# A Quality Measure Model for Microarray Images

Pan-Gyu Kim[1, 2], Kiejung Park[2], and Hwan-Gue Cho[1]

[1]Department of Computer Science and Engineering, Pusan National University, Busan 609-735, Korea ;
[2]SmallSoft Co., Ltd., Jang-Dong 59-5, Yusung-Gu, Daejeon 305-811, Korea;

{pgkim, hgcho}@pearl.cs.pusan.ac.kr

## Abstract

The DNA microarray method is important to determining the complete expression profile of a cell. Microarray image processing is becoming more important as the amount of microarray data increases so very rapidly. Because microarray image quality can be low, experimenters need to control image quality in microarray experiments. We devised five functions as quality measures for signal noise, background noise, scale invariant, spot regularity, and spot alignment, which can check the quality of microarray and validate the correctness of experiments. We also suggest a linear-weighted integration model combining the P-values of the five quality measures. In evaluating practical microarray images with these measures, each function provided good measure values.

**Keyword**: Microarray, Image quality, Quality measure, Auto gridding

## I. Introduction

The DNA microarray method is a very powerful tool to simultaneously show the gene expression of hundreds of thousands of test genes in a cell. As experimentation using microarray increases, the request of auto analysis increases. Microarray images have several faults due to an experimental environment or conditions. And if highly up-regulated genes or control genes are concentrated in any region of a chip, neighboring pixels might be affected by interference. These obstacles present a significant difficulty in microarray image analysis. Generally, a microarray experiment has three processes: design, experimentation and analysis. A good design helps to grid spots easily and to reduce inferences between spots. A good experiment produces good spots. If spot positions are irregular, background intensity can contain errors in image processing.

Although auto analysis systems for microarray images have been introduced recently [8, 9, 11], they assume that image quality is good and that the proportion of expressed genes to total genes is regular. Major commercial systems for microarray image analysis, such as GenePix[1] and ImaGene [2] among others, support several features for quality measures such as error flags. As these flags usually focus on the correctness of the analysis system rather than the quality of images, they cannot represent the quality of microarray images. Gabriel *et al.* defined a simple measure for noises to find the regularity of a microarray image in the auto gridding process, which was the first attempt to evaluate noise levels for microarray images [6]. Although Kuklin *et al*. developed more various and concrete quality measures, they offered just a separate quality for each measure (for example, signal to noise, the roundness of spots) instead of a quality value for an image [10]. By analyzing various

intensities of images statistically, Brown *et al*. showed that features of images vary as intensities of spots increase [3]. Whereas the integrated quantitative measure of Hautaniemi *et al*., which combines various quality measures using a Bayesian network, shows what factors have an effect on the qualities of microarray images, it does not suggest a quantitative value for an image [7].

It is necessary to develop measure functions to measure microarray image quality, so we can determine if we can analyze a microarray image automatically or cannot. When a client requests quantitative values for the quality of a DNA microarray experiment, those measure function values can be used as criteria.

In this paper, we compared the qualities of images using a few fundamental statistics and quantified those qualities using five measures. To estimate the quality of an image, we calculated statistics for inter-blocks, and inter-spots distances. An integrated model, merging five measures into a single value helped to evaluate the results of the image analysis.

## II. Materials and Methods

### A. *Primary Statistics for Microarray Images*

A few statistical values for spots and blocks are required before microarray quality measures can be defined. Those statistic show the general characteristics of chip image quality. We define a few statistical values for spot and block regularity as below:

- $ISa = \dfrac{\text{The sum of distances between adjancent spots in a same column}}{\text{The number of spots in a chip}} + \dfrac{\text{The sum of distances between adjancent spots in a same row}}{\text{The number of spots in a chip}} = \dfrac{\sum_i \sum_j (S_{i,j+1} - S_{i,j}) + \sum_i \sum_j (S_{i+1,j} - S_{i,j})}{\text{The number of spots in a chip}}$
- $IBa = \dfrac{\text{The sum of distances between adjancent blocks in a same column}}{\text{The number of blocks in a chip}} + \dfrac{\text{The sum of distances between adjancent blocks in a same row}}{\text{The number of blocks in a chip}} = \dfrac{\sum_i \sum_j (B_{i,j+1} - B_{i,j}) + \sum_i \sum_j (B_{i+1,j} - B_{i,j})}{\text{The number of blocks in a chip}}$
- $RS = \dfrac{\text{The sum of radii of spots}}{\text{The number of spots in a chip}} = \dfrac{\sum_{\text{each spot}} \text{Radius size of a spot}}{\text{The number of spots in a chip}}$

$S_{i,j}$ is the center position of a spot located in row $i$ and column $j$ in a spot grid, and $B_{i,j}$ is the center position of a block located in row $i$ and column $j$ in a block grid.

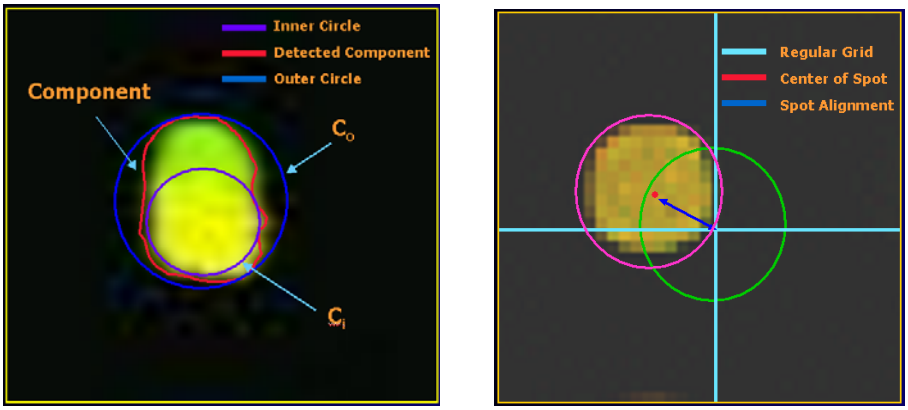
Generally, the larger the intervals between spots or blocks become the smaller the standard deviation of intervals is. And the larger the size of a spot radius becomes, the better the quality of microarray is.

### B. *Quality Measure Functions*

The expression rate of each spot is acquired by examining the intensity of the red and green pixels. Therefore, to evaluate the quality of a microarray experiment, it is required to quantify the quality of microarray images.

We define five quality measures that can represent the quality of a microarray image: signal noise, background noise, scale invariant, size regularity, and spot alignment. Signal and background noise lead to errors in calculating expression rate of a spot. Unlike a contamination or a scanning error, the fault of a spotter or experimenter can generate the problem of spot roundness. Scale invariant and size regularity can represent spot roundness. Whereas scale

invariant depends on spot radius, size regularity depends on spot shape. Spot alignment represents the regularity of spots and blocks to know whether auto gridding is possible.




(a) (b)

**Fig. 1**. **(a)**. Scale invariant and size regularity: Suppose the center of a signal component in a spot is the center of the spot. Let $C_o$ denote the smallest circle containing the signal component and $C_i$ denote the largest circle contained by the signal component. Scale invariant is defined as the ratio of the radius of $C_i$ to $C_o$ and size regularity is defined as the ratio of the area of the signal component area to $C_o$. **(b)**. Spot alignment: When we divide a block into a spot grid structure using even spaces, we calculate the spot alignment error as the difference between the real spot positions and the ideal spot positions in an ideal spot grid. If the spot alignment is 1, the real spot position is equal to the ideal spot position, and if 0, the real spot is located in the center of two adjacent ideal spots.

**Signal Noise for a Spot**: Contamination of a spot area such as a blob can result in an error in the image analysis process. We define the signal noise for a spot $Spot_{ij}$ as the ratio of contaminated signal spot area to signal area.

$$SN_{i,j} = \frac{\text{Contaminated Signal Area of } Spot_{i,j}}{\text{Signal Area of } Spot_{i,j}}$$

**Background Noise for an Image**: The intensity of the background influences the calculation of spot expression rates. We define the background noise for $Image_i$ as the ratio of contaminated background area to background area.

$$BN_i = \frac{\text{Contaminated Background Area of } Image_i}{\text{Background Area of } Image_i}$$

**Scale Invariant for a Spot**: A good spot is a circular shape. We define scale invariant representing the roundness of $Spot_{ij}$ as the ratio of the radius of the largest circle contained by $Spot_{ij}$ to that of the smallest circle containing $Spot_{ij}$.

$$SI_{i,j} = \frac{\text{Radius of The Largest Circle Contained by } Spot_{i,j}}{\text{Radius of The Smallest Circle Containing } Spot_{i,j}}, \in [0,1]$$

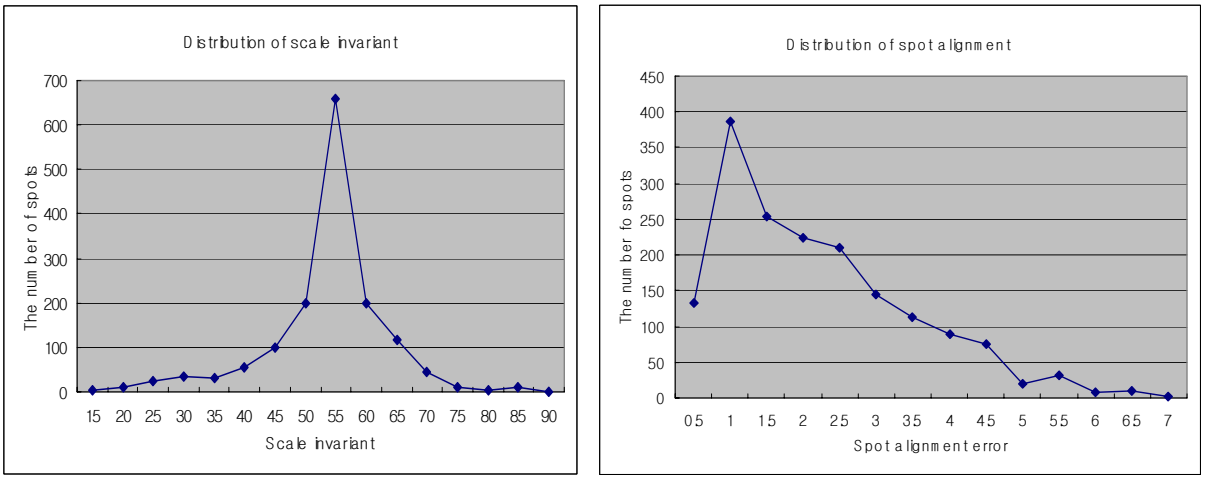
**Size Regularity for a Spot**: Size regularity is similar to scale invariant. It is focused not on the shape but the size of a spot. We define size regularity as the ratio of signal (component) area to

the area of the smallest circle containing $Spot_{ij}$(Fig. 1 (a)). Size regularity is influenced by the size of the spot radius because the area of a circle is in inverse proportion to the square of its radius. Unlike scale invariant, size regularity is affected by the radius of a spot rather than its shape.

$$SR_{i,j} = \frac{\text{Signal Area of } Spot_{i,j}}{\text{Area of The Smallest Circle Containing } Spot_{i,j}}, \in [0,1]$$

**Spot Alignment for an Image**: The position of each spot in a block must be at a regular interval. We define the spot alignment error as the difference between the spot positions in an ideal grid structure and the real spot positions. If spot alignment quality is bad, auto gridding is difficult and adjusting spots automatically is very time-consuming. Suppose each interval between adjacent spots in chip $Image_i$ is the same, and let $PS_{j,k}$ denote a spot located in microarray grid position ($j$, $k$), and $RS_{j,k}$ a spot located in real micorarray grid position ($j$, $k$). We define spot alignment as below (Fig. 1 (b)).

$$SA_i = 1 - \text{spot alignment error}$$
$$= 1 - \frac{\sum_j \sum_k | RS_{j,k} - PS_{j,k} |}{\text{Average of Distance Between Adjacent Spots} \cdot \text{Number of Spots}}, \in [0,1]$$




(a) (b)

**Fig. 2. (a)**. The distribution of scale invariant (1,715 spots): Scale invariant has a normal distribution like that of signal noise, background noise, and size regularity. **(b)**. The distribution of spot alignment (1,700 spots): Spot alignment has a Possion distribution unlike the others.

## C. *Linear Combination of P-values*

T Whereas signal noise, background noise, scale invariant, and size regularity have normal distributions, spot alignment has a Poisson distribution (Fig. 2). This helps to integrate the five measure values into a single value. To represent the quality of a microarray image as a single value, we derive an integrated function from the five measures. As each measure has a different distribution, each measure is transformed into a normalized quality measure that has a (0,1) domain.

Signal noise and background noise are inversely proportional to the quality of a microarray image. If measure $M$ has a distribution $P$, we define the normalized measure $NM$ for signal noise or background noise as below, where $x$ is a measured value.

$$NM = P(x > M)$$

And for the other three measures,

$$NM = P(x < M).$$

We make an integrated model by summing five weighted normalized measures,

$$TQ = \sum_{M} NM \cdot W_M$$

where $W_M$ is the weight of each measure $M$, and $\sum W_M = 1$. We can control the effect of a specific measure by changing its weight.

## III. Results and Discussion

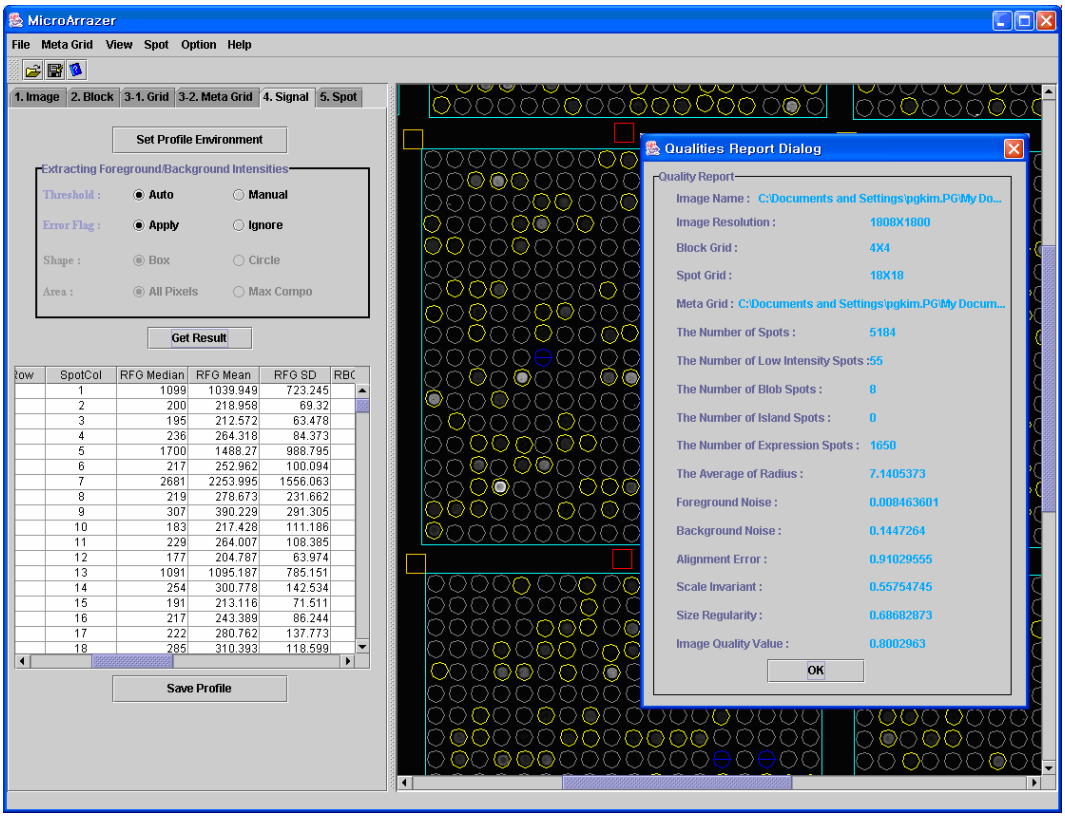



**Fig. 3.** A snapshot of the microarray image analysis system, Arrayzer: It supports special functions as five quality measures, an integrated quality value, and various error flags, among other features.

We implemented the quality measure model into Arrayzer, a microarray image analysis system (Fig. 3).We have tested the model with six chip images (*A* to *F*). First, the primary statistics were calculated as shown in Table 1. Arrayzer utilizes the primary statistics as a template in analyzing a microarray image with the same grid structure. Table 2 shows the quality values for six test images

and Fig. 4 shows six color-mapped test images of gray scale. Whereas scale invariant is relatively constant, the other four measures show varying values among the test samples. For example, *A* can have many erroneous expression rates due to bad background noise values. As shown in Fig. 4, in actuality the separation of the foreground from the background in *A* is very difficult. *D* has good values for background and signal noises. *E* has bad values for size regularity and spot alignment, the shape of many spots diverge from the circular, and the location of spots are irregular, as shown in Fig. 4. *C* and *D*, which appear as good images, have good quality values. *F*, which seems not suitable for auto gridding, has the worst quality values.

The test results for the images show that the value differences between images for signal noise and spot alignment are smaller than those for the others due to the development of microarrayer technology. There are quality differences of signal and background noises between images from a laser scanner and from a CCD camera for the same slide, and background noise shows a larger variation than others. Scale invariant and size regularity are influenced by experimental conditions. The amount of dye is an important factor. The more dyes is used, the better scale invariant and size regularity are.

## IV. Conclusions

In this paper, we defined five measure functions that check the quality of microarray and validate the correctness of experiments: signal noise, background noise, scale invariant, spot regularity, and spot alignment. Because each quality measure expresses the microarray quality very well, we can estimate how well a microarray image is created. We also suggested a linear-weighted integration model that combines P-values of the five quality measures.

We will try to develop a threshold value for each measure, which would be helpful in determining if a microarray images is practically useful. For example, if the background noise of *A* in Fig. 4 is more than 50% and the threshold for background noise is 50%, we can reject the result of A because the background noise of A exceeds the threshold.

Information and the software used in this research are available via http://pearl.cs.pusan.ac.kr/~arrayzer/index.html.

**Table 1.** Chip specification: This shows the expression rate and grid structure of spots and blocks in a chip. The information can be used as a meta-grid, which helps to analyze a microarray image that has the same grid structure. B : number of blocks; S : number of spots; BD : block distance; SD : spot distance; C : number of component detected.

| Chip | B | S | Image Resolution | BD | SD | Radius | C(%) |
|---|---|---|---|---|---|---|---|
| *A* | 16 | 16×100 | 1886×2321 | 102.41 | 35.97 | 8.16 | 141(8.8%) |
| *B* | 16 | 16×324 | 1900×1860 | 29.73 | 23.03 | 6.63 | 995(19.2%) |
| *C* | 16 | 16×324 | 1808×1800 | 31.33 | 23.02 | 7.14 | 1650(31.8%) |
| *D* | 16 | 16×168 | 1984×2004 | 122.79 | 24.89 | 7.49 | 1289(48.0%) |
| *E* | 16 | 16×342 | 2000×2000 | 22.17 | 22.93 | 8.82 | 656(12.0%) |
| *F* | 4 | 4×1596 | 1024×1024 | 23.17 | 10.43 | 4.13 | 715(11.2%) |

**Table 2.** Quality measure values for six test images.

| Chip | Background Noise | Signal Noise | Scale Invariant | Size Regularity | Spot Alignment |
|---|---|---|---|---|---|
| *A* | 51.94% | 0.53% | 0.5648 | 0.6931 | 0.9223 |
| *B* | 36.39% | 2.34% | 0.5238 | 0.5534 | 0.9016 |
| *C* | 21.35% | 0.85% | 0.5575 | 0.7315 | 0.9456 |
| *D* | 9.85% | 0.85% | 0.5608 | 0.6211 | 0.9526 |
| *E* | 30.13% | 1.58% | 0.5411 | 0.4512 | 0.8721 |
| F | 46.24% | 5.92% | 0.4714 | 0.5076 | 0.8445 |

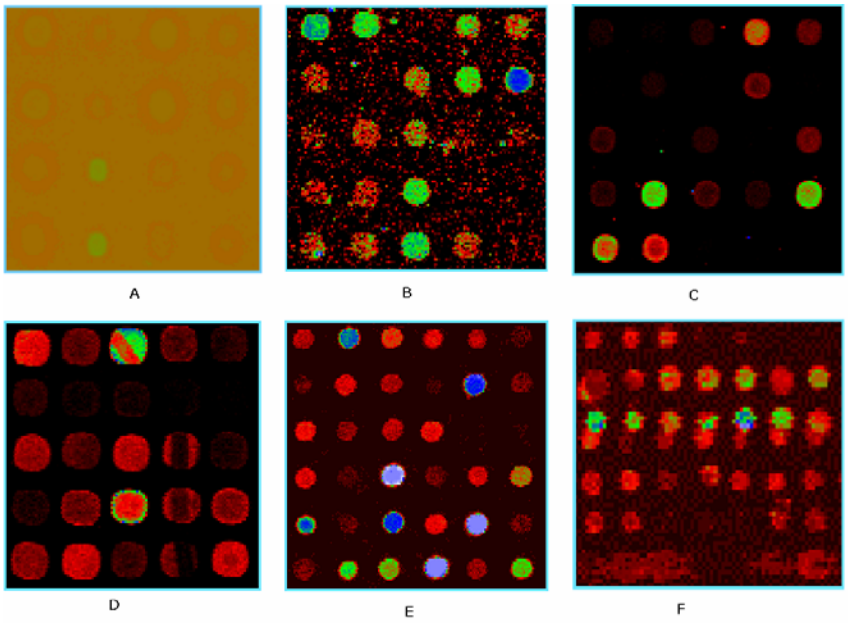



**Fig. 4.** Color-mapped images of six test gray images: *A* can have many erroneous expression rates due to bad background noise values. The separation of foreground from background in *A* is very difficult. *D* has good values for background and signal noises. *E* has bad values for size regularity and spot alignment, the shapes of many spots diverge from the circular, and the locations of spots are irregular. *C* and *D*, which appear as good images, have good quality values. *F*, which seems not suitable for auto gridding, has the worst quality values.

## References

[1] Axon Instruments Inc (2003) GenePix® Pro 5.0. http://www.axon.com/gn_GenePixSoftware.html.

[2] BioDiscovery (2003) ImaGene 5.6. http:// www.biodiscovery.com/imagene.asp.

[3] Brown, C.S., Goodwin, P.C., Sorger, P.K.: Image metrics in the statistical of DNA microarray data. PNAS, Vol. 98, No. 4. (2001) 8944-8949.

[4] Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P., Trent, J.M.: Expression profiling using cDNA microarrays. Nature Genetics, Vol. 21. (1999) 10-14.

[5] Ekins, R., Chu, F.: Microarrays: their origins and applications. Trends Biotechnology, Vol. 17. (1999) 217-218.

[6] Gabriel, R. B., Sethi, B.S.: On detecting spatial regularity in noisy images. Information Processing Letters, Vol. 12, No. 4. (1999).

[7] Hautaniemi, S., Edgren, H., Vesanen, P., Wolf, M., Jarvinen, A.K., Yli-Harja, O., Astola, J., Kallioniemi, O., Monni, O.: A novel strategy for microarray quality control using Bayesian networks. Bioinformatics, Vol. 19. (2003) 2031-2038.

[8] Jung, H.Y., Cho, H.G.: An automatic block and spot indexing with k-nearest neighbors graph for microarray image analysis. Bioinformatics, Vol. 18. (2002) 141S-151S.

[9] Jung, H.Y., Hwang, M.N., You, Y.J., Cho, H.G.: A graph model and analysis algorithm for cDNA microarray images. Journal of Korea Information Science Society, Vol. 29. (2002) 411-421.

[10] Kuklin, A.P.A., Shams, S.: Quality control in microarray image analysis. Microscopy Research. (2001).

[11] Steinfath, M., Wruck, W., Seidel, H., Lehrach, H., Radelof, U., O. Brien, J.: Automated image analysis for array hybridization experiments. Bioinformatics, Vol. 17. (2001) 634-641.

Pan Gyu Kim is a researcher in SmallSoft Co., Ltd. and Ph.D. student in Pusan National University, South Korea. He received B.S. and M.S. degree from the Department of Computer Science of Pusan National University. His research interests are image processing and data mining.

Kiejung Park is a senior researcher and CEO at SmallSoft Co., Ltd. He received B.S. degree from the Department of Computer Science and Engineering of Seoul National University in 1986, and M.S. and Ph.D. degrees from the Department of Biological Science and Engineering of Korea Advanced Institute of Science and Technology in 1989 and 2002 respectively. His research interests are sequence analysis algorithms and biological information system development.

Hwan-Gue Cho received B.S. degree from Seoul National University and M.S. degree and Ph.D. in Korea Advanced institute of Science and Technology, South Korea. Currently, he is a professor in Pusan National University, South Korea. His research interests are visualization in graphics and sequence alignment and bio network analysis in bioinformatics.