International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 4 Issue 10|| October 2015 || PP.60-70

In silico Drug Design: Prospective for Drug Lead Discovery

Le Anh Vu[1], Phan Thi Cam Quyen[2], Nguyen Thuy Huong[3]
[1,3]*Faculty of Chemical Engineering, HoChiMinh City University of Technology, Vietnam*
[2]*KienGiang Center for Seeds and Seedlings of Agriculture, Forestry and Fisheries, Vietnam*

ABSTRACT: *The field of in silico drug design is a rapidly growing area in which many successes have occurred in recent years. The explosion of bioinformatics, cheminformatics, genomics, proteomics, and structural information has provided hundreds of new targets as well as new ligands. Therefore, in silico drug design represents computational methods and resources that are used to facilitate the opportunities for future drug lead discovery. This review reported a brief history of drug design and summarized the most important steps of in silico drug design strategy for the discovery of new molecular entities. The workflow of the entire virtual designing campaign is discussed, from the choice of a target, the evaluation of a structure of that target, ligand search, receptor theory to molecular docking, virtual high-throughput screening, the pivotal questions to consider in choosing a method for drug lead discovery and evaluation of the drug leads.*

KEYWORDS – *in silico drug design, computational modelling, virtual screening, molecular docking, drug lead discovery.*

I. INTRODUCTION

Drug discovery process is a critical issue in the pharmaceutical industry since it is a very costly and time consuming process to produce new drug potentials and enlarge the scope of diseases incurred [1]. Two different methods are widely used in the pharmaceutical industry for finding hits are: high throughput screening and virtual screening. In high throughput screening (HTS), the chemical compounds are synthesized, and screened against protein based or cell based assays. This process is commonly used in all major pharmaceutical industries. However, the cost in synthesis of each compound, *in vitro* testing and low hit rate are posing huge problems for pharmaceutical industries. Current efforts within the industry are directed to reduce the timeline and costs. Besides, HTS campaigns to identify compounds exerting a desired phenotype or entire pathways, many of these drugs are failing in clinical development either because of poor pharmacokinetic characteristics or to intolerable side effects, which may reflect insufficient specificity of the compounds [2]. At present, hundreds of thousands to millions of molecules have to be tested within a short period for finding novel hits, therefore, highly effective screening methods are necessary for today's researchers.

In view of the above problems in finding new drugs by HTS; cost effective, reliable virtual screening procedures are in practice. The so-called *in silico* approaches, using computational environments as their experimental laboratories [3].This review is intended to provide an overview of the process of *in silico* drug design from the selection of a target to the generation and evaluation of lead compounds. An in-depth discussion or evaluation of the computational methods involved in drug discovery will not be provided here, since that subject has been covered in reviews elsewhere [4–9].

II. DRUG DISCOVERY

Drugs are chemicals that prevent disease or assist in restoring health to diseased individuals. As such they play an indispensable role in modern medicine. Medicinal chemistry is that branch of science that provides these drugs either through discovery or through design. The classical drugs of antiquity were primarily discovered by empirical observation using substances occurring naturally in the environment. During the last two centuries, drugs increasingly were also prepared by chemical alteration of natural substances. In the century just past many novel drugs were discovered entirely by chemical synthesis. In the third millennium, all of these techniques are still in use and a researcher of drug design and development must appreciate their relative value. Added to this picture are novel opportunities made possible by deeper understanding of cell biology and genetics [10]. Drug discovery is one of the most crucial components of the pharmaceutical industry's Research and Development (R&D) process and is the essential first step in the generation of any robust, innovative drug pipeline [11]. The process of drug development aims towards the identification of compounds with pharmacological interest to assist in the treatment of diseases and ultimately to improve the quality of life. The compounds used in pharmacology are mainly small organic molecules (ligands) which interact with specific biomolecules (receptors) [12].

2.1 Traditional Drug Discovery Limitations

In the distant past, designing a new drug by changing the molecular structure of an existing drug was a slow process of trial and error. Now, a computer can display the molecular structure of any drug from a list of thousands in a database. With only very slight molecular changes, the original drug may be significantly changed in a variety of ways that influence absorption, metabolism, half-life, therapeutic effect, or side effects. The computer can also identify those chemicals that would probably not be successful in treating a particular disease before time and money are invested in extensive testing. Using computers to manipulate chemicals at the molecular level and design new drugs is based on molecular pharmacology, the study of the chemical structures of drugs and their interactions at the molecular level within a cell and even within DNA in the nucleus. Traditionally, drugs are discovered by synthesizing compounds in a time consuming multi-step process against a battery of *in vivo* biological screens and further investigating the promising candidates for their pharmacokinetic properties, metabolism and potential toxicity. Such a development process has resulted in high attrition rates with failures attributed to poor pharmacokinetics (39%), lack of efficacy (30%), animal toxicity (11%), adverse effects in humans (10%) and various commercial and miscellaneous factors [13].

There are an estimated 35,000 open reading frames in the human genome, which, in turn, generate an estimated 500,000 proteins in the human proteome. About 10,000 of those proteins have been characterized crystallographically. In the simplest terms, that means that there are about 490,000 unknowns that may potentially foil any scientific effort. This means that drug design is a very difficult task. A pharmaceutical company may have from 10 to 100 researchers working on a drug design project, which may take from 2 to 10 years to get to the point of starting animal and clinical trials. Even with every scientific resource available, the most successful pharmaceutical companies have only one project in ten succeed in bringing a drug to market. Drug design projects can fail for a myriad of reasons. Some projects never even get started because there are no adequate assays or animal models to test for proper functioning of candidate compounds. Some diseases are so rare that the cost of a development effort would never be covered by product sales (as in the case of orphan drugs). Even when the market exists, and assays exist, every method available may fail to yield compounds with sufficiently high activity. On the other hand, compounds that are active against the disease may be too toxic, not bioavailable, or too costly to manufacture. Recent estimates of how much it costs to bring a drug to the market have ranged from $300 million to $1.7 billion. A single laboratory researcher's salary, benefits, laboratory equipment, chemicals, and supplies can cost in the range of $200,000 to $300,000 per year. Some typical costs for various types of experiments are listed in Table 1, owing to the enormous costs involved, the development of drugs is primarily undertaken by big pharmaceutical companies. Indeed, the dilution of investment risk over multiple drug design projects pushes pharmaceutical companies to undertake many mergers in order to form massive corporations. Because of all these reasons, it is necessary to effectively leverage every computational tool that can help to achieve successful results [14].

Table 1. Typical costs of experiments [14].

Experiment	Typical Cost per Compound ($)
Computer modeling	10
Biochemical assay	400
Cell culture assay	4,000
Rat acute toxicity	12,000
Protein crystal structure	100,000
Animal efficacy trial	300,000
Rat 2-years chronic oral toxicity	800,000
Human clinical trial	500,000,000

2.2 Drug Design

Drug design, sometimes referred to as rational drug design (or more simply rational design), is the inventive process of finding new medications based on the knowledge of biological targets [11]. Rational drug design can be broadly divided into two categories: development of small molecules with desired properties toward targets, biomolecules (proteins or nucleic acids), whose functional roles in cellular processes and 3D structural information are known. This approach in drug design is well established, being applied extensively by the pharmaceutical industries. Another approach is development of small molecules with predefined properties toward targets, whose cellular functions and their structural information may be known or unknown [15]

In the most basic sense, drug design involves design of small molecules that are complementary in shape and charge to the biomolecular target to which they interact and will, therefore, bind to it. The identification of a potential drug target is valuable and significant in the research and development of drug molecules at early stages. Due to the limitation of throughput, accuracy and cost, experimental techniques cannot be applied

widely. Therefore, the development of *in silico* target identification algorithms, as a strategy with the advantage of fast speed and low cost, has been receiving more and more attention worldwide. It has been of great importance to develop a fast and accurate target identification and prediction method for the discovery of targeted drugs, construction of drug-target interaction network as well as the analysis of small molecule regulating network [16].

III. IN SILICO DRUG DESIGN

3.1 *In Silico* Drug Design

In silico is a term that means "computer aided". The phrase was coined in 1989 as an analogy to the Latin phrases *in vivo*, *in vitro*, and *in situ*. So *in silico* drug design means rational design by which drugs are designed/discovered by using computational methods. According to Kubinyi [17], most of the drugs in the past were discovered by coincidence or trial and error method, or in other words, serendipity played an important role in finding new drugs.

Current trend in drug discovery is shifted from discovery to design, which needs understanding the biochemistry of the disease, pathways, identifying disease causative proteins and then designing compounds that are capable of modulating the role of these proteins. This has become common practice in biopharmaceutical industries. Both experimental and computational methods play significant roles in the drug discovery and development and most of the times run complementing each other [18].

The main aim of computer aided drug design (CADD) is to bring the best chemical entities to experimental testing by reducing costs and late stage attrition [19]. CADD involves:

a. Computer based methods to make more efficient drug discovery and development process.

b. Building up chemical and biological information databases about ligands and targets/proteins to identify and optimize novel drugs.

c. Devising *in silico* filters to calculate drug likeness or pharmacokinetic properties for the chemical compounds prior to screening to enable early detection of the compounds which are more likely to fail in clinical stages and further to enhance detection of promising entities.

There are various computational techniques which are capable of producing the desired effect at various stages of the drug discovery process [20]. The two major disciplines of CADD which can manipulate modern day drug discovery process and which are capable of accelerating drug discovery are bioinformatics and cheminformatics. In general:

a. Bioinformatic techniques hold a lot of prospective in target identification (generally proteins/enzymes), target validation, understanding the protein, evolution and phylogeny and protein modeling [21].

b. Cheminformatic techniques hold a lot of prospective in storage management and maintenance of information related to chemical compounds and related properties, and importantly in the identification of novel bioactive compounds, and further in lead optimization. Besides, cheminformatic methods are extensively utilized in *in silico* ADME (Absorption, Distribution, Metabolism and Elimination) prediction and related issues that help in the reduction of the late stage failure of compounds [20].

3.2 Why Computer Aided Drug Discovery?

Besides the significant costs and time associated in bringing a new drug to the market [22], some of the major reasons for the pharmaceutical industries to look for alternative or complementary methods to experimental screening are:

a. In a survey study, five of the 40,000 compounds tested in animals reach human testing and only one out of these five reaching the clinical trials is finally approved [19].

b. On the other hand, the tremendous increment in chemical space and target proteins/receptors increases the demand for the HTS and will in turn call for new lead identification strategies (rational approaches) to reduce costs and enhance efficacy.

c. Advances in computing technologies on software and hardware have enabled reliable computational methods.

IV. OVERVIEW OF THE PROCESS

The process of *in silico* drug design is an iterative one (see Fig. 1) and often proceeds through multiple cycles before an optimized lead goes into clinical assay. The first cycle includes the cloning, purification and structure determination of the target protein or nucleic acid by one of three principal methods: X-ray crystallography, nuclear magnetic resonance (NMR), or homology modeling. Using computer algorithms, compounds or fragments of compounds from a database are positioned into a selected region of the structure (docking). These compounds are scored and ranked based on their steric and electrostatic interactions with the target site, and the best compounds are tested with biochemical assays. In the second cycle, structure determination of the target in complex with a promising lead from the first cycle, one with at least micromolar inhibition *in vitro*, reveals sites on the compound that can be optimized to increase potency. Additional cycles include synthesis of the optimized lead, structure determination of the new target: lead complex, and further

optimization of the lead compound. After several cycles of the drug design process, the optimized compounds usually show marked improvement in binding and, often, specificity for the target [23].

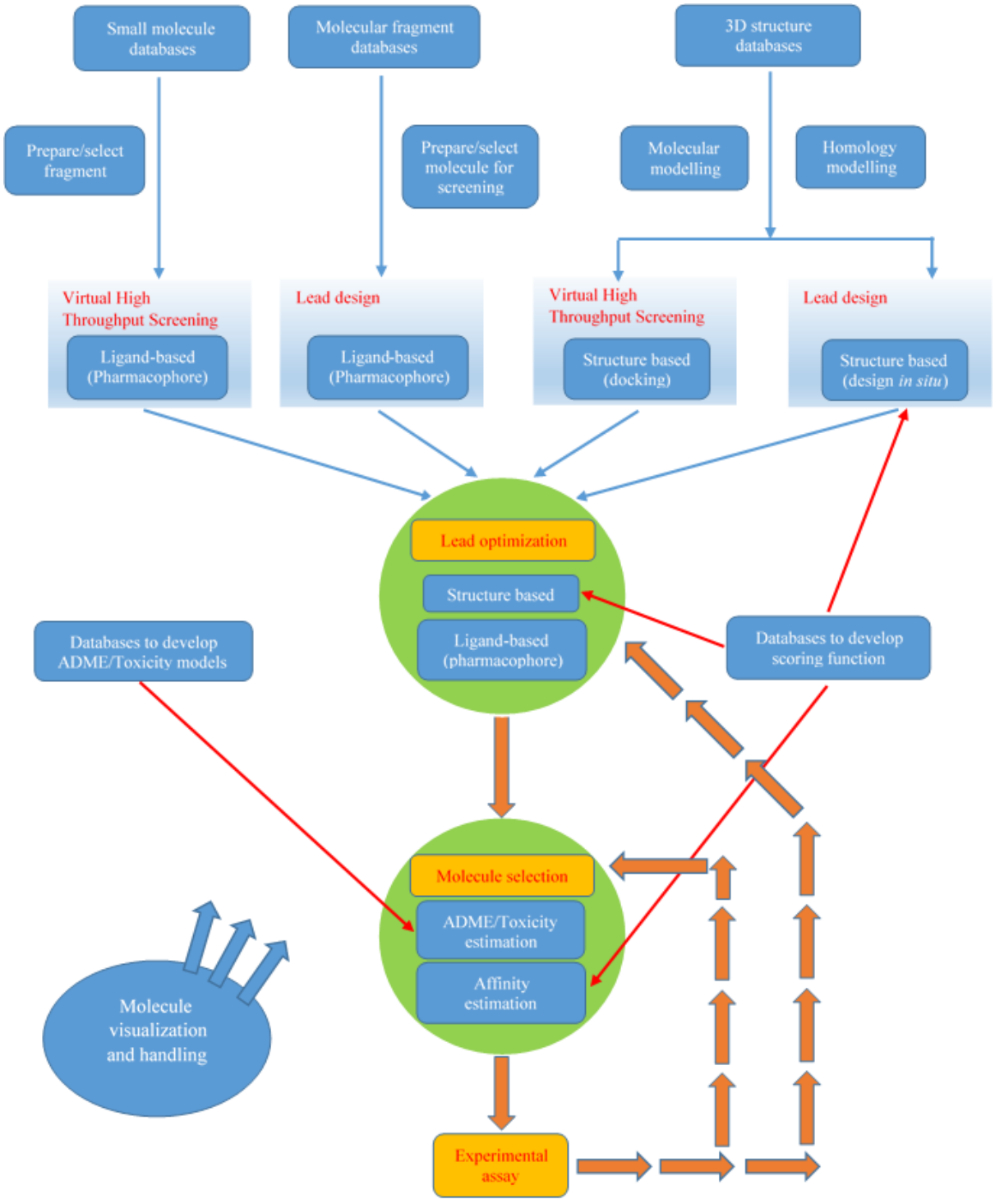

Figure 1. The iterative process of *in silico* drug design.

V. STRATEGIES OF IN SILICO DESIGN

In silico drug design can be applied by either of two strategies of design depending on the knowledge of the target, presence of the primary sequence and 3D structure. These two strategies are:

5.1 Structure Based Drug Design

Structure based drug design (SBDD) is one of the earliest techniques used in drug design. 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 NMR techniques. X-ray and NMR methods can resolve the structure of proteins to a resolution of a few angstroms [1].

However structure based drug design is not a single tool or technique. It is a process that incorporates both experimental and computational techniques. This is generally the preferred method of drug design, since it has the highest success rate. In the drug design stage of SBDD, docking is the preferred tool for giving a computational prediction of compound activity [14]. The following steps are mostly used in SBDD:

Target Determination

Drug Target: is a biomolecule which is involved in signaling or metabolic pathways that are specific to a disease process. Biomolecules play critical roles in disease progression by communicating through either protein–protein interactions or protein–nucleic acid interactions leading to the amplification of signaling events and/or alteration of metabolic processes. In structure based drug design, a known 3D structure of the target is the initial step in target identification. This is usually determined either by X-ray crystallography or by NMR to identify its binding site, the so called active site [15].

Homology Modelling: if crystallographic coordinates or a 2D NMR models are not available, then a homology model is usually the next best way for determining the protein structure. A homology model is a three-dimensional protein structure that is built up from fragments of crystallographic models. Thus, the shape of an α-helix may be taken from one crystal structure, the shape of a β-sheet taken from another structure, and loops taken from other structures. These pieces are put together and optimized to give a structure for the complete protein. Often, a few residues are exchanged for similar residues, and some may be optimized from scratch. Homology models may be very accurate or very marginal, depending upon the degree of identity and similarity that the protein bears to other proteins with known crystal structures. Since the homology model building process is dependent upon utilizing crystal structure coordinates for similar proteins (called the template), a crucial factor to consider is how similar the unknown sequence should be to the template protein. A number of metrics have been suggested for this. One of the most conservative metrics suggests that there should be over 70% sequence identity (not similarity) with the template, in order to get a homologous model that can be trusted. Other metrics suggest having over 30% or 40% sequence identity with the template. One study showed that having 60% or more sequence identity gave a success rate greater than 70%. With higher sequence identities, the percent of error is decreased, where as many as 10% of homology models may have a root mean square deviation (RMSD) greater than 5A° (which represent error cutoff) [14]. In order to clarify the seemingly disparate metrics mentioned in the previous paragraph, Rost carried out an extensive study looking at how much sequence identity is needed to get a good homology model as a function with the number of aligned residues. For a small sequence of 25 aligned residues, 60% identity was necessary. For a large region of 250 aligned residues, templates with over 20% identity could give good homology models [24]. The metrics used by Rost are somewhat less conservative than some of the other metrics. Rost's results also reflect improvements in homology model software and methodology compared with earlier work. Percent similarity is also a useful metric to examine. If several potential templates have essentially the same percent identity, then the one with the highest percent similarity may be chosen. Researchers may also choose the one in which the crystal structure has the best resolution [14].

Protein Folding: another method for target identification is protein folding. This is a difficult process which starts with the primary sequence only and runs a calculation that tries an incredibly large number of conformers. This is an attempt to compute the correct shape of the protein based on the assumption that the correct shape has the lowest energy conformer. This assumption is not always correct, since some proteins are folded to conformers that are not at the lowest energy with the help of chaperones. It is also difficult to write an algorithm that can determine when disulfide bonds should be formed. So, sometimes protein folding gives an accurate model, and sometimes it gives a rather poor model. The real problem with protein folding is that there is no reliable way to tell whether it has given an accurate model. There are only some checks that provide some circumstantial evidence that the model might be good or bad. For example, one can check if hydrophilic residues are on the exterior of the protein and hydrophobic residues are on the interior. So, pharmaceutical

companies can be justifiably hesitant to spend millions of dollars on research and development based on a folded protein model when there is no way to have confidence in the accuracy of that model. For this reason, protein folding tends to be the last resort for building three-dimensional protein models. Homology model building has two important advantages over protein folding. First, it is more accurate on average. Second, and more importantly, the researcher can get a better estimate of whether the homology model is likely to be qualitatively and quantitatively accurate, based on the degree of similarity to a known structure. The role and reliability of homology model building is increasing as the number of available crystal structures increases. Knowing the three-dimensional structure of a protein is only the beginning of understanding it. It is also important to understand the mechanism of chemical reactions involving that protein, where it is expressed in the body, the pharmacophoric description, and the mechanism of binding with chemical inhibitors [14].

Ligands Search

Much of drug design is a refinement process. In this process, successive changes are made on molecular structures in order to improve activity. However, the process needs to get started with some compounds having at least marginal activity. There are often a couple of known inhibitors from previous studies on the target, or very similar targets. There often needs to be at least one known inhibitor in order to provide a reference for the development of an assay [14].

In silico screening of chemical compound databases for the identification of novel chemotypes is termed as Virtual Screening (VS). VS is generally performed on commercial, public or private 2-dimensional or 3-dimensional chemical structure databases. Virtual screening is employed to reduce the number of compounds to be tested in experimental stages, thereby allowing focusing on more reliable entities for lead discovery and optimization [25-28]. The costs associated with the virtual screening of chemical compounds are significantly lower when compared to screening of compounds in experimental laboratories. Virtual screening methods are mainly driven by the availability of the existing knowledge. Depending on already existing knowledge on the drug targets and potential drugs, these methods fall mainly in these two categories (Fig. 2):

a. Structure based virtual screening (SBVS).
b. Ligand based virtual screening (LBVS).

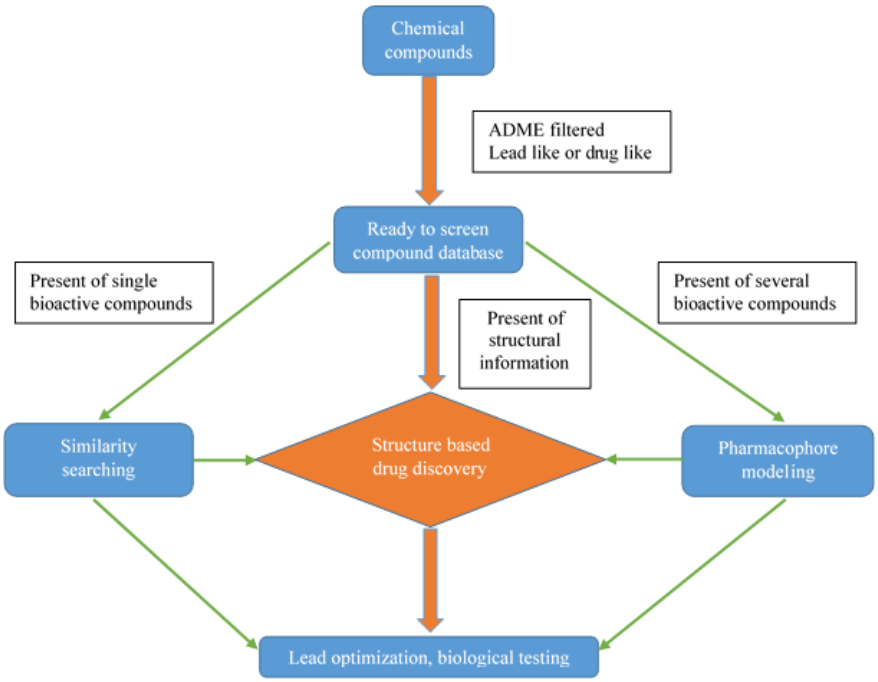

Figure 2. Schematic representation of virtual screening methods [29].

In the absence of receptor structural information and when one or more bioactive compounds are available, ligand based virtual screening (LBVS) is generally utilized [30-32]. This screening method can be carried out by either of the following approaches:

a. Similarity search: similarity searching is performed when a single bioactive compound is available. The basic principle behind similarity searching is to screen databases for similar compounds with the backbone of the lead molecule.

b. Pharmacophore-based virtual screening: pharmacophore is the three-dimensional geometry of interaction features that a molecule must have in order to bind in a protein's active site. These include such

features as hydrogen bond donors and acceptors, aromatic groups, and bulky hydrophobic groups. When one or several bioactive compounds are available, pharmacophore-based virtual screening is performed. The principle behind the pharmacophore is a set of chemical features; their arrangement in a 3-Dimensional space is responsible for the bioactivity of the compound [14]. By utilizing the chemical features of already known bioactive compounds, a pharmacophore model is built, which later is used to screen against database of unknown compounds for finding chemical compounds with similar chemical features (Fig. 3).

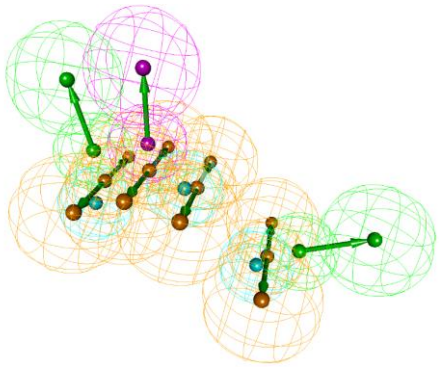

Figure 3. Example of a pharmacophore model [33].

Molecular Docking

Docking is an automated computer algorithm that determines how a compound will bind in the active site of a protein. This includes determining the orientation of the compound, its conformational geometry, and the scoring (Fig. 4). The scoring may be a binding energy, free energy, or a qualitative numerical measure. In some way, every docking algorithm automatically tries to put the compound in many different orientations and conformations in the active site, and then computes a score for each. Some programs store the data for all of the tested orientations, but most only keep a number of those with the best scores [14]. In general, there are two key components of molecular docking [29], as follows:

a. Accurate pose prediction or binding conformation of the ligand inside the binding site of the target protein.

b. Accurate binding free energy prediction, which later is used to rank the order of the docking poses.

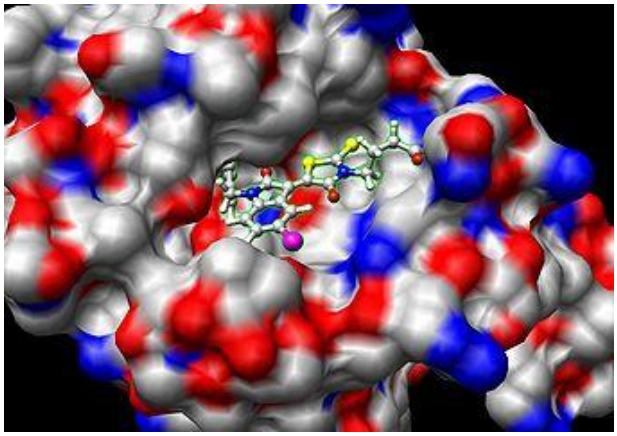

Figure 4. Example of molecular docking between chemical compound and a protein.

The docking algorithm usually carries out the first part of the docking (predicting binding conformation) and the scoring function associated with the docking program carries out the second part that is binding free energy calculations. The key components of molecular docking are further displayed below.

a. Pose prediction: docking algorithms usually perform pose predictions which aim to identify molecular features that are responsible for molecular recognition. Pose predictions are very complex and often difficult to understand when simulated on a computer [34].

b. Activity prediction: after the pose prediction by docking algorithm, the immediate step in the docking process is activity prediction, which is also termed scoring. Docking score is achieved by the scoring functions associated with the particular docking software. Scoring functions are designed to calculate biological activity by estimating the interactions between the compound and protein target. During the early stages of the

docking experiments, scoring was performed based on the simple shape and electrostatic complementarities. However, currently, the docking conformers are often treated with sophisticated scoring methods that include the Van der Waals interactions, electrostatic interactions, solvation effects and entropic effects [35], [14].

Docking Algorithms

Depending on the flexibility of protein and ligand, docking algorithms can be divided into 3 types [36], as follows:

a. Rigid docking: protein and ligand are consider to be rigid
b. Semi-flexible docking: protein is fixed and ligand is flexible
c. Flexible docking: both protein and ligand are flexible

Based on the principle of conformation generation, the search methods are categorized into:

a. Stochastic
b. Systematic
c. Deterministic

The search algorithm positions molecules in various locations, orientations, and conformation within the active site. Some of the earliest docking programs positioned a molecule in the active site, holding it rigid with respect to conformational changes, but all modern docking algorithms include ligand conformational changes. The choice of search algorithm determines how thoroughly the program checks possible molecule positions, and how long it takes to run. The search algorithm does not determine whether the docking program gives accurate results. But the scoring function is responsible for determining whether the orientations chosen by the search algorithm are the most energetically favorable, and is responsible for computing the binding energy. Thus, a search algorithm that does not sample the space thoroughly will give inaccurate results if the correct orientation is not sampled. However, most search functions will sample the space adequately if they are given the correct input parameters. Many search algorithms have been developed depending on the principles of conformation generation. One of the earliest used algorithms was Monte Carlo search algorithm, which is built around a random number generator. In the simplest implementation, position, orientation, and conformation are all chosen at random. Sometimes, position and conformation are checked independently. Thus, a position is chosen and many conformations are tested while in that position; then a new position is chosen, and the process repeats. Another important algorithm is the tabu search algorithm. Most tabu searches are implemented as a modified version of the Monte Carlo search. Like the Monte Carlo search, the tabu search chooses orientations and confirmations randomly. However, the Monte Carlo algorithm utilizes no knowledge of what positions have already been sampled, and thus sometimes results in recomputing positions that have already been computed. The tabu algorithm keeps track of which positions have already been sampled, and avoids sampling those positions again. Thus, it can give the same results with fewer iterations, by eliminating any duplication of work. Genetic algorithms can sample a space thoroughly, if the parameters are chosen wisely, and can run very quickly. Many docking algorithms were originally developed to simulate the ligand binding in a crevice in the surface of the folded protein. Some programs have difficulty in docking compounds in the active site that is completely enclosed. This can happen when the protein folds down over the active site or the entire active site opens and closes via a clam shell movement of two large sections of the protein. When this occurs, additional inputs are needed that will allow the docking program to function correctly with an encapsulated active site [14].

Scoring Functions

One of the two important components of molecular docking is scoring. While docking aims at reproducing binding conformation close to the X-ray crystal structure, the aim of scoring is to quantify the free energy associated with protein and ligand in the formation of the protein-ligand interactions. Most of the docking softwares are equipped with scoring functions, which enable computing free energy associated with protein-ligand interactions (docking score). The docking score is used to rank the chemical compounds in a virtual screening campaign. Wide ranges of scoring functions are available to calculate the binding between the protein and virtual ligand. These methods range from estimating binding by a simple shape and electrostatic complementarities to the estimation of free energy of protein and ligand complex in aqueous solutions. Only few of them are capable of addressing the thermodynamic process involved in the binding process. However, methods based on thermodynamic parameters require an extensive simulating time, and consequently significant central processor unit (CPU) time. Therefore, these methods are restricted to a smaller set of compounds, making it impractical to use them in large-scale virtual screening experiments. Currently, three main types of scoring functions are applied: force field-based, empirical scoring functions and knowledge based scoring functions [37].

Force field-based scoring functions: This type relies on the molecular mechanics methods. Force field-based methods calculate both the protein-ligand interaction energy and ligand internal energy and later sum both the energies. The following represents total energy equation based on force field:

$E_{total} = E_{bonded} + E_{nonbonded}$ where the components of the covalent and non-covalent contributions are given by the following summations:

$E_{bonded} = E_{bond} + E_{angle} + E_{dihedral}$

$E_{nonbonded} = E_{electrostatic} + E_{Van\ der\ Wals}$

where:

E_{bond} represent potential energy of covalent bonds.

E_{angle} represent potential energy between angled bonds.

$E_{dihedral}$ represent potential energy of tortion of bonded atoms.

$E_{electrostatic}$ represent potential energy of electrostatic forces.

$E_{Van\ der\ Waals}$ represent potential energy of Van der Waals forces.

Different force field functions are based on different force field parameter sets. For example, AutoDock relies on the Amber force field and G-Score relies on the Tripos force field [37]. Van der Waals and electrostatic energy terms describe both the internal energy of the ligand and the interactions between the protein and ligand. The van der Waals energy term is described by the Lennard Jones potential. Electrostatic terms are described by the Coulombic formula with a distance dependent dielectric constant for charge separation. Advantages of force field-based scoring functions include accounting of solvent, and disadvantages include over-estimation of binding affinity [37] and arbitrarily choosing non bonded cutoff terms [34].

Knowledge based scoring functions: It uses atom pair interaction potentials as in potential of mean force (PMF). Atom pair interaction potentials are usually derived from structural information stored in the databases (ChemBridge structural database and Protein Data Bank) of protein-ligand complexes. It relies on the assumption that repeated occurrence of close intermolecular interactions between certain types of functional groups or atom types are energetically more favorable than the randomly occurring interactions, thus complementarily contribute to the binding affinity. The robust nature of this scoring function makes it usable in virtual screening. Knowledge based scoring functions rely on existing intermolecular interaction databases. One major limitation of this method is the limited availability of such structural information in the intermolecular interaction databases. D-score [38] and PMF scoring functions rely on knowledge based scoring functions [39].

Empirical scoring functions: The score in the empirical scoring function is derived from the individual energy contributions of each component involved in the intermolecular interactions, as shown in the equation below:

$$\Delta G_{bind} = \Delta G_{desolvation} + \Delta G_{motion} + \Delta G_{configuration} + \Delta G_{interaction}$$

where:

desolvation – enthalpic penalty for removing the ligand from solvent.

motion – entropic penalty for reducing the degrees of freedom when a ligand binds to its receptor.

configuration – conformational strain energy required to put the ligand in its "active" conformation.

interaction – enthalpic gain for "resolvating" the ligand with its receptor [40].

Empirical scoring functions are easier to apply and are subjected to less computational error. For example, Kuntz in his early work emphasized on the molecular shape, because shape complementarity is certainly essential for a ligand to be placed in the binding site and can be easily and accurately computed. However, in his later work he added chemical information, molecular mechanical energies, and empirical hydrophobicities to make the scoring function more accurate [41-42]. Böhm developed another empirical scoring function that takes into account hydrogen bonding, ionic interactions, lipophilic contact surface and a number of rotatable bonds [40], [2]. Due to their robust nature, empirical scoring functions are widely used in virtual screening experiments along with knowledge based scoring functions. One of the major limitations of the empirical scoring function is that it works very well with rigid ligands, but the results are not satisfying with flexible ligands. This is because most of the empirical scoring functions ignore the internal energy of the ligand. Scorings such as ChemScore (docking tool) [43] and Ludi (de novo design tool) [40] rely on the empirical scoring function.

5.2 Ligand Based Drug Design

In the past century many drug's target proteins were unknown. This is still fairly common, although it is slowly becoming less so as the body of knowledge about biological systems expands. The success of the design is greater if the target is known and a structure-based drug design process can be followed. However, there are times when there is a good reason for using a drug design without a known target. For example, cell surface receptors make excellent drug targets, but are very difficult to crystallize. So if homology modelling was unreliable or low identity score for the homolog protein was observed, in this case the techniques used for structure-based drug design cannot be used. Pharmacophore models and 3D-QSAR models can be used instead. A 3D-QSAR is a computational procedure used for quantitatively predicting the interaction between a molecule and the active site of a specific target. The great advantage of a 3D-QSAR is that it is not necessary to know what the active site looks like. Thus, it is possible to use this technique when the target is unknown. A 3D-QSAR is a mathematical attempt to define the properties of the active site without knowing its structure. This is

done by computing the electrostatic and steric interactions that an imaginary probe atom would have if it were placed at various positions on a grid surrounding a known active compound. In some cases, other interactions, such as hydrogen bonding, will also be included. After doing this for multiple active compounds, a partial least squares algorithm can be used to determine what spatial arrangement of features there could be in an active site that interacts with the known active molecules [14].

VI. CONCLUSION

In silico drug design is a powerful method, especially when used as a tool within an apparatus, for discovering new drug leads against important targets. After a target and a structure of that target are defined, new leads can be designed from chemical principles or chosen from a subset of small molecules that scored well when docked *in silico* against the target. After a preliminary evaluation of bioavailability, the candidate leads continue in an iterative process of reentering structural determination and reevaluation for optimization. Focused libraries of synthesized compounds based on *in silico* strategy can create a very promising lead which can continue to clinical trials. As structural genomics, bioinformatics, cheminformatics, proteomics and computational power continue to explode with new advances, further successes in *in silico* drug design are likely to follow. Each year, new targets are being diagnosed, structures of those targets are being determined at an amazing rate, and our capability to capture a quantitative picture of the interactions between macromolecules and ligands is accelerating.

ACKNOWLEDGEMENTS

This review is funded by HoChiMinh University of Technology under grant number TNCS-2015-KTHH-05.

REFERENCES

[1]. V. Rao and K. Srinivas. Modern drug discovery process: An *in silico* approach. *Journal of Bioinformatics and Sequence Analysis, 2*, 2011, 89-94.
[2]. H. Böhm, G. Schneider, H. Kubinyi, R. Mannhold and H. Timmerman. *Virtual screening for bioactive molecules (Methods and principles in medicinal chemistry)* (Wiley-VCH, Weinheim, Germany, 2000).
[3]. H. Noori and R. Spanagel. *In silico* pharmacology: drug design and discovery's gate to the future. *In silico Pharmacology*, *1*, 2013, 1-2.
[4]. R. Taylor, P. Jewsbury and J. Essex. A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des., 16*, 2002, 151–166.
[5]. D. Joseph-McCarthy. Computational approaches to structure-based ligand design. *Pharmacol. Ther., 84*, 1999, 179–191.
[6]. R. Bohacek and C. McMartin. Modern computational chemistry and drug discovery: structure generating programs. *Curr. Opin. Chem. Biol., 1*, 1997, 157–161.
[7]. H. Carlson and J. McCammon. Accommodating protein flexibility in computational drug design. *Mol. Pharmacol., 57*, 2000, 213–218.
[8]. B. Shoichet, S. McGovern, B. Wei and J. Irwin. Lead discovery using molecular docking. *Curr. Opin. Chem. Biol., 6*, 2002, 439–446.
[9]. G. Klebe and H. Bohm. Energetic and entropic factors determining binding affinity in protein-ligand complexes. *J. Recept. Signal Transduct. Res.*, 17, 1997, 459–473.
[10]. U. Madsen, P. Krogsgaard-Larsen and T. Liljefors. *Textbook of drug design and discovery* (Taylor and Francis, Washington, USA, 2002).
[11]. S. Arlington. An industrial revolution in R&D. *Pharmaceutical Executive*, *20*, 2000, 74-84.
[12]. D. Plewczynski, A. Philips, M. von Grotthuss, L. Rychlewski and K. Ginalski. HarmonyDOCK: the structural analysis of poses in protein-ligand docking. *Journal of Computational Biology*, *18*, 2010, 1-10.
[13]. K. Gunjan, S. Dinesh, V. Yogesh and S. Vishal (2013). A review on drug designing, methods, its applications and prospects. *International Journal of Pharmaceutical Research and Development*, *5*, 2013, 15-30.
[14]. D. Young. *Computational drug design* (John Wiley & Sons, New Jersey, USA, 2009).
[15]. S. Mandal, M. Moudgil and S. Mandal (2009). Rational drug design. *European Journal of Pharmacology*, *625*, 2009, 90-100.
[16]. H. Markus, J. Seifert and K. Bernd (2007). Virtual high-throughput screening of molecular databases. *Journal of Current Opinion in Drug Discovery and Development*, *10*, 2007, 298-307.
[17]. H. Kubinyi. Chance favors the prepared mind from serendipity to rational drug design. *Journal of Receptors and Signal Transduction Research*, *19*, 1999, 15–39.
[18]. J. Bajorath. Integration of virtual and high-throughput screening. *Nature Reviews Drug Discovery*, *1*, 2002, 882–894.
[19]. I. Kapetanovic. Computer-aided drug discovery and development (caddd): *In silico* chemico-biological approach. *Chemico-Biological Interactions*, *171*, 2008, 165–176.
[20]. H. Hoeltje, W. Sippl, D. Rognan and G. Folkers. *Molecular modeling, basics principles and applications* (John Wiley & Sons, Chichester, U.K, 2003).
[21]. T. Lengauer. *Bioinformatics from genome to drugs* (Wiley- VCH, Weinheim, Germany, 2001).
[22]. K. Bleicher, H. Böhm, K. Müller and A. Alanine. Hit and lead generation: beyond high-throughput screening. *Nature Reviews. Drug Discovery*, *2*, 2003, 369–378.
[23]. Amy C. Anderson. The process of structure-based drug design. *Chemistry & Biology*, *Vol. 10*, 2003, 787–797.
[24]. B. Rost. Twilight zone of protein sequence alignments. *Protein Engineering*, *12*, 1999, 85-94.
[25]. J. Koh. Making virtual screening a reality. *Proceedings of the National Academy of Science USA*, *100*, 2003, 6902–6903.
[26]. H. Köppen. Virtual screening - what does it give us? *Current Opinion in Drug Discovery & Development*, *12*, 2009, 397–407.

[27]. S. Subramaniam, M. Mehrotra and D. Gupta. Virtual high throughput screening (vhts) - a perspective. *Bioinformation*, *3*, 2008, 14–17.

[28]. U. Rester. From virtuality to reality - virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective. *Current Opinion in Drug Discovery and Development*, *11*, 2008, 559–568.

[29]. A. Leach and V. Gillet. *An introduction to cheminformatics* (Kluwer Academic publishers, Netherlands, 2003)

[30]. C. McInnes.Virtual screening strategies in drug discovery. *Current Opinion in Chemical Biology*, *11*, 2007, 494–502.

[31]. A. Reddy, S. Pati, P. Kumar, H. Pradeep and G. Sastry. Virtual screening in drug discovery – a computational perspective. *Current Protein and Peptide Science*, *8*, 2007, 329–351.

[32]. A. Jain. Virtual screening in lead discovery and optimization. *Current Opinion in Drug Discovery & Development*, *7*, 2004, 396–403.

[33]. S. Yang. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discovery Today*, *15*, 2010, 444-450.

[34]. D. Kitchen, H. Decornez, J. Furr and J. Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*, *3*, 2004, 935–949.

[35]. H. Gohlke and G. Klebe. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie (International Edition in English)*, *41*, 2002, 2644–2676.

[36]. V. Kasam. *In silico drug discovery on computational Grids for finding novel drugs against neglected diseases*. PhD Thesis. University of Bonn, Germany, 2009.

[37]. N. Moitessier, P. Englebienne, D. Lee, J. Lawandi and C. Corbeil. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *British Journal of Pharmacology*, *153*, 2008, S7–S26.

[38]. H. Gohlke, M. Hendlich and G. Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *Journal of Molecular Biology*, *295*, 2000, 337–356.

[39]. I. Muegge and Y. Martin. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *Journal of Medicinal Chemistry*, *42*, 1999, 791–804.

[40]. H. Böhm. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *Journal of Computer-Aided Molecular Design*, *8*, 1994, 243–256.

[41]. I. Kuntz, J. Blaney, S. Oatley, R. Langridge and T. Ferrin. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, *161*, 1982, 269–288.

[42]. I. Kuntz. Structure-based strategies for drug design and discovery. *Science*, *257*, 1992, 1078–1082.

[43]. M. Eldridge, C. Murray, T. Auton, G. Paolini and R. Mee. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Computer Aided Molecular Design*, *11*, 1997, 425–445.