Fractal Characteristics of Mass Spectrometry based Cancer Data

TUAN D. PHAM
James Cook University
Bioinformatics Applications Research Centre
Information Technology Discipline
School of Mathematics, Physics, and Information Technology
Townsville, QLD 4811
AUSTRALIA
tuan.pham@jcu.edu.au

Abstract: This paper addresses the fractal analysis of mass spectrometry data for the prediction of complex diseases. We studied ovarian and prostate cancers as examples of the analysis. Experimental results show that the fractal dimensions of cancer states distinctively tend to have higher values than those of the control states. High values of the Hurst exponent of the mass spectrometry data under study suggest the persistent behavior of the datasets and the reliability of the fractal dimensions.

Key–Words: Fractal analysis, mass spectrometry data, cancer prediction.

1 Introduction

Advances in proteomics technology offer tremendous opportunities for far more better understanding of the biology of many complex human diseases at molecular levels, which will hopefully lead to the early detection and treatment of such diseases – by early recognition of symptoms, one can get the most effective clinical treatment for the best outcome.

Proteomics is defined as the study of proteome which refers to the entire set of expressed protein in a cell. Major research areas of proteomics include structural, functional, and interaction studies [1]. Structural proteomics uses X-ray crystallography, nuclear magnetic resonance, or even both to study the final three dimensional shape of proteins. Functional proteomics involves the use of mass spectrometry (MS) to study the regulation, timing, and location of protein expression. Interaction studies seek to understand how protein pair between themselves and other cellular components interact to constitute to more complex models of the molecular machines.

In particular, protein expression profiles or expression proteomics can be used for large-scale protein characterization or differential expression analysis that has many applications such as disease classification and prediction, new drug treatment and development, virulence factors, and polymorphisms for genetic mapping, and species determinants [2, 3]. In comparison with transcriptional profiling in functional genomics, proteomics has some obvious advantages in that it provides a more direct approach to studying cellular functions because most gene functions are characterized by proteins.

The identities of expressed proteins in a proteome can be determined by protein separation, identification, and quantification. One of many approaches for separating proteins involves two-dimensional gel electrophoresis followed by gel image processing. Once proteins are separated, protein differential expression can be characterized using mass spectrometry, which is a high-resolution technique for determining molecular masses and provides rapid and accurate measurement of protein profiling in complex biological and chemical mixture. Protein profiling of plasma and serum can be prepared with a matrix-assisted laser desorption ionization (MALDI) ion source or the surface-enhanced laser desorption ionization (SELDI) ion source coupled to a time-of-flight (TOF) mass analyzer with a chevron microchannel plate detector.

In regards to recent applications of proteomic technology, proteomic patterns have recently been utilized for early detection of cancer progressions [4, 5, 6]. Furthermore, methods for classification of normal and cancerous states using mass spectrometry data have been recently developed. Petricoin et al. [5] applied cluster analysis and genetic algorithms to detect early stage ovarian cancer using proteomic spectra. Ball et al. [7] applied integrated approach based on neural networks to study SELDI-MS data for classification of human tumors and identification of biomarkers. Lilien et al. [8] applied principal component analysis and a linear discriminant function


Although fractal approach has been applied for the analysis of complex signals including imaging, environmental, and time-series data [15], it has rarely been used for the study of mass spectrometry data. In fact, fractal modeling can be a useful tool for explaining the dynamic behavior of time series in terms of scale invariant long-term correlations. As a departure from other existing methods mentioned above and a precursor to the robust classification of diseases using mass spectrometry data, we present in this paper a pioneering application of the fractal analysis of this type of proteomic data. A distinct view of this study is that the values of fractal dimensions can be used as an indicator for expressing the risk of a patient having the disease and a potential feature for early prediction of fatal diseases using mass spectrometry data. The rest of this paper will begin by reviewing the two essential concepts of fractals known as the fractal dimensions and the Hurst exponent. We next illustrate the fractal analysis by studying two popular public mass spectrometry datasets of ovarian and prostate cancers.

2 Measure of Self-Similarity by Fractal Dimensions

The fractal dimension is a mathematical expression of the space filling properties of an object whose concept leads to many different types of fractal dimensions [16]. The simplest form of fractal dimensions is the self-similarity dimension. The self-similarities of the line, square, and cube are equal 1, 2, and respectively. Consider a geometrically self-similar fractal object which consists of line segments. If each line segment is divided into \( M \) smaller line segments, then \( N \) smaller objects are produced. Furthermore, if the object is geometrically self-similar, then each of the objects of smaller sizes is an exact but reduced size copy of the whole object. The self-similarity dimension \( d \) is then expressed as [16]

\[
N = M^d
\]

which can be written in another form as

\[
d = \frac{\log(N)}{\log(M)}
\]

Because the self-similarity dimension requires that each smaller subject formed by the division of the whole object must be an exact copy of the whole object, it can only be used to study objects that are geometrically self-similar. Such a fractal dimension is not very useful for analyzing many real objects that usually have irregular shapes. Thus more general fractal dimensions have been developed as more general forms of the fractal dimension. Two such popular forms are known as the capacity and the Hausdorff-Besicovitch dimensions.

The capacity of an object can be determined by covering it with balls of a radius \( r \). The smallest number of balls \( N(r) \) that covers all the parts of the object is counted. Then the radius of the previous balls is reduced and again \( N(r) \) is counted. The capacity is the value of \( \log N(r)/\log(1/r) \) in the limit as \( r \) shrinks to 0:

\[
d = \lim_{r \to 0} \frac{\log N(r)}{\log(1/r)}
\]

The relationship of the capacity and the self-similarity dimensions is that if \( M = 1/r \), then \( N = M^d \).

The Hausdorff-Besicovitch dimension is similar, but not identical, to the capacity dimension. In the capacity dimension, the object is covered with the number of balls \( N(r) \) of a given radius \( r \). In the Hausdorff-Besicovitch dimension, the object is covered with sets.

The capacity dimension \( d \) expression in (3), in the limit where \( r \) approaches 0, can be determined using the box counting method [16] that works as follows. An object is covered with a grid that is divided into a number of boxes. The number of boxes of the grid containing at least some part of the object is then counted. The counting procedures are repeated where each time the current box size is reduced by half the size of the previous boxes. The capacity dimension is then determined as the slope of the plot of \( \log N(r) \) versus \( \log(1/r) \), or alternatively, the negative of the slope of the plot of \( \log N(r) \) versus \( \log(r) \). The basic idea is that if an object is self-similar, then the slope of \( \log N(r) \) versus \( \log(1/r) \) is the same as the limit of \( \log N(r)/\log(1/r) \) as \( r \) approaches 0. However, the determination of the slope is much easier than the limit.
3 Measure of Roughness by Hurst Exponent

The Hurst exponent is a measure of roughness of self-affine curves [17]. This exponent is usually denoted as $H$. The proper range for the exponent is from 0 corresponding to very rough random fractal series, to 1 corresponding to rather smooth-looking random fractals. A useful information of the Hurst exponent for the study of time series is that when $H > 0.5$, the series is considered to be persistent and has low level of noise or even shows periodicities; if $H < 0.5$, the series is considered anti-persistent, has high level of noise and high complexity; whereas $H = 0.5$ indicates a statistical independence with finite variance [18].

A method for determining $H$ is by the use of the power spectrum method. If the spectral density $P(k) = k^{-B}$ then [19]

$$H = \frac{B - 1}{2}$$  \hspace{1cm} (4)

Another method for computing $H$ is called the width method [20]. Consider a with $w(l)$ of a single valued function $y(x)$, which is a self-affine fractal with dimension $d$, on the scale $l$ defined as

$$w(l) = \sqrt{\langle y^2(x) \rangle_l - \langle y(x) \rangle_l^2}$$  \hspace{1cm} (5)

can be scaled as

$$w(l) = l^H$$  \hspace{1cm} (6)

where $\langle \cdot \rangle$ denoted an average of all sections with length $l$.

The Hurst exponent $H$ can be determined by fitting a straight line to the $w(l)$ versus $l$ data in a log-log plot. Using the limit formula for the box-counting dimension and some calculus, the fractal dimension can be determined as [17]

$$d = \lim_{k \to \infty} \frac{\log[(2^{2-H})^k N]}{\log(2^k/r)} = 2 - H$$  \hspace{1cm} (7)

4 Fractal Analysis of Ovarian and Prostate Cancer Data

A fractal is defined an object in space or a process in time that has a fractal dimension that is greater than its topological dimension. According to definition of fractals, mass spectrometry data, that is a function of relative image intensitives in time ($m/z$), can be considered as a fractal. Thus, the investigation of the fractal characteristics of mass spectrometry data can be helpful for the interpretation of this type of complex biological signals for disease detection at the molecular level.

In this study, fractal analysis was performed using two public MS-based cancer datasets: (1) ovarian high-resolution SELDI-TOF mass spectrometry dataset, and (2) prostate cancer mass spectrometry dataset (PC-H4). The ovarian high-resolution SELDI-TOF mass spectrometry dataset and the prostate cancer mass spectrometry dataset (PC-H4) can be obtained from the FDA-NCI Databank (http://home.ccr.cancer.gov/ncifdaproteomics/patterns.asp). The ovarian data consists of 100 control samples and 170 cancer samples. The MS-based prostate cancer dataset consists of 190 samples diagnosed as benign prostate hyperplasia (with PSA levels greater than 4), 63 samples considered as no evidence of disease (with PSA level less than 1), and 26 samples diagnosed as prostate cancer (with PSA levels from 4 to 10), and 43 samples as prostate cancer (with PSA level greater than 10). Figures 1 and 2 show the plots of typical ovarian, and control samples respectively; whereas Figures 3 and 4 show the plots of typical prostate and control samples respectively.

We applied the box counting method to compute the fractal dimensions for all the samples of the ovarian and cancer datasets. We then computed the Hurst exponents to determine the behavior and the levels of noise of the same datasets.

Figures 5-6 show the plots of $N(r)$ versus $(1/r)$ on a log-log scale of ovarian-cancer and control samples respectively. Similarly, Figures 7-9 show the
Table 1: Fractal characteristics of MS-based ovarian cancer data

<table>
<thead>
<tr>
<th></th>
<th>$d_{top}$</th>
<th>$d_{mean}$</th>
<th>$\sigma_r^2$</th>
<th>$H$</th>
<th>$\sigma_H^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>1.312</td>
<td>1.370</td>
<td>0.00008</td>
<td>0.885</td>
<td>0.00002</td>
</tr>
<tr>
<td>Cancer</td>
<td>1.394</td>
<td>1.383</td>
<td>0.00005</td>
<td>0.885</td>
<td>0.00006</td>
</tr>
</tbody>
</table>

Table 2: Fractal characteristics of MS-based prostate cancer data

<table>
<thead>
<tr>
<th></th>
<th>$d_{top}$</th>
<th>$d_{mean}$</th>
<th>$\sigma_r^2$</th>
<th>$H$</th>
<th>$\sigma_H^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>1.226</td>
<td>1.261</td>
<td>0.00007</td>
<td>0.867</td>
<td>0.00002</td>
</tr>
<tr>
<td>Benign (PSA &gt; 4)</td>
<td>1.275</td>
<td>1.278</td>
<td>0.0002</td>
<td>0.891</td>
<td>0.00002</td>
</tr>
<tr>
<td>Cancer (PSA 4-10)</td>
<td>1.263</td>
<td>1.264</td>
<td>0.004</td>
<td>0.894</td>
<td>0.00004</td>
</tr>
<tr>
<td>Cancer (PSA &gt; 10)</td>
<td>1.260</td>
<td>1.262</td>
<td>0.0002</td>
<td>0.880</td>
<td>0.00002</td>
</tr>
</tbody>
</table>
plots of $N(r)$ versus $(1/r)$ on a log-log scale of prostate-cancer, benign, and control samples respectively. The straight lines are the linear regression lines through the data.

To distinguish the differences of the fractal dimensions between cancer, benign, and normal states, we applied the $k$-nearest neighbor, where $k = 1$, to select the top fractal dimension that has the best hits. Tables 1 and 2 show the value of the fractal dimensions and the Hurst components of the best hit samples of the ovarian and the prostate datasets respectively. For the ovarian cancer dataset, the average fractal dimensions of the cancer and control samples are 1.383 and 1.370 respectively; whereas the fractal dimensions of the cancer and control samples that give the best classification rates are 1.394 and 1.312 respectively. In both cases, the fractal dimensions of the cancer samples are higher than those of the control samples.

The Hurst components for all samples are close to one showing the persistency and low level of noise in the mass spectrometry samples and hence confirms the reliability of the fractal dimensions. The fractal analysis of mass spectrometry data for ovarian and prostate cancers in this study suggests the tendency of higher values of fractal dimensions for cancer states while lower values of fractal dimensions for normal states.

5 Conclusion

We have applied the fractal approach for the study of mass spectrometry cancer data which have become a popular means of the proteomics-based approach for early detection of fatal diseases and new drug discovery. The fractal dimensions of the experimental data suggest some distinctive trends between normal and cancer states. The results reported herein show the potential application of fractal dimensions for further study and analysis of the complexity of the mass spectrometry data and can be used as a useful feature for disease classification.

References:


