MATERIAL AND METHODS

Methodologies

1. Pharmacophore Searching

Pharmacophore searching is an application of ligand-based method. If the structural information about the target receptor is not available, it is possible to derive the important features from the ligand, as called pharmacophore. The pharmacophore can be characterized directly by analyzing either receptor–ligand complexes or the binding site of the receptor. In case of unknown receptor structures, a pharmacophore model is generated from known active compounds, usually by calculating all of the possible superpositions of predefined chemical groups. The features used to define a pharmacophore are typical to be the positions of specific atoms or groups within each molecule, e.g. aromatic group, hydrogen bond donors and acceptors, positively and negatively charged group and hydrophobic group.

To combine the relationships between the pharmacophore, it is called a three dimensional (3D) pharmacophore. Each pharmacophore can be combined by using the distance or distance ranges (average interfeature distances and their tolerances), angles, or planes (Van Drie, 1997). A 3D pharmacophore is the 3D arrangement of chemical groups that is required for the biological activity of a molecule. During database searching, the 3D pharmacophore model serves a template for the selection of molecules having the specified geometrical constraints (Seifert *et. al.*, 2003). An example of 3D pharmacophore based on the description of NNRTIs series is shown in Figure 8. The requirements of this pharmacophore model are a conjugated system between two-hydrogen bonding acceptor atoms (Xs) within plane A, an aromatic center (AC), and two hydrocarbon centers (HC).

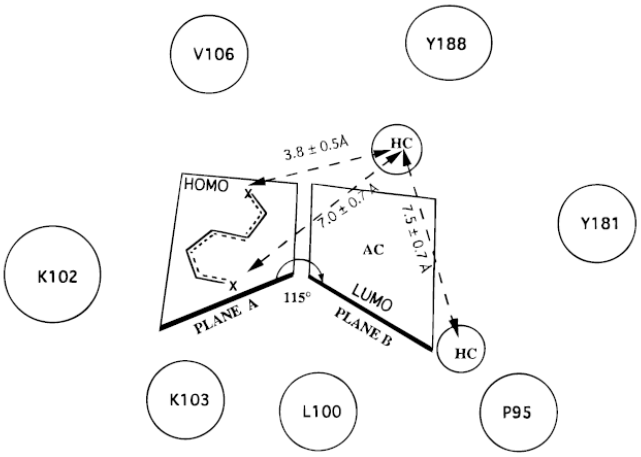


Figure 8 3D pharmacophore model derived from a series of NNRTIs.
Source: Gussio *et al*. (1996).

2. Molecular Docking

Molecular docking is a method to identify correct poses of ligands in the binding pocket of a protein and to predict the affinity between the ligand and the protein (Krovat *et al.*, 2005). A general docking procedure consists of three steps: identification of the binding site, a search algorithm to effectively sample the set of orientation and conformation of ligands, and a scoring function (McConkey *et al.*, 2002). More details of molecular docking method are described in Appendix A. In this study, some docking tools, FlexX, GOLD and Surflex, have been applied for performing virtual screening. Their docking algorithms and scoring functions are shown in Table 2.

Table 2 Docking algorithm and scoring functions of the docking tools.

Method	Docking Algorithm	Scoring Function
FlexX	Incremental construction	Empirical scoring function
GOLD	Genetic algorithm	Force field scoring function (GoldScore) and Empirical scoring function (ChemScore)
Surflex	A molecular similarity method (morphological similarity)	Empirical scoring function (Hammerhead scoring function)

The details of each docking method are described as following.

2.1 FlexX

FlexX (Rarey *et. al.*, 1996, 1999; Hoffmann *et al.*, 1999) is a fast, flexible docking method that uses an incremental construction algorithm. It rapidly docks a conformationally flexible ligand into a binding site. The conformational flexibility of the ligand is included by generating multiple conformations for each fragment and including all in the ligand building steps. The base fragment of the ligand is automatically selected and placed into the active site. After that, the remainder of ligand is incrementally reconstructed from other fragments and the placement of the ligand is scored on the basis of protein-ligand interactions. Protein-ligand interactions in FlexX include both polar (Hydrogen bond and charge-charge) and non-polar (hydrophobic) interactions. FlexX used the following model to describe molecular interactions based on spherical surfaces. Each interacting group is assigned both an interaction type and interaction geometry. The geometry contains an interactions center c, an interaction radius r, and an interaction surface, which is part of the spherical surface with radius r about c. An example of interaction center and surface of the carbonyl group is shown in Figure 9a. The interaction surfaces can be spheres, cones, capped cones and spherical rectangles (Figure 9b).

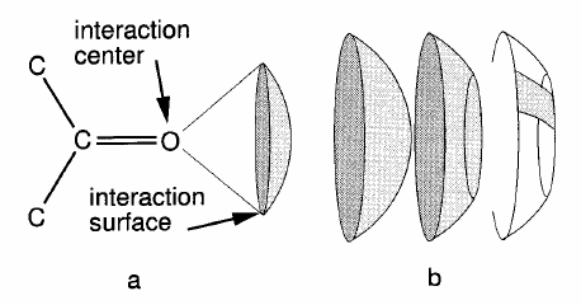


Figure 9 Example of carbonyl group's interaction model in FlexX program.

a). Interaction center and surface of the carbonyl group.

b). Spheres, cones, capped cones and spherical rectangles interaction surfaces.

Source: Rarey (1996).

Docking Algorithms

In FlexX, an incremental construction strategy was used to be a docking algorithm. The FlexX's docking algorithm procedure consists of three steps (Kramer *et. al.*, 1999). They are base selection, base placement, and complex construction. Initially, the base selection is the selection of a connected part of the ligand, the base fragment. For the better computing times and docking results, the number of potential interaction groups should be maximized while the number of alternative conformations of the base fragment should be minimized. Once the base fragment is selected, the remaining part of the ligand is automatically divided into fragments. The base fragments are the first parts of the ligand that are placed into the active site, as called the base placement. The goal of the fragment placement algorithm is to find positions of the base fragment in the active site such that a sufficient number of favorable interactions between the fragment and the protein can occur simultaneously. The fragment placement algorithm is shown in Figure 10. The interaction surfaces on the receptor site are represented by the interaction points. To place the base fragment into the receptor site, three interaction center of the fragment are mapped onto three interaction points of the receptor by superposing the three point pairs onto each other. After the base fragment has been located into favorable placement, the complex

construction step is started by using incremental construction algorithms. The remaining fragments are added to the base fragment.

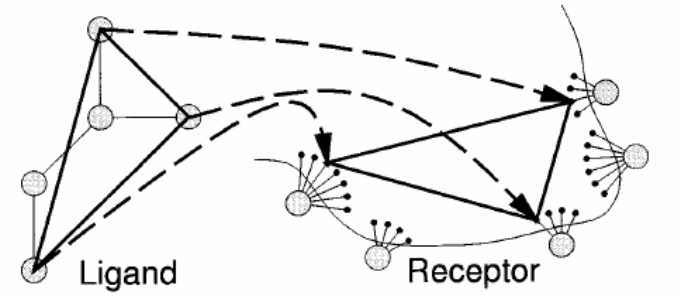


Figure 10 Fragment placement algorithm in FlexX.
Source: Rarey (1996).

Scoring function

The estimated free energy of binding was calculated from empirical scoring function, as shown in the follow equation.

$$\begin{aligned}\Delta G = \Delta G_0 &+ \Delta G_{rot} \times N_{rot} \\ &+ \Delta G_{hb} \sum_{neutral\ H-bonds} f(\Delta R, \Delta\alpha) \\ &+ \Delta G_{io} \sum_{ionic\ \mathrm{int.}} f(\Delta R, \Delta\alpha) \\ &+ \Delta G_{aro} \sum_{aro\ \mathrm{int.}} f(\Delta R, \Delta\alpha) \\ &+ \Delta G_{lipo} \sum_{lipo.cont.} f^*(\Delta R) \end{aligned} \quad (1)$$

Where N_{rot} is the number of rotatable bonds that are immobilized in the complex. ΔG_{hb}, ΔG_{io}, ΔG_{rot}, and ΔG_0 are adjustable parameters. ΔG_{aro} accounts for the interactions of aromatic groups. ΔG_{lipo} is a modified term that is calculated as a pairwise sum over all atom-atom contacts. $f(\Delta R,\Delta\alpha)$ is a scaling function that penalizes deviations from ideal geometry. $f^*(\Delta R)$ is a scaling factor accounts for contacts with ideal distance but penalizes for close contacts. The scaling factors are shown in equation (2).

$$f^*(\Delta R) = \begin{cases} 0 & \Delta R > 0.6\ \text{Å} \\ 1 - \frac{\Delta R - 0.2}{0.4} & 0.2\ \text{Å} < \Delta R \leqslant 0.6\ \text{Å} \\ 1 & -0.2\ \text{Å} < \Delta R \leqslant 0.2\ \text{Å} \\ 1 - \frac{-\Delta R - 0.2}{0.4} & -0.6\ \text{Å} < \Delta R \leqslant -0.2\ \text{Å} \\ \frac{\Delta R + 0.6}{0.2} & \Delta R \leqslant -0.6\ \text{Å} \end{cases} \quad (2)$$

Where $\Delta R = R - R_0$, R is the distance between the atom centers. R_0 is the ideal value, which is assumed to be the sum of both van der Waals radii plus 0.6 Å. From the scoring function, the first term (ΔG_0) is a fixed ground term. The second term ($\Delta G_{rot} \times N_{rot}$) is a term taking into account the loss of entropy during ligand binding due to the hindrance of rotatable bonds. The third-fifth terms (terms contributing of hydrogen bond, ionic interaction, and aromatic interaction) are sums overall pairwise interactions. The last term rates the atom-atom contacts between protein and ligand, which are hydrophobic contacts and forbiddingly close contacts (clashes). The functions $f(\Delta R,\Delta\alpha)$ and $f^*(\Delta R)$ are heuristic distance and angle dependent penalties.

Advantages and disadvantages

In FlexX program, the advantages and disadvantages are shown in Table 3.

Table 3 Advantages and disadvantages of FlexX program.

Advantages/ Disadvantages	FlexX
Advantages	- FlexX is fast. The computing time depends on the size of the active site, the size of the ligand and the degree of ligand symmetry and lies in the range from a few seconds up to a few minutes. - FlexX is automatically positioning the ligand. - The conformational flexibility of the ligand is taken into account by considering both the torsion angle flexibility as well as the conformational flexibility of ring systems. - Enantiomerism can be handled as a degree of freedom. - The placement algorithm in FlexX is based on the evaluation of interactions occurring between the protein and ligand. This ensures that the search is limited to low-energy structures improving the quality of the results in a given amount of computing time. - The protein interaction is fully specification of active site including oxidation states, metal ions and sidechain protonation states.
Disadvantages	- The protein is considered rigid. However, the FlexE module can be employed to consider flexibility of the protein. - FlexX scoring function was not as effective due to the multiple lipophilic contacts of inhibitors. - Tends to place conformationally flexible ligands outside the binding site (if topological constraints exist).

2.2 GOLD

GOLD (Genetic Optimization for Ligand Docking) is a docking method using a genetic algorithm (GA) for docking flexible ligands into protein binding sites (Jones *et al.*, 1997; Nissink *et al.*, 2002; Taylor, 2002; Verdonk *et al.*, 2003). Ligand was treated as fully flexible and protein was treated as partially flexible in the neighborhood of the protein active site i.e. OH and NH_3 groups of protein residues. Specifically, each Ser, Thr and Tyr OH will be allowed to rotate to optimise its hydrogen-bonding to the ligand. A possible docked orientation of the ligand is encoded as a chromosome. A chromosome contains the information about the mapping of ligand H-bond atoms onto protein H-bond atoms, mapping of hydrophobic points on the ligand onto protein hydrophobic points, and the conformation around flexible ligand bonds and protein OH groups. A fitness score of each chromosome was assigned based on its predicted binding affinity. GOLD has been tested on a data set bigger than 100 of complexes extracted from the Protein Data Bank. These tests revealed that GOLD achieved about 70-80% success rate in identifying the experimental binding mode. GOLD will only produce reliable results if it is used properly and correct atom typing for both protein and ligand is particularly important.

Docking Algorithms

Genetic algorithm (GA) approach has been used to explore conformation space and ligand binding modes. It was inspired by the Darwinian principles of evolution. GA operates on the relation of two spaces – the space that encodes possible discrete solutions (genotype) and the property space (phenotype) of these solutions. In the GA, the chemical structure of a compound can be understood as the genotype or chromosome which is constructed from genes – atoms or building blocks like amino acids or nucleotides – whereas the actual molecule with its physical and biological properties represents the phenotype. GA uses several different genotypes or individuals at the same time – the population – and investigate their properties using a given selection function. After such evaluation of the population, the individual

members are rank-ordered according to their fitness. The chromosomes of the rank-ordered individuals are then subjected to changes (genetic operators) that generate new chromosomes according to predefined rules. These genetic operators are inspired by those of DNA like genetics: death, replication, insertion, mutation, and crossover. Replication regenerates an equivalent chromosome or individual. Mutation sets one or more elements in the parent gene to a different value, based on a predefined mutation rate between 0 and 100 percent. Crossover takes two or more chromosomes to build new chromosomes by mixing them according to various rules. Deletion deletes an element from the parent gene. Insertion introduces new elements. In nature, these mechanisms enable life forms to adapt to a particular environment over successive generations. The fitness function measures each individual phenotype's fitness within this population. The higher the fitness, the greater should be the probability of passing the genomic information onto the next generation.

GOLD uses a genetic algorithm for optimizing the fitness score. A population of potential solutions (i.e. possible docked orientations of the ligand) is set up at random. Each member of the population is encoded as a chromosome, which contains information about the mapping of ligand H-bond atoms onto (complementary) protein H-bond atoms, mapping of hydrophobic points on the ligand onto protein hydrophobic points, and the conformation around flexible ligand bonds and protein OH groups. Each chromosome is assigned a fitness score based on its predicted binding affinity and the chromosomes within the population are ranked according to fitness.

Scoring function

GOLD offers a choice of fitness functions: GoldScore, ChemScore and User Defined Score. The original GOLD scoring function is GoldScore and it is selected to be default scoring function in GOLD.

1) GoldScore fitness function

The GoldScore fitness function is a force field scoring function consisting of the sum of protein-ligand complexation term, a hydrogen-bonding term and an internal energy term. Force-field scoring function that was used to calculate these contribution terms is shown in the following equation.

$$
\begin{aligned}
E_{total} &= E_{complex} + E_{H-bond} + E_{\text{int}ernal} \\
&= \sum_{protein} \sum_{ligand} \left(\frac{A_{ij}}{d_{ij}^{8}} - \frac{B_{ij}}{d_{ij}^{4}} \right) + \\
&\quad \sum_{protein} \sum_{ligand} [(E_{da} + E_{ww}) - (E_{dw} + E_{aw})] + \\
&\quad \left\{ \sum_{ligand} \left(\frac{C_{ij}}{d_{ij}^{12}} - \frac{D_{ij}}{d_{ij}^{6}} \right) + \sum_{ligand} \frac{1}{2} V \left[1 + \frac{n}{|n|} \cos(|n| \omega \right] \right\}
\end{aligned}
\tag{3}
$$

From the scoring function, a 4-8 potential was used to determine the energy of interaction between the ligand and the protein. This potential is proved to be particularly effective in reproducing experimental ligand binding modes. A 4-8 potential with a linear cut-off is much softer than the 6-12 potential that is traditionally used, allowing the GA to form close contacts with the protein more easily. The hydrogen bonding term is a sum of the individual energies from all the donor-acceptor pairs between the complexes. The energy of each hydrogen bond is calculated with a complicated function considering the type and the geometry of the donor-acceptor pair. The internal energy of the ligand includes a dispersion-repulsion energy and a torsional energy, both of which are calculated according to the Tripos force field. This scoring function was originally calibrated by reproducing the three-dimensional structures of 100 protein-ligand complexes. The fitness score is taken as the negative of the sum of the component energy terms, so that larger fitness scores are better.

2) ChemScore fitness function

ChemScore was derived empirically from a set of 82 protein-ligand complexes for which measured binding affinities were available. Unlike GoldScore, the ChemScore function was trained by regression against measured affinity data (empirical scoring function), although there is no clear indication that it is superior to GoldScore in predicting affinities. ChemScore function is defined in equation (4).

$$\text{ChemScore} = \Delta G_{\text{binding}} + E_{\text{clash}} + E_{\text{int}} + E_{\text{cov}}, \quad (4)$$

The estimated free energy of binding ($\Delta G_{\text{binding}}$) is defined as shown in equation (5).

$$\Delta G_{binding} = \Delta G_0 + \Delta G_{h\,bond} S_{h\,bond} + \Delta G_{metal} S_{metal} + \Delta G_{lipo} S_{lipo} + \Delta G_{rot} H_{rot} \quad (5)$$

Where S_{hbond}, S_{metal} and S_{lipo} are scores for hydrogen-bonding, acceptor-metal, and lipophilic interactions, respectively. H_{rot} is a score representing the loss of conformation entropy of the ligand upon binding to the protein. ΔG terms are coefficients derived from a multiple linear regression analysis on a training set of 82 protein-ligand complexes from the PDB.

The clash term (E_{clash}) is summed over all non-hydrogen protein-ligand atom pairs, as defined in equation (6).

$$E_{clash} = \sum \varepsilon_{clash}(r, r_{clash}) \quad (6)$$

Where r is the distance between a protein-ligand atom pair and r_{clash} is the clash distance for that pair. The clash energy for each atom pair depends on the nature of the protein and ligand atom; it is zero for $r > r_{clash}$, and for $r \leq r_{clash}$.

The internal energy (E_{int}) of the ligand is the sum of a torsional term (E_{tors}) and a clash term (E_{clash}). The latter is calculated analogously to the protein–ligand clash energy, but only for ligand atoms that are separated by at least four

bonds. The torsional term is a summation over the ligand rotatable bonds (RB). Because GOLD also flips ring corners, which affects ring torsion angles, our implementation of the ligand torsional energy also includes a summation over free ring corners (RC) (equation (7)).

$$E_{tors} = \sum_{RB} \varepsilon_{tors}(\theta_{RB}) + \sum_{RC} \sum_{RCB} \varepsilon_{tors}(\theta_{RCB}) \quad (7)$$

The second summation in the right-hand term is over the ring bonds RCB affected by the ring flipping of ring corner RC.

The covalent energy term only applies to ligands bound covalently to the protein. It consists of a torsional part and a bond-angle part, as shown in equation (8).

$$E_{\text{cov}} = \sum_{CB} \varepsilon_{tors}(\theta_{CB}) + C_{\text{cov}} \sum_{BA} k_{BA}(\varphi_{BA} - \varphi_{0,BA})^2 \quad (8)$$

The first summation is over all torsion angles and θ_{CB} involved in the covalent linkage. The second summation is over the covalent bond angles (φ_{BA}) around the covalent linkage. The force constants (k_{BA}) and the ideal bond angles ($\varphi_{0,BA}$) are taken from GOLD. C_{cov} is a constant used to balance the covalent bond term against the rest of the Chemscore function.

The design of GOLD favors the docking of hydrophilic ligands. There are two reasons for this. Firstly, the chromosome encoding in GOLD means that the GA samples binding modes by searching patterns of hydrogen-bonding motifs. Thus the algorithm is directed to find hydrogen-bond networks, whereas it is not guided to find hydrophobic interactions. Secondly, the fitness function contains a term for dispersive interactions but does not have a term for desolvation. With a significant part of the hydrophobic contribution to binding missing from the fitness score, the algorithm is likely to underestimate the contribution to binding from hydrophobic interactions.

Advantages and disadvantages

In GOLD program, the advantages and disadvantages are shown in Table 4.

Table 4 Advantages and disadvantages of GOLD program.

Advantages/ Disadvantages	GOLD
Advantages	- The conformational flexibility of the ligand is taken into account. - Partial flexibility in the neighborhood of the protein active site i.e. OH and NH_3 groups of protein residues was treated. - GOLD offers a choice of scoring functions, GoldScore, ChemScore and User Defined Score. - GOLD allows specification of a protein-ligand covalent bond. - GOLD is able to predict binding to seven metal ions i.e. Mg, Zn, Fe, Mn, Ca, Co and Gd.
Disadvantages	- GOLD required properly and correct starting model for both protein and ligand, e.g. correct atom typing and add all hydrogen atoms, including those necessary for defining the correct ionisation and tautomeric states of the residues such as Asp, Glu and His. - GOLD is not too related to ligand size or flexibility but that performance of the algorithm is highly dependent on ligand hydrophobicity. GOLD is most likely to fail given a hydrophobic ligand.

2.3 Surflex

Surflex (Jain, 2003) is a docking method using a surface-based molecular similarity method to generate suitable putative poses for molecular fragments. The ideas of Surflex are that it firstly assumed an active site of ligand, as called protomol. The protein's surface is coated with a collection of molecular fragments that could potentially interact with the protein. The molecular fragments, as called probe, represent as a potential alignment point for atoms in a ligand (Ruppert *et al.*, 1997). Three types of molecular fragments, i.e. CH_4, C=O and N-H, are placed into the active site in multiple positions. An example of the protomol generated for streptavidin is shown in Figure 11. The protomol is identified based on identification of the protein residues containing any atoms whose surface was within 2.0 Å of any atom of the native ligand biotin.

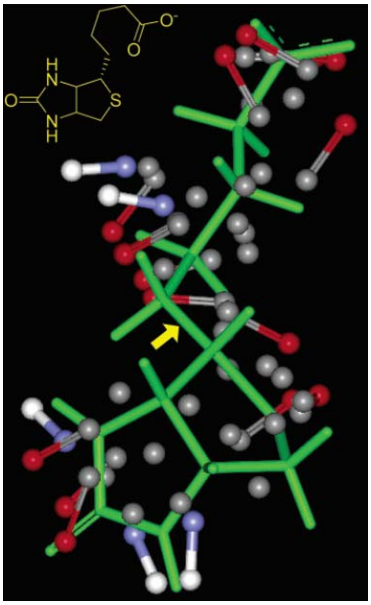


Figure 11 Protomol for streptavidin (1stp) compared with the native pose of biotin (green). The protomol consists of CH_4 (hydrogens not shown), N-H, and C=O molecular fragments.

Source: Jain (2003).

There are two parameters controlling the extent of the protomol. They are proto_thresh and proto_bloat. The proto_thresh determines the degree of buried-ness for the primary volume used to generate the protomol. The proto_bloat indicates how far beyond the primary volume that the protomol volume should be expanded. Flexible docking precedes either by incremental construction from high-scoring fragments as in Hammerhead or by a crossover procedure that combines pieces of poses from intact molecules. Hammerhead's empirical scoring function is then used to predict binding affinities of docked ligands. The docking and scoring procedures of Surflex are shown in Figure 12. The details of docking algorithms and scoring function are described as follows.

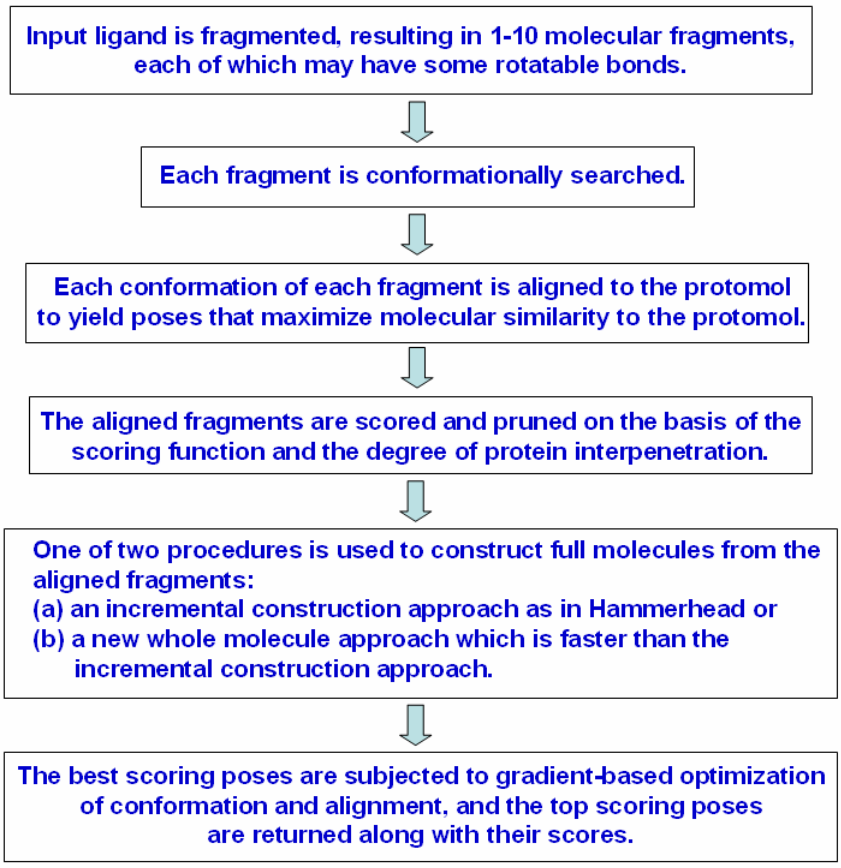


Figure 12 Docking and scoring procedures of Surflex program.

Docking Algorithms

In Surflex, the docking algorithms consist of a process of fragmentation, conformational search, alignment and scoring, and reconstruction from highscoring aligned fragments. Molecules are fragmented by breaking non-ring rotatable bonds, avoiding fragmentation into fragments smaller than 6 heavy atoms. Each fragment is conformationally searched. Following completion of fragmentation and conformational search (and fast internal clash relaxation), the resulting molecular fragments are used in the next step. Surflex utilizes the morphological similarity (MS) function and fast pose generation techniques to generate putative alignments of molecules or molecular fragments to the protomol (Jain, 2000). Morphological similarity is defined as a Gaussian function of the differences in molecular surface distances of two molecules at weighted observation points on a uniform grid. Optimization of the similarity of two molecules is performed by finding sets of observers of each molecule that form triangles of the same size, where each pair of corresponding points in the triangles are observing similar features. The transformation that yields a superposition of the triangles will tend to yield high-scoring superpositions of the molecules. In Surflex's docking search algorithm, poses of molecular fragments that tend to maximize similarity to protomols. Molecular fragments having highest scoring, as called heads, use incremental construction algorithm to add a directed alignment of tail (next molecular fragment) by aligning each conformation based on the similarity to the protomol.

Scoring function

In Surflex, Hammerhead's empirical scoring function is a smooth non-linear function used as scoring function to estimate a binding affinity for the protein-ligand complex in units of $-\log K_d$ (Welch *et al.*, 1996). It was validated with binding affinities of 34 protein/ligand complexes represented a broad range of binding affinities and variety of functional groups (Jain, 1996). The parametrization of the function effectively models the noncovalent interactions of organic ligands with proteins, including proteins with bound metal ions in their active sites. Each atom on

the protein and ligand is labeled as being nonpolar (e.g., the H of a C-H) or polar (e.g., the H of an N-H or the O of a C=O), and polar atoms are also assigned a formal charge if present.

The important terms in scoring function include hydrophobic complimentarily, polar complementarity with additional terms for entropic, and solvation effects. The full scoring function is the sum of each of these terms over the atom pairs (i, j), as shown in equation (9). It is a weighted sum of non-linear functions of exposed protein-ligand atomic van der Waals surface distances. The non-linear functions consist of the Gaussian-like function (f) and sigmoid function (s) of pairwise surface distances.

$$Binding\ \ Affinity = \sum_{i,j} f_0(d(i,j)) + \sum_{i,j} f_1(d(i,j),i,j) + \sum_{i,j} f_2(d(i,j),i,j) + [(l_5.phbe) + (l_6.lhbe)] + [(l_7.n\_rot) + (l_8 \log(mol\ weight))] \quad (9)$$

where

$$d(i,j) = ((x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2)^{1/2} - r_i - r_j \quad (10)$$

The atomic coordinates of atom i are denoted by x_i, y_i, and z_i. The van der Waals radius of atom i is denoted by r_i. The functions f_0, f_1, and f_2 define the hydrophobic, polar and repulsive contributions, respectively, to the binding affinity. The l_5, l_6, l_7, and l_8 represent the tuning parameters of each contribution. Solvation effects are represented by the difference between the total number of potential protein hydrogen-bond equivalents and the actual polar interaction amount (*phbe*) and the difference between the total number of ligand hydrogen-bond equivalents and the actual polar interaction amount (*lhbe*). For the entropic effect, it depends on the number of freely rotatable bonds in the ligand (*n_rot*) and the log of the molecular weight of the ligand (*mol weight*).

Advantages and disadvantages

In Surflex program, the advantages and disadvantages are shown in Table 5.

Table 5 Advantages and disadvantages of Surflex program.

Advantages/ Disadvantages	Surflex
Advantages	- Docking time was roughly linear in number of rotatable bonds, beginning with a few seconds for rigid molecules and adding approximately 3 seconds per rotatable bond on standard Windows hardware. - Protein structure preparation is simple and no special requirement of cofactors/metals/waters is required.
Disadvantages	- Surflex has been designed primarily as a screening tool of small molecule libraries, and over 80% of ligands from commercial small-molecule screening libraries have 15 or fewer rotatable bonds. - Surflex's scoring function was developed strictly on noncovalent complexes, and the utility of screening hits that are reactive is generally thought to be minimal. - The Surflex scoring function does not account intramolecular ligand nonbonded contacts toward a ligand's docking score, and this also contributed the problem.

Computational Details of Calculations

1. CHAPTER I: Virtual Screening of K103N and Y181C HIV-1 Reverse Transcriptase Mutants Based on Nevirapine and Some NNRTIs

The aim of this work is to find novel inhibitors insensitive to the K103N and Y181C mutations of HIV-1 Reverse Transcriptase, based on nevirapine, PNU-142721 and some NNRTIs. Some docking methods, FlexX, GOLD and Surflex, have been performed with the nevirapine-RT complex to validate and select the best possible strategy. Before performing the molecular docking with the database, the compounds in the database were filtered by using pharmacophore searching. The 3D pharmacophore models were constructed based on the known important interactions between the amino acid in the binding pocket and NNRTIs. After filtering the database, the selected compounds were applied to docking. The hits from docking were selected and classified. The details of each step are shown as following sections. Finally, the selected compounds from the classification were tested for HIV-1 RT inhibition. The test for HIV-1 RT inhibition is performed by using ultrasensitive reverse transcriptase assay (as described in Appendix B) with wild-type, K103N, and Y181C HIV-1 RT at Pr615-Virus, Department of Microbiology, Faculty of Science, Mahidol University.

1.1 Starting Geometry

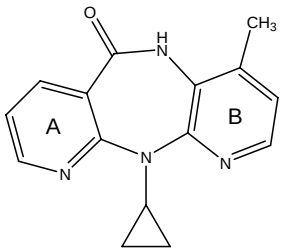
Preparation of protein coordinates

The X-ray structures of K103N and Y181C HIV-1 RT mutants in complex with nevirapine were obtained from the Protein Data Bank (pdb code 1fkp and 1jlb, Ren *et al.,* 2000, 2001). Nevirapine was removed from each structure. The complex structure of K103N HIV-1 RT with PNU-142721 was obtained from the PDB (code 1ikx, Lindberg *et al.,* 2002). PNU-142721 and water atoms were removed. All-hydrogen atoms were added to the proteins using standard SYBYL (Clark *et al.,*

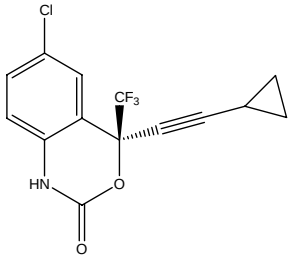
1989) geometries. The angle between the hydroxyl group of Tyr318 and the oxygen atom of His235 was adjusted to form a strong hydrogen bond in both protein mutants.

Preparation of NNRTIs coordinates

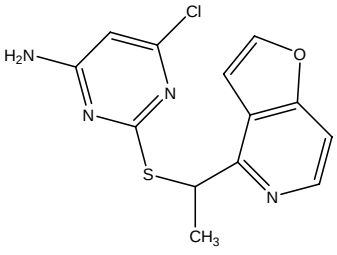
Starting geometry of ligands was also taken from the PDB. Non-nucleoside reverse transcriptase inhibitors (NNRTIs) of three complexes, nevirapine (pdb code 1fkp), efavirenz (pdb code 1fko, Ren *et al.,* 2000) and PNU-142721 (pdb code 1ikx), were studied for K103N mutation. Nevirapine (pdb code 1jlb), efavirenz (pdb code 1jkh, Ren *et al.,* 2001) and 8-Cl TIBO (pdb code 1uwb, Das et. al., 1996) were studied for Y181C mutation. Hydrogen atoms were added to each of ligand. All ligands were minimized by using the default parameter of Tripos force field in SYBYL 6.9 program. The structures of these NNRTIs are shown in Figure 13.



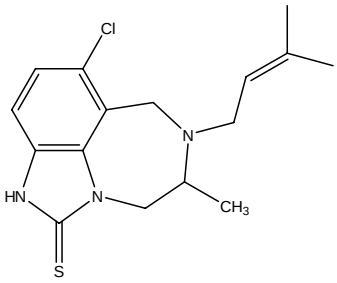
nevirapine



efavirenz



PNU-142721



8-Cl TIBO

Figure 13 Structures of NNRTIs.

Preparation of the screening compounds library

A database of about 501,750 drug-like compounds was obtained from 12 commercially-available screening collections (e.g. Asinex, Chemdiv, etc.) (Table 6) as already described by Bissantz *et al.*, 2000. From here on, this dataset will be referred to "Bioinfo" database.

Table 6 Contents of "Bioinfo" database.

Database	No. of compounds[a]
Asinex	121,344
Bionet	10,959
Biospecs	73,684
Chembridge	77,227
Chemdiv	46,577
Chemstar	7,039
Enamine	51,004
InterBioScreen	45,009
Maybridge	18,329
Timtec	1,762
Tripos	39,999
Vitasm	8,817

a) Number of compounds obtained after filtering by drug-likeness and removing duplicates.

1.2 Molecular Docking

Validation of the docking method

To determine the ability of docking method to reproduce the X-ray orientation of known ligands, three docking methods (FlexX, GOLD and Surflex) were applied to the nevirapine-RT complexes of K103N and Y181C mutations and

the PNU-142721-RT complex of K103N mutation. The root-mean-square deviation (rmsd) of the top-scored pose from the X-ray pose was used to determine the validation of the docking method. The detail of each docking method was described as following.

For FlexX, the default parameters of the FlexX program version 1.11.1 was used to dock nevirapine and PNU-142721 back to its binding pocket. The binding site is defined by using the amino acid residues located 6.5 Å apart from ligand. In case of GOLD, GOLD version 2.1 was used. The binding site was defined by using the center of mass of ligand and the radius of the binding site was set to 12 Å. For setting the GA parameters, the default parameters of library screening settings were used. The number of chromosome in each population and the number of operation were set to 50 and 1000, respectively. The GoldScore fitness function was used to determine the fitness score. For parameters controlling the extent of the protomol in Surflex, proto_thresh and proto_bloat were set to 0.4 and 0.0, respectively. Default parameters were used for matching ligand atoms with the previously-generated protomol.

Docking of NNRTIs and Database docking

The GOLD method has been used to dock previously-selected NNRTIs. The NNRTIs, nevirapine (pdb code 1fkp), efavirenz (pdb code 1fko) and PNU-142721 (pdb code 1ikx), were docked into the K103N protein (pdb code 1fkp) and the NNRTIs, nevirapine (pdb code 1jlb), efavirenz (pdb code 1jkh) and 8-Cl TIBO (pdb code 1uwb), were docked into the Y181C protein (pdb code 1jlb). PNU-142721 (pdb code 1ikx) was docked into the K103N protein (pdb code 1ikx). The conformations of docked NNRTIs were used to construct the pharmacophore models for 3D database searching. GOLD method was also performed with the compounds selecting from 3D database searching. From 3D database searching, the compounds matching K103N and Y181C pharmacophores were docked into the K103N and Y181C binding pockets, respectively.

1.3 Pharmacophore database searching

Pharmacophore-based database searching is an effective and fast method to discover the lead compounds from known binders (Sheridan *et al.*, 1989; Martin, 1992). In this study, 3D database searching using the UNITY4.4 program was previously applied to the "Bioinfo" database for filtering the compounds in the database. The UNITY program is designed to examine the database of compounds and return those compounds that match a specified query. The Directed Tweak algorithm (Hurst, 1994) in UNITY quickly finds molecules that could match the query. In 3D searching, the 3D queries can be based on molecules, molecular fragments or receptor sites. A set of molecular features and constraints can also be included to the query models. The molecular features can be atom centers, lines, planes, centroids, extension points, hydrogen bond sites and hydrophobic sites. Distance, angle, excluded volume, surface volume and spatial constraints define the geometric relationships between features. In this study, the conformations of docked NNRTIs were used to construct the pharmacophore models for 3D searching. The pharmacophore models were constructed based on the important interactions of NNRTIs in the binding pocket. The distance constraints were defined to the pharmacophore models. These models were used as constraints for 3D searching with the compound database of about 501,750 compounds by using the default parameters of 3D search in UNITY4.4 program.

1.4 Classification of the hits

Because of the massive output from virtual screening, it is necessary to find a method to organize or cluster the results. The ClassPharmer program (Bioreason Inc.) was used to analyze the hits and then organize them into classes based on a common scaffold. The first step was to classify the compounds into the homogeneous chemical families. After that, some statistical operations were used to prioritize the classes. In this study, two statistical analyses were applied to prioritize and select the compounds after classifying with two different procedures. They are the probability distribution (PrbDst) and the percent of compounds with a value for the

specified attribute (%Val). The PrbDst calculation provides the probability that the distribution of the attribute (here the GoldScore) for a class is different from the distribution of that attribute for all compounds in the classification. This calculation is useful to distinguish the class from noise. A value of 0 indicates that the distribution of values in the class is the same as the data set and that class contains the noise. A value of 1 indicates that the distribution of attribute values is trended either to higher or lower values. A PrbDst was not calculated for the class with less than 4 compounds. In the first procedure, classes were prioritized for selection based on a PrbDst value higher than 0.2, the attribute being here the GoldScore obtained by docking. Only compounds showing a GoldScore higher than 45 were analyzed herein. In a second procedure, classes presenting a minimal percent of compounds (here set to 20%) exhibiting a GoldScore value higher than 50 were selected in the final hitlist.

Both of the procedures were classified by using a medium homogeneity level, a medium redundancy level and the allowance of fuzziness ring closures. In the first procedure, the compounds having GoldScore higher than 45 were classified and the probability distribution of each class was calculated. Then, the classes having the probability distribution higher than 0.2 and the distribution of GoldScore trending to higher GoldScore were studied. The CompoundSelector module in ClassPharmer was used to select the lists of compounds based upon the compound having the highest GoldScore for each class. In the second procedure, the compounds having a GoldScore higher than 50 were classified. To compare with the first procedure, the classes with less than 4 compounds were removed. The compounds having GoldScore less than 50 were imported and classified into the earlier classified classes. The percent of compounds having the GoldScore higher than 50 in the class (%Val) was calculated. The classes with %Val higher than 20 were studied and the CompoundSelector module was used to select the lists of compounds based upon the compound having the highest GoldScore of the class. The common compounds from both procedures were further selected for biological evaluation.

2. CHAPTER II: Design of Nevirapine Derivatives Insensitive to the K103N and Y181C HIV-1 Reverse Transcriptase Mutants

Because nevirapine showed a lack of affinity upon two important mutations, the K103N and Y181C mutations, the aim of this work is to find novel nevirapine analogue insensitive to these mutations. Nevirapine derivatives were designed using a combinatorial library design approach (Krier *et al.,* 2005). Nevirapine derivatives were docked into the binding pocket of K103N and Y181C HIV-1 RT using GOLD program. The hits from docking were further performed post-docking procedure by topologically analyzing with the SILVER program. Furthermore, quantum chemical calculation was performed to calculate the interaction energy between nevirapine derivatives and mutant amino acid residue. The results will give more information on the interaction between nevirapine derivatives and N103 or C181.

2.1 Starting Geometry

Preparation of protein and nevirapine coordinates

The X-ray structures of K103N and Y181C HIV-1 RT mutants in complex with nevirapine were obtained from the Protein Data Bank (pdb code 1fkp and 1jlb). Nevirapine was removed from each structure. All-hydrogen atoms were added to the proteins using standard SYBYL geometries. For nevirapine of each structure, hydrogen atoms were added to them. Then, each nevirapine was minimized by using the default parameter of Tripos force field in SYBYL 6.9 program.

Preparation of nevirapine derivatives

From the nevirapine-RT complexes of K103N and Y181C protein mutants, some nevirapine derivatives were designed to find the compounds having the interaction with N103 and C181. Compounds were constructed using a combinatorial library design approach. Starting coordinates was taken from the minimized nevirapine. Substituents based on ten fragments were attached or replaced to

nevirapine by using SYBYL6.9 program. The fragments consist of the necessary functional groups for drug-like compound (Figure 14). These functional groups include diversity types of interaction, i.e. hydrophobic, hydrophilic, H-bond acceptor, H-bond donor, aliphatic and aromatic. To design the new compounds having the interaction with the side-chain of N103, some substituent groups (R1, R2 and R3) were designed by substituting or adding to nevirapine. In case of the designed compounds having the interaction with C181, R3 and R4 substituent groups were added or replaced to nevirapine. All designed compounds are shown in Table 7.

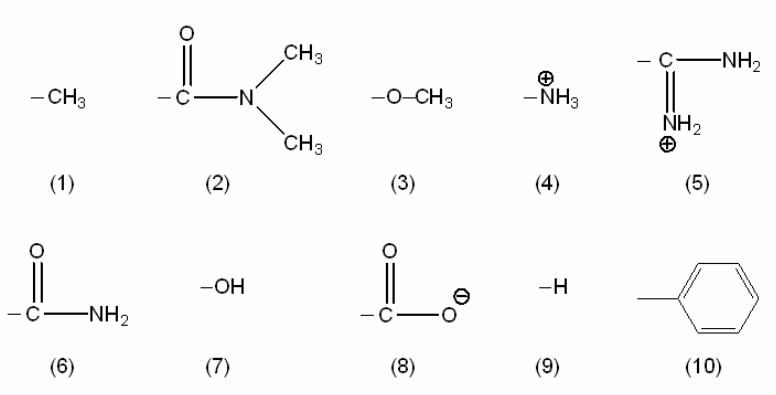


Figure 14 Structure of fragments, used in the design of nevirapine derivatives.

2.2 Docking of designed nevirapine derivatives

From Chapter I, the validation of three docking methods, FlexX, GOLD and Surflex, showed that GOLD method revealed a good ability to reproduce the X-ray bound conformation with rmsd less than 1.0 Å for both K103N and Y181C mutations. Therefore, the GOLD method was selected for docking designed nevirapine derivatives into K103N and Y181C HIV-1 RT. The docked compounds having score higher than that of nevirapine were selected and supposed to have tighter binding than nevirapine. Then, SILVER program was used to perform post-process docking results for large numbers of ligands. The hits from docking can be prioritized by the following topological analysis. Compounds presenting a significant percentage of the surfaces buried upon binding (>80%) and exhibiting hydrogen bonds to either N103 or C181 residues of the HIV-RT were selected.

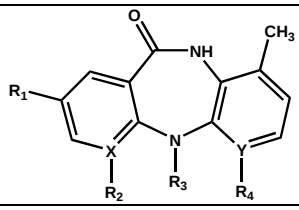
2.3 Quantum chemical calculations

To ensure that the selected compounds have attractive interactions with the mutated residues, interaction energies between selected compounds and N103 or C181 were calculated by using quantum chemical calculations. These were performed at B3LYP/6-31G(d) and MP2/6-31G(d) levels of theory by using Gaussian03 program and the results were compared with Nevirapine's results. The interaction energy (INT) of the selected compounds with N103 or C181 is defined as shown in equation (11).

$$INT = E_{(ligand+N103\ or\ C181)} - [\ E_{ligand} + E_{(N103\ or\ C181)}\ ] \qquad (11)$$

Where $E_{(ligand+N103\ or\ C181)}$ is the energy of the complex structure of ligand and N103 or C181. E_{ligand} and $E_{(N103\ or\ C181)}$ are the energies of ligand and N103 or C181, respectively. All energies were obtained from the single point calculation at B3LYP/6-31G(d) and MP2/6-31G(d) levels of theory. The structures of ligands were taken from their docked conformation in the binding pocket. The geometries of N103 and C181 were taken from K103N and Y181C mutants, respectively, which are starting geometries for docking.

Table 7 Structures of designed nevirapine derivatives, using a combinatorial library design approach.



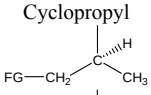
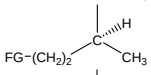
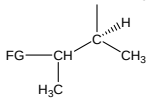
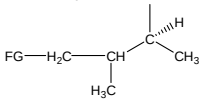
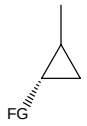
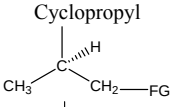
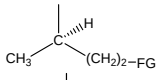
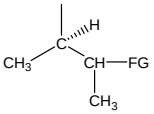
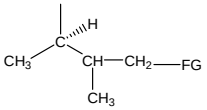
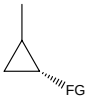
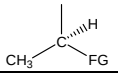
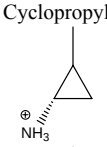
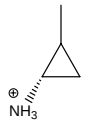
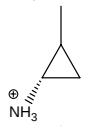
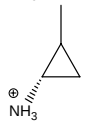
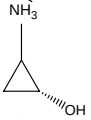
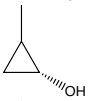
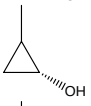
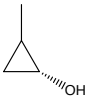
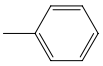
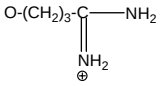
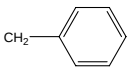
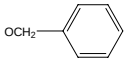
No.	X	Y	R_1	R_2	R_3	R_4
Str 1-10	C	N	H	O-CH_2-FG	Cyclopropyl	-
Str 11-20	C	N	H	O-$(CH_2)_2$-FG	Cyclopropyl	-
Str 21-30	C	N	H	O-$(CH_2)_3$-FG	Cyclopropyl	-
Str 31-40	C	N	H	FG	Cyclopropyl	-
Str 41-50	N	N	H	-		-
Str 51-60	N	N	H	-		-
Str 61-70	N	N	H	-		-
Str 71-80	N	N	H	-		-
Str 81-90	N	N	H	-		-
Str 91-100	N	N	H	-		-
Str 101-110	N	N	H	-	CH_2-FG	-
Str 111-120	N	N	H	-	$(CH_2)_2$-FG	-
Str 121-130	N	N	H	-		-
Str 131-140	N	C	H	-	Cyclopropyl	O-CH_2-FG
Str 141-150	N	C	H	-	Cyclopropyl	O-$(CH_2)_2$-FG
Str 151-160	N	C	H	-	Cyclopropyl	O-$(CH_2)_3$-FG
Str 161-170	N	C	H	-	Cyclopropyl	FG
Str 171-180	N	N	H	-		-
Str 181-190	N	N	H	-		-
Str 190-200	N	N	H	-		-
Str 201-210	N	N	H	-		-
Str 211-220	N	N	H	-		-
Str 221-230	N	N	H	-		-
Str 231-240	N	N	H	-		-

Table 7 (cont'd)

No.	X	Y	R_1	R_2	R_3	R_4
Str 241-250	N	N	FG	-	Cyclopropyl	-
Str 251	N	N	CH_2-NH_3^+	-	Cyclopropyl	-
Str 252	N	N	CH_2-OH	-	Cyclopropyl	-
Str 253	N	N	CH_2-COO^-	-	Cyclopropyl	-
Str 254-263	C	N	H	O-CH_2-FG		-
Str 264-273	C	N	H	O-$(CH_2)_2$-FG		-
Str 274-283	C	N	H	O-$(CH_2)_3$-FG		-
Str 284-293	C	N	H	FG		-
Str 294-303	C	N	H	O-CH_2-FG		-
Str 304-313	C	N	H	O-$(CH_2)_2$-FG		-
Str 314-323	C	N	H	O-$(CH_2)_3$-FG		-
Str 324-333	C	N	H	FG		-
Str 334-338	C	N	NH_3^+, OH, COO^-, $CH_2NH_3^+$, CH_2OH		Cyclopropyl	-
Str 339-343	C	N	NH_3^+, OH, COO^-, $CH_2NH_3^+$, CH_2OH	O-$(CH_2)_3$-COO^-	Cyclopropyl	-
Str 344-348	C	N	NH_3^+, OH, COO^-, $CH_2NH_3^+$, CH_2OH		Cyclopropyl	-
Str 349-353	N	N	NH_3^+, OH, COO^-, $CH_2NH_3^+$, CH_2OH	-		-
Str 354-358	C	N	NH_3^+, OH, COO^-, $CH_2NH_3^+$, CH_2OH		Cyclopropyl	-
Str 359-363	N	C	NH_3^+, OH, COO^-, $CH_2NH_3^+$, CH_2OH	-	Cyclopropyl