Structure Based Activity Prediction of HIV-1 Reverse Transcriptase Inhibitors

Marc R. de Jonge,* Lucien M. H. Koymans, H. Maarten Vinkers, Frits F. D. Daeyaert, Jan Heeres, Paul J. Lewi, and Paul A. J. Janssen

Center for Molecular Design, Johnson & Johnson Pharmaceutical Research and Development, Janssen Pharmaceutica NV, Antwerpsesteenweg 37, B-2350 Vosselaar, Belgium

Received June 16, 2004

We have developed a fast and robust computational method for prediction of antiviral activity in automated de novo design of HIV-1 reverse transcriptase inhibitors. This is a structure-based approach that uses a linear relation between activity and interaction energy with discrete orientation sampling and with localized interaction energy terms. The localization allows for the analysis of mutations of the protein target and for the separation of inhibition and a specific binding to the enzyme. We apply the method to the prediction of pIC_{50} of HIV-1 reverse transcriptase inhibitors. The model predicts the activity of an arbitrary compound with a q^2 of 0.681 and an average absolute error of 0.66 log value, and it is fast enough to be used in high-throughput computational applications.

Introduction

Reverse transcriptase (RT) is a key enzyme for the causative agent of AIDS, the human immunodeficiency virus (HIV). Inhibition of the function of this enzyme by means of a modified nucleoside substrate, the nucleoside reverse transcriptase inhibitor (NRTI) AZT, was the basis for the first AIDS therapy.[1,2] Later it was found that the function of RT could also be blocked by noncompetitive inhibitors that show no chemical resemblance to nucleosides, the non-nucleoside reverse transcriptase inhibitors (NNRTI) like TIBO[3] and nevirapine.[4] This second mode of inhibition is attractive for drug design because, unlike for the nucleoside inhibitors, little interference with human systems and, therefore, low toxicity and side effects were expected. It turned out, however, that a large number of errors occurs during the viral replication,[5] resulting in a high mutation rate of RT. Patients treated with current NRTIs[6] or NNRTIs[7] in monotherapeutic regimens develop resistance to treatment, because of the emergence of these mutated strains.

Therefore there is an interest in new compounds that are highly potent and less susceptible to mutations. In our design effort for such compounds we need a computational approach that can distinguish inhibitors with such properties from less useful compounds. As we intend to assess the effects of different resistance inducing mutations in the future, a target directed approach is the most obvious choice because of its explicit description of such mutations. However, many target directed methods, like for instance the variants of free energy perturbation[8] and molecular dynamics,[9] have the disadvantage that they are computationally very expensive. In our automated drug design method[10] we need to overcome this problem. A faster computational approximation was developed that can still give a quantitative binding strength. The method is similar in nature to the linear interaction energy method,[11] but it focuses only on the most common possible geometries of the inhibitor−enzyme complex, rather than using a full dynamics simulation. The method provides an estimate of inhibitory strength toward the virus (pIC_{50}) for an arbitrary compound, and a qualitative indication of potential resistance inducing mutations.

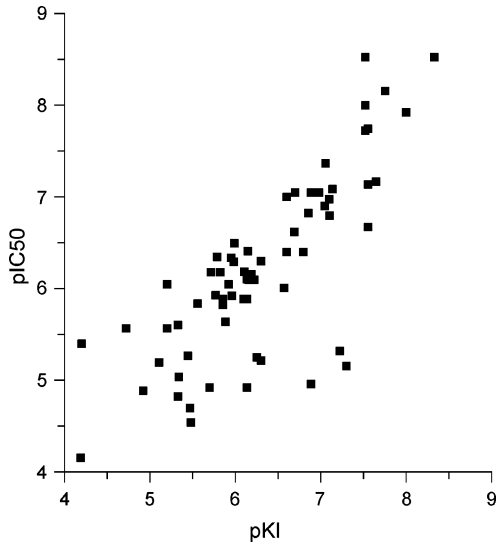

Figure 1. pIC_{50} vs pk_i of the 68 NNRTI records in the NIAID/OI Therapeutics Database where both observations are given, $r^2 = 0.624$.

The method generates a number of low energy inhibitor−enzyme complexes by means of conformational searching and docking. Subsequently a statistical model is derived that relates inhibition of viral replication with the localized nonbonded interaction energy contributions between inhibitor and enzyme residues. Localizing the interaction energy allows the method to discriminate between strong binding on an arbitrary part of the enzyme, strong inhibition of the bare enzyme, and possibly strong inhibition of the viral replication. The latter two are not necessarily the same as strong binding as is illustrated for NNRTIs in the NIAID database[12] in Figure 1. This figure shows that k_i and pIC_{50} do not

* Corresponding author currently at MolMo Services BVBA. E-mail: marc@molmo.be. Phone: +31135717960. Fax: +31135717961.

10.1021/jm049534r CCC: $30.25
Published on Web 01/15/2005

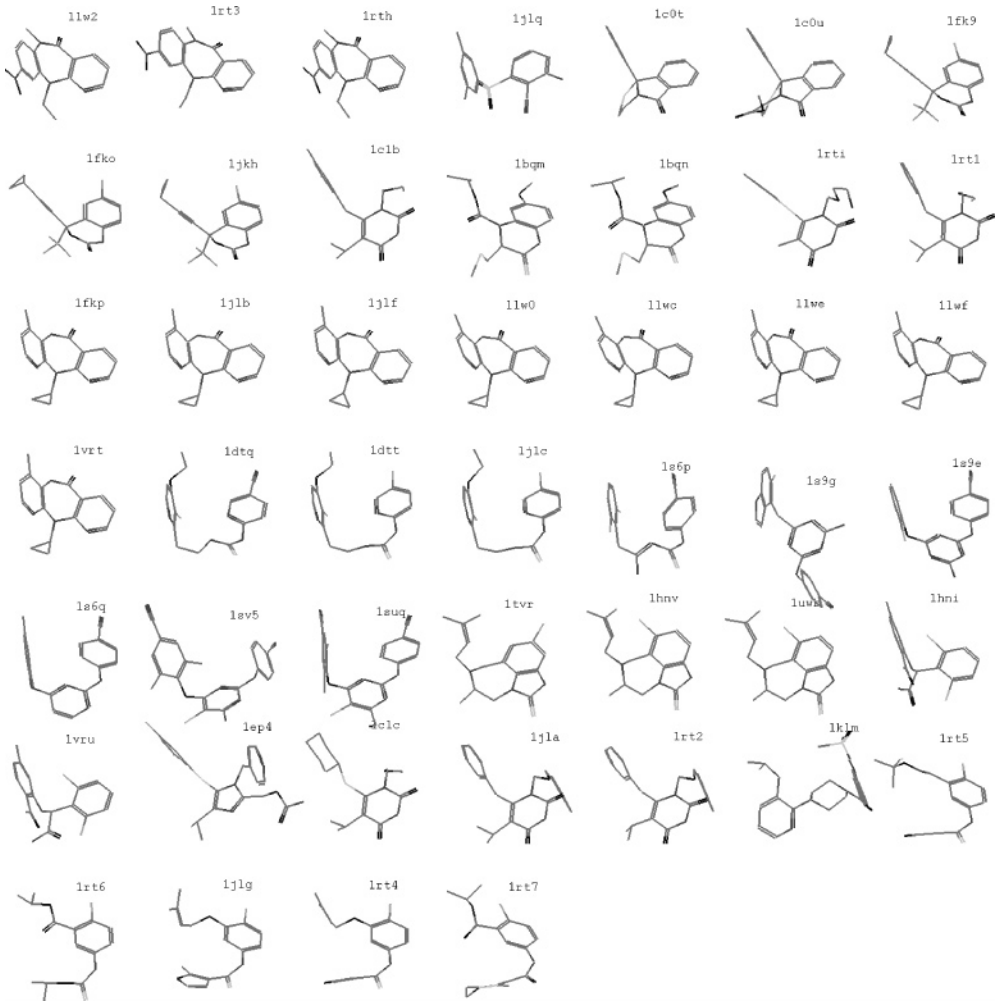

Figure 2. Test set for conformation search and docking: 46 NNRTIs with the conformation and orientation with which they are cocrystallized in HIV-1 RT.

correlate perfectly, even though enzyme and mechanism of action are identical for all data points.

Materials and Methods

In order to evaluate the interaction energy between RT and a newly designed NNRTI we first need to generate the relevant 3D structures of the complex. This is done in separate steps: (1) generate low energy conformers of the inhibitors; (2) dock the conformers in the enzyme; (3) relax the resulting complexes.

For each of these steps it is required that the method can progress without user intervention to facilitate automatic drug design and optimization strategies. It is also important that the computations can be applied to compounds that may not belong to existing series. This means that manual docking and template based activity computations such as those given by Rizzo et al.[13] cannot be used.

Low energy conformers are generated by setting the free rotatable bonds of the compounds at 60° intervals combined with inversions of chiral atoms, and local optimization with a MMFF94[14,15]-based force-field parametrization. All combinations of up to 7 degrees of freedom are generated exhaustively. If a compound has more degrees of freedom, the 7 with the largest effect on the geometry (identified by their rms difference resulting from either an inversion or a small torsional variation) are evaluated exhaustively while the remainder of conformational space is scanned nonexhaustively by means of a genetic algorithm (GA) with real-valued coefficients for torsion angles and a single bit chirality. The GA will try up to 10000 attempts to improve one of the best 100 conformers found at each instant.

This procedure results in a list of up to 1000 low energy conformers that differ at least 0.25 Å in rmsd. In order to increase computational speed we will not consider all conformers for the subsequent stage; instead we will test how many conformers, and what level of internal energy, are needed to find the geometry of the X-ray structure, on the set of well-known NNRTIs cocrystallized with RT available from the Protein Data Bank[16] shown in Figure 2. The molecules in this figure are plotted with the conformation of the X-ray structure and are all placed in a comparable orientation, by superimposing the surrounding peptide C-α atoms.

The second stage of the modeling takes a list of low energy conformers and docks them into the non-nucleoside binding pocket of wild-type RT. The docking target consists of a 25 Å region around the NNRTI binding pocket, for which the coordinates of pdb entry 1s6q[17] were used. The binding region of this structure was supplemented with hydrogen atoms, and the geometry was relaxed by means of a force-field minimization of all atoms except the α-carbons. In order to allow rapid docking of large numbers of compounds, an interaction grid method[18] that contains contributions of Coulomb and van der Waals energies was employed. The grid-based energy was supplemented with a separate hydrogen-bonding term. The grid dimensions are 200 × 200 × 200 cells at 0.25 Å spacing. The protein part of the Coulomb interaction is precomputed for each grid point by adding all the point-charge contributions. For the van der Waals energy two grids are generated; first a soft potential function with $1/r^4$ attraction and $1/r^8$ repulsion terms for an initial Monte Carlo steric scan, and second a standard Lennard-Jones function with $1/r^6$ attraction and $1/r^{12}$ repulsion for the subsequent simulated annealing optimiza-

tion. The parameters for both potentials were chosen to produce an atom-type based geometric mean function that replicates the potential-well depth and position of our MMFF94[14,15] based force field. The docking hydrogen-bond terms are implemented by identifying "sweet spots" in the protein structure, i.e., locations where the interaction energy can improve by matching a donated hydrogen to an acceptor atom. This is achieved by adding a negative valued Gaussian function on these points that approximates the extra energy contribution for a well-formed hydrogen donor−acceptor pair.

The performance of the docking algorithm is tested by docking the set of ligands used in the test of the conformational search into the NNRTI binding pocket grid. Note that we will use one target structure for cross-docking all these ligands, not their respective native crystal structures. Being able to dock in a structure that is not cocrystallized with the ligand is necessary because we intend to use the method for the design of novel compounds, where evidently no crystallographic data is available. Because of the focus on new compounds, for which we do not know a priori which (if any) pharmacophore is present, we have chosen not to use the explicit pharmacophore model for docking that we have presented previously,[19] and that we have used in lead optimization, but rather a method that performs an unbiased rigid multiple conformation interaction energy optimization by means of a simulated annealing.[20]

The third part of the procedure, the relaxation of the ligand-binding site complex, is needed to derive a more accurate energy-based scoring: The docked molecules are still in their local optimum structure of the conformation search, as is the target enzyme. It is necessary to relax the geometry, and to allow the two systems to adapt to each other into a geometry that more closely resembles one of the low energy states of a dynamics simulation. This is performed by means of a local geometry optimization[21] of all complexes where the total energy of interaction after docking is sufficiently negative. The method allows any number of complexes to be generated, but we expect that the contribution to the measured activity of the compound is dominated by the lowest energy states, as the Boltzmann factor for an energy difference of only 3 kcal/mol between two conformers already results in a less than 1% occupancy of the higher level. For this reason we will limit the number of complexes we consider for the energy prediction to those that are within the relatively narrow energy range of 6 kcal/mol from the lowest, and we will generate the final pIC_{50} prediction in the design application based on the combined contribution of the complexes within this energy range.

The intended use of the modeling procedure as a high volume design tool has important implications: It means we are less interested in avoiding false negatives than normal docking-screening strategies (as we are not mining databases for existing potential lead compounds) but we need a very robust function with respect to false positives (as we do synthesize the promising compounds we have designed). We will evaluate the robustness of the method in this respect by applying it to all 4387 compounds that show a quantifiable activity in in vitro antiviral testing,[22] encompassing all published NNRTIs, and a large number of unpublished variants synthesized in-house since 1989.

Results: Performance of the Conformation Search

After isolating the inhibitors from the pdb entries, adding hydrogen atoms (all ligands are considered neutral), and local optimization of the geometry we obtain a list of conformers of each NNRTI. An important condition for a useful conformation search is that it has to find the locally optimized conformer isolated from the X-ray structure of the bound inhibitor. We tested this aspect by randomizing the molecule geometries and searching for all low energy conformers. The results are shown in Table 1: We find that all initial NNRTI structures were reproduced independently in this manner (these are characterized by their less than 0.01 Å difference with the locally optimized X-ray geometry). Furthermore, all these structures are within 6 kcal/mol of internal strain energy relative to the lowest energy conformer. In 18 out of 46 cases the ligand conformer taken from the crystal structure was the also the computed global minimum of the compound. Columns 4, ΔE_{cryst}, and 5, $rank_{cryst}$, of Table 1 show the energy difference between the optimized X-ray geometry and the lowest energy conformer, and the rank of the X-ray geometry in the list of conformers. However, for a number of flexible compounds, which have a large number of energetically close conformers, the closest geometry to the X-ray structure has a very high rank in the list of conformers, despite its only moderately worse internal energy. In these cases it turns out that a very similar conformer, where the rms difference with the crystal is less than 1 Å, is often found at a lower energy level. Columns 6 and 7 of Table 1 show the internal energy and rank of these similar conformers. By using this more relaxed criterion we find 27 out of 46 compounds within 0.1 kcal/mol of the global optimum, and all conformers but one, within 2 kcal/mol. The two most noteworthy exceptions are the structures from pdb entry 1rt7 (UC-84), which seems to be in a strained conformation, and 1s9g (R120394), which, although not highly strained, seems to be in its highest energy minimum.

As we will relax the geometry after docking we will make use of the presence of these conformers close to the complexed optimum and reduce the computation time by only docking a limited subset of low energy forms. Especially for the de novo method[10] this poses no problem, as the algorithm will design another optimum structure as long as the fitness function provides enough evolutionary pressure. For benchmarking and database screening, the number of conformers to evaluate can be increased to attain a required level of exclusion of false negative, at the cost of extra computational effort, the worst case in our test set being entry 1rt7, where the first conformer closely resembling the crystal has rank number 135. Run times of the conformation search depend strongly on the number of degrees of freedom and range from 2 s to 150 min for the full search on a 500MHZ MIPS R14000 CPU for the current dataset.

Results: Performance of the Docking Algorithm

The second part of our modeling approach consists of docking the designed ligand into the target grid. The same set of NNRTIs that is used for the conformation search is used as a test set for the docking algorithm. For each X-ray structure a corresponding region around the NNRTI-binding site is isolated and optimized. This optimized structure is used as a reference for the quality of the docking: The heavy atoms of the protein backbone atoms of the optimized X-ray structure are superimposed on those of the docked and optimized complex, and the rms difference between the ligand heavy atoms is computed for the superimposed complexes.

The results are shown in Table 2. The first two columns again show inhibitor name and pdb code; column 3, $E_{dock}(min)$, shows the lowest docking energy

Table 1. Conformational Search of 46 NNRTI Crystal Structures from the Protein Data Bank[a]

compound	pdb	rms_{cryst} (Å)	ΔE_{cryst} (kcal/mol) rms = 0	$rank_{cryst}$ rms = 0	$\Delta E_{similar}$ (kcal/mol) rms < 1	$rank_{similar}$ rms < 1	N_{conf}
1051U91	1lw2	0.90	1.32	2	0.00	1	3
1051U91	1rt3	0.91	1.32	2	0.00	1	3
1051U91	1rth	0.00	0.00	1	0.00	1	3
739W94	1jlq	1.84	0.00	2	0.00	2	4
BM+21.1326	1c0t	1.26	0.00	2	0.02	2	2
BM+50.0934	1c0u	0.00	0.00	1	0.00	1	3
DMP-266	1fk9	0.00	0.00	1	0.00	1	2
DMP-266	1fko	0.00	0.00	1	0.00	1	2
DMP-266	1jkh	0.00	0.00	1	0.00	1	2
GCA-186	1c1b	2.13	3.20	37	0.59	3	101
HBY097	1bqm	1.41	5.53	77	0.23	7	84
HBY097	1bqn	1.27	3.75	65	0.23	7	88
HEPT	1rti	1.92	0.65	7	0.65	7	144
MKC-442	1rt1	1.98	3.65	42	0.36	3	93
nevirapine	1fkp	0.00	0.00	1	0.00	1	1
nevirapine	1jlb	0.00	0.00	1	0.00	1	1
nevirapine	1jlf	0.00	0.00	1	0.00	1	1
nevirapine	1lw0	0.00	0.00	1	0.00	1	1
nevirapine	1lwc	0.00	0.00	1	0.00	1	1
nevirapine	1lwe	0.00	0.00	1	0.00	1	1
nevirapine	1lwf	0.00	0.00	1	0.00	1	1
nevirapine	1vrt	0.00	0.00	1	0.00	1	1
PETT-1	1dtq	1.31	5.94	152	0.72	4	459
PETT-2	1dtt	1.38	0.83	4	0.83	4	436
PETT-2	1jlc	1.38	0.83	5	0.83	5	458
R100943	1s6p	1.22	0.42	5	0.00	2	35
R120394	1s9g	2.19	0.71	13	0.71	13	13
R129385	1s9e	0.00	0.04	1	0.00	1	5
R147681	1s6q	0.00	0.00	1	0.00	1	10
R165335	1sv5	0.00	0.00	1	0.00	1	6
R185545	1suq	0.31	0.13	3	0.00	1	21
R82913	1tvr	0.00	0.00	1	0.00	1	13
R86183	1hnv	1.30	2.02	6	0.06	2	12
R86183	1uwb	1.12	2.02	7	0.06	2	14
R95845	1hni	0.90	0.19	3	0.00	1	54
R95845	1vru	0.76	0.08	2	0.08	2	48
S-1153	1ep4	2.14	2.74	76	1.62	31	196
TNK-6123	1c1c	1.54	4.75	72	0.53	5	143
TNK-651	1jla	3.00	3.53	59	0.96	16	126
TNK-651	1rt2	2.99	0.96	17	0.96	17	133
U-90152	1klm	2.72	4.36	175	1.45	43	280
UC-10	1rt5	2.13	1.06	37	1.06	37	114
UC-38	1rt6	1.10	0.55	8	0.22	4	188
UC-781	1jlg	2.44	2.08	108	0.82	25	167
UC-781	1rt4	2.48	2.08	110	1.01	34	178
UC-84	1rt7	2.60	5.86	142	5.59	135	249

[a] Column 1 shows the compound name, and column 2 shows the originating pdb entry; rms_{cryst} is the rms difference between the crystal geometry and the best conformer, ΔE_{cryst} is the internal energy of the crystal geometry minus the internal energy of the best conformer, $rank_{cryst}$ is the rank order of the geometry of the crystal structure in the list of computed conformers, $\Delta E_{similar}$ is the internal energy of the first conformer within 1 Å of the crystal structure, $rank_{similar}$ is its rank in the conformation list, and N_{conf} is the total number of conformers found within 7 kcal/mol.

found; and column 4, rms E_{dock}(min), shows the rms difference between the ligand atoms after the protein parts of the docked complex and original pdb structure are superimposed. The fifth column shows the rank number of the docking that most closely resembles the X-ray structure (with rms best in column 6 being the rms difference of the inhibitors, rms BB in column 7 being the difference of the peptide backbone atoms, and E_{dock} best in column 8 being the accompanying docking energy). The table shows that 31 out of 46 X-ray structures are reproduced to within the best resolution quoted in the pdb-data file; we consider these to be correctly docked. The 15 rows where this is not the case are in boldface type in the table. We allow quite a large difference between docked and crystal structure because both structures are optimized geometries starting from different X-ray structures that differ substantially in geometry, even after relaxation (column7). Figure 3 illustrates a common trait of the poorly docked compounds: Their higher rms deviation is accompanied by a systematically higher docking energy than the correctly docked compounds. This means that we can apply a "low pass" filter to the docking result for novel compounds by only accepting molecules docked below a certain threshold energy, thus preferentially weeding out badly docked ones and thereby reducing the false positive rate of the prediction.

Another noteworthy feature of the docking results is the fact that there may be a correlation with protein mutation and our ability to predict the correct ligand orientation (which is based on only wild-type geometry). It appears that mutations of residues M184 and Y188 change the geometry of the binding site so much that our wild-type model no longer can be applied, while for the K103, Y181, and K219 mutations this effect seems to be so much less pronounced that we can successfully

Table 2. Cross-Docking of 46 Locally Optimized NNRTIs from the Protein Data Bank into the Generalized Wild-Type NNRTI Binding Site[a]

compound	pdb	E_{dock}(min) (kcal/mol)	rms E_{dock}(min) (Å)	rank best match	rms best match (Å)	rms BB (Å)	E_{dock} best match (kcal/mol)	mutations
1051U91	1lw2	−29.7	0.49	1	0.49	1.73	−29.7	
1051U91	1rt3	−29.7	5.62	5	1.77	2.09	−26.5	K219Q
1051U91	1rth	−30.3	1.87	4	1.69	1.79	−30.2	
739W94	1jlq	−29.6	1.11	1	1.11	1.87	−29.6	
BM+21.1326	1c0t	−27.8	1.31	3	1.03	1.83	−27.2	
BM+50.0934	1c0u	−28.4	1.86	1	1.86	1.99	−28.4	
DMP-266	1fk9	−30.2	0.93	1	0.93	1.78	−30.2	
DMP-266	1fko	−30.6	1.38	1	1.38	2.17	−30.6	K103N
DMP-266	1jkh	−30.1	0.89	4	0.89	2.11	−26.9	Y181C
GCA-186	**1c1b**	**−24.9**	**7.59**	**10**	**5.80**	**2.10**	**−18.8**	
HBY097	1bqm	−21.4	7.13	2	0.90	1.48	−21.1	
HBY097	**1bqn**	**−21.8**	**5.85**	**7**	**3.38**	**1.67**	**−16.5**	**Y188L**
HEPT	1rti	−27.0	6.33	8	2.97	2.01	−26.4	
MKC-442	**1rt1**	**−24.8**	**6.11**	**6**	**4.78**	**2.13**	**−19.5**	
nevirapine	**1fkp**	**−30.1**	**5.83**	**9**	**4.89**	**2.04**	**−24.4**	**K103N**
nevirapine	1jlb	−28.3	3.87	2	1.36	1.89	−27.9	Y181C
nevirapine	**1jlf**	**−29.3**	**6.67**	**3**	**4.89**	**2.15**	**−27.5**	**Y188C**
nevirapine	1lw0	−30.0	5.81	9	1.57	1.83	−23.4	
nevirapine	**1lwc**	**−29.9**	**5.79**	**10**	**3.83**	**1.89**	**−23.7**	**M184V**
nevirapine	1lwe	−27.5	1.37	1	1.37	1.96	−27.5	
nevirapine	**1lwf**	**−30.1**	**5.87**	**6**	**3.53**	**2.08**	**−27.2**	**M184V**
nevirapine	**1vrt**	**−30.0**	**5.76**	**6**	**3.80**	**1.86**	**−27.1**	
PETT-1	**1dtq**	**−31.2**	**6.10**	**6**	**3.57**	**1.89**	**−25.8**	
PETT-2	1dtt	−35.1	1.07	1	1.07	1.94	−35.1	
PETT-2	1jlc	−36.3	1.19	2	1.09	1.95	−36.2	Y181C
R100943	1s6p	−32.2	0.82	2	0.80	1.62	−32.0	
R120394	**1s9g**	**−28.9**	**3.45**	**1**	**3.45**	**2.08**	**−28.9**	
R129385	1s9e	−37.3	1.00	3	0.91	1.74	−37.1	
R147681	1s6q	−38.5	0.66	1	0.66	0.67	−38.5	
R165335	1sv5	−39.1	1.04	2	0.99	1.56	−39.0	
R185545	1suq	−39.5	0.58	1	0.58	1.71	−39.5	
R82913	1tvr	−31.5	1.61	2	1.57	1.30	−31.2	
R86183	1hnv	−33.2	2.57	6	2.44	1.29	−33.1	
R86183	1uwb	−33.2	2.17	3	1.98	1.31	−33.1	Y181C
R95845	**1hni**	**−28.7**	**5.95**	**7**	**5.30**	**1.51**	**−23.0**	
R95845	1vru	−28.2	7.01	4	1.10	1.80	−26.0	
S-1153	**1ep4**	**−21.4**	**7.94**	**3**	**4.20**	**2.37**	**−5.9**	
TNK-6123	1c1c	−34.8	1.33	1	1.33	1.99	−34.8	
TNK-651	**1jla**	**−24.5**	**6.56**	**1**	**6.56**	**2.22**	**−24.5**	**Y181C**
TNK-651	**1rt2**	**−24.2**	**8.05**	**2**	**6.57**	**2.21**	**−24.2**	
U-90152	**1klm**	**−22.5**	**12.20**	**2**	**10.81**	**2.80**	**−15.4**	
UC-10	1rt5	−36.7	1.19	2	1.08	1.97	−36.7	
UC-38	1rt6	−27.0	5.29	2	1.39	1.98	−26.4	
UC-781	1jlg	−29.6	1.11	1	1.11	2.01	−29.6	
UC-781	1rt4	−31.8	1.58	1	1.58	1.89	−31.8	
UC-84	1rt7	−29.0	1.58	3	1.37	1.96	−28.9	

[a] The first column indicates compound name, and the second column indicates the pdb entry code; E_{dock}(min) is the lowest (grid-based) nonbonded interaction energy, rms E_{dock}(min) is the rms difference between this docked conformer and crystal structure (after local optimization of docked ligand in complex with the wild-type target and with its original crystal), rank best is the rank of the docked orientation that is most similar to the ligand's crystal structure, rms best is the rms difference between this docking and the crystal, rms BB is the difference between the peptide backbone atoms (C-α, N, O, C) of the optimized X-ray structure and the optimized pocket after docking, E_{dock} best is the nonbonded interaction energy of the best docked ligand, and mutations indicates which amino acids in the crystal structure (within the subset of residues used for the docking grid) are different from the wild-type structure used for the docking. Table rows in boldface type denote the incorrectly docked orientations, which differ more from the crystal structure than the best X-ray resolution given for the pdb entry.

match the ligand position docked in the wild-type pocket to that in the mutant X-ray structures.

Results and Conclusion

We will use the complex geometries generated by combining conformation search and docking to predict the pIC_{50} of an arbitrary compound. As we want to be able to deal with any possible structure, we need as wide a range of structures in the derivation of the model as possible. To this end we took 4387 compounds with experimentally determined anti-HIV activity from our database (comprising protease inhibitors, NNRTIs, NRTIs, and others) and produced the docked complexes of these compounds in our HIV-1 RT model. As is to be expected, a number of compounds could not be docked at all ($E_{dock} > 0.0$) because they are not NNRTIs; this is especially obvious in the case of very large HIV protease inhibitors (about 500 out of the 1500 nondocked molecules). Also a fair amount of complexes showed indications of suspected poorly docked orientations ($E_{dock} > -20$). Such a high docking energy is a negative signal for the quality of docking in the case of compounds that have appreciable NNRT inhibitory activity (appreciable enough to warrant the cocrystallization of the compound

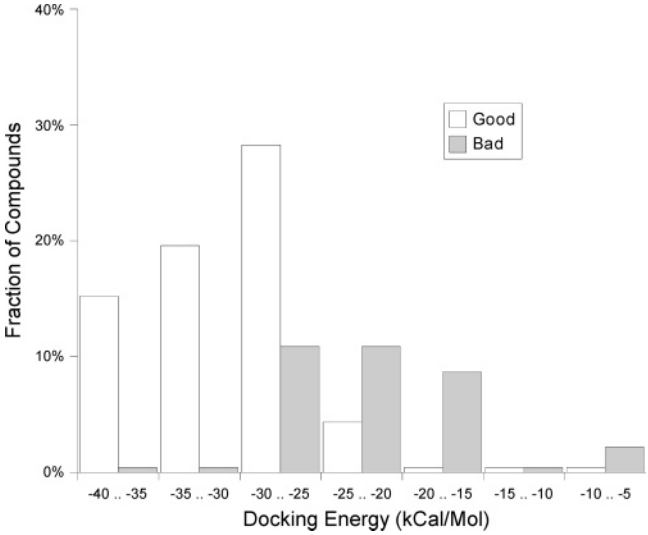

Figure 3. Distribution of docking energies, in white the correctly docked NNRTIs, in gray the incorrectly docked ones (see also Table 2).

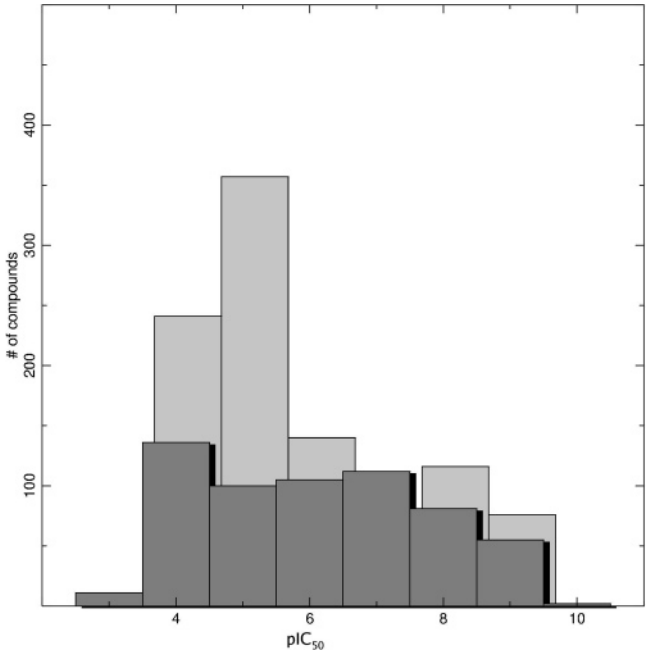

Figure 4. Number of compounds in the training set that allow self-consistent models in light gray, and number of compounds that do not always allow this in darker gray, both binned by their measured antiviral activity.

with RT), but we need to allow for the possibility that marginal inhibitors, that were never cocrystallized with RT, actually bind in such energetically less favorable orientations.

We will filter out complexes that are incorrectly or unfavorably oriented by means of a self-consistent analysis approach. One hundred subsets of approximately 200 compounds are randomly generated by picking 30 compounds in 6 activity bins between pIC_{50} of 4 and 10 from the total dataset. For every subset we generate a linear fit of interaction energy contributions to the observed data. This fit is generated by means of a genetic algorithm that selects specific interaction energy contributions from the total matrix of ligand−protein residue interactions. The fitness function that is minimized by the GA is the average absolute deviation, which is a more robust estimator than the average squared deviation in data that may have large outliers.[23] The gene of a population member consists of a fixed length list of binary values, indicating either "on" or "off" state of one particular interaction term. This approach differs from the stepwise feature variable selection methods in that the genetic algorithm is able test a large number of different patterns of variables simultaneously, and can create crossovers between different

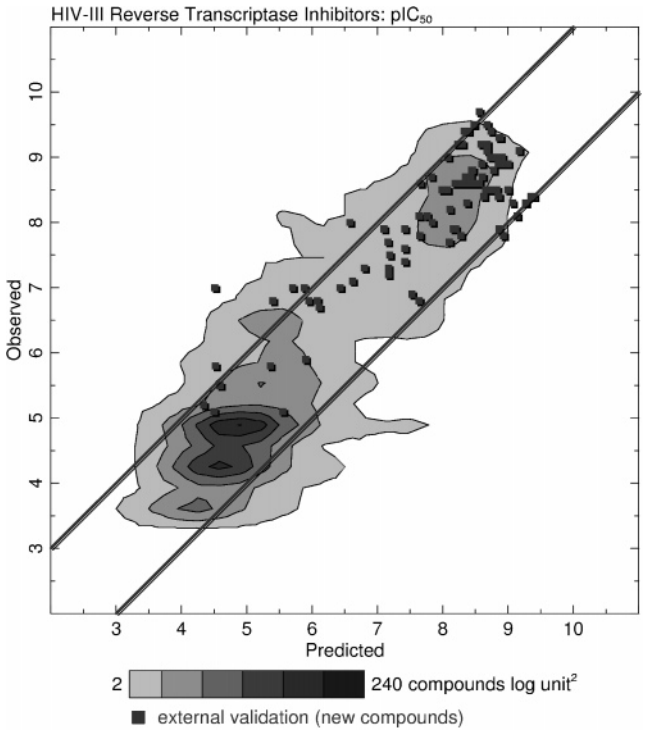

Figure 5. Final self-consistent model, training set contoured by measurement density and external reference as square markers. Training set $q^2 = 0.681$; solid lines indicate a +1 or −1 log-value range.

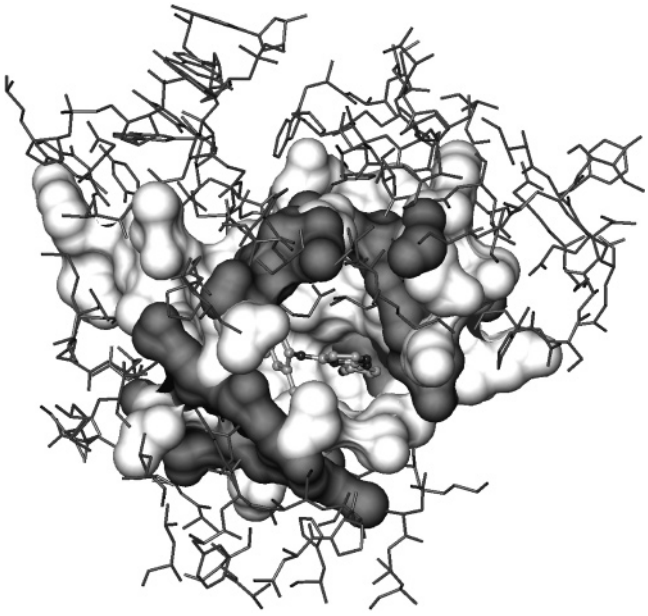

Figure 6. NNRTI binding site, with residues involved in the inhibition model. Light gray area indicates the protein side chains, dark gray the backbone atoms. The thin lines show residues that do not contribute energy terms to the prediction. The molecule in the pocket is a typical potent NNRTI.

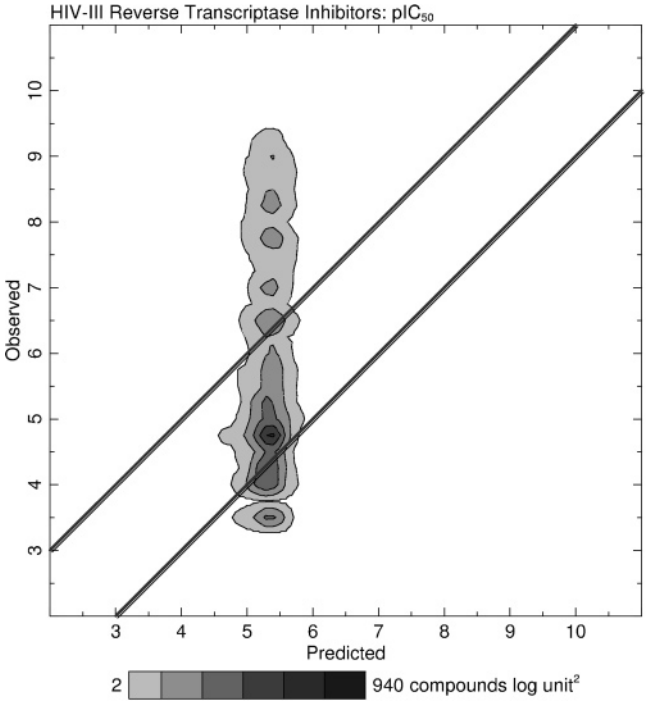

Figure 7. Test for overfitting. The "best" model found by the genetic algorithm after destruction of correlation of the computed interaction matrix and pIC_{50} by randomizing the observation column (training set $q^2 = 0.011$).

promising sets of variables. This helps to avoid the problems often associated with stepwise regression; the GA, for instance, converges to the same optimum, whether the initial guess for the model consists of all-on or of all-off genes, essentially simulating stepwise selection and stepwise removal of variables. Another feature of this approach is that there is no unphysical relative scaling of energy contributions involved, as often occurs in linear interaction energy methods, and which bears with it a large risk of over-fitting the data. In this approach a particular contribution either pertains to the inhibition region of RT or it does not. We consider this a more realistic description for this non-competitive binding site. As a result, for a chosen binding pocket described by the set of residues, only the two parameters defining a straight line are optimized.

The GA run is repeated for each of the 100 subsets of points, resulting in 100 straight-line models where a number of predicted activities will be far away from the best line. These points are marked "inconsistent" for this particular run. They need not be incorrectly predicted, as we do not know a priori how good the fitted line through a subset of points actually is, and whether the current subset is biased by the data selection. Finally we derive a consensus model based on those complexes where the best binding ligand conformer was consistent in at least in 50% of the runs. Figure 4 shows the numbers of consistent (light gray) and inconsistent (dark gray) models at each activity bin. The final model for HIV-1 RT activity prediction is shown in Figure 5; the training set points are so heavily clustered around the diagonal that the point density is shown with the contours (up to 240 data points in a 1×1 pIC_{50} cell). In this figure 100 independent verification compounds are plotted, mostly synthesized and all tested after the generation of the model, for instance from the novel series of NNRTIs shown in Heeres et al.[24]

It is not obvious how many degrees of freedom are really available for the binary residue selection, so the q^2 statistic should be used with caution. To be certain that we are not overfitting the data by means of this variable selection approach, we repeated the analysis of the same interaction energy matrix with all computed energy values, but with the observed pIC_{50} rearranged in random order. The resulting best model after this randomization is shown in Figure 7. For this randomized model, R^2 is 0.014 and q^2 is 0.011, and all predicted activities are approximately equal to the average pIC_{50} of the training set (with a resulting average absolute error of 1.3). This clearly shows that our modeling approach does not generate a spurious energy−activity correlation by means of the residue selection.

We therefore conclude that the model approach that we have derived is capable of predicting the antiviral activity of an arbitrary NNRTI by means of a rapid computation and in a robust manner that does not require any user intervention. These are the exact characteristics we need for the automatic drug design approach of the SYNOPSIS[10] program while still being useful in more traditional drug design methods.[24]

References

(1) Gallo, R. C. The AIDS virus. *Sci. Am.* **1987**, *256*, 46−56.
(2) Mitsuya, H.; Weinhold, K. J.; Furman, P. A.; St Clair, M. H.; Nusinoff-Lehrman, S.; Gallo, R. C.; Bolognesi, D.; Barry, D. W.; Broder, S. 3′-azido-3′deoxythymidine (bwa509u): an antiviral agent that inhibits the infectivity and cytopathic effect of human T-lymphotropic virus III/lymphadenopathy-associated virus in vitro. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 7096−7100.
(3) Pauwels, R.; Andries, K.; Desmyter, H.; Schols, D.; Kukla, M. J.; Breslin, H. J.; Raeymaeckers, A.; Van Gelder, J.; Woestenborghs, R.; Heykants, J.; Schellekens, K.; Janssen, M. A. C.; De Clercq, E.; Janssen, P. A. J. Potent and selective inhibition of HIV-1 replication in vitro by a novel series of TIBO derivatives. *Nature* **1990**, *343*, 470−474.
(4) Merluzzi, V. J.; Hargrave, K. D.; Labadia, M.; Grozinger, K.; Skoog, M.; Wu, J. C.; Shih, C.-K.; Eckner, K.; Hattox, S.; Adams, J.; Rosenthal, A. S.; Faanes, R.; Eckner, R. J.; Koup, R. A.; Sullivan, J. L. Inhibition of HIV-1 replication by a nonnucleoside reverse transcriptase inhibitor. *Science* **1990**, *250*, 1411−1413.
(5) Wei X.; Ghosh, S. K.; Taylor M. E.; Johnson, V. A.; Emini, E. A.; Deutsch, P.; Lifson, J. D.; Bonhoeffer, S.; Nowak, M. A.; Hahn, B. H. Viral dynamics in human immunodeficency virus type 1 infection. *Nature* **1995**, *373*, 117−122.
(6) Frost, S. D. W.; McLean, A. R. Quasispecies dynamics and the emergence of drug resistance during zidovudine therapy of HIV infection. *AIDS* **1994**, *8*, 323−332.
(7) Eshleman, S. H.; Guay, L. A.; Mwatha, A.; Brown, E. R.; Cunningham, S. P.; Musoke, P.; Mmiro, F.; Jackson, J. B. Characterization of nevirapine resistance mutations in women with subtype A vs. D HIV-1 6-8 weeks after single-dose nevirapine (HIVNET 012). *JAIDS, J. Acquired Immune Defic. Syndr.* **2004**, *35*, 126−130.
(8) Eriksson, M. A. L.; Pitera, J.; Kollman, P. A. Prediction of the binding free energies of new TIBO-like HIV-1 reverse transcriptase inhibitors using a combination of PROFEC, PB/SA, CMC/MD, and free energy calculations. *J. Med. Chem.* **1999**, *42*, 868−881.
(9) Smith, R. H. J.; Jorgensen, W. L.; Tirado-Rives, J.; Lamb, M. L.; Janssen, P. A. J.; Michejda, C. J.; Smith, M. B. K. Prediction of binding affinities for TIBO inhibitors of HIV-1 reverse transcriptase using Monte Carlo simulations in a linear response method. *J. Med. Chem.* **1998**, *41*, 5272−5286.
(10) Vinkers, H. M.; De Jonge, M. R.; Daeyaert, F. F. D.; Heeres, J.; Koymans, L. M. H.; Van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Van Aken, K.; Janssen, P. A. J. SYNOPSIS: synthesize and optimize system in silico. *J. Med. Chem.* **2003**, *46*, 2765−2773.
(11) Aqvist, J.; Medina, C.; Samuelson J. E. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385−391.
(12) NIAID (2004) Database for anti-HIV compounds, url: http://apps1.niaid.nih.gov/struct_search/deafult.htm.

(13) Rizzo, R. C.; Udier-Blagovic, M.; Wang, D.; Watkins, E. K.; Smith, M. B. K.; Smith, R. H. S.; Tirado-Rives, J.; Jorgensen, W. L. Prediction of activity for nonnucleoside inhibitors with HIV-1 reverse transcriptase based on Monte Carlo simulations. *J. Med. Chem.* **2002**, *45*, 2970−2987.
(14) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, 17, 490−519.
(15) Halgren, T. A. MMFF94s option for energy minimization studies. *J. Comput. Chem.* **1999**, *20*, 720−729.
(16) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(17) Das, K.; Clark, A. D.; Lewi, P. J.; Heeres, J.; De Jonge, M. R.; Koymans, L. M. H.; Vinkers, H. M.; Daeyaert, F. F. D.; Ludovici, D. W.; Kukla, M. J.; De Corte, B.; Kavash, R. W.; Ho, C. Y.; Lichtenstein, M. A.; Andries, K.; Pauwels, R.; Boyer, P. L.; Clark, P.; Hughes, S. H.; Janssen, P. A. J.; Arnold, E. Roles of conformational and positional adaptability in structure based design of tmc125-r165335 (Etravirine) and related non-nucleoside reverse transcriptase inhibitors that are higly potent and effective against wild-type and drug-resistant HIV-1 variants. *J. Med. Chem.* **2004**, *47*, 2550−2560.
(18) Tomioka, N.; Itai, A.; Iitaka, Y. A method for fast energy estimation and visualization of protein-ligand interaction. *J. Comput.-Aided Mol. Des.* **1987**, *3*, 197−210.
(19) Daeyaert, F. F. D.; De Jonge, M. R.; Heeres, J.; Koymans, L. M. H.; Lewi, P. J.; Vinkers, H. M.; Janssen, P. A. J. A pharmacophore docking algorithm and its application to the cross-docking of 18 HIV-NNRTI's in their binding pockets. *Proteins: Struct., Funct., Bioinf.* **2004**, *54*, 526−533.
(20) Corana, A.; Marchesi, M.; Martini, C.; Ridella, S. Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm. *ACM Trans. Math. Software* **1987**, *13*, 262−280.
(21) Nash, S. G. Newton-type minimization via the lanczos method. *SIAM J. Numer. Anal.* **1984**, *21*, 770−778.
(22) Hertogs, K.; de Béthune, M. P.; Miller, V.; Ivens T.; Schel, P.; Van Cauwenberge, A.; Van Den Eynde, C.; Van Gerwen, V.; Azijn, H.; Van Houtte, M.; Peeters, F.; Staszewski, S.; Conant, M.; Bloor, S.; Kemp, S.; Larder, B.; Pauwels, R. A rapid method for simultaneous detection of phenotypic resistance to inhibitors of protease and reverse transcriptase in recombinant HIV-1 isolates of patients treated with antiretroviral drugs (PR-RT-Antivirogram™). *Antimicrob. Agents Chemother.* **1998**, *42*, 269−276.
(23) Press, H. P.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical recipes in C*; Cambridge University Press: Cambridge, 1988.
(24) Heeres, J.; Lewi, P. J.; De Jonge, M. R.; Koymans, L. M. H.; Daeyaert, F. F. D.; Vinkers, H. M.; Van Aken, K. J. A.; Arnold, E.; Das, K.; Hoornaert, G. J. C.; Compernolle, F. J. C.; Kilonda, A.; Ceglia, M.; Azzam, R. A.; Andries, K.; de Béthune, M.-P.; Azijn, H.; Pauwels, R.; Janssen, P. A. J. Design, synthesis, and sar of a novel pyrazinone series with non nucleoside HIV-1 reverse transcriptase inhibitory activity (NNRTI). *J. Med. Chem.*, in press.

JM049534R