Gene, 137 (1993) 127–132

GENE 07426

Computer-assisted drug discovery — a review

(Protein structure; computer modeling; AIDS; proteases; epitopes; ligand docking; malaria; mefloquine resistance; schistosomiasis)

Eugene Sun[a] and Fred E. Cohen[a,b,c]

Departments of [a]Medicine, [b]Biochemistry and Biophysics, and [c]Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA 94143-0446, USA

Received by J.W. Larrick/R.A. Houghten: 1 February 1993; Accepted: 15 March 1993; Received at publishers: 20 July 1993

SUMMARY

A structure-based approach to new drug development is an attractive alternative to the traditional paradigm of drug discovery through screening. The elements of this approach are reviewed, with emphasis on the use of homology-built model structures. Two examples, proteases from the organisms that cause malaria and schistosomiasis, illustrate both the model-building process and the feasibility of using such models to computationally screen and identify compounds that inhibit their targets in the low micromolar range.

INTRODUCTION

The tools of molecular biology have facilitated the more rapid investigation of normal and diseased states. With this improved understanding of physiology and pathophysiology, our desire to intervene pharmacologically in disease processes has grown. One method of rapidly translating biological insight into pharmaceutical development exploits the field of structural biology and the tools of structure based drug design. The need for these methods and examples of their application to common infectious diseases are presented. The role of computational chemistry in this process is highlighted.

NEW DRUG DEVELOPMENT

(a) The need for new drug development

The need for ongoing development of new drugs needs no emphasis in light of the current global situation of health and disease. Nowhere is this more evident than in the arena of infectious diseases. The AIDS virus continues its unchecked spread in many parts of the world, with an estimated global prevalence approaching ten million (Mann, 1991). The lack of effective antiviral therapies is unfortunately not limited to the AIDS virus. Age-old scourges such as tuberculosis, malaria and other parasitic diseases continue to exact annual tolls of millions of lives, and in some areas appear to have spread further despite the existence of effective chemotherapy (Bloom and Murray, 1992). While political and economic considerations weigh heavily in the battle against these diseases, the continuous emergence of antibiotic resistance bespeaks the biological problems in eradicating them, and mandates a continual and vigorous effort to develop new agents. Even among bacterial infections routinely regarded as easily treatable, the use and abuse of antibiotics has bred multiply resistant organisms, forcing the use of more potent drugs, thus fueling the cycle of the development of resistance (Neu, 1992). An optimistic U.S.

Correspondence to: Dr. F.E. Cohen, Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA 94143-0446, USA. Tel. (1-415) 476-8519; Fax (1-415) 476-6515; e-mail: cohen@cgl.ucsf.edu

Abbreviations: AIDS, acquired immune deficiency syndrome; CPU, central processor unit; DOCK, see section e; IC_{50}, 50% inhibitory concentration; K_i, inhibition constant; MIDAS, molecular interactive display and simulation; NMR, nuclear magnetic resonance; *P.*, *Plasmodium*.

Surgeon General proclaimed in 1969 that it was time to 'close the book on infectious diseases' (Bloom, 1992). A quarter century later, this statement remains absurdly premature.

(b) Traditional methods of drug discovery

Traditionally, the process of drug development has revolved around a screening approach. If possible, a disease is represented by a bioassay that is easily automated; more cumbersome live cell assays, and even animal models, are often required. Both synthetic compounds and natural products such as soil or microbial broths serve as screening material. As the process is essentially a large-scale adventure in serendipity, an effort to start with more diverse biologic material has led pharmaceutical firms to forage in ever more exotic domains of fauna and flora. Efforts to purify, characterize, and synthesize active compounds often prove difficult. Even after an efficacious compound has emerged, its mechanism of action is often unknown, though it may eventually provide insights into disease physiology. The entire process is laborious, inefficient, expensive, and conceptually inelegant, yet the undeniable fact remains that it has resulted in the prototype compound for most drugs in use today (Ganellin, 1992).

(c) Structure-based drug development

The shortcomings of traditional drug discovery, as well as the allure of a more deterministic approach to combating disease has led to the concept of 'rational drug design' (Kuntz, 1992). The terms 'mechanistic' or 'structure-based' are preferable to 'rational'; the latter implies that the historically successful screening approach is irrational. In its simplest formulation, the principle of structural complementarity is exploited to yield target-specific antagonists. A necessary first step is the identification of a molecular target critical to a disease process or an infectious pathogen. Optimally, it is unique to the pathologic process; in this regard, infectious diseases are particularly suited for the exploitation of divergent or completely novel metabolic pathways (Cohen, 1977). Implication of the target molecule in pathogenesis or virulence is an a priori condition, since the ensuing exercise is nontrivial in both time and cost. An adaptation of Koch's postulates should be applied, namely, that specific inhibition of the target, or mutational inactivation of its gene, leads to reversion to normal physiology or avirulent phenotype.

The next step is the determination of the molecular structure of the target. The validity of structure-based drug discovery rests largely on a high-resolution target structure of sufficient molecular detail to allow selectivity in the screening of compounds. Frequently, this is the main bottleneck of the entire process. Although technologic advances in molecular biology have resulted in thousands of gene sequences encoding proteins, the structures of fewer than 1% of them have been solved by X-ray crystallography or multidimensional NMR spectroscopy. Compared with gene cloning, the techniques of heterologous gene expression, protein purification and crystallography are largely empirical, and often lengthy processes. De novo prediction of protein structure based on primary sequence remains insufficiently accurate, and awaits conceptual advances in the protein folding problem (McGregor and Cohen, 1991). Is this approach to drug discovery then limited to the few hundred proteins for which structures are known?

(d) Homology-built models: imitation crystals

Homology-based modeling is an approach that circumvents the absence of empirical structural data. If a protein sequence is sufficiently homologous to another protein of known structure, a model can be constructed by making reasonable assumptions regarding structural conservation. The accuracy of the model correlates directly with the degree of aligned pairwise sequence identity (Chothia and Lesk, 1986). Examples of useful model structures include antibodies (Bruccoleri et al., 1988), serine proteases (Greer, 1990), and aspartyl proteases (Blundell et al., 1983). Also in favor of this approach is the relatively high degree of conserved structure in the immediate vicinity of the active sites of homologous enzymes. Obviously, this part of the enzyme surface is often the target for drug discovery.

Homology modeling begins by aligning the aa sequence of the target enzyme to those of the homologous proteins of known structure, with an emphasis on aligning the structurally conserved regions. The structurally conserved regions of the model protein are assumed to adopt a structure that is identical to its homologs. Using interactive molecular graphics, the model structure is 'fleshed out' by replacing sidechains, and incorporating loops that bridge gaps in the core structure. Sidechain orientations are matched to homologous structures where possible, or selected from rotamer libraries that are constructed from the observed distributions of sidechain dihedral angles found in proteins (Ponder and Richards, 1987; Sutcliffe et al., 1987). Loop conformations are more problematic, though 'canonical' structures and prediction algorithms can be used with some success (Jones and Thirup, 1986; Chothia and Lesk, 1987).

It is necessary to translate structurally plausible models into energetically reasonable ones. A number of computer programs based on molecular mechanics have been developed for this purpose, such as AMBER (Assisted Model Building with Energy Refinement) (Weiner and Kollman, 1981) and CHARMM (Chemistry at Harvard

Macromolecular Mechanics) (Brooks et al., 1983). These programs optimize macromolecular structures by searching for potential energy minima defined by a 'force field' that incorporates covalent geometry, electrostatic interactions, van der Waals forces and hydrogen bonding. A potential pitfall to this approach is that it is likely to identify a local rather than the global energy minimum (McGregor and Cohen, 1991). This was made abundantly clear by the inability of energy calculations to distinguish a correct protein structure from its misfolded counterpart (Novotny et al., 1988). Although molecular dynamics can be used to increase the radius of convergence of energy calculations, the success of homologous modeling builds from the assumption that related sequences will adopt very similar structures.

(e) Looking for a match: DOCK

With an experimental or hypothetical structure in hand, the challenge is to find a complementary structure that will translate into a good ligand or antagonist. Quantitation of bimolecular interaction energetics is complicated by the introduction of further variables such as binding mode, conformation, and solvent effects. However, since the ultimate objective is the generation of novel structures as potential lead compounds, a rapid screening algorithm with high throughput is preferable to a thermodynamically stringent but time-consuming one. The program DOCK exploits a geometric description of the surface of the target molecule (Connolly, 1983) to define plausible binding pockets. Receptor-ligand interactions are scored on their shape complementarity; alternatively, force field-based scoring that incorporates electrostatic effects can be performed (Kuntz et al., 1982; Meng et al., 1992). With such an approach, it is obvious that the wider the search, the more fruitful it is likely to be. In this respect, structure-based drug discovery resembles the traditional screening approach. The crucial difference is that the screening actually takes place on computer screens and in microprocessors, rather than in bioassays and animal systems.

A number of chemical compound databases are available for screening. These include the Cambridge Structural Database that contains 100000 crystallographically determined structures (Allen et al., 1979) and the Fine Chemicals Directory of more than 55000 commercially available compounds (Molecular Designs, San Leandro, CA, USA). The computational generation of three-dimensional structures from small molecule chemical structures adds significantly to the pool of possible ligands. The programs CONCORD (Rusinko et al., 1988) and COBRA (Leach et al., 1990) use a set of rules to generate energetically favorable 3-D conformations of existing compounds. Thus, vast repositories of both public domain and proprietary compounds can be screened in this way.

A final step is the visual inspection of proposed receptor-ligand complexes. The experienced medicinal chemist possesses intuitive skills that have been difficult to translate into computer algorithms. Many proposed interactions, though computationally sensible, may be chemically or medicinally unrealistic. In addition, related structures may be suggested by the creative user that can be built and compared, in a qualitative manner, to the proposed ligand in the context of the receptor. The process of homology model-building and subsequent structure-based search for lead compounds is schematized in Fig. 1, and illustrated by the examples that follow (Ring et al., 1993).

(f) Parasite proteases as potential drug targets

Proteases are a diverse group of enzymes involved in many physiologic and pathologic processes. They are particularly amenable to homology modeling as they are well represented in the Brookhaven Protein Data Bank of crystallographic structures (Bernstein et al., 1977). This

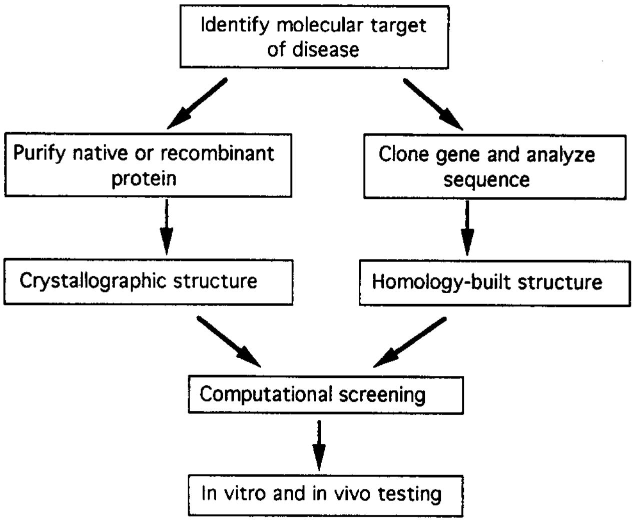
Identify molecular target of disease
Purify native or recombinant protein
Clone gene and analyze sequence
Crystallographic structure
Homology-built structure
Computational screening
In vitro and in vivo testing

Fig. 1. General scheme for structure-based drug discovery. The identification of an appropriate molecular target is fundamental to the ultimate success of this approach. Once identified, a detailed structure is obtained; traditionally, this has been accomplished by crystallography of purified native or recombinant protein. Alternatively, primary sequence may be used to construct a model based on structural conservation to homologous proteins. Databases of chemical structures are computationally screened in the context of the target structure, and potentially active compounds subjected to in vitro and in vivo testing. After identifying a series of potential lead compounds that are active at concentrations less than 10 μM, the analogs are developed following the principles of medicinal chemistry and assayed for activity. From this information, a structure-activity relationship is developed that helps to direct the synthesis of additional analogs until compounds with low nanomolar potency are identified. Ultimately, toxicologic and pharmacologic characterization of the most potent compounds must be completed before clinical testing is initiated.

is especially true for the serine proteases (trypsin, chymotrypsin, elastase) and the cysteine proteases (papain, actinidin, cathepsin B). A serine protease secreted by larvae of the helminthic parasite *Schistosoma mansoni* (McKerrow et al., 1985), and a vacuolar cysteine protease of the malarial organism *Plasmodium falciparum* (Rosenthal et al., 1988) were modeled after their respective homologous structures.

The schistosome protease model was built using seven serine protease structures as reference, and the malaria protease was modeled on two cysteine protease structures, using the MIDAS molecular modeling package (Ferrin et al., 1988). Approximately 60% of the amino acids in each protease could be aligned to structurally conserved regions, which comprised the molecular core, and included the active site and substrate binding domains. Sidechains were oriented and loops built as described above to complete the models. The refinement programs QPACK (Gregoret and Cohen, 1990) and AMBER (Weiner and Kollman, 1981) were applied to optimize atomic packing and potential energy, respectively.

The model structures of the malaria and schistosome proteases were used by the DOCK program to screen the Fine Chemicals Directory. For each enzyme, 4400 high scoring compounds were selected from the database of over 55 000, and visually inspected in the context of their active site and substrate binding clefts. Fifty-two candidate inhibitors for the schistosome protease and thirty-one for the malaria protease were selected, and tested for their ability to inhibit the respective enzymes in vitro. Of those selected, two compounds inhibited the schistosomal protease with K_i values of 3 and 5 μM, and one compound inhibited the malaria protease with an IC_{50} of 6 μM.

These compounds were then tested in more biological contexts. The schistosomal enzyme is implicated in facilitating skin invasion by cercarial larvae (McKerrow et al., 1985). In a direct skin invasion assay using live *S. mansoni* cercariae (Cohen et al., 1991), pretreatment of skin with either of the two active compounds significantly inhibited (90% and 50%) larval penetration compared with controls. The malaria protease is involved in intracellular hemoglobin degradation, serving a digestive role for the parasite (Rosenthal et al., 1988). Trophozoites of *P. falciparum* were grown in the presence of the inhibitor and tritiated hypoxanthine as a marker of metabolic activity. Hypoxanthine uptake was 50% inhibited by the inhibitor at 7 μM, comparable to the IC_{50} found for the protease. These results indicate that homology-built models are reasonable templates for structure-oriented searches of chemical databases, and can generate compounds of sufficient potency to serve as leads in new drug development.

(g) New versus old approaches

What are the costs of this exercise in terms of material, time, manpower, and computational expense? The building of the model is the most labor-intensive and time-consuming segment; it is least subject to automation, and involves judgments requiring a fair amount of expertise in protein structure. By the same token, the finite amount of information available to guide the modeler limits the effort to a period of one or two weeks, and can be performed on workstations such as the Silicon Graphics IRIS. Because the process is heavily interactive, the friendliness of the graphical interface is an important factor. Refinement and optimization programs vary widely in their computational needs, depending on the algorithm applied (Cohen et al., 1990). Those using molecular mechanics generally require minutes to hours on a workstation or minicomputer. A DOCK search of a 50 000-compound database can consume several days of workstation CPU time. The final visual screening is again individualized, and depends on the cutoff criteria applied, as well as the experience of the modeler. In all, the identification of a 'short list' of potential lead compounds using a homology-built model consumes less than a month's time with relatively modest computational cost by today's standards.

Trivial though this may seem compared to the laborious process of actually screening thousands of compounds in vitro, several caveats apply: (*1*) a considerable amount of scientific effort must be devoted to the identification of suitable targets for this approach; (*2*) the target molecule can only be modeled if homologous structures are available; otherwise, the arduous route of crystallography must be followed; (*3*) the in compuo identification of lead compounds must be validated in vitro; and (*4*) the bioactivity of a compound is but one necessary component of an effective pharmaceutical. The road to successful drugs is littered with highly active compounds which failed to satisfy the long list of criteria that includes bioavailability, metabolic stability, synthetic tractability, toxicity, and last but not least, profitability.

In the search for antimalarial compounds, the U.S. Army screened more than 250 000 compounds beginning in the 1960s using a murine model of disease (Strube, 1975). The cost of the entire project is not known, but almost certainly was an enormous sum. The result was the introduction of a single effective antimalarial, mefloquine, in 1983. Although it is difficult to generalize, this kind of return on investment in the drug discovery process is not atypical. Mefloquine, undeniably, has had a significant impact on the treatment and prevention of malaria. In affluent nations, it is now the recommended prophylaxis for travel to most malarious areas (Centers for Disease Control, 1991). However, in only a disturbingly few years after its introduction, mefloquine resistance has been reported in Southeast Asia (Nosten et al., 1991). It is important to recognize that this is not a failure of the drug per se, but serves as another reminder that the book on infectious diseases is most certainly not closed.

It is hoped that structure-based methodologies can significantly shorten the drug development cycle by accelerating the pace of compound discovery. Although the acquisition of useful structural data remains a major stumbling block, it may be sidestepped using homology-based models. The ability to accurately simulate molecular interactions on a manageable computational scale is now a reality, and offers the potential of exploiting biomolecular diversity in calculated ways to interrupt pathologic processes. Ultimately, perhaps the most rational aspect of structure-based approaches to the discovery and development of new drugs is the recognition for their continuing need.

(h) Conclusions

(*1*) Structure-based drug design is an approach to developing therapeutic agents that exploit molecular structures specific to pathogens and pathophysiologic processes.

(*2*) High resolution crystallographic protein structures are ideal targets for the selection of highly specific antagonists, but are not always easy to obtain.

(*3*) Homology-built models of proteins can be used as plausible templates in the absence of structural data, if there is sufficient homology to a of protein of known structure.

(*4*) Current technology permits the computational screening of sizable molecular databases in the context of ligand-receptor interactions, at relatively modest expenditures of both time and computational power.

(*5*) Examples of this approach have been given with proteases from two medically important pathogens, demonstrating the utility of homology-built models as targets in computational screening to identify potential lead compounds for drug development.

ACKNOWLEDGEMENTS

This work was supported by grants from the Defense Advanced Research Projects Agency (MDA-972-91-J1013 to F.E.C.) and the National Institutes of Health (GM39900 to F.E.C., and AI01095 to E.S.). We thank our colleagues James H. McKerrow, Philip J. Rosenthal, and Christine S. Ring for helpful discussions.

REFERENCES

Allen, F.H., Bellard, S., Brice, M.D., Cartwright, B.A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B.G., Kennard, O., Motherwell, W.D.S., Rodgers, J.R. and Watson, D.G.: The Cambridge crystallographic data centre: computer-based search, retrieval, analysis and display of information. Acta Crystallogr. Sect. B 35 (1979) 2331–2339.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M.: The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. 112 (1977) 535–542.

Bloom, B.R.: Tuberculosis, back to a frightening future. Nature 358 (1992) 538–539.

Bloom, B.R. and Murray, C.J.L.: Tuberculosis: commentary on a reemergent killer. Science 257 (1992) 1055–1064.

Blundell, T., Sibanda, B.L. and Pearl, L.: Three-dimensional structure, specificity, and catalytic mechanism of renin. Nature 304 (1983) 273–275.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M.: CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. 4 (1983) 187–217.

Bruccoleri, R.E., Haber, E. and Novotny, J.: Structure of antibody hypervariable loops reproduced by a conformational search algorithm. Nature 335 (1988) 564–568.

Centers for Disease Control: Change of dosing regimen for malaria prophylaxis with mefloquine. Morbid Mortal Weekly Rep. 40 (1991) 72–73.

Chothia, C. and Lesk, A.M.: The relation between the divergence of sequence and structure in proteins. EMBO J. 5 (1986) 823–826.

Chothia, C. and Lesk, A.M.: Canonical structures for the hypervariable regions of immunoglobulins. J. Mol. Biol. 196 (1987) 901–917.

Cohen, F.E., Gregoret, L.M., Amiri, P., Aldape, K., Railey, J.F. and McKerrow, J.H.: Arresting tissue invasion of a parasite by protease inhibitors chosen with the aid of computer modeling. Biochemistry 30 (1991) 11221–11229.

Cohen, N.C., Blaney, J.M., Humblet, C., Gund, P. and Barry, D.C.: Molecular modeling software and models for medicinal chemistry. J. Med. Chem. 33 (1990) 883–894.

Cohen, S.S.: A strategy for the chemotherapy of infectious disease. Science 197 (1977) 431–432.

Connolly, M.L.: Solvent-accessible surfaces of proteins and nucleic acids. Science 221 (1983) 709–713.

Ferrin, T.E., Huang, C., Jarvis, L. and Langridge, R.: The MIDAS display system. J. Mol. Graphics 6 (1988) 13–37.

Ganellin, C.R.: Past approaches to discovering new drugs. In: Wermuth, C.G. (Ed.), Medicinal Chemistry for the 21st Century. Blackwell, Oxford, 1992, pp. 3–12.

Greer, J.: Comparative modeling methods: application to the family of the mammalian serine proteases. Proteins 7 (1990) 317–334.

Gregoret, L.M. and Cohen, F.E.: Novel method for the rapid evaluation of packing in protein structures. J. Mol. Biol. 211 (1990) 959–974.

Jones, T.A. and Thirup, S.: Using known substructures in protein model building and crystallography. EMBO J. 5 (1986) 819–822.

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

Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E.: A geometric approach to macromolecule-ligand interactions. J. Mol. Biol. 161 (1982) 269–288.

Leach, A.R., Dolata, D.P. and Prout, K.: Automated conformational analysis and structure generation: algorithms for molecular perception. J. Chem. Inf. Comput. Sci. 30 (1990) 316–324.

McGregor, M.J. and Cohen, F.E.: Analysis of conformational tendencies in proteins. Curr. Opin. Struct. Biol. 1 (1991) 345–350.

McKerrow, J.H., Pino-Heiss, S., Lindquist, R. and Werb, Z.: Purification and characterization of an elastinolytic proteinase secreted by cercariae of *Schistosoma mansoni*. J. Biol. Chem. 260 (1985) 3703–3707.

Mann, J.M.: Global AIDS: critical issues for prevention in the 1990s. Int. J. Health Serv. 21 (1991) 553–559.

Meng, E.C., Shoichet, B.K. and Kuntz, I.D.: Automated docking with grid-based energy evaluation. J. Comput. Chem. 13 (1992) 505–524.

Neu, H.C.: The crisis in antibiotic resistance. Science 257 (1992) 1064–1073.

Nosten, F., ter Kuile, F., Chongsuphajaisiddhi, T., Luxemburger, C., Webster, H.K., Edstein, M., Phaipun, L., Thew, K.L. and White, N.J.: Mefloquine-resistant *falciparum* malaria on the Thai-Burmese border. Lancet 337 (1991) 1140–1143.

Novotny, J., Rashin, A.A. and Bruccoleri, R.E.: Criteria that discriminate between native proteins and incorrectly folded models. Proteins 4 (1988) 19–30.

Ponder, J.W. and Richards, F.M.: Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193 (1987) 775–791.

Ring, C.S., Sun, E., McKerrow, J.H., Lee, G.K., Rosenthal, P.J., Kuntz, I.D. and Cohen, F.E.: Structure-based inhibitor design using homology built structures. Proc. Natl. Acad. Sci. USA (1993) in press.

Rosenthal, P.J., McKerrow, J.H., Aikawa, M., Nagasawa, H. and Leech, J.: A malarial cysteine proteinase is necessary for hemoglobin degradation by *Plasmodium falciparum*. J. Clin. Invest. 82 (1988) 1560–1566.

Rusinko III, A., Skell, J.M., Balducci, R., McGarity, C.M. and Pearlman, R.S.: CONCORD, a Program for the Rapid Generation of High Quality Approximate 3-Dimensional Molecular Structures. Tripos Associates, St. Louis, MO, 1988.

Strube, R.S.: The search for new antimalarial drugs. J. Trop. Med. Hyg. 78 (1975) 171–185.

Sutcliffe, M.J., Hayes, F.R.F. and Blundell, T.L.: Knowledge based modeling of homologous proteins, part II: rules for the conformations of substituted sidechains. Protein Eng. 1 (1987) 385–392.

Weiner, P.K. and Kollman, P.A.: AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. J. Comput. Chem. 2 (1981) 287–303.