Biotechnology and Molecular Biology Review Vol. 6(4), pp. 88-93, April 2011
Available online at http://www.academicjournals.org/BMBR
ISSN 1538-2273 ©2011 Academic Journals

Review

In silico modeling in conjunction with natural products: Paving the way for rational drug-design

Shailza Singh[1]* and D. K. Sharma

Center for Energy Studies, Indian Institute of Technology Delhi, Hauz Khas, New Delhi-110016, India.

Accepted 16 February, 2011

Genome sequencing projects has produced a vast wealth of data describing the protein coding regions of the genome under study. However, only a minority of the protein sequences identified has a clear sequence homology to a known protein. In such cases valuable three-dimensional models of the protein coding sequence can be constructed by homology modeling methods. Threading methods uses specialized schemes to relate protein sequences to a library of known structures. Even in cases where there is no clear sequence homology, they have been shown to be able to identify the likely protein fold. As collaborative efforts in systematic structure determination begins to develop in the future, the number of protein sequences that cannot be assigned to a structural class by homology or threading methods will decrease, simply because they belong to a previously unidentified protein folding class. Moreover, the differences in substrate specificity may be explained on the basis of the predicted structures of the protein and its complex with the substrate. Natural products have been a rich source in providing leads for the development of drugs for the treatment of different infections. For this reason, *in silico* modeling methods in conjunction with natural products are likely to become increasingly useful in the near future for structure-based drug design approaches.

Key words: homology modeling, natural products, drug discovery.

INTRODUCTION

It is generally recognized that drug discovery and development are very time and resources consuming processes. There is an ever growing effort to apply computational power to the combined chemical and biological space in order to streamline drug discovery, design, development and optimization. In biomedical arena, computer-aided or *in silico* design is being utilized to expedite and facilitate hit identification, hit-to-lead selection, optimize the absorption, distribution, metabolism, excretion, toxicity profile (ADME/Tox), and avoid safety issues. Commonly used computational approaches include ligand-based drug design, structure-based drug design and quantitative structure-activity relationships. The structure-based design methods used to optimize these leads into drugs are now often applied much earlier in the drug discovery process. Protein structure is used in target identification and selection (the assessment of the 'druggability' or tractability of a target), in the identification of hits by virtual screening and in the screening of fragments. Additionally, the key role of structural biology during lead optimization to engineer increased affinity and selectivity into leads remains as important as ever.

Significant improvements in the era of genomics and proteomics and concurrent progresses in bioinformatics, X-ray crystallography and Nuclear Magnetic Resonance techniques have given rise to the expectation that the three dimensional structure or reliable homology modeling of target proteins can be achieved in a reasonably short time. Furthermore, considerable effort is also spent on sequencing the genome of various organisms. The human genome is expected to be fully characterized in the next few years. Metabolic pathway analysis of complete genome sequences enables identification of all members of essential pathways that are present as well as genes that are missing from individual pathogens. It is likely that structural genomics will carry on present efforts and that substantial knowledge will be gathered in the near future about the increasing number of drug targets. This will considerably increase the demand for the design of new specific inhibitors tailored

*Corresponding author. E-mail: shailza_iitd@yahoo.com. Tel: 011-26591256. Fax: +91-11-26581121.

to a particular target. Automation of lead compound design *in silico* given the structure of the protein target and a definition of its active site vies for the top of the wish in any drug discovery programme. The advances in this area are rapid but are constantly confronted with the question of viability due mostly to compromises necessitated by computational expediencies. Thus, what is required is a practical scheme based in the theoretical rigors of first principles and assured of transferability across systems at least as a benchmark for enabling a systematic growth of the field.

The majority of drugs available today were discovered either from chance observations or from the screening of synthetic or natural product libraries. Natural products, either as pure compounds or as standardized plant extracts, provide unlimited opportunities for new drug leads because of the unmatched availability of chemical diversity. The chemical modification of lead compounds, on a trial-and-error basis, typically led to compounds with improved potency, selectivity and bioavailability and reduced toxicity. However, this approach is labor and time-intensive and researchers in the pharmaceutical industry are constantly developing methods with a view to increasing the efficiency of the drug discovery process (Giersiefen et al., 2003). The 'rational', protein structure-based approach relies on an iterative procedure of the initial determination of the structure of the target protein, followed by the prediction of hypothetical ligands for the target protein from molecular modeling and the subsequent chemical synthesis and biological testing of specific compounds (the structure-based drug design cycle).

The rational approach is severely limited to target proteins that are amenable to structure determination. Although the protein data bank (PDB) (Berman et al., 2000, http://www.rcsb.org/pdb) is growing rapidly (~13 new entries daily), the 3D structure of only 1 to 2% of all known proteins has as yet been experimentally characterized. However, advances in sequence comparison, fold recognition and protein-modeling algorithms have enabled the partial closure of the so-called 'sequence-structure gap' and the extension of experimental protein structure information to homologous proteins. Threading methods which relate protein sequences to a library of known structures have been shown to be able to identify the likely protein fold even in cases where there is no clear sequence homology. As collaborative efforts in systematic structure determination began to develop the number of protein sequences that cannot be assigned to a structural class by homology or threading methods, simply because they belong to a previously unidentified protein folding class, started decreasing. The quality of these homology models, and thus their applicability to, for example, drug discovery predominantly depends on the sequence similarity between the protein of known structure (template) and the protein to be modeled (target). Despite the numerous uncertainties that are associated with homology modeling, recent research has shown that this approach can be used to significant advantage in the identification and validation of drug targets, as well as for the identification and optimization of lead compounds.

DRUG DISCOVERY: LESSONS FROM THE PAST

It may be very useful to record a brief summary of some of the historical approaches to drug design and discovery to learn from whence this art had evolved (Oprea et al., 2001). The use of natural products with therapeutic properties is as ancient as human civilization and, for a long time, mineral, plant and animal products were the main sources of drugs. Many ancient populations made use of the medicinal properties of plant extracts, all as a result of several trial and error searches for remedies of specific ailments. Nature has been and is still the best source for drug and its precursors (Harvey, 2000). Natural products and their derivatives discovered years ago are still considered as useful therapeutics even today. For decades, natural products have been a wellspring of drugs and drug leads. Beyond the discovery of natural product thienamycin and the synthetic lead oxazolidinone in the 1970s, there has been a dearth of new compounds (Singh and Barrett, 2006). According to a survey, 61% of the 877 molecule new chemical entities introduced as drug worldwide during 1981 to 2002 can be traced to or were inspired by natural products (Harvey, 2000). These include natural products (6%), natural product derivatives (27%), synthetic compounds with natural-product derived pharmacophores (5%), and synthetic compounds designed on the basis of knowledge gained from a natural product (that is, a natural product mimic; 23%). In certain therapeutic areas, the productivity is higher: 78% of antibacterial and 74% of anticancer compounds are natural products or have been derived from, or inspired by a natural product. These numbers are not surprising if it is assumed that natural products evolved for self-defense. But the influence of natural products is significant even in therapeutic areas for which they might not seem relevant, such as cholesterol management, diabetes, arthritis and depression. On average, natural products have higher molecular weights; incorporate fewer nitrogen, halogen, or sulfur atoms but more oxygen atoms; and are sterically more complex, with more bridgehead tetrahedral carbon atoms, rings, and chiral centers.

TIME AND COST FACTORS INVOLVED IN DRUG DISCOVERY AND DESIGN

Compound discovery and development is an intense and lengthy process (Cunningham, 2000). For a pharmaceutical industry, the number of years to bring a

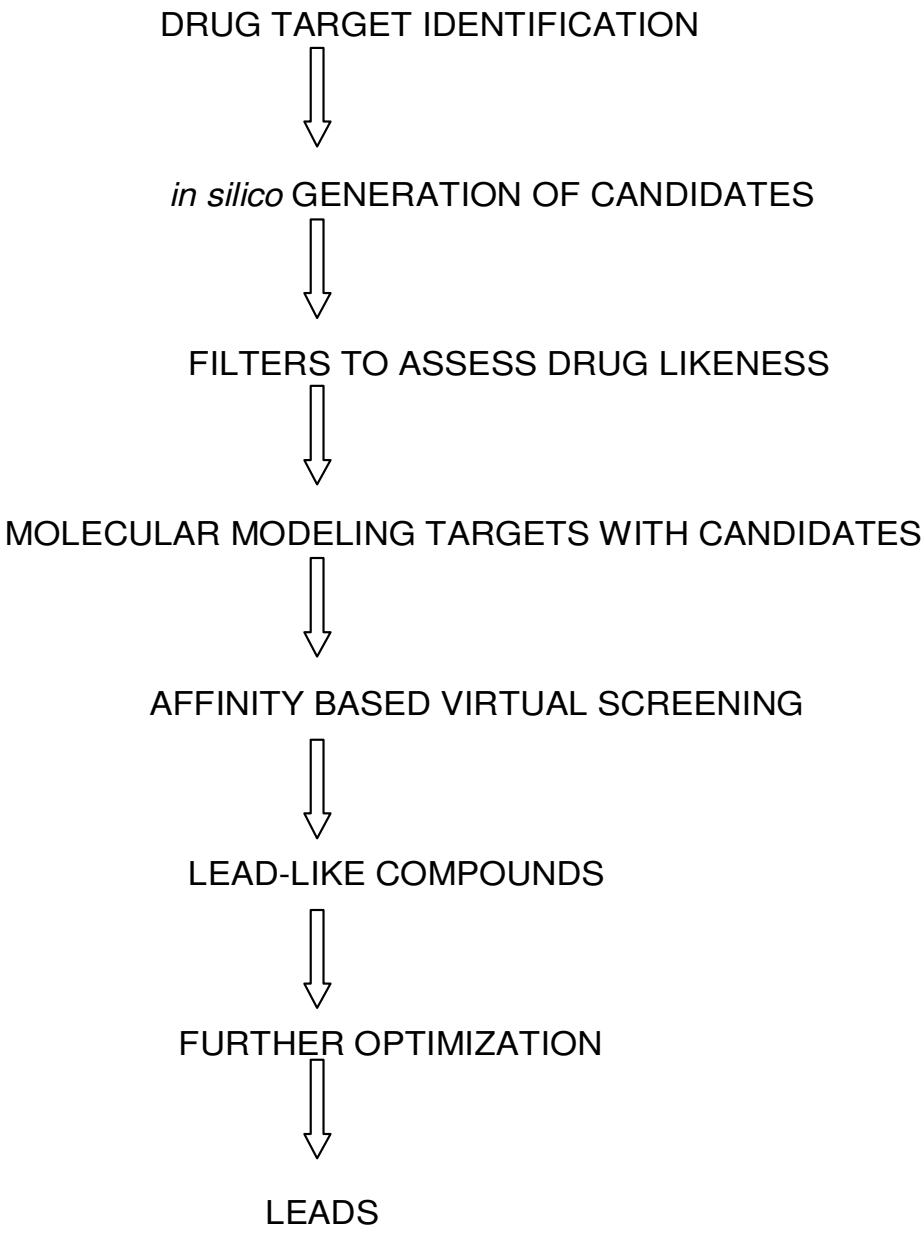

Figure 1. Potential areas for *in silico* intervention in drug discovery.

drug from discovery to market varies and may go upto 15 years costing upto US$900 million per individual drug (Brennan, 2000; DiMasi, 1995). For every 5000 chemicals evaluated as a part of the drug discovery and preclinical trials, only five are allowed to human trials and of these five, only one is approved for the market. A total of 40% of the compounds fail due to poor pharmacokinetics and 11% due to preclinical toxicity. In the past, it was difficult to almost impossible to predict these characteristics for a specific compound. Drug discovery in the new millennium is armed with not only new and efficient techniques for producing and screening new entities, but with computing power that was imaginable a decade ago. It is now possible to design algorithms and empirical screens to predict a *priori* absorption and distribution properties of lead molecules *in silico.* Modern computers give us the opportunity to submit new compounds to a rigorous virtual screening to assess their druggability given the reliable filters. This can potentially save research from pursuing wrong 'leads'. The investment of time and resources that can be directed to more promising new agents will allow the lead-to-market time to shorten considerably in the coming years. Combined with experiment and informatics, computer modeling is expected to accelerate drug discovery (Figure 1) and to find solution to most of the above-cited problems within

the next decade.

PRESENT STATE OF THE ART: COMPUTER-AIDED DRUG DESIGN

Given the vast size of organic chemical space (Kuntz, 1992), drug discovery cannot be reduced to a simple "synthesize and test" drudgery. There is an urgent need to identify and/ or design drug-like molecules (Walters et al., 1998; Lipkowitz et al., 1997) from the vast expanse of what could be synthesized. *In silico* methods have the potential to reduce both time and cost in developing suggestions on drug/ lead-like molecules. Computational tools have the advantage for delivering new lead candidate more quickly and at lower cost. Drug discovery in the 21st century is expected to be different in at least two distinct ways: Development of individualized medicine departing from genomic information and extensive use of *in silico* simulations to facilitate target identification, structure prediction and lead/drug discovery. The expectations from computational methods for reliable and expeditious protocols for developing suggestions on potential leads are continuously on the increase. Several conceptual and methodological concerns remain before an automation of drug design *in silico* could be contemplated.

Computational methods are needed to exploit the structural information to understand specific molecular recognition events and to elucidate the function of the target macromolecule. This information should ultimately lead to the design of small molecule ligands for the target, which will block/activate its normal function and thereby act as improved drugs.

Three-dimensional protein structures are key to a detailed understanding of the molecular basis of protein function. Combining sequence information with 3D structure gives invaluable insights for the development of effective rational strategies for experiments such as site directed mutagenesis, studies of disease related mutations, or the structure based design of specific inhibitors. Techniques for experimental structure solution have made great progress in recent years. However, experimental structure determination is still a time-consuming process without guaranteed success. This is reflected by the fact that the number of structurally characterized proteins is about two orders of magnitude smaller than the number of known protein sequences in the UniProt database (Apweiler et al., 2004), which holds more than one million entries. Thus, no experimental structural information is available for the vast majority of protein sequences. Therefore, theoretical methods for protein structure prediction aiming to bridge this structure knowledge gap have gained much interest in recent years. Moreover, the application of molecular genetics techniques has permitted the manipulation of biosynthetic pathways at a more functional level for the generation of novel chemical species. It also renders uncultivatable microorganisms accessible for secondary metabolite generation (Nisbet and Moore, 1997).

STEPS IN HOMOLOGY/ COMPARATIVE MODELING

Among all current theoretical approaches, comparative modeling is the only method that can reliably generate a 3D model of a protein (target) from its amino acid sequence (Tramontano et al., 2001). Successful model building requires at least one experimentally solved 3D structure (template) that has a significant amino acid sequence similarity to the target sequence. Various structural genomics initiatives were started in the last few years, aiming to speed up the elucidation of new protein structures (Brenner, 2001). Experimental structure elucidation and comparative modeling complement one another in the exploration of the protein structure space. Protein structure similarity clustering is a novel synergistic strategy. It has enabled the identification of biologically relevant starting points in structural space. It has provided guiding structures for the development of focused compound libraries that has yielded biologically prevalidated hits with high fidelity and that too at comparably small library size (Koch and Waldmann, 2005). A key to the efficient coverage will be the careful and optimal selection of the proteins for structural genomics (Vitkup et al., 2001). The growing number of structural templates brings a steadily increasing number of sequences into 'modeling distance' for comparative modeling.

Homology, or comparative, modeling uses experimentally determined protein structures to predict the conformation of another protein that has a similar amino acid sequence. The method relies on the observation that in nature the structural conformation of a protein is more highly conserved than its amino acid sequence and that small or medium changes in sequence typically result in only small changes in the 3D structure (Lesk and Chothia, 1986). Generally, the process of homology modeling involves four steps: Fold assignment, sequence alignment, model building and model refinement The fold assignment process identifies proteins of known 3D structure (template structures) that are related to the polypeptide sequence of unknown structure (the target sequence; this is not to be mistaken with drug target). Next, a sequence database of proteins with known structures (e.g. the PDB-sequence database) is searched with the target sequence using sequence similarity search algorithms or threading techniques (Godzik, 2003). Following identification of a distinct correlation between the target protein and a protein of known 3D structure, the two protein sequences are aligned to identify the optimum correlation between the residues in the template and target sequences. The next stage in the homology modeling process is the model- building phase.

Here, a model of the target protein is constructed from the substitution of amino acids in the 3D structure of the template protein and the insertion and/or deletion of amino acids according to the sequence alignment. Finally, the constructed model is checked with regard to conformational aspects and is corrected or energy minimized using force-field approaches.

For the rational design of new drugs, structural information about the target protein and specifically binding ligands is of utmost importance. Such information can be derived from known ligands with gradually increasing affinity towards the target protein (Moro et al., 1999). Commonly applied docking methods mostly provide binding sites which may necessarily provide pockets that are appropriate in shape and size to reasonably accommodate the ligands known to bind to the modeled protein. Sometimes, the placement of the ligands into the binding site is usually performed manually, often involving some arbitrary assumptions, followed by the relaxation of the complex by molecular dynamics (MD) simulations (Moro et al., 1999). Thus, it can be said that protein modeling and docking are an integrated step and search for new chemical entities are underway towards drug discovery process.

Usually, ligands are placed into binding sites following strategies applied in docking programs (Joseph-McCarthy, 1999). The initial docking is often followed by a minimization using interaction potentials to optimize the mutual protein-to-ligand orientation.

The domain of protein docking desires to determine if there exists some protein or ligand that can dock to a particular protein. Current approaches generally assume the two bodies to be rigid; that is, docking is attempted by exploring translational and rotational configurations of the two bodies. However, a protein often has some degree of flexibility, admitting self-motions of that protein. These different conformations must also be taken into account for docking. Generating such motions, however, is a very difficult problem, due to the high-dimensionality associated with protein structures. More detailed information about ligand features influencing the binding affinity can, at most, be incorporated manually by adjusting conformations of amino acid side-chains or selecting the most consistently explaining representative from a set of protein models.

HOMOLOGY MODELING AND NATURAL PRODUCTS

Ligand-supported homology modeling of protein binding sites of different therapeutic areas such as infectious diseases is gaining a major concern (Hazai et al., 2006). These complex pathologies clearly represent a worthwhile pursuit for the research to be undertaken in this direction.

The emergence of multiple drug resistance to human pathogens has necessitated the search for newer molecules or compounds from other antimicrobial substances or from other sources such as plants (secondary metabolites). The race is on to find new low-cost tools that can be used not only to step up the prevention of diseases, but also to improve and accelerate their diagnosis and treatment as well. Traditionally, the use of medicinal plants for the treatment of human diseases is well known and is practiced in Ayurveda since ancient times. The medicinal plants have provided leads for antiparasitic, antifungal, antiviral, anticancer and antibacterial diseases including flavonoids, coumarins, napthoquinones, terpenoids, alkaloids, steroids etc (Kayser, 2000). Recognition of the biological properties of myriad natural products has fueled the current focus for the search for new drugs. Newer molecular structures as isolated from natural products may be suitably modified to obtain designer molecules for drug designing. Pharmacological testing, modifying, derivatising and research on these natural products represent a major strategy for discovering and developing new drugs (Kayser et al., 2000). The combinatorial chemistry has helped in the development of a series of similar but homologous structural compounds for testing.

However, drug development is very laborious, time consuming, costly, energy intensive and lengthy job. Though, nature continues to baffle scientists when they discover newer molecules having widely different chemical structures, still the synthetic organic chemists and pharmacologists modify these into a multiple of structures. It would require a bioinformatics approach to handle such a gigantic operation of designing and testing so many molecules for the drug development. Biosynergistic use of these compounds formulations may lead to fight different diseases through a series of metabolic biochemical reactions including bioenergetics. It is time when it should be thought how diseases should be treated more rationally and flavonoids and its derivatives form an important choice for the same. Flavonoids inhibit or kill many bacterial strains, inhibit important viral enzymes, such as reverse transcriptase and protease, destroy some pathogenic protozoans (Havsteen, 2002). It also displays anticarcinogeic effects (Hirano et al., 1994). Thus, it can be stated that the pace of natural product research and the level of global interest in the particular area of our environment has risen dramatically in the past few years. This period is projected to continue for the future as the interface between biology and chemistry becomes even more blurred and the public demand rises for the cost effective medications and biological agents from sustainable resources.

The research approach should focus on how to discover novel plant-derived natural products through molecular docking as new lead compounds for potential agents, and to modify these compounds to find still more potent agents with focus being on the application of homology modeling. Another dimension of research used is virtual parallel screening which is a multitarget

computational tool resulting in an *in silico* profile for each compound screened. Based on this, a predicted bioactivity profile (Rollinger, 2009) can be extrapolated to prioritize targets for experimental studies. However, experimental studies using the compounds in laboratory against these diseases would be laborious, costly and time consuming. Therefore, presently the bioinformatics approach should be selected to cut short the cost, labor and time involved in discovering herbal based drugs for various diseases. Hence, it can be said that the structure-based design of a new compound is almost never a *de novo* process, but more often a modification of existing leads, either naturally occurring (ATP, peptide substrates, natural products) or found by conventional biochemical screening. There are number of examples to showcause the aforementioned theory, one of them being staurosporine and quercetin as ATP competitive inhibitors for IKKß inhibition (Avila et al., 2009). Another is the use of crocacin which can inhibit mitochondrial respiration and has shown activity on several plant pathogens (Crowley et al., 2008). Thus, it can be concluded that homology modeling applied for genome-wide prediction of drug target protein structures represent a real chance for the growth and development of pharmaceutical industry.

REFERENCES

Apweiler R, Bairoch A Wu CH (2004). Protein sequence databases Curr. Opin. Chem. Biol., 8: 76- 80.

Avila CM, Romeiro NC, Sant'Anna CMR, Barriero EJ, Fraga CAM (2009). Structural insights into IKKß inhibition by natural products staurosporine and quercetin. Bioorganic Med. Chem. Lett. 19: 6907-6910.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PN (2000). The Protein Data Bank. Nuc Acids Res., 28: 235-242.

Brenner SE (2001). A tour of structural genomics. Nature Rev. Genet., 2: 801–809.

Brennan MB (2000). Drug discovery: Filtering out failures early in the game. Chem. Engg. News, 78: 63-73.

Crowley PJ, Berry EA, Cromartie T, Daldal F, Godfrey CRA, Lee DW, Phillips JE, Taylor A, Viner R. (2008). The role of molecular modeling in the design of analogues of the fungicidal natural products rocacins A and D. Biorganic Med. Chem., 16: 10345-10355

Cunningham MJ (2000). Genomics and proteomics.The new millennium of drug discovery and development. J. Pharma. Toxicol. Methods, 44: 291-300.

DiMasi JA (1995). Trends in Drug Development. Costs, Time and Risks. Drug Inf. J., 29: 375-384.

Giersiefen H, Hilgenfeld R, Hillish A (2003). Modern Methods of Drug Discovery: An Introduction. In Modern Methods of Drug Discovery (Hillisch, A. and Hilgenfeld, R., eds), pp. 1-18

Godzik A (2003). Fold Recognition Methods. In Structural Bioinformatics (Bourne, P. and Weissig, H., eds), pp. 525–546

Harvey A (2000). Strategies for discovering drugs from previously unexplored natural products. Drug Disc Today. 5: 294-300.

Hazai E, Bikadi Z, Zsila, F, Lockwood SF (2006). Molecular modeling of the noncovalent binding of the dietary tomato carotenoids lycopene and lycophyll, and selected oxidative metabolites with 5-lipoxygenase. Bioorg. Med. Chem., 14: 6859-6867.

Havsteen BH (2002). The biochemistry and medical significance of flavonoids. Pharmacol Ther., 96: 67-202.

Hirano T, Gotoh M, Oka K (1994). Natural flavonoids and lignans are potent cytostatic agents against human leukemic HL-60 cells. Life Sci., 55: 1061–1069.

Joseph-McCarthy D (1999). Computational approaches to structure-based ligand design. Pharmacol Ther., 84: 179-191.

Kayser O, Kidderlen AF, Croft SL (2000). Natural products as potential anti-parasitic drugs. Acta Tropica, 77: 307-314.

Koch MA, Waldmann H (2005). Protein structure similarity clustering and natural product structure as guiding principle in drug discovery. Drug Discovery Today, 10: 471-483.

Kuntz ID (1992). Structure-based strategies for drug design and discovery. Science, 257: 1078-1082.

Lipkowitz KB, Boyd DB (1997). Rev. Comput. Chem., 11: 1-60.

Lesk AM, Chothia C (1986). The response of protein structures to amino-acid sequence changes. Philos. Trans. R. Soc. Lond. B Biol. Sci., 317: 345-356.

Moro S, Hoffmann C, Jacobson KA (1999). Role of the extracellular loops of G protein-coupled receptors in ligand recognition: a molecular modeling study of the human P2Y1 receptor. Biochemistry, 38: 3498-3507.

Nisbet LJ, Moore M (1997). Will natural products remain an important source of drug research in the future? Curr. Opinion Biotechnol., 8: 708-712.

Oprea TI, Davis AM, Teague SG, Leeson PD (2001). Is there a difference between leads and drugs? A historical perspective. J Chem Inf. Comput Sci., 41: 1308-1315.

Rollinger JM (2009). Assessing target information by virtual parallel screening- The impact on natural product research. Phytochem. Lett., 2: 53-58.

Singh SB, Barrett JF (2006). Empirical antibacterial drug discovery-Foundation in natural products. Biochem. Pharmacol., 71:1006-1015.

Tramontano A, Leplae R, Morea V (2001). Analysis and assessment of comparative modeling predictions in CASP4. Proteins, 45(5): 22–38.

Vitkup D, Melamud E, Moult J, Sander C (2001). Completeness in structural genomics. Nature Struct. Biol., 8: 559–566.

Walters PW, Stahl MT, Murcko MA (1998). Virtual Screening- an overview. Drug Dis. Today, 3: 160-178.