Ligand–Protein Inverse Docking and Its Potential Use in the Computer Search of Protein Targets of a Small Molecule

Y.Z. Chen* **and D.G. Zhi**
Department of Computational Science, National University of Singapore, Singapore

ABSTRACT **Ligand–protein docking has been developed and used in facilitating new drug discoveries. In this approach, docking single or multiple small molecules to a receptor site is attempted to find putative ligands. A number of studies have shown that docking algorithms are capable of finding ligands and binding conformations at a receptor site close to experimentally determined structures. These algorithms are expected to be equally applicable to the identification of multiple proteins to which a small molecule can bind or weakly bind. We introduce a ligand–protein inverse-docking approach for finding potential protein targets of a small molecule by the computer-automated docking search of a protein cavity database. This database is developed from protein structures in the Protein Data Bank (PDB). Docking is conducted with a procedure involving multiple-conformer shape-matching alignment of a molecule to a cavity followed by molecular-mechanics torsion optimization and energy minimization on both the molecule and the protein residues at the binding region. Scoring is conducted by the evaluation of molecular-mechanics energy and, when applicable, by the further analysis of binding competitiveness against other ligands that bind to the same receptor site in at least one PDB entry. Testing results on two therapeutic agents, 4H-tamoxifen and vitamin E, showed that 50% of the computer-identified potential protein targets were implicated or confirmed by experiments. The application of this approach may facilitate the prediction of unknown and secondary therapeutic target proteins and those related to the side effects and toxicity of a drug or drug candidate. Proteins 2001;43:217–226.** **© 2001 Wiley-Liss, Inc.**

Key words: drug target; flexible docking; ligand–protein binding; ligand–protein interaction; molecular recognition; protein cavity; rational drug design; tamoxifen; target of drug; vitamin E

INTRODUCTION

Ligand–protein docking has been developed as a useful tool in facilitating drug design.[1,2] In this approach, docking single or multiple small molecules in single or multiple conformations to a receptor site is attempted to find putative ligands. A number of flexible docking algorithms have been introduced. These include multiple-conformer shape matching,[3,4] genetic algorithm,[5–7] evolutionary programming,[8] simulated annealing,[9] fragment-based docking,[10–14] and other novel algorithms.[15–19] Testing results have shown that these algorithms are capable of finding ligands and binding conformations at a receptor site close to experimentally determined structures.[4–19]

Because of their capability in identifying potential ligands and binding conformations, these algorithms are expected to be equally applicable to an inverse-docking process for finding multiple putative protein targets to which a small molecule can bind or weakly bind. This may be applied to the identification of unknown and secondary therapeutic targets of drugs, drug leads, natural products, and other ligands. Interactions of a drug or a drug candidate with some of the proteins in the human body have implications for unwanted side effects and toxicity.[20] Therefore, ligand–protein inverse docking may find an application in facilitating the prediction of protein targets related to the side effects and toxicity of a drug or drug candidate. Although good biological activity against an intended target is a necessity, a drug candidate needs to pass additional tests and clinical trials for side effects, toxicity, bioavailability, and efficacy.[21] A substantial portion of the $350 million and 12 years spent on average for commercial drugs has been squandered on many drug candidates that failed to ever reach the market.[21,22] Therefore, new tools for low-cost and fast-speed drug testing, particularly in the early stages of development, are needed. Ligand–protein inverse docking may perhaps be developed into such a tool. Rapid progress is being made in structural genomics,[23] functional genomics,[24] proteomics,[25] and pharmacokinetics and drug metabolism.[26] This is expected to provide information about protein structure and function, cellular profiles of proteins, molecular distribution, and metabolism that is needed for accurate predictions of ligand–protein interactions and related physiological effects.

This inverse-docking strategy requires a sufficient number of proteins of known three-dimensional (3D) structures. At present there are 12,282 protein entries in the

*Correspondence to: Y. Z. Chen, Department of Computational Science, National University of Singapore, Blk S17, Level 7, 3 Science Drive 2, Singapore 117543. E-mail: yzchen@cz3.nus.edu.sg

Received 20 September 2000; Accepted 11 December 2000

Protein Data Bank (PDB), and the number increases at a rate of well over 100 per month.[27] About 17% of these have unique sequences.[28] The introduction of high-throughput methods is expected to enable the structural determination of 10,000 proteins with unique sequences within 5 years.[23] Thus, the number of proteins is approaching a meaningful level to cover a diverse set of potential targets.

All small-molecule drugs appear to bind to cavities of proteins and nucleic acids. Thus, to facilitate computer-automated inverse-docking searches for putative protein targets of a small molecule, we have developed a protein cavity database from protein entries in the PDB. This database is composed of models of individual cavities in each protein. The model of a cavity is a cluster of overlapping spheres that fill up that cavity.[3]

An inverse-docking procedure called INVDOCK has been introduced to conduct computer-automated inverse-docking searches of this database to identify potential protein targets of a small molecule. A small molecule is flexibly docked into each cavity by a procedure involving multiple-conformer shape-matching alignment of the molecule to the cavity followed by molecular-mechanics torsion optimization and energy minimization on both the ligand and the binding region of the receptor. A new scoring method is used that performs binding competitive analysis in addition to the evaluation of molecular-mechanics ligand–protein interaction energy. The procedure is tested on a number of ligand–protein complexes from the PDB. The root-mean-square deviation (RMSD) of docked ligands with respect to the corresponding ligand in the original PDB structure is computed to evaluate the quality of our inverse-docking procedure. The procedure is then used to search potential protein targets of two drugs, 4H-tamoxifen and vitamin E, and the results are compared to available experimental data.

METHODS

Protein Cavity Database

We followed Kuntz et al.[3] to model a cavity by a group of overlapping spheres that fill up that cavity. Each cavity entry is derived from the corresponding PDB entry by the following procedure: Ligands and water in a PDB protein structure are first removed. The surface of this protein, as defined by Richards,[29] is then generated. The van der Waals surface of a solvent-accessible atom is generated with the respective parameter from the AMBER force field.[30] The inward-facing surface covering the interface of van der Waals surfaces is computed with a probe sphere 1.4 Å in radius. The whole protein surface is then coated and filled with a cluster of spheres by a method similar to that of Kuntz et al. Each sphere is checked for the extent of its surrounding space covered by protein atoms. The surrounding space is defined as a region within 15 Å of the center of the sphere. A sphere is considered covered if more than 50% of the direction around the sphere is covered by protein atoms. The covered spheres are then divided into separate cavity groups on the basis of the spatial separation of nearest neighboring spheres.

In addition, noncovered spheres close to a covered sphere in a group are included in that group. This is to ensure that a cavity model sufficiently covers the cavity surface region where a tail of a ligand might be located. In some ligand–protein PDB structures, such as HIV-1 protease complexed to its inhibitor, one or both ends of the ligand are found to be located in such a region. Figures 1 and 2 show the cavity models of a protein HIV-1 protease (PDB Id: 1hsg) and an estrogen receptor (PDB Id: 1a52), respectively. The cavity in 1hsg is well represented. In the structure of 1a52, the entrance of the cavity shown in Figure 2 is connected to a groove extending to different parts of the surface of the protein. The computer-

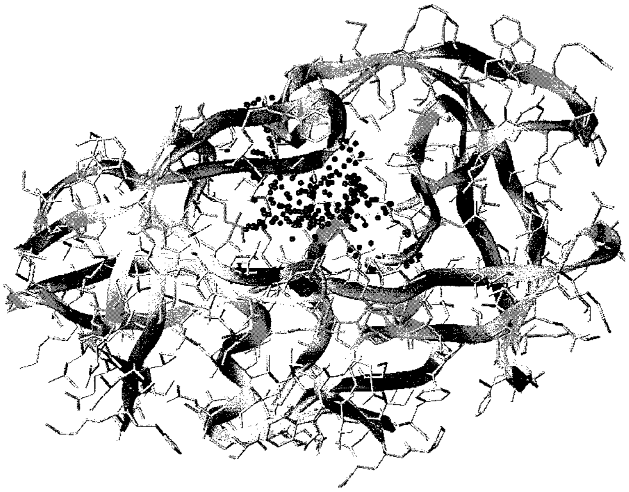

Fig. 1. Computer-generated model of a cavity in HIV-1 protease (PDB Id: 1hsg). The dark balls are the centers of spheres representing the cavity. There are 228 spheres in this model.

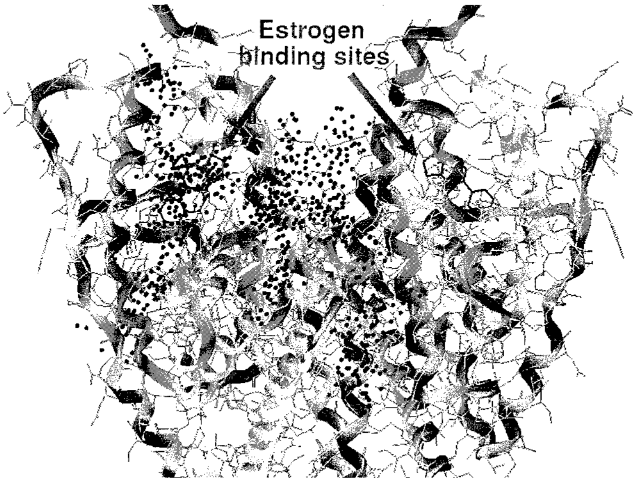

Fig. 2. Computer-generated model of one of the two cavities in the estrogen receptor (PDB Id: 1a52). The dark balls are the centers of spheres representing the cavity and adjacent grooves connected to it. There are 704 spheres in this model. The two arrows point to the location of two estrogen molecules found in the PDB structure. These two estrogen molecules are also shown (dark stick structures) at the respective binding sites in the structure.

generated model includes both the cavity and the groove to which it is connected. Because a groove may also be a binding site, no attempt has been made to remove grooves from a cavity model. A visual inspection of these and a large number of other proteins showed that all the cavities and their adjacent grooves were reasonably represented.

3D structures of proteins from different species have been deposited in the PDB. Because of species-related variation in protein sequences, ligand binding is often specific to a protein of a particular species, or a species class, and those of close phylogenetic distance. In general, the main interest of target identification is to understand the molecular mechanism of the therapeutic and physiological effect of a ligand with respect to one species or a few particular species. Therefore, cavity entries can be divided into classes defined by disease-related species, sequence similarity, and phylogenetic distance to a particular species or species class such as human and mammal. Additional information such as the known ligand-binding region and the extent of each sphere being buried inside a cavity is also included in a cavity entry to facilitate the prioritized docking strategy described next.

Inverse-Docking Procedure

The inverse-docking procedure INVDOCK is designed for the automated search of every entry of the protein cavity database for potential protein targets of a small molecule. Because of the large number of cavity entries, across-the-board specification of possible binding sites within each cavity is difficult. Hence, all parts of each cavity are subject to docking. Docking to sites inside each cavity starts from known ligand-binding sites, followed by more interior sections and then the remaining part. To save central processing unit (CPU) time, the program proceeds to the next protein entry when the first successful dock is obtained without an exhaustive search for the optimum binding mode within each cavity. Although the optimum binding mode is not specifically sorted, the prioritized search strategy seems to dock a ligand to a site reasonably close to the experimentally observed positions in most of the ligand–protein structures studied. All cavity entries are subject to searching unless the related protein has been identified as a potential target.

Flexible ligand docking is similar to the multistep strategy approach proposed by Wang et al.[19] Ligand conformation is sampled at a resolution similar to that of Wang et al. The docking of a particular conformer to a cavity is as follows: First, the ligand is aligned within the selected site by the position of each ligand atom being matched with the center of the spheres. Because of the relatively low-resolution nature of ligand conformation sampling, a certain degree of structural clash is allowed at this stage. A molecular-mechanics conformation optimization is then conducted by a limited torsion space sampling of rotatable bonds in the ligand and those in the side-chain of the receptor amino acid residues at the binding site. Each rotatable bond is sampled at ±15°. This is followed by 50 iterations of Cartesian coordinate energy minimization on all ligand and protein atoms at the binding site to further optimize the ligand–protein complex. Energy minimization is by a steepest decent method.

In both torsion optimization and energy minimization, AMBER force fields[30] are used for covalent-bond, bond-angle, torsion, and nonbonded van der Waals and electrostatic interactions. A large number of crystal structures in PDB contain only non-hydrogen atoms. To save computing time and avoid the difficulty in modeling hydrogens, the Morse potential,[31] which is a function of the donor–acceptor distance, is used to represent hydrogen-bond terms. This potential has been shown to give a reasonable description of hydrogen-bond energy and dynamics in biomolecules.[32,33] The published Morse potential parameters[32,33] are used in this work.

Protein flexibility is known to influence ligand binding, and thus methods such as side-chain conformational sampling and protein ensemble generation have been explored in docking studies.[34] Limited side-chain conformation sampling is considered in INVDOCK along with additional energy minimization for all atoms at the binding site. However, no attempt is made to explicitly sample the ensemble of protein conformations. A substantial number of proteins in PDB have multiple entries submitted by different groups; some of them are associated with different binding ligands or different mutants. Moreover, there are multiple conformations in most NMR structures. To a certain extent, INVDOCK searching of these multiple forms of structures/conformations may serve as a partial sampling of the conformation for some proteins.

Scoring

The scoring of docked structures is based on a ligand–protein interaction energy function, ΔE_{LP}, composed of the same hydrogen-bond and nonbonded terms as those used for structure optimization. An analysis of a large number of PDB ligand–protein complexes shows that the computed ΔE_{LP} is generally less than $\Delta E_{\mathrm{Threshold}} = -\alpha N$ kcal/mol, where N is the number of ligand atoms and α is a constant ($\sim$1.0). The exact value of α can be determined by the fitting of the linear equation $\Delta E_{\mathrm{Threshold}} = -\alpha N$ to the computed ΔE_{LP} for a large set of PDB structures. This statistically derived energy value can be used empirically as a threshold for screening likely binders. A polynomial form of $\Delta E_{\mathrm{Threshold}}$ involving more parameters may also be introduced to derive an energy threshold. ΔE_{LP} can be required to be less than $\Delta E_{\mathrm{Threshold}}$ when successfully docked structures are selected.

Ligand binding is competitive in nature. A drug is less likely to be effective if it binds to its receptor noncompetitively against natural ligands and, to some extent, other drugs that bind to the same receptor site. This binding competitiveness may be partially taken into consideration for cavities known to be ligand-bound in at least one PDB entry. Ligands in PDB structures are known binders. Therefore, PDB ligands bound to the same receptor site as that of a docked molecule may thus be considered competitors of that molecule. In INVDOCK selection of a potential protein target, the computed ΔE_{LP} is not only evaluated against $\Delta E_{\mathrm{Threshold}}$ but also compared to the ligand–

protein interaction energy of the corresponding PDB ligands that bind in the same cavity in this or other relevant PDB entries. The ligand–protein interaction energy for the relevant PDB structures is computed by the same energy functions as that for the docked molecule. In addition to the condition that it be lower than $\Delta E_{\text{threshold}}$, ΔE_{LP} of a docked molecule is required to be lower than a competitor energy threshold $\Delta E_{\text{Competitor}}$ when a potential target is selected. $\Delta E_{\text{Competitor}}$ can be taken as the highest ligand–protein interaction energy of the corresponding PDB ligands multiplied by a factor β. For finding weak binders as well as strong binders, a factor $\beta \leq 1$ is introduced to scale the ligand–protein interaction energy of PDB ligands. This is because a weak binder may have a slightly higher interaction energy than that of a PDB binder. We have not found experimental data to determine the value of β. Here, β is tentatively determined by an analysis of the computed energy for a number of compounds. Our study suggests that a value of 0.8 for β leads to reasonable results statistically.

INVDOCK Search Setup

In this work, the capability of INVDOCK in identifying putative protein targets is tested on two molecules. One is 4H-tamoxifen, which is an active metabolite of the clinical anticancer agent tamoxifen. The other is vitamin E. These two molecules have been the subjects of extensive investigations, and a number of their protein targets have been probed. Therefore, they are suitable for testing the INVDOCK program. The 3D structure of 4H-tamoxifen is from the ACD3D database of MDL (http://www.mdli.com/cgi/dynamic/welcome.html), and that of vitamin E is from Woodcock's molecular model database (http://www.sci.ouc.bc.ca/chem/molecule/molecule.html). These two structures were selected because they were model-built. Computer 3D model building programs typically generate structures that are virtually identical to the corresponding X-ray crystal structure in 38–57% of the cases.[35] Because there are substantially more model-built molecular structures than X-ray crystal structures, the use of these model-built structures can better evaluate the capability of the INVDOCK program for the type of molecular structures more likely to be used in future studies.

Most experimental studies of tamoxifen and vitamin E have been focused on their effect in systems relevant to humans. Therefore, in this work only human and mammalian proteins are searched for the identification of potential protein targets of these two molecules. Observed effects of a drug relevant to target identification are usually associated with an alteration in the protein activity and a change in the protein level due to the binding of the drug to an active site or sites used by other ligands to regulate protein function. These sites are known ligand-binding sites. Hence, we only focus on the search of cavities known as ligand-bound in at least one PDB entry. The number of these cavity entries is 2,700. The program has been tested on both Windows and UNIX platforms. The CPU times for searching these entries are approximately 8 days on an 800-MHz AMD Athlon PC and 20 days on a 250-MHz SGI R10000 Octane workstation.

RESULTS AND DISCUSSION

Testing of the INVDOCK Docking Algorithm

The flexible docking algorithm of INVDOCK is tested on nine ligand–protein complexes from the PDB. Some of these structures have been used in testing different docking programs.[4–19] Our selected structures include several proteins bound either by a clinical drug, a natural product, or a natural ligand. The ligand in each of these structures is first removed. It is then docked to the corresponding ligand-depleted structure with INVDOCK. Table I lists the computed RMSD of each of the docked ligands with respect to the corresponding ligand in the original PDB structure. Figure 3 shows a structural comparison between docked and original ligands in four of these proteins. Six of the ligands are docked into the binding site with an RMSD in the range of 0.80–2.05 Å, which is comparable to those obtained in other docking studies.[4–19]

Although they are docked to the same position as that in the corresponding crystal structure, the RMSD of the other three ligands is substantially larger than that obtained in other studies. Each of these docked molecules largely overlaps the crystal ligand. However, they are either flipped or have one end oriented in a different direction with respect to the crystal structure. As shown in Figure 3, the folate molecule (corresponding PDB Id: 1dhf) is docked such that its tail is oriented away from the crystal position. However, the glycyl-L-tyrosine molecule (PDB Id: 3cpa) is docked in such a way that not only its position is shifted by about 2 Å, but it also flips by 180° along its short axis. The reason for such a deviation in the orientation of docked molecule is that INVDOCK does not attempt an exhaustive search for the optimum binding mode in a cavity. As a result, when a molecule is docked into a location with a similar steric contact as that of a PDB ligand, its ligand–protein interaction energy may well pass the scoring threshold. This problem may be resolved by the introduction of a search for the optimum binding mode and the refinement of the scoring function in the INVDOCK algorithm. Nonetheless, our results seem to suggest that the INVDOCK algorithm is capable of docking a molecule to a site reasonably close to the original binding location.

To test whether docking quality can be improved by an extended INVDOCK algorithm that includes the search for the optimum binding mode, we carried out a separate study for all the complexes tested. As shown in Table I, the RMSD of the three ligands, with a larger RMSD in the original INVDOCK run, is significantly improved from 3.56Å–6.55Å in the original INVDOCK computation to 0.97Å–2.41Å in the new computation. This indicates that docking quality can be improved to a certain extent by the extension of the INVDOCK algorithm to include a search for the optimum binding mode. The same computation on the other six ligands shows no significant change between the RMSD of the optimum binding mode and that from the original INVDOCK computation. However, in Table I the

TABLE I. Testing Results of INVDOCK[†]

Molecule	Docked protein	PDBld	RMSD	E	Description of docking quality	RMSD_{opt}	E_{opt}	E_{cryst}
Indinavir	HIV-1 protease	1hsg	1.38	−70.25	Match	1.23	−84.29	−88.68
Xk263 of Dupont Merck	HIV-1 protease	1hvr	2.05	−70.20	Match	1.48	−96.08	−101.92
Vac	HIV-1 protease	4phv	0.80	−94.51	Match	1.01	−95.09	−113.74
Folate	Dihydrofolate reductase	1dhf	6.55	−48.67	One end match, the other in a different orientation	2.41	−63.02	−76.60
5-Deazafolate	Dihydrofolate reductase	2dhf	1.48	−65.49	Match	1.35	−77.12	−72.79
Estrogen	Estrogen receptor	1a52	1.30	−45.86	Match	1.18	−49.25	−47.12
4-Hydroxytamoxifen	Estrogen receptor	3ert	5.45	−55.15	Complete overlap, flipped along short axis	0.97	−55.31	−53.55
Guanosine-5′-[B,G-methylene] triphosphate	H-Ras P21	121p	0.94	−80.20	Match	0.94	−80.20	−74.47
Glycyl-*L-tyrosine	α-Carboxypeptidase A	3cpa	3.56	−40.63	Overlap, flipped along short axis	2.19	−44.84	−45.48

[†] The computed RMSD and the ligand–protein interaction energy E are given along with a description of a structural comparison between docked and crystal ligands. Also included are the computed RMSD (RMSD_{opt}) and energy (E_{opt}) for the corresponding optimum docked structure and the energy (E_{cryst}) for the corresponding crystal structure. RMSDs are in angstroms, and energies are kilocalories per mole.

ligand–protein interaction energy of all the ligands in the optimum binding mode is substantially decreased with respect to that derived from the original INVDOCK computation. In most cases, this energy is closer to that in the original PDB structure. This seems to indicate that the INVDOCK optimization and scoring scheme is capable of guiding the system to a configuration fairly close to the native state.

Potential Protein Targets of 4H-Tamoxifen

Tamoxifen is an anticancer drug widely used for the treatment of breast cancer,[36] and it has been approved as the first cancer-preventive drug. The tamoxifen metabolite 4H-tamoxifen is believed to be the major contributor to the antiestrogenic effects of tamoxifen inside the human body.[36] Hence, 4H-tamoxifen is investigated in this study. Potential human and mammalian protein targets of 4H-tamoxifen identified by INVDOCK are given in Table II along with the respective clinical implications from experiments. Figure 4 shows the 4H-tamoxifen (dark ball-and-stick structure) docked to one of the identified protein targets, estrogen receptor (PDB Id: 1a52). The crystal structure of estrogen (gray stick structure), together with the INVDOCK docked structure of estrogen (dark line-drawing structure), is also included for comparison. Docked 4H-tamoxifen has a more extensive van der Waals contact and a stronger hydrogen-bond interaction. The computed ligand–protein van der Waals and hydrogen-bond energies for the docked 4H-tamoxifen–receptor complex are −27.63 and −4.20 kcal/mol, respectively. In contrast, the corresponding energies of the crystal estrogen-receptor complex 1a52 are −24.16 and −2.42 kcal/mol, respectively. This enhanced steric and hydrogen-bond interaction might contribute to the binding competitiveness of 4H-tamoxifen against estrogen.[36]

A number of known protein targets of tamoxifen are found in the table. These include estrogen receptor,[36] protein kinase C,[37] collagenase,[38] 17β-hydroxysteroid dehydrogenase,[39] alcohol dehydrogenase,[40] and prostaglandin synthetase.[41] It has been observed that two other INVDOCK identified proteins, glutathione transferase and 3α-hydroxysteroid dehydrogenase, exhibited altered activity by tamoxifen,[42,43] which may be indicative of direct binding of tamoxifen to these proteins. Experiments showed that the level of another two identified proteins, dihydrofolate reductase and immunoglobulin, is changed by tamoxifen.[44,45] Ligand binding is known to self-regulate the protein level in certain cases.[46] It remains to be seen whether these two proteins are also the target of tamoxifen as implicated by the INVDOCK search.

Known targets of tamoxifen, such as calmodulin,[37] are not identified by INVDOCK. One possible reason might be that the available PDB structures of calmodulin are not sufficiently close to tamoxifen-bound conformation. None of these PDB structures is bound by a ligand similar in structure to tamoxifen. The conformation of calmodulin is known to change significantly with the binding of ligands.[47] Because of the intrinsic flexibility of this protein, it is highly likely that ligand binding to this protein involves an induced fit. Our analysis of two PDB structures of calmodulin bound by a different ligand (PDB Id: 1a29 and 2bbm) shows that the conformation of this protein is dependent on its binding ligand. In a recent molecular docking study, a ligand was docked into calmod-

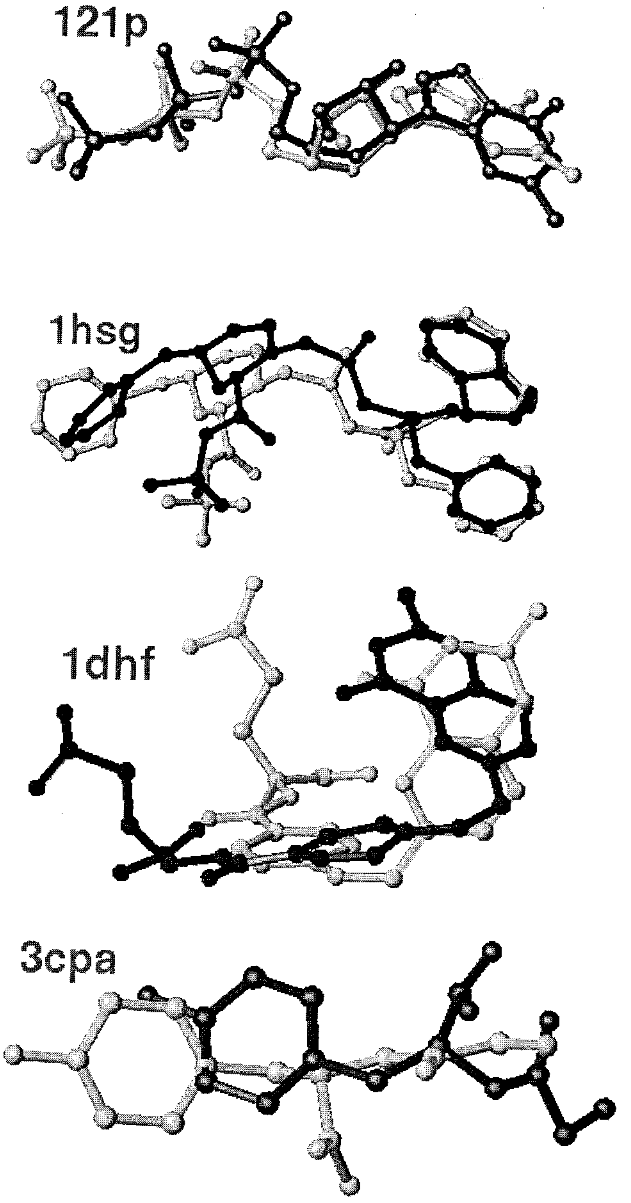

Fig. 3. Docked (dark) and crystal (gray) structures of ligands in some PDB ligand–protein complexes. The PDB Id of each structure is shown. The structure of docked ligands in 121p and 1hsg shows a reasonable match with the corresponding crystal structure. While overlapping with the crystal ligands, the other two docked ligands orient differently. The tail of the docked ligand in 1dhf orients in a different direction than that in the crystal structure. The docked ligand in 3cpa is shifted and flipped with respect to the crystal ligand.

ulin by consideration of conformation changes that mimic an induced fit.[48] We also found that 4H-tamoxifen can be placed in calmodulin with a slightly less favorable steric interaction than allowed by the INVDOCK scoring function. An appropriate conformation change in calmodulin might allow for the removal of this unfavorable interaction.

The limited number of protein entries available in our cavity database is also expected to result in missed hits. For instance, the known tamoxifen metabolizing protein cytochrome P450[49] is not identified in this work because no corresponding human or mammalian entry is available in the cavity database. A search of bacterial proteins in the database identified this protein (PDB Id: 5cp4 and 1cpt) as a potential target.

As shown in Table II, apart from the 10 potential protein targets that have been implicated or confirmed experimentally, INVDOCK identified 10 other proteins as potential targets of 4H-tamoxifen. These include aldose reductase, arginase, carbonic anhydrase, macrophage migration inhibitory factor, purine nucleoside phosphorylase, DNA polymerase β, hypoxanthine-guanine phosphoribosyltransferase, retinoic acid oxidase, sepiapterin reductase, and tyrosine 3-monooxygenase. We have not found in the literature a report that links tamoxifen to each of these proteins. There is also no report that indicates each of these proteins is not a target of tamoxifen or its analogs. Further investigation is, therefore, needed to determine whether or not 4H-tamoxifen can bind to these proteins.

Potential Protein Targets of Vitamin E

Vitamin E is a widely used supplement, and many experimental studies have indicated its therapeutic effect for a number of diseases.[50–61] It is, therefore, of interest to probe multiple protein targets of this molecule. Potential human and mammalian protein targets of vitamin E identified by INVDOCK are given in Table III together with respective clinical implications from experimental finding. The identified potential protein targets include known targets of vitamin E, such as glutathione *S*-transferase[50] and acetylcholinesterase.[51] Experiments have shown that vitamin E alters the activity of each of the following seven INVDOCK identified proteins: 17-β-hydroxysteroid-dehydrogenase,[52] alcohol dehydrogenase,[53] glutathione synthetase,[54] D-amino acid oxidase,[55] guanylyl cyclase,[56] prostaglandin endoperoxide synthase,[57] and nitric oxide synthase.[58] Thus, these experimental findings might indicate direct binding of vitamin E to each of these proteins. It has been observed that the administration of vitamin E changes the level of each of the four other INVDOCK-identified proteins: aldose reductase,[59] immunoglobulin,[60] purine nucleoside phosphorylase,[61] and C-H-Ras 21 protein.[62] As ligand binding is known to self-regulate the protein level in certain cases,[46] there is a possibility that each of these three proteins is also a target of vitamin E as implicated by the INVDOCK search.

There are 11 other potential protein targets identified by INVDOCK. These are inositol monophosphatase, collagenase, lactoferrin, transducin, estrogen sulfotransferase, fructose-2,6-bisphosphatase, FR-1 protein, procarboxypeptidase A, 70-kDa heat shock protein, insulin, and α-chymotrypsin A. We have found no report that either implicates or excludes each of these proteins as a possible target of vitamin E. Hence, further validation tests are needed to clarify these proteins.

Factors That Might Affect the Quality of a Ligand–Protein Inverse-Docking Search

Overall, about 50% of INVDOCK-identified potential protein targets of 4H-tamoxifen and vitamin E have been

TABLE II. Putative Protein Targets of 4H-Tamoxifen Identified From an INVDOCK Search of Human and Mammalian Proteins

PDB	Putative protein target	Experimental finding	Target status	Clinical implication	Reference
1a52	Estrogen receptor	Drug target	Confirmed	Treatment of breast cancer	36
1akz	Uracil-DNA glycosylase				
1ayk	Collagenase	Inhibited activity	Confirmed	Tumor cell invasion and cancer metastasis	38
1az1	Aldose reductase				
1bnt	Carbonic anhydrase				
1boz	Dihydrofolate reductase	Decreased level		Combination therapy for cancer	44
1d3v	Arginase				
1d6n	Hypoxanthine-guanine phosphoribosyltransferase				
1dda	Alcohol dehydrogenase	Inhibition	Confirmed	Enhanced ethanol's sedative effect	40
1dht, 1fdt	17β-Hydroxysteroid dehydrogenase	Inhibitor	Confirmed	Promotion of tumor regression	39
1gsd, 3ljr	Glutathione transferase A1-1, glutathione *S*-transferase	Suppressed enzyme and activity		Genotoxicity and carcinogenicity	42
1mch	Immunoglobulin λ light chain	Temerarily enhanced lg level		Modulation of immune response	45
1p1g	Macrophage migration inhibitory factor				
1ulb	Purine nucleoside phosphorylase				
1zqf	DNA polymerase β				
2nll	Retinoic acid receptor				
1a25	Protein kinase C	Inhibition	Confirmed	Anticancer	37
1aa8	D-Amino acid oxidase				
1afs	3α-Hydroxysteroid dehydrogenase	Effect on androgen-induced activity		Hepatic steroid metabolism	43
1pth	Prostaglandin H2 synthase-1	Direct inhibition	Confirmed	Prevention of vasoconstriction	41
1sep	Sepiapterin reductase				
2toh	Tyrosine 3-monooxygenase				

implicated or confirmed by experiments. Several reasons might contribute to the discrepancy between INVDOCK screening results and experimental data. It is not expected that exhaustive experiments have been done to determine all protein targets of a drug. Thus, a discrepancy might arise for those identified proteins that lack relevant experimental information. Some of the PDB structures may be of little relevance to a binding study for a particular molecule. These include entries containing an incomplete section or a chain, protein mutants that are structurally different from the corresponding proteins investigated in experiments, ligand-bound proteins whose conformation is relevant only to a specific set of ligands, and macromolecular complexes unrelated to a particular biological process studied experimentally. The docking of a molecule to such an irrelevant structure may thus generate a false hit. Anticipated rapid progress in structural genomics[23] is expected to provide a more diverse set of relevant structures. Knowledge from the study of protein functions[25] also facilitates the selection of relevant structures in the determination of putative protein targets related to a particular cellular or physiological condition.

Another possible cause of discrepancy between INVDOCK screening results and experimental data is a lack

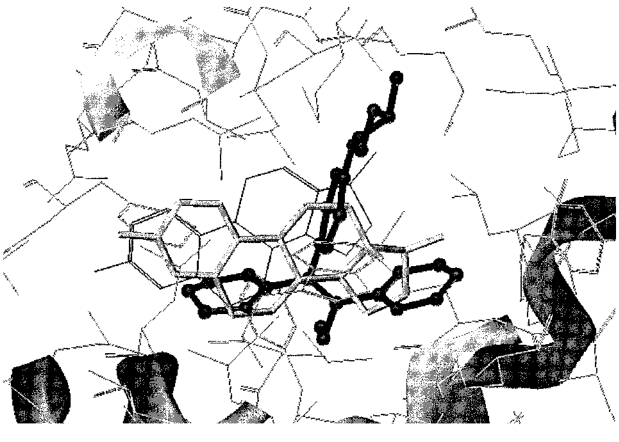

Fig. 4. Comparison of the docked structure of 4H-tamoxifen (dark ball-and-stick structure) with the crystal structure of estrogen (gray stick structure) in the protein estrogen receptor (PDB Id: 1a52). For comparison, the docked structure of estrogen (dark line-drawing structure) is also included. Docked tamoxifen has a more extensive van der Waals contact and a stronger hydrogen-bond interaction, which might contribute to its binding competitiveness against estrogen.

TABLE III. Putative Protein Targets of Vitamin E Identified From an INVDOCK Search of Human and Mammalian Proteins

PDB	Putative protein target	Experimental finding	Target status	Clinical implication	Reference
1a27	17β-Hydroxysteroid-dehydrogenase	Stimulated activity		Testicular steroidogenesis	52
1az1	Aldose reductase	Increased level		Treatment of cataract development	59
1bmk	Map kinase P38				
1crr	C-H-Ras P21 protein	Decreased H-Ras expression		Improved cancer therapy	62
1dda	Alcohol dehydrogenase	Potentiated ethanol effect on activity		Ethanol metabolism in liver	53
1imc	Inositol monophosphatase				
1mcd	Immunoglobulin λ light chain	Increased IgM level		Effect on immune system	60
20gs	Glutathione *S*-transferase (GST)	Direct binding	Confirmed	Liver GST activity inhibition	50
2clj	Acetylcholinesterase	Inhibition of activity	Confirmed	Treatment of poisoning, alzheimer's	51
2hgs	Glutathione synthetase	Increased activity		Antioxidative effect	54
4ayk	Collagenase				
1a9o	Purine nucleoside phosphorylase (PNP)	Attenuated PNP release		Attenuated cell injury	61
1aa8	D-Amino acid oxidase	Effect of Vitamin E Deficiency on activity		Effect on hepatic peroxisomes	55
1awn	Guanylyl cyclase	Reduced activity			56
1b7u	Lactoferrin				
1b9x	Transducin				
1bo6	Estrogen sulfotransferase				
1djj	Prostaglandin endoperoxide synthase	Effect on activity			57
1fbt,	Fructose-2,6-bisphosphatase				
1frb	FR-1 protein				
1pca	Procarboxypeptidase A				
1qqm	D199S mutant of bovine 70 kDa heat shock protein				
1wav	Insulin				
2nse	Nitric oxide synthase	Increased activity		Blood-pressure reduction	58
6cha	α-Chymotrypsin A				

of consideration of protein profiles such as gene expression patterns and protein levels. Some experimental studies of ligand–protein interactions are based on the investigation of cell lines or other assays. The observation of molecular events related to a particular ligand–protein interaction requires that the protein under study be at a sufficient level in the system being investigated. If such a level is not reached at a particular setting, the corresponding experiment is not useful in probing the binding of a molecule to that protein. Proteins not expressed or ones expressed at excessively low levels in a particular biological process are unlikely a detectable and thus therapeutically meaningful target. Advances in proteomics are providing a rapidly growing body of information about the profiles of proteins inside cells.[24] The incorporation of this information into an inverse-docking procedure can enable the prediction of more relevant protein targets.

The pharmacokinetic and metabolic profile of a molecule may also need to be considered in an inverse-docking procedure. The action of a therapeutic molecule requires it to achieve an adequate concentration in the fluid bathing the target tissue. The concentration of a molecule is determined by its pharmacokinetic and metabolic profile. Therefore, information about this profile is important in the prediction of potential protein targets that a small molecule can reach at a sufficient concentration. Developments in pharmacokinetics and drug metabolism[26] are providing more and more information in this regard.

The capability of an inverse-docking approach in identifying potential protein targets of a small molecule is constrained by the relatively limited number of available protein 3D structures. This is particularly true for membrane-bound receptor proteins that are key therapeutic and toxicity targets for a variety of diseases or symptoms. This problem can be partially alleviated by structural information generated from rapid progress in high-throughput X-ray crystallography of protein structures and in the development of new structural determination methods.[23]

The quality of an inverse-docking procedure also depends on the algorithm and force fields used in docking and scoring. Ideally, in an inverse-docking procedure the optimum binding mode should be searched in a cavity. Although more CPU-time demanding, such a search may become practical if one can improve the overall inverse-docking search speed. For instance, the search speed of

INVDOCK can be significantly improved by the distribution of a database search onto multiple computers. Cavity models may also be further refined to reduce the search space. CPU time saved from these and other improvements may be used for searching optimum binding modes. There has been progress in refining docking/scoring algorithms and force fields, particularly in the areas of protein flexibility[34] and ligand flexibility.[4–19] Knowledge and new methods and force fields generated from these studies can also be incorporated into an inverse-docking procedure.

In a ligand–protein docking study, potential ligands are typically selected from the lowest energy docked structures. On the other hand, in an inverse-docking procedure potential protein targets are selected on the basis of an energy threshold. Docked structures with ligand–protein interaction energies lower than that threshold are considered putative targets. However, this energy threshold is more difficult to define. A stricter condition on the energy threshold can easily result in missed hits. Likewise, an excessively relaxed energy threshold may lead to the overproduction of false hits. The empirical formula for the energy threshold used in this study seems to give reasonable results for the two agents studied. However, additional study is needed to more extensively evaluate and further refine the energy threshold.

CONCLUSION

Ligand–protein inverse-docking is proposed as an approach for the computer-aided identification of potential protein targets of a small molecule. The testing results of the computer program INVDOCK show the potential of this approach. The performance and applicability of the ligand–protein inverse-docking program need to be further enhanced by the refinement of docking and scoring algorithms, the development of more efficient cavity models, and the incorporation of new information from advances in structural genomics, proteomics, protein function, pharmacokinetics, and drug metabolism. The improvement of search speed is also a key in improving docking quality, which allows for the introduction of more sophisticated docking algorithms such as the search for the optimum binding mode in a cavity and more accurate modeling of protein flexibility. One method is to distribute the task of a database search onto multiple computers. Further development of this inverse-docking approach may bring interesting applications such as the determination of unknown and secondary therapeutic targets of drugs, drug leads, natural products, and synthetic chemicals and the identification of protein targets related to the side effects and toxicity of these molecules.

ACKNOWLEDGEMENTS

This work was supported in part by the Singapore Academic Research Fund.

REFERENCES

1. Kuntz ID. Structure-based strategies for drug design and discovery. Science 1992;257:1078–1082.
2. Blundell TL. Structure-based drug design. Nature 1996;384(*Suppl* 6604):23–26.
3. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule–ligand recognitions. J Mol Biol 1982;161:269–288.
4. Miller MD, Kearsley SK, Underwood DJ, Sheridan RP. FLOG: a system to select "quasi-flexible" ligands complementary to a receptor of known three dimensional structure. J Comput Aided Mol Des 1994;8:153–174.
5. Judson RS, Jaeger EP, Treasurywala AM. A genetic algorithm method for molecular docking molecules. J Mol Struct 1994;308: 191–206.
6. Oshiro CM, Kuntz ID, Dixon JS. Flexible ligand docking using a genetic algorithm. J Comput Aided Mol Des 1995;9:113–130.
7. Jones G, Willet P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol 1997;267:727–748.
8. Gehlhaar DK, Verkhivker GM, Rejto PA, Sherman CJ, Fogel DB, Fogel LJ, Freer ST. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. Chem Biol 1995;2:317–324.
9. Morris GM, Goodsell DS, Huey R, Olson AJ. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. J Comput Aided Mol Des 1996;10:293–304.
10. Bohm HJ. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. J Comput Aided Mol Des 1992;6:61–78.
11. Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm. J Mol Biol 1996;261:470–489.
12. Ewing T, Kuntz ID. Critical evaluation of search algorithms for automated molecular docking and database screening. J Comp Chem 1997;18:1175–1189.
13. Welch W, Ruppert J, Jain AN. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. Chem Biol 1996;3:449–462.
14. Eisen MB, Wiley DC, Karplus M, Hubbard RE. Hook—a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a molecular binding site. Proteins 1994;19:199–221.
15. McMartin C, Bohacek RS. QXP: powerful, rapid computer algorithms for structure-based drug design. J Comput Aided Mol Des 1997;11:333–344.
16. Blom NS, Sygusch J. High resolution fast quantitative docking using fourier domain correlation techniques. Proteins 1997;27:493–506.
17. Baxter CA, Murray CW, Clark DE, Westhead DR, Eldridge MD. Flexible docking using Tabu search and an empirical estimate of binding affinity. Proteins 1998;33:367–382.
18. Lorber DM, Shoichet BK. Flexible ligand docking using conformational ensembles. Protein Sci 1998;7:938–950.
19. Wang J, Kollman PA, Kuntz ID. Flexible ligand docking: a multistep strategy approach. Proteins 1999;36:1–19.
20. Royer RJ. Mechanism of action of adverse drug reactions: an overview. Pharmcoepidemiol Drug Saf 1997;6(*Suppl* 3):S43–S50.
21. DiMasi JA, Bryant NR, Lasagna L. New drug development in the United States from 1963 to 1990. Clin Pharmacol Ther 1991;50: 471–486.
22. Drews J. Strategic choices facing the pharmaceutical industry: a case for innovation. Drug Discov Today 1997;2:72–78.
23. Sali A. 100,000 protein structures for the biologist. Nat Struct Biol 1998;5:1029–1032
24. Persidis A. Proteomics. An ambitious drug development platform attempts to link gene sequence to expressed phenotype under various physiological states. Nat Biotechnol 1998;16:393–394.
25. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan YJ. Predicting function: from genes to genomes and back. J Mol Biol 1998;283:707–725.
26. Lin JH, Lu AY. Role of pharmacokinetics and metabolism in drug discovery and development. Pharmacol Rev 1997;49:404–449.
27. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
28. Rost B, Sander C. Bridging the protein sequence–structure gap by structure predictions. Annu Rev Biophys Biomol Struct 1996;25: 113–136.
29. Richards FM. Areas, volumes, packing and protein structure. Annu Rev Biophys Bioeng 1977;6:151–176.
30. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr,

Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins and nucleic acids. J Am Chem Soc 1995;117:5179–5197.

31. Baird NC. Simulation of hydrogen bonding in biological systems: ab initio calculations for NH3–NH3 and NH3–NH4+. Int J Quantum Chem Symp 1974;1:49–53.
32. Chen YZ, Prohofsky EW. The role of a minor groove spine of hydration in stabilizing poly(dA).poly(dT) against fluctuational interbase H-bond disruption in the premelting temperature regime. Nucleic Acids Res 1992;20:415–419.
33. Chen YZ, Prohofsky EW. Premelting base pair opening probability and drug binding constant of a daunomycin—Poly d(GCAT)–Poly d(ATGC) complex. Biophys J 1994;66:820–826.
34. Carlson HA, McCammon JA. Accommodating protein flexibility in computational drug design. Mol Pharmacol 2000;57:213–218.
35. Nicklaus MC, Milne GW, Zaharevitz D. Chem-X and CAMBRIDGE. Comparison of computer generated chemical structures with X-ray crystallographic data. J Chem Inf Comput Sci 1993;33: 639–646
36. Favoni RE, Cupis AD. Steroidal and nonsteroidal oestrogen antagonists in breast cancer: basic and clinical appraisal. Trends Pharmacol Sci 1998;19:406–415.
37. Rowlands MG, Budworth J, Jarman M, Hardcastle IR, McCague R, Gescher A. Comparison between inhibition of protein kinase C and antagonism of calmodulin by tamoxifen analogues. Biochem Pharmacol 1995;50:723–726.
38. Abbas Abidi SM, Howard EW, Dmytryk JJ, Pento JT. Differential influence of antiestrogens on the in vitro release of gelatinases (type IV collagenases) by invasive and non-invasive breast cancer cells. Clin Exp Metastasis 1997;15:432–439.
39. Santner SJ, Santen RJ. Inhibition of estrone sulfatase and 17 beta-hydroxysteroid dehydrogenase by antiestrogens. J Steroid Biochem Mol Biol 1993;45:383–90.
40. Messiha FS. Leu-enkephalin, tamoxifen and ethanol interactions: effects on motility and hepatic ethanol metabolizing enzymes. Gen Pharmacol 1990;21:45–48.
41. Ritchie GA. The direct inhibition of prostaglandin synthetase of human breast cancer tumor tissue by tamoxifen. Recent Results Cancer Res 1980;71:96–101.
42. Nuwaysir EF, Daggett DA, Jordan VC, Pitot HC. Phase II enzyme expression in rat liver in response to the antiestrogen tamoxifen. Cancer Res 1996;56:3704–3710.
43. Lax ER, Rumstadt F, Plasczyk H, Peetz A, Schriefers H. Antagonistic action of estrogens, flutamide, and human growth hormone on androgen-induced changes in the activities of some enzymes of hepatic steroid metabolism in the rat. Endocrinology 1983;113: 1043–1055
44. Levine RM, Rubalcaba E, Lippman ME, Cowan KH. Effects of estrogen and tamoxifen on the regulation of dihydrofolate reductase gene expression in a human breast cancer cell line. Cancer Res 1985;45:1644–1650.
45. Paavonen T, Aronen H, Pyrhonen S, Hajba A, Andersson LC. The effect of toremifene therapy on serum immunoglobulin levels in breast cancer. APMIS 1991;99:849–853.
46. Schmidt TJ, Meyer AS. Autoregulation of corticosteroid receptors. How, when, where, and why? Receptor 1994;4:229–257.
47. Meador WE, Means AR, Quiocho FA. Modulation of calmodulin plasticity in molecular recognition on the basis of X-ray structures. Science 1993;262:1718–1721.
48. Sandak B, Wolfson HJ, Nussinov R. Flexible docking allowing induced fit in proteins: insights from an open to closed conformational isomers. Proteins 1998;32:159–174
49. Crommentuyn KM, Schellens JH, van den Berg JD, Beijnen JH. In-vitro metabolism of anti-cancer drugs, methods and applications: paclitaxel, docetaxel, tamoxifen and ifosfamide. Cancer Treat Rev 1998;24:345–366.
50. Arita M, Sato Y, Arai H, Inoue K. Binding of alpha-tocopherylquinone, an oxidized form of alpha-tocopherol, to glutathione-S-transferase in the liver cytosol. FEBS Lett 1998;436:424–426.
51. Chelliah J, Smith JD, Fariss MW. Inhibition of cholinesterase activity by tetrahydroaminoacridine and the hemisuccinate esters of tocopherol and cholesterol. Biochim Biophys Acta 1994;18:17–26.
52. Reddy GP, Prasad M, Sailesh S, Kumar YV, Reddanna P. Arachidonic acid metabolites as intratesticular factors controlling androgen production. Int J Androl 1993;16:227–233.
53. Tyopponen JT, Lindros KO. Combined vitamin E deficiency and ethanol pretreatment: liver glutathione and enzyme changes. Int J Vitam Nutr Res 1986;56:241–245.
54. Makar TK, Nedergaard M, Preuss A, Gelbard AS, Perumal AS, Cooper AJ. Vitamin E, ascorbate, glutathione, glutathione disulfide, and enzymes of glutathione metabolism in cultures of chick astrocytes and neurons: evidence that astrocytes play an important role in antioxidative processes in the brain. J Neurochem 1994;62:45–53.
55. Suga T, Watanabe T, Matsumoto Y, Horie S. Effects of long-term vitamin E deficiency and restoration on rat hepatic peroxisomes. Biochim Biophys Acta 1984;794:218–224.
56. Sobolev AS, Tertov VV, Rybalkin SD. Activation of guanyl cyclase during lipid peroxidation of biomembranes. Biokhimiia 1982;47: 1251–1261.
57. Greenberg-Levy SH, Budowski P, Grossman S. Lipoxygenase and other enzymes of arachidonic acid metabolism in the brain of chicks affected by nutritional encephalomalacia. Int J Biochem 1993;25:403–409.
58. Newaz MA, Nawal NN, Rohaizan CH, Muslim N, Gapor A. Alpha-tocopherol increased nitric oxide synthase activity in blood vessels of spontaneously hypertensive rats. Am J Hypertens 1999;12:839–844.
59. Libondi T, Menzione M, Iuliano G, Della Corte M, Latte F, Auricchio G. Changes of some biochemical parameters of the lens in galactose-treated weaned rats with and without vitamin E therapy. Ophthalmic Res 1985;17:42–48.
60. Reddy PG, Morrill JL, Minocha HC, Morrill MB, Dayton AD, Frey RA. Effect of supplemental vitamin E on the immune system of calves. J Dairy Sci 1986;69:164–171.
61. Bailey SM, Reinke LA. Antioxidants and gadolinium chloride attenuate hepatic parenchymal and endothelial cell injury induced by low flow ischemia and reperfusion in perfused rat livers. Free Radic Res 2000;32:497–506.
62. Prasad KN, Kumar A, Kochupillai V, Cole WC. High doses of multiple antioxidant vitamins: essential ingredients in improving the efficacy of standard cancer therapy. J Am Coll Nutr 1999;18: 13–25.