3D-Garden: a system for modelling protein–protein complexes based on conformational refinement of ensembles generated with the marching cubes algorithm

Victor I. Lesk* and Michael J. E. Sternberg
Division of Molecular Biosciences, Imperial College London, South Kensington, London SW72AZ, UK

Received on November 27, 2007; revised on February 19, 2008; accepted on March 5, 2008
Advance Access publication March 7, 2008
Associate Editor: Burkhard Rost

ABSTRACT
Motivation: Reliable structural modelling of protein–protein complexes has widespread application, from drug design to advancing our knowledge of protein interactions and function. This work addresses three important issues in protein–protein docking: implementing backbone flexibility, incorporating prior indications from experiment and bioinformatics, and providing public access via a server. 3D-Garden (Global And Restrained Docking Exploration Nexus), our benchmarked and server-ready flexible docking system, allows sophisticated programming of surface patches by the user via a facet representation of the interactors’ molecular surfaces (generated with the marching cubes algorithm). Flexibility is implemented as a weighted exhaustive conformer search for each clashing pair of molecular branches in a set of 5000 models filtered from around ~340,000 initially.

Results: In a non-global assessment, carried out strictly according to the protocols for number of models considered and model quality of the Critical Assessment of Protein Interactions (CAPRI) experiment, over the widely-used Benchmark 2.0 of 84 complexes, 3D-Garden identifies a set of ten models containing an acceptable or better model in 29/45 test cases, including one with large conformational change. In 19/45 cases an acceptable or better model is ranked first or second out of 340,000 candidates.

Availability: http://www.sbg.bio.ic.ac.uk/3dgarden (server)
Contact: v.lesk@ic.ac.uk
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION
Interest in the modelling of protein complexes and their binding sites dates back to the very first research into protein structure. Computational protein–protein docking was first studied in 1978 (Wodak and Janin, 1978); recent advances in computational power and in the protein data bank deposition rate of both apo and complexed protein structures (Berman et al., 2000) have catalysed efforts in the field. Progress is scrutinized periodically in the Critical Assessment of Protein Interactions (CAPRI) experiment (Janin, 2005; Lensink et al., 2007), revealing a rich variety of approaches for generating initial ensembles of putative structures, for scoring, filtering and refinement (Halperin et al., 2002; Mendez et al., 2005).

Ensembles may be generated using exhaustive schemes, which sample evenly the six-dimensional space of models arising from the application of Euclidean transformations to one of the interactors while the other is held fixed. The most popular exhaustive scheme is the Fourier correlation technique (Katchalski-Katzir et al., 1992), which takes advantage of the fast Fourier transform algorithm. A correlation approach using spherical polar basis functions is also possible (Ritchie and Kemp, 2000). In contrast, guided schemes for ensemble generation, for example based on geometric hashing (Sandak et al., 1998) or a geometric prefilter for the potential to form hydrogen bonds (Meyer et al., 1996), use biased or restricted sampling of the six-dimensional space, with the aim of vastly reducing the number of models which need to be scored. The ensemble generation algorithm used by 3D-Garden (Global And Restrained Docking Exploration Nexus), based on facet-matching of the interactors’ molecular surfaces, belongs to this class.

The choice of scoring function is constrained by computational cost. While more complexes remain in the pool, coarse but computationally efficient scoring functions are used out of necessity; for example, only scores expressible as a convolution (or sums thereof) are compatible with the Fourier correlation method (Chen et al., 2003; Gabb et al., 1997; Mandell et al., 2001), in which billions of models can be processed. As the ensemble size is reduced through filtering, more sophisticated and computationally intensive scores become accessible: exact calculation of pairwise shape complementarity and physics inspired electrostatic energy terms (3D-Garden and many others), hydrogen bonding potentials (Kortemme et al., 2003), statistical potentials for residue contacts (Moont et al., 1999), or for atomic contacts (Zhang et al., 1997), clustering coefficients (Kozakov et al., 2005) and external-data-driven scores originally devised for analysis of NMR output (Dominguez et al., 2003).

A small number of models may be subjected to the most intensive step, structural refinement. This may constitute nothing more than a fine-tuning of the rigid transformation which defines the model. In general, however, refinement

*To whom correspondence should be addressed.
involves some search of the internal degrees of freedom of each interactor, either of the atomic positions, as in classic minimization approaches (Li et al., 2003) or of the dihedral angles. Refinement based on dihedral angles is usually restricted to sidechains, in which case rotamer libraries (Dunbrack and Cohen, 1997; Tuffery et al., 1991) may be used to guide the search. Methods used include an iterative mean-field approximation to the ensemble of sidechain rotamers (Jackson et al., 1998) or Monte Carlo (Gray et al., 2003). In more recent work, backbone coordinates are candidates for refinement as well; some methods focus on identifying significant hinges (Schneidman-Duhovny et al., 2005a), others on refining and remodelling loops (Bastard et al., 2006; Wang et al., 2007). 3D-Garden refines the backbone fragments near to an N- or C-terminus.

The CAPRI experiment, started in 2001, has created standard criteria for what constitutes a good model of a protein complex. CAPRI’s core assessment protocol classifies each of a set of ten models as good, medium, acceptable or incorrect based on root mean square deviation (RMSD) of backbone heavy atoms from the reference structure, and on interface contacts (Mendez et al., 2003). CAPRI has been central in evaluating and disseminating advances in protein–protein docking, and acts as a point of reference for the docking community. Accordingly, 3D-Garden is assessed here under the protocol used in the CAPRI experiment. We also form a secondary test set drawn from past CAPRI experiment targets.

For those CAPRI targets where neither of the interactors undergoes conformational change of more than ~2Å (‘rigid body docking’), better quality models are almost always present after pooling the various predictors’ contributions. However it is the opposite class of interactions, those involving larger conformational change, which are often associated with unusual biomolecular mechanisms and functions. There is thus demand for ‘flexible docking’ to be brought within reach. To date, only a few docking software systems model any backbone flexibility (Verbitsky et al., 1999; Wang et al., 2007), let alone search conformational space effectively. 3D-Garden combines an ensemble generator based on the marching cubes algorithm (Lorensen and Cline, 1987) with a conformational refinement engine to form a comprehensive framework for flexible docking.

Our assessment of 3D-Garden is based on the non-redundant Benchmark 2.0 of 84 complexes (Mintseris et al., 2005). We restrict the surface searched on each interactor to a 100 Å² patch within the interface, centered around a point of close contact. We partition Benchmark 2.0 into a training set and a test set. A further test set is composed of 24 targets from past CAPRI assessments.

In over a third of Benchmark 2.0 test cases, 3D-Garden’s primary prediction from an ensemble of size ~340 000 is of acceptable or better quality. In the majority of test cases, at least one of 3D-Garden’s ten preferred models is acceptable or better. Results for the CAPRI test set are broadly similar. In general the few highly flexible cases in the benchmark remain poorly modelled, indicating a need for extensive tuning of 3D-Garden’s versatile refinement algorithm, and for a flexible benchmark of adequate size.

3D-Garden is available for academic use as a web server at http://www.sbg.bio.ic.ac.uk/~3dgarden; a tutorial is provided.

2 SYSTEM AND METHODS

3D-Garden is a protocol implemented in a suite of programmes for generating and assessing models for blind or non-blind protein docking. ‘Global and Restrained’ refers to 3D-Garden’s functionality for transforming information supplied by the user into a restriction of the searched area of molecular surface on one or both interactors. We will now describe in detail what happens in a 3D-Garden docking run and the benchmark we use for assessment.

Figure 1 shows a simplified operational flowchart of 3D-Garden. An initial ensemble is populated using surfaces constructed from the marching cubes method, taking account of any restrictions indicated by the user. This ensemble is scored on the fly using a re-weighted Lennard–Jones potential. The top 5000 structures are refined using an exhaustive conformational search of clashing molecular branches (parts of a molecule which can be disconnected by breaking only one bond), to identify the best conformers according to a refinement score, which is a weighted sum of the two Lennard–Jones terms and an electrostatic term.

2.1 Molecular surfaces

3D-Garden uses the molecular surface when constructing the initial ensemble to ensure that the interactors make at least one glancing contact, with the aim of reducing the number of models in the ensemble with substantial interpenetration. Viewed mathematically, the space searched is four-dimensional, rather than six-dimensional.

Historically, molecular surface construction algorithms have focused on the locus of the centre (‘solvent accessible surface’) or inward-pointing boundary (‘solvent excluded surface’) of a fictitious probe sphere as it rolls over atoms in the molecule (Connolly, 1983). Properties of the surface are derived and any singularities are dealt with; the output is post-processed into surface dots, curved patches or a polyhedral representation. A more recent algorithm dedicated to the triangular facet representation (Sanner et al., 1995) achieves large improvements in speed, but the hard-sphere approximation remains.
We try a different method, based on *marching cubes*, for triangulating the interactors' molecular surface. *Marching cubes* has no singularities, and allows us to encode a realistic density profile for each atom. *Marching cubes* requires as input

- a density function $\rho(\vec{r})$,
- an 'isovalue' of that density function identifying an 'isosurface' (the higher dimensional equivalent of a contour) to represent the molecular surface and
- the spacing and extent of a finite cuboidal grid containing the surface.

Our density function has the interpretation of distance from the molecular surface, which we construct by invoking an isovalue of $\rho = 0$:

$$
\rho(\vec{r}) = -d \ln \left( \sum_i \exp \left( -\frac{||\vec{r} - \vec{r}_i||}{a_i} \right) \right)
$$

where $a_i$ is the CHARMM22 hard-sphere radius for atom $i$ (MacKerrel et al., 1998) and $\vec{r}_i$ is the position vector of its centre. $d$ governs how steeply an atom falls off from its van der Waal's surface; our choice $d = 0.2$ Å gives a molecular surface which closely follows the parts of the atoms' van der Waals surfaces exposed to solvent while smoothing away interstices, internal channels and voids. We use a cubic lattice with parameter 0.6 Å; at this resolution the average facet area is about 0.08 Å², and the deviation between the true isosurface and the discretized construction is negligible. For a visualization of the output, see Figure 2.

### 2.2 User input

The user adjusts the computational effort through two parameters, active facet pair count and azimuthal steps per active facet pair (explained in next section). The product of these two is, naively, the number of rigid models to be generated and scored in what is 3D-Garden’s most computationally intensive stage; by comparison, the surface generation takes very little time, and refinement is also quicker with reasonable parameters. The active facet pair count nominated by the user is only an upper bound; the heuristic algorithm which selects active facets is designed to stop short of the nominated count in order to preserve the facets’ even distribution.

Where high-quality information exists restricting the location of the binding site on one or both interactors, 3D-Garden may be directed to confine its search to the indicated regions, which are input as follows:

- for each interactor, the user specifies a set of surface patches;
- the user specifies each patch by its area in Å² and an atom or residue (user choice) identifying its centre;
- models are generated using facets from the union of all the specified patches on each interactor and
- the user may invert their total selection on either or both interactors, switching the excluded and included regions.

Each patch is constructed by accumulating neighbours of already included facets on a breadth-first basis until the requested area is achieved. The average patch shape is thus circular; other shapes can be designed using multiple overlapping patches. Binding-site information is not compulsory; if none is provided, models will simply be generated using facets sampled from the entire surfaces of both interactors.

Other significant parameters which may be specified by the user are

- grid spacing $a$ and distance parameter $d$ for surface creation,
- how many models, if any, will be computationally refined,
- the angular resolution for conformational refinement and
- how many of models to output.

### 2.3 Ensemble generation

The initial ensemble is generated using a set of rigid transformations of the smaller interactor relative to the larger. Each transformation has the property of setting some pair of facets, one from each interactor, flush with each other (normals antiparallel) and with coinciding centroids. Since combining every pair of facets would lead to an impossibly large and highly redundant ensemble, the first step is to identify a subset of facets on each interactor which will be active, i.e. participate in contacting-facet pairs. The number of active facets requested on each interactor is an integer close to the square root of the total facet-pair count indicated by the user. The active facets are then selected evenly, according to an algorithm described in the next section, from the region of surface remaining after exclusion of any areas identified by the user as not being relevant to the interaction. All possible pairings of active facets are used to generate models. Each pairing gives rise to a family of models differing by regular rotations around the flush facet normal; how many is governed by the user's specification of the number of azimuthal steps per facet pair. The models are created in turn and scores are evaluated on the fly. For each model, the score and Euclidean transformation applied to the smaller interactor are written to disk, and the coordinates discarded. The whole process is illustrated schematically in Figure 3.

For the benchmark runs we nominate an active facet pair count of 10,000; the actual number of facet pairs produced by the sampling algorithm varies between complexes, with a mean of 8,400 and a SD of 500. This corresponds to ~90 active facets for each interactor, selected from the ~1200 facets which lie in the 100 Å² sampled region. Forty azimuthal rotation steps are used, leading to an average ensemble size of about 340,000 models.

### 2.4 Effect of search density and size on model quality

We examine how the quality of rigid models depends on how small an area the search is restricted to, and on how intensive a search is carried out, by restricting the search to a single patch on each interactor which includes the known binding site, and varying its size from 50 to 10,000 Å². We perform this scan on selected proteins, firstly while keeping the number of active facet pairs fixed, and then while...
maintaining a constant search resolution (by increasing the number of active facet pairs in proportion to the square of the patch size). In both cases we record the number of acceptable rigid models among those scoring in the top 5000 which would qualify for refinement in a full run of 3D-Garden. Typical results like those shown in Figure 4 (Supplementary material) for the small complex 1grn suggest that, while a more intensive search is rewarding, 3D-Garden’s ability to identify acceptable models is compromised for patch sizes more than a few hundred Å². This leads to adopting a patch size of 100 Å² and facet pair count of 10,000, to ensure that the pattern of 3D-Garden’s strengths and weaknesses is revealed; a more challenging setup is likely to be uninformative.

2.5 Selection of active facets

The first active facet is the lowest serial number facet from the region to be sampled. Starting from this root, facets are then traversed according to edge sharing, in breadth first order. Traversed facets are permanently excluded from the sampled region, until the total area of excluded facets is the initial area of the sampled region divided by the target number of active facets (the square root of the nominated active facet pair count). This area condition triggers the next facet traversed to be marked as active, at which point the area accumulator is set to zero, and the cumulative area to trigger the next active facet is re-calculated as the quotient of the remaining untraversed area and the number of requested active facets yet to be identified. The root for the breadth first search is set to the new active facet, and the traversal starts afresh. This continues until the target number of active facets have been identified, or until all facets have been traversed in which case fewer active facets than nominated will generally be returned.

It often happens, for example when the sampled region contains disconnected patches, that the breadth first traversal hits a dead end before either of the stopping conditions is met. In this case, the traversal cursor jumps to the lowest serial number untraversed facet, and continues on a breadth first basis; the new root facet is not otherwise treated specially and the area accumulator is not reset.

2.6 Scoring

The scoring function used for the initial ensemble is an all-atom modified Lennard–Jones potential with explicit hydrogens but no electrostatic term:

$$f = \sum_{ij} \epsilon_{ij} \cdot \left\{ \begin{array}{ll}
\frac{2}{R_{min,ij}^{12}} - 1.38 \times 2 \frac{R_{min,ij}^{6}}{R_{ij}^{6}} & \text{if } R_{ij} > 0.77 \times R_{min,ij} \\
9.78 & \text{otherwise}
\end{array} \right.$$

The well-depth parameters $\epsilon_{ij}$ and the atomic radii $R_{min,ij}$ are taken from the par_all22_prot.inp file of CHARMM22. For unattested atom types (e.g. in many heterogens), $\epsilon_{ij}$ and $R_{min,ij}$ are constructed by averaging over all atom types for the same chemical element, or, failing that, the lightest element in the same chemical group. The scoring function is continuous in $R_{ij}$ differing from the canonical Lennard–Jones potential only in the upweighting of the attractive term, and a flat region at inter-atomic separations smaller than 0.77 of the minimum hard-sphere separation $R_{min,ij}$.

The upweighting factor of 1.38 and the radius of the flat region are tuned using the ensembles of the training set. Our tuning procedure consists of a simple grid scan of parameter space; the ideal parameter vector is characterized by the highest number of training set complexes with acceptable, medium and high quality models as primary predictions and among the preferred 10 models.

The serial numbers of the 5000 highest scoring models are passed on to the refinement engine, which has its own scoring function and weight parameters.

2.7 Refinement

3D-Garden’s method for refining a model begins by identifying every atom in the model which clashes sterically ($R_{ij} > 0.77 \times R_{min,ij}$) with the other interactor. For each such clashing atom, 3D-Garden determines the smallest molecular branch containing a rotatable bond whose dihedral coordinate governs that clashing atom’s position. The other rotatable bonds within this branch are identified. Clashing atom pairs are then resolved in turn, by evaluating every joint conformation of the two relevant molecular branches (or of just one of them, if the other is too large). Each dihedral angle is searched in even steps. The angular resolution used for a given dihedral takes account of the bulk of atoms downstream of the corresponding bond: the torsion angles associated with rotatable bonds near the root of a branch are incremented finely, whereas the ends of branches undergo only coarse, if any, conformational exploration. Branches with more than 14 rotatable bonds or 500 atoms are kept fixed. A candidate branch for refinement must therefore

---

Figure 3. A low-resolution cartoon of the ensemble generation process. (i) Molecular surfaces for interactors A and B are constructed in the facet representation. (ii) Atom and residue masking information is translated into masked out facets (black); the active facets (white) are sampled from among the remaining facets (grey). (iii) Each pair of active facets $f_A, f_B$ are superposed. (iv) The superposition defines a family of $N$ structures generated by rotations of $\Delta \phi = 2\pi/N$ around the common normal, which are evaluated in turn.
consist of no more than two terminal arginine residues, or seven alanines or glycines. The scoring function used to identify the best conformers and subsequently re-rank the models includes both a newly re-weighted Lennard–Jones term and an electrostatic term:

\[ f = f_{LJ} + 132 \times f_{ES}, \]

with

\[ f_{LJ} = \sum_{i,j} \varepsilon_{ij} \begin{cases} 12 \left( \frac{r_{ij}}{\sigma_{ij}} \right)^{12} & -1.2 \times 2 \left( \frac{r_{ij}}{\sigma_{ij}} \right)^{6} & r_{ij} > 0.77 \times R_{\text{min},ij} \\ 11.5 & \text{Otherwise}; \end{cases} \]

\[ f_{ES} = \sum_{i,j} q_i q_j \begin{cases} \frac{1}{6} & r_{ij} > 1.9 \text{Å} \\ 0.277 & \text{Otherwise}; \end{cases} \]

where partial charges \( q_i, q_j \) come from CHARMM22.

The limit on refinable branch size means that the only backbone torsion angles to be refined are those near the N- or C-terminus of a chain, and that loop configurations never change.

3D-Garden avoids the use of rotamer libraries, allowing a continuous variation of the search resolution and seamless handling of molecular branches for which rotamer libraries have not been developed (such as arbitrary backbone or heterogen inclusions). Furthermore, in a rotamer library-based approach, interactions of the entire branch must be calculated for each conformation; 3D-Garden only needs to calculate the interactions of that part of the branch which is changed from the previous conformation. The number of atoms which change position is minimized by arranging for more extremal dihedrals to move faster within the conformation loop.

2.8 Dataset and parameters for assessment

The authors of the updated Benchmark 2.0 provide classifications for each complex based on tractability [rigid-body, medium difficulty, difficult] and on generalized biological role [enzyme + inhibitor, antigen + antibody, other]. We partition the Benchmark 2.0 into training and test sets such that complexes of each role and difficulty category are fairly represented both for training and testing (cf. Table 1). Complexes are assigned to the training set if their PDB code is lexicographically in the top half among their role category and to the test set (42%). Sixty nine interactor structures had at least one residue annotated as missing, which constituted roughly equal proportions of the training set (40%) and test set (42%).

Our secondary test set consists of CAPRI targets T1–2, T4–9, T11–23 and T25–27. Omitted targets are T3 (too large), T10 (trimer) and T24 and T28–31 (experimental structures not released). We implement the two proposed solutions of the ambiguous T27 as two separate cases, giving a total size of 25 for the CAPRI test.

To summarize: in all cases, we simulate the existence of supplementary information by restricting the contact points to patches of size 100 Å² within the known binding site. We generate rigid ensembles of ~340,000 models, of which the top 5000 are refined and re-scored with an electrostatic term included. These refined scores form the basis for result evaluation.

For each complex, we examine the properties of the best scoring 10 models, following the convention adopted by the CAPRI experiment. Models are categorized as incorrect, acceptable, medium and high quality according to their CAPRI L_RMS, L_RMS and f\textsubscript{nat} statistics. To be high quality, a model must have f\textsubscript{nat} \geq 0.3 and L_RMS \leq 1 Å or L_RMS \leq 1 Å. A medium quality model must have f\textsubscript{nat} \geq 0.1 and L_RMS \leq 5 Å or L_RMS \leq 2 Å. An acceptable model must have f\textsubscript{nat} \geq 0.1 and L_RMS \leq 10 Å or L_RMS \leq 4 Å. A model which does not qualify as acceptable is classed as incorrect.

The f\textsubscript{nat} of a model refers to the ‘fraction of correctly predicted native residue contacts’, specifically, the fraction of those inter-component residue pairs whose nearest heavy atoms are separated by \( \leq 5 \) Å in the experimental structure, which are also in contact by that criterion in the model. The L_RMS of a model is the RMSD between the model and the experimental structure of the smaller interactor’s heavy atoms, after a superposition based on the larger interactor’s heavy atoms. The f\textsubscript{nat} statistics.

3 RESULTS AND DISCUSSION

Table 2 shows the performance of 3D-Garden over the 45 complexes in the test set and its various sub-categories, and over

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Incorrect</th>
<th>Acceptable</th>
<th>Medium</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test (45):</td>
<td>Quality of primary prediction</td>
<td>29</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Best quality in preferred 10</td>
<td>16</td>
<td>8</td>
<td>19</td>
</tr>
<tr>
<td>Rigid-body test (35)</td>
<td>22</td>
<td>8</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Medium difficulty test (6)</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Difficult test (4)</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Enzyme-Inhibitor test (12)</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Antibody-Antigen test (11)</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>CAPRI targets (25)</td>
<td>17</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>9</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

All models belong to exactly one model quality category.
the CAPRI test set, with an average of 8400 active facet pairs, the active facets on each interactor being sampled from a 100 Å² patch around a point of close contact with the other interactor using the even sampling algorithm described in Section 2.5.

3D-Garden places at least one acceptable model in its 10 preferred structures in over two thirds of rigid-body cases, and in over three quarters of cases involving enzyme–inhibitor or antibody–antigen complexes (a likely reflection of most interactors in these categories in Benchmark 2.0 being relatively rigid). In a substantial majority of antibody–antigen complexes the primary prediction is of acceptable quality, and at least one of the preferred 10 models is of medium quality, which reflects the fact that in about half of antibody–antigen cases the antibody starts in its bound form (although it still undergoes refinement in our trial).

The CAPRI test set shows a similar identification rate of acceptable models to the Benchmark 2.0 test set; the CAPRI targets have a larger fraction without any medium model among the preferred 10, but also more cases with a high quality model present. Conformational change continues to pose a challenge; an acceptable model is present in the preferred 10 models in only two of the six medium flexibility cases of the Benchmark 2.0 test set, and only one of the four difficult cases. These results reflect the absence of a published flexible benchmark of sufficient size, which restraints us from developing dedicated parameter sets for the flexible regime, or recipes for scaling the parameters based on predicted flexibility. We hope that someone will publish a larger flexible dataset soon. Alternatively, Dockground (Gao et al., 2007), a database of complex structures, introduces a server-based paradigm which could lead to flexible and other benchmarks being generated ‘on the fly’.

Our analysis protocol is aimed at facilitating future comparisons with 3D-Garden. We note that among the recent literature, only Wang et al.’s fold-tree based extension to RosettaDock uses definitions of model quality based on CAPRI’s, although they modify the native residue contact/ interface distance cutoffs from 5/10 to 4/8 Å. Also, Wang et al.’s emphasis on exploring binding funnels leads them to present results only for medium structures, and to use prediction sets of size three rather than 10. Other studies introduce their own model quality schemes, each with different, albeit reasonable, RMSD cutoffs and nomenclature, and without reference to correct prediction of residue contacts. While such autonomy is natural in the context of method validation, we feel it impedes making inter-methodology comparisons which are transparent enough to benefit practitioners and users.

To this end we adhere to the prominent external standard, both for model quality (the CAPRI scheme), and for benchmarking (by using Benchmark 2.0).

Table 3 compares the quality of the ideal-I_RMS bound-conformation structure for each complex with the best model quality present among the ~340 000 pool structures after 5000 have been conformationally refined. Boxes above the diagonal contain complexes where the 3D-Garden algorithm’s best generated model was substantially worse than the ideal-I_RMS rigid-body structure. Boxes below the diagonal contain complexes where there was a pool structure of substantially better quality than the ideal-I_RMS structure, suggesting a successful conformational refinement of the backbone or the existence of a bound-conformation pose of higher model quality than that with ideal-I_RMS.

Table 3. Ideal-I_RMS structure quality vs. best pool model quality from all Benchmark 2.0 and CAPRI complexes (total 109)

<table>
<thead>
<tr>
<th>Ideal-I_RMS model is</th>
<th>Quality of best model in pool is</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Incorrect</td>
</tr>
<tr>
<td>Medium</td>
<td>0</td>
</tr>
<tr>
<td>Acceptable</td>
<td>0</td>
</tr>
<tr>
<td>Incorrect</td>
<td>1</td>
</tr>
</tbody>
</table>

The pool content data also reveals important information about the significance of our results. Since this is a perturbation study, the pool is already enriched in native-like structures. For a predicted model of a given quality to be significant, it must constitute an enrichment beyond that already present in the pool. As an outlying worst-case example, the fact that an acceptable structure is found in the preferred 10 models for Benchmark 2.0 rigid-body test set structure 1ktz is not significant for the reliability of the protocol; with 28 000/340 000 of the structures in this pool are acceptable (Table 4a in supplementary information), the probability of at least one being acceptable among a set of 10 randomly chosen from the 1ktz ensemble is therefore as high as 0.6. In other words, this result would be expected by chance. It is essential to quote pool content statistics, particularly in a perturbation study.

The highest acceptable model ranks in Table 4a in supplementary information (dealing with the ‘rigid-body’ complexes of the test set) confirm how important it is that the assessment scheme should arise from an external source; had we, like many other studies, permitted ourselves the free choice of the number of structures in the prediction set, then by choosing a set size of 12 we could have improved our reported reliability for this category by 15%. We also note the importance of transparently separating training and test sets; the true reliability must be inferred from the test sets alone, and a comparison between Tables 4 and 5 in supplementary information reveals this to be substantially worse across almost all categories than the same statistic calculated with training complexes (wrongly) included.

We note the broad agreement between the ideal rigid model quality and the categorization of complexes within Benchmark 2.0. Generally speaking, the ideal-I_RMS rigid superposition
for Benchmark 2.0 ‘rigid-body’ complexes is a high quality model; for ‘medium difficulty’ complexes it is a medium quality model and for ‘difficult’ complexes the ideal superposition tends to be only acceptable. There is substantial variation from this correspondence; however, with ideal-I_RMS rigid superpositions for 19/63 of the ‘rigid-body’ complexes being only medium quality, 4/13 of the ‘medium difficulty’ complexes have high quality ideal-I_RMS superpositions and half of the ‘difficult’ complexes having ideal-I_RMS superpositions as good as medium quality. It should also be remembered that model qualities of the ideal-I_RMS superpositions are in general only lower bounds for the ideal rigid model quality; a better model quality can sometimes be achieved by adjusting the model such that, although the I_RMS worsens, the L_RMS or f_nat improves crossing a boundary between model quality categories. We could not think of a good way to adjust for this.

Loops are regarded as important features of interfaces due to their conformational freedom, and loop movements on binding have been documented in a comprehensive study of protein–protein recognition sites (Lo Conte et al., 1999). 3D-Garden does not adjust loop conformations, but still singles out acceptable models of 1mlc and 1fg1, complexes where Wang et al. identify interface loops significantly affecting docking. For CAPRI T20, another mobile loop example, 3D-Garden’s preferred models are all incorrect. Higher consistency may require explicit loop adjustment, a feature from which our approach would benefit.

Finally Table 6 (Supplementary information) shows that, for the CAPRI test set excluding target nine, the distribution of ideal-I_RMS model qualities is broadly consistent with the ‘rigid-body’ Benchmark 2.0 complexes.

4 CONCLUDING REMARKS

In this work we introduce and assess the 3D-Garden protocol for modelling protein complexes. 3D-Garden uses the molecular surfaces in facet representation to generate a relatively small initial ensemble of models where the interactors are guaranteed to make a glancing contact, and spends more time scoring each initial model than the classic correlation engines, DOT, GRAMM, ZDOCK and 3D-DOCK. Sidechain and backbone dihedral refinement is carried out by a weighted exhaustive search of the conformers relevant to each clashing coordinate pair. 3D-Garden’s scoring procedure is more basic than most other docking protocols; in particular there is no solvation term, unlike ClusPro, ZRANK (Pierce and Weng, 2007) and RosettaDock. We implemented clustering and solvation initially but removed them when no qualitative improvement was observed. 3D-Garden’s server complements existing web facilities for ClusPro, GRAMM-X (Tovchigrechko and Vakser, 2006), PatchDock (Schneidman-Duhovny et al., 2005b), Hex and RosettaDock among others.

Training and testing of 3D-Garden has been carried out on complementary subsets of the widely-used Benchmark 2.0, strictly following the ensemble examination and model quality classification scheme of the CAPRI experiment. We establish, within our protocol and within the framework of a non-global search, that shape complementarity measured by a weighted Lennard–Jones term is an effective filter for the initial ensemble and that introducing an electrostatic term at that stage is unrewarding within 3D-Garden’s facet-matching framework.

While the 3D-Garden server’s (v1.2 at time of writing) primary function is predicting protein complexes, users may instead request decoys to be generated by perturbing a solution which they supply.

Structures may be either uploaded or retrieved from a server-side PDB mirror using four-character id; users may then mask out chains by label as required, and 3D-Garden will remove their heterologous. From a single pair of interactors structures, the user may initiate multiple docking runs with different parameters and surface masks. 3D-Garden uses a powerful PDB parser, correctly assimilating coordinate information from hydrogen atoms, modified residues, and any molecules in an error-corrected version of the PDB heterogen dictionary. Non-heterogen nucleotide data is not handled in v1.2, although this is a high priority for future development.

ACKNOWLEDGEMENTS

Dr Lesk was funded by the BBSRC to carry out this work. The authors would like to thank Dr Raphael Chaleil, Sara Dobbins, Pier Federico Gherrardini, Dr Suhail Islam, Mark Wass and Thomas Waterer for reading the draft. We also acknowledge Keith Sephton of the London e-Science Centre, and Dr Islam, for managing the computer systems used by 3D-Garden. We used Dr Islam’s PREPIX application for visualizing molecular surfaces.

Conflict of Interest: none declared.

REFERENCES


