EcoGenomics: analysis of complex systems via fractal geometry

Robert W. Chapman,[1,*] Javier Robalino,[†] and Harold F. Trent III[§]

*South Carolina Department of Natural Resources, Hollings Marine Laboratory, 331 Fort Johnson Road, Charleston South Carolina. 29412, USA; [†]Medical University of South Carolina, Hollings Marine Laboratory, 331 Fort Johnson Road, Charleston, SC 29412, USA; [§]The National Ocean Service, Hollings Marine Laboratory, 331 Fort Johnson Road, Charleston, SC 29412, USA

Synopsis Ecogenomics is a convenient descriptor for the application of advanced molecular technologies to studies of organismal responses to environmental challenges in their natural settings. The development of molecular tools to survey changes in the transcript profile of thousands of genes has presented scientists with enormous analytical challenges. In the main, these center about the reduction of massively paralleled data to statistics or indices comprehensible to the human mind. Historically, scientists have used linear statistics such as ANOVA to accomplish this task, but the sheer volume of information available from microarrays severely limits this approach. In addition, important information in microarrays may not reside solely in the up or down regulation of individual genes, but rather in their dynamic, and probably nonlinear, interactions. In this presentation, we will explore alternative approaches to extracting of these signals using artificial neural networks and fractal geometry. The goal is to produce predictive models of gene dynamics in individuals and populations under environmental stress and reduce the number of genes that must be surveyed in order to recover transcript profile patterns of environmental challenges.

Introduction

The advent of microarray technology nearly a decade ago presented biologists with unprecedented access to the transcription profiles of organisms and unparalleled access to information on the molecular responses to environmental stress and disease (Brown and Botstein 1999; Young 2000; Waters and others 2003; Williams and others 2003). This technological power did not come without a cost as methods to assess reproducibility of the data and analytical means to turn the data into information lagged far behind the development of molecular tools. In the main, the issues have been normalization methods to account for nonbiological sources of variation including starting material, labeling, hybridization, and local and global bias in background due to differences between dyes (Cy 3 versus Cy 5) (van de Peppel and others 2003). Further analysis of the data, once "normalized" is complicated by its massively paralleled nature, which, in the main, precludes the application of linear statistics.

While the technical and analytical methods employed to assess changes in transcript profiles are important to understanding microarray data, this work seeks to explore a different territory. Functional genomics as currently understood, seeks to identify genes that are significantly affected by stressors and place the changes within the context of their metabolic pathways. Hence, the emphasis is upon analytical methods that identify significant changes and discard the genes that change little or not at all. EcoGenomics as originally described by Chapman (2001) has a different perspective. It is to understand the "gestalt" of transcription signatures; their patterns, if you will. The differences between functional genomics and EcoGenomics are equivalent to the difference between ecotoxicologists that monitor keystone species as indicators of environmental conditions and ecologists that wish to understand the function of the ecosystem. Clearly there is some overlap, but ecotoxicology would emphasize organisms at the extremes (highly sensitive to change), while ecologists would consider those species that show no changes as important features of ecosystem function. In this context EcoGenomics would place value upon transcription signatures that do not change as a result of environmental conditions, while functional genomics might not.

So how does one recover the patterns? Ecology has a long history of describing systems in terms of diversity, evenness, and patchiness. These terms have no counter parts in molecular biology even though the frequency

From the symposium "Genomic and Proteomic Approaches in Crustacean Biology" presented at the annual meeting of the Society for Integrative and Comparative Biology, January 4–8, 2006, at Orlando, Florida.

[1] E-mail: chapmanr@mrd.dnr.state.sc.us

Integrative and Comparative Biology, volume 46, number 6, pp. 902–911
doi:10.1093/icb/icj049
Advance Access publication May 10, 2006

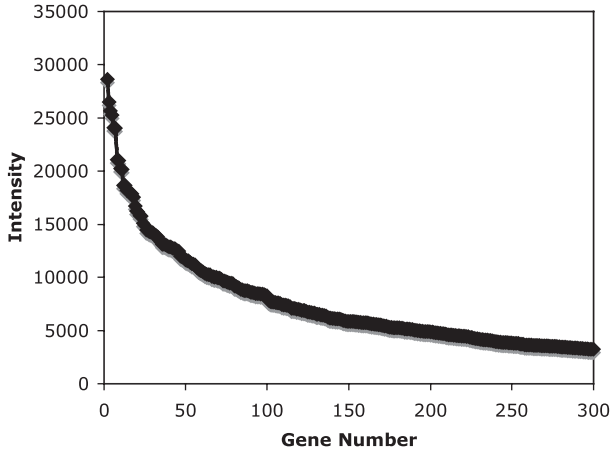

Fig. 1 Plot of the 300 most intensely fluorescent genes. On the *x*-axis the genes were ranked according to their intensity values (*y*-axis).

distributions of species in ecosystems (MacAuthur and Wilson 1967) bear an uncanny similarity to the frequency distributions of mRNA transcripts in tissues, assuming that fluorescence is proportional to transcript abundance (Fig. 1). More recently, ecologists have been using fractal geometry to study a variety of phenomena related to pattern recognition ranging from the distributions and abundances of phytoplankton and benthic communities (Azovsky and others 2000; Lovejoy and others 2001), and the influence of physical parameters on those distributions, to the influence of water movements on the branching patterns of marine sponges (Abraham 2001). The motivation for this development has been the recognition of the relationship between patterns and scale in ecology (Turner and others 1991) and the difficulties in scaling measures of species distributions in time and space. Fractal geometry solves many of the problems.

Fractals are objects whose topological dimensions (points, lines, and planes) are different and usually greater than the capacity dimension. The capacity dimension is a phase space in which noninteger dimensions are permitted and usually defined as

$$D_{\text{cap}} = -\lim[(\ln N)/\ln(e), \text{as } e \text{ approaches } 0,$$

where N is the number of elements which cover the metric space and e is the diameter of N. In addition, true fractals, such as the well-known Mandelbrot and Sierpinski fractals, also satisfy the inequality

$$D_{\text{corr}} < D_{\text{inf}} < D_{\text{cap}},$$

where D_{corr} is the correlation dimension and is defined in http://mathworld.wolfram.com/CorrelationDimension.html, while D_{inf} is the well-known Shannon information index (or diversity index in ecology). These measures are often referred to as the fractal moments analogous to the mean, variance, and skewness of objects in Euclidean space. Fractals are also self-similar, meaning that they contain copies of themselves within themselves, and, while they may not exhibit the same details at all scales, they exhibit the same type of structure. Hence, they are considered to be scale invariant. As such they are ideal for the study of ecological patterns, such as species abundances, which can display patterns at a variety of spatial scales (cf. Azovsky 2000; Lovejoy and others 2001). The usual method for estimating D_{cap} is the box-counting method, in which the metric space is progressively divided in half and counting the number of boxes, which contain data points. The slope of the line based on a plot of $\ln(e)$ versus $\ln(N)$, is the capacity dimension (cf. http://www.ees.nmt.edu/~davew/P362/boxcnt.htm, for example). An additional benefit of extracting D_{cap} via the box-counting method is that one can estimate the number of boxes necessary to recover the original metric space and their sizes. This is important in ecology as it informs us as to the number of samples and spatial scales over which they should be collected, in order to recover the original geometry of distributions and abundances of the species.

The benefits of employing fractal geometry to ecological situations should be obvious, but its application to microarray data may be less clear-cut. Most methods that compare microarray data involve normalization to some standard or reference material. This is necessary to compensate for the nonbiological sources of variation mentioned above. Normalization of microarray data has taken several forms, but generally relies upon linear transformations of the data. Some investigators have employed housekeeping genes under the assumption that these genes were largely stable across the range of experimental conditions. For the most part this has been dismissed in favor of "all genes" or quantile normalizations under the assumption that relatively few genes vary in expression levels between samples. This assumption has been shown, in some cases to be unrealistic (van de Peppel and others 2003) and these normalizations mask general changes in transcription levels from various stressors. Others have employed "spike in" controls at a single or linear gradient of concentrations to establish standards against which to measure experimental data (van de Peppel and others 2003). While the latter approach is sound and offers great promise, it is technically difficult to execute. Regardless of the means by which the data from different arrays are brought to the same scale, most investigators have sought to

identify genes that are up-regulated or down-regulated by some degree of fold changes. Placing confidence limits of fold changes, challenges the application of linear statistics due to the degrees of freedom (number of variables = genes), even if one accepts the validity of current normalization protocols. This is the "Curse of Dimensionality" as recognized by Bellman (1961) where the available output hyperspace expands exponentially with linear increases in the number of input variables. As a concrete illustration, to examine all possible single-gene impacts with a very modest 1000 gene array would require 1001 observations or arrays. To examine all possible linear pairwise interactions of the same array would require nearly 500,000 observations and the problem grows exponentially worse as higher dimensional interactions are considered. Yet it is the interaction, and presumably nonlinear interaction, between genes that initially prompted our interest in the technology. By virtue of their ability to identify the number of variables necessary to describe the original space, fractals can provide a means of addressing this problem.

Machine learning tools such as Artificial Neural Networks (ANNs, Hagan and others 1996; Khan and others 2001), Support Vector Machines (SVMs, Vapnik 1998) and Genetic Algorithms (GA, Holland 1975), have been used as alternatives to linear statistical methods to deconvolve patterns in complex data. Unlike linear approaches they require little or no mechanistic understanding of the system, but do require large amounts of data. As such, they would seem to be ideally suited for the massively paralleled nature of microarray data. These methods, however, are computationally intense and for modest arrays containing, say, 10,000 features might require extensive computation time to find optimal solutions. Further, the training set for ANNs needs to contain large numbers of records (individual microarray slides), and while there is no set minimum, it is our experience that 100–150 are required for adequate training. We do not know of a microarray dataset this large. The ability of fractals to reduce dimensionality and retain original geometry would vastly accelerate the discovery of patterns and processes using these tools (Wang and others 2005).

In this article, we will discuss the application of some tools from fractal geometry to microarray analysis and the implications for EcoGenomics. As we will show, the analysis provides some unique insights into patterns and processes, in addition to reducing the information necessary to recover those patterns. In the main, however, our intention is to use these tools as a prelude to more sophisticated analyses such as ANNs.

Methods and materials

Microarrays

The development of ESTs for printing the microarrays used in this study was presented elsewhere (Gross and others 2001; Robalino and others manuscript in preparation, www.marinegenomics.org) and the details of printing and quality control will also be presented in a separate publication (J. Robalino and others manuscript in preparation). In brief, the arrays contain 13,056 individual features, comprising 2469 unigenes along with landing lights, positive and negative controls, and viral genes derived from the white-spot syndrome virus (WSSV). WSSV is a lethal pathogen of the Pacific white shrimp, *Litopenaeus vannamei*, and these genes have been included in the arrays as a means of testing for viral gene expression in challenged individuals.

RNA samples and microarray hybridizations

Total RNA was obtained from tissues stored in RNA later (Ambion) using RNeasy kits (Qiagen) according to the manufacturer's instructions. For hemocytes, cells were freshly recovered from hemolymph by centrifugation, and used for RNA extraction as described above. To generate labeled target RNA, total RNA (1 μg) from gills, hepatopancreas, or muscle obtained from individual shrimp was used in 1 round of linear RNA amplification using the Amino Allyl MessageAmp II aRNA kit (Ambion). For hemocyte samples, essentially all the RNA extracted from one individual shrimp was used, and 2 rounds of amplification were applied. Amino allyl-modified RNA (aRNA) (10 μg) labeled with Cy3 (Ambion) was used for subsequent hybridizations in every case. For microarray pre-treatment, slides were rinsed in 0.2% SDS for 1 min, rinsed with water, boiled for 1 min, rinsed again with water, and dipped in 70% ethanol before drying. Arrays were prehybridized in 50% formamide, 2.5× Denhardt's solution, 4× SSPE, 2.4% SDS, and 100 μg/ml salmon sperm DNA for 1 h at 50°C. Labeled aRNA was boiled for 1 min and prehybridized at 50°C for 1 h in 33% formamide, 2.6× SSPE, 1.6% SDS, 1.7× Denhardts, poly dA (1 mg/ml), and mouse cot-1 DNA (1 mg/ml). Hybridization was performed overnight at 50°C in an air incubator (SlideOut, Boekel). Washes were as follows: once in 2× SSC–0.1% SDS for 5 min, twice in 0.2× SSC–0.1%SDS for 5 min, twice in 0.2× SSC for 5 min, and once in 0.1× SSC for 5 min. After a brief rinse with water, the slides were dried and scanned using a ScanArray Express instrument (Perkin-Elmer). Images scanned at 67 PMT gain and 90% laser power were used for tissue-specific profiling analyses. Raw images were

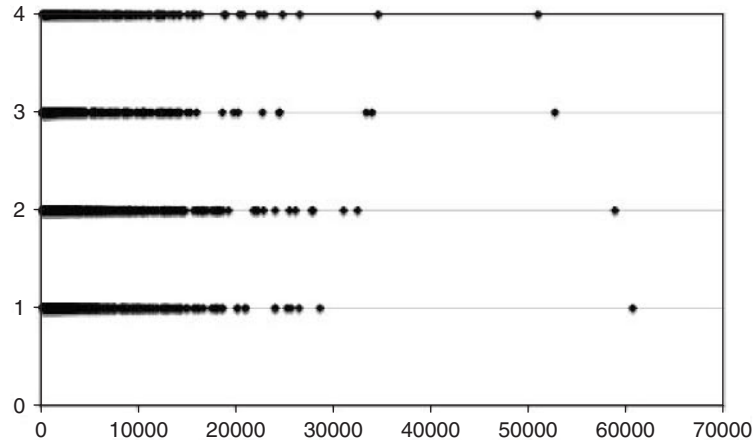

Fig. 2 Raw intensity values after background subtraction of the unigenes in gill (1,2) and muscle (3.4) from two *L. vannamei* individuals plotted along a single dimensional axis.

processed and mapped using the GridGrinder tool integrated into the microarray analysis pipeline at www.marinegenomics.org.

To analyze the effects of WSSV infection in gene expression, the hepatopancreas of 8 uninfected and 8 WSSV-infected animals was used to interrogate the microarray, using the same methods described above. The infection was allowed to progress for 40 h after injecting a dose of WSSV sufficient to cause 100% mortality in 4–7 days (Prior and others 2003). The data from this experiment were analyzed from images scanned at 75 PMT gain and 90% laser power. All of the data from the microarray scans can be found at and retrieved from www.marinegenomics.org. The tissue-comparison samples are MGMA# 386–397 and the viral-challenge samples MGMA# 362–378.

Fractal analyses

Previous studies employing fractal geometry to microarray analyses have relied upon either the information (Cazalis and others 2004) or capacity dimension (Wang and others 2005) estimates as baselines for "good" solutions to probe selection for clustering. However, there are three measures of fractal dimensions that have value in the biological interpretation of microarray data. The capacity dimension (D_{cap}), which describes the space occupied, the information dimension (D_{inf}) which is the Shannon Information or Entropy measure, commonly used as a measure of community diversity in ecology, and finally, the correlation dimension (D_{cor}) which measures the correlation between points in phase space. It is also known as the attractor space or dimension as it describes the tendency of fractal objects to exhibit chaotic orbits in an equilibrium space. While the biological meaning of the measures will become clearer later in this report, it is valuable at this point to illustrate the points via graphics.

In Figure 2, we illustrate the intensity plots for four microarray slides where each spot on the line represents the average intensity of each unigene on the microarray. The unigenes were average solely for the purpose of illustration. We used the program, FD3 (Sarraille and DiFalco 1993) to estimate the fractal dimensions, which divides the space into 2^{32} line segments and counts the number of segments that contain a spot. The program then halves the number of segments and repeats the counting. The program continues halving the number of segments until the entire line is treated as one segment. The program is based upon the methods of Leibovitch and Toth (1989). For two axes (comparing two microarrays) the segments are boxes, for three axes cubes, and so on. For the purpose of this article we will limit our analyses to one and two embedding dimensions (individual slide and pairwise analyses). We have written a program in MATLAB that can run multidimensional comparisons as well as individual analyses, run FD3 and then assign coordinates to individual features. This work will be reported elsewhere.

We generated means and standard deviation for replicates of each tissue type from the tissue-comparison experiment and for the 8 control and 8 viral samples in the WSSV challenge experiment. An alternative to estimating fractal dimensions was employed. The approach is shown in Figure 3, where the values for the individual runs for gill samples were combined (3 samples) and the capacity dimension was determined by plotting log base 2 of line segment size, Ln (e), against log base 2 of the number of occupied segments, Ln N(e). The negative slope of the linear regression is D_{cap}. The D_{inf} and D_{cor} were

determined in a similar manner using the appropriate values generated by FD3 to replace ln$N(e)$.

To estimate the number of features on the microarrays necessary to recover the original geometry, we fitted a second-order polynomial to the D_{cap} points and evaluated the first derivative at a point equal to D_{cap} (Fig. 3, poly. D_{cap}). This returned the value of Ln $N(e)$ which was used to estimate the number of required features from an additional plot of Ln $N(e)$ versus the actual count returned by FD3.

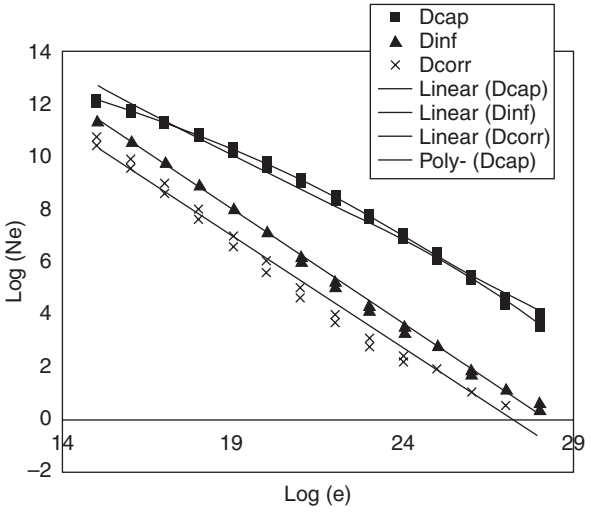

Fig. 3 Plots of log of segments size (e) versus the log of the number of segments occupied for the capacity, information and correlation dimensions from gill tissue. The negative slopes are reported in Table 1 as direct estimates. (See text for details).

Results

Inspection of Figure 2 shows that muscle has fewer genes with intensities above 15,000 than does gill. Intuitively, we would expect that the increase in density of intensities about 15,000 in gill tissue would increase the D_{cap} in gill relative to D_{cap} in muscle. This is reflected in Table 1 where the fractal dimensions for the 4 tissues from 3 shrimp are summarized. The top half of the table presents the average and standard deviations for each dimension for each tissue type calculated from individual runs of FD3. Here, the fractal dimensions of hemocytes and gill tend to be larger than those of hepatopancreas and muscle. In Table 2, tests for significant differences in these dimensions are shown and indicate that muscle differs significantly from the other tissue in all dimensions except for the D_{corr} comparison to hemocytes. The fractal dimension obtained by combining all data points prior to computing the dimensions are presented in the bottom half of Table 1. Here, D_{cap} is smaller than the average value computed from individual slides in all tissues. The opposite is the case for D_{inf} and D_{corr}. The number of features required to recover D_{cap} for the individual tissues is shown in the last row of Table 1 and in general indicate that larger fractal dimensions require more features to recover the space. Regardless of the method used to estimate the fractal dimensions, the tissues ranked as hemocytes > gill > hepatopancreas > muscle.

Progressing to pairwise comparisons permits the fractal dimensions to expand, and these comparisons are presented in Table 3. Here we see the same general

Table 1 Means and standard deviations of fractal dimensions for comparisons of expression profiles within and between tissues of shrimp, *L. vannamei*

	Hemocytes	Muscle	Gill	Hepatopancreas
(Capacity)				
Avg	0.7036	0.5896	0.6842	0.6581
SD	0.0171	0.0281	0.0068	0.0238
(Information)				
Avg	0.8079	0.7305	0.8042	0.7883
SD	0.0344	0.0150	0.0075	0.0217
(Correlation)				
Avg	0.7436	0.6934	0.7558	0.7434
SD	0.0441	0.0116	0.0146	0.0269
Capacity direct	0.6462	0.5472	0.6505	0.6287
Information direct	0.8482	0.7948	0.8580	0.8454
Correlation direct	0.8270	0.7840	0.8371	0.8223
Required features	456	385	419	404

Avg = average over three slides and SD = standard deviation over three slides.

Table 2 Student's *t*-test values for comparisons of dimensions (from Table 1) for *L. vannamei* tissues

	Muscle	Gill	Hepatopancreas
Hemocytes			
Capacity	**6.0027**	1.8259	2.6891
Information	**3.5723**	0.1820	0.8347
Correlation	1.9068	0.4549	0.0067
Muscle			
Capacity		**5.6674**	**3.2219**
Information		**7.6117**	**3.7951**
Correlation		**5.7960**	**2.9563**
Gill			
Capacity			1.8264
Information			1.1995
Correlation			0.7017

Values in bold are significant at $P < 0.05$.

pattern of hemocytes, hepatopancreas and gill having larger fractal dimensions than does muscle, except for D_{corr}. In these pairwise comparisons D_{corrz} remained approximately the same as in the individual slide analysis presented in the previous paragraph. These comparisons also distinguished D_{cap} in hepatopancreas from all other tissues and the significance tests are presented in Table 4. The averages values of D_{inf} and D_{corr} were significantly different only in comparisons of gill versus muscle and hemocytes versus muscle due to the large standard deviations associated with these estimates.

Pairwise comparisons can also be made between tissues (Table 3). The fractal dimensions in these cross-tissue comparisons tend toward intermediate values between those of individual tissues. For example, the cross comparison of gill and muscle generated D_{cap} of 0.9154 while the within-gill value was 1.0256 and the muscle value 0.7779. In Figure 4, we plotted the intensities of gene expression in muscles and gills. In the upper graph, one muscle sample was used on the ordinate and another muscle sample was used to derive the coordinates highlighted in yellow. In blue are the coordinates generated by a comparison of muscle versus gill. In the lower graph the reverse is illustrated. These plots reveal one of the limitations in using fractal geometry; while D_{cap} in the gill versus gill comparison is larger than the corresponding value in gill versus muscle, the fractal values do not suggest the skewed distribution of points that indicates the number of genes that are more highly expressed in gill than in muscle. The fractals do recover the dramatic differences in the dispersion of points observed from within-tissue comparisons.

Table 3 Fractal dimension from the pairwise comparsions of intensity values from shrimp (*L. vannamei*) tissues computed from averaging each pairwise combination

	Hemocytes	Gill	Hepatopancreas	Muscle
Hemocytes				
(Capacity)				
Avg	1.0694	1.0536	1.0469	0.9239
SD	0.0337	0.0306	0.0278	0.0178
(Information)				
Avg	1.0600	1.0755	1.0528	0.9741
SD	0.1062	0.0583	0.0656	0.0621
(Correlation)				
Avg	0.6689	0.7353	0.6894	0.6770
SD	0.1326	0.0758	0.0806	0.0779
Gill				
(Capacity)				
Avg		1.0256	1.0203	0.9154
SD		0.0284	0.0190	0.0197
(Information)				
Avg		1.0649	1.0693	0.9743
SD		0.0439	0.0332	0.0325
(Correlation)				
Avg		0.7596	0.7517	0.7139
SD		0.0841	0.0519	0.0435
Hepatopancreas				
(Capacity)				
Avg			0.9483	0.8843
SD			0.0168	0.0286
(Information)				
Avg			1.0164	0.9574
SD			0.0364	0.0460
(Correlation)				
Avg			0.7657	0.7056
SD			0.0628	0.0726
Muscle				
(Capacity)				
Avg				0.7779
SD				0.0292
(Information)				
Avg				0.8911
SD				0.0325
(Correlation)				
Avg				0.7110
SD				0.0573

Abbreviations as in Fig. 1.

The fractal dimensions generated by computing slopes from the combined pairwise outputs of FD3 are presented in Table 5. In general, these values are virtually identical to those obtained from averaging

Table 4 Tests for significant differences in the fractal dimensions of the indicated *L. vannamei* tissues

	Muscle	Gill	Hepatopancreas
Hemocytes			
(Capacity)	**11.32**	1.72	**5.58**
(Information)	2.63	0.07	0.14
(Correlation)	0.50	1.00	1.14
Muscle			
(Capacity)		**10.52**	**8.76**
(Information)		**5.51**	**4.44**
(Correlation)		0.83	1.11
Gill			
(Capacity)			**4.06**
(Information)			1.47
(Correlation)			0.10

Values employed were the diagonal elements in Table 3 and in bold are the significant values at $P < 0.05$.

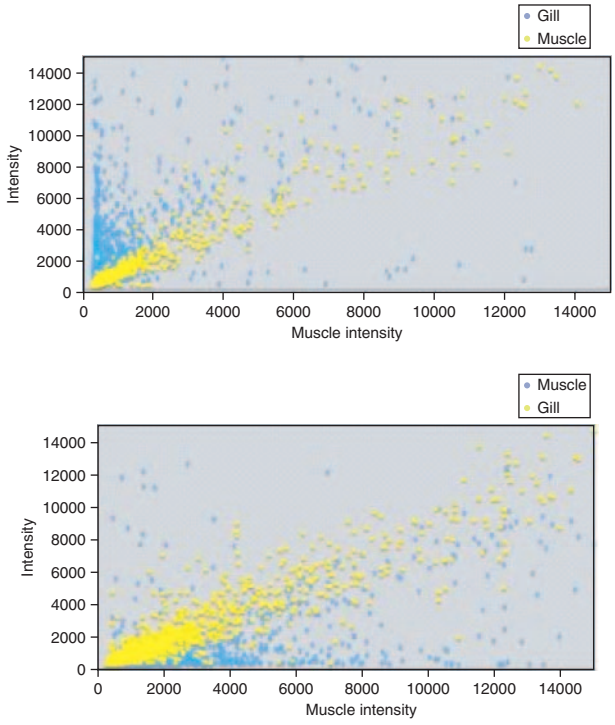

Fig. 4 Plots of background subtracted intensities for each feature on the arrays. Muscle tissue is on the ordinate and muscle (yellow) and gill (blue) values on the abscissa in the upper panel. The lower panel is a reverse plot using gill on the ordinate and gill (yellow) and muscle (blue) on the abscissa.

the individual estimates. The number of features necessary to recover the capacity dimensions of each comparison (Table 5) ranged from a low of 332 (gill versus hepatopancreas) to a high of 426 (hemocytes versus muscle). These values are consistent with the range identified in the individual slide analyses presented above (Table 1).

The second dataset analyzed in the present report is a viral challenge of shrimp summarized in Tables 6 and 7. Here, hepatopancreas tissue was used in the microarray analysis and four salient points should be noted. First, the fractal dimensions in the control individuals are remarkably similar to those for hepatopancreas in the individual tissue comparisons (Tables 1 and 6); however, the number of features required to cover the space increases relative to the tissue data, due to a small difference in D_{cap}. Second, the standard deviations for the fractal dimensions are smaller (Tables 6 and 7) than those in tissue-comparison data (Tables 1 and 3), which is as it should be due to large sample sizes. Third, we do not see any large difference in the fractal dimensions between the controls and virally challenged groups, except for D_{cap} and D_{inf} (Table 6) when the slopes were calculated from the combined individual slide outputs, resulting in a 3-fold difference in the number of features required to cover the space (Table 6). This difference is not evident in the pairwise comparisons (Table 7) and a plot (data not shown) of controls versus controls and controls versus virally challenged individuals showed a slight tendency for suppression of expression in virally challenged individuals.

Discussion

The analyses presented in this article are qualitatively distinct from those used in most microarray studies in that we employ background-corrected raw intensity values and recover all the microarray features necessary to recover the original geometry. This contrasts with the normalization of microarrays and identification of genes with significantly different transcriptional signatures between experimental and control conditions. It, thus, avoids the complications introduced by normalization procedures, the need for internal standards and titration discussed by Van de Peppel and others (2003), and recovers information on general suppression of transcription due to a stressor or tissue type.

While the above attributes are important to our understanding and analyses of microarray data, we are also of the opinion that comparisons of fractal dimensions offer some important biological insights. We noted above that the fractal dimensions in tissue comparisons ranked as hemocytes > gill > hepatopancreas > muscle. Given the functions of these tissues, this ranking makes some sense. Hemocytes are the primary defense cells in shrimp and respond quickly

Table 5 Fractal dimensions from pairwise comparisons of *L. vannamei* tissues computed by taking the slopes after combining individual comparisons

	Hemocytes	Gill	Hepatopancreas	Muscle
Hemocytes				
Capacity	1.0694	1.0535	1.0467	0.9195
Information	1.06	1.0687	1.0545	0.9662
Correlation	0.6689	0.713	0.7023	0.6687
Required Features	350	405	380	426
Gill				
Capacity		1.0159	1.0202	0.9114
Information		1.0698	1.0542	0.9727
Correlation		0.7818	0.727	0.7133
Required Features		357	332	400
Hepatopancreas				
Capacity			0.9483	0.8872
Information			1.0164	0.9594
Correlation			0.7657	0.7089
Required Features			368	411
Muscle				
Capacity				0.7733
Information				0.8938
Correlation				0.7174
Required Features				398

See Fig. 2 for more details.

Table 6 Fractal dimensions for individual slides from viral-challenge experiment

	Control	Virus
Capacity		
Avg	0.6744	0.6735
SD	0.0071	0.0097
Information		
Avg	0.7861	0.7928
SD	0.0121	0.0159
Correlation		
Avg	0.7382	0.7510
SD	0.0179	0.0258
Capacity direct	0.6744	0.6155
Information direct	0.7861	0.8410
Correlation direct	0.8382	0.8318
Required features	641	192

Abbreviations as in Fig. 1.

to a diverse array of invading pathogens whereas gills perform a variety of functions in addition to oxygen exchange. Hepatopancreas tissue performs a host of digestive functions, but interacts less directly with the environment. It is intuitive that these tissues would exhibit a more diverse array of transcriptional signatures than does muscle. We must be careful not to overinterpret the tissue analyses as these are based upon only three test animals, and larger samples sizes are clearly in order. It should be noted that the muscle transcriptome is probably underrepresented in this microarray, as the EST-mining strategy that led to the generation of this tool focused on hemocytes, gills, and hepatopancreas as sources of genes. It is conceivable that this bias influenced the rank of the fractal dimensions described above.

The viral-challenge experiment also provides some important, if somewhat unproven, suggestions. In comparing estimates of D_{cap} using the output of all 8 controls and all 8 experimentals (Table 6), we see that D_{cap} and the number of features necessary to recover the fractal space decline in the viral-exposed group. We interpret this as an indication that transcription and individual variation in expression decline under a viral exposure. We note that this is a very slight decline in the pairwise comparisons of control and virally challenged shrimp. We could address this hypothesis in a number of ways. One way would be to extend the analyses into even higher dimensional embedding space (more slides compared at once); however, this approach would come at a cost. The minimum number of data points required

Table 7 Average and standard deviation of fractal dimension for pairwise comparisons of virus challenged and control shrimp (upper portion)

	Control versus Control	Virus versus Virus	Control versus Virus
Capacity			
Avg	0.9489	0.9476	0.9467
SD	0.0122	0.0112	0.0205
Information			
Avg	1.0018	1.0455	1.0266
SD	0.0390	0.0495	0.0417
Correlation			
Avg	0.6729	0.7380	0.7086
SD	0.0473	0.0760	0.0609
Capacity	0.9495	0.947	0.9471
Information	1.0335	1.0427	1.0256
Correlation	0.7183	0.7324	0.7073
Box	571	577	585

The lower portion are the fractal dimension computed by taking slopes from the combined outputs.

for estimation is 2^{4Fd}, where Fd is the fractal dimension (Liebovitch and Toth 1989) and as we do not know what Fd is (it is what the analysis is trying to find), the number of axes (embedding dimensions) is usually taken as a surrogate (Sarraille and DiFalco 1993). In this case, where our arrays contain 13,056 features, we are limited to 3 slide comparisons ($2^{12} = 4096 < 13{,}056 < 2^{16} = 65{,}536$). This also suggests that more than 4 slide comparisons using fractal geometry are unlikely to ever be achieved due to space limitations on even the most densely populated microarrays.

The analyses in this article also indicate that all of the comparisons, whether done in one or two embedding dimensions, are not true fractals where $D_{cap} > D_{inf} > D_{corr}$, because in general $D_{inf} > D_{cap}$.in these analyses. This inequality is usually satisfied (asymptotically) when rather simple processes underpin the overall shape of the object under investigation. For multifractal sets, where multiple processes acting at a variety of scales may underpin the overall geometry, this inequality may not hold. The failure of our analyses to satisfy the inequality most likely indicates that the transcript profiles are responding to multiple driving forces and is consistent with our current (albeit limited) understanding of the genome and metabolic processes.

We are not unaware that the reader is probably curious about the genes identified by fractals as important in recovering the capacity dimension. We have remained deliberately silent on the issue for two reasons. First, most of the line segments in individual slide analyses or boxes in pairwise comparisons contain many genes, any one of which would serve as a surrogate for the others in subsequent analyses such as ANNs or clustering. Second, and more important, is that gene selection for subsequent analysis or re-designing microarrays as monitoring tools for specific (or even general) environmental challenges is a field ripe for the application of genetic algorithms as has been done for probe selection for oligonucleotide arrays (Cazalis and others 2004). Genetic algorithms take principles from biology to find optimal solutions to a variety of optimization problems, including those in engineering (Holland 1975). In this case, we would construct line segments (chromosomes), equal to the number of features necessary to recover the geometry, that contain various combinations of genes from each line segment identified by the fractal analysis. These chromosomes will then be allowed to mutate (swap one gene from a segment with another from the same segment) and recombine (swap contiguous segments). Finally, the chromosomes are subject to selection where superior solutions to the problems are allowed to survive to the next generation. Over multiple generations of mutation, recombination, and selection, the genetic algorithm can find the optimal selection of genes that best solve the problem. The challenge of designing a program for this purpose is a subject for future work.

We view fractal analysis as largely a front end or selection tool for other analyses such as ANNs, Self-Organizing Maps, and cluster analyses. In the future, this may lead to analysis by metabolic control networks or biochemical systems networks approaches outlined by (Voit 2000; Voit and Almeida 2004; Leibovitch and others 2005; Shehadeh and others 2006). For the moment, the potential for fractals to address important questions concerning the application to systems biology is intriguing. In a recent book, Wagner (2005) discussed the notion of a neutral space where organisms can adopt a variety of solutions to the challenges of life, all of which are roughly equivalent. Some parts of the neutral space may be more densely populated than others. We are interested in the potential for fractal geometry to characterize this neutral space and address one of the questions Wagner poses in the epilogue, "Is it possible to infer the global structure of a neutral space from a small sample, a small number of biological systems within that space?" We suspect that fractal analysis of microarray data can describe the size of the space, how it is populated and via the correlation dimension, how it will behave.

Acknowledgments

We thank the organizers of the symposium, Drs Don Mykles and David Towle for inviting our participation. We also thank the Genomics Core Facility at the Hollings Marine Lab (HML) for their diligence in constructing the microarrays (Dr Paul Gross, Director), Dr J. Almeida and Mr D. McKillian for their efforts in constructing the microarray pipeline at www.marinegenomics.org and Dr Greg Warr for simply being himself. We also thank the members of the Marine Infectious Disease Laboratory, Dr C. Browdy director, for assistance in the disease challenge experiments and providing specimens for the research. The project was supported by NOAA/NMFS grant # NA03NMF4720362 (to RWC). J.R. was also supported by Escuela Superior Politécnica del Litoral and Fundación para la Ciencia y la Tecnología, Ecuador, This is contribution number 581 to the Marine Resources Division of the South Carolina Department of Natural Resources, and #31 to the Marine Genomics Group at the Center for Marine Biomedicine and Environmental Science at the Medical Univeristy of South Carolina.

Conflict of interest: None declared.

References

Abraham ER. 2001. The fractal branching of an arborescent sponge. Mar Biol 138:503–10.

Azovsky AI, Chertoprood MV, Kucheruk NV, Rybnikov PV, Sapozhnikov FV. 2000. Fractal properties of spatial distribution of intertidal benthic communities. Mar Biol 136:581–90.

Bellman R. 1961. Adaptive control processes: a guided tour. Princeton, NJ: Princeton University Press.

Brown PO, Botstein D. 1999. Exploring the new world of the genome with DNA microarrays. Nat Genet 21:33–7.

Cazalis D, Milledge T, Narasimhan G. 2004. Probe selection algorithms. Proceedings of the SCI Conference, OR, USA.

Chapman RW. 2001. EcoGenomics—a consilence for comparative immunology? Dev Comp Immunol 25:549–51.

Gross PS, Bartlett TC, Browdy CL, Chapman RW, Warr GW. 2001. Immune gene discovery by expressed sequence tag analysis of hemocytes and hepatopancreas in the Pacific White Shrimp, *Litopenaeus vannamei*, and the Atlantic White Shrimp, *L. setiferus*. Dev Comp Immunol 25:565–77.

Hagan MT, Demuth HB, Beale M. 1996. Neural network design. Boston, MA: Prindle, Weber & Schmidt Pub.

Holland JH. 1975. Adaptation in natural and artificial systems. Ann Arbor: University of Michigan Press.

Khan J, Wei S, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 7:673–9.

Liebovitch L, Toth T. 1989. A fast algorithm to determine fractal dimensions by box counting. Phys Lett 141: 386–90.

Liebovitch LS, Shehadeh LA, Jirsa VK. 2005. Patterns of genetic interactions: analysis of mRNA levels from cDNA microarrays. In: Reeke GN, Poznanski RR, Lindsay KA, Rosenberg JR, Sporns O, editors. Modelling in the neurosciences 2nd edition. Boca Raton, FL: Taylor & Francis. p 9–24.

Lovejoy S, Currie WJS, Tessier Y, Claereboudt MR, Bourget E, Roff JC, Schertzer D. 2001. Universal multifractals and ocean patchiness: phytoplankton, physical fields and coastal heterogeneity. J Plankton Res 23:117–41.

MacAuthur RH, Wilson EO. 1967. The theory of island biogeography. Princeton, NJ: Princeton University Press.

Prior S, Browdy CL, Shepard EF, Laramore R, Parnell PG. 2003. Controlled bioassay systems for determination of lethal infective doses of tissue homogenates containing Taura syndrome or white spot syndrome virus. Dis Aquat Org 54:89–96.

Sarraille J, DiFalco P. 1993. Available from /hpux.connect.org.uk/hppd/hpux/Physics/fd3-0.3/.

Shehadeh LA, Liebovitch LS, Jirsa VK. 2006. Relationship between global structures of genetic networks and mRNA levels measured by cDNA microarrays. Physica 364: 297–314.

Turner SJ, O'Neill RV, Conley W, Conley MR, Humphries HC. 1991. Pattern and scale: statistics for landscape ecology. In: Turner MG, Gardner RH, editors. Ecological studies. Volume 82, Quantitative methods in landscape ecology. Berlin: Springer-Verlag. p 17–49.

Van de Peppel J, Kemmeren P, van Bake H, Radonjic M, van Leenen D, Holstege FCP. 2003. Monitoring global messenger RNA changes in externally controlled microarray experiments. EMBO Rep 4:387–93.

Vapnik V. 1998. Statistical learning theory. New York: Wiley-Interscience.

Voit EO. 2000. Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists. New York: Cambridge University Press.

Voit EO, Almeida J. 2004. Decoupling dynamical systems for pathway identification from metabolic profiles. Bioinformatics 20:1670–81.

Wagner A. 2005. Robustness and evolvability in living systems. Princeton, NJ: Princeton University Press.

Wang L-Y, Balasubramanian A, Chakraborty A, Comaniciu D. 2005. Fractal Clustering for Microarray Data Analysis. Proceedings of the IEEE Computational Systems Bioinformatics Conference. p 97–98.

Waters MD, Olden K, Tennant RW. 2003. Toxicogenomic approach for assessing toxicant-related disease. Mutat Res 544:415–24.

Williams TD, Gensberg K, Minchin SD, Chipman JK. 2003. A DNA expression array to detect toxic stress response in European flounder (*Platichthys flesus*). Aquat Toxicol 65:41–157.

Young RA. 2000. Biomedical discovery with DNA arrays. Cell 102:9–15.