MB&B 452a.
Prof. Mark Gerstein.
Zhang Yang

Microarray Analysis of Host-Pathogen Interactions: An Array Expression Database Enabling Cross-Comparison of Pathogenesis

Introduction

Viral and bacterial pathogenicity depends on the capacity of these infectious agents to adapt to and persist in the host microenvironment. Pathogenesis results from the imbalance between factors that promote pathogen multiplication and host-tissue damages over those needed for microbial removal and protection of host tissues [1]. Although still in their infancy, DNA microarrays have been shown to be highly effective for studying host-pathogen interactions in two ways: 1) identifying mechanisms in microorganisms contributing to pathogenicity, and 2) surveying host responses to infection. Conventional studies have involved elucidating mechanisms behind a small number of genes in individual experiments. Microarrays allow for the simultaneous measurement of the expression patterns of thousands of genes in a genome-wide framework. Functional genomics – such as large-scale expression studies – and computational biology have powerful implications in the study of host-pathogen interactions. Currently, techniques to monitoring whole-genome, tissue-specific responses to different pathogen infections are not readily available. Progresses in the combined use of microbial and host microarrays suggest the potential to uncover host-pathogen dialogue in a gene-by-gene and condition-, site- and time-specific manner.

Methods and Applications

Since 1997 when bacteriophage ϕx174 became the first organism to be fully sequenced, there have been an exponential growth in the amount of sequence information available [2]. As of 2001, 41 microbial genomes having been fully sequenced and published, and more than 120 more genomes are in

the process of being sequenced. The small sizes of viral and bacterial genomes make it relatively easy to manufacture pathogen arrays. Numerous studies have been conducted, for example, using arrays representing nearly all of the virally encoded genes from human cytomegalorius (HCMV) and herpes simplex virus type 1 (HSV-1) [3,4]. A variety of human oligonucleotide microarrays are available as well. Due to the size of the human genome and its complexity, these arrays are composed largely of expressed sequence tags (ESTs) and are on average of more than 50% representation of the predicted human coding regions. Various arrays made based on genomes of other organisms – yeast, C. elegans, and drosophila - have been used as alternative host arrays [5].

Detailed theory and background of microarray technology is not given due to the scope of this paper. This information has been given elsewhere [6]. In summary, different types of microarrays – spotted glass slide microarray and high-density oligonucleotide arrays – are essentially microscopic representations of thousands of different DNA sequences. They allow for the measurement of the relative abundance of DNA or RNA, through which expression profiles can be derived. Clustering methods - hierachical, k-means, fuzzy k-means, neural network, etc – are applied to the raw data such that patterns in gene expression profiles can be observed. The basic assumption behind these clustering approaches is that genes whose transcription profiles are similar are more likely to be found in the same or closely related pathways [7]. This "guilt-by-association" approach is a novel way to identify gene function through uncovering co-expression at specific times or under certain environmental and physiological conditions .

Comparative Genomics of Pathogens.

Pathogen arrays allows for the efficient and global comparison between related genomes. To identify strain-specific genes, Salama *et. al*. compared the genomic contents of 15 different virulent *H. pylori* strains, using a fully sequenced *H. pylori* genome as the reference. He concluded that some 22% of the *H. pylori* genes were dispensable in one or more strains [8]. Similar microarray studies on *E. coli* strains have shown that the degree of diversity within prokaryotic 'species' is far greater than that within animal species [9]. By comparing of isogenic mutant pathogen strains lacking single virulence genes, or virulence factor-associated biologic activities, one can isolate candidate genes for specific virulence attributes, which in turn might yield mechanistic insight into those virulence factors.

Host-Pathogen Interaction

How microorganisms establish pathogenesis in an eukaryotic host is a complex process for which our current understanding is very limited. Effects of viral infections on the host have been studies through microarray for retroviruses, herpesviruses, orthomyxoviruses, enteroviruses, adenoviruses, hepatitis B and C viruses, etc [10, 11, 12, 13]. These studies share the common model of measuring the gene expression of host cells before and after infection and elucidating patterns of host-genes regulation

throughout the stages of infection. Inconsistencies in the functional annotation of host genes and different experimental designs have made it difficult to discern common patterns of host response to pathogens. Researchers found some consistency between patterns of host gene induction and conventional knowledge of ant-infection mechanisms. For example, in the study of human cytomegalovirus infection using human foreskin fibroblasts as the host, 3.9% of the total of 6600 genes arrayed changed their expression level by at least four-fold. Those differentially expressed genes include interferon-inducible genes and protein degradation genes. As expected, researchers observed a close association between viral replication and the up-regulation of certain host transcription, translation, and protein-synthesis genes [14].

Future directions

Cross-comparisons of Viral and Bacterial Microarray Data: Array Expression Database with New Analysis Tools

A global understanding of how the presence of pathogens remodels the host's transcriptome holds the potential of elucidating mechanisms of viral or bacterial infection and host response. Current theory holds that eukaryotic hosts discriminate between and tailor their responses to different forms of infections (a single-stranded mRNA sense viral genome vs. double-stranded DNA genomes, and Gram-positive versus Gram-negative bacteria). There is also an underlying, broad mechanism responding to infections in general [15]. The cross-comparison of viral and bacterial microarray expression data would reveal common trends in infection response as well as distinguish pathogen-specific mechanisms. As of now, due to the lack of adequate analysis tools, drawing crude trends from different microarray data sets is a risky endeavor. For example, inconsistencies in the expression pattern of an interferon-regulating gene across different experiments might indicate that this gene is infection-specific and not part of the general anti-infection mechanism. It is equally possible that this reflects the different time points post-infection at which the microarray experiments were conducted. Varying stages of infections have been known present different pathologies – one could expect analogous differences in expression profiles throughout these stages as well. The inconsistency may also reflect differences in the arrays used (e.g. if the gene of interest is represented) or in the arraying methods (e.g. different reference RNA samples and varying analysis and clustering methods). Moreover, the lack of uniform functional annotation for microbial genes adds more complexity to the cross-comparison of microarray data [16].

The ultimate goal of these cross-comparison studies is to examine those host-pathogen interactions in a gene-by-gene, and tissue-, site-, and time-specific manner. To overcome the limitations previously described, efficient and standardized means of microarray data analysis, data storage, and retrieval systems need to be put in place. An array expression database could be set-up for microbial array studies. To build such a database, expression data would be given an accession number upon

publication and be broadly organized in terms of the model system from which it is derived. A similar approach as the Gene Ontology project could be used to annotate microbial and host genes in three levels - molecular function, biological process, and cellular component [17]. There is an added need to define those genes represented on host arrays with organism-specific tissue information. For example, expression data derived from an infected liver Kupffer cell-line (macrophage cells involved in innate immunity) would be expected to differ significantly from data of a study on the same infection on histiocytes.

For this expression database, specialized analysis tools need to be developed for the unique properties of pathogen genomes, which differ in significant ways from genomes of large, eukaryotic organisms. For example, microbial genomes are known for pathogencity islands and operons that are tightly co-regulated [18]. It is therefore important to correlate genomic structure information with the co-expression of genes. This might present a novel way to identify such islands or operons. Moreover, most current clustering analysis been performed within the scope of the same strain of pathogen under different environmental conditions. Because of the presence of many closely related strains of pathogens, it is important to adapt clustering methods to the study of expression profiles derived from these very similar strains. For example, yellow fever virus and hepatitis C virus present very different pathologies although they are quite closely related [19]. An cross-analysis of differentially expressed genes between these two viruses upon the infection of same host might reveal those virulence factors that lead acute infections as with yellow fever and those that lead to more persistent effects as in hepatitis C. Furthermore, incorporating host and pathogenic genes into the same array – an approach rarely taken – could help determining the coordinated interactions between host and pathogen.

Presently, array expression data are not yet suitable for this kind of large-scale database. Relationships between different array methodologies are not known, and different quantification and clustering methods yield results that cannot be readily cross-compared. The integrated expression analysis of pathogenic genes and host immune-response genes require overcoming these limitations. These sorts of studies – conducted within the scope an expression database – hold the potential of revealing complex processes of pathogenesis and helping to develop anti-infective therapeutics and prevention strategies.

REFERENCES

[1] Tzou P, De Gregorio E, Lemaitre B. How Drosophila combats microbial infection: a model to study innate immunity and host-pathogen interactions. Curr Opin Microbiol. 2002 ;5(1):102-10.

[2] Sanger F, Air GM, Barrell BG, *et al*. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 1977; 265: 687–695.

[3] Chambers J, Angulo A, Amaratunga D, *et al*. DNA microarrays of the complex human cytomegalovirus genome: profiling kinetic class with drug sensitivity of viral gene expression. *J Virol* 1999; 73: 5757–5766.

[4] Stingley SW, Ramirez JJ, Aguilar SA, *et al*. Global analysis of herpes simplex virus type 1 transcription using an oligonucleotide-based DNA microarray. *J Virol* 2000; 74: 9916–9927.

[6] Lipshutz RJ. Applications of high-density oligonucleotide arrays. *Novartis Found Symp*. 2000;229:84-90; discussion 90-3.

[7] Kaminski N, Friedman N. Practical approaches to analyzing results of microarray experiments. *Am J Respir Cell Mol Biol*. 2002; 27(2):125-32.

[8] Joyce EA, Chan K, Salama NR, Falkow S. Redefining bacterial populations: a post-genomic reformation. *Nat Rev Genet*. 2002 Jun;3(6):462-73

[9] Ochman H, Jones IB. Evolutionary dynamics of full genome content in Escherichia coli. *EMBO J*. 2000 Dec 15;19(24):6637-43.

[11] Geiss GK, Bumgarner RE, An MC, *et al*. Large-scale monitoring of host cell gene expression during HIV- 1 infection using cDNA microarrays. *Virology* 2000; 266: 8–16.

[12] Mossman KL, Macgregor PF, Rozmus JJ, Goryachev AB, Edwards AM, Smiley JR. Herpes simplex virus triggers and then disarms a host antiviral response. *J Virol* 2001; 75: 750–758.

[13] Zhu H, Cong JP, Mamtora G, Gingeras T, Shenk T. Cellular gene expression altered by human cytomegalovirus: global monitoring with oligonucleotide arrays. *Proc Natl Acad Sci USA* 1998; 95: 14470–14475.

[14] Chambers J, Angulo A, Amaratunga D, *et al*. DNA microarrays of the complex human cytomegalovirus genome: pro.ling kinetic class with drug sensitivity of viral gene expression. *J Virol* 1999; 73: 5757–5766.

[15] Manger ID, Relman DA. How the host 'sees' pathogens: global gene expression responses to infection. *Curr Opin Immunol*. 2000 ;12(2):215-8. Review.

[16] Kaminski N, Friedman N. Practical approaches to analyzing results of microarray experiments. *Am J Respir Cell Mol Biol*. 2002; 27(2):125-32.

[17] Lewis S, Ashburner M, Reese MG. Annotating eukaryote genomes. *Curr Opin Struct Biol*. 2000;10(3):349-54.

[18] Odenbreit S, Haas R. Helicobacter pylori: impact of gene transfer and the role of the cag pathogenicity island for host adaptation and virulence. *Curr Top Microbiol Immunol*. 2002; 264(2):1-22.

[19] Monath TP. Yellow fever: an update. *Lancet Infect Dis*. 2001