Electrophoretic Signal Comparison Applied to mRNA Differential Display Analysis

ABSTRACT

Gene expression analysis by electrophoretic methods is currently limited by the labor-intensive visual evaluation of the electrophoretic signal profiles. For this purpose, we present a flexible approach to computer-assisted comparison of quantitative electrophoretic patterns between multiple expression signals. Gaussian curves are first fitted to the complex peak mixtures, and the resulting approximate signals are then aligned and compared on a peak-by-peak basis with respect to specific patterns defined by the investigator. The rationale of the method is to produce a compressed list of exceptional expression patterns quantified by a set of associated numeric features. A score value is attached to each pattern in such a way that large values identify the most potential findings to be focused on in visual analysis instead of the vast amount of original electrophoretic results. The validity of the method is demonstrated by analyzing a large set of electrophoretic data from mRNA differential display experiments monitoring changes in gene expression patterns in human colonic carcinoma. The automated identification of variously defined gene expression patterns agrees well with the visual evaluation of the same electropherograms. The general comparison approach may also be found useful with other gene expression profiling instruments.

INTRODUCTION

Recent advances in electrophoretic methods and laboratory automation have provided rapid means for separating and analyzing large numbers of components from complex mixtures of macromolecules. Applications include gene expression profiling of healthy and diseased tissues, both at the mRNA and protein level. However, utilization of the full potential of quantitative electrophoresis requires accurate and efficient analysis of electropherograms by computerized methods. While substantial effort has been devoted over the past two decades to the development of computer-assisted analysis of 2-D electrophoretic images (18), software packages available for 1-D electrophoretic signals still lack the versatility to allow advanced analysis.

The carcinoma of the large intestine is a common cancer of increasing incidence. The molecular pathology of colonic carcinoma has been studied extensively (23). However, it is obvious that several genes involved in the pathogenesis of this highly malignant neoplasm are still unknown. Several methods are available to identify cancer-related changes in gene expression, including cDNA microarrays and mRNA differential display (DD) (9). In mRNA DD, differences in gene expression between two or more samples are identified by comparing the relative amount of electrophoretically separated cDNA fragments produced in an arbitrarily primed PCR (10). In contrast to microarrays, the information obtained with DD-PCR is not restricted to a chosen set of genes. In fact, the gene fragments recovered in DD can be conveniently used to produce relevant probes to be spotted on cDNA microarrays (3,24). Several genes associated to colonic carcinoma have been identified as a result of DD analysis (13,14,26).

The significance of bioinformatics has recently increased rapidly with the use of high-throughput methods relying on powerful data analysis. Modern automated multicapillary electrophoretic devices rapidly produce large amounts of 1-D electrophoretic data. Several steps in the analysis of this data can be automated. There are a plenty of tools available for the detection of fragments in electrophoretic signal profiles, and the resulting numeric peak data can easily be retrieved and exported as tabular data in a standard format for subsequent computational analysis. However, more detailed analysis by means of peak-based comparison between multiple signals is yet less developed. Consequently, the comparison of electrophoretic patterns currently relies on subjective visual examination of autoradiograms or electropherograms, which is a time- and labor-intensive step in the analysis that is also prone to errors (11,12).

We hypothesized that numeric electrophoretic data can be used to quantify specific changes in gene expression profiles. Accordingly, we have previously shown that electrophoretic signals enable statistical analysis of similarity of expression patterns (1). We then developed an automated method for aligning the peaks of electrophoretic signals and locating peak pairs showing dissimilar intensity in otherwise similar samples (2). The aim of the present study is to elucidate the usefulness of this automatic approach by quantifying variously defined gene expression patterns in human colonic carcinoma, as analyzed...
by DD-PCR. It is presumed that such a computerized preprocessing of the patterns gives benefit in the form of excluding insignificant results from the visual confirmation step, hence saving a substantial amount of human labor. We demonstrate the performance of the method using specific score functions for characterizing differences between normal and diseased tissue samples on three levels of increasing complexity.

**MATERIALS AND METHODS**

**Patients and Tissues**

Samples were obtained from resected large intestinal tissues of four patients operated on for colorectal adenocarcinoma as described previously (16). The samples subjected to mRNA DD were normal colonic mucosa (designated as n1 in the text) and carcinoma (c1) from patient 1 (male, 51 years of age), normal colonic mucosa (n2) and carcinoma (c2) from patient 2 (female, 49 years), carcinoma (c3) from patient 3 (female, 75 years), and carcinoma (c4) from patient 4 (male, 58 years).

**DD-PCR Analysis**

Total RNA was isolated as described previously (4). The six RNA samples were each reverse transcribed to 15 different cDNA pools by the use of the three possible combinations of one base anchored and the 12 possible combinations of two base anchored cDNA primers (15). The DD-PCR using a common secondary T7 sequence tagged Cy5-labeled primer (Cy5-T7: 5′-Cy5-GTAATACGACTCACTATAGGG-3′; APBiotech, London, UK) as a downstream primer and a set of gene specific primers as upstream primers was done as described earlier (16).

**Electrophoretic Analysis**

The fluorescently labeled PCR products were analyzed by electrophoresis through a 0.5 mm thick denaturing 4% T/3% C polyacrylamide gel with a separation distance of 20 cm in an automated sequencer (ALFexpress II™; APBiotech). An internal Cy5-labeled size standard series was added to the electrophoretic samples to aid in size determination. In the molecular weight determination, the primer peak and the internal and external size standards were utilized (15).

**Signal Analysis**

Automated scanning of expression patterns within the given set of electrophoretic signals was based on the following three steps: (i) identification of valid peaks of each signal, (ii) matching of signal pairs with respect to a given distance function on peaks, and (iii) computation of a score function characterizing the rank of observed patterns. The first step can be done using a variety of analysis software packages. In the present study, we used the ALFwin™ Fragment Analyser (APBiotech), which quantifies each peak (fragment) by its location (size s in bp) and intensity (area A and height h in arbitrary units). The matching step was done by finding the optimum alignment across the peaks of the signals (2,25). Briefly, a global alignment of a signal pair \((u, v)\) is obtained by inserting zero peaks (with \(A=0\)) either into the middle or to the ends of the peak sequences \((u_k)_{k=1}^{n_u}\) and \((v_k)_{k=1}^{n_v}\), representing the signals, where the peaks are originally numbered with respect to their sizes. The resulting sequence of the aligned peak pairs with length \(L_{u,v}\) is denoted by \((u_k, v_k)_{k=1}^{L_{u,v}}\).

The distance \(D(u,v)\) between the signals \(u\) and \(v\) is given by

\[
D(u,v) = \min \sum_{k=1}^{L_{u,v}} d(u_k, v_k) \quad \text{[Eq. 1]}
\]

where the minimum is computed over the finite set of global alignments \((u_k, v_k)_{k=1}^{L_{u,v}}\), satisfying \(n_u + n_v \leq L_{u,v}\). The distance \(d(u_k, v_k)\) between the aligned peak pairs was specified by modeling the electrophoretic signals as Gaussian mixtures, where each peak \(u_k\) at size \(t\) inside the given size range is approximated by

\[
u_k(t) = h_k^u \exp \left( -\frac{(t-s_k^u)^2}{4A_k^u} \right) \quad \text{[Eq. 2]}
\]

By virtue of the approximation we can derive an analytical formula for the (Matusita) distance \(d(u_k, v_k) = \left[ \frac{1}{n} \sum_{i=1}^{n} (u_i(t) - v_i(t))^2 \right]^{1/2}\) and an efficient dynamic pro-
programming algorithm with the time and space complexity of the order of $n \times n$ for determining the optimum pairwise alignment of the electrophoretic signals (2).

Score Functions

Given the optimum alignment and the corresponding distance (Equation 1), the third step involves the specification and computation of a score function that characterizes the objects of the signal comparison in question. In this study, we considered three score functions operating at different levels of the signal analysis: comparison of signal pairs, peak pairs, and multiple peak pairs. Each score function is defined so that the larger the score value, the more significant was the object under comparison on that level. The objects are finally ranked according to decreasing score values, and the visual evaluation by the investigator will be restricted to the subset of objects with the highest ranks. Because the signals are introduced in the present study in groups of size six, we refer to the objects using the signal names $n_1$ and $n_2$ (normal samples) and $c_1$ to $c_4$ (diseased samples). Note that we use this fixed layout of data to keep the discussion of the paper concrete. However, the method generalizes immediately to other layouts as well.

At the first comparison level, the score function depends directly on the distances between the aligned signal pairs ($n_1, c_1$) and ($n_2, c_2$) and indirectly on the average pairwise distances of all the signals in the same group ($n_1$ to $c_4$). Formally, $S_1$ assesses the significance of objects of all six signals as follows:

$$S_1(n_1, n_2, \ldots, c_4) = \frac{D(n_1, c_1)D(n_2, c_2)}{[(D(n_1, n_2) + D(n_1, c_1) + \ldots + D(c_3, c_4))/15]^2}$$

[Eq. 3]

The numerator characterizes the within-subject dissimilarities between the normal and diseased samples, whereas the denominator characterizes the average dissimilarity between the 15 signal pairs among the group. Both distances, $D(n_1, c_1)$ and $D(n_2, c_2)$, must be relatively large to obtain exceptional $S_1$ values. The inter-subject correlation in the numerator accounts only for the variation between the first two subjects because

Table 1. Pairwise Distances between the Signals Illustrated in Figure 1

<table>
<thead>
<tr>
<th>Signal</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>$c_3$</th>
<th>$c_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1$</td>
<td>0.00</td>
<td>64.13</td>
<td>127.03</td>
<td>133.24</td>
<td>122.46</td>
<td>124.42</td>
</tr>
<tr>
<td>$n_2$</td>
<td>0.00</td>
<td>97.43</td>
<td>123.15</td>
<td>117.41</td>
<td>106.94</td>
<td></td>
</tr>
<tr>
<td>$c_1$</td>
<td>0.00</td>
<td>13.72</td>
<td>12.75</td>
<td>7.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c_2$</td>
<td>0.00</td>
<td>3.54</td>
<td>10.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c_3$</td>
<td>0.00</td>
<td></td>
<td></td>
<td>7.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c_4$</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The overall distance $D(u, v)$ is computed using Equation 1, where $D(u, v) = D(v, u)$. The normal samples $n_1$ and $n_2$ differ markedly from the diseased samples $c_1$ to $c_4$ according to the distance values, whereas the diseased samples are relatively similar to each other.

![Figure 1. Expression profiles of the group with the largest 1st-level score. The score $S_1(n_1, n_2, \ldots, c_4) = 3.06$ was evaluated using Equation 3. Colonel carcinoma manifests as large within-subject distances $D(n_1, c_1) = 127.03$ and $D(n_2, c_2) = 123.15$, whereas the average pairwise distance within the group is only 71.46. The scales are the same among the profiles of each figure.](image)

![Figure 2. Expression profiles of the group with the largest second-level score. The extreme score $S_2(n_1, c_1) = 23.75$ was evaluated using Equation 4. Only the interesting part (100–180 bp) of the signal is shown. The overall distance between the signal pair is $D(n_1, c_1) = 1148.28$ with $L_{n_1, c_1} = 34$ aligned peak pairs. Notice the high signal intensities among this group.](image)
the lack of normal samples for the other subjects in the current experiment.

The second level score function characterizes the extent of which each individual aligned peak pair \((u_k, v_k)\) contributes to the overall signal distance \(D(u, v)\). The score function \(S_2\) is given by

\[
S_2(u_k, v_k) = \frac{d(u_k, v_k)}{D(u, v)} \quad \text{[Eq. 4]}
\]

where the objects under comparison are the aligned peak pairs \((u_k, v_k)\) originating from \((n_1, c_1)\) or \((n_2, c_2)\). Because the score function is designated to identify the differentially expressed peaks, \(S_2\) value is proportional to the individual peak pair distance \(d(u_k, v_k)\), whereas the requirement of similar background is considered by normalizing the distance by the average difference of all aligned peak pairs. Normally, one could scan each peak pair in all signal pairs, but in this study we concentrate only on the signal pairs \((n_1, c_1)\) and \((n_2, c_2)\) that may reveal within-subject variation with respect to the presence or absence of cancer cells in the tissue sample in question.

At the third level, inter-subject variation is characterized as well. The objects represent the aligned peak pairs of the four signals \((n_1 \text{ to } c_2)\) whose sizes are within a predefined range. Function \(S_3\) is defined as

\[
S_3(n_{1k}, n_{2l}, c_{1k}, c_{2l}) = \text{sgn}(A_{1k}^{n_1} - A_{1k}^{c_1}) S_2(n_{1k}, c_{1k}) \text{sgn}(A_{2l}^{n_2} - A_{2l}^{c_2}) S_2(n_{2l}, c_{2l}) \quad \text{[Eq. 5]}
\]

for the aligned peak pairs \(k\) and \(l\) satisfying the inequality \(\max\{s_{1k}^{n_1}, s_{1k}^{n_2}, s_{1k}^{c_1}, s_{1k}^{c_2}\} - \min\{s_{1k}^{n_1}, s_{1k}^{n_2}, s_{1k}^{c_1}, s_{1k}^{c_2}\} < T\). The pre-specified threshold \(T\) fixes the maximal size difference of the four aligned peaks within the same group. Function \(S_3\) obtains large values when both peak pairs \(k\) and \(l\) differ significantly with respect to score function \(S_2\) and the direction of the differences is the same, that is, the presence of colonic carcinoma is associated in both subjects either in up- or down-regulated fragment intensities.

RESULTS

The electrophoretic analysis of the normal and carcinoma samples produced altogether 1224 signals to be compared. The signal-level comparison with respect to the score function \(S_1\) (Equation 3) ranks the 204 groups each consisting of the six signals \((n_1, n_2, c_1, c_2, c_3, \text{ and } c_4)\) according to the decreasing score values. Figure 1 shows the smooth approximation of the expression signals by Gaussian functions (Equation 2) for the group with the maximum

![Figure 3. Expression profiles of the group with the largest third-level score.](image)

The score \(S_3(n_{124}, n_{224}, c_{121}, c_{222}) = 284.31\) was evaluated using Equation 5. The corresponding \(S_2\)-ranks were 35 for \((n_{124}, c_{121})\) and 5 for \((n_{224}, c_{222})\). The largest peak height is \(h_{12}^{c_1}\). The minor down-regulation in \((n_{115}, n_{215}, c_{10}, c_{213})\) is just outside the 5% threshold.
score value. Table 1 gives the pairwise signal distances within this particular group. As expected, the group is represented by similar signals in general but with large variation at certain sizes between the normal and diseased samples of the first two subjects (Figure 1).

The search for differentially expressed peaks using the function $S_2$ (Equation 4) results in a ranked list of 7128 peak pairs matched between the signal pairs $(n_1, c_1)$ and $(n_2, c_2)$. The group containing the peak pair with the highest $S_2$-rank is shown in Figure 2. A large peak $n_{14}$ is seen in the normal sample, whereas the corresponding peak $c_{12}$ in the diseased sample is significantly smaller (Table 2). Since $n_1$ and $c_1$ are very similar in other respects, the normalized score $S_2$ of $(n_{14}, c_{12})$ obtains an extreme value. A particular group may include several exceptional findings [e.g., the peak pair $(n_{29}, c_{21})$ with $S_2$-rank 68 belongs to the top 1% of all the matched peak pairs in the data (Figure 2)].

The score function $S_3$ (Equation 5) identifies the aligned peak four-tuples in signal pairs $(n_1, c_1)$ and $(n_2, c_2)$ such that the sizes of the peaks are within $T=7$ bp from each other. A total of 4206 such objects were automatical-

### Table 2. Partial Alignment of One Signal Pair Illustrated in Figure 2

<table>
<thead>
<tr>
<th>Pair</th>
<th>Signal $n_1$</th>
<th>Signal $c_1$</th>
<th>Distance $d(n_{1k}, c_{1k})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>Index</td>
<td>Size</td>
<td>Area</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>105</td>
<td>83</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>109</td>
<td>88</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>113</td>
<td>90</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>131</td>
<td>1315</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>147</td>
<td>116</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>169</td>
<td>63</td>
</tr>
</tbody>
</table>

The peak quantities $(s, A, h)$ of the signals $n_1$ and $c_1$ were produced by the ALFwin Fragment Analyser software. The detected peaks are indexed according to increasing sizes, and the artificial zero peaks have index 0 and sizes marked by –.

**Figure 4. Differentially expressed genes in the colonic carcinoma samples.** Automatically discovered objects among the data expressing (A) up-regulation of phosphoserine aminotransferase detected as a pattern $(n_{15}, n_{215}, c_{117}, c_{215})$ and (B) down-regulation of α-tropomyosin identifiable from the peak pair $(n_{232}, c_{231})$.

**Figure 5. Evaluation of the second-level score function in an example sample.** (A) Artificially modified intensity of the transcriptional regulator gene (histone H3.3), multiplied by factors from 2 to 10 (dotted curves). (B) Score function $S_3$-value as a function of the multiplication factor for the normal-carcinoma pair $(n_{117}, c_{114})$ (filled squares), normal-normal pair $(n_{117}, n_{217})$ (open squares), and for a reference peak pair $(n_{28}, c_{26})$ (dotted line).
ly detected and ranked. Figure 3 represents the group with the maximal $S_3$-value resulting from the pairs $(n1_{24}, c1_{21})$ and $(n2_{24}, c2_{22})$ at approximately 290 bp. Because the pairs show systematic up-regulation, the $S_3$-value of the object $(n1_{24}, n2_{24}, c1_{21}, c2_{22})$ is exceptional. In the same group, another expression pattern $(n1_{14}, n2_{14}, c1_{12}, c2_{12})$ shows up-regulation near 110 bp with rank 15; hence, 0.4% of the $S_3$-ordered list need to be scanned for finding this pattern (Figure 3).

DISCUSSION

We have briefly demonstrated the power of the comparison approach by illustrating the signal groups with the best ranking expression patterns at the three levels of increasing particularity. In addition to the small-scale graphical results presented above, we discovered a number of visually confirmed up- or down-regulated expression patterns in the current data by using the score functions given in Equations 3–5, fixed for the current experiment. Naturally, score functions can have any form appropriate for the experiment and for the desired patterns queried from the electrophoretic peak data. As the detection and quantification of the peaks are an important prerequisite for efficient signal alignment and comparison, the proposed method is a natural adjunct to the spectrum of analysis tools already developed for 1-D electrophoretic peak data (6,19,20,22). Subsequently, the identified expression patterns can be applied to the classification of electrophoretic signals by pattern recognition methods such as neural networks (7). One should notice that we are not building an automated expert system for classification and diagnostic prediction of cancers (5,8,17), but our method was developed to facilitate the selection of electrophoretic patterns showing the greatest interest.

Automated methods, in general, can provide only a set of potential findings for confirmation and further study. The true significance of the candidate patterns must be confirmed visually, and the biological significance of the results must be verified by subsequent testing. In mRNA DD experiments, the cDNA fragments resulting from differentially expressed mRNAs are purified, and the genes are identified by nucleic acid sequencing. In the current experiment, we studied the expression patterns of certain genes purified by a fluorescent PAGE-based method and identified by cyclic sequencing (16). As an example, Figure 4A shows up-regulation of the phosphoserine aminotransferase gene at approximately 140 bp, which was automatically detected as the 19th largest $S_2$-value among 7128 objects found. The change is also systematic, as judged by the joint rank of the four-tuple $(n1_{15}, n2_{15}, c1_{17}, c2_{15})$, which obtains the 26th largest score $S_3$ of all the 4207 four-tuples found in the data. As another example, down-regulation of the α-tropomyosin gene in colonic tumor tissue at 475 bp was detected automatically by the seventh largest $S_2$-value (Figure 4B). Notice the inconsistency in the regulation pattern between the two subjects, resulting in the 211th lowest $S_3$-rank. Whether this is a methodological issue or whether it reflects the true inter-subject bias in relation to the colonic carcinoma remains to be studied.

Figure 5A illustrates a signal group with the 24th largest $S_1$-value, where the pair $(n1, c1)$ shows several differences and the pair $(n1, n2)$ is almost identical. To study the sensitivity and specificity of the methodology within this group, we artificially varied the measured intensity of a control gene, histone H3.3 at approximately 290 bp, and reevaluated the score function $S_2$ after each repetition for all six signal pairs. Initially, the pair $(n2_{38}, c2_{60})$ obtains the largest value among the resulting 75 peak pairs matched (Figure 5B, dotted line). Figure 5B shows that the observed down-regulation of the control gene in the normal-carcinoma pair yields positive detection of equal score not until at factor 9 (filled squares), while the normal-normal pair of the same gene gives presumably false-positive detection already at factor 4 (open squares). This demonstrates the fact that distortion in expression levels between two similar signals may decrease the specificity of the detection significantly. On the other hand, if the signals being compared show variation at several locations, then sensitivity may be decreased because of the inability to detect all the differences. Multiple dif-
ferentially expressed peaks could be identified simultaneously by replacing the denominator of Equation 4 with a trimmed mean of the distances (2). However, this modification conflicts with the basic idea of mRNA DD analysis, where rare differences are searched between otherwise similar signals.

Although we observe that a score function obtains an extreme value, we cannot immediately conclude that there is a genuine differentially expressed pattern in the electropherogram. The investigator must eventually make the decision between a true change and the extreme value being an artifact, but automated identification of candidate patterns greatly contributes to the prescreening process. Disturbances of signals, both in amplitude and in mobility, were identified as the main cause for inadequate comparison efficiency. Whereas the differences in the mobility between the signals of a group can be corrected in the sizing step using internal molecular standards (15) or unsupervised methods (21), substantial variations are expected to remain in signal intensities, even when the same mRNA DD experiment is repeated. This is because the method is semiquantitative by nature, and uncontrollable mispriming events may occur (1,7). In the absence of an efficient correction scheme for the intensities, we currently leave the electrophoretic data unscaled in amplitude.

If several repeated measurements were available, then the sensitivity and reproducibility of the method could be investigated, and the statistical significance of the observed differences could be assessed. However, even with a limited number of repetitions of electrophoretic experiment, distinguishable changes in the gene expression profiles can be prescreened automatically using the proposed signal comparison method.

REFERENCES


Received 12 June 2002; accepted 17 October 2002.

Address correspondence to:
Dr. Tero Aittokallio
Turku Centre for Computer Science
Lemminkäisenkatu 14A
FIN-20520 Turku, Finland
e-mail: tero.aittokallio@utu.fi