IJPSR (2011), Vol. 2, Issue 10 (Review Article)

INTERNATIONAL JOURNAL OF PHARMACEUTICAL SCIENCES AND RESEARCH

Received on 07 July, 2011; received in revised form 04 August, 2011; accepted 21 September, 2011

ROLE OF BIOINFORMATICS, CHEMOINFORMATICS AND PROTEOMIC IN BIOMARKER IDENTIFICATION AND DRUG TARGET VALIDATION IN DRUG DISCOVERY PROCESSES

Tara Shankar Basuri* and Anwar S. Meman

Department of Pharmaceutical Chemistry, SSR College of pharmacy, Sayli road , Silvassa, India

Keywords:

G protein-coupled receptors, Diabetes mellitus

Correspondence to Author:

Tara Shankar Basuri
Department of Pharmaceutical Chemistry, SSR College of pharmacy, Sayli road , Silvassa, India

ABSTRACT

Novel biomarker identification and drug target validation are highly complex and resource-intensive processes, requiring an integral use of various tools, approaches and information. The recently developed proteomic technology features high-throughput parallel analysis of thousands of proteins in individual patients and amount populations and thus opens up the possibility of providing more details at a global level on the molecular mechanisms. With regularly updated public databases, bioinformatics can contribute to these processes by providing functional information of target candidates and correlating this information to the biological pathways. In this review, we outline recent advances of bioinformatic application in proteomic research on biomarker discovery and drug target validation. Specifically, we highlight how bioinformatics can facilitate the proteomic studies of biomarker identification and drug target validation, rating valuable data for the development of new drug candidates. Chemoinformatics has evolved over the last 30 years into a scientific discipline that now is in full bloom. It covers many areas such as chemical structure representation, chemical reaction manipulation, data processing and data analysis, property prediction, chemometrics, data mining, structure elucidation, and synthesis design. Chemoinformatics methods have successfully been applied in all fields of chemistry. The future will bring a rapid expansion of the use of Chemoinformatics to further our understanding of chemistry and to process the flood of chemical information.

INTRODUCTION: Bioinformatics is a field of information technology concerning the storage, retrieval, visualization, prediction and analysis of molecular data with biological or clinical significance. Bioinformatics [1] can accelerate proteomic studies in data mining, integrated data management and network modeling. Data mining is a process which is now recognized as a key tool in proteomics. This is due to the development of a wide range of software programs dedicated to mining data obtained at different stages of proteomic experiments.

For example, using software analysis, MS results can be compared with sets of theoretical protein sequences available in databases. Integrated data management is used to integrate data acquired from proteomics of various areas of specialization within a software environment in order to improve the reliability and to allow better understanding of results. Network modeling and systems biology provide information for better understanding of large molecular networks in their cellular context by in silico modeling of the complexity of biological processes involving functionally interacting molecules.

Chemo informatics [2] is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information.

Application of Bioinformatics in Drug Designing: Bioinformatics plays an important role in the design of new drug compounds. Rational Drug Design (RDD): Rational drug design [3] is a process used in the biopharmaceutical industry to discover and develop new drug compounds. RDD uses a variety of computational methods to identify novel compounds, design compounds for selectivity, efficacy and safety, and develop compounds into clinical trial candidates. These methods fall into several natural categories Structure -based drug design [4], ligand-based drug design, de novo design and homology modeling, depending on how much information is available about drug targets and potential drug compounds. We shall focus on structure-based drug design in this article and describe a few of its salient features.

Structure-Based Drug Design (SBDD): Structure-based drug design [5] is one of several methods in the rational drug design toolbox. Drug targets are typically key molecules involved in a specific metabolic or cell signaling pathway that is known, or believed, to be related to a particular disease state. Drug targets are most often proteins and enzymes in these pathways. Drug compounds are designed to inhibit, restore or otherwise modify the structure and behavior of disease-related proteins and enzymes.

SBDD uses the known 3D geometrical shape or structure of proteins to assist in the development of new drug compounds. The 3D structure of protein targets is most often derived from x-ray crystallography or nuclear magnetic resonance (NMR) techniques. X-ray and NMR methods can resolve the structure of proteins to a resolution of a few angstroms (about 500,000 times smaller than the diameter of a human hair). At this level of resolution, researchers can precisely examine the interactions between atoms in protein targets and atoms in potential drug compounds that bind to the proteins.

This ability to work at high resolution with both proteins and drug compounds makes SBDD one of the most powerful methods in drug design. SBDD methods have been used in designing drugs for a well known cancer-related protein complex. Two protein targets that have been studied extensively in cancer research are p53 and MDM2. These two proteins form a single p53-MDM2 complex as part of a cell-signaling pathway that regulates cell division. Mutated forms of p53-MDM2 result in various forms of tumors and cancers. Several decades of research have been aimed at designing small-molecule compounds that restore the normal function of p53-MDM2, and consequently reduce or eliminate certain forms of cancer.

One well-known anticancer drug 'nutria¢' - has been developed by Roche Pharmaceuticals to restore the normal functioning of MDM2. SBDD methods played an important role in this development. The beauty of the SBDD method is the extremely high level of detail that it reveals about how drug compounds and their protein targets interact. We can identify the exact location of all five nutlinÃ¢ compounds, their individual 3D orientations relative to MDM2 surface and interior amino acids, and how deeply embedded each nutlinÃ¢ compound is in the interior of MDM2. This information is useful in designing the 3D shape of the nutlinÃ¢ parent compound or various analogues of the drug. This information also assists researchers in designing drug compounds that bind selectively and tightly to MDM2, thus leading to more potent and safer cancer drugs.

Docking Ligands: One of the key benefits of SBDD [6] methods is the exceptional capability it provides for docking putative drug compounds (ligands) in the active site of target proteins. Most proteins contain pockets, cavities, surface depressions and other geometrical regions where small-molecule compounds can easily bind. With high-resolution x-ray and NMR structures for proteins and ligands, researchers can show precisely how ligands orient themselves in protein active sites. Open source bioinformatics tools such as VMD and NAMD.

Drug Lead Optimization [7, 8]: When a promising lead candidate has been found in a drug discovery program, the next step (a very long and expensive step) is to optimize the structure and properties of the potential drug. This usually involves a series of modifications to the primary structure (scaffold) and secondary structure (moieties) of the compound. This process can

be enhanced using software tools that explore related compounds (bioisosteres) to the lead candidate. Open Eye's WABE is one such tool. Lead optimization tools such as WABE offer a rational approach to drug design that can reduce the time and expense of searching for related compounds (**Fig. 1**).

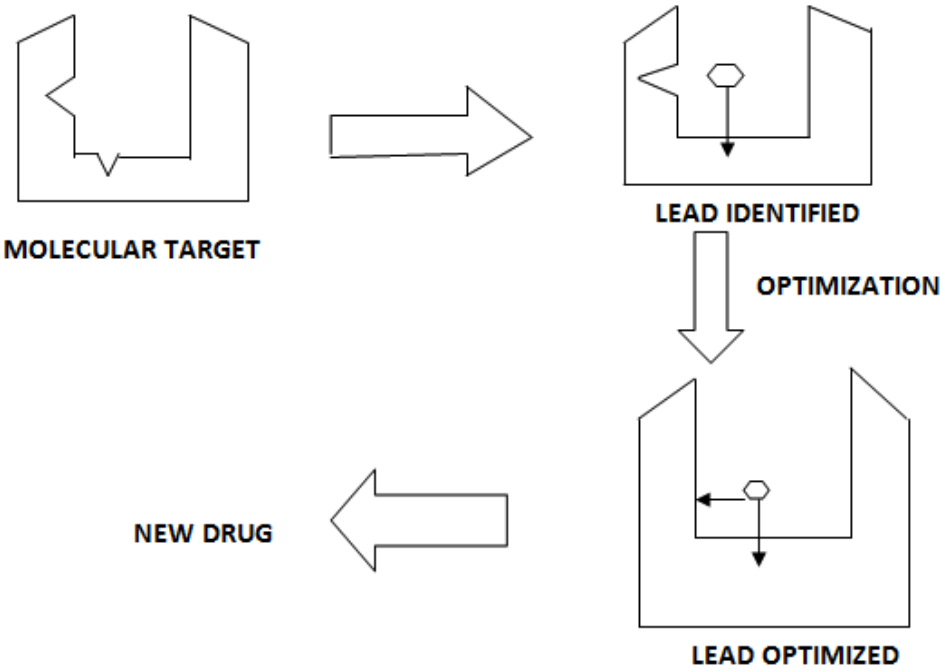

FIG 1: LEAD OPTIMIZATION PROCESS

Homology Modeling: Homology modeling is one method used to predict 3-D structure. In homology modeling, the amino acid sequence of a specific protein (target) is known, and the 3-D structures of proteins related to the target (templates) are known. Bioinformatics software tools are then used to predict the 3-D structure of the target based on the known 3-D structures of the templates. MODELLER [9] is a well-known tool in homology modeling, and the SWISS-MODEL Repository is a database of protein structures created with homology modeling. A common activity in biopharmaceutical companies is the search for drug analogues. Starting with a promising drug molecule, one can search for chemical compounds with similar structure or properties to a known compound. There are a variety of methods used in these searches, including sequence similarity, 2D and 3D shape similarity, substructure similarity, electrostatic similarity and others.

Physicochemical Modeling: Drug-receptor interactions occur on atomic scales. To form a deep understanding of how and why drug compounds bind to protein targets, we must consider the biochemical and biophysical properties of both the drug itself and its target at an atomic level. Swiss-PDB is an excellent tool for doing this. Swiss-PDB can predict key physicochemical properties, such as hydrophobicity and polarity that have a profound influence on how drugs bind to proteins.

Drug Bioavailability and Bioactivity: Most drug candidates fail in Phase III clinical trials after many years of research and millions of dollars have been spent on them. And most fail because of toxicity or problems with metabolism. The key characteristics for drugs are Absorption, Distribution, Metabolism, Excretion, Toxicity (ADMET) and efficacy in other words bioavailability and bioactivity. Although these properties are usually measured in the lab, they can also be predicted in advance with bioinformatics software.

Drug Design: Drug design is the approach of finding drugs by design, based on their biological targets. Typically a drug target is a key molecule involved in a particular metabolic or signaling pathway that is specific to a disease condition or pathology, or to the infectivity or survival of a microbial pathogen.

Computer-assisted Drug Design: Computer-assisted drug design uses computational chemistry to discover, enhance, or study drugs and related biologically active molecules. Methods used can include simple molecular

modeling, using molecular mechanics, molecular dynamics, semi-empirical quantum chemistry methods, ab initio quantum chemistry methods and density functional theory. The purpose is to reduce the number of targets for a good drug that have to be subjected to expensive and time-consuming synthesis and trialing.

Drug Design based on Bioinformatics Tools: The processes of designing a new drug using bioinformatics tools have open a new area of research. However, computational techniques assist one in searching drug target and in designing drug in silico, but it takes long time and money. In order to design a new drug one need to follow the following path.

Identify Target Disease: Target identification [10] alone is not sufficient in order to achieve a successful treatment of a disease. A real drug needs to be developed. This drug must influence the target protein in such a way that it does not interfere with normal metabolism. One way to achieve this is to block activity of the protein with a small molecule. Bioinformatics methods have been developed to virtually screen the target for compounds that bind and inhibit the protein. Another possibility is to find other proteins that regulate the activity of the target by binding and forming a complex.

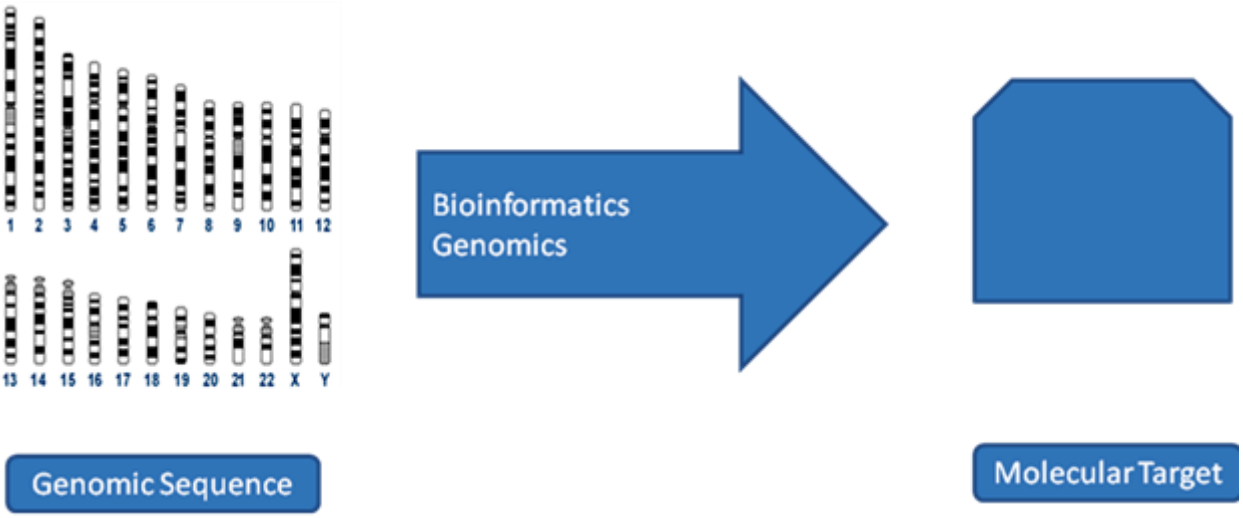

FIG. 2: BIO INFORMATICS GENOMICS

Study interesting compounds: One needs to identify and study the lead compounds [11] that have some activity against a disease. These may be only marginally useful and may have severe side effects. These compounds provide a starting point for refinement of the chemical structures.

Detect the molecular bases for disease [12]: If it is known that a drug must bind to a particular spot on a particular protein or nucleotide then a drug can be tailor made to bind at that site. This is often modeled computationally using any of several different techniques. Traditionally, the primary way of determining what compounds would be tested computationally was provided by the researchers\understanding of molecular interactions. A second method is the brute force testing of large numbers of compounds from a database of available structures.

Rational Drug Design Techniques: These techniques attempt to reproduce the researchers understanding of how to choose likely compounds built into a software package that is capable of modeling a very large number of compounds in an automated way. Many different algorithms have been used for this type of testing, many of which were adapted from artificial intelligence applications. The complexity of biological systems makes it very difficult to determine the structures of large biomolecules.

Ideally experimentally determined (x-ray or NMR) structure is desired, but biomolecules are very difficult to crystallize.

Refinement of Compounds: Once you got a number of lead compounds have been found, computational and laboratory techniques have been very successful in refining the molecular structures to give a greater drug activity and fewer side effects. This is done both in the laboratory and computationally by examining the molecular structures to determine which aspects are responsible for both the drug activity and the side effects.

Quantitative Structure Activity Relationships (QSAR): This computational technique should be used to detect the functional group in your compound in order to refine your drug. This can be done using QSAR that consists of computing every possible number that can describe a molecule then doing an enormous curve fit to find out which aspects of the molecule correlate well with the drug activity or side effect severity. This information can then be used to suggest new chemical modifications for synthesis and testing.

Solubility of molecule: One need to check whether the target molecule is water soluble or readily soluble in fatty tissue will affect what part of the body it becomes concentrated in. The ability to get a drug to the correct part of the body is an important factor in its potency. Ideally there is a continual exchange of information between the researchers doing QSAR studies, synthesis and testing. These techniques are frequently used and often very successful since they do not rely on knowing the biological basis of the disease which can be very difficult to determine.

Drug Testing: Once a drug has been shown to be effective by an initial assay technique, much more testing must be done before it can be given to human patients. Animal testing is the primary type of testing at this stage. Eventually, the compounds, which are deemed suitable at this stage, are sent on to clinical trials. In the clinical trials, additional side effects may be found and human dosages are determined.

Applications of Chemoinformatics:

Prediction of Properties: It has been realized in recent years that during the development of a new drug [12] increasing attention has to be given not only to the optimization of its biological activity but also to ensure that is has favorable physical, chemical, and biological properties such as adsorption, distribution, metabolism, excretion, and toxicity (ADME-Tox). Methods are being developed for the prediction of these properties prior to the synthesis of the respective compounds in order to use these methods in the virtual screening of large sets of compounds. One of the properties that deserve special attention is aqueous solubility because this property has to be in a certain range in order for a drug to be orally administered and, on the other hand, also to be absorbed into the body.

Analysis of Analytical Chemistry Data [13]: The analysis of samples to assign their quality, their place of origin, or their age has high commercial interest. As the relationships between the composition of a sample and its quality, origin, or age are highly complex, chemo metrics methods and other inductive learning methods have been employed since a long time.

Computer-Assisted Structure Elucidation (CASE): The elucidation of the structure of a compound [14] is presently residing nearly exclusively on spectral data of various sorts (NMR, IR, and MS). The derivation of the structure of a compound from spectroscopic data involves the processing of large amounts of information and many decisions have to be made between hosts of alternatives. These efforts are continuing, and still a lot is to be done.

Computer-Assisted Synthesis Design (CASD): The design of a synthesis [15] for an organic compound involves the consideration of many alternatives, has to draw from a broad knowledge of organic reactions, has to focus on a large selection of available starting materials, and has to consider a variety of economic effects.

Thus, it is one of the most challenging problems in organic chemistry and has attracted early on interest as a field of exercise for artificial intelligence techniques. Our group has worked on this challenging project since 30 years and we have arrived at a version of the WODCA system (Workbench for the Organization of Data for Chemical Applications) that we consider mature enough to be able to be of practical use for synthesis chemists WODCA comprises a series of tools that the user/chemist can employ for planning the synthesis of individual compounds or of combinatorial libraries (**Fig. 3**)

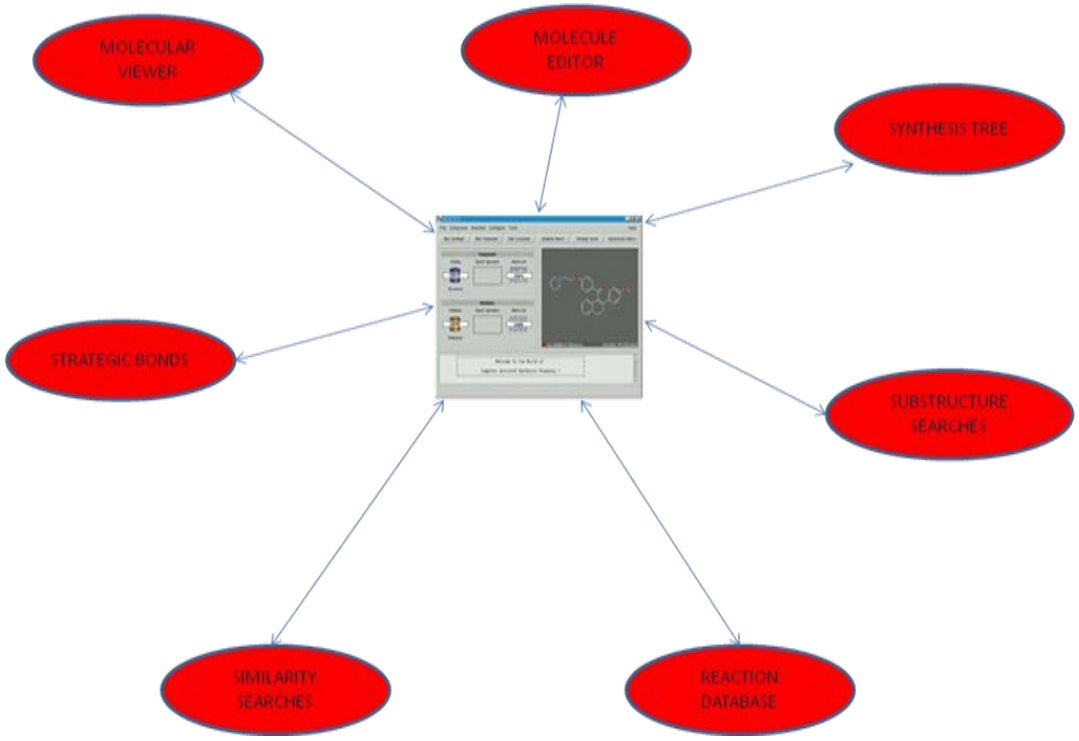

FIG. 3: TOOLS OF THE SYNTHESIS DESIGN SYSTEM WODCA FOR PLANNING THE SYNTHESIS OF ORGANIC COMPOUNDS AND COMBINATORIAL LIBRARIES

- A molecule editor and a molecule viewer allow the chemist to communicate with the system in the language he/she is used in the form of structure diagrams.
- Similarity searches can be used to drive the design of a synthesis as quickly as possible to available starting materials. Different similarity criteria, specifically designed for focusing on synthesis reactions, are provided that the user can choose from. Catalogs of available compounds are attached to the system but methods are also provided to append one's own proprietary catalogs.
- Strategic bonds can be perceived where a molecule should be dissected in order to simplify a synthesis problem and work with smaller precursor molecules.
- Such a dissection of a molecule corresponds to a retro reaction and searches in a reaction database can be invoked to verify whether such a retro reaction corresponds to a known reaction or reaction type.

Drug Design: The area of drug design [16] is presently undoubtedly the most important field for using Chemoinformatics methods. The reasons are several: first, there is enormous economic pressure to reduce the high costs needed for developing a new drug and to reduce the time needed for this process. Secondly, experimental methods recently introduced in the drug design process such as combinatorial chemistry and high-throughput docking. Screenings produce enormous amounts of data that have to be analyzed. And lastly, it is clear that the biological activity of a chemical compound cannot yet be predicted from first principles and therefore still needs methods that learn from available data.

Fig. 4 shows a simplified outline of the drug design process [17] and thus highlights the interplay of bioinformatics and chemoinformatics in this process. Bioinformatics methods should assist in identifying from genetic information the target protein that is the focus of a certain disease. The next steps are the identification of a new lead structure, the optimization of this lead structure to increase the biological activity, and then, or better simultaneously, optimizing the ADME-Tox properties to convert the highly active compound into a drug with advantageous physical, chemical, and biological properties. For all those three tasks chemoinformatics methods have been developed to increase the efficiency in achieving these goals.

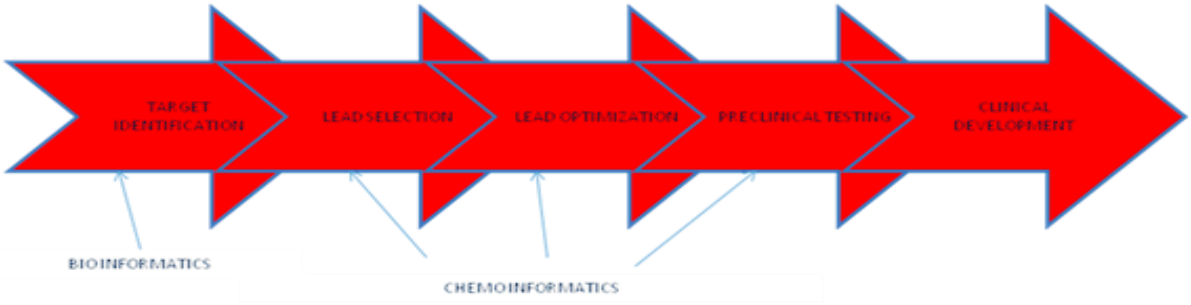

FIG. 4: THE DRUG DESIGN PROCESS

Fig. 5 gives an overview of the methods developed for lead discovery. They fall into two categories, target-based methods [18] that need the 3D structure of the target protein and ligand-based methods that do not need the 3D structure of the target protein but do their job by analyzing a series of ligands that bind to this target protein.

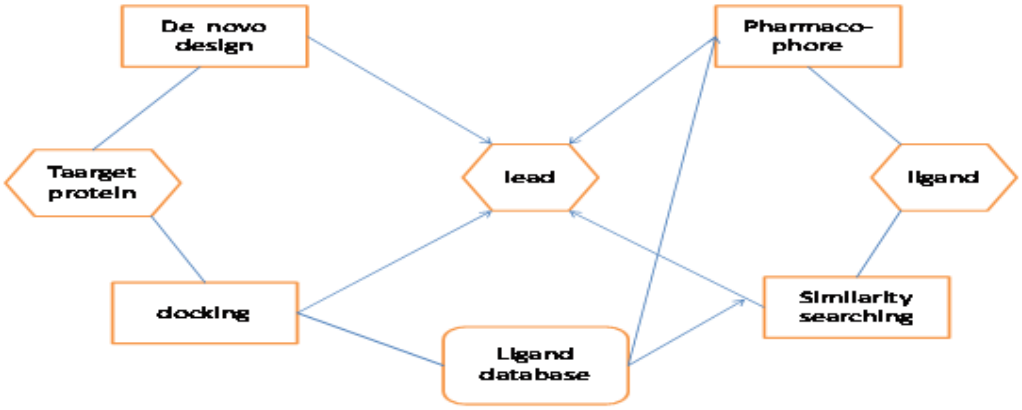

FIG. 5: METHODS DEVELOPED FOR LEAD DISCOVERY

Fig. 6 shows the major methods for lead optimization. Their first task is to expand a lead into a set of potential ligand [19]. This can be achieved through similarity searching in compound databases, by lead hopping, by analyzing the outcome of high-throughput experiments, or by performing compound screening on virtual libraries. Once a set of ligands has been obtained those can either be docked into the target protein, if its 3D structure is known, or one can try to establish quantitative structure - activity relationships for this set of ligands in order to pinpoint the highly active compounds. To summarize, chemo informatics [20] has developed a large arsenal of methods that can be utilized to make the drug design process more efficient. Much has been done but it should be emphasized that we are just at the beginning and much more has to be developed in order that we achieve the goals that can be envisaged.

FIG. 6: MAJOR METHODS FOR LEAD OPTIMIZATION

Bioinformatics application in Proteomic Research on Biomarker Discovery and Drug Target Validation: The proteome [20] is defined as the entirely expressed protein complement of a cell, organ or organism and it includes allisoforms and post-translational variants. Proteomic technology attempts to separate, identify and characterize a global set of proteins in an effort to provide information about protein abundance, location, modification and protein-protein interaction in a proteome of a given biological system. By studying the interrelationships of protein expression and modification in health and disease, or drug treatment, proteomics can be applied to biomarker discovery and drug target validation. In this review article, we aim to outline recent advances in this emerging field and highlight a number of successful applications.

Major Technological Platforms for Proteomics: A number of complementary technologies have been developed to analyze proteomes [21] in a global scale. Currently, the most commonly used proteomic platforms include two-dimensional gel electrophoresis (2DE), protein chip arrays and liquid chromatography, incorporated with matrix- assisted laser desorption/ionization time of flight (MALDI-TOF), surface enhanced laser desorption ionization time of flight (SELDI-TOF) and/or tandem mass spectrometry (MS/MS). 2DE separates protein complexes according to their isoelectric points and molecular weights. This technique is suitable for analyzing proteomes of cell, tissue and serum and can detect proteins with post-translational modification, including phosphorylation, glycosylation, etc.

Although low abundant proteins may not be able to be well detected by 2DE, this problem can be overcome by using sample pre-fractionation or -purification. Differently expressed protein samples are subjected to digestion with a protease, which yields peptides prior to MALDI-TOF MS. MALDI-TOF MS enables peptides to be ionized at high sensitivity without excessive fragmentation. Two steps in MALDI-TOF MS are peptides ionization by a laser coupled with time of flight and detection as ion mass to charge (*m/z*) ratio with a mass analyzer. The ionized and separated peptides can be further characterized by tandem mass spectrometry (MS/MS) by a second MS analyzer to generate an MS/MS spectrum representing a series of ion fragments of the specific peptide, which can construct a partial amino acid sequence according to the MS/MS spectrum. MS/MS based protein identification is often more accurate with higher outputs, since it provides information not only about peptide masses but also peptide sequences.

Protein chip arrays can identify altered proteins by surface affinity and molecular weight. This is a very useful technique since the experimental conditions can be well controlled.

For example, different cofactors or inhibitors can be introduced in the binding assays to adjust the stringency of the binding activities. Another advantage of this technique is that these highly parallel assays are suitable for analysis of low abundant proteins. In addition, with proper detection methods, protein chip arrays can be used to identify the downstream targets of various enzymes such as kinases, phosphatases, methyl transferases and proteases.

Protein chip arrays are most often combined with SELDI-TOF MS. SELDI-TOF MS is an evolving proteomic platform that allows rapid and sensitive analysis of complex protein mixtures. It is highly sensitive to analyze small amounts of raw protein samples and can detect proteins with molecular weights lower than 6 kDa (the lowest detection range of 2DE). However, SELDI-TOF MS is not a quantitative approach and protein identification cannot be directly determined through this method. In addition, protein chip arrays can be multifaceted by changing the capture ligand on the protein chip surface.

For example, small molecule array is used to investigate small-molecule-protein interaction, AB array uses specific antibody as the capture ligand to study cell-signaling and peptide array uses peptides as capture ligands to study protein-protein interaction. In particular, as a big achievement in protein chip arrays, Schreiber and MacBeath have reported that proteins in their native state can be successfully attached on microarray surfaces. In this study, they demonstrated three applications of the protein chip arrays, including the screening for protein-protein interactions, the identification of the substrates of protein kinases and the identification of the protein targets of small molecules.

Liquid chromatography/tandem mass spectrometry (LC-MS/MS) is an analytical method for identifying multiple components of a protein mixture. The peptide mixtures in very complex protein samples are physically resolved by chromatographic separation prior to injection into the mass spectrometer to generate a more informative map, consisting of both the unique elution characteristics (column retention times) as well as *m/z* ratios of individual peptide. LC-MS/MS is well-suited to examine complex protein samples, since peptides with the same nominal *m/z* are less likely to be introduced to MS/MS at the same time, and fewer artifacts arise due to ion suppression or ion-ion interference.

LC-MS/MS can also overcome the difficulties of 2DE in the identification of very large and basic proteins by pre-fractionation using 1DE. Besides these major technological platforms, a number of differential approaches based on isotope-coded affinity tags (ICAT), isobaric tags for relative and absolute quantitation (iTraq), and non-labeling technologies have been developed. ICAT labels proteins at cysteine residues with light and heavy tags carrying a biotin moiety for quantitative proteomic analysis [19]. iTraq employs a 4-plex set of amine reactive isolaric tags to derivatize peptides at the N-terminus and the lysine side chains, thereby labeling all peptides in a digest mixture [20].

iTraq is a new LC based technique, which is more susceptible to errors in precursor ion isolation. In MS, peptides labeled with any of the isotopic tags are indistinguishable (isobaric). A number of non-labeling technologies have thus also been developed. Zurbig *et al.* has applied capillary electrophoresis (CE) coupled with MS/MS for biomarker identification in different body fluids. Beckman- Coulter PF2DTM system [22] and Agilent® High Capacity Multiple Affinity Removal System have all been used in the separation of human serum proteins for biomarker identification.

Challenges in Proteomic Approaches: Proteomics [23] provides a large number of validated targets for drug design and thus optimal methods have to be created to handle this challenge. This high dimensionality of data generated from these studies requires the development of advanced bioinformatic tools for efficient and accurate data analyses.

For proteome profiling of a particular system or organism, a number of specialized software tools and advanced informatics are needed to support the analysis and management of these massive amounts of data. The rapidly emerging field of bioinformatics has the capacity to greatly enhance treatment efforts by serving as a bridge between proteomic raw data and applicable output. By correlating genetic variation and potential changes in protein structure with clinical risk factors, disease presentation and differential response to treatment and drug candidates, it may be possible to obtain valuable new insights to support and guide rational decision-making, both at the clinical and public health levels.

Fig. 1 illustrates the application of bioinformatics integrated proteomics approaches in the biomarker discovery and drug development pipeline. Of particular interests, the integrated approaches can be applied to preclinical or clinical analysis of samples, laboratory work of biomarker discovery, quantification and validation of potential biomarkers, followed by the preclinical and clinical development. Application of this emerging integrated technology in drug development can be divided into three categories: target discovery and validation, illustration of efficacy and toxicity of compounds and identification or prediction of drug response.

Bioinformatic tools for structure determination of Drug Targeted Proteins: In drug discovery pipeline, one of the most important steps is the determination of three dimensional structure of a target protein or nucleic acid. Bioinformatic software can use the three-dimensional structural information of the unliganded target to design entirely new lead compounds *de novo*. This software allows rapidly and accurately docking large numbers of candidate molecules into the binding site of the target macromolecule prior to actual synthesis and biological studies. The earliest and most widely used bioinformatic software is the program DOCK. DOCK has been successfully applied to the *in silico* virtual high throughput screen for high affinity cytochrome p450cam substratesand imidazole inhibitors of cytochrome p450 enzymes. Besides DOCK, software programs such as ADAM, AutoDOCK, Flex X, SLIDE, DISCO, MCSS, and 3D-QSAR can also score candidate molecules according to their interactions with the selected site of target proteins.

Databases for Protein-Protein Interaction: Bioinformatics plays a critical role in the analysis of protein-protein interaction [24]. Several databases that accumulate these data are currently available. They play an essential role in visualizing and integrating experimental data with the information on protein-protein interactions available in the database. Recently, the most well-established databases in this specific area include Database of Interacting Proteins (DIP), the General Repository for Interaction Datasets (GRID), the Bimolecular Interaction Network Database (BIND), the expression profile reliability (EPR) index and the paralogous verification method (PVM), the mitochondrial protein sequence database and annotation system MitoProteome and the Human Protein Reference Database (HPRD) .

Drug Metabolism and Drug Interaction Databases: Drug metabolism and drug interaction databases [25] attempt to link information of proteomics with drug compounds. The information needed by these databases is not electronically accessible; it resides in monographs, books, FDA filings and journal's articles. Several commercial databases of drug metabolism or drug interaction are available now, including Metabolism and Transport Drug Interaction Database or DIDbase offered by the University of Washington. DIDbase is a web-based database designed to evaluate the interaction potential of drugs in development. It contains the detailed information, such as therapeutic range, PK parameters, kinetic parameters, *V*max data, AUC data, dosage parameters, dosage intervals, administration method and experimental parameters that are extracted and manually verified from thousands of pharmaceutical research papers.

Other databases include VITIC toxicity database, which is designed to allow users to predict and assess toxicological effects of more chemicals while reducing the need for animal testing ; PharmGKB, an integrated resource designed to archive experimental and literature data on drug-gene and drug-protein interactions and DrugBank, which is a freely available and fully downloadable drug interaction and drug metabolism dataset. DrugBank allows users to scan sequences of a new protein or even a new proteome against DrugBank's entire chemical database.

Predictive Software Tools for ADME-Tox: The use of predictive software, particularly in drug toxicological research, is leading to the advent of *in silico* ADME-Tox. *In silico* ADME-Tox allows usually expensive and time consuming bench work to be done in seconds with a computer, at a tiny fraction of the cost. There are numerous predictive software and databases; for example, LogP for predicting the activity, absorption, distribution and metabolism of drug candidates META and METEOR are also used for *in silico* biotransformation predictions.

Standardization of data exchange and concerted Bioinformatic Analysis of Proteomic Data: The quest for high-throughput proteomics [26] has revealed a number of critical issues. For example, reliable high throughput spot matching and quantification in 2DE remain a significant bottleneck in the bioinformatics pipeline. To this end, it is important to establish a full multi-site standardization infrastructure for processing, archival storage, standardization and retrieval of proteomic data and metadata.

The Protein Standards Initiative (PSI) group of the Human Proteome Organization (HUPO) has been tasked with the establishment of this kind of common standards for data exchange in the field of proteomics. The HUPO PSI has developed standards for the representation of proteomic data by initially focusing on mass spectrometry and protein- protein interactions.

Current achievements and potential application of Bioinformatics in Proteomic Research on Biomarker Discovery and Drug Target Validation:

Biomarker (Proteomic Signature) Discovery: A biomarker [27] is a "laboratory measurement or physical sign used as a substitute for a clinically meaningful marker that measures directly how a patient feels, functions or survives". "Changes induced by a therapy on related markers are expected to reflect changes in a clinically meaningful marker". In this regard, ideal biomarkers are those which are easily accessible and amenable to serial monitoring, i.e., they can be obtained by non-invasive or minimally-invasive techniques and have limited potential for sampling error. Biomarkers can provide a basis for the selection of lead candidates for clinical trials and for the

understanding of candidates' pharmacology. They can also help in the characterization of the subtypes of diseases for which a therapeutic intervention is most appropriate. One approach used for the establishment of the linkage of a biomarker to a clinical endpoint is to estimate the proportion of treatment effect that is accounted for by the surrogate endpoint. The newly developed bioinformatics-integrated proteomics will be able to perform high throughput parallel analysis of thousands of proteins in individual patients and large amount of populations. This integrated technique opens up the possibility of providing more details at the global level on the molecular mechanisms associated with diseases.

Using bioinformatics combined with protein chip arrays, Chan and his co-workers have identified CA15.3 and CA 27.29 as unique markers for monitoring therapy or recurrence of advanced breast cancer. These two markers have been approved by the Food and Drug Administration of the United States. The diagnosis of breast cancer has also been performed using artificial neural network to analyze the large amount of data acquired from a protein chip array study. CA 19-9 is a biomarker successfully identified by bioinformatics-integrated proteomic approach for the diagnosis of pancreatic adenocarcinoma. Recently, the high-resolution serum proteomic profiling revealed a disease- specific, carrier-protein-bound mass signature for Alzheimer's disease.

Molecular Drug Target: Target identification [28] and validation are the first key steps in the drug discovery pipeline. Reliable technologies for addressing target identification and validation are the foundation of successful drug development. Proteomics has been well utilized in protein expression profiling and tissue/ cell-scale target validation. A direct application of bioinformatics-integrated proteomic analysis in drug development is to validate the presence or absence of targets in a tempo-spatial manner. Nearly half of the molecular drug targets fall into six families: G-protein-coupled receptors (GPCRs), serine/threonine and tyrosine protein kinases, zinc metallopeptidases, serine proteases, nuclear hormone receptors and phosphodiesterases. Comparative proteomic studies in mammalian cells have resulted in important insights into novel potential drug targets or indirect drug effecting molecules.

The study by Oh and his colleagues on the endothelial cell surface proteins has led to the discovery of two proteins, amino peptidase-P and annexin A1 that serve as *in vivo* targets for antibodies in lungs and solid tumors, respectively. Most drugs bind to discrete binding sites, which can be identified readily by structural analysis. It is possible to filter

Role of Bioinformatics in Target Validation: Genomic, transcriptomic and proteomic technologies [29] are currently driving the pharmaceutical industry's search for novel targets that will result in innovative therapies. Building up the case that drug modulation of a target is likely to have a beneficial effect in a given disease (target validation) is a key step in this process and combines data from molecular biology, cell biology, bioinformatics [24] and *in vitro* and *in vivo* experiments, with the amount of work needed for validation increasing dramatically for ''novel'' targets with no known biological function or link to disease.

This review focuses on the increasingly sophisticated *in silico* [25] approaches that are being used to support target validation. Predicting function from sequence and structure the most commonly used approach to assign function to proteins is by sequence similarity. So, attention has focused on complementing and extending this approach by the development of complementary methods to function prediction using sequence and structural information. Sequence-based approaches .the identification of signatures of domains and functional sites in amino acid sequences has played an important and complementary role to similarity searching methods in the functional characterization of proteins (**Fig. 7**).

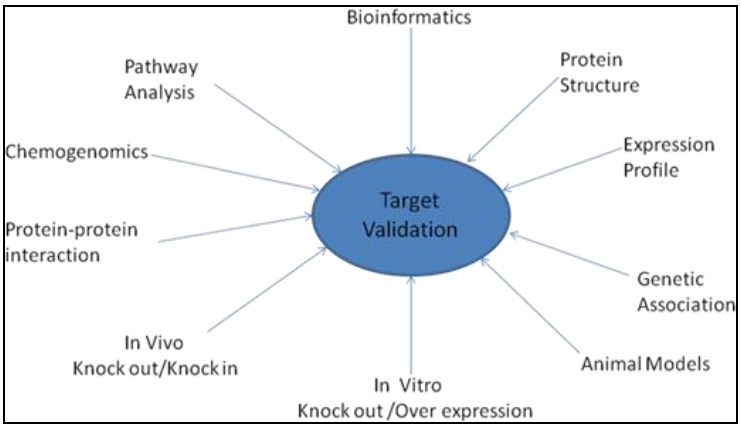

FIG. 7: TARGET VALIDATION INVOLVES LINKING PUTATIVE TARGETS TO BIOLOGICAL FUNCTION IN HEALTHY AND DISEASED STATES

In an extension of this approach, the prediction of sequence motifs associated with post-translational modifications and sub cellular localization of proteins has the ability to transfer predict functional categories, whereas Proteome Analyst predicts sub cellular location using database text annotations from homologs in addition to sequence information. The Eukaryotic Linear Motif (ELM) server is a resource for investigating short peptide linear motifs which are used for cell compartment targeting, protein–protein interaction, regulation by phosphorylation, acetylation, glycosylation and a range of other post-translational modifications.

Scan site identifies short sequence motifs within query proteins that regulate protein - protein interactions [43] in cell signaling and functional information between sequences that are unrelated at the primary sequence or evolutionary level. The key principle here is that functionally-related proteins will have similar posttranslational modifications and sorting signals even if they are unrelated at the sequence level. The Prot Fun method integrates individual [42] attributes (e.g. glycosylation, phosphorylation, signal peptides etc.) to can be used to generate biochemical tools that enable the identification of interaction partners.

The availability of the sequenced genomes of a wide range of organisms has facilitated the development of protein function prediction methods based on viewing this data in an evolutionary context Phylogenomic profiling focuses on how proteins became similar in sequence through evolution rather than on the sequence similarity itself. In this approach, the evolutionary history of genes is used to predict the function of uncharacterized genes. The re-sampled inference of orthologs (RIO) web server has been developed to automate phylogenomic analysis.

Proteomics: A Major New Technology for the Drug Discovery Process:

Use of Proteomics to identify Disease Specific Proteins: In most cases, the drug discovery process [30] is initiated by the identification of a novel candidate target - almost always a protein - that is believed to be instrumental in the disease process. To date, there is a variety of means whereby drug targets have been forthcoming. These include molecular, cellular and genomic approaches, mostly centered upon DNA and mRNA analysis. The gene in question is isolated, and expression and characterization of its coded protein product - i.e. the drug target - is invariably a secondary event.

With the proteomic approach, the starting point is at the other end of the 'telescope'. Here there is direct and immediate comparison of the proteomes from paired normal and disease materials. Examples of these pairs are:

1. Purified epithelial cell populations derived from human breast tumours, matched to purified normal populations of human breast epithelial cells.

2. The invading pathogenic hyphal form of *C. albicans*, matched to the non invading yeast form of *C. albicans.* When the proteome images from each pair are aligned, the Proteograph™ software is able to rapidly identify those proteins (each referenced as having a unique molecular cluster index, or MCI) that are either unique, or those that are differentially expressed.

SUMMARY: By combining with bioinformatics, proteomic technology has progressed substantially from the simple concept of 2DE into a series of technologies capable of investigating the total protein content of a biological system and its response to changing conditions. This technology has revolutionized the way in which researchers analyze the presence and relative abundance of proteins and expedite the screening and validation process for drug discovery.

Integrative functional informatics has an intrinsic framework of accuracy, quantification and reliability for biomarker analysis and can facilitate the assessment of various key factors in drug development pipeline, such as mapping of molecular drug targets and drug action mechanisms. This novel, integral use of experimental model will significantly save time and money, owing to the high-through output data acquisition and analysis systems at early stages of each exercise. However, application of this approach is rather limited at present due to the lack of well established model systems and accuracy.

Additionally, in analysis of a growing amount of data in proteomics, the problem with standardization and comparability should be addressed. Furthermore, generalized data formats should be created, since they play an important role in interpretation of data originating from different studies. Nevertheless, with the development of relevant databases, rapidly emerging information generated by bioinformatics will be integrated into more aspects of proteomic studies in biomarker discovery and drug target validation. The key point is the efficiency of the combination of bioinformatics with proteomics to use the rapidly emerging information and techniques available for accelerating biomarker discovery and drug development.

REFERENCES:

1. Reginald H. Garrett and Charles M. Grisham ,Textbook of Biochemistry,4th Edition,
2. Johann Gasteiger. The central role of Chemoinformatics. Chemo metrics and Intelligent Laboratory Systems 82 (2006) 200 – 209.
3. David A. Morrison. Application of bioinformatics to parasitology. International Journal for Parasitology 35 (2005) 463–464.
4. Mauno Vihinen. Review Bioinformatics in proteomics Biomolecular Engineering .18 (2001) 241–248.
5. Derek J. Smith Applications of bioinformatics and computational biology to influenza surveillance and vaccine strain selection Journal of Vaccine 21 (2003) 1758–1761.
6. M.N. Shuaibu et al. Selection and identification of malaria vaccine target molecule using bioinformatics and DNA vaccination / Vaccine *28 (2010) 6868–6875.*
7. Mark S Boguski Bioinformatic Current Opinion in Genetics and Development 1994, 4:383-388
8. Chemical proteomic and bioinformatic strategies for the identification and quantification of vascular antigens in cancer. *Journal of proteomics*: 73 1954 – 1973. (2010)
9. J. C. Alvarez, "High-throughput docking as a source of novel drug leads," *Curr. Opin. Chem. Biol.*, vol. 8, no.4, pp. 365-370, Aug. 2004.
10. Jose D. Debes, MD, and Raul Urrutia, MD, Rochester, Minn Bioinformatics tools to understand human diseases Surgical research review Surgery 2004; 135:579-85.
11. Jan C. Wiemer, Alexander Prokudin Bioinformatics in proteomics: application, terminology, and pitfalls Pathology – Research and Practice 200 (2004) 173–178.
12. Verena Strassberger, Tim Fugmann, Dario Neri, Christoph Roesli Chemical proteomic and bioinformatic strategies for the identification and quantification of vascular antigens in cancer .Journal of Proteomics73 (2010)1954 - 1973.
13. Tony Maschio and Tom Kowalski. Bioinformatics – a patenting view TRENDS in Biotechnology Vol.19 No.9 September 2001.
14. Brian Desany and Zemin Zhang .Bioinformatics and cancer target discovery. DDT Vol. 9, No. 18 September 2004
15. Robert Molidor, Alexander Sturn, Michael Maurer, Zlatko Trajanoski New trends in bioinformatics: from genome sequence to personalized medicine Experimental Gerontology .38 (2003) 1031–1036.
16. Sergey E. Ilyin, Albert Pinhasov, Anil H.Vaidya, Frank A. Amato, Jack Kauffman, Hong Xin, Patricia Andrade-Gordon, Emerging paradigms in applied bioinformatics.2BIOSILICO Vol. 1, No. 3 July 2003.
17. Jose C.M. Mombach, Ney Lemke, Norma M. da Silva, Rejane A. Ferreira, Eduardo Isaia, Cláudia K. Barcellos... Bioinformatics analysis of mycoplasma metabolism: Important enzymes, metabolic similarities, and redundancy.Computers in Biology and Medicine 36 (2006) 542–552.
18. 6Manchao Zhang , Xueliang Fang , Hongpeng Li, Ribo Guo ,Xiaojin Wu , Bihua Li , Feng Zhu, Yan Ling , Brian N. Griffith , Shaomeng Wang , Dajun Yang Bioinformatics-based discovery and characterization of an AKT-selective inhibitor 9-chloro-2-methylellipticinium acetate (CMEP) in breast cancer cells Cancer Letters 252 (2007) 244–258.
19. Taizo Hanai, Hiroyuki Hamada Masahiro Okamoto. Application of bioinformatics for DNA microarray data to bioscience, bioengineering and medical fields journal of bioscience and bioengineering, the society for biotechnology, Japan vol. 101, no. 5, 377–384. 2006.
20. Mauno Vihinen Bioinformatics in proteomics Biomolecular Engineering.18 (2001) 241–248.
21. Stefano Moretti, Athanasios V. Vasilakos An overview of recent applications of Game Theory to bioinformatics. Information Sciences 180 (2010) 4312–4322.
22. Verena Strassberger, Tim Fugmann, Dario Neri, Christoph Roesli Chemical proteomic and bioinformatic strategies for the identification and quantification of vascular antigens in cancer.
23. David Edwards and Jacqueline Batley.Plant bioinformatics: from genome to phenome. Trends in Biotechnology Vol.22 No.5 May 2004.
24. Jose D. Debes, MD, and Raul Urrutia, MD, Rochester, Minn Bioinformatics tools to understand human diseases Journel of Surgery 2004;135:579-85.
25. Jan C. Wiemer, Alexander Prokudin. Bioinformatics in proteomics: application, terminology, and pitfalls Pathology – Research and Practice 200 (2004) 173–178.
26. Michele R. Forman, Sarah M. Greene, Nancy E. Avis, Stephen H. Taplin, Paul Courtney, MS, Peter A. Schad, Bradford W. Hesse, Deborah M. Winn. Bioinformatics Tools to Accelerate Population Science and Disease Control Research. American Journal of Preventive Medicine. 2010; 38(6):646–651.
27. Timothy M.D. Ebbels, Rachel Cavill. Bioinformatic methods in NMR-based metabolic profiling Progress in Nuclear Magnetic Resonance Spectroscopy 55 (2009) 361–374.
28. M.N. Shuaibua, M. Kikuchia, M.S. Cherifa, G.K. Helegbea, T. Yanagib, K. Hirayamaa Selection and identification of malaria vaccine target molecule using bioinformatics and DNA vaccination Vaccine 28 (2010) 6868–6875.
29. Jensen, L.J. et al. (2002) Prediction of human protein function from posttranslational modifications and localization features. J. Mol. Biol. 319, 1257–1265.
30. Turk, B.E. and Cantley, L.C. (2003) Peptide libraries: at the crossroads of proteomics and bioinformatics. Curr. Opin. Chem. Biol. 7, 84–90.

****