7

Database Mining: Defining the Pathogenesis of Inflammatory and Immunological Diseases

Fan Yang, Irene Hwa Yang, Hong Wang and Xiao-Feng Yang
Department of Pharmacology, Cardiovascular Research Center, Temple University School of Medicine, Philadelphia, U.S.A.

1. Introduction

Cardiovascular disease (CVD) is a leading cause of mortality in developed countries (Jan et al., 2010; Yang et al., 2008). Despite a long held understanding and strong characterization of the traditional and non-traditional risk factors for CVD, some mechanisms of CVD onset have only recently been uncovered. As a chronic inflammatory autoimmune disease, atherosclerosis and its progression involve innate and adaptive immune systems. Using new concepts and technologies to improve the current understandings of the molecular pathogenesis of inflammatory and immune responses would lead to the future development of novel therapeutics for these diseases.

Biomedical literature and databases, available in electronic forms, contain a vast amount of knowledge resulting from experimental research (Ishii et al., 2007; Palakal et al., 2007). In the past decade, both traditional hypothesis-driven research and discovery-driven "-omics" research, including genomics, transcriptomics (Liang et al., 2005), proteinomics, metabolomics, glycomics, lipidomics, localizomics, protein-DNA interactomics, protein-protein interactomics, fluxomics, phenomics (Joyce & Palsson, 2006), and antigen-omics (http://www.cancerimmunity.org/links/databases.htm) (Houle et al., 2010; Shimokawa et al., 2010; Weinstein, 1998;2002), has generated a tremendous amount of data and established many experimental data-based searchable databases. These databases include PubMed, nucleotide database, protein database, and other databases generated by the National Institutes of Health (NIH)/National Center for Biotechnology Information (NCBI) (see the NCBI handbook at http://www.ncbi.nlm.nih.gov/books/NBK21101/) and other institutions. This development has not only provided resources, but also raised unprecedented challenges and opportunities for biomedical scientists to develop more systemic and panoramic approaches to analyze the data contained in the databases and generate new hypotheses. The inconsistency between the vast amount of experimental data, various searchable databases, and relatively smaller numbers of database-mining research papers (< 50 papers on database mining in inflammation and immune responses listed in the PubMed) indicate the challenges that experimental biomedical scientists face, which include both technical/methodological difficulties and out-of-date concepts.

Traditionally, medical literature search using the Index Medicus was the major approach for biomedical scientists to identify knowledge gaps and preparing new hypotheses. However, this approach has been significantly enhanced by more systemic approaches such as *1)*

NCBI-PubMed search and Google Scholar search; *2)* experimentally screening cDNA libraries and various arrays (nucleic acid arrays, antibody arrays, protein arrays and metabolic arrays) (King et al., 2005; Loza et al., 2007; Pandey et al., 2004; Warner & Dieckgraefe, 2002); and *3)* mining experimental databases (Chen et al., 2010; Jan et al., 2010; Ng et al., 2004; Yang et al., 2006a; Yang et al., 2006b; Yin et al., 2009). The screening analysis of microarray data often requires bioinformatic methods, algorithms, and expertise. In comparison, database mining offers many advantages. First, database mining requires much less bioinformatic assistance in each laboratory when compared to the generation of algorithms required in microarray analyses, since the purpose of generating databases is to use bioinformatic approaches to mine easily organize the experimental data for biomedical scientists to mine (Spasic et al., 2005). Second, database mining enables full-value extraction from costly experimental data, and third, it provides panoramic analyses on existing knowledge gaps by generating new hypotheses for further experimental research. However, database mining requires biomedical scientists to have more conceptual advances than technical assistances. The purpose of database mining is to analyze experimental data deposited by various research projects, rather than predicting theoretic results based on pure theoretical bioinformatic studies. Thus, database mining is not limited to sequence comparisons of nucleic acids and proteins (Mount, 2004), sequence alignments, analysis of hydrophobicity index and functional domain prediction of proteins. Additionally, database mining has not generally been listed as a required course for graduate and postdoctoral studies, which presents a challenge of properly training young biomedical scientists with essential database mining techniques. On top of these aforementioned challenges, reviewers from peer-reviewed database mining publications often mistakenly regard the experimental data in electronic forms deposited in databases as "non-experimental or theoretical" and demand ridiculous additional verifying experiments to be performed, even requiring the use of outdated experimental techniques or methods. To overcome these difficulties, bioinformatic scientists will have to work together with biomedical colleagues and delve into the biological significance of database mining projects, rather than sticking to an argument of "no algorithms means no bioinformatics". Already, more and more database mining papers have been published as scientists put aside their differences. For example, the 2011 (18th) database issue of the journal "Nucleic Acid Research" features descriptions of 96 new and 83 updated online databases covering various areas of molecular biology (Galperin & Cochrane, 2011). The Nucleic Acids Research online Database Collection, available at: http://www.oxfordjournals.org/nar/database/a/, now lists 1330 carefully selected molecular biology databases. In addition, 32 databases and analysis resources of immunological interest have been established (Salimi et al., 2010). Moreover, our recent invited review lists 11 B cell antigen epitope databases and 13 T cell antigen epitope analysis resources (Jan et al., 2010). These progresses suggest that a data mining approach has gradually been accepted as mainstream practice in analyzing experimental data and generating new hypotheses for various projects (Salimi et al., 2010).

Our lab has successfully pioneered major advances in database mining in the fields of adaptive immune reactions, innate immune responses, and inflammation (Chen et al., 2010; Jan et al., 2010; Ng et al., 2004; Virtue, 2011; Yang et al., 2006a; Yang et al., 2006b; Yin et al., 2009). In this chapter, we will summarize the general approaches, principles, and databases used and new working models proposed in our database mining research. This discussion will prove to be important and useful for most biomedical scientists, since many are not

often involved in the bioinformatic algorithm generation, but may want to use database mining methods in their research either as parts of existing experimental studies or as free-standing projects. Of note, the database mining concept is not "brand new". Medical research has a long history in full-value extraction from costly data. For example, a meta-analysis uses a statistical approach to combine the results of several epidemiological studies that address a set of related research hypotheses. This practice started well over 100 years ago and has been widely used in various disease-related researches (http://en.wikipedia.org/wiki/Meta-analysis) (Egger & Smith, 1997; Egger et al., 1997). We believe that the practice of database mining will become a routine exercise to identify existing knowledge gaps and to generate new hypotheses.

2. Principles of database mining

In recent years, many databases regarding immune responses and inflammation have been established (Jan et al., 2010; Yang et al., 2006a), which have expanded the scope and depth of a publicly searchable online repertoire of tools. The results derived from the database mining analyses have become parts of many research papers or free-standing papers. Although projects may vary in format, database mining approaches follow the same set of principles (Fig. 1): 1) Hypothesis: A clearly-presented hypothesis based on the current biomedical literature search in a given field and previous experimental data in the lab is required to carry on a database mining project as we reported (Ng et al., 2004; Yan et al., 2004), which is similar to that of experimental projects. Of note, the database mining referred here focuses on database mining as a free standing project rather than as a part of experimental research; 2) Scope: Database mining scopes in terms of gene numbers are far more than that examined in experimental approaches. For example, our own research will examine mRNA transcript expressions of about 30 genes including all the reported toll-like receptors, NOD-like receptors, and inflammatory caspases in more than ten tissues. This scope allows us to obtain a panoramic view on the expressions of inflammatory pathways without focusing on a single gene in many tissues (Yin et al., 2009); 3) Suitable databases: Databases that are suitable for examining the hypothesis are available for online analytic search, which is also similar to the methods and reagents for experimental projects; 4) Sizable experimentally verified data for generating confidence intervals with statistical significance: To consolidate the results generated from database mining, the experimentally verified data are published by various laboratories, which can be used to generate statistically significant confidence intervals by using the same online analysis tools as we reported (Virtue, 2011). In this study, our analysis in the TargetScan yielded 524 microRNAs, which were predicted to participate in 1368 unique interactions with the 33 inflammatory gene mRNAs. To ensure relevance, we examined the context value and percentage of experimentally verified microRNAs. Confidence intervals were generated from 45 interactions between 28 experimentally verified human microRNAs and 36 genes found within the Tarbase, an online database of experimentally verified microRNAs (http://diana.cslab.ece.ntua.gr/tarbase/) (Papadopoulos et al., 2009; Sethupathy et al., 2006). These experimental interactions were also selected based on their confirmation by luciferase reporter assays and single site specificity. The 45 microRNA-mRNA interactions that met these criteria were then evaluated in TargetScan to determine the microRNA

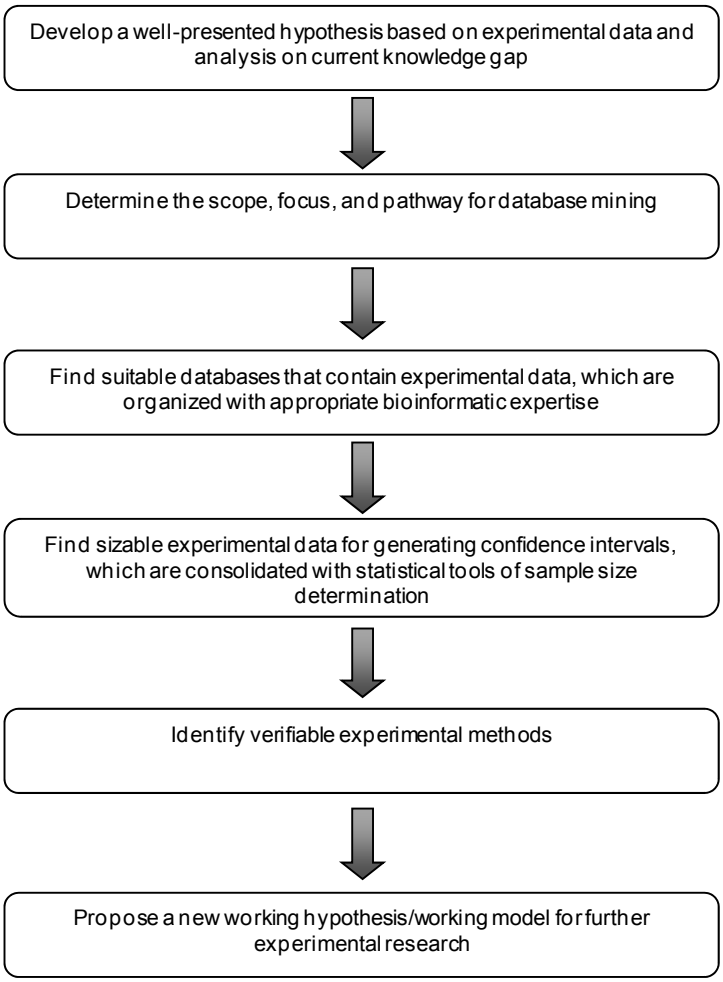


Fig. 1. Detabase mining flow-chart and principles.

context values and percentages. Analysis of this data yielded a mean and standard deviation (SD) of -0.25 ± 0.12 and 76.07 ± 19.07 for context value and context percentage, respectively. The intervals were then constructed and the lower limits (the mean - 2 x standard deviations) were calculated for context percentage (76.07-1.96 (19.07/SQRT (46)) = 76.07 - 5.51 = 70.56) and context value (-0.25-1.96(0.12/SQRT (46) = - 0.25 - 0.04= -0.22). All predicted microRNAs interactions with a context value ≤-0.22 and context percentage ≥70 were accepted. Using the lower limit thresholds for context value and percentage, 297 out of the 524 predicted microRNAs met the criteria and were considered equivalent to the experimentally verified microRNAs. In order to generate valid confidence intervals, sample sizes have to be estimated with statistical tools of sample size determination (Rosner, 2000) as we reported (Ng et al., 2004); *5)* Verifiable methods: Experimental methods are available

to verify the data generated by the database mining (Yan et al., 2004); and *6)* A new working model/hypothesis: Through database mining, a new knowledge gap will be identified, and a new hypothesis will be proposed to test fewer, much more-focused genes in further experiments. The following sections will illustrate these principles in our own publications (Chen et al., 2010; Jan et al., 2010; Ng et al., 2004; Virtue, 2011; Yang et al., 2006a; Yang et al., 2006b; Yin et al., 2009).

3. Database mining example 1: Stimulation-responsive alternative splicing is an important mechanism in generating self-antigen epitopes (Ng et al., 2004; Xiong et al., 2006; Yan et al., 2004; Yang et al., 2006a; Yang, 2007)

In our invited review, we pointed out that the identification and molecular characterization of self-antigens expressed by human malignancies, that are capable of elicitation of anti-tumor immune responses in patients, have been an active field in tumor immunology (Yang & Yang, 2005). More than 2,000 tumor antigens have been identified, and most of these antigens are self-antigens (Yang & Yang, 2005). Despite this, the important question of how non-mutated self-protein antigens, generated from normal cells and tumor cells, gain immunogenicity and trigger immune recognition remained unanswered (Yang & Yang, 2005). Mutations may be responsible for some aspects of elevated immunogenicity underlying certain tumor-specific antigens (p53 and Ras), while chromosome translocations and abnormalities, such as expression of the fusion oncogene Bcr-Abl in chronic myelogenous leukemia (Clark et al., 2001; Pinilla-Ibarz et al., 2000; Yotnda et al., 1998; Zorn, 2001) (Yang et al., 2002; Yang et al., 2001) are responsible for other aspects. However, the mechanism underlying the immunogenicity of most non-mutated self-tumor antigens is their aberrant overexpression in tumors (Yang & Yang, 2005). Zinkernagel *et al* (Zinkernagel & Hengartner, 2001) suggested that the overexpression of self-antigens or novel antigenic structure, overcomes the threshold of antigen concentration at which an immune response is initiated (Shlomchik et al., 2001). This threshold might be lower for certain untolerized regions of certain antigen epitopes. Overexpressed genes, often encode tumor antigens up to 100 fold. These genes are identified by serological identification of self-antigens by screening a cDNA library with patients' sera (SEREX) (Sahin et al., 1995), which may reflect the inherent methodological bias for the detection of abundant transcript (Preuss et al., 2002). The overexpression of tumor antigens in tumors may result from transcriptional and post-transcriptional mechanisms. We recently demonstrated that overexpression of tumor antigen CML66L in leukemia cells and tumor cells via alternative splicing is the mechanism for its immunogenicity in patients with tumors (Yan et al., 2004; Yang et al., 2001). This not only illustrates the principle of overexpression of tumor antigen, but also elucidated alternative splicing as its molecular mechanism (Yan et al., 2004). A significant proportion of the SEREX-defined self-tumor antigens are autoantigens (Chen, 2004), for example, CML28 that we identified is autoantigen Rrp46p (Yang et al., 2002). Using this information gathered from SEREX, we hypothesized that alternative splicing is a general mechanism for the overexpression of untolerized self-antigen epitopes in tumors and autoimmune diseases. In order to test this hypothesis, we database mined the NIH-NCBI AceView database to examine the potential mechanisms of how non-mutated self-proteins gain new untolerized structures that trigger immune recognition (Ng et al., 2004). The AceView database provides a curated, comprehensive, and non-redundant sequence representation of all public mRNA sequences (mRNAs from GenBank or RefSeq, and single pass cDNA sequences from dbEST and Trace). These experimental cDNA sequences are first

co-aligned on the genome, and then clustered into a minimal number of alternative transcript variants and grouped into genes (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/). Our results showed that alternative splicing occurs in 100% of autoantigen transcripts. This is significantly higher than the approximately 42% rate of alternative splicing observed in the 9554 randomly selected human gene transcripts (p<0.001). Within the isoform-specific regions of the autoantigens, 92% and 88% encoded MHC class I and class II-restricted T-cell antigen epitopes, respectively, and 70% encoded antibody binding domains. Alternative splicing can be canonical or non-canonical. Canonical splicing removes introns that have 5'GT and 3'AG consensus flanking sequences (GT-AG rule) (Lewin, 2000). Our results demonstrated that 80% of the autoantigen transcripts undergo non-canonical alternative splicing, which is significantly higher than the less than 1% rate in randomly selected gene transcripts (p<0.001). These studies suggest that non-canonical alternative splicing may be an important mechanism for the generation of untolerized epitopes that may lead to autoimmunity. Furthermore, the product of a transcript that does not undergo alternative splicing is unlikely to be a target antigen in autoimmunity (Ng et al., 2004). To consolidate this finding, we also examined the effect of proinflammatory cytokine tumor necrosis factor-α (TNF-α) on the prototypic alternative splicing factor (ASF)/SF2 in the splicing machinery. Our results show that TNF-α downregulates ASF/SF2 expression in cultured muscle cells. This result correlates with our finding of reduced expression of ASF/SF2 in inflamed muscle cells from patients with autoimmune myositis (Xiong et al., 2006). Based on our and others' data, we recently proposed a new model of stimulation-responsive splicing for the selection of autoantigens and self-tumor antigens (Yang et al., 2006a) [also see Fig. 1 at (http://preview.ncbi.nlm.nih.gov/pubmed/16890493)]. Our new model theorizes that the significantly higher rates of alternative splicing of autoantigen and self-tumor antigen transcripts that occur in response to stimuli, such as proinflammatory cytokines, could induce extra-thymic expression of untolerized antigen epitopes to elicit autoimmune and anti-tumor responses. By using B lymphocyte (B cell) antigen epitope analysis databases and T cell antigen epitope analysis databases listed in Tables in our recent invited review (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2858284/pdf/JBB2010-459798.pdf) (Jan et al., 2010), we showed that protein sequences encoded by alternatively spliced exons are sufficient to equip antibody-binding antigen epitopes and major histocompatibility complex (MHC) class I- and MHC II-restricted T cell antigen epitopes to stimulate B lymphocytes and T lymphocytes, respectively (Ng et al., 2004). Of note, our model not only applies to non-mutated self-tumor antigens associated tumors and autoantigens associated with various autoimmune diseases, but also to the composition and expansion of the self-antigen repertoire of stem cells. Our additional database mining study has generated a new model of differential epitope processing for MHC class I-restricted viral antigen epitopes and tumor antigen epitopes (Yang et al., 2006b). Our reports have demonstrated the principles of database mining in adaptive immune responses.

4. Database mining example 2: Three-tier model for inflammasome/caspase-1 activation and inflammation privilege of tissues are important mechanisms underlying the differences in the readiness of inflammation initiation in tissues

Atherosclerosis is the leading cause of morbidity and mortality in industrialized society. Several "traditional" risk factors have been identified for atherosclerosis including

hyperlipidemia, oxidized low density lipoprotein, cigarette smoking, diabetes, hypertension, obesity (Ross, 1992), and hyperhomocysteinemia (HHcy), etc. Chronic vascular inflammation is an essential requirement for the progression of atherosclerosis in patients (Hansson, 2005). Recent progress in characterizing pathogen-associated molecular patterns' (PAMPs) receptor families (PAMP-Rs) and inflammasomes (the protein complex for activation of caspase-1) has further emphasized the importance of proinflammatory cytokine interleukin-1β (IL-1β) signaling in bridging proatherogenic risk factors to initiate inflammation (Yang et al., 2008). However, constitutive expression levels and expression readiness of PAMP-Rs, inflammasome components and proinflammtory caspases in tissues remained poorly defined. We hypothesized that PAMP-Rs, inflammasome components, proinflammatory caspases, IL-1, and IL-18 are differentially expressed in cardiovascular tissues. To examine this hypothesis, we mined the NCBI-UniGene database, analyzed cDNA cloning and DNA sequencing data from tissue cDNA libraries and studied expression profiles of Toll-like receptors (TLRs), cytosolic nucleotide binding and oligomerization domain (NOD)-like receptors (NLRs), inflammasome components, inflammatory caspases, and caspase-1 cleavable inflammatory cytokines. The UniGene database provides an organized view of the transcriptome with information on protein similarities, gene expression, cDNA clone reagents, and genomic location (http://www.ncbi.nlm.nih.gov/unigene), in which each UniGene entry is a set of transcript sequences that appear to come from the same transcription locus (gene or expressed pseudogene). After analyzing the data from the UniGene database, we made several important findings: (1) Among 11 tissues examined, vascular tissues and heart express fewer types of TLRs and NLRs than immune system tissues including blood, lymph nodes, thymus, and trachea; (2) Brain, lymph nodes, and thymus do not express proinflammatory cytokines IL-1β and IL-18 constitutively, suggesting that these two cytokines need to be upregulated in response to inflammatory stimuli in the tissues; and (3) based on the expression data of three characterized inflammasomes (NALP1, NALP3 and IPAF inflammasomes), the examined tissues can be classified into three tiers: the first tier tissues including brain, placenta, blood, and thymus express inflammasome(s) in constitutive status; the second tier tissues have inflammasome(s) in nearly-ready expression status (with the requirement of upregulation of one component); and the third tier tissues like heart and bone marrow, require upregulation of at least two components in order to assemble functional inflammasomes. Based on the expression readiness of inflammasomes in tissues, we propose a new working model of three-tier responsive expression of inflammasomes in tissues and suggest a new concept of third tier tissues' inflammatory privilege, which provides an insight on the differences of tissues in initiating acute inflammations. This model suggests that *(a)* first-tier tissues with constitutively expressed inflammasomes initiate inflammation quicker than second and third-tier tissues; and *(b)* second tier tissues (requiring one component of upregulation) including vascular tissue, and third tier tissues including heart (requiring more than one component upregulation) are in an inducible expression status of inflammasomes. The inducible expressions of inflammasomes are presumably mediated through various signal pathways that initiate inflammation, and the interplay between the signal pathways, may take a longer time and overcome a higher threshold than first tier tissues. Traditional concepts of immune privilege suggests a protective mechanism from autoimmune destruction based on the lack of expression of antigen-presenting self-major compatibility complex (MHC) molecules in tissues (Yang & Yang, 2005). The lack of expression of self-MHCs in immune privileged tissues including

testis results in the failure of self-antigen presentation that stimulates the hosts' immune system, thereby protecting immune privileged tissues from autoimmune destruction. Similarly, we proposed a new concept of tissues' inflammatory privileges that emphasize a protective mechanism against tissue destruction mediated by inflammasome/IL-1β-based innate immune responses. In our new concept of tissues' inflammatory privilege, vascular tissue and heart disproportionally express fewer types of TLRs and NLRs and may only inducibly express inflammasomes, thus preventing against uncontrolled inflammatory destruction mediated by inflammasome-based innate immune responses (Streilein & Stein-Streilein, 2000). Our new concept and model may also explain the potential differences between cardiovascular tissues and other tissues in initiating acute inflammation. The first-tier tissues may have a higher probability of experiencing acute inflammation than the second-tier and third-tier tissues.

We and others showed that elevated levels of plasma homocysteine (Hcy), termed hyperhomocysteinemia (HHcy), is an independent risk factor, equivalent to hyperlipidemia, for cardiovascular diseases (CVD) including coronary heart disease and stroke (Maron & Loscalzo, 2009; Wang et al., 2003; Zhang et al., 2009). Recently, we performed an additional database mining study using to examine the expression of more than 20 homocysteine metabolic enzymes and methylation enzymes in >20 tissues in humans and mouse (Chen et al., 2010). We generated a new model of how hypomethylation (a post-translational protein modification) modulates the expressions of homocysteine-metabolizing enzymes (Chen et al., 2010). Taken together, our studies have demonstrated the principles of database mining in innate immune reactions.

5. Database mining example 3: A group of anti-inflammatory microRNAs may play critical roles in inhibiting the expression of proatherogenic molecules

Previous research has established that numerous genes are upregulated in atherogenesis through epigenetic or genetic transcriptional mechanisms (Turunen et al., 2009). However, transcription-independent mechanisms have received far less scrutiny. Recent publications suggest that microRNAs, a newly characterized class of short (18-24 nucleotide long), endogenous, non-coding RNAs (Bartel, 2009), contribute to the development of particular disease states by regulating diverse biological processes such as cell growth, differentiation, proliferation, and apoptosis (Zhang, 2008). This biological control is accomplished by post-transcriptional gene silencing (Naeem et al., 2010) through Watson and Crick base-pairing predominately at the 3'-untranslated region (3'UTR) of messenger RNAs (mRNAs) (Cordes et al., 2009; Rasmussen et al., 2010). This pairing can be further characterized as "perfect" or "near perfect", leading to target mRNA cleavage and degradation, or "imperfect", causing the inhibition of mRNA translation (Naeem et al., 2010). With the identification and sequencing of more than 800 human microRNAs thus far, it is thought that up to 30% of human genes may be regulated by microRNAs (Cheng et al., 2010; Zhang, 2008). Supporting evidence suggests that microRNAs function as key players during critical stages of cellular development and finely tune gene expression in the maintenance of routine cellular functioning (Baek et al., 2008). Furthermore, microRNAs can act on transcription factors, which lead to a broad indirect cellular effect as a result of their widespread gene modulating nature. In addition, the recent research has demonstrated that changes in microRNAs expression patterns are connected to several pathological conditions including cardiovascular disease and atherosclerosis. These studies primarily focused on

characterizing microRNAs in atherosclerosis disease models, which had been previously reported to have elevated expression in disease conditions (Haver et al., 2010; Rink & Khanna, 2010). Thus, current microRNAs research has failed to provide a panoramic view of how microRNAs regulate proatherogenic inflammatory genes in a panoramic view and whether upregulation of proatherogenic inflammatory genes is the result of anti-inflammatory microRNA downregulation. To address these issues, we hypothesized that a group of anti-inflammatory microRNAs may regulate the expressions of proatherogenic molecules (Virtue, 2011). We then developed a novel database mining approach using three types of databases including the online microRNA target prediction software TargetScan (http://www.targetscan.org/) (Dong et al., 2010; Rosero et al., 2010; Vickers & Remaley, 2010), the Tarbase, an online database of experimentally verified microRNAs (http://diana.cslab.ece.ntua.gr/tarbase/) (Papadopoulos et al., 2009; Sethupathy et al., 2006), and the online microRNA.org expression database (http://www.microrna.org/microrna/home.do) (Betel et al., 2008), in concert with a statistical analysis strategy established in our previous database mining publications (Chen et al., 2010; Ng et al., 2004; Shen et al., 2010; Yang et al., 2006b; Yin et al., 2009). Our unique research using database mining yielded several key findings. First, we discovered that the expression of 33 inflammatory genes (mRNAs) is upregulated in atherosclerotic lesions and second, that the mRNAs of those genes contain structural features in their 3'UTR for potential regulation by microRNAs. Furthermore, these structural features are statistically identical to experimentally verified 3'UTR microRNAs binding sites. Third, 21 out of the 33 inflammatory genes (64%) are targeted by highly expressed microRNAs while the remaining 12 inflammatory genes (36%) are targeted by normally expressed microRNAs. Fourth, it was also established that 10 of the 21 highly expressed microRNA-targeted inflammatory genes (48%) were targeted by a single microRNA, suggesting the specificity of microRNA regulation. Meanwhile, 12 out of the 25 highly expressed microRNAs (48%) targeted single inflammatory genes while the other 13 microRNAs targeted multiple inflammatory genes. Finally, it was determined that the microRNAs targeting atherosclerotic inflammatory genes use statistically higher numbers of "poorly conserved" binding interactions than the control group of microRNAs from the confidence interval. These results suggest that the microRNAs regulating atherosclerotic inflammatory genes possess special features (Virtue, 2011).

Previous research has shown that microRNAs participate in modulating atherosclerosis-related processes including hyperlipidemia (microRNA-33, microRNA-125a-5p), hypertension (microRNA-155), plaque rupture (microRNA-222, microRNA-210), and atherosclerosis itself (microRNA-21, microRNA-126) (Rink & Khanna, 2010). However, whether certain microRNAs play a role in preventing the disease development remains unknown. One of the most interesting findings from our study is that the 25 microRNAs that are highly expressed under normal untreated conditions target 21 out of the 33 (64%) atherosclerosis-upregulated inflammatory genes. The important result suggests a novel mechanism where a group of highly expressed anti-inflammatory microRNAs suppress the upregulation of proatherogenic inflammatory genes under normal physiological conditions. It has been well established that microRNAs play important roles in fine-tuning developmental processes and participate in the development of diseases such as cancer. Our results are the first to suggest that microRNAs may play a protective role by suppressing proatherogenic genes to maintain healthy arteries. Our conclusion is supported by other publications, which show that 7 out of the 20 microRNAs identified in this study were

downregulated in the experimental studies by various proatherogenic factors (Chen et al., 2009; Elia et al., 2009; Ji et al., 2007). Together, our studies have demonstrated the principles of database mining in inflammation.

6. Conclusion

Active research in human and mouse genomes, transcriptomes, microRNAs transcriptomes, proteomes, and antigen-omes in the past decade has generated a tremendous amount of data and established many experimental data-based searchable databases. This provides unprecedented opportunities for biomedical scientists to develop more systemic and panoramic approaches to analyze the databases and generate new hypotheses. In this chapter, we briefly summarize our pioneering efforts in using our new database mining methods to address important questions in inflammatory and immunological diseases. The new principles and basic methodologies of database mining developed in our laboratories are elucidated in the following studies: 1) stimulation-responsive alternative splicing model for the generation of untolerized autoantigen epitopes; 2) a three-tier model for inflammasome/caspase-1 activation and inflammatory privileges of tissues; and 3) a group of anti-inflammatory microRNAs in inhibiting proatherogenic gene expression during atherogenesis. With recent technological breakthroughs, database mining has provided significant new insights and hypotheses in specifying the novel directions for experimental research.

7. Acknowledgements

This work was partially supported by the National Institutes of Health Grants HL094451 and HL108910 (XFY), HL67033, HL82774, and HL77288 (HW). FY and IHY contribute equally to this work. Correspondence: Prof. Yang at xfyang@temple.edu.
Disclosures: none declared.

8. References

Baek, D., J. Villen, C. Shin, F. D. Camargo, S. P. Gygi, and D. P. Bartel. "The Impact of MicroRNAs on Protein Output." *Nature* 455, no. 7209 (2008): 64-71.

Bartel, D. P. "MicroRNAs: Target Recognition and Regulatory Functions." *Cell* 136, no. 2 (2009): 215-33.

Betel, D., M. Wilson, A. Gabow, D. S. Marks, and C. Sander. "The MicroRNA.Org Resource: Targets and Expression." *Nucleic Acids Res* 36, no. Database issue (2008): D149-53.

Chen, N. C., F. Yang, L. M. Capecci, Z. Gu, A. I. Schafer, W. Durante, X. F. Yang, and H. Wang. "Regulation of Homocysteine Metabolism and Methylation in Human and Mouse Tissues." *Faseb J* 24, no. 8 (2010): 2804-17.

Chen, T., Z. Huang, L. Wang, Y. Wang, F. Wu, S. Meng, and C. Wang. "MicroRNA-125a-5p Partly Regulates the Inflammatory Response, Lipid Uptake, and Orp9 Expression in oxLDL-Stimulated Monocyte/Macrophages." *Cardiovasc Res* 83, no. 1 (2009): 131-9.

Chen, YT. "Serex Review." *Cancer Immunity* http://www.cancerimmunity.org/SEREX/ (2004).

Cheng, Y., N. Tan, J. Yang, X. Liu, X. Cao, P. He, X. Dong, S. Qin, and C. Zhang. "A Translational Study of Circulating Cell-Free MicroRNA-1 in Acute Myocardial Infarction." *Clin Sci (Lond)* 119, no. 2 (2010): 87-95.

Clark, R. E., I. A. Dodi, S. C. Hill, J. R. Lill, G. Aubert, A. R. Macintyre, J. Rojas, A. Bourdon, P. L. Bonner, L. Wang, S. E. Christmas, P. J. Travers, C. S. Creaser, R. C. Rees, and J. A. Madrigal. "Direct Evidence That Leukemic Cells Present HLA-Associated Immunogenic Peptides Derived from the Bcr-Abl B3a2 Fusion Protein." *Blood* 98, no. 10 (2001): 2887-93.

Cordes, K. R., N. T. Sheehy, M. P. White, E. C. Berry, S. U. Morton, A. N. Muth, T. H. Lee, J. M. Miano, K. N. Ivey, and D. Srivastava. "MiR-145 and MiR-143 Regulate Smooth Muscle Cell Fate and Plasticity." *Nature* 460, no. 7256 (2009): 705-10.

Dong, H., M. Paquette, A. Williams, R. T. Zoeller, M. Wade, and C. Yauk. "Thyroid Hormone May Regulate Mrna Abundance in Liver by Acting on Microrna." *PLoS One* 5, no. 8 (2010).

Egger, M., and G. D. Smith. "Meta-Analysis. Potentials and Promise." *Bmj* 315, no. 7119 (1997): 1371-4.

Egger, M., G. D. Smith, and A. N. Phillips. "Meta-Analysis: Principles and Procedures." *Bmj* 315, no. 7121 (1997): 1533-7.

Elia, L., M. Quintavalle, J. Zhang, R. Contu, L. Cossu, M. V. Latronico, K. L. Peterson, C. Indolfi, D. Catalucci, J. Chen, S. A. Courtneidge, and G. Condorelli. "The Knockout of MiR-143 and -145 Alters Smooth Muscle Cell Maintenance and Vascular Homeostasis in Mice: Correlates with Human Disease." *Cell Death Differ* 16, no. 12 (2009): 1590-8.

Galperin, M. Y., and G. R. Cochrane. "The 2011 Nucleic Acids Research Database Issue and the Online Molecular Biology Database Collection." *Nucleic Acids Res* 39, no. Database issue (2011): D1-6.

Hansson, G. K. "Inflammation, Atherosclerosis, and Coronary Artery Disease." *N Engl J Med* 352, no. 16 (2005): 1685-95.

Haver, V. G., R. H. Slart, C. J. Zeebregts, M. P. Peppelenbosch, and R. A. Tio. "Rupture of Vulnerable Atherosclerotic Plaques: MicroRNAs Conducting the Orchestra?" *Trends Cardiovasc Med* 20, no. 2 (2010): 65-71.

Houle, D., D. R. Govindaraju, and S. Omholt. "Phenomics: The Next Challenge." *Nat Rev Genet* 11, no. 12 (2010): 855-66.

Ishii, N., K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P. Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori, and M. Tomita. "Multiple High-Throughput Analyses Monitor the Response of E. Coli to Perturbations." *Science* 316, no. 5824 (2007): 593-7.

Jan, M., S. Meng, N. C. Chen, J. Mai, H. Wang, and X. F. Yang. "Inflammatory and Autoimmune Reactions in Atherosclerosis and Vaccine Design Informatics." *J Biomed Biotechnol* 2010 (2010): 459798.

Ji, R., Y. Cheng, J. Yue, J. Yang, X. Liu, H. Chen, D. B. Dean, and C. Zhang. "MicroRNA Expression Signature and Antisense-Mediated Depletion Reveal an Essential Role of Microrna in Vascular Neointimal Lesion Formation." *Circ Res* 100, no. 11 (2007): 1579-88.

Joyce, A. R., and B. O. Palsson. "The Model Organism as a System: Integrating 'Omics' Data Sets." *Nat Rev Mol Cell Biol* 7, no. 3 (2006): 198-210.

King, J. Y., R. Ferrara, R. Tabibiazar, J. M. Spin, M. M. Chen, A. Kuchinsky, A. Vailaya, R. Kincaid, A. Tsalenko, D. X. Deng, A. Connolly, P. Zhang, E. Yang, C. Watt, Z. Yakhini, A. Ben-Dor, A. Adler, L. Bruhn, P. Tsao, T. Quertermous, and E. A. Ashley. "Pathway Analysis of Coronary Atherosclerosis." *Physiol Genomics* 23, no. 1 (2005): 103-18.

Lewin, Benjamin. "Nuclear Splicing." In *Genes Vii*, edited by Benjamin Lewin. Cambridge: Oxford University Press Inc., New York, 2000.

Liang, M., A. W. Cowley, Jr., M. J. Hessner, J. Lazar, D. P. Basile, and J. L. Pietrusz. "Transcriptome Analysis and Kidney Research: Toward Systems Biology." *Kidney Int* 67, no. 6 (2005): 2114-22.

Loza, M. J., C. E. McCall, L. Li, W. B. Isaacs, J. Xu, and B. L. Chang. "Assembly of Inflammation-Related Genes for Pathway-Focused Genetic Analysis." *PLoS One* 2, no. 10 (2007): e1035.

Maron, B. A., and J. Loscalzo. "The Treatment of Hyperhomocysteinemia." *Annu Rev Med* 60 (2009): 39-54.

Mount, DW. "Historical Introduction and Overview." In *Bioinformatics. Sequence and Genome Analysis*, edited by DW Mount, 1-27. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press, 2004.

Naeem, H., R. Kuffner, G. Csaba, and R. Zimmer. "Mirsel: Automated Extraction of Associations between Micrornas and Genes from the Biomedical Literature." *BMC Bioinformatics* 11 (2010): 135.

Ng, B., F. Yang, D. P. Huston, Y. Yan, Y. Yang, Z. Xiong, L. E. Peterson, H. Wang, and X. F. Yang. "Increased Noncanonical Splicing of Autoantigen Transcripts Provides the Structural Basis for Expression of Untolerized Epitopes." *J Allergy Clin Immunol* 114, no. 6 (2004): 1463-70.

Palakal, M., J. Bright, T. Sebastian, and S. Hartanto. "A Comparative Study of Cells in Inflammation, Eae and Ms Using Biomedical Literature Data Mining." *J Biomed Sci* 14, no. 1 (2007): 67-85.

Pandey, R., R. K. Guru, and D. W. Mount. "Pathway Miner: Extracting Gene Association Networks from Molecular Pathways for Predicting the Biological Significance of Gene Expression Microarray Data." *Bioinformatics* 20, no. 13 (2004): 2156-8.

Papadopoulos, G. L., M. Reczko, V. A. Simossis, P. Sethupathy, and A. G. Hatzigeorgiou. "The Database of Experimentally Supported Targets: A Functional Update of Tarbase." *Nucleic Acids Res* 37, no. Database issue (2009): D155-8.

Pinilla-Ibarz, J., K. Cathcart, and D. A. Scheinberg. "CML Vaccines as a Paradigm of the Specific Immunotherapy of Cancer." *Blood Rev* 14, no. 2 (2000): 111-20.

Preuss, K. D., C. Zwick, C. Bormann, F. Neumann, and M. Pfreundschuh. "Analysis of the B-Cell Repertoire against Antigens Expressed by Human Neoplasms." *Immunol Rev* 188 (2002): 43-50.

Rasmussen, K. D., S. Simmini, C. Abreu-Goodger, N. Bartonicek, M. Di Giacomo, D. Bilbao-Cortes, R. Horos, M. Von Lindern, A. J. Enright, and D. O'Carroll. "The MiR-144/451 Locus Is Required for Erythroid Homeostasis." *J Exp Med* 207, no. 7 (2010): 1351-8.

Rink, C., and S. Khanna. "MicroRNA in Ischemic Stroke Etiology and Pathology." *Physiol Genomics* (2010).

Rosero, S., V. Bravo-Egana, Z. Jiang, S. Khuri, N. Tsinoremas, D. Klein, E. Sabates, M. Correa-Medina, C. Ricordi, J. Dominguez-Bendala, J. Diez, and R. L. Pastori. "MicroRNA Signature of the Human Developing Pancreas." *BMC Genomics* 11, no. 1 (2010): 509.

Rosner, B. "Estimation of Sample Size and Power for Comparing Two Means." In *Fundamentals of Biostatistics*, edited by B. Rosner, 307-29. Australia, Canada, Mexico, Singapore, Spain, United Kingdom, United States, 2000.

Ross, R. "Atherosclerosis." In *Cecil Textbook of Medicine*, edited by JB Wyngaarden, Smith, LH, Bennett, JC., 293-98. Philadelphia, London, Toronto, Montreal, Sydney, Tokyo: W.B. Saunders Company, 1992.

Sahin, U., O. Tureci, H. Schmitt, B. Cochlovius, T. Johannes, R. Schmits, F. Stenner, G. Luo, I. Schobert, and M. Pfreundschuh. "Human Neoplasms Elicit Multiple Specific Immune Responses in the Autologous Host." *Proc Natl Acad Sci U S A* 92, no. 25 (1995): 11810-3.

Salimi, N., W. Fleri, B. Peters, and A. Sette. "Design and Utilization of Epitope-Based Databases and Predictive Tools." *Immunogenetics* 62, no. 4 (2010): 185-96.

Sethupathy, P., B. Corda, and A. G. Hatzigeorgiou. "Tarbase: A Comprehensive Database of Experimentally Supported Animal Microrna Targets." *Rna* 12, no. 2 (2006): 192-7.

Shen, J., Y. Yin, J. Mai, X. Xiong, M. Pansuria, J. Liu, E. Maley, N. U. Saqib, H. Wang, and X. F. Yang. "Caspase-1 Recognizes Extended Cleavage Sites in Its Natural Substrates." *Atherosclerosis* 210, no. 2 (2010): 422-29.

Shimokawa, K., K. Mogushi, S. Shoji, A. Hiraishi, K. Ido, H. Mizushima, and H. Tanaka. "Icod: An Integrated Clinical Omics Database Based on the Systems-Pathology View of Disease." *BMC Genomics* 11 Suppl 4 (2010): S19.

Shlomchik, M. J., J. E. Craft, and M. J. Mamula. "From T to B and Back Again: Positive Feedback in Systemic Autoimmune Disease." *Nat Rev Immunol* 1, no. 2 (2001): 147-53.

Spasic, I., S. Ananiadou, J. McNaught, and A. Kumar. "Text Mining and Ontologies in Biomedicine: Making Sense of Raw Text." *Brief Bioinform* 6, no. 3 (2005): 239-51.

Streilein, J. W., and J. Stein-Streilein. "Does Innate Immune Privilege Exist?" *J Leukoc Biol* 67, no. 4 (2000): 479-87.

Turunen, M. P., E. Aavik, and S. Yla-Herttuala. "Epigenetics and Atherosclerosis." *Biochim Biophys Acta* 1790, no. 9 (2009): 886-91.

Vickers, K. C., and A. T. Remaley. "MicroRNAs in Atherosclerosis and Lipoprotein Metabolism." *Curr Opin Endocrinol Diabetes Obes* 17, no. 2 (2010): 150-5.

Virtue, A, J. Mai, Y. Yin, S. Meng, T. Tran, X. Jiang, H. Wang, and X-F Yang. "Structural Evidence of Anti-Atherogenic MicroRNAs. " *Frontiers in Bioscience* 17 (2011): 3133-45.

Wang, H., X. Jiang, F. Yang, J. W. Gaubatz, L. Ma, M. J. Magera, X. Yang, P. B. Berger, W. Durante, H. J. Pownall, and A. I. Schafer. "Hyperhomocysteinemia Accelerates Atherosclerosis in Cystathionine Beta-Synthase and Apolipoprotein E Double Knock-out Mice with and without Dietary Perturbation." *Blood* 101, no. 10 (2003): 3901-7.

Warner, E. E., and B. K. Dieckgraefe. "Application of Genome-Wide Gene Expression Profiling by High-Density DNA Arrays to the Treatment and Study of Inflammatory Bowel Disease." *Inflamm Bowel Dis* 8, no. 2 (2002): 140-57.

Weinstein, J. N. "Fishing Expeditions." *Science* 282, no. 5389 (1998): 628-9.

———. "'Omic' and Hypothesis-Driven Research in the Molecular Pharmacology of Cancer." *Curr Opin Pharmacol* 2, no. 4 (2002): 361-5.

Xiong, Z., A. Shaibani, Y. P. Li, Y. Yan, S. Zhang, Y. Yang, F. Yang, H. Wang, and X. F. Yang. "Alternative Splicing Factor Asf/Sf2 Is Down Regulated in Inflamed Muscle." *J Clin Pathol* 59, no. 8 (2006): 855-61.

Yan, Y., L. Phan, F. Yang, M. Talpaz, Y. Yang, Z. Xiong, B. Ng, N. A. Timchenko, C. J. Wu, J. Ritz, H. Wang, and X. F. Yang. "A Novel Mechanism of Alternative Promoter and Splicing Regulates the Epitope Generation of Tumor Antigen Cml66-L." *J Immunol* 172, no. 1 (2004): 651-60.

Yang, F., I. H. Chen, Z. Xiong, Y. Yan, H. Wang, and X. F. Yang. "Model of Stimulation-Responsive Splicing and Strategies in Identification of Immunogenic Isoforms of Tumor Antigens and Autoantigens." *Clin Immunol* 121, no. 2 (2006a): 121-33.

Yang, F., and X. F. Yang. "New Concepts in Tumor Antigens: Their Significance in Future Immunotherapies for Tumors." *Cell Mol Immunol* 2, no. 5 (2005): 331-41.

Yang, X. F. "Immunology of Stem Cells and Cancer Stem Cells." *Cell Mol Immunol* 4, no. 3 (2007): 161-71.

Yang, X. F., D. Mirkovic, S. Zhang, Q. E. Zhang, Y. Yan, Z. Xiong, F. Yang, I. H. Chen, L. Li, and H. Wang. "Processing Sites Are Different in the Generation of HLA-A2.1-Restricted, T Cell Reactive Tumor Antigen Epitopes and Viral Epitopes." *Int J Immunopathol Pharmacol* 19, no. 4 (2006b): 853-70.

Yang, X. F., C. J. Wu, L. Chen, E. P. Alyea, C. Canning, P. Kantoff, R. J. Soiffer, G. Dranoff, and J. Ritz. "CML28 Is a Broadly Immunogenic Antigen, Which Is Overexpressed in Tumor Cells." *Cancer Res* 62, no. 19 (2002): 5517-22.

Yang, X. F., C. J. Wu, S. McLaughlin, A. Chillemi, K. S. Wang, C. Canning, E. P. Alyea, P. Kantoff, R. J. Soiffer, G. Dranoff, and J. Ritz. "CML66, a Broadly Immunogenic Tumor Antigen, Elicits a Humoral Immune Response Associated with Remission of Chronic Myelogenous Leukemia." *Proc Natl Acad Sci U S A* 98, no. 13 (2001): 7492-7.

Yang, X. F., Y. Yin, and H. Wang. "Vascular Inflammation and Atherogenesis Are Activated Via Receptors for Pamps and Suppressed by Regulatory T Cells." *Drug Discov Today Ther Strateg* 5, no. 2 (2008): 125-42.

Yin, Y., Y. Yan, X. Jiang, J. Mai, N. C. Chen, H. Wang, and X. F. Yang. "Inflammasomes Are Differentially Expressed in Cardiovascular and Other Tissues." *Int J Immunopathol Pharmacol* 22, no. 2 (2009): 311-22.

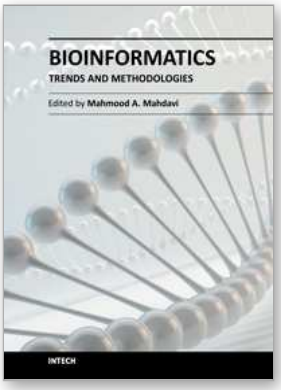
Yotnda, P., H. Firat, F. Garcia-Pons, Z. Garcia, G. Gourru, J. P. Vernant, F. A. Lemonnier, V. Leblond, and P. Langlade-Demoyen. "Cytotoxic T Cell Response against the Chimeric P210 Bcr-Abl Protein in Patients with Chronic Myelogenous Leukemia." *J Clin Invest* 101, no. 10 (1998): 2290-6.

Zhang, C. "Micrornas: Role in Cardiovascular Biology and Disease." *Clin Sci (Lond)* 114, no. 12 (2008): 699-706.

Zhang, D., X. Jiang, P. Fang, Y. Yan, J. Song, S. Gupta, A. I. Schafer, W. Durante, W. D. Kruger, X. Yang, and H. Wang. "Hyperhomocysteinemia Promotes Inflammatory Monocyte Generation and Accelerates Atherosclerosis in Transgenic Cystathionine Beta-Synthase-Deficient Mice." *Circulation* 120, no. 19 (2009): 1893-902.

Zinkernagel, R. M., and H. Hengartner. "Regulation of the Immune Response by Antigen." *Science* 293, no. 5528 (2001): 251-3.

Zorn, E., Orsini, E., Wu, CJ., Stein, B., Chillemi, A., Canning, C., Alyea, EP, Soiffer, RJ., and Ritz, J. "A CD4+ T Cell Clone Selected from a Cml Patient after Donor Lymphocyte Infusion Recognizes Bcr-Abl Breakpoint Peptides but Not Tumor Cells." *Transplantation* 71, no. 8 (2001): 1131-7.



Bioinformatics - Trends and Methodologies
Edited by Dr. Mahmood A. Mahdavi

ISBN 978-953-307-282-1
Hard cover, 722 pages
Publisher InTech
Published online 02, November, 2011
Published in print edition November, 2011

Bioinformatics - Trends and Methodologies is a collection of different views on most recent topics and basic concepts in bioinformatics. This book suits young researchers who seek basic fundamentals of bioinformatic skills such as data mining, data integration, sequence analysis and gene expression analysis as well as scientists who are interested in current research in computational biology and bioinformatics including next generation sequencing, transcriptional analysis and drug design. Because of the rapid development of new technologies in molecular biology, new bioinformatic techniques emerge accordingly to keep the pace of in silico development of life science. This book focuses partly on such new techniques and their applications in biomedical science. These techniques maybe useful in identification of some diseases and cellular disorders and narrow down the number of experiments required for medical diagnostic.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821