FunHunt: model selection based on energy landscape characteristics

Nir London and Ora Schueler-Furman
Department of Molecular Genetics and Biotechnology, Institute of Medical Research, Hadassah Medical School, The Hebrew University, Jerusalem, Israel

Abstract
Protein folding and binding is commonly depicted as a search for the minimum energy conformation in a vast energy landscape. Indeed, modelling of protein complex structures by RosettaDock often results in a set of low-energy conformations near the native structure. Ensembles of low-energy conformations can appear, however, in other regions of the energy landscape, especially when backbone movements occur upon binding. What then characterizes the energy landscape near the correct orientation? We have applied a machine learning algorithm to distinguish ensembles of low-energy conformations around the native conformation from other low-energy ensembles. FunHunt, the resulting classifier, identified the native orientation for 50/52 protein complexes in a test set, and for all of 12 recent CAPRI targets. FunHunt is also able to choose the near-native orientation among models created by algorithms other than RosettaDock, demonstrating its general applicability for model selection. The features used by FunHunt teach us about the nature of native interfaces. Remarkably, the energy decrease of trajectories toward near-native orientations is significantly larger than for other orientations. This provides a possible explanation for the stability of association in the native orientation. The FunHunt approach, discriminating models based on ensembles of structures that map the nearby energy landscape, can be adapted and extended to additional tasks, such as ab initio model selection, protein interface design and specificity predictions.

Introduction
Recent years have seen a significant improvement in structure prediction, approaching atom-level resolution in many cases [1]. The Rosetta framework has been particularly successful, partly due to its emphasis on a common framework for a range of different tasks [2–4]. In our initial studies of RosettaDock, we observed that a local search around the native conformation results in an energy funnel: a sample of low-energy conformations near the native structure that is distinguished from the background conformations of higher energy values [5] (Figure 1a). Thus, in consequent application of RosettaDock to specific systems (e.g. [6]), as well as in CAPRI blind predictions [7,8], the funnel characteristic of a local search has been taken as one of the decisive and successful criteria for the selection of high-resolution models.

Standard RosettaDock runs model conformational flexibility upon binding at the side chain level, while the backbone is kept fixed. When the starting conformation differs significantly in its backbone from the final bound conformation, this protocol will also produce energy funnels around other (FALSE) orientations, in addition to the correct orientation (e.g. Figure 1b; [9]). What then distinguishes the correct orientation from alternative orientations? Does the energy landscape around the native conformation show special features that are absent from alternative locations? Can these features be used to select the correct orientation, and, importantly, what may they teach us about the process of protein–protein association and binding?

In order to address these questions, we developed the FunHunt classifier (see Figure 2). We trained an SVM (support vector machine [10]) classifier to distinguish correct orientations from alternative conformations, based on a range of different features. We focus on those features that are orthogonal to the optimization of the Rosetta protocol. The promising results indicate that near-native orientations are indeed distinguishable from alternative funnels in the energy landscapes, and the contributing features provide new insights about protein–protein association [9,11].

Development of FunHunt
The flow chart of the FunHunt classifier creation and application is shown in Figure 2. In order to assess different features of protein complex ensembles, we created a non-redundant set of 52 protein complexes (no two complexes in the set show more than 70% sequence identity for both monomers). For each of the complexes, we sampled two energy funnels using a RosettaDock local docking run: a TRUE funnel originating from the native complex, and a FALSE funnel originating from an alternative, non-native, low-energy conformation that had been detected previously in a global run (the RosettaDock protocol searches the local minimum in a rugged landscape by using an MCM (Monte Carlo minimization) procedure [12]; in a local run, 500 small
Figure 1 | Energy funnels appear near the native orientation in the Rosetta energy landscape (A), but also elsewhere, as for example in (B): how can the correct orientation be selected?

The central plot shows models according to their energy (y-axis, measured by the RosettaDock scoring function) and distance to the native orientation (x-axis, measured by the backbone root mean square deviation), demonstrating two distinct funnels. The plots in (A) and (B) show models according to their distance to the starting structure of a local run, similar to plots in [5]. FunHunt evaluates the local energy landscape from ensembles of models in funnel tips (corresponding cartoon representations above each plot).

Figure 2 | Development flow chart of FunHunt classifier

- Create dataset
- Local docking
- True Funnel
- False Funnel
- Select representative models
- Calculate features
- Normalize values
- Train Support Vector Machine (SVM) classifier
- Define a small set orthogonal to the RosettaDock energy function
- Reduce number of features (RFE)
- FunHunt Score = Σ (Feature weight * Feature value)

Perturbations around a given starting structure are optimized, whereas in a global run the minimum of 10^4–10^6 random orientations is computed [5,13]. Each orientation was represented by 25 structures at the funnel tip that were characterized by a set of features. Each feature was normalized over the 50 structures (25 structures from both the TRUE and FALSE funnels) and represented as Z-scores. We trained an SVM [10] classifier on the resulting normalized feature vectors. We used the RFE (repeated feature elimination [14]) feature selection method to reduce the number of features, and from the resulting set we also removed features that are part of the RosettaDock energy function. The underlying rationale is that during the docking process the structures have already been optimized in the direction of the energy function, and, since this was not enough for discrimination, selection of features orthogonal to the energy function are expected to maximally improve discrimination. The final set of seven features and the associated weights are given in Table 1. The sign of the weight indicates the characteristics of models from the near-native population, and the size of the weight indicates its relative importance.

Performance of FunHunt

The performance of a classifier can be assessed (i) at the model level, i.e. the percentage of the 50 structures for each complex that were labelled correctly; and (ii) at the complex level, i.e. using the top ranked model as predictor for the
Table 1 | Features that characterize near-native ensembles of conformations

The weight indicates the relative contribution of each feature to the classifier. The FunHunt score is the weighted sum of the features (see Figure 2).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Weight</th>
<th>Characteristics of the near-native funnel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denv: RosettaDock environment score</td>
<td>—0.81</td>
<td>Better interface environment</td>
</tr>
<tr>
<td>Δenergy: decrease in energy during the RosettaDock full atom MCM protocol</td>
<td>0.67</td>
<td>Larger energy decrease during optimization trajectory</td>
</tr>
<tr>
<td>Average sequence conservation of interface residues (based on CONSURF [22])</td>
<td>—0.6</td>
<td>Higher sequence conservation of interface residues</td>
</tr>
<tr>
<td>SASA: solvent-accessible surface area</td>
<td>0.55</td>
<td>Smaller area buried at interface (larger SASA)</td>
</tr>
<tr>
<td>ncont: number of contacts at interface</td>
<td>0.3</td>
<td>Compact interfaces (more atom-atom contacts)</td>
</tr>
<tr>
<td>Dhb-bb: number of buried unsatisfied backbone hydrogen donors/acceptors at interface</td>
<td>—0.23</td>
<td>Less unsatisfied buried hydrogen bond donors and acceptors from the peptide backbone</td>
</tr>
<tr>
<td>Distance between the two monomer centres of mass</td>
<td>0.2</td>
<td>Centres of mass are further apart</td>
</tr>
</tbody>
</table>

Features of FunHunt

If FunHunt is able to distinguish the native orientation from others, what then characterizes near-native conformational ensembles? The features that compose the classifier can shed light on protein association and binding (Table 1). The classifier consists of a weighted sum of seven different contributions. In the following, we describe the different features of the classifier, starting from the most important features (those with largest weight).

Residue interface propensity: Denv

The most important feature is the interface environment propensity of residues at the protein–protein interface [5]: not surprisingly, native interfaces are composed of amino acids with stronger preference for interfaces (compared with surface preference) than non-native orientations. This parameter is part of the low-resolution scoring function in RosettaDock that is optimized during the initial, fast, low-resolution part of the docking protocol (with centroids as coarse side chain representations). It is, however, neither used for the subsequent high-resolution optimization, nor for scoring of the different RosettaDock decoys. Thus the high weight of this parameter indicates that, in addition to the high-resolution modelling of packing, more general, low-resolution propensities are critical to a native orientation. Thus a low-resolution feature can be used as an additional selection source, especially when the model is not perfect due to backbone conformational changes that will occur upon binding.

Decrease in energy during full-atom optimization protocol: Δenergy

After low-resolution local optimization of an initial orientation (see, for example, Denv above), side chain atoms are added back to the structure and 50 full-atom MCM cycles are performed to locate the local minimum on a very rugged energy landscape [13]. The energy decrease during these cycles is significantly larger for near-native orientations. The larger energy gap might prevent the subsequent dissociation of the protein pair, and therefore lead to a preference for the native orientation over alternative orientations.

Thus, by achieving a good score at the initial coarse-grained level, the proteins might be able to associate in a correct orientation, and, by decreasing significantly in energy in the subsequent local atomic resolution refinement step, the proteins might be prevented from dissociation back to individual monomers. This is similar to two-state models proposed earlier [16], and a model proposed for binding to specific DNA-sites by transcription factors [17].

The combination of good low-resolution scores with optimal high-resolution scores has been shown elsewhere to improve model selection [18]. Technically this indicates a balance between features that can be observed at low resolution, with features that can be modelled only at the later, time-intensive steps of atomic detail.
SASA (solvent-accessible surface area) and ncont (number of contacts)
Native orientations show larger SASAs, indicating that less surface is buried by the interface. These smaller interfaces are denser, since they also contain a larger ncont. In fact, these two parameters mainly allow exclusion of false positives with large, holy interfaces.

Sequence conservation
Interfaces tend to be more conserved throughout evolution compared with wrong interfaces (e.g. [19,20]). Such constraints apply to interfaces, as interaction with a certain partner might be critical for function.

Buried unsatisfied backbone hydrogen-bond donors and acceptors: Dhb-bb
Near-native orientations contain fewer unsatisfied backbone hydrogen bond donors and acceptors at the interface. It has been suggested that protein interfaces contain extensive hydrogen bond networks, and all hydrogen bond acceptors and donors of polar buried residues are, in general, satisfied [21]. Indeed unsatisfied buried donors and acceptors are very rare, since they would result in a significant entropy penalty that is not (yet) appropriately treated in Rosetta. This feature is emphasized even more by the classifier.

We conclude that native orientations can indeed be selected by a simple classifier that evaluates a small set of features on an ensemble of models representing an energy funnel in the conformation landscape. This promising result indicates that the near-native energy landscape does indeed show specific features. The main bottleneck of the presented methodology is its assumption that the set of candidates includes a near-native orientation. Current efforts should therefore aim at improved and reliable protocols for the initial selection of candidates to verify that near-native orientations are included.

Extension of FunHunt beyond docking of proteins
The approach presented here draws from an assumption that near-native features can be described better by a set of individual models that sample the tip of energy funnels. In principle, FunHunt can be adapted to other modelling tasks, such as ab initio modelling and binding specificity. The encouraging results shown here demonstrate that a simple model can significantly improve model selection and, beyond this, suggest new insights into protein modelling in general.

Availability
A server that performs the FunHunt analysis on two candidate orientations is available at http://funhunt.furmanlab.cs.hujii.ac.il.

References

Received 7 August 2008
doi:10.1042/BST0361418