J|A|C|S
COMMUNICATIONS

Published on Web 04/11/2002

Arrayed Primer Extension Computing with Variant mRNA Splice Forms. Multiple Isoforms of CD44 in a Human Breast Tumor

Hyunsoo Kim and Michael C. Pirrung*

Department of Chemistry, Levine Science Research Center, Duke University, Durham, North Carolina 27708-0317

Received September 4, 2001

The completion of the human genome sequence[1] has revealed that the total number of genes is less than earlier estimated, perhaps 30 000. Because this value is also significantly less than the number of human cDNAs[2] that have been identified, perhaps 120 000, it is apparent that a portion of protein sequence diversity is due not to genomic sequences but to alternative splicing of initial RNA transcripts. Of 14 000 known human genes, >40% possess multiple variant spliced forms.[3] Understanding the biological functions of the genome will therefore require methods to examine each exon in variantly spliced mRNAs. Ideally, such methods would permit many RNAs to be examined simultaneously. Microarrays are ideal for highly parallel analysis, but microarray techniques with fidelity adequate to analyze complex mRNA for individual exons are unknown. APEX microarray RNA analysis[4] has therefore been developed. Microarrays of exon-specific and junction-specific primers for the variantly spliced exons of the CD44 locus accurately detect the presence of included exons in mature RNA transcripts when used with RNAseH (−) reverse transcriptase. This microarray technique was applied to a human breast tumor sample, enabling the certain detection and assignment of four alternatively spliced CD 44 transcripts.

Understanding the spliced forms of a mature mRNA begins with identification of the exon usage. The presence or absence of an exon is a Boolean variable, and the structure of an mRNA formed from a genomic sequence of *n* exons can be represented by an *n*-bit binary number. For example, CD44 is a cell adhesion molecule[5] found on the surface of all mammalian cells. Its genetic locus comprises 20 exons, 10 of which are alternatively spliced (Figure 1). Thus, a molecular code for the structure of a CD44 mRNA is

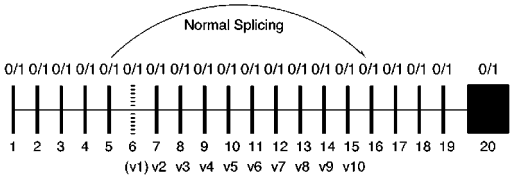

Figure 1. The CD44 locus comprises 20 exons, of which exons 6−15 are omitted in normal splicing. The variably spliced exons have alternative descriptors v2−v10 (in man, v1 has an in-frame stop codon and is silent).

simply a 20-bit number. So far as is known, exons 1−5 and 16−20 are always included, so codes for mRNAs for all functional CD44s begin and end with 11111. The normally spliced mRNA would be coded 11111000000000011111; known variants will have one or more of the 0's switched to 1's. This results in >1000 (2^{10}) possible spliced transcript sequences. Determination of the structure of a mature CD44 mRNA is thus a combinatorial analytical problem.[6] When multiplied by the vast number of genes, analysis of RNA splicing across the genome would become intractable (like RNA folding,[7] it may be NP-complete[8]). Powerful methods based on parallel processing, such as microarray techniques, are needed to enable global investigation of splicing.

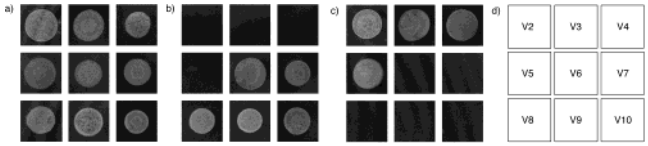

Figure 2. RNA APEX on a CD44 exon-specific microarray. The spot size is ~80 μm. The APEX reaction mix includes 1× RT buffer, 10 mM DTT, 18 μM ddNTP (minus T), 20 pM Fl-ddUTP, 0.4 M trehalose, 10 U RNAse inhibitor, and 400 U RNAse H(−) MMLV RT in 50 μL. (a) APEX with RNA template prepared from CD44 v2−v10. The average signal-to-background (S/B) ratio (the spot compared to adjacent sites without primer) across all primers is 18. (b) APEX with v6−v10 template; (c) APEX with v2−v5 template; (d) key.

Valuable microarray methods have been devised for studying mRNA or cDNA levels in parallel.[9] Hybridization data, which are in analog form, can be used for differential gene expression analysis.[10] Hybridization to "exon arrays" has also been used in the annotation of the human genome.[11] As described above, the analysis of mRNA splicing variants calls for digital information. The APEX (arrayed primer extension) method[12] provides high-fidelity, essentially digital detection of nucleic acid sequences with high parallelism. APEX exploits single-nucleotide extension of microarrays of DNA primers that are complementary to a template. APEX has been used for the solution of Boolean problems, including those of the NP-complete class[4a] and thus seemed ideal for the analysis of variantly spliced mRNAs. The application of APEX to RNA templates and DNA primers requires extension by a reverse transcriptase (RT). We earlier investigated RNA APEX but had not applied it to microarrays or discovered an optimal RT.[13]

Initial experiments to analyze splice variants of CD44 RNA by APEX used primers addressing unique sites within each exon v2 through v10. Oligonucleotides were synthesized with 5′-T_{10} linker sequences, 5′-phosphitylated, and sulfurized with Beaucage reagent. Microarrays were prepared by spotting the resulting 5′-phosphorothioate primers onto bromoacetamide glass using our previously described method.[14] Templates were prepared from cDNA[15] by PCR with forward primers bearing T7 RNA polymerase promoters followed by in vitro transcription. The full-length RNA was partially hydrolyzed by treatment with concentrated NH_4OH at 70 °C for 3 min, resulting in average template length of 40 nt. Template (~1 μg) was hybridized to the microarray, and APEX reactions were performed with MMLV RNAse H(−) reverse transcriptase. The use of the RNAse H(−) RT and the fragmentation step[16] are crucial to RNA APEX.

With a RNA template including each of the nine variable exons, APEX signals are seen for all primers (Figure 2a). A template bearing only exons v6−v10 gives an average APEX S/B of 12−

10.1021/ja012102a CCC: $22.00

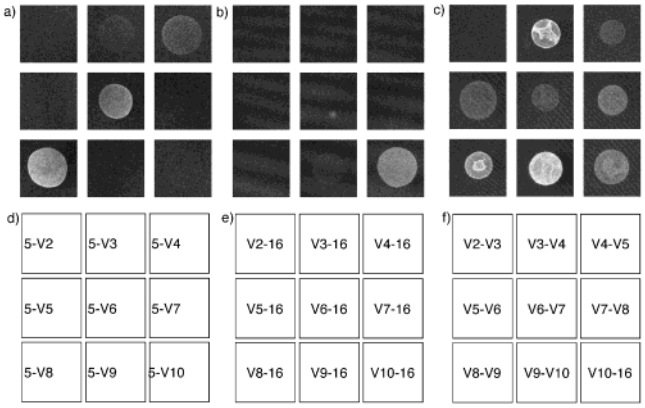

Figure 3. RNA APEX with a human breast tumor CD44 template. (a) Exon 5−variable exon border primers. (b) Variable exon−exon 16 border primers. (c) Variable exon−variable exon border primers. (d) Key for (a). (e) Key for (b). (f) Key for (c).

20 at primers addressing exons included in the template, and of <1 at primers addressing exons absent from the template (Figure 2b). Similar high fidelity is observed with template bearing only exons v2−v5 (Figure 2c). The Boolean codes, or "splicotypes", for the variable region v2−v10 of this locus are: (a) 111 111 111, (b) 000 011 111, (c) 111 100 000.

Many human diseases are related to RNA splicing.[17] Variant splicing and up-regulation of the expression of CD44, which generally relate to a poor prognosis, can occur in cancer.[18] The functional reasons for this observation seem to relate to enhancement of metastatic motility[19] and adhesion power of transformed cells. One particular CD44 antigen, v6, is found in many breast,[20] bladder,[21] and colon[22] cancers. Alternative splicing of the CD44 mRNA is due not to genomic changes but to trans-acting splicing factors.[23] CD44 has been examined as a molecular diagnostic in cancer. Protein levels[24] or RNA expression levels[25] have been studied, but the absence of variant-specific reagents has limited the data that can be obtained and the conclusions that can be drawn. We therefore examined the utility of the RNA APEX microarray assay for variant splice forms of CD44 in breast cancer.

Total cDNA was obtained from the primary breast tumor of a single patient. The variable CD44 locus was amplified by two rounds of PCR (*Pfu* polymerase followed by Thermosequenase) with exon 5 and reverse exon 16 primers. The amplification product showed at least five bands (of 160, 550, 800, 1200, and 1300 bp). The 160 bp amplicon is standard CD44 (without variant exons); the other amplicons represent variant splice forms.[26] Analysis of this mixed RNA population (resulting from multiple cell types within the tumor sample) with the exon-specific microarray gave the tentative variant splicotype of 0- - - - - -1,[27] with mixed signals for exons v3−v9. An additional microarray was prepared with primers specific for the junctions of exon 5 with each variant exon and for the junctions of each variant exon with exon 16. For the latter, only junctions 5−16 (standard) and v10−16 are observed, while the former shows junctions 5−v3, 5−v4, 5−v6, and 5−v8 (Figure 3). The variant splicotypes in this sample, which are consistent with amplicon sizes, were thereby established as 011 111 111, 001 111 111, 000 011 111, and 000 000 111.[28]

To our knowledge, there have been no previous reports of the presence of more than two CD44 isoforms within the same tissue. This microarray method can detect any type of variant among known exons, even when multiple forms exist in the same tissue sample. Application of these methods to analysis of human tumors, tissues, and cell lines should provide a much better picture of the diversity in CD44 splicing.

This RNA APEX analysis method provides a high-fidelity, digital signal of the variantly spliced exons within an RNA. With RNA samples containing only one spliced form, splicotypes can be determined using microarrays bearing only exon-specific primers. With heterogeneous RNAs, multiple splicotypes can be assigned using additional splice-specific primers. Applications of this technique should be broad, given an extensive database[29] of alternatively spliced exons.

Acknowledgment. Financial support was provided by NIH GM 46720 and NSF EIA-0086015. We thank Dr. U. Grünthert for the CD44 cDNA.

Supporting Information Available: Sequences of oligonucleotide primers, protocols for amplification, and analysis of breast tumor CD44 (PDF). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Venter, J. C.; et al. *Science* **2001**, *291*, 1304.
(2) Claverle, J.-M. *Science* **2001**, *291*, 1255.
(3) Sakharkar, M. K.; Kangueane, P.; Woon, T. W.; Tan, T. W.; Kolatkar, P. R.; Long, M.; de Souza, S. J. *Bioinformatics* **2000**, *16*, 1151−2. Sakharkar, M. K.; Long, M.; Tan, T. W.; de Souza, S. J. *Nucleic Acids Res.* **2000**, *28*, 191−2.
(4) (a) Pirrung, M. C.; Connors, R. V.; Montague-Smith, M. P.; Odenbaugh, A. L.; Walcott, N. G.; Tollett, J. J. *J. Am. Chem. Soc.* **2000**, *122*, 1873. (b) Tollett, J. J.; Kurg, A.; Shah, A.; Roa, B. B.; Richards, C. S.; Nye, S. H.; Pirrung, M.; Metspalu, A.; Shumaker, J. M. *Am. J. Hum. Gen.* **1997**, *61*(S), 1322.
(5) Goodison, S.; Urquidi, V.; Tarin, D. *Mol. Pathol*. **1999**, *52*, 189−96.
(6) Smith, C. W. J.; Valcárcel, J. *Trends Biochem. Sci.* **2000**, *25*, 381−8.
(7) Lyngso, R. B.; Pedersen, C. N. *J. Comput. Biol.* **2000**, *7*, 409−27.
(8) Garey, M. R.; Johnson, D. S. *Computers and Intractability: A Guide to the Theory of NP-completeness*; Freeman: San Francisco, 1979. Gaspin, C.; Westhof, E. In *Advances in Molecular Bioinformatics*; Schulze-Kremer, S., Ed.; IOS Press: Amsterdam, 1994; p 103.
(9) Hughes, T. R.; Shoemaker, D. D. *Curr. Opin. Chem. Biol*. **2001**, *5*, 21−5.
(10) Schena, M.; Shalon, D.; Davis, R. W.; Brown, P. O. *Science* **1995**, *270*, 467−70.
(11) Shoemaker, D. D.; Schadt, E. E.; Armour, C. D.; He, Y. D.; Garrett-Engele, P.; McDonagh, P. D.; Loerch, P. M.; Leonardson, A.; Lum, P. Y.; Cavet, G.; Wu, L. F.; Altschuler, S. J.; Edwards, S.; King, J.; Tsang, J. S.; Schimmack, G.; Schelter, J. M.; Koch, J.; Ziman, M.; Marton, M. J.; Li, B.; Cundiff, P.; Ward, T.; Castle, J.; Krolewski, M.; Meyer, M. R.; Mao, M.; Burchard, J.; Kidd, M. J.; Dai, H.; Phillips, J. W.; Linsley, P. S.; Stoughton, R.; Scherer, S.; Boguski, M. S. *Nature* **2001**, *409*, 922−7.
(12) Shumaker, J. M.; Tollett, J. J.; Filbin, K. J.; Montague-Smith, M. P.; Pirrung, M. C. *Bioorg. Med. Chem*. **2001**, *9*, 2269.
(13) Pirrung, M. C.; Davis, J. D.; Labriola, J. P.; Montague-Smith, M. P.; Weislo, L. J. *Bioorg. Med. Chem. Lett*. **2001**, *11*, 2437.
(14) Pirrung, M. C.; Odenbaugh, A. L.; Davis, J. D. *Langmuir* **2000**, *16*, 2185.
(15) Grünthert, U. *Curr. Top. Microbiol. Immunol.* **1993**, *184*, 47−63.
(16) Similar observations have been made in microarray hybridization of cRNA: Chee, M.; Yang, R.; Hubbell, E.; Berno, A.; Huang, X. C.; Stern, D.; Winkler, J.; Lockhart, D. J.; Morris, M. S.; Fodor, S. P. *Science* **1996**, *274*, 610−4.
(17) Philips, A. V.; Cooper, T. A. *Cell. Mol. Life Sci*. **2000**, *57*, 235−49. Cooper, T. A.; Mattox, W. *Am. J. Hum. Genet*. **1997**, *61*, 259−66.
(18) Sneath, R. J.; Mangham, D. C. *Mol. Pathol*. **1998**, *51*, 191−200. Zoller, M. *J. Mol. Med*. **1995**, *73*, 425−38. Gunthert, U.; Stauder, R.; Mayer, B.; Terpe, H. J.; Finke, L.; Friedrichs, K. *Cancer Surv*. **1995**, *24*, 19−42.
(19) Ponta, H.; Sleeman, J.; Dall, P.; Moll, J.; Sherman, L.; Herrlich, P. *Invasion Metastasis* **1994**−**95**, *14*, 82−6.
(20) Herrera-Gayol, A.; Jothy, S. *Exp. Mol. Pathol.* **1999**, *66*, 149−56.
(21) Cooper, D. L. *J. Pathol*. **1995**, *177*, 1−3.
(22) Herrlich, P.; Pals, S.; Ponta, H. *Eur. J. Cancer* **1995**, *31A*, 1110−2.
(23) Graveley, B. R. *RNA* **2000** *6*, 1197−211.
(24) Woodman, A. C.; Goodison, S.; Drake, M.; Noble, J.; Tarin, D. *Clin. Cancer Res*. **2000**, *6*, 2381−92.
(25) Goodison, S.; Yoshida, K.; Sugino, T.; Woodman, A.; Gorham, H.; Bolodeoku, J.; Kaufmann, M.; Tarin, D. *Cancer Res*. **1997**, *57*, 3140−4.
(26) van Weering, D. H.; Baas, P. D.; Bos, J. L. *PCR Methods Appl*. **1993**, *3*, 100−6.
(27) The normally spliced form gives no signal with the variant exon-specific microarray. The "-" indicates a mixture of forms.
(28) Additional, weaker signals were also observed. RNA APEX shows extremely low nonspecific signals, suggesting that even these weak signals should be regarded as positives, and that further CD44 splice forms may exist in this tissue (e.g., a v9−16 junction is possible).
(29) Stamm, S.; Zhu, J.; Nakai, K.; Stoilov, P.; Stoss, O.; Zhang, M. Q. *DNA Cell Biol*. **2000**, *19*, 739−56.

JA012102A