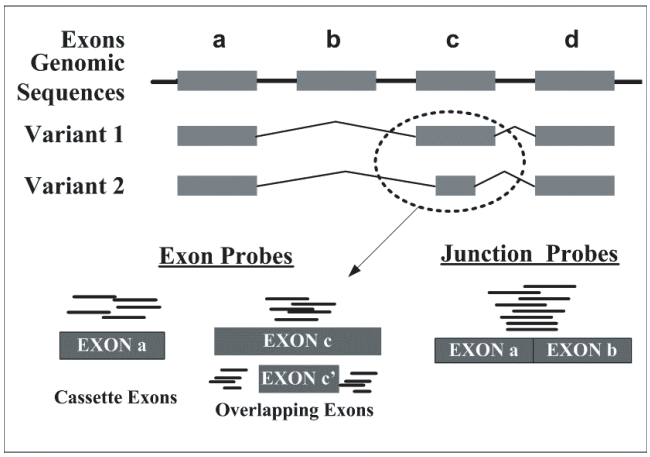

Fig. 1. Probe tiling.

Probe selection

There are two main types of probes: exon probes and junction probes. Exon probes are selected using Affymetrix's expression probe selection software (Mei *et al.*, 2003). If two exons overlap, probes are selected from the overlapping regions and the unique regions. Junction probe tiling is position-constrained. We choose eight symmetrically positioned probes across junctions. The center position of these probes relative to the junction are $-5,-3,-2,-1,+1,+2,+3,+5$. Figure 1 summarizes the probe tiling strategy.

Clones

Three CD44 splice variants represented by IMAGE clone ID: 588908 (clone 1), 118372 (clone 2) and 3638681 (clone 3) were purchased from *Invitrogen Inc.* The simplified structures of these clones are shown in Figure 5B.

ALGORITHM

Probe models

Based on the Li and Wong reduced model (Li and Wong, 2001), the relationship between probe intensity and target transcript concentration measured by probes and probe affinities can be expressed by the following formula:

$$y = PM - MM = \alpha x + \varepsilon \quad (1)$$

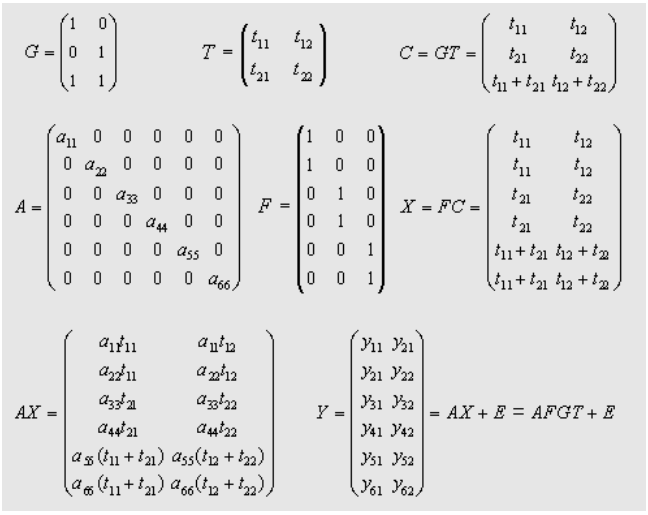

Fig. 2. Example of matrix representation. The example gene has two variants with 3 features and each feature contains 2 probes. Variant 1 has feature 1 and 3, while variant 2 has feature 2 and 3. Two experiments are performed.

Here, *PM* and *MM* are probe intensities for perfect match and mismatch respectively, the target transcript concentration measured by a probe is denoted by x, α denotes the probe affinity term and we use ε to denote the error term where $\varepsilon \sim N(0, \sigma^2)$.

Since a transcript is usually represented by multiple probes and has different concentrations in different experiments, the above formula is generalized to:

$$y_{ij} = a_i x_j + \varepsilon_{ij} \tag{2}$$

where i is the index for probes and j is the index for experiments. We extend Equation (2) to the problem of measuring the concentrations of several splice variants.

Models in the context of gene structure and application to alternative splicing

A transcript may be uniquely identified by a set of features, each of which may be represented by a series of probe sequences. A gene feature can be either an exon, intron, partial exon, intron, or a junction (exon-exon junction, exon-intron junction, intron-exon junction). Exon features can be partitioned further depending on whether the exon is a cassette exon or an exon overlapping with others. Intron features may be treated the same way. Probes can be mapped to the features that contain them, and in turn, the features can be mapped to the transcripts that contain them. We represent these relationships via matrices.

Typically, a gene structure contains all known transcripts of each gene and the feature composition for each transcript, but it also can contain only a subset of features of interest. The relationship between features and transcripts can be represented by a *q-by-t* matrix $G = (g_{lk})$ containing binary values of 1 or 0, where $g_{lk} = 1$ means feature l is present in transcript k, while $g_{lk} = 0$ means this feature is absent. The total number of transcripts is t and q is the total number of features. The transcript concentrations of a given gene in all experiments are represented by a *t-by-x* matrix $T = (t_{kj})$, where t_{kj} represents the concentration of transcript k in experiment j. Here x is the total number of experiments. Let $C = (c_{lj})$ be the *q-by-x* matrix defined by $C = GT$. It is easily seen that c_{lj} is the concentration of feature l in experiment j.

The mapping of probes to features is represented in a similar way by a matrix F. Multiple probes can be chosen to represent each gene feature and some probes can be in more than one feature. Matrix $F = (f_{il})$ is a *p-by-q* matrix with values 0 or 1, where p is the total number of probes, q is the total number of features, f_{il} equals 1 if probe i belongs to feature l, and f_{il}equals 0 otherwise. Let $X = FC$. Thus $X = (x_{ij})$ is a *p-by-x* matrix and x_{ij} is the sum of the concentrations in experiment j of all the features to which probe i belongs. By the definition of C and F, x_{ij} is the actual concentration of all the target transcripts in experiment j interrogated by probe i.

We develop an equation analogous to (1) that relates the matrix X of actual concentrations to the matrix Y of observed probe intensities. Let $A = (a_{ii})$ be a *p-by-p* diagonal matrix where a_{ii} represents the probe affinity term. The predicted probe intensities can then be expressed as $AX = AFGT$. The observed probe intensities are given by a *p-by-x* matrix $Y = (y_{ij})$, where y_{ij} is the intensity of probe i for experiment j. The observed probe intensities will equal the predicted probe intensities plus experimental error denoted by $E = (\varepsilon_{ij})$ as shown in Equation (2). Thus the matrix version of Equation (2) is $Y = AX + E = AFGT + E$. To illustrate this formulation, Figure 2 shows all matrices of a simple gene with 2 transcripts, 3 features and 2 probes per feature.

Model fitting and minimization

We want to minimize the differences between the predicted and observed intensities for all probes using a maximum likelihood framework. Since we are assuming Gaussian noise, this leads to a standard regression framework, so we use the squared error loss function.

The squared difference between predicted and observed intensity values for all probes of each gene can be written as function $f(A, T) = (||Y - AFGT||_2)^2$. We want to minimize f over the unknowns A and T.

Some constraints or penalty terms are needed in order to solve this minimization problem because it is under-constrained as stated. Thus the following constraints are

added:

$$\sum_{i=1}^{z} a_{ii}^2 = \text{constant} \quad (3)$$

$$a_{ii} \geqslant 0 \quad (4)$$

$$t_{mj} \geqslant 0 \quad (5)$$

Where z in equation (3) is the total number of probes used in the constraint. Equations (4) and (5) reflect the fact that concentration and affinity terms must be non-negative.

Alternatively to (3), we can add $\gamma(||A||_2)^2$ to f, where γ is a small positive constant.

Solving the minimization problem with constraints (3)–(5) corresponds to maximum likelihood estimation (MLE). This can be approached by alternately fixing A and solving for T, then fixing T and solving for A until convergence. Each step in this procedure is a linear least squares minimization with linear constraints. The final values of T and A yield the relative concentration of each transcript variant and the relative affinity term of each probe.

RESULTS

We validated our model using two approaches. First, we applied the model to a set of controlled experiments with spiked clones, and compared predicted concentrations to actual concentrations. Second, we applied it to the analysis of endogenous tissue samples, confirming the results with the TaqMan PCR assay. All experiments used a custom-designed Affymetrix microarray for detecting the 21 well-documented human genes that exhibit splice variation.

Two-variant spike experiments

We tested the accuracy and sensitivity of the algorithm with dilution experiments (using yeast complex background) using target preparations derived from pairs of cDNA clones representing two splice variants from the same gene. In one set, we mixed target derived from two CD44 variants (clone 1 and clone 2) with differing concentrations: the first variant ranged from 0 to 64 pM and the second variant ranged from 64 pM to 0 pM with the total concentration held constant at 64 pM. By diluting the whole set 4 and 16 times, we obtained further results for titration experiments with total concentrations of 16 pM and 4 pM respectively. The variant concentrations as well as the results from the algorithm are detailed in Figure 3.

In all three sets of experiments, the predicted concentration of each variant (indicated by bars in Fig. 3) is similar to the actual concentration (indicated by lines in Fig. 3). Furthermore, the individual concentrations are consistent between different experiments. For instance, the 8 pM concentration of variant 2 in the 64 pM set of experiments is comparable to the 8 pM sample in the 16 pM set, and the predictions for 4 pM concentration are similar in all three sets of experiments. Each ratio of the two variants was tested three times: at the 64 pM, 16 pM, and 4 pM total concentration levels. In each case, the predicted concentration mirrored the actual concentrations. Thus, we are able to compare the relative abundance of the targets in different samples. The results indicate that the algorithm is very sensitive, as it can detect concentrations as low as 0.5 pM.

This two-variant spike experiment was also done with different sets of genes, including ACHE, TPM2, MYLK and MAPT. Similar results were obtained for each of the different variant pairs (data not shown). Figure 4 shows the correlation of the predicted concentration with the actual concentrations of the two variants of CD44 and TPM2. The R^2 scores between the predicted concentrations and the actual concentrations for these tested pairs are greater than 0.94.

Three-variant spike experiments

In order to test a more general case, a third CD44 variant (clone 3) was added. The experiment was designed to test all possible combinations of clones at 0 and 4 pM under simple background. In general, the predicted concentrations are consistent with the actual concentration of each variant as shown in Figure 5A.

TPM2 tissue experiments

Further validation was performed on tissue samples, studying the gene TPM2. Beta-tropomyosin gene contains in its central portion two mutually exclusive exons (A and B). Variants containing exon A (TPM2-A) are mainly present in skeletal muscle, while variants containing exon B (TPM2-B) are present in non-muscle and smooth muscle tissues (MacLeod *et al.*, 1985; Helfman *et al.*, 1986; Widada *et al.*, 1988; Clouet d'Orval *et al.*, 1991; Lees-Miller and Helfman, 1991; Novy *et al.*, 1993; Beisel and Kennedy, 1994; Pittenger *et al.*, 1995; Gallego *et al.*, 1996). Figure 6A shows the predicted relative concentrations of TPM2-A and TPM2-B of 7 human tissues. Based on the prediction, TPM2-A is observed in adult and fetal skeletal muscle, as well as esophagus and fetal heart. TPM2-B is not observed in skeletal muscle, as expected, but is observed in esophagus, stomach, uterus, and fetal umbilical cord (Helfman *et al.*, 1986). The result is consistent with Taqman quantitative PCR validation for selected tissues (Fig. 6B).

DISCUSSION

This work demonstrates that our gene structure-based approach can be used to estimate the relative abundance of splice variants. The algorithm generates the relative concentration of each variant and an affinity term associated

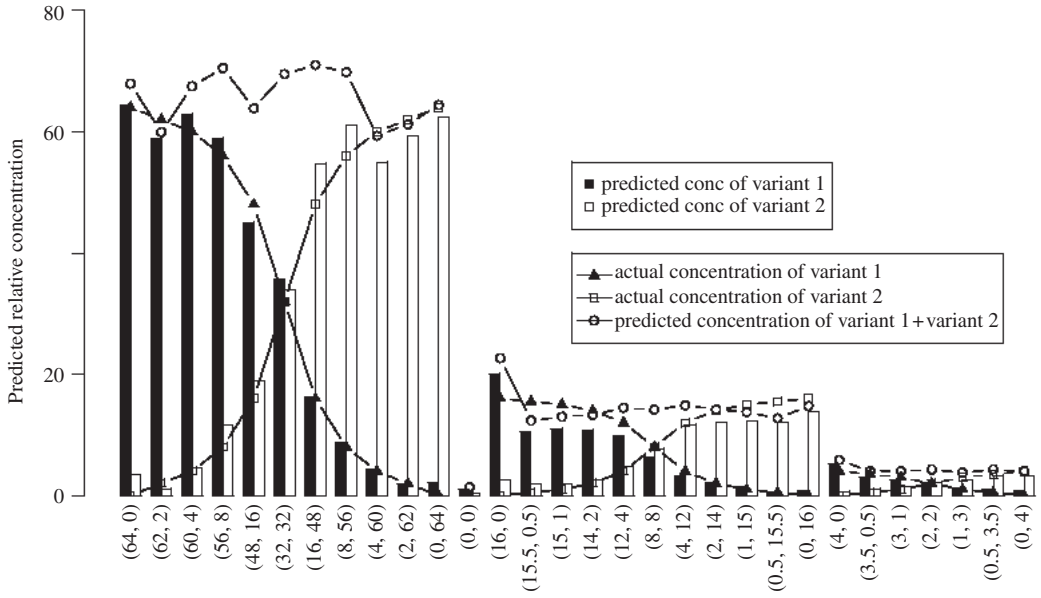

Fig. 3. Predicted relative concentration in two-variant titration experiments. Two CD44 variants were mixed at 30 different known concentrations. The experiments as well as the actual concentrations of each variant pair are indicated by *X*-axis. There are three sets of experiments. In each set, we vary the concentration of each variant while keeping the total concentration fixed. The total concentration of each set is 64 pM (samples 1–11), 16 pM (samples 13–23) and 4 pM (samples 24–30) respectively. Sample 12 is a control experiment. The *Y*-axis indicates the scaled predicted concentration of each variant as well as the total concentrations. For easy comparison, the actual concentrations are also plotted in the same chart.

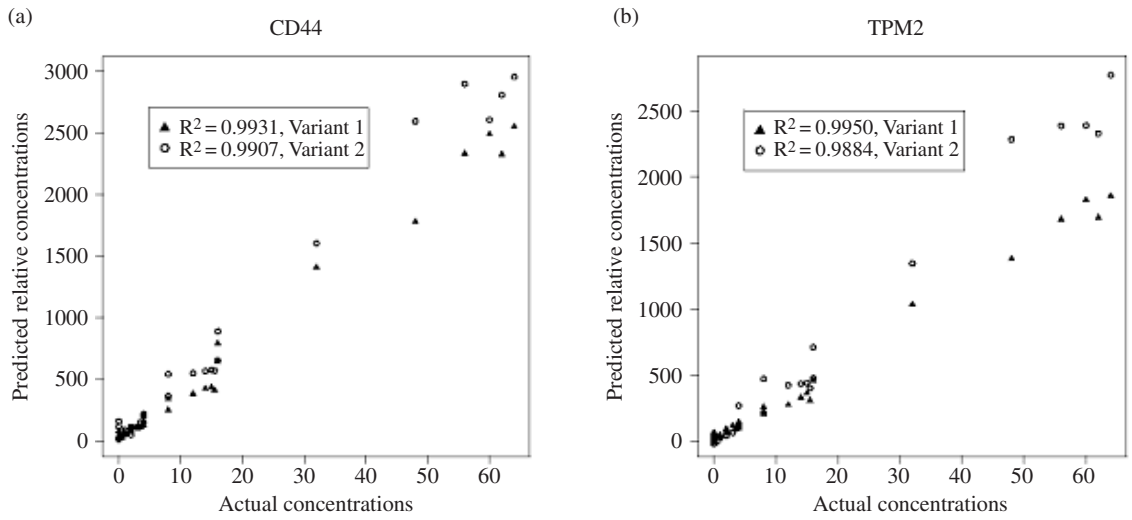

Fig. 4. Correlations between predicted concentrations and actual concentrations. *X*-axis is the actual concentrations of each of the two variants (indicated by bullets and triangles) in 30 experiments. *Y*-axis is the predicted relative concentration of each variant in these experiments.

with each probe. The predicted concentrations can be used to compare the expression level of multiple variants of the same gene in a sample as well as expression changes of the same variant across multiple samples.

Generic model

As described above, the reduced probe model assumes that mismatch probes account for all non-specific hybridizations. However this is often not true. A more

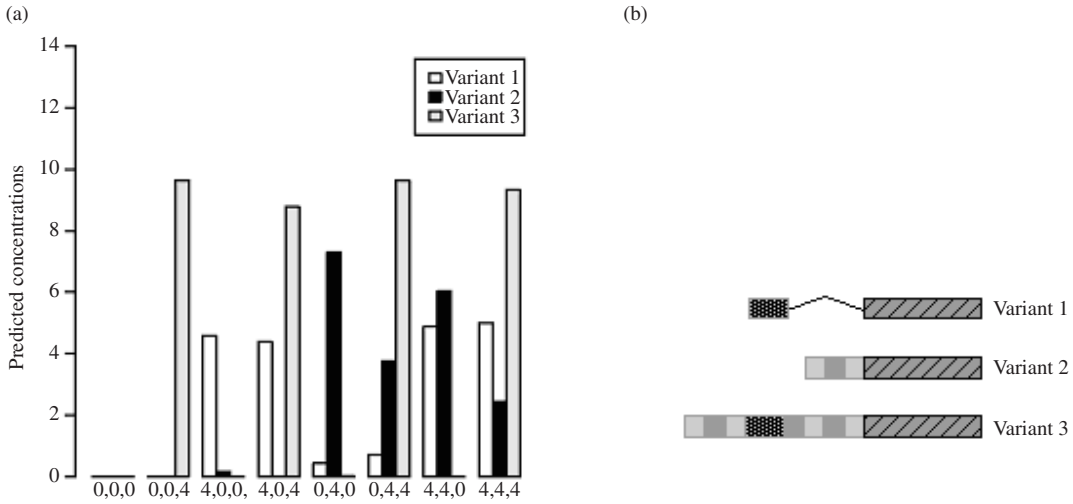

Fig. 5. (A) Predicted relative concentrations in three-variant titration experiments. Three CD44 variants are spiked in with different concentrations. The concentration of each variant is shown along the X-axis. The Y-axis indicates the predicted relative concentration of each variant. **(B) Cartoon representation of CD44 variants.** Variant 1 has a unique exon-exon junction compared with variant 2 and 3. Variant 2 is totally contained within variant 3.

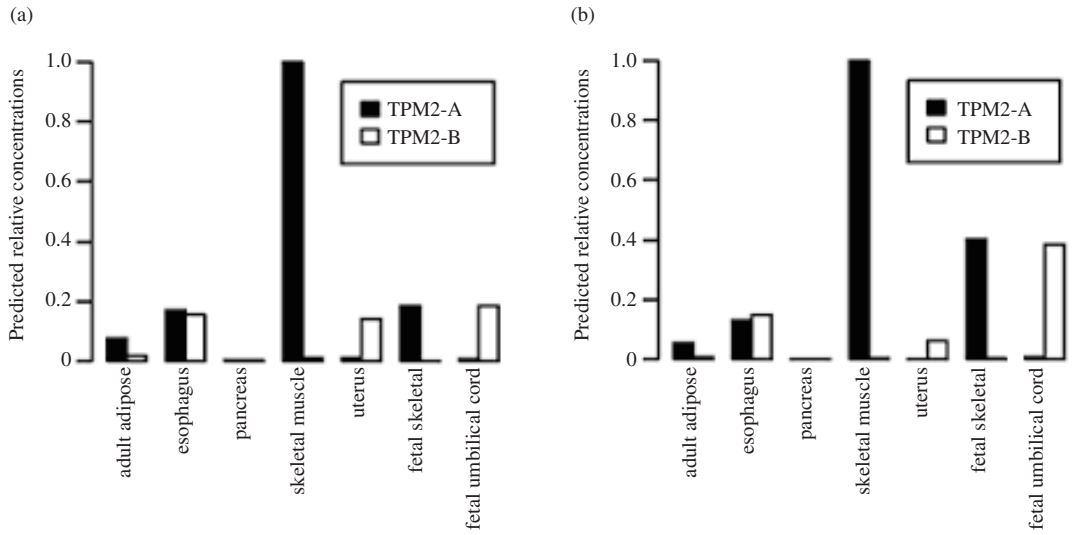

Fig. 6. (A) Predicted relative concentrations for TPM2 in Human tissues. Tissue samples are indicated along the X-axis. The Y-axis indicates the predicted relative concentration of each variant. **(B) TaqMan results of TPM2-A and TPM2-B in Human tissues.** The Y-axis is the scaled molar amount of the variants measured by TaqMan technique.

generic model includes a background term for each probe. The probe model in formula (1) is then expressed as:

$$y = ax + b + \varepsilon. \tag{6}$$

If we let the column vector $\vec{b} = (b_i)$ represent probe-specific background terms, $\vec{1} = (1_j)$ be the row vector of 1s, and $B = \vec{b}\vec{1}$ be the outer product of these, then as above we can generalize (6) to

$$Y = AD + B + \varepsilon = AFGT + B + \varepsilon \tag{7}$$

Since B is treated as a property of probe, in the minimization process we solve for B at the same time we solve for the affinity term A.

Limitations of the algorithm

Degeneracy occurs when there is no unique solution for each of the variants. As mentioned above, the G matrix represents the relationship between transcripts and features of interest. It is obvious, for example, if the number of features is less than the number of transcripts, there is no unique solution. A simple alternative is to combine and solve for the concentration of several transcripts altogether. Other complications such as the 'ill-conditioned' situation, a classical matrix computation problem, can make computation quite difficult. Many techniques such as orthogonal transformation can be applied to help solve the problem.

This algorithm is intended for splice variant typing, not discovery. A limitation exists when the input gene structure is incorrect, which can happen when there are unknown transcripts present in the test samples. The robustness of the method is a topic of ongoing research.

Three-variant spike experiments

Even though the predicted concentrations are consistent with the variants' actual concentrations, some inconsistencies are evident (Fig. 5A). First, the concentration of variant 1 appears to be lower than that of variant 2 and 3 (experiments 2, 3 and 5). Given careful examination and gel analysis, it appears that the actual spiked concentrations of variants 2 and 3 are higher than 4 pM due to a consistent error in estimation of the molar amount of spiked transcripts. This error is probably caused by the greater efficiency of full length transcript synthesis for the shorter variant transcripts in our *in vitro* transcription reactions.

Second, both experiment 5 (0,4,0) and 6 (0,4,4) show non-zero concentrations of variant 1. We hypothesize that it is related to a splice variant specific junction effect: cross hybridization from partially-overlapping junctions, specifically those beginning or ending at the same exon. In this example, the junction probes of variant 1 partially overlap with those of variant 2 and 3 (Fig. 5B). We call these partially-overlapping junctions *competitive junctions*. Future work will include development of a model for this junction-specific effect.

In conclusion, we have developed an efficient algorithm for estimating the relative concentrations of splice variants. This algorithm can potentially help in obtaining a more accurate interpretation of microarray data and thus a better understanding of biological functions.

ACKNOWLEGEMENTS

The authors would like to thank Tom Ryder for scientific discussions of the project. We are also grateful to Keith Jones for careful reading and critical comments on the manuscript.

REFERENCES

Beisel,K. and Kennedy,J. (1994) Identification of novel alternatively spliced isoforms of the tropomyosin-encoding gene, TMnm, in the rat cochlea. *Gene*, **143**, 251–256.

Clark,T.A., Sugnet,C.W. *et al.* (2002) Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*, **296**, 907–910.

Clouet d'Orval,B., d'Aubenton Carafa,Y. *et al.* (1991) RNA secondary structure repression of a muscle-specific exon in HeLa cell nuclear extracts. *Science*, **252**, 1823–1828.

Croft,L., Schandorff,S. *et al.* (2000) ISIS, the Intron Information System, reveals the high frequency of alternative splicing in human genome. *Nature Genet.*, **24**, 340–341.

Gallego,M., Sirand-Pugnet,P. *et al.* (1996) Tissue-specific splicing of two mutually exclusive exons of the chicken beta-tropomyosin pre-mRNA: positive and negative regulations. *Biochimie*, **78**, 457–465.

Helfman,D., Cheley,S. *et al.* (1986) Nonmuscle and muscle tropomyosin isoforms are expressed from a single gene by alternative RNA splicing and polyadenylation. *Mol. Cell Biol.*, **6**, 3582–3595.

Hodges,D. and Bernstein,S. (1994) Genetic and biochemical analysis of alternative RNA splicing. *Adv. Genet.*, **31**, 207–281.

Hu,G.K., Madore,S.J. *et al.* (2001) Predicting splice variant from DNA chip expression data. *Genome Res.*, **11**, 1237–1245.

Kapranov,P., Cawley,S.E. *et al.* (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, **296**, 916–919.

Kent,W.J. and Haussler,D. (2001) Assembly of the working draft of the human genome with GigAssembler. *Genome Res.*, **11**, 1541–1548.

Kochiwa,H., Suzuki,R. *et al.* (2002) Inferring alternative splicing patterns in mouse from a full-length cDNA library and microarray data. *Genome Res.*, **12**, 1286–1293.

Lander,E.S., Linton,L.M. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lees-Miller,J. and Helfman,D. (1991) The molecular basis for tropomyosin isoform diversity. *Bioessays*, **13**, 429–437.

Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Liu,W., Qian,C. *et al.* (1997) Silent mutation induces exon skipping of fibrillin-1 gene in Marfan syndrome. *Nature Genet.*, **16**, 328–329.

MacLeod,A., Houlker,C. *et al.* (1985) A muscle-type tropomyosin in human fibroblasts: evidence for expression by an alternative RNA splicing mechanism. *Proc. Natl Acad. Sci. USA*, **82**, 7835–7839.

Mei,R., Hubbell,E. *et al.* (2003) Probe selection for high-density oligonucleotide arrays. Submitted.

Miki,R., Kadota,K. *et al.* (2001) Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18 816 full-length enriched mouse cDNA arrays. *Proc. Natl Acad. Sci. USA*, **98**, 2199–2204.

Novy,R., Lin,J. *et al.* (1993) Human fibroblast tropomyosin isoforms: characterization of cDNA clones and analysis of tropomyosin isoform expression in human tissues and in normal and transformed cells. *Cell Motil Cytoskeleton*, **25**, 267–281.

Pittenger,M., Kistler,A. *et al.* (1995) Alternatively spliced exons of the beta tropomyosin gene exhibit different affinities for F-actin

and effects with nonmuscle caldesmon. *J. Cell Sci.*, **108**, 3253–3265.

Shoemaker,D.D., Schadt,E.E. *et al.* (2001) Experimental annotation of the human genome using microarray technology. *Nature*, **409**, 922–927.

Siffert,W., Rosskopf,D. *et al.* (1998) Association of a human G-protein beta 3 subunit variant with hypertension. *Nature Genet.*, **18**, 45–48.

Smith,C., Patton,J. *et al.* (1989) Alternative splicing in the control gene expression. *Annu. Rev. Genet.*, **23**, 527–577.

Stallings-Mann,M., Ludwiczak,R. *et al.* (1996) Alternative splicing of exon 3 of the human growth hormone receptor is the result of an unusual genetic polymorphism. *Proc. Natl Acad. Sci. USA*, **93**, 12394–12399.

Sutcliffe,J.G. and Milner,R.J. (1988) Alternative mRNA splicing: the shaker gene. *Trends Genet.*, **4**, 297–299.

Venter,J.C., Adams,M.D. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Widada,J., Ferraz,C. *et al.* (1988) Complete nucleotide sequence of the adult skeletal isoform of human skeletal muscle beta-tropomyosin. *Nucleic Acids Res.*, **16**, 3109.

Yeakley,J.M., Fan,J.B. *et al.* (2002) Profiling alternative splicing on fiber-optic arrays. *Nat. Biotechnol.*, **20**, 353–358.