Chapter 2

Materials and Methods

GEMDOCK is a nearly automatic tool that was modified and enhanced from our original technique[16,21] for virtual screening (Figure 1). GEMDOCK can be sequentially applied to prepare target proteins and ligand databases, predict docked conformations and binding affinity using flexible ligand docking, and rank a series of candidates for post-docking analysis. The target protein is first prepared by specifying the atomic coordinates from the Protein Data Bank (PDB), the ligand binding area, atom formal charge (Table 1), and atom types (Table 2). When active ligands of the target protein are available, GEMDOCK evolves a pharmacological consensus (e.g., hot spots) and ligand preferences from the target protein and these ligands by overlapping the docked ligand conformations or superimposing X-ray structures. The pharmacological consensus and ligand preferences were incorporated into our scoring function to improve screening accuracy. The ligand database was constructed from the public compound databases, e.g., the MDL Drug Data Report (MDDR) or the Available Chemical Directory (ACD), according to the characteristics of the target protein and ligand preferences mined from known active compounds. After the ligand database and the target protein are prepared and the pharmacological preferences are evolved, GEMDOCK sequentially predicts the binding conformation and estimates the binding affinity for each ligand in the compound database. Finally, GEMDOCK ranks these docked ligand conformations for use in the post-docking analysis.

2.1 Scoring function

We developed a new scoring function that simultaneously serves as the scoring function for both molecular docking and the ranking of screened compounds for post-docking analysis. This function consists of a simple empirical binding score and a pharmacophore-based score to reduce the number of false positives. The energy function can be dissected into the following terms:

$$E_{tot} = E_{bind} + E_{pharma} + E_{ligpre} \quad (1)$$

where E_{bind} is the empirical binding energy, E_{pharma} is the energy of binding site pharmacophores (hot spots), and E_{ligpre} is a penalty value if a ligand does not satisfy the ligand preferences. E_{pharma} and E_{ligpre} are especially useful in selecting active compounds from hundreds of thousands of non-active compounds by excluding ligands that violate the characteristics of known active ligands, thereby improving the number of true positives. The values of E_{pharma} and E_{ligpre} are determined according to the pharmacological consensus derived from known active compounds and the target protein. In contrast, the values of E_{pharma} and E_{ligpre} are set to zero if active compounds are not available.

The empirical-binding energy (E_{bind}) is given as

$$E_{bind} = E_{inter} + E_{intra} + E_{penal} \qquad (2)$$

where E_{inter} and E_{intra} are the intermolecular and intramolecular energies, respectively, and E_{penal} is a large penalty value if the ligand is out of the range of the search box. For our present work, E_{penal} was set to 10,000. The intermolecular energy is defined as

$$E_{inter} = \sum_{i=1}^{lig} \sum_{j=1}^{pro} \left[F\left(r_{ij}^{B_{ij}}\right) + 332.0 \frac{q_i q_j}{4 r_{ij}^2} \right] \qquad (3)$$

where $r_{ij}^{B_{ij}}$ is the distance between atoms i and j with interaction type B_{ij} formed by pair-wise heavy atoms between ligands and proteins, B_{ij} is either a hydroge n bond or a steric state, q_i and q_j are the formal charges and 332.0 is a factor that converts the electrostatic energy into kilocalories per mole. The terms *lig* and *pro* denote the number of heavy atoms in the ligand and receptor, respectively. $F\left(r_{ij}^{B_{ij}}\right)$ is a simple atomic pair-wise potential function (Figure 2). In this atomic pair-wise model, the interactive types include only hydrogen bonding and steric potentials having the same function form but different parameters, $V_1, \ldots, V_6$. The energy value of hydrogen bonding should be larger than that for steric potential. In this model, atoms are divided into four different atom types [16]: donor, acceptor, both, and nonpolar. A hydrogen bond can be formed by the following pair-atom types: donor-acceptor (or acceptor-donor), donor-both (or both-donor), acceptor-both (or both-acceptor), and both-both. Other pair-atom combinations are used to form the steric state. We used the atom formal charge to calculate the electrostatic energy [16], which is set to 5 or –5, respectively, if the electrostatic energy is more than 5 or less than –5.

The intramolecular energy of a ligand is

$$E_{intra} = \sum_{i=1}^{lig} \sum_{j=i+2}^{lig} \left[F\left(r_{ij}^{B_{ij}}\right) + 332.0 \frac{q_i q_j}{4 r_{ij}^2} \right] + \sum_{k=1}^{dihed} A[1 - \cos(m q_k - q_0)] \qquad (4)$$

where $F\left(r_{ij}^{B_{ij}}\right)$ is defined as for Equation 3 except the value is set to 1000 when $r_{ij}^{B_{ij}}$ < 2.0 Å, and *dihed* is the number of rotatable bonds in a ligand. We followed the work [4] to set the values of *A*, *m*, and q_0. For the sp^3-sp^3 bond, *A* = 3.0, *m* = 3, and q_0 = p; for the sp^3-sp^2 bond, *A* = 1.5, *m* = 6, and q_0 = 0.

2.2 Mining Pharmacological Consensuses

GEMDOCK evolves the binding site pharmacological consensus and ligand preferences from both known active ligands and the target protein to improve screening accuracy. We used the premise that previously acquired interactions (hot spots) between ligands and the target protein can be used to guide the selection of lead compounds for subsequent investigation and refinement. For each known active ligand, GEMDOCK first yielded 10 docked ligand conformations by docking the ligand into the target protein, and only the ligand with the lowest docked conformation energy was retained for pharmacological consensus analysis. The protein-ligand interactions were extracted by overlapping these lowest-energy docked conformations, and the interactions were classified into three different types, including hydrogen bonding, hydrogen-charged interactions, and hydrophobic interactions. After all of the protein-ligand interactions were calculated, the atom interaction-profile weight of the target protein representing the pharmacological consensus of a particular interaction was given as

$$Q_j^k = \frac{f_j^k}{N} \qquad (5)$$

where f_j^k is the number of an atom *j* (in a protein) interacting with ligands with the interaction type *k*, and *N* is the number of known active ligands. In our present work, an atom *j* was considered a hot-spot atom when Q_j^k was more than 0.5.

The pharmacophore-based interaction energy (E_{pharma}) between the ligand and the protein is calculated by summing the binding energies of all hot-spot atoms:

$$E_{pharma} = \sum_{i=1}^{lig} \sum_{j=1}^{hs} CW(B_{ij}) F\left(r_{ij}^{B_{ij}}\right) \quad (6)$$

where $CW(B_{ij})$ is a pharmacological-weight function of a hot-spot atom j with interaction type B_{ij}, $F\left(r_{ij}^{B_{ij}}\right)$ is defined as in Equation 3, *lig* is the number of heavy atoms in a screened ligand, and *hs* is the number of hot-spot atoms in the protein. The $CW(B_{ij})$ is given as

$$CW(B_{ij}) = \begin{cases} 1.0 & \text{if } Q_j^k \le 0.5 \text{ or } B_{ij} \ne k \\ 1.5 + 5(Q_j^k - 0.5) & \text{if } Q_j^k > 0.5 \text{ and } B_{ij} = k \end{cases}. \quad (7)$$

Q_j^k is the atomic pharmacological-profile weight (Equation 5), and k is the interaction type (e.g., hydrogen bonding, hydrogen-charged interactions, or hydrophobic interactions) of the hot-spot atom j.

We evolved the ligand preferences (E_{ligpre}) from known ligands to reduce the deleterious effects of screening ligand structures that are rich in charged or polar atoms. Docking methods using energy-based scoring functions are often biased toward such compounds, which abound with charged and polar atoms (i.e., hydrogen donor or acceptor atoms) because the pair-atom potential of the electrostatic energy and hydrogen bonding energy is always larger than the steric energy. For our purpose, the atomic pair-wise potential energies of the electrostatic, hydrogen bond, and steric potential were set to –5, –2.5, and –0.4, respectively (Figure 2). If the binding site of a target protein is hydrophobic, the ligand preference (E_{ligpre}) is a penalty value for those screened ligands having many charged and polar atoms. The E_{ligpre} is given as

$$E_{ligpre} = LP_{elec} + LP_{hb} \quad (8)$$

where LP_{elec} and LP_{hb} are the penalties for the electrostatic (i.e., the number of charged atoms of a screened ligand) and hydrophilic (i.e., the fraction of polar atoms in a screened ligand) constraints, respectively. LP_{elec} is defined as

$$LP_{elec} = \begin{cases} 10NA_{elec} & \text{if } NA_{elec} > UB_{elec} \\ 0 & \text{if } NA_{elec} \le UB_{elec} \end{cases} \quad (9)$$

where $UB_{elec} = q_{elec} + s_{elec}$

, NA_{elec} is the number of charged atoms of a screened ligand and UB_{elec} is the upper bound number of charged atoms derived from known active compounds. $?_{elec}$ is the maximum number of charged atoms among known active compounds, and s_{elec} is the standard derivation of the charged atoms of known active compounds. LP_{hb} is defined as

$$LP_{hb} = \begin{cases} 5NA_{hb} & \text{if } r_{hb} > Ur_{hb} \\ 0 & \text{if } r_{hb} \le Ur_{hb} \end{cases} \qquad (10)$$

$$\text{where } r_{hb} = \frac{NA_{hb}}{NA_t} \text{ and } Ur_{hb} = q_{hb} + s_{hb}$$

, r_{hb} is the fraction of polar atoms (i.e., the atom type is both, donor, or acceptor) in a screened ligand and Ur_{hb} is the upper bound of the fraction of polar atoms calculated from known active ligands. NA_{hb} and NA_t are the number of polar atoms and the total number of the heavy atoms of a screened ligand, respectively. $?_{hb}$ and s_{hb} are the maximum ratio and the standard derivation of the ratios of polar atoms evolved from known ligands, respectively.

In order to reduce the deleterious effects of biasing toward the selection of high molecular weight compounds, we formulated a normalization strategy defined as

$$E_{tot} = \frac{E_{tot}}{(NA_t)^K} \text{ where } K = \begin{cases} 0.5 & \text{if } m_{mw} \le 15 \\ 0.5 - \dfrac{0.45(m_{mw} - 15)}{25} & \text{if } 15 < m_{mw} \le 40 \\ 0.05 & \text{if } m_{mw} > 40 \end{cases} \qquad (11)$$

where E_{tot} is the empirical binding energy (Equation 1), NA_t is the total number of the heavy atoms in a screened ligand, and μ_{mw} is the mean of the number of heavy atoms in known active compounds.

2.3 Flexible Docking Search Method

The search algorithm of GEMDOCK is a generic evolutionary method[16]. The core idea of our evolutionary approach was to design multiple operators that cooperate using the family competition model, which is similar to a local search procedure. The rotamer-based mutation operator, a discrete operator, is used to reduce the search space of ligand structure conformations. The Gaussian and Cauchy mutations, continuous genetic operators, efficiently search the orientation and conformation of the ligand relating to the center of the target protein. GEMDOCK randomly generates a starting population with N solutions by initializing the orientation and conformation of the ligand relating to the center of the receptor. Each solution is represented as a set of three n-dimensional vectors (x^i, s^i, $?^i$), where n is the number of adjustable variables of a docking system and i = 1, . . ., N where N is the population size. The vector x

represents the adjustable variables to be optimized in which x_1, x_2, and x_3 are the 3-dimensional location of the ligand; x_4, x_5, and x_6 are the rotational angles; and from x_7 to x_n are the twisting angles of the rotatable bonds inside the ligand. s and $?$ are the step-size vectors of decreasing-based Gaussian mutation and self-adaptive Cauchy mutation. In other words, each solution x is associated with some parameters for step-size control. The initial values of x_1, x_2, and x_3 are randomly chosen from the feasible box, and the others, from x_4 to x_n, are randomly chosen from 0 to $2p$ in radians. The initial step sizes s is 0.8 and $?$ is 0.2. After GEMDOCK initializes the solutions, it enters the main evolutionary loop which consists of two stages in every iteration: decreasing-based Gaussian mutation and self-adaptive Cauchy mutation. Each stage is realized by generating a new quasi-population (with N solutions) as the parent of the next stage. These stages apply a general procedure "FC_adaptive" with only different working population and the mutation operator.

The FC_adaptive procedure employs two parameters, namely, the working population (P, with N solutions) and mutation operator (M), to generate a new quasi-population. The main work of FC_adaptive is to produce offspring and then conduct the family competition. Each individual in the population sequentially becomes the "family father". With a probability p_c, this family father and another solution that is randomly chosen from the rest of the parent population are used as parents for a recombination operation. Then the new offspring or the family father (if the recombination is not conducted) is operated by the rotamer mutation or by differential evolution to generate a quasi offspring. Finally, the working mutation is operates on the quasi offspring to generate a new offspring. For each family father, such a procedure is repeated L times called the family competition length. Among these L offspring and the family father, only the one with the lowest scoring function value survives. Since we create L children from one "family father" and perform a selection, this is a family competition strategy. This method avoids the population prematureness but also keeps the spirit of local searches. Finally, the FC_adaptive procedure generates N solutions because it forces each solution of the working population to have one final offspring.

2.3.1 Recombination Operator

GEMDOCK implemented modified discrete recombination and intermediate recombination. A recombination operator selected the "family father (*a*)" and another solution (*b*) randomly selected from the working population. The former generates a child as follows:

$$x_j^c = \begin{cases} x_j^a \text{ with probabilit y } 0.8 \\ x_j^b \text{ with probabilit y } 0.2 \end{cases}$$

The generated child inherits genes from the "family father" with a higher probability 0.8. Intermediate recombination works as:

$$w_j^c = w_j^a + b\left(w_j^b - w_j^a\right)/2$$

where *w* is *s* or *?* based on the mutation operator applied in the FC_adaptive procedure. The intermediate recombination only operated on step-size vectors and the modified discrete recombination was used for adjustable vectors (*x*).

2.3.2 Mutation Operators

After the recombination, a mutation operator, the main operator of GEMDOCK, is applied to mutate adjustable variables (*x*).

Gaussian and Cauchy Mutations are accomplished by first mutating the step size (*w*) and then mutating the adjustable variable *x*:

$$w_j^{'} = w_j^{'} A(\cdot)$$

$$x_j^{'} = x_j + w_j^{'} D(\cdot)$$

where w_j and x_j are the *i*th component of *w* and *x*, respectively, and w_j is the respective step size of the x_j where *w* is *s* or *?*. *A*(.) is evaluated as exp[*t'N*(0, 1) +N_j(0, 1)] if the mutation is a self-adaptive mutation, where *N*(0, 1) is the standard normal distribution, N_j(0, 1) is a new value with distribution *N*(0, 1) that must be regenerated for each index *j*. When the mutation is a decreasing-based mutation *A*(.) is defined as a fixed decreasing rate ? = 0.95. *D*(.) is evaluated as *N*(0, 1) or *C*(1) if the mutation is, respectively, Gaussian mutation or Cauchy mutation. For example, the self-adaptive Cauchy mutation is defined as

$$y_j^c = y_j^a \exp\left[t^{'} N(0,1) + tN_j(0,1)\right]$$

$$x_j^c = x_j^a + y_j^c C_j(t)$$

We set t and t' to $\left(\sqrt{2n}\right)^{-1}$ and $\left(\sqrt{2\sqrt{2n}}\right)^{-1}$, respectively, according to the suggestion of evolution strategies. A random variable is said to have the Cauchy distribution ($C(t)$) if it has the density function: $f(y; t) = \frac{t/p}{t^2 + y^2}$, $-\infty < y < \infty$. In this paper t is set to 1. Our decreasing-based Gaussian mutation uses the step-size vector s with a fixed decreasing rate ? = 0.95 and works as

$$s^{c} = g s^{a}$$

$$x_j^c = x_j^a + s^{c} N_j(0,1)$$