METHOD OF GENERATING AND REPRESENTING GENE DIFFERENTIAL EXPRESSION PROFILES BY APPLYING FOURIER TRANSFORM

Field of the Invention

[0001] The present invention generally relates to bioinformatics, and more particularly to a method of generating and representing gene differential expression profiles by applying Fourier Transform.

Background of the Invention

[0002] Researches for bioinformatics in the post-genome sequence era have been shifting towards a functional understanding of the roles and relationships between different genes in different conditions. While the advances in genetic expression profiling techniques including microarrays enable detailed and genome-scale measurements, extracting of meaningful information from large datasets is still a challenge.

[0003] The principle of microarrays is to measure gene differential expression levels by direct comparison in a single slide of the fluorescence ratio between experimental mRNA and control mRNA labeled with different fluorescent markers. The commonly used fluorescent markers are Cy3 (green) and Cy5 (red). One of the basic applications of microarrays is the identification of genes that are differentially expressed in compared conditions. The common process for identification of differentially expressed genes includes normalization, and calculation of variation folds between the two fluorescent signals so as to generate gene differential express profiles. Then, the possible candidates of differentially expressed genes are selected from the gene differential expression profile by their large changes of variation folds. If a selection is based on a single slide, the confidence is low because it cannot distinguish real changes from experimental artifacts. Then, the selected candidates of differentially expressed genes have to be verified by repeating the microarray experiment or conducting other types of experiments such as Northern blot. It is apparent that this single slide-based approach has many drawbacks including time-consuming, high cost, and low confidence.

[0004] Therefore, there is an imperative need to develop a method of generating gene differential expression profiles so that the candidates of differentially expressed genes can be selected with confidence. This invention satisfies this need by disclosing a method of generating gene differential expression profile from the gene expression profiles obtained from at least two microarray slides by applying Fourier Transform. Other advantages of this invention will be apparent with reference to the detailed description.

Summary of the Invention

[0005] The present invention provides a method of generating a gene differential expression profile for a pair of conditions (T/N). In one aspect, the method comprises acquiring n sets (S) of gene expression profiles for the pair of conditions, wherein n >1, and each gene with each set of gene expression profile has two expression values, one for T and another for N; wherein each set of gene expression profiles contains m genes; and wherein m >1; arranging the n sets of gene expression profiles in a linear order, i.e., S1(T1N1)S2(T2N2)...Sn-1(Tn-1Nn-1)Sn(TnNn), so the gene expression values from one condition are alternatively interposed by the ones from another condition; and applying the Discrete Fourier Transform to the arranged gene expression values for each gene of the set of gene expression profiles; thereby generating the gene differential expression profile for the pair of condition. In an embodiment, the gene differential expression profile contains at least one gene; thereby candidate genes that are either overexperssed or underexpressed in either of the pair of conditions may be selected from the gene differential expression profile. In a further embodiment, the gene differential expression profile is comprised of overexpressed genes and underexpressed genes. In another embodiment, the pair of conditions may be any paired conditions in which at least one gene is differentially expressed. In another further embodiment, the paired conditions may be any one selected from the group consisting of tumor/normal, tumor x/tumor y, disease/normal, disease x/disease y, development stage x/development stage y, treatment/non-treatment, treatment x/treatment y, and infected/non-infected. In yet another embodiment, when the Discrete Fourier Transform is applied to the arranged gene expression profiles, it gives out values for parameters including Nyquist Frequency component, absolute amplitude, normalized amplitude, and phase.

[0006] In another aspect, the present invention provides a method of representing the gene differential expression profile. In one embodiment, the representing method comprises plotting the normalized amplitude versus absolute amplitude; thereby the gene differential expression profile is represented graphically.

[0007] One advantage of the present method is the elimination of the process of normalization.

[0008] The objectives and advantages of the invention will become apparent from the following detailed description of preferred embodiments thereof in connection with the accompanying drawings.

Brief Description of the Drawings

[0009] Preferred embodiments according to the present invention will now be described with reference to the Figures, in which like reference numerals denote like elements.

[0010] FIG 1 shows an exemplary graph of a single gene expression value in a disease-normal-disease-normal manner with a distinctive high-low feature.

[0011] FIG 2a shows the magnitude of the Fourier Transform of the graph when a 256-point DFT was performed onto the data shown in FIG 1, after removing the mean value from these data points.

[0012] FIG 2b is the plot of the phase for the DFT of data shown in FIG 1, after removing the mean value from these data points.

[0013] FIG 3 shows an exemplary graph of a single gene expression value in a disease-normal-disease-normal manner with no distinctive high-low feature.

[0014] FIG 4 shows the magnitude of the Fourier Transform of the graph when a 256-point DFT was performed onto the data shown in FIG 3, after removing the mean value from these data points.

[0015] FIG 5 shows an exemplary graph of a single gene expression value in a disease-normal-disease-normal manner with a distinctive high-low feature similar to FIG 1, but the gene is expectedly under-expressed in the disease tissue.

[0016] FIG 6a shows the magnitude of the Fourier Transform of the graph when a 256-point DFT was performed onto the data shown in FIG 5, after removing the mean value from the data points.

[0017] FIG 6b is the plot of the phase for the DFT of data shown in FIG 5, after removing the mean value of these data points.

[0018] FIG 7 shows a flowchart of a program for generating a gene differential expression profile in accordance with one embodiment of the present invention.

[0019] FIG 8a shows a plot of magnitude of a 92-point DFT for the 10 genes that are regularly differentially expressed in the 92 slides.

[0020] FIG 8b shows a plot of magnitude of a 92-point DFT for the 10 genes that are not regularly differentially expressed in the 92 slides.

[0021] FIG 9a shows a conventional scatter plot of microarray data where the Red (R) intensity is plot versus the Green (G) intensity.

[0022] FIG 9b shows another conventional scatter plot of microarray data where the Log2(Red intensity) or (Log2R) is plot versus the Log2(Green intensity) or (Log2G).

[0023] FIG 9c shows yet another conventional scatter plot of microarray data where M is plot versus A.

[0024] FIG 10 shows the plot of the Nyquist Frequency Normalised Amplitude versus Nyquist Frequency Absolute Amplitude for all the genes in the GeneChip used for prostate cancer.

Detailed Description of the Invention

[0025] The present invention may be understood more readily by reference to the following detailed description of certain embodiments of the invention.

[0026] Throughout this application, where publications are referenced, the disclosures of these publications are hereby incorporated by reference, in their entireties, into this application in order to more fully describe the state of art to which this invention pertains.

[0027] The present invention provides a method for generating gene differential expression profiles by applying Fourier Transform. As defined herein, the "gene differential expression profile" refers to a database containing at least one gene that is

differentially expressed in a pair of conditions. A pair of conditions refer to any situation in which two conditions affect at least one gene in a different manner. The exemplary pairs of conditions include tumor/normal, tumor x/tumor y, disease/normal, disease x/disease y, development stage x/development stage y, treatment/non-treatment, and treatment x/treatment y. It is generally assumed that when cells are treated with different stimuli or chemicals, or in different stages such as developmental and pathological stages, the gene expression profile of the cells will be different accordingly. Thus, a gene differential expression profile is very useful in identifying the genes that are differentially expressed in certain conditions. For example, the genes that are differentially expressed in tumors will be useful targets for developing diagnostic kits and pharmaceutical drugs. While the pair of tumor and normal tissues will be employed herein to illustrate the principles of the present invention, it is to be appreciated that the principles can be applied to any pair of conditions.

[0028] Prior to the detailed description of generating gene differential expression profile, there is provided a brief description of Sampling Theorem and Discrete Fourier Transform.

[0029] The Sampling Theorem states that for a limited bandwidth (band-limited) signal with maximum frequency *fmax*, the equally spaced sampling frequency *fs* must be greater than twice of the maximum frequency *fmax* (*i.e., fs* > *2·fmax*) in order to have the signal be uniquely reconstructed without aliasing. The frequency *2·fmax* is called the Nyquist sampling rate. Half of this value, *fmax*, is sometimes called the Nyquist frequency. The Sampling Theorem is considered to have been articulated by Nyquist in 1928 and mathematically proven by Shannon in 1949. Some books use the term "Nyquist Sampling Theorem", and others use "Shannon Sampling Theorem". They are in fact the same Sampling Theorem. *See, e.g.,* http://lanoswww.epfl.ch/studinfo/courses/cours_nonlinear_de/extras/Hoffman(1997)_An_Introduction_to_Fourier_Theory.pdf, An Introduction to Fourier Theory by Forrest M. Hoffman, 1997.

[0030] The Fourier transform, in essence, decomposes or separates a waveform or function into sinusoids of different frequency which sum to the original waveform. It identifies or distinguishes the different frequency sinusoids and their respective amplitudes. The Fourier transform of f(x) is defined as

$$F(s) = \int_{-\infty}^{\infty} f(x) \exp(-i\, 2\pi x s)\, dx.$$

[0031] Because a digital computer works only with discrete data, numerical computation of the Fourier transform of f(t) requires discrete sample values of f(t), which is called f_k. In addition, a computer can compute the transform F(s) only at discrete values of s, that is, it can only provide discrete samples of the transform, F_r. If f(kT) and $F(rs_0)$ are the kth and rth samples of f(t) and F(s), respectively, and N_0 is the number of samples in the signal in one period T_0, then

$$f_k = T\, f(kT) = T_0 N_0^{-1}\, f(kT)$$

[0032] and

$$F_r = F(rs_0)$$

[0033] where

$$s_0 = 2\pi\, \mathcal{F}_0 = 2\pi\, T_0^{-1}.$$

[0034] The discrete Fourier transform (DFT) is defined as

$$F_r = \sum_{k=0}^{N_0-1} f_k \exp(-i\, r\Omega_0 k) \qquad \text{where } \Omega_0 = 2\pi\, N_0^{-1}.$$

[0035] The Fast Fourier Transform (FFT) is a DFT algorithm that reduces the number of computations from something on the order of N_0^2 to $N_0 \log N_0$. There are basically two types of Tukey-Cooley FFT algorithms in use: decimation-in-time and decimation-in-frequency. The algorithm is simplified if N_0 is chosen to be a power of 2, but it is not a requirement. See, e.g., http://astronomy.swin.edu.au/~pbourke/analysis/dft/, DFT, (Discrete Fourier Transform), FFT (Fast Fourier Transform), Written by Paul Bourke, June 1993.

[0036] The application of the Discrete Fourier Transform in the microarray technologies has been very limited so far. The Discrete Fourier Transform has been employed to identify the periodically expressed genes based on time course microarray or the cyclically expressed genes in cell-cycle specific genes. See, Luan and Li, *Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data.* BIOINFORMATICS 20: 332–339 (2004); Wichertl et al., *Identifying periodically expressed transcripts in microarray time series data.*

BIOINFORMATICS 20: 5–20 (2004); Zhang and Zhang, *Fourier Harmonic Approach for Visualizing Temporal Patterns of Gene Expression Data.* Murali Ramanathan. The Computational Systems Bioinformatics Conference, CSB2003; and Shedden and Cooper; *Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization.* Proc Natl Acad Sci U S A. 99(7): 4379–4384 (2002). It is obvious that the cell-cycle specific genes would express in a waveform manner so that the Fourier Transform may apply. Thus, the limitation may be due to lack of naturally formed waveforms within the microarray data

[0037] Now there is provided a more detailed description of the present invention in accordance with one aspect of the present invention. The gene differential expression profile is generated by applying the well known Discrete Fourier Transform (DFT) to gene expression profiles. The method disclosed herein is simple, fast and eliminating of the process of normalization.

[0038] When a gene differential expression profile for a pair of conditions between a disease tissue and a normal tissue is to be generated, the gene expression profile of the disease tissue may be represented by T_i and the gene expression profile of the normal tissue by N_i. The intensity of the gene expression could be extracted from microarray data or GeneChip array data that are either publicly or privately available. Then, the gene expression profiles are arranged alternatively between the diseased tissue and normal tissue. For example, the gene expression intensity data from M slides may be arranged as illustrated in Table 1.

[0039] Table 1. An alternative arrangement of the gene expression intensity data of the disease tissue and normal tissue from M for N genes.

[0040]

Genes	T_1	N_1	T_2	N_2	....	T_{M-1}	N_{M-1}	T_M	N_M
Gene1	393.15	280.4	323.75	316.9		327	300.5	255.15	263.9
Gene2	151.4	102.5	108.5	117		112	88.4	108.65	103
Gene3	137.65	103.65	101.75	107.9		113	86.15	102.5	91.15
Gene4	217.65	189.75	213.5	179.65		223.15	199.05	177.4	170.4
Gene5	191.4	135	152.5	143.9		150.65	146.15	146.15	134
Gene6	929.25	122	141.25	197.15		143.75	152.15	171.25	129.5
...									

Gene N-1	141.15	98	98.65	95.15	112.15	82.4	118.75	100.15
Gene N	582.65	210.75	259	387.25	234	285.25	360.65	215.75

[0041] As shown in FIG 1, plotting of a single gene expression value in a disease-normal-disease-normal manner generates a graph with a distinctive high-low feature. Now, the graph becomes a waveform so that the Fourier Transform may be applied to find out the fundamental frequency, phase/angle, absolute amplitude and normalized amplitude of the graph. If all the gene expression values are plotted the same way, then the expression datasets for all the genes can be computed by the Fast Fourier Transform. As discussed hereinafter, prior knowledge of a gene whether it is overexpressed or underexpressed is not a concern for the present invention because the phase of the Fourier Transform will determine whether the gene is overexpressed or underexpressed. For the purpose of illustration, FIG 1 shows if a gene were consistently over-expressed in a disease tissue, the expression intensity (or the value) of the gene in the disease tissue would be higher than that in the paired normal tissue. It is to be noted that it may not be the case for real expression data.

[0042] When a 256-point DFT was performed onto the graph shown in FIG 1, the magnitude of the Fourier Transform of the graph is shown in FIG 2a. It is obvious from FIG2a that if the gene expression is regularly differentially expressed across the slides, there is a significant higher frequency component in the transform. In particular, FIG 2a shows the higher frequency component as the Nyquist Frequency.

[0043] However, if one gene is not consistently differentially expressed in a pair of disease/normal tissues, it is expected that plotting of the gene expression value in a disease-normal-disease-normal manner will not generate a graph with a distinctive high-low feature as shown in FIG 3. Therefore, the graph would not have a single fundamental frequency identifiable from the plot, as shown in FIG 4. To be more specific, the Nyquist Frequency component that is expected has a lower magnitude than other frequency components.

[0044] From these observations, it is apparent that the Nyquist Frequency component gives an excellent indication on the differential feature of gene expressions in a

pair of conditions, i.e., whether a gene is differentially expressed on one of the paired conditions. Therefore, in one embodiment of the present invention, a gene differential expression profile can be generated by calculating the Nyquist Frequency of each gene of gene expression profiles from a pair of conditions by applying the Fourier Transform.

[0045] Now, when a gene is under-expressed in the disease tissue within a pair of disease/normal tissues, the angle/phase of the graph formed by using the same plot manner as shown in FIG 1 will be different from the one of the over-expressed genes. FIG 5 shows an exemplary graph of a single gene expression value in a disease-normal-disease-normal manner with a distinctive high-low feature similar to FIG 1, but the gene is expectedly under-expressed in the disease tissue. When a 256-point DFT was performed onto the graph shown in FIG 5, the magnitude of the Fourier Transform of the graph is shown in FIG 6a. It is obvious from FIG 6a that there is a significant higher frequency component in the transform at the Nyquist Frequency component, indicating the gene is differentially expressed. FIG 2b is the plot of the phase for the DFT of data shown in FIG 1 while FIG 6b is the plot of the phase for the DFT of data shown in FIG 5. It is obvious now that the phase of the Nyquist Frequency in FIG 2b is different from that in the FIG 6b. To be precise, the FIG 2b shows a zero radian (0°) phase shift at the Nyquist Frequency component whereas the FIG 6b shows a $-\pi$ radians (180°) phase shift.

[0046] Then whether a gene is differentially expressed can be determined indirectly by measuring the amplitude of the transform, and whether the gene is over or under expressed can be measured from the phase of the transform; without knowing the ratio or log(ratio). More interestingly, the inventors of the present invention discovered that the "Nyquist Frequency" alone gives excellent results so that the DFT formula is simplified. In one embodiment of the present invention, only one Fourier coefficient $(N/2)^{th}$ is needed, or that of the Nyquist Frequency. Hence the whole computation time is dramatically reduced. For example, for 92 slides of 12,500 genes, the computation time is reduced from 650secs to a mere 4 secs due to the simplification.

[0047] Now referring to FIG 7, there is provided a flowchart of a program for generating a gene differential expression profile in accordance with one embodiment of the present invention. When the program starts **10**, it reads input file/data **20**, wherein the input file/data may be manually input by a user or downloaded from a public or private database. Then the program will arrange the input file/data in such a manner that the two

values from one slide are positioned together, while the values from all slides for one condition is alternatively positioned by the values from all slides for another condition **30**. Then the program removes the mean values across the slides for each gene expression **40.** Then the program computes the DFT on gene expression profile and extracts only the Nyquist Frequency component **50**. Then the program extracts other three parameters: the absolute amplitude, the normalized amplitude and the phase **60**. Then the program sorts gene expression using the normalized amplitude in the descending order **70**. Then the program extracts the top 10-20 genes' expression values for clustering as the most likely candidates for differentially expressed genes in one of the paired conditions **80** before the program comes to an end **90**.

[0048] Examples

[0049] These examples are described solely for purposes of illustration and are not intended to limit the scope of the invention. Although specific terms have been employed herein, such terms are intended in a descriptive sense and not for purposes of limitations.

[0050] A prostate dataset was downloaded from Broad Institute as a benchmark for verification of the present invention. Broad Institute is a research collaboration of the Massachusetts Institute of Technology and Harvard University. The dataset is downloadable from the following URL:http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=75

[0051] The prostate dataset was downloaded from the cited address because their work has been published so that it would serve as a good control for the present invention. See, Singh et al., Gene expression correlates of clinical prostate cancer behavior. Cancer Cell. 2002 Mar;1(2):203-9. In this study, microarray expression analysis was used to identify genes that might anticipate the clinical behavior of this disease. Singh *et al.* concluded that a set of genes was identified that strongly correlated with the state of tumor differentiation as measured by Gleason score. Its results are shown in Table 2.

[0052] Table 2. The results from Singh *et al.*

	Raw Data				Normalized Data			
	Prediction Results			Success Rate	Prediction Results			Success Rate
4 Gene model		TT	TN	**85.7%** (30/35) p = 0.006*		TT	TN	**77.2%** (27/35) p = 0.0005
	PT	26	4		PT	19	0	
	PN	1	4		PN	8	8	
16 Gene model		TT	TN	**82.9%** (29/35) p = 0.047		TT	TN	**85.7%** (30/35) p = 0.0001
	PT	27	6		PT	22	0	
	PN	0	2		PN	5	8	

TT=True Tumor, TN=True Normal, P=Predicted Tumor, PN=Predicted Normal
*All p values calculated by Fisher's Exact Test

[0053] For the demonstration of the method of the present invention, we downloaded the Affymetrix datasets. From this dataset, we found 46 slides of Affymetrix gene expression for prostate cancer tumor tissues and another 46 for normal tissues. The signal values of each slide for the prostate cancer tissue and normal tissue were arranged in the same manner as discussed in FIG 1. Then the arranged expression profiles were subjected to the Fourier Transform of the present invention. We identified top 10 genes with high amplitude on the fundamental frequency that we were looking for. These genes together with their expression intensities, without normalization, were used for classification using Eisen's K-Means Cluster Program. See, Eisen *et al.*, Cluster analysis and display of genome-wide expression patterns. Genetics, Vol. 95: 14863-14868 (1998). We successfully differentiated the tumor slides from the normal slides with high prediction rates, as shown in result section.

[0054] Prior to discussion of the classification results, the terms and methods for verifying the performance of the present invention are set forth below.

[0055] Bayes Theorem

	Disease	**No Disease**
True Test	**True Positives**	**False Positives**
False Test	**False Negative**	**True Negative**

[0056] SENSITIVITY

[0057] Definition: Percent with the disease who test positive.

[0058] Formula: Sensitivity = True Positives / (True Positives + False Positives).

[0059] SPECIFICITY

[0060] Definition: Percent without the disease who test negative.

[0061] Formula: Specificity = True Negatives / (True Negatives + False Negatives)

[0062] POSITIVE PREDICITIVE VALUE

[0063] Definition: Percent of positive tests who actually have the disease

[0064] Formula: POSITIVE PREDICITIVE VALUE = True Positives / (True Positives + False Negatives)

[0065] NEGATIVE PREDICITIVE VALUE

[0066] Definition: Percent of negative tests who do not have the disease

[0067] Formula: NEGATIVE PREDICITIVE VALUE = True Negatives / (True Negatives + False Positives)

[0068] The classification results of the 46 prostate cancer tissues and 46 normal tissues from Affymetrix are summarized in Table 3 and Table 4.

[0069] Table 3. The classification results of the 46 prostate/normal pairs for un-normalized data.

	Disease	**No Disease**
True Test	39	2
False Test	7	44

[0070] We then used the above-discussed formulas to measure the performance of the present invention. The results for Table 3 are as follows:

[0071] Sensitivity = 39/(39+2) = 95.12%

[0072] Selectivity =44/(44+7) = 86.27%

[0073] Positive Predictive Value = 39/(39+7)=84.78%

[0074] Negative Predictive Value = 44/(44+2) = 95.65%

[0075] Prediction Result = (39+44)/(46+46) = 90.21%

[0076] Table 4. The classification results of the 46 prostate/normal pairs for mean normalized data.

[0077]

	Disease	**No Disease**
True Test	42	2
False Test	4	44

[0078] We then used the above-discussed formulas to measure the performance of the present invention. The results for Table 4 are as follows:

[0079] Sensitivity = 42/(42+2) = 95.45%

[0080] Selectivity =44/(44+4) = 91.67%

[0081] Positive Predictive Value = 42/(42+4) = 91.30%

[0082] Negative Predictive Value = 44/(44+2) = 95.65%

[0083] Prediction Result =(44+42)/(46+46) = 93.48%

[0084] In comparison with the data from Singh et al., the classification results of present invention are significantly better than the published work. In addition, the method of the present invention is much simpler and easier to deploy.

[0085] A more detailed comparison of classification results between the present invention and the Singh el al. also verified that 9 out of the first 18 genes from the present invention are identical to the ones used in Singh's 16-gene model. Table 5 shows the results.

[0086] Table 5. The corresponding genes identified by the present invention and Singh et al.

[0087]

1	37639_at	x
2	38634_at	x
3	41152_f_at	
4	32598_at	x
5	41706_at	
6	37366_at	
7	33656_at	
8	41468_at	x
9	1740_g_at	
10	39315_at	
11	40282_s_at	x
12	1804_at	
13	31527_at	
14	38406_f_at	x
15	39755_at	x

16	40024_at	
17	39054_at	x
18	38028_at	x

[0088] To illustrate the identified top 10 genes graphically, FIG 8a shows highly repetitive gene expression profile with an easily identifiable fundamental frequency. However, as shown in FIG 8b, the gene expression profile without any regularity does not have an identifiable fundamental frequency.

[0089] For further verification of the method of the present invention, a second dataset was downloaded from the following URL: http://microarray.cnmcresearch.org/ListProjExp.asp?ProjectName=WSilk+Macular+Degeneration. This is an experiment on Macular Degeneration using Affymetrix GeneChip HG_U95 Av2 conducted by Karl Csaky at National Eye Institute. Experimental Rationale: Age-related macular degeneration has a strong epidemiologic association with cardiovascular disease. One pathogenic hypothesis that applies to both diseases is the concept of an abnormal cellular response to injury resulting in a disease phenotype. Furthermore, it has been hypothesized that this abnormal cellular response is also present in dermal fibroblasts. It analyzed this hypothesis by using microarray technology.

[0090] We used the same microarray data from the published paper, and performed the clustering and achieved 100% success prediction rate. The top 10 genes are: FABP4, C14orf132, COX7A1, AKAP2, DLC1, ICAM2, ICAM2, PECAM1, PLA2G1B and ICAM2 (genes are represented by their symbols).

[0091] Another aspect of the present invention is the graphic representation of gene differential expression profiles by using the parameters derived from the Discrete Fourier Transform (DFT). The parameters include absolute amplitude, normalized amplitude, and phase/angle.

[0092] For a Discrete Fourier Transform (DFT), the results are always in complex values, i.e. $F = R+Ji$, where R is the real component and J is the imaginary component. i is defined as $sqrt(-1)$. The number of points in DFT depends on the input data. It is usually selected as the same number of points as the input data, or larger. In one embodiment of the present invention, we selected the number to be equal.

[0093] Absolute Amplitude

[0094] Let the number of point for the input data, f, to be N_0, the DFT of this data is given by:

$$F_r = \sum_{k=0}^{N_0-1} f_k \exp(-i\, r\Omega_0 k)$$

where $\Omega_0 = 2\pi N_0^{-1}$.

[0095] So F_r will be a complex number in the form of:

[0096] $F_r = R_r + J_r\, i$

[0097] The absolute amplitude is defined as:

[0098] $|F_r| = \text{sqtr}(|R_r|^2 + |J_r|^2)$

[0099] Normalised Amplitude

[00100] Normalised amplitude is defined as the absolute amplitude divided by the sqrt(total DFT amplitude).

[00101] Let F_{total} be the Total DFT amplitude, which is defined as:

[00102] $$F_{Total} = \sum_{n=0}^{N_0-1} |F_n|^2$$

[00103] Then the Normalised amplitude is defined as:

[00104] $$F_{r_N} = \frac{|F_r|}{\sqrt{F_{Total}}}$$

[00105] According to Parseval's Theorem, F_{Total} can also be calculated from the input data points as:

[00106] $$F_{Total} = N_0 \sum_{k=0}^{N_0-1} |f_k|^2$$

[00107] The Phase/Angle of the Discrete Fourier Transform is defined as:

[00108] $$\theta_k = \tan^{-1}\left(\frac{J_k}{R_k}\right)$$

[00109] For the purpose of comparison, three conventional representations of gene differential expression profiles are shown in FIGS 9a-9c. FIG 9a shows a conventional scatter plot of microarray data where the Red (R) intensity is plot versus the Green (G) intensity. For example, the R or G intensities may represent genes from disease tissues or normal ones. FIG 9b shows another conventional scatter plot of microarray data where the Log2(Red intensity) or (Log2R) is plot versus the Log2(Green intensity) or (Log2G). For

example, the Logarithm two of Disease gene intensity is plot versus Logarithm two of Non-disease intensity. FIG 9c shows yet another conventional scatter plot of microarray data where M is plot versus A. In this representation, the M is defined as: M = Log2C, where C = R/G; the A is defined as: A = (Log2R + Log2G)/2. It is obvious that FIG 9c provides a better representation of gene differential expression profiles. However, the majority of the genes that are not differentially expressed are condensed around the zero horizontal line so that it is hard to visualize the differentially expressed genes.

[00110] Now referring to FIG 10, the Nyquist Frequency Normalized Amplitude is plot versus Nyquist Frequency Absolute Amplitude, where the normalized amplitude of over-expressed genes which the Nyquist Frequency phase is zero, is set positive, and the normalized amplitude of under-expressed genes which the Nyquist Frequency phase is $-\pi$, is set negative. This representation makes the differentially expressed genes prominent.

[00111] While the present invention has been described with reference to particular embodiments, it will be understood that the embodiments are illustrative and that the invention scope is not so limited. Alternative embodiments of the present invention will become apparent to those having ordinary skill in the art to which the present invention pertains. Such alternate embodiments are considered to be encompassed within the spirit and scope of the present invention. Accordingly, the scope of the present invention is described by the appended claims and is supported by the foregoing description.

CLAIMS

What is claimed is:

1. A method of generating a gene differential expression profile for a pair of conditions (T/N), said method comprising steps of:

acquiring n sets (S) of gene expression profiles for the pair of conditions, wherein n >1, and each gene with each set of gene expression profile has two expression values, one for T and another for N; wherein each set of gene expression profiles contains m genes; and wherein m >1;

arranging the n sets of gene expression profiles in a linear order, i.e., S1(T1N1)S2(T2N2)...Sn-1(Tn-1Nn-1)Sn(TnNn), so the gene expression values from one condition are alternatively interposed by the ones from another condition; and

applying the Discrete Fourier Transform to the arranged gene expression values for each gene of the set of gene expression profiles; thereby generating the gene differential expression profile for the pair of condition.

2. The method of claim 1, wherein the gene differential expression profile contains at least one gene; thereby candidate genes that are either overexperssed or underexpressed in either of the pair of conditions may be selected from the gene differential expression profile.

3. The method of claim 2, wherein the gene differential expression profile is comprised of overexpressed genes and underexpressed genes.

4. The method of claim 1, wherein the pair of conditions may be any paired conditions in which at least one gene is differentially expressed.

5. The method of claim 4, wherein the paired conditions may be any one selected from the group consisting of tumor/normal, tumor x/tumor y, disease/normal, disease

x/disease y, development stage x/development stage y, treatment/non-treatment, treatment x/treatment y, and infected/non-infected.

6. The method of claim 1, wherein when the Discrete Fourier Transform is applied to the arranged gene expression profiles, it gives out values for parameters including Nyquist Frequency component, absolute amplitude, normalized amplitude, and phase.

7. The method of claim 6, further comprising the step of:

plotting the normalized amplitude versus absolute amplitude; thereby the gene differential expression profile is represented graphically.