CovalentDock: Automated covalent docking with parameterized covalent linkage energy estimation and molecular geometry constrains

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Journal of Computational Chemistry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>JCC-12-0144.R1</td>
</tr>
<tr>
<td>Wiley - Manuscript type:</td>
<td>Software News and Updates</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>06-Aug-2012</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Ouyang, Xuchang; Nanyang Technological University, BioInformatics Research Centre, School of Computer Engineering Zhou, Shuo; Peking University, State Key Laboratory of Natural and Biomimetic Drugs, School of Pharmaceutical Sciences Su, Chinh Tran To; Nanyang Technological University, BioInformatics Research Centre, School of Computer Engineering Ge, Zemei; Peking University, State Key Laboratory of Natural and Biomimetic Drugs, School of Pharmaceutical Sciences Li, Runtao; Peking University, State Key Laboratory of Natural and Biomimetic Drugs, School of Pharmaceutical Sciences Kwoh, Chee Keong; Nanyang Technological University, BioInformatics Research Centre, School of Computer Engineering</td>
</tr>
<tr>
<td>Key Words:</td>
<td>covalent docking, covalent drug, molecular docking, virtual screening, computer-aided drug design</td>
</tr>
</tbody>
</table>
CovalentDock: Automated covalent docking with parameterized covalent linkage energy estimation and molecular geometry constrains

Xuchang Ouyang‡*, Shuo Zhou†*, Chinh Tran To Su†, Zemei Ge†, Runtao Li†, Chee Keong Kwoh‡

† State Key Laboratory of Natural and Biomimetic Drugs, School of Pharmaceutical Sciences, Peking University, Beijing, China 100191
‡ BioInformatics Research Centre, School of Computer Engineering, Nanyang Technological University, Singapore 639798
* These authors contributed equally to this work.
Correspondence to: Chee Keong Kwoh (asckkwoh@ntu.edu.sg) and Runtao Li(lirt@bjmu.edu.cn)

ABSTRACT

Covalent linkage formation is a very important mechanism for many covalent drugs to work. However, partly due to the limitations of proper computational tools for covalent docking, most covalent drugs are not discovered systematically. In this paper, we present a new covalent docking package, the CovalentDock, built on the top of the source code of Autodock. We developed an empirical model of free energy change estimation for covalent linkage formation, which is compatible with existing scoring functions used in docking, while handling the molecular geometry constrains of the covalent linkage with special atom types and directional grid maps. Integrated preparation scripts are also written for the automation of the whole covalent docking workflow. The result tested on existing crystal structures with covalent linkage shows that CovalentDock can reproduce the native covalent complexes with significant improved accuracy when compared to the default covalent docking method in Autodock. Experiments also suggest that CovalentDock is capable of covalent virtual screening with satisfactory enrichment performance. In addition, the investigation on the results also shows that the chirality and target selectivity along with the molecular geometry constrains are well preserved by CovalentDock, showing great capability of this method in the application for covalent drug discovery.

Keywords

covalent docking; covalent drug; molecular docking; virtual screening; computer-aided drug design
INTRODUCTION

Molecular docking is a computational procedure to predict the binding conformation as well as the binding affinity of a complex of two molecules, typically one receptor and one ligand. Well-known docking programs and software include Autodock, Autodock Vina, GOLD, and FlexX. However, these and many other methods mainly focus on the docking between the two molecules through non-covalent interactions, such as van der Waals interaction, the electrostatics interaction and hydrogen bonding, or using other empirical or knowledge-based scoring functions to characterize these non-covalent interactions. The focus on non-covalent interactions is quite understandable since the main stream of rational drug design relies on these non-covalent interactions as the mechanism of the functionality of the drugs.

However, another important category of drugs, namely the covalent drugs, adopts covalent bond formation as part of its binding mechanism. Distinct from conventional non-covalent drugs, the covalent drugs usually have much stronger binding affinity with the targets because of the covalent linkage formed in between. This higher binding affinity of typical covalent drugs brings a risk of toxicity due to the difficulty of disassociation if off-target binding happens. Thus highly specified selectivity profiles of the covalent drugs are required. However, this characteristic also enables covalent drug to have a stronger potency while maintaining a pharmaceutically favored small molecule size, which makes the covalent drugs quite efficient and widely used. In fact, many famous drugs, including aspirin and penicillin, are covalent drugs. And it is reported that three of the ten top-selling drugs in U.S. in 2009 are also covalent drugs, suggesting a great potential in this area.

Despite the popularity of covalent drugs, however, most of them are discovered by serendipity instead of systematical design. Nowadays, with advanced and specialized computational tools, it is possible to design and simulate covalent drugs in silico, and to perform large scale investigation on different target proteins to get the selectivity profile of the designed covalent drug in order to reduce of risk of toxicity, thus offering a new perspective to covalent drug design. However, the existing computational tools do not account for this kind of important chemical event very well. For example, two of the most popular molecular docking packages, Autodock and GOLD, do support “covalent docking” feature, but with major limitations in their functionality, accuracy and usefulness: Autodock uses a naive “Gaussian well” on the score to
reward the covalent bonding atoms. However, the “Gaussian well” must be set to a specified coordinate instead of on any target link atoms. The over-simplification of the scoring function as well as the lack of established guideline in parameter setting also compromises its performance. GOLD uses a “link atom” added to both ligand and receptor structure to ensure the geometry of covalent docking. However, the scoring functions are modified drastically to reduce clash penalty but not to reward the linkage formation over non-covalent docking, and the covalent bonding is “forced” to happen despite the overall docking formation. Moreover, since most of the molecular docking programs are designed for non-covalent docking, preparation steps for covalent docking requires much effort, and it is even more difficult to perform convenient virtual screening on covalent drug library due to the lack of automation in setting up the experiment. This gap between the need and practical difficulty in computational tools for covalent docking leads to the research of this project.

In this project, we developed a model to account for the energy contribution from covalent binding as a term compatible to the scoring function of non-covalent docking. We implemented this method as CovalentDock based on the source code of Autodock. Inspired by both Autodock and GOLD, we devised a method that utilizes a dummy atom to serve the similar purpose as the “link atom” in GOLD, integrating the proposed energy term into the scoring function of Autodock, and use the built-in grid map calculation and Genetic Algorithm to search for the docking results. Meanwhile, we developed a procedure to automatically recognize and prepare the covalently bondable chemical groups, enabling large scale automated covalent docking and covalent virtual screening. This implementation was tested on 76 covalently bound complexes collected from the Protein Data Bank (PDB). The results show that our method is able to reproduce the conformations in native PDB structures with an average RMSD=1.68Å, outperforming result given by the default covalent docking method of Autodock with average RMSD=2.49Å and of GOLD with average RMSD=3.69Å on the same data set. Results of the covalent virtual screening test also suggest that our method has a better capability to enrich true active covalent binders from randomly selected decoys compared to the default covalent docking method of Autodock. In addition, the investigation on the results shows that the chirality and target selectivity of the covalent drugs, along with the molecular geometry constrains of the covalent linkages, are well preserved by CovalentDock, showing great capability of this method.
METHOD

Overall workflow

As a special version of the docking package featuring covalent docking, the workflow is fundamentally similar to the conventional molecular docking. Similar to the procedures in conventional non-covalent docking, preprocessing of the structures is needed to make them suitable for covalent docking. The preparation work needed for covalent docking in preprocessing starts with identifying the bond forming groups. Then the processed structures will be docked with covalent linkage expected. Finally the result can be examined and compared with non-covalent docking to identify possible covalent binding. The whole procedure is shown in Figure 1.

![Figure 1. Overall workflow for covalent docking.](image)

Modeling the contribution of covalent bond formation

The change of Gibbs free energy of a chemical system is given by:

\[ \Delta G = \Delta H - T \Delta S \]  \hspace{1cm} (1)

where \( \Delta H \) is the enthalpy change, \( T \) is the temperature of the system and \( \Delta S \) is the entropy change.

Consider a complex of two molecules at a constant temperature, forming a covalent bond between them will change the free energy of the system. The change in free energy comes from two sources:
1. Formation of the covalent bond between the two molecules, sometimes coupled with bond breaking as well, gives a change in enthalpy. This enthalpy change can be estimated by existing chemistry references\textsuperscript{17,18}. However, considering that the standard enthalpy change only provides an estimation on the optimal bond length, while in molecular docking, that length might not be perfectly optimal. Thus, we formulated the Morse potential\textsuperscript{19} with parameters fitting to simulated result from Gaussian\textsuperscript{20} to estimate the enthalpy change brought by covalent bond formation on sub-optimal bond length. The enthalpy change of a given covalent bonding pair given by the Morse potential is:

\[
\Delta H = D \left( e^{-2\alpha(r-r_0)} - 2e^{-\alpha(r-r_0)} \right) \quad \ldots (2)
\]

where D is the dissociation energy, \(\alpha\) is a parameter controlling the well width, r is the bond length and \(r_0\) is equilibrium bond length. The parameterization of this equation is subjected to the specified type of covalent bonding pairs.

2. Forming such a covalent bond will also influence the conformation of the two molecules, putting some additional constrains on both structures to satisfy the molecular geometry, and bringing a conformational entropy change as well. This part of energy contribution can also be estimated with simulated result from Gaussian\textsuperscript{20}.

With these two components estimated, the energy change \(\Delta G\) of forming a covalent bond could be calculated from the conformation of the complex during docking.

In order to make the energy estimation compatible with the scoring function in the docking program used to estimate the entropy change, an additional empirical term is also included to correct the difference between the predicted energy score and empirical data. The overall additional energy term for scoring functions is:

\[
E = \begin{cases} 
D \left( e^{-2\alpha(r-r_0)} - 2e^{-\alpha(r-r_0)} \right) - T\Delta S_{\text{est}} + C, & r \leq r_m \\
0, & r > r_m 
\end{cases} \quad \ldots (3)
\]

where \(r_m\) is the maximum bond length without disassociation (which is defined as the distance when \(E=0\) when \(r>r_0\), \(\Delta S_{\text{est}}\) is the conformation entropy estimated by Gaussian, C is a correcting empirical constant. An example of this energy term with parameters determined for \(\beta\)-lactam family is shown in Figure 2.

As a term in the scoring function, this energy change appears as a well, with maximum reward at the optimal bond length, rising sharply when the distance approaches to zero to penalty this
tendency, and smoothing to zero when distance goes too large. Since the energy term has already been corrected with an empirical constant according to the original scoring function, it could be directly embedded to the corresponding docking program.

**Implementation of the proposed model in Autodock**

Autodock\(^4^6\) is one of the most popular and effective molecular docking packages\(^2\). The software package includes two major components: *autogrid* and *autodock* (to discriminate this component from the whole package, Autodock with a capital initial is used to refer to whole package). The molecular docking is carried out in a two-step manner: *autogrid* pre-calculates the energy grid maps for each atom type in the ligand, later *autodock* utilizes the grid maps to estimate the energy of the complex conformation during docking by interpolation.

In order to use the docking scheme of Autodock to perform covalent docking, the *autogrid* is modified to take two additional atom types: the atom type R in receptor and L in ligand, as shown in Figure 3, where R and L are the atom pairs identified that the covalent bond could be formed. The atom type R and L will be treated as normal atom types when interacting with other atom, while the interaction between R and L will be treated specially with the model described in the previous section. Therefore, the docking of the two molecules could be considered as a complex with a “conceptual” covalent linkage between R and L. Unlike the implementation in GOLD\(^8^9\), this docking procedure dose not force R and L to end up with a favorable position in
order to form a covalent linkage, but instead it takes the all contribution between the ligand and the receptor to produce the final result conformation.

Some special tweaks are introduced to ensure the optimal geometry of the covalently docked complex is obtained. As shown by the example of Michael addition in Figure 3, two bond angles $\gamma$ and $\theta$ at the two ends of the “conceptual” covalent linkage should have correct bending in the complex. During the structural preparation steps, which will be detailed in the following section, a dummy atom will be attached to the bond forming atom $L$ in the ligand. The structural optimization during that process will ensure angle $\gamma'$ is arranged as in Figure 3 to satisfy proper molecular geometry constrains. This dummy atom will end up with approaching $R$, should the covalent bond formed, giving an angle $\gamma$ approaching $\gamma'$. Meanwhile, angle $\theta$ is rewarded to the optimal angle by taking the energy potential between $R$ and $L$ multiplied with a directional term, which is the same term used to ensure the directionality of H-bond in Autodock. The covalent bond forming atom $L$ with angle $\theta$ will have an energy potential:

$$E_{\text{angle}} = E_{\text{cov}} \cdot \cos^4(\theta - \theta_0)$$

where $E_{\text{cov}}$ is the covalent binding free energy estimated with Equation(3), $\theta_0$ is the optimal angle bending.

**Automated structure preparation with chirality consideration**

For practical use of the covalent docking, the structures, especially the structures of the ligand, are presented commonly to the docking program in an “intact form”, i.e. the form before covalently bound to the receptor. During covalent docking, formation of the covalent linkage would cause structural changes in both molecules, especially in the ligand. An example is shown in Figure 4, where (a) and (c) are the structure of the same ligand before and after the covalent linkage formed with the receptor respectively. In order to adapt these structural changes, the structure of both the receptor and the ligand need to be carefully prepared with automated procedures. The potential bond-forming chemical groups and specific atom pair in both receptor and ligand are first identified. For the receptor protein, it mainly involves locating nucleophilic terminals within the binding site, such as the oxygen on serine residues and sulfur on cysteine residues. For the ligand, the main task is to recognize special patterns of electrophilic chemical groups based on the topology of the compound. When the bond forming groups have been identified, it is necessary to alter the structures in order to perform covalent docking. While the
changes in the receptor structure are minor, which usually involves just losing a hydrogen atom, the changes in the ligand structure can be relatively significant. For example, the basic mechanism of penicillin is breaking the four-membered $\beta$-lactam ring between the nitrogen and carbonyl carbon, followed by carbonyl carbon forming a covalent linkage to the serine on the receptor in the active site.\textsuperscript{21,22} The structure altering of the penicillin starts with an intact $\beta$-lactam ring, as shown in Figure 4(a). Then the bond between C1 and N on the $\beta$-lactam ring is broken. A hydrogen atom is transferred from the serine nucleophilic oxygen to the N. Since C1 now has an available valence, it is able to form a covalent bond with the receptor. A “dummy” oxygen atom is artificially attached to C1 to temporarily occupy the empty valence of C1, as shown in Figure 4(b). Now the topology of the structure becomes identical with the covalently bonded ligand (with one extra “dummy atom” substituting the corresponding bond-forming atom in receptor), but its molecular geometry is not optimal. A short structural optimization is then performed for this structure using Amber 10\textsuperscript{23} with force field ff99sb and GAFF, to obtain a final optimized structure ready to perform covalent molecular docking, as shown in Figure 4(c).

During this step, chirality is carefully considered and handled. As shown in Figure 5, after covalently bound to the receptor, two new chiral centers are generated on the ligand and there will be 4 different possible configurations. Without prior knowledge, all four stereoisomers are
passed to covalent docking. The one with highest score in result is regarded as the correct configuration.

In the package of CovalentDock, this preparation procedure is automated with integrated scripts. It therefore enables large scale automated virtual screening without human interference. Do note that the protocol we offered relies on the molecular dynamic package Amber 10\textsuperscript{23} for structure optimization. However, the covalent docking method is also compatible with structures processed by other means, or directly obtained from existing structure databases. However, the chirality should still be considered.

**RESULT AND DISCUSSION**

The CovalentDock package for covalent docking is implemented based on the source code of Autodock 4.2. We tested this package on a dataset of 76 complexes discussed in literature, all of them are experimentally observed with covalent linkage formed between the ligand and the receptor. The structures are collected from the Protein Data Bank\textsuperscript{16}. For each of the complex, the corresponding original ligand structure before covalent bond formed is prepared using PyMol\textsuperscript{24}. The covalently bound ligand structure is deleted from the complex, leaving the receptor structure for our covalent docking experiment.

Among the 76 complexes, 13 are with Michael acceptors, and the rest are in $\beta$-lactam family. The basic chemical reactions of the covalent binding for both Michael acceptor and $\beta$-lactam family along with corresponding atom labeling are shown in Figure 6.

![Figure 6. Reactions of covalent binding. The bond forming atom pairs are labeled R and L for the atom in receptor and in ligand, respectively. The covalent bond formed between the receptor and ligand is illustrated with a dash line. (a) Reaction between the Michael acceptor group in ligand and Cysteine residue in receptor; (b) Reaction between the $\beta$-lactam ring in ligand and Serine residue in receptor.](image)

<table>
<thead>
<tr>
<th></th>
<th>Michael addition</th>
<th>$\beta$-lactam</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>10.16</td>
<td>21.54</td>
</tr>
<tr>
<td>$r_0$</td>
<td>1.84</td>
<td>1.36</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>3.41</td>
<td>3.78</td>
</tr>
<tr>
<td>$\Delta S_{\text{est}}$</td>
<td>-9.36</td>
<td>-10.96</td>
</tr>
<tr>
<td>$C$</td>
<td>4.74</td>
<td>0.00</td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>100.83°</td>
<td>119.62°</td>
</tr>
</tbody>
</table>

Table 1. Parameter used for the Michael addition group and $\beta$-lactam group. $r_0$ and $\theta_0$ are extracted from Amber Force Field parameters\textsuperscript{25}. $D$, $\alpha$ and $\Delta S_{\text{est}}$ is estimated by Gaussian simulation under 298.15K.
Parameters used in the energy estimation, as described in Equation (3), are listed in Table 1. For Michael additions, since the binding affinity is directly available in literature, the constant C in Equation (3) is estimated empirically to correct the final binding affinity score, while it is set to zero for β-lactams due to the lack of experimentally measured binding affinity.

**Native structure reproduction**

The capability of CovalentDock to reproduce the native ligand structure is tested against the covalent docking protocols in original Autodock and GOLD. The experiment is carried out by feeding all 76 experimentally determined complexes with covalent linkages to the docking programs for the prediction of docking conformations. The main criterion is the minimum Root-Mean-Square Deviation (RMSD) of the heavy atoms (all atoms except for hydrogen) between the result and the native structure achieved on each complex over 10 repeated but independent docking experiments, as defaulted in Autodock and GOLD.

The results show that CovalentDock can reproduce the covalently bound native structure very accurately, if the correct chirality is chosen. The 76 complexes have an overall average minimum RMSD=1.68Å, with an average minimum RMSD= 1.73Å and 1.67Å for Michael additions and β-lactams, respectively.

This result is compared with the result produced by the default covalent docking method in Autodock. With the Gaussian well set to be centered at the position of the corresponding bond
forming atom in receptor, and both well width and depth set as default values, Autodock obtained a result with average RMSD of 2.49Å for all the complexes tested. The comparative result is shown in Figure 7. The performance of CovalentDock is also compared with GOLD, benchmarked on the same data set with default covalent docking protocol. The result shows that GOLD doesn’t perform very well on most complexes, given an average minimum RMSD of 3.69Å, as shown in Figure 8. However, detailed analysis suggests that GOLD performs exceptionally well on kinases, which will be elaborated in one of the following case studies.

For the capability to reproduce the native structure, another criterion used is the “hit rate”, calculated as the percentage of the complexes with a minimum RMSD within a pre-defined cutoff. The hit rates for CovalentDock, Autodock and GOLD with different cutoff RMSD are shown in Figure 9. The result shows that for all given cutoff RMSD, as exemplified with 1Å, 2Å and 3Å, CovalentDock is more likely to produce “hits” than the other two covalent docking protocol. Even with the cutoff RMSD set to 1Å, which will only consider the results highly similar to the native structure as hits, CovalentDock still gives a hit rate at 30.26%, while Autodock and GOLD produced “hit” on 17.11% and 6.58% on the same dataset of 76 complexes.

![Figure 9. Hit rate achieved by CovalentDock, Autodock and GOLD on the dataset of 76 complexes, with different cutoff RMSD.](image)

**Ranking of the near-native structures**

Ideally, molecular docking programs, either covalent or non-covalent, should always put the “correct conformation” on top ranking positions. However, with imperfections in the scoring functions in use, it is found that the most near-native structures (the ones with minimum RMSD)
are not always properly ranked by the docking methods to-date.\textsuperscript{25-27} We examined the capability to rank the near-native structures of CovalentDock and compared with the covalent docking method in Autodock and GOLD as well. For this test, the number of repeated docking experiment is increased to 100 for each complex in our dataset.

For each complex in the dataset, the ranking of the conformation with minimum RMSD is recorded among all the 100 conformations. On average, the result structures with minimum RMSD are found at top 30.22, 41.04, and 63.78 ranking positions for CovalentDock, Autodock and GOLD, respectively. Although ranking is not the primary goal in the design of the energy estimation scheme, by implementing this energy estimation in the covalent docking protocol, the ranking performance of CovalentDock is also an evidence of its superiority.

We also examined the minimum RMSD and hit rate for these three covalent docking protocols when only a given number of top ranking conformations are considered. Figure 10 shows the average minimum RMSD that can be found in the result for each complex among a given number of top ranking conformations, and Figure 11 plots the corresponding hit rate achieved by CovalentDock, Autodock and GOLD, with a RMSD cutoff set at 2Å. The result clearly shows that CovalentDock is with an advantage in terms of both RMSD and hit rate, regardless the acceptance cutoff set in rank.

![Figure 10. Average minimum RMSD found among a pre-defined top ranking cutoff number.](image1)

![Figure 11. Hit rate achieved by considering only a pre-defined number of top ranking conformations. Results with RMSD less than 2Å are considered “hits” in this plot.](image2)
Covalent Virtual Screening

The automated preparation procedure we developed enables CovalentDock to perform large scale virtual screening on covalent drug candidates. To illustrate this capability, we performed a virtual screening test on Michael acceptors against the target EGFR kinase domain.

The receptor structure used in this virtual screening test is extracted from the PDB complex 1XKK\textsuperscript{28}. The terminal sulfur atom on Cys797 is identified as the covalent binding site. The library used in this virtual screening test comes from the ZINC database\textsuperscript{29,30}, using the “Lead-Like” subset containing 5,037,475 chemical compounds, which is prepared, filtered and recommended by ZINC for discovery projects. Open Babel\textsuperscript{31} is used to add missing hydrogen to the structure. All the compounds with Michael acceptors are identified from the raw library using Open Babel with the SMARTS pattern “[\text{S}(\text{CH1,CH2}=\text{CC}=\text{O})]”, resulting 160,957 compounds in total ready for screening. For testing purpose, we randomly selected 19,949 compounds from all the Michael acceptor containing compounds, which will be considered as decoys in this test, although it is understood that some of them may be potential active leads. On the other hand, 51 active ligands are collected from literature\textsuperscript{32-39}. As a result, the library we screened has 20,000 compounds in total, with 51 positive controls.

The virtual screening experiment is carried out by performing a quick covalent docking using CovalentDock. The preparation scripts will preprocess the structure and prepare relevant configuration files to enable automated testing. For comparison, the same virtual screening test is also carried out using the default covalent docking protocol in Autodock. GOLD is left out from this test due to the difficulty of its automation because the structures and configuration files prepared by the scripts developed for CovalentDock are not compatible with GOLD and cannot be easily translated.

The Enrichment Factor (EF) is used as a performance indicator for the virtual screening tests. The enrichment factor at the top p% is calculated with:

\[
\text{EF}_{p\%} = \frac{H_{p\%}}{H_{\text{total}} \cdot p\%}
\]

where \(H_{p\%}\) is the number of positive controls compounds found in the top p% ranking results among all the compounds screened, and \(H_{\text{total}}\) is the total number of positive controls, which will be 51 in this virtual screening test. In addition, the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) is also examined. The results are shown in Figure 12 and
Figure 12 shows that for any given value of p%, CovalentDock gives better EF compared to the default covalent docking methods in Autodock. For example, the EFs calculated at top 5% are 17.25 and 4.31 for CovalentDock and Autodock, respectively, meaning that it is much more likely to find positive controls in the top 5% ranking compounds by using CovalentDock than using Autodock. The ROC plotted in Figure 13 shows that CovalentDock achieved an AUC of 0.97, while the AUC for the Autodock ROC is 0.68. Both performance indicators show that the covalent docking method we proposed as CovalentDock outperforms the default docking method in Autodock in virtual screening test with a significant advantage.

**Case Study: Chirality selectivity**

2AX0\(^{40}\) is the complex of a novel non-nucleotide NS5b RNA polymerase inhibitor and its receptor. The ligand has a Michael acceptor inside and in the structure a covalent bond is formed between the beta carbon of the Michael acceptor and the sulfur atom of Cys366. As a result, the alpha and beta carbon of the Michael acceptor become chiral, and the ligand adopts the alpha(S)-beta(R) configuration. In our experiments, all the four possible configurations of the ligand after bonding are investigated. As expected, the top ranking conformation adopts the alpha(S)-beta(R) configuration. The RMSD of the result is 1.24Å while all the important lipophilic interactions and hydrogen bonds in the original complex are preserved, as shown in Figure 14(a). However,
Figure 14. Covalent docking result structures for case studies. In all subfigures, the receptors are shown as cyan cartoon, the native ligand structures are shown as green sticks and docking results given by CovalentDock are shown as magenta sticks. Special alternative ligand structures discussed by each case study are shown as yellow sticks. (a) 2AX0 with a Michael acceptor in the ligand, the magenta ligand structure has the same configuration with the green native structure, while the yellow one adopts a different configuration. (b) 3C9W with a Michael acceptor in the ligand (c) 2E14 with a Michael acceptor in the ligand. The covalent linkage shown in the green native structure has distorted bond length (2.35Å) and angle (154.8°). (d) 1IYP with a β-lactam ring in the ligand and three serine residues located in the binding site. While the highest ranking result shown in magenta formed a linkage with the same serine (SER-70) of the green native ligand, covalent linkages with alternative serine residues are also found in lower ranking results. An example of a covalent linkage formed between the ligand and SER-237 are shown in yellow. (e) 2ZC3 with a β-lactam ring in the ligand, the magenta ligand structure agree with the green native structure very well around the covalent binding site, while the yellow structure (produced by Autodock) has a distorted molecular geometry on the covalent linkage. (f) 2HWO with Michael acceptor in the ligand. The geometry constrains on the covalent linkage are well satisfied. The yellow structure (produced by GOLD) has better alignment with the green native structure than the magenta one, mainly due to the conformation of the non-covalent parts of the ligand, although both yellow and magenta conformations are with acceptable accuracy.
the docking modes of the three alternative configurations do not agree with the native structure at all, one example of the binding mode of alpha(S)–beta(S) configuration is also shown in Figure 14(a). This selectivity in chirality of the ligand has also been observed in many other complexes, and all of the top ranking conformations are observed to adopt the correct configuration, expect the complex 2E14 which will be discussed in next case study.

**Case Study: Chirality alert**

The complexes of 3C9W\(^{41}\) and 2E14\(^{42}\) are ERK with hypothemycin and its derivative FR148083 as their ligands. The only difference between them is that FR148083 is an analog with a double bond in place of the epoxide, thus we could expect that their docking mode will be very similar. Good results are given by CovalentDock for both complexes, with RMSD of 0.98Å and 1.26Å respectively, and both the covalent linkage between the ligand and receptor are in the R configurations, as shown in Figure 14(b) and Figure 14(c), respectively. While the linkage in the native structure of 3C9W indeed adopted the R configuration,\(^{41}\) however, this prediction conflicts with the native structure of 2E14, which adopts the S configuration.\(^{42}\) It has already been discussed by Rastelli et al.\(^{41}\) that R configuration was in better agreement with both the electron density and the expected tetrahedral geometry around the carbon, while the geometry of FR148083 bond to ERK as shown in 2E14 with S configuration has distorted bond lengths, angles and torsions. With this supporting statement, it is suggested that CovalentDock could give a reasonable prediction of covalent complex formation in which covalent linkage is formed in more favorable configuration.

**Case Study: Target selectivity**

1IYP\(^{43}\) is the Toho-1 β-lactamase in complex with cephalothin. There are three serine residues in the binding site so there is a chance that the ligand may not bind to the correct serine as in the native structure. In our result, the highest ranking conformation is highly identical with the original structure (covalent linkage formed on Ser70 with RMSD of 0.69Å), while in the lower-ranking conformations, covalent linkages are also found with alternative serines (Ser237 and Ser130), as shown in Figure 14(d). There are some other examples where multiple target residues are found in the binding site in the data set. Our docking results always give the best ranking to the results linked to the proper residue as in the native structures, suggesting the capability of our
method to search the correct receptor amino acid regardless of the existence of other nucleophilic residues.

**Case Study: Molecular geometry**

2ZC3\(^{44}\) is the PBP-2X acyl-enzyme complex with biapenem from Streptococcus pneumoniae. In this case, the default covalent docking method in Autodock results better RMSD than CovalentDock (with RMSD of 1.76Å and 2.25Å respectively). However, it is found that the result given by Autodock achieved this result by compromising proper molecular geometry of the covalent linkage. As shown in Figure 14(e), the Autodock produced a docking mode with a heavily distorted covalent linkage between the ligand and the receptor with bond length of 1.95Å and bond angle (equivalent to angle \(\theta\) in Figure 3) of 165.9°, while the corresponding bond length and angle in native structure are 1.37Å and 122.4° respectively. Meanwhile, the result given by CovalentDock satisfies the molecular geometry of the covalent linkage very well, with bond length of 1.38Å and angle of 128.7°. In many other cases where Autodock achieved a better result in terms of overall RMSD, the distorted covalent linkages are also observed. We suspect that it is caused by the relatively wide “Gaussian well” on the default parameters so that the covalent linkage geometry constrains can be compromised and trade-off for overall better conformation. This strategy may work well when the covalent bond formation only contribute to a small portion of the total binding affinity so that the contribution on other parts dominate the final score. However, it does not apply for many covalent drugs in which the covalent bond formation contributes significantly to the binding affinity.\(^{11,14}\) In this sense, we do not see any advantage in this strategy of default covalent docking method in Autodock over CovalentDock to compromise the molecular geometry of the covalent linkage.

**Case study: Kinase efficiency**

2HWO\(^{45}\) is the Crystal structure of Src kinase domain in complex with a covalent inhibitor. The ligand is a typical 4-anilinequinazoline with a Michael acceptor. In this case, acceptable result was given by CovalentDock, with a RMSD of 1.27Å. What’s interesting is that GOLD gave even better result on this particular complex with an RMSD of 0.66Å. A closer look at the covalent linkages formed by CovalentDock and GOLD show that both methods enforced the molecular geometry constrains properly, as shown in Figure 14(f). Thus the bias in the result RMSD is not contributed by the covalent linkage, but rather by the other parts adopting conventional non-
covalent docking. Considering that the scoring function in GOLD was specially optimized for some kinds of targets such as kinases and heme-containing proteins\textsuperscript{46}, it is not to our surprise that it can achieve good results on these targets. In more general cases where the target proteins are not the specially treated, as shown in Figure 8, this advantage of GOLD is no long there.

**AVAILABILITY**

The CovalentDock package, including the source files, binary executables, preparation scripts and brief instructions to use are publicly available on \url{http://code.google.com/p/covalentdock/}.

**CONCLUSION**

In this paper, we presented a formulation to estimate the energy contribution of covalent linkage formation in protein-ligand binding. And a new software package, CovalentDock and its accompanying scripts are developed to allow convenient and accurate covalent docking. The performance of the CovalentDock outperforms the default covalent docking method of Autodock and GOLD in terms of better structural agreement of the result compared to the native structures in PDB, as well as the ranking for near native conformations. CovalentDock also outperforms Autodock in virtual screening test to retrieve the true active controls from a large library of decoys. In addition, some findings in the docking result of CovalentDock, such as the chirality selectivity observed in 2AX0, alternative configurations observed in 2E14 and the target selectivity of the β-lactam in 1IYP are also supported by existing scientific observations\textsuperscript{40-43}. Compared to the original Autodock covalent docking, molecular geometry can be well preserved by CovalentDock.

For the future work, we will explore better tuning of the energy estimation model as more knowledge of the structure and binding energy for covalent binding become available. The possibility to couple the CovalentDock with some other non-covalent docking scoring functions are another direction to explore. The parameterization will be done and made available for more covalent binding structure patterns. We are also interested in bring side-chain flexibility into CovalentDock in the future, especially the flexibility on covalent-bond-forming residues, which is expected to further boost the performance of covalent docking.
ACKNOWLEDGEMENT

The authors would like to thank Stephanus Daniel Handoko and Mostafa Mostafa Hashim Ellabaan from Centre for Computational Intelligence, Nanyang Technological University for delightful discussion and help in this project. This work was supported by Singapore MOE AcRF Grant (No. MOE2008-T2-1-074) and the National Natural Science Foundation of China (No. 21172011).

References

19. Morse, P. M. Physical Review 1929, 34(1), 57-64.
CovalentDock: Automated covalent docking with parameterized covalent linkage energy estimation and molecular geometry constrains

Xuchang Ouyang‡*, Shuo Zhou†*, Chinh Tran To Su†, Zemei Ge†, Runtao Li†, Chee Keong Kwoh‡

† State Key Laboratory of Natural and Biomimetic Drugs, School of Pharmaceutical Sciences, Peking University, Beijing, China 100191
‡ BioInformatics Research Centre, School of Computer Engineering, Nanyang Technological University, Singapore 639798
* These authors contributed equally to this work.
Correspondence to: Chee Keong Kwoh (asckkkwob@ntu.edu.sg ) and Runtao Lirtirt@bjmu.edu.cn)

ABSTRACT

Covalent linkage formation is a very important mechanism for many covalent drugs to work. However, partly due to the limitations of proper computational tools for covalent docking, most covalent drugs are not discovered systematically. In this paper, we present a new covalent docking package, the CovalentDock, built on the top of the source code of Autodock. We developed an empirical model of free energy change estimation for covalent linkage formation, which is compatible with existing scoring functions used in docking, while handling the molecular geometry constrains of the covalent linkage with special atom types and directional grid maps. Integrated preparation scripts are also written for the automation of the whole covalent docking workflow. The result tested on existing crystal structures with covalent linkage shows that CovalentDock can reproduce the native covalent complexes with significant improved accuracy when compared to the default covalent docking method in Autodock. Experiments also suggest that CovalentDock is capable of covalent virtual screening with satisfactory enrichment performance. In addition, the investigation on the results also shows that the chirality and target selectivity along with the molecular geometry constrains are well preserved by CovalentDock, showing great capability of this method in the application for covalent drug discovery.

Keywords

covalent docking; covalent drug; molecular docking; virtual screening; computer-aided drug design
INTRODUCTION

Molecular docking is a computational procedure to predict the binding conformation as well as the binding affinity of a complex of two molecules, typically one receptor and one ligand.\textsuperscript{1-3} Well-known docking programs and software include Autodock\textsuperscript{4-6}, Autodock Vina\textsuperscript{7}, GOLD\textsuperscript{8,9} and FlexX\textsuperscript{10}. However, these and many other methods mainly focus on the docking between the two molecules through non-covalent interactions, such as van der Waals interaction, the electrostatics interaction and hydrogen bonding, or using other empirical or knowledge-based scoring functions to characterize these non-covalent interactions.\textsuperscript{3,5-7,9} The focus on non-covalent interactions is quite understandable since the main stream of rational drug design relies on these non-covalent interactions as the mechanism of the functionality of the drugs.\textsuperscript{11,12}

However, another important category of drugs, namely the covalent drugs, adopts covalent bond formation as part of its binding mechanism. Distinct from conventional non-covalent drugs, the covalent drugs usually have much stronger binding affinity with the targets because of the covalent linkage formed in between.\textsuperscript{11,12} This higher binding affinity of typical covalent drugs brings a risk of toxicity due to the difficulty of disassociation if off-target binding happens. Thus highly specified selectivity profiles of the covalent drugs are required.\textsuperscript{13} However, this characteristic also enables covalent drug to have a stronger potency while maintaining a pharmaceutically favored small molecule size, which makes the covalent drugs quite efficient and widely used.\textsuperscript{14} In fact, many famous drugs, including aspirin and penicillin, are covalent drugs.\textsuperscript{11} And it is reported that three of the ten top-selling drugs in U.S. in 2009 are also covalent drugs\textsuperscript{11}, suggesting a great potential in this area.

Despite the popularity of covalent drugs, however, most of them are discovered by serendipity instead of systematical design.\textsuperscript{11} Nowadays, with advanced and specialized computational tools, it is possible to design and simulate covalent drugs \textit{in silico}, and to perform large scale investigation on different target proteins to get the selectivity profile of the designed covalent drug in order to reduce of risk of toxicity, thus offering a new perspective to covalent drug design. However, the existing computational tools do not account for this kind of important chemical event very well. For example, two of the most popular molecular docking packages, Autodock and GOLD, do support “covalent docking” feature, but with major limitations in their functionality, accuracy and usefulness: Autodock uses a naive “Gaussian well” on the score to
reward the covalent bonding atoms. However, the “Gaussian well” must be set to a specified coordinate instead of on any target link atoms. The over-simplification of the scoring function as well as the lack of established guideline in parameter setting also compromises its performance. GOLD uses a “link atom” added to both ligand and receptor structure to ensure the geometry of covalent docking. However, the scoring functions are modified drastically to reduce clash penalty but not to reward the linkage formation over non-covalent docking, and the covalent bonding is “forced” to happen despite the overall docking formation. Moreover, since most of the molecular docking programs are designed for non-covalent docking, preparation steps for covalent docking requires much effort, and it is even more difficult to perform convenient virtual screening on covalent drug library due to the lack of automation in setting up the experiment. This gap between the need and practical difficulty in computational tools for covalent docking leads to the research of this project.

In this project, we developed a model to account for the energy contribution from covalent binding as a term compatible to the scoring function of non-covalent docking. We implemented this method as CovalentDock based on the source code of Autodock. Inspired by both Autodock and GOLD, we devised a method that utilizes a dummy atom to serve the similar purpose as the “link atom” in GOLD, integrating the proposed energy term into the scoring function of Autodock, and use the built-in grid map calculation and Genetic Algorithm to search for the docking results. Meanwhile, we developed a procedure to automatically recognize and prepare the covalently bondable chemical groups, enabling large scale automated covalent docking and covalent virtual screening. This implementation was tested on 76 covalently bound complexes collected from the Protein Data Bank (PDB). The results show that our method is able to reproduce the conformations in native PDB structures with an average RMSD=1.68Å, outperforming result given by the default covalent docking method of Autodock with average RMSD=2.49Å and of GOLD with average RMSD=3.69Å on the same data set. Results of the covalent virtual screening test also suggest that our method has a better capability to enrich true active covalent binders from randomly selected decoys compared to the default covalent docking method of Autodock. In addition, the investigation on the results shows that the chirality and target selectivity of the covalent drugs, along with the molecular geometry constrains of the covalent linkages, are well preserved by CovalentDock, showing great capability of this method.
METHOD

Overall workflow

As a special version of the docking package featuring covalent docking, the workflow is fundamentally similar to the conventional molecular docking. Similar to the procedures in conventional non-covalent docking, preprocessing of the structures is needed to make them suitable for covalent docking. The preparation work needed for covalent docking in preprocessing starts with identifying the bond forming groups. Then the processed structures will be docked with covalent linkage expected. Finally the result can be examined and compared with non-covalent docking to identify possible covalent binding. The whole procedure is shown in Figure 1.

Modeling the contribution of covalent bond formation

The change of Gibbs free energy of a chemical system is given by:

\[ \Delta G = \Delta H - T \Delta S \quad \text{...(1)} \]

where \( \Delta H \) is the enthalpy change, \( T \) is the temperature of the system and \( \Delta S \) is the entropy change.

Consider a complex of two molecules at a constant temperature, forming a covalent bond between them will change the free energy of the system. The change in free energy comes from two sources:

1. Formation of the covalent bond between the two molecules, sometimes coupled with bond breaking as well, gives a change in enthalpy. This enthalpy change can be estimated by existing chemistry references\textsuperscript{17,18}. However, considering that the standard enthalpy change only provides an estimation on the optimal bond length, while in molecular docking, that length might not be perfectly optimal. Thus, we formulated the Morse potential\textsuperscript{19} with parameters fitting to simulated result from Gaussian\textsuperscript{20} to estimate the enthalpy change brought by covalent bond formation on sub-optimal bond length. The enthalpy change of a given covalent bonding pair given by the Morse potential is:

\[ \Delta H = D \left( e^{-2a(r-r_0)} - 2e^{-a(r-r_0)} \right) \quad \text{...(2)} \]
where D is the dissociation energy, α is a parameter controlling the well width, r is the bond length and \( r_0 \) is equilibrium bond length. The parameterization of this equation is subjected to the specified type of covalent bonding pairs.

2. Forming such a covalent bond will also influence the conformation of the two molecules, putting some additional constrains on both structures to satisfy the molecular geometry, and bringing a conformational entropy change as well. This part of energy contribution can also be estimated with simulated result from Gaussian\(^{20}\).

With these two components estimated, the energy change ∆G of forming a covalent bond could be calculated from the conformation of the complex during docking.

In order to make the energy estimation compatible with the scoring function in the docking program used to estimate the entropy change, an additional empirical term is also included to correct the difference between the predicted energy score and empirical data. The overall additional energy term for scoring functions is:

\[
E = \begin{cases} 
D \left( e^{-2\alpha(r-r_0)} - 2e^{-\alpha(r-r_0)} \right) - T\Delta S_{\text{est}} + C, & r \leq r_m \\
0, & r > r_m 
\end{cases} \quad \cdots (3)
\]

where \( r_m \) is the maximum bond length without disassociation (which is defined as the distance when \( E = 0 \) when \( r > r_0 \)), \( \Delta S_{\text{est}} \) is the conformation entropy estimated by Gaussian, C is a correcting empirical constant. An example of this energy term with parameters determined for β-lactam family is shown in Figure 2.

As a term in the scoring function, this energy change appears as a well, with maximum reward at the optimal bond length, rising sharply when the distance approaches to zero to penalty this tendency, and smoothing to zero when distance goes too large. Since the energy term has already been corrected with an empirical constant according to the original scoring function, it could be directly embedded to the corresponding docking program.

**Implementation of the proposed model in Autodock**

Autodock\(^4\)\(^-6\) is one of the most popular and effective molecular docking packages\(^2\). The software package includes two major components: autogrid and autodock (to discriminate this component from the whole package, Autodock with a capital initial is used to refer to whole package). The molecular docking is carried out in a two-step manner: autogrid pre-calculates the energy grid
maps for each atom type in the ligand, later *autodock* utilizes the grid maps to estimate the energy of the complex conformation during docking by interpolation.

In order to use the docking scheme of Autodock to perform covalent docking, the *autogrid* is modified to take two additional atom types: the atom type R in receptor and L in ligand, as shown in Figure 3, where R and L are the atom pairs identified that the covalent bond could be formed. The atom type R and L will be treated as normal atom types when interacting with other atom, while the interaction between R and L will be treated specially with the model described in the previous section. Therefore, the docking of the two molecules could be considered as a complex with a “conceptual” covalent linkage between R and L. Unlike the implementation in *GOLD*\(^8,9\), this docking procedure does not force R and L to end up with a favorable position in order to form a covalent linkage, but instead it takes the all contribution between the ligand and the receptor to produce the final result conformation.

Some special tweaks are introduced to ensure the optimal geometry of the covalently docked complex is obtained. As shown by the example of Michael addition in Figure 3, two bond angles \(\gamma\) and \(\theta\) at the two ends of the “conceptual” covalent linkage should have correct bending in the complex. During the structural preparation steps, which will be detailed in the following section, a dummy atom will be attached to the bond forming atom L in the ligand. The structural optimization during that process will ensure angle \(\gamma'\) is arranged as in Figure 3 to satisfy proper molecular geometry constraints. This dummy atom will end up with approaching R, should the covalent bond formed, giving an angle \(\gamma\) approaching \(\gamma'\). Meanwhile, angle \(\theta\) is rewarded to the optimal angle by taking the energy potential between R and L multiplied with a directional term, which is the same term used to ensure the directionality of H-bond in Autodock\(^6\). The covalent bond forming atom L with angle \(\theta\) will have an energy potential:

\[
E_{\text{angle}} = E_{\text{cov}} \times \cos^4(\theta - \theta_0) \quad \ldots(4)
\]

where \(E_{\text{cov}}\) is the covalent binding free energy estimated with Equation(3), \(\theta_0\) is the optimal angle bending.

**Automated structure preparation with chirality consideration**

For practical use of the covalent docking, the structures, especially the structures of the ligand, are presented commonly to the docking program in an “intact form”, i.e. the form before
covalently bound to the receptor. During covalent docking, formation of the covalent linkage would cause structural changes in both molecules, especially in the ligand. An example is shown in Figure 4, where (a) and (c) are the structure of the same ligand before and after the covalent linkage formed with the receptor respectively. In order to adapt these structural changes, the structure of both the receptor and the ligand need to be carefully prepared with automated procedures. The potential bond-forming chemical groups and specific atom pair in both receptor and ligand are first identified. For the receptor protein, it mainly involves locating nucleophilic terminals within the binding site, such as the oxygen on serine residues and sulfur on cysteine residues. For the ligand, the main task is to recognize special patterns of electrophilic chemical groups based on the topology of the compound. When the bond forming groups have been identified, it is necessary to alter the structures in order to perform covalent docking. While the changes in the receptor structure are minor, which usually involves just losing a hydrogen atom, the changes in the ligand structure can be relatively significant. For example, the basic mechanism of penicillin is breaking the four-membered β-lactam ring between the nitrogen and carbonyl carbon, followed by carbonyl carbon forming a covalent linkage to the serine on the receptor in the active site.\textsuperscript{21,22} The structure altering of the penicillin starts with an intact β-lactam ring, as shown in Figure 4(a). Then the bond between C1 and N on the β-lactam ring is broken. A hydrogen atom is transferred from the serine nucleophilic oxygen to the N. Since C1 now has an available valence, it is able to form a covalent bond with the receptor. A “dummy” oxygen atom is artificially attached to C1 to temporarily occupy the empty valence of C1, as shown in Figure 4(b). Now the topology of the structure becomes identical with the covalently bonded ligand (with one extra “dummy atom” substituting the corresponding bond-forming atom in receptor), but its molecular geometry is not optimal. A short structural optimization is then performed for this structure using Amber\textsuperscript{10} with force field ff99sb and GAFF, to obtain a final optimized structure ready to perform covalent molecular docking, as shown in Figure 4(c).

During this step, chirality is carefully considered and handled. As shown in Figure 5, after covalently bound to the receptor, two new chiral centers are generated on the ligand and there will be 4 different possible configurations. Without prior knowledge, all four stereoisomers are passed to covalent docking. The one with highest score in result is regarded as the correct configuration.
In the package of CovalentDock, this preparation procedure is automated with integrated scripts. It therefore enables large scale automated virtual screening without human interference. Do note that the protocol we offered relies on the molecular dynamic package Amber $^{10,23}$ for structure optimization. However, the covalent docking method is also compatible with structures processed by other means, or directly obtained from existing structure databases. However, the chirality should still be considered.

**RESULT AND DISCUSSION**

The CovalentDock package for covalent docking is implemented based on the source code of Autodock 4.2. We tested this package on a dataset of 76 complexes discussed in literature, all of them are experimentally observed with covalent linkage formed between the ligand and the receptor. The structures are collected from the Protein Data Bank $^{16}$. For each of the complex, the corresponding original ligand structure before covalent bond formed is prepared using PyMol $^{24}$. The covalently bound ligand structure is deleted from the complex, leaving the receptor structure for our covalent docking experiment.

Among the 76 complexes, 13 are with Michael acceptors, and the rest are in β-lactam family. The basic chemical reactions of the covalent binding for both Michael acceptor and β-lactam family along with corresponding atom labeling are shown in Figure 6.

Parameters used in the energy estimation, as described in Equation(3), are listed in Table 1. For Michael additions, since the binding affinity is directly available in literature, the constant C in Equation(3) is estimated empirically to correct the final binding affinity score, while it is set to zero for β-lactams due to the lack of experimentally measured binding affinity.

**Native structure reproduction**

The capability of CovalentDock to reproduce the native ligand structure is tested against the covalent docking protocols in original Autodock and GOLD. The experiment is carried out by feeding all 76 experimentally determined complexes with covalent linkages to the docking programs for the prediction of docking conformations. The main criterion is the minimum Root-Mean-Square Deviation (RMSD) of the heavy atoms (all atoms except for hydrogen) between the result and the native structure achieved on each complex over 10 repeated but independent docking experiments, as defaulted in Autodock and GOLD.
The results show that CovalentDock can reproduce the covalently bound native structure very accurately, if the correct chirality is chosen. The 76 complexes have an overall average minimum RMSD=1.68Å, with an average minimum RMSD= 1.73Å and 1.67Å for Michael additions and β-lactams, respectively.

This result is compared with the result produced by the default covalent docking method in Autodock. With the Gaussian well set to be centered at the position of the corresponding bond forming atom in receptor, and both well width and depth set as default values, Autodock obtained a result with average RMSD of 2.49Å for all the complexes tested. The comparative result is shown in Figure 7. The performance of CovalentDock is also compared with GOLD, benchmarked on the same data set with default covalent docking protocol. The result shows that GOLD doesn’t perform very well on most complexes, given an average minimum RMSD of 3.69Å, as shown in Figure 8. However, detailed analysis suggests that GOLD performs exceptionally well on kinases, which will be elaborated in one of the following case studies.

For the capability to reproduce the native structure, another criterion used is the “hit rate”, calculated as the percentage of the complexes with a minimum RMSD within a pre-defined cutoff. The hit rates for CovalentDock, Autodock and GOLD with different cutoff RMSD are shown in Figure 9. The result shows that for all given cutoff RMSD, as exemplified with 1Å, 2Å and 3Å, CovalentDock is more likely to produce “hits” than the other two covalent docking protocol. Even with the cutoff RMSD set to 1Å, which will only consider the results highly similar to the native structure as hits, CovalentDock still gives a hit rate at 30.26%, while Autodock and GOLD produced “hit” on 17.11% and 6.58% on the same dataset of 76 complexes.

**Ranking of the near-native structures**

Ideally, molecular docking programs, either covalent or non-covalent, should always put the “correct conformation” on top ranking positions. However, with imperfections in the scoring functions in use, it is found that the most near-native structures (the ones with minimum RMSD) are not always properly ranked by the docking methods to-date.25-27 We examined the capability to rank the near-native structures of CovalentDock and compared with the covalent docking method in Autodock and GOLD as well. For this test, the number of repeated docking experiment is increased to 100 for each complex in our dataset.
For each complex in the dataset, the ranking of the conformation with minimum RMSD is recorded among all the 100 conformations. On average, the result structures with minimum RMSD are found at top 30.22, 41.04, and 63.78 ranking positions for CovalentDock, Autodock and GOLD, respectively. Although ranking is not the primary goal in the design of the energy estimation scheme, by implementing this energy estimation in the covalent docking protocol, the ranking performance of CovalentDock is also an evidence of its superiority.

We also examined the minimum RMSD and hit rate for these three covalent docking protocols when only a given number of top ranking conformations are considered. Figure 10 shows the average minimum RMSD that can be found in the result for each complex among a given number of top ranking conformations, and Figure 11 plots the corresponding hit rate achieved by CovalentDock, Autodock and GOLD, with a RMSD cutoff set at 2Å. The result clearly shows that CovalentDock is with an advantage in terms of both RMSD and hit rate, regardless the acceptance cutoff set in rank.

**Covalent Virtual Screening**

The automated preparation procedure we developed enables CovalentDock to perform large scale virtual screening on covalent drug candidates. To illustrate this capability, we performed a virtual screening test on Michael acceptors against the target EGFR kinase domain.

The receptor structure used in this virtual screening test is extracted from the PDB complex 1XKK. The terminal sulfur atom on Cys797 is identified as the covalent binding site. The library used in this virtual screening test comes from the ZINC database, using the “Lead-Like” subset containing 5,037,475 chemical compounds, which is prepared, filtered and recommended by ZINC for discovery projects. Open Babel is used to add missing hydrogen to the structure. All the compounds with Michael acceptors are identified from the raw library using Open Babel with the SMARTS pattern “[$([CH1,CH2]=CC=O)]”, resulting 160,957 compounds in total ready for screening. For testing purpose, we randomly selected 19,949 compounds from all the Michael acceptor containing compounds, which will be considered as decoys in this test, although it is understood that some of them may be potential active leads. On the other hand, 51 active ligands are collected from literature. As a result, the library we screened has 20,000 compounds in total, with 51 positive controls.
The virtual screening experiment is carried out by performing a quick covalent docking using CovalentDock. The preparation scripts will preprocess the structure and prepare relevant configuration files to enable automated testing. For comparison, the same virtual screening test is also carried out using the default covalent docking protocol in Autodock. GOLD is left out from this test due to the difficulty of its automation because the structures and configuration files prepared by the scripts developed for CovalentDock are not compatible with GOLD and cannot be easily translated.

The Enrichment Factor (EF) is used as a performance indicator for the virtual screening tests. The enrichment factor at the top p% is calculated with:

\[ \text{EF}_{p\%} = \frac{H_{p\%}}{H_{\text{total}} \cdot p\%} \]

where \( H_{p\%} \) is the number of positive controls compounds found in the top p% ranking results among all the compounds screened, and \( H_{\text{total}} \) is the total number of positive controls, which will be 51 in this virtual screening test. In addition, the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) is also examined. The results are shown in Figure 12 and Figure 13, respectively. Note that the vertical axis of Figure 12 is plotted in the logarithmic scale.

Figure 12 shows that for any given value of p%, CovalentDock gives better EF compared to the default covalent docking methods in Autodock. For example, the EFs calculated at top 5% are 17.25 and 4.31 for CovalentDock and Autodock, respectively, meaning that it is much more likely to find positive controls in the top 5% ranking compounds by using CovalentDock than using Autodock. The ROC plotted in Figure 13 shows that CovalentDock achieved an AUC of 0.97, while the AUC for the Autodock ROC is 0.68. Both performance indicators show that the covalent docking method we proposed as CovalentDock outperforms the default docking method in Autodock in virtual screening test with a significant advantage.

**Case Study: Chirality selectivity**

2AX0\(^{40}\) is the complex of a novel non-nucleotide NS5b RNA polymerase inhibitor and its receptor. The ligand has a Michael acceptor inside and in the structure a covalent bond is formed between the beta carbon of the Michael acceptor and the sulfur atom of Cys366. As a result, the alpha and beta carbon of the Michael acceptor become chiral, and the ligand adopts the alpha(S)-beta(R) configuration. In our experiments, all the four possible configurations of the ligand after bonding are investigated. As expected, the top ranking conformation adopts the alpha(S)-beta(R) configuration. The RMSD of the result is 1.24Å while all the important lipophilic interactions
and hydrogen bonds in the original complex are preserved, as shown in Figure 14(a). However, the docking modes of the three alternative configurations do not agree with the native structure at all, one example of the binding mode of alpha(S)–beta(S) configuration is also shown in Figure 14(a). This selectivity in chirality of the ligand has also been observed in many other complexes, and all of the top ranking conformations are observed to adopt the correct configuration, expect the complex 2E14 which will be discussed in next case study.

Case Study: Chirality alert

The complexes of 3C9W$^{41}$ and 2E14$^{42}$ are ERK with hypothemycin and its derivative FR148083 as their ligands. The only difference between them is that FR148083 is an analog with a double bond in place of the epoxide, thus we could expect that their docking mode will be very similar. Good results are given by CovalentDock for both complexes, with RMSD of 0.98Å and 1.26Å respectively, and both the covalent linkage between the ligand and receptor are in the R configurations, as shown in Figure 14(b) and Figure 14(c), respectively. While the linkage in the native structure of 3C9W indeed adopted the R configuration,$^{41}$ however, this prediction conflicts with the native structure of 2E14, which adopts the S configuration.$^{42}$ It has already been discussed by Rastelli et al.$^{41}$ that R configuration was in better agreement with both the electron density and the expected tetrahedral geometry around the carbon, while the geometry of FR148083 bond to ERK as shown in 2E14 with S configuration has distorted bond lengths, angles and torsions. With this supporting statement, it is suggested that CovalentDock could give a reasonable prediction of covalent complex formation in which covalent linkage is formed in more favorable configuration.

Case Study: Target selectivity

1IYP$^{43}$ is the Toho-1 β-lactamase in complex with cephalothin. There are three serine residues in the binding site so there is a chance that the ligand may not bind to the correct serine as in the native structure. In our result, the highest ranking conformation is highly identical with the original structure (covalent linkage formed on Ser70 with RMSD of 0.69Å), while in the lower-ranking conformations, covalent linkages are also found with alternative serines (Ser237 and Ser130), as shown in Figure 14(d). There are some other examples where multiple target residues are found in the binding site in the data set. Our docking results always give the best ranking to the results linked to the proper residue as in the native structures, suggesting the capability of our
method to search the correct receptor amino acid regardless of the existence of other nucleophilic residues.

**Case Study: Molecular geometry**

2ZC3 \(^{44}\) is the PBP-2X acyl-enzyme complex with biapenem from Streptococcus pneumoniae. In this case, the default covalent docking method in Autodock results better RMSD than CovalentDock (with RMSD of 1.76Å and 2.25Å respectively). However, it is found that the result given by Autodock achieved this result by compromising proper molecular geometry of the covalent linkage. As shown in Figure 14(e), the Autodock produced a docking mode with a heavily distorted covalent linkage between the ligand and the receptor with bond length of 1.95Å and bond angle (equivalent to angle \(\theta\) in Figure 3) of 165.9°, while the corresponding bond length and angle in native structure are 1.37Å and 122.4° respectively. Meanwhile, the result given by CovalentDock satisfies the molecular geometry of the covalent linkage very well, with bond length of 1.38Å and angle of 128.7°. In many other cases where Autodock achieved a better result in terms of overall RMSD, the distorted covalent linkages are also observed. We suspect that it is caused by the relatively wide “Gaussian well” on the default parameters so that the covalent linkage geometry constrains can be compromised and trade-off for overall better conformation. This strategy may work well when the covalent bond formation only contribute to a small portion of the total binding affinity so that the contribution on other parts dominate the final score. However, it does not apply for many covalent drugs in which the covalent bond formation contributes significantly to the binding affinity.\(^{11,14}\) In this sense, we do not see any advantage in this strategy of default covalent docking method in Autodock over CovalentDock to compromise the molecular geometry of the covalent linkage.

**Case study: Kinase efficiency**

2HWO \(^{45}\) is the Crystal structure of Src kinase domain in complex with a covalent inhibitor. The ligand is a typical 4-anilinequinazoline with a Michael acceptor. In this case, acceptable result was given by CovalentDock, with a RMSD of 1.27Å. What’s interesting is that GOLD gave even better result on this particular complex with an RMSD of 0.66Å. A closer look at the covalent linkages formed by CovalentDock and GOLD show that both methods enforced the molecular geometry constrains properly, as shown in Figure 14(f). Thus the bias in the result RMSD is not contributed by the covalent linkage, but rather by the other parts adopting conventional non-
covalent docking. Considering that the scoring function in GOLD was specially optimized for some kinds of targets such as kinases and heme-containing proteins, it is not to our surprise that it can achieve good results on these targets. In more general cases where the target proteins are not the specially treated, as shown in Figure 8, this advantage of GOLD is no long there.

AVAILABILITY

The CovalentDock package, including the source files, binary executables, preparation scripts and brief instructions to use are publicly available on http://code.google.com/p/covalentdock/.

CONCLUSION

In this paper, we presented a formulation to estimate the energy contribution of covalent linkage formation in protein-ligand binding. And a new software package, CovalentDock and its accompanying scripts are developed to allow convenient and accurate covalent docking. The performance of the CovalentDock outperforms the default covalent docking method of Autodock and GOLD in terms of better structural agreement of the result compared to the native structures in PDB, as well as the ranking for near native conformations. CovalentDock also outperforms Autodock in virtual screening test to retrieve the true active controls from a large library of decoys. In addition, some findings in the docking result of CovalentDock, such as the chirality selectivity observed in 2AX0, alternative configurations observed in 2E14 and the target selectivity of the β-lactam in 1IYP are also supported by existing scientific observations. Compared to the original Autodock covalent docking, molecular geometry can be well preserved by CovalentDock.

For the future work, we will explore better tuning of the energy estimation model as more knowledge of the structure and binding energy for covalent binding become available. The possibility to couple the CovalentDock with some other non-covalent docking scoring functions are another direction to explore. The parameterization will be done and made available for more covalent binding structure patterns. We are also interested in bring side-chain flexibility into CovalentDock in the future, especially the flexibility on covalent-bond-forming residues, which is expected to further boost the performance of covalent docking.
ACKNOWLEDGEMENT

The authors would like to thank Stephanus Daniel Handoko and Mostafa Mostafa Hashim Ellabaan from Centre for Computational Intelligence, Nanyang Technological University for delightful discussion and help in this project. This work was supported by Singapore MOE AcRF Grant (No. MOE2008-T2-1-074) and the National Natural Science Foundation of China (No. 21172011)

References

19. Morse, P. M. Physical Review 1929, 34(1), 57-64.
Overall work flow for covalent docking.

85x48mm (300 x 300 DPI)
Estimated free energy change against bond length. Parameterized for β-lactam family.

85x72mm (300 x 300 DPI)
Special atom type labeling on the covalent bond forming atom pairs and molecular geometry constrains. (R, L) are the bond-forming pairs with R in receptor and L in ligand. Dum is a dummy atom added to the ligand during preprocessing. θ and γ are the angles on the two ends of the bond between R and L, and γ' is the angle when R is conceptually substituted by the dummy atom.

85x68mm (300 x 300 DPI)
Illustration of the structure preparation of a penicillin molecule. (a) An intact penicillin molecule with the β-lactam ring identified and labeled. (b) Breaking the β-lactam ring, with temporary dummy atoms. But the structure is not minimized so that the newly added temporary atoms do not satisfy molecular geometry. (c) After minimization, molecular geometry is satisfied.

85x55mm (300 x 300 DPI)
The original ligand (a) of 2AX1 and four possible configurations (b-e) after bond formation. The yellow atom is the "link atom" on the cysteine of the receptor. The configurations on the same row are enantiomers.
Reactions of covalent binding. The bond forming atom pairs are labeled R and L for the atom in receptor and in ligand, respectively. The covalent bond formed between the receptor and ligand is illustrated with a dash line. (a) Reaction between the Michael acceptor group in ligand and Cysteine residue in receptor; (b) Reaction between the β-lactam ring in ligand and Serine residue in receptor.
Results with minimum RMSD of each complex found by CovalentDock and Autodock covalent docking protocol.

297x420mm (300 x 300 DPI)
Results with minimum RMSD of each complex found by CovalentDock and GOLD covalent docking protocol.

297x420mm (300 x 300 DPI)
Hit rate achieved by CovalentDock, Autodock and GOLD on the dataset of 76 complexes, with different cutoff RMSD.

297x420mm (300 x 300 DPI)
Average minimum RMSD found among a pre-defined top ranking cutoff number.

297x420mm (300 x 300 DPI)
Hit rate achieved by considering only a pre-defined number of top ranking conformations. Results with RMSD less than 2Å are considered “hits” in this plot.
Enrichment Factor calculated with different p% of the covalent screening test done by CovalentDock and Autodock. Note the vertical axis is in the logarithmic scale.
297x420mm (300 x 300 DPI)
Receiver Operating Characteristic (ROC) of the covalent screening test. The AUC is shown in the legend.

297x420mm (300 x 300 DPI)
Covalent docking result structures for case studies. In all subfigures, the receptors are shown as cyan cartoon, the native ligand structures are shown as green sticks and docking results given by CovalentDock are shown as magenta sticks. Special alternative ligand structures discussed by each case study are shown as yellow sticks. (a) 2AX0 with a Michael acceptor in the ligand, the magenta ligand structure has the same configuration with the green native structure, while the yellow one adopts a different configuration. (b) 3C9W with a Michael acceptor in the ligand (c) 2E14 with a Michael acceptor in the ligand. The covalent linkage shown in the green native structure has distorted bond length (2.35Å) and angle (154.8°). (d) 1IYP with a β-lactam ring in the ligand and three serine residues located in the binding site. While the highest ranking result shown in magenta formed a linkage with the same serine (SER-70) of the green native ligand, covalent linkages with alternative serine residues are also found in lower ranking results. An example of a covalent linkage formed between the ligand and SER-237 are shown in yellow. (e) 2ZC3 with a β-lactam ring in the ligand, the magenta ligand structure agree with the green native structure very well around the covalent binding site, while the yellow structure (produced by Autodock) has a distorted molecular geometry on the covalent linkage. (f) 2HWO with Michael acceptor in the ligand. The geometry constrains on the covalent linkage are well satisfied. The yellow structure (produced by GOLD) has better alignment with the
green native structure than the magenta one, mainly due to the conformation of the non-covalent parts of
the ligand, although both yellow and magenta conformations are with acceptable accuracy.
<table>
<thead>
<tr>
<th></th>
<th>Michael addition</th>
<th>β-lactam</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>10.16</td>
<td>21.54</td>
</tr>
<tr>
<td>r₀</td>
<td>1.84</td>
<td>1.36</td>
</tr>
<tr>
<td>α</td>
<td>3.41</td>
<td>3.78</td>
</tr>
<tr>
<td>ΔSₑₓₜ</td>
<td>-9.36</td>
<td>-10.96</td>
</tr>
<tr>
<td>C</td>
<td>4.74</td>
<td>0.00</td>
</tr>
<tr>
<td>θ₀</td>
<td>100.83°</td>
<td>119.62°</td>
</tr>
</tbody>
</table>

Table 1. Parameter used for the Michael addition group and β-lactam group. r₀ and θ₀ are extracted from Amber Force Field parameters²³, D, α and ΔSₑₓₜ is estimated by Gaussian simulation under 298.15K.
Covalent linkage formation is a very important mechanism for many covalent drugs to work. However, the existing computational tools do not account for this kind of important chemical event very well. In this article, we present a new covalent docking package, the CovalentDock, with parameterized covalent linkage energy estimation and molecular geometry constrains for accurate and automated covalent docking.