BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 10 2009, pages 1271–1279
doi:10.1093/bioinformatics/btp150

Structural bioinformatics

Artefacts and biases affecting the evaluation of scoring functions on decoy sets for protein structure prediction

Julia Handl[1], Joshua Knowles[2] and Simon C. Lovell[1,*]

[1]Faculty of Life Sciences and [2]School of Computer Science, University of Manchester, Manchester, UK

Received on October 29, 2008; revised on March 6, 2009; accepted on March 14, 2009

Advance Access publication March 17, 2009

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Decoy datasets, consisting of a solved protein structure and numerous alternative native-like structures, are in common use for the evaluation of scoring functions in protein structure prediction. Several pitfalls with the use of these datasets have been identified in the literature, as well as useful guidelines for generating more effective decoy datasets. We contribute to this ongoing discussion an empirical assessment of several decoy datasets commonly used in experimental studies.

Results: We find that artefacts and sampling issues in the large majority of these data make it trivial to discriminate the native structure. This underlines that evaluation based on the rank/*z*-score of the native is a weak test of scoring function performance. Moreover, sampling biases present in the way decoy sets are generated or used can strongly affect other types of evaluation measures such as the correlation between score and root mean squared deviation (RMSD) to the native. We demonstrate how, depending on type of bias and evaluation context, sampling biases may lead to both over- or under-estimation of the quality of scoring terms, functions or methods.

Availability: Links to the software and data used in this study are available at http://dbkgroup.org/handl/decoy_sets.

Contact: simon.lovell@manchester.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

A prime requirement for protein structure prediction is the ability to assess the accuracy of a set of candidate protein conformations. The scoring functions traditionally used for assessment derived directly from physical principles or from the properties of known protein conformations. More recently, the term MQAP (model quality assessment program) has been introduced to refer, generally, to the class of methods aimed at the ranking of structures within a given set of candidate models. In addition to 'traditional' scoring functions and assessment scores (which can score individual models in isolation) (Laskowski *et al.*, 1993; Lazaridis and Karplus, 1999; Lu and Skolnick, 2001; Luthy *et al.*, 1992; Tress *et al.*, 2003), this class also comprises consensus and clustering techniques, which rely on the availability of a set of candidate models, as the structural density within the set is taken into account (Ginalski *et al.*, 2003; Lundstrom *et al.*, 2001; Shortle *et al.*, 1998).

To test scoring functions and MQAP methods and to make further progress towards the design of more effective ones, it is now common practice to employ so-called protein decoy sets. These sets are various and differ in the properties of the decoys comprising them and the ways the decoys have been generated. Existing decoy sets include collections of protein conformations obtained using various methods of *de novo* prediction, comparative modelling, molecular dynamics (MD) simulations and loop modelling (Eramian *et al.*, 2006; Fogolari *et al.*, 2005; Keasar and Levitt, 2003; Park and Levitt, 1996; Simons *et al.*, 1997; Tsai *et al.*, 2003), several of which have been made available in the Decoys 'R' Us repository (Samudrala and Levitt, 2000). More recently, decoy sets composed of the server models submitted during the biennial CASP exercise (Critical Assessment in Structure Prediction; Moult *et al.*, 2007) have also increased in popularity. In fact, CASP7 and CASP8 have seen the introduction of a separate MQAP category, so that the blind comparison of MQAP methods on the CASP server models is now a routine part of the exercise.

A decoy set for a given protein typically includes the experimental structure (as determined by nuclear magnetic resonance or X-ray crystallography), so that the quality of each decoy can be verified objectively using measures such as the root mean squared deviation (RMSD) from the experimental structure or the Global Distance Test (Zemla, 2003). For an accurate scoring function, the experimental structure should correspond to the energy minimum and accuracy is, therefore, commonly tested by ranking the decoys, and assessing whether the native structure reliably ranks first (Park and Levitt, 1996). A related method considers the *z*-score of the experimental structure, which assesses how well its score is separated from the average of the decoys. Both approaches remain in common use (Fujitsuka *et al.*, 2004; Hsieh and Luo, 2004; Hu *et al.*, 2004; Krishnamoorthy and Tropsha, 2003; Lee and Duan, 2004; Li and Liang, 2007; McConkey *et al.*, 2003; Yang and Chen, 2004; Zhang *et al.*, 2004). Other, more powerful evaluation methods have been proposed, which aim to verify whether a scoring function provides a meaningful ranking among the decoys. For example, Tsai *et al.* (2003) consider the enrichment score, i.e. the proportion of near-native structures within the highest ranking decoys. Alternatively, correlation coefficients or scatterplots may be used to establish the degree of correlation between the score assigned by the scoring function and the accuracy of a structure.

*To whom correspondence should be addressed.

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

For the evaluation of physics-based potentials, it has become a common practice to precede the above analyses by a preprocessing step in which all conformations (both the experimental structure and the decoys) are locally minimized (Verma and Wenzel, 2007) or even relaxed using MD simulations (Wroblewska and Skolnick, 2007). This is primarily aimed at removing obvious artefacts (such as van der Waals clashes) and to obtain meaningful energy values (Verma and Wenzel, 2007), but it has recently been shown that relaxation may generally help in obtaining a more objective picture of a scoring function's performance at identifying the native structure (Wroblewska and Skolnick, 2007). Such preprocessing steps have not typically been employed during the evaluation of the knowledge-based scoring functions.

1.1 Confounding factors in decoy-based evaluation of scoring functions

The design of good decoy sets is challenging. Park and Levitt (1996) proposed a number of properties required of good decoy structures. They suggested that: 'Decoy structures must: (1) include structures that are close to the native X-ray structure; (2) be native-like in all properties of the real polypeptide chain except the overall folded conformation; otherwise they could be distinguished by trivial tests; (3) be diverse so as to sample all possible arrangements and (4) be numerous for more sensitive testing.' Tsai *et al.* (2003) supplemented this list with a further important criterion: Decoy structures must (5) 'be produced by a relatively unbiased procedure that does not use information from the native structure during the conformational search.'

Meanwhile, decoy sets of protein structures have increased in importance and popularity. Nevertheless, several of the deficiencies pointed out by Park and Levitt and others are still occurring. In this article, we argue that many popular decoy sets remain deficient with regard to the second, third or fifth of the above criteria. We further demonstrate empirically some of the experimental biases that arise as a consequence of these deficiencies. In particular, we provide evidence that trivial discrimination of the native remains a problem in several commonly used decoy collections and that the use of a range of different sets or the minimization of all conformations prior to the analysis is not always sufficient to avoid this problem. We further show that even experimental studies based on the more global measures of scoring function performance (e.g. correlation analyses or enrichment scores) can be significantly affected by biases intrinsic to particular decoy sets. In this context, we differentiate between three different kinds of biases, which arise due to (i) a lack of independence between sampling points; (ii) the use of the knowledge of the native during decoy generation; and (iii) the use of scoring functions during decoy generation.

2 MATERIALS AND METHODS

2.1 Decoy sets and preparation of structures

We used a diverse and representative collection of popular decoy sets from the literature, which are described in more detail in the Supplementary Material. Briefly, five out of 10 decoy sets from the Decoys 'R' Us repository were used, namely the 4state (Park and Levitt, 1996), lmds (Keasar and Levitt, 2003), lattice (Samudrala and Levitt, 2000), fisa (Simons *et al.*, 1997) and vhp_mcdm (Fogolari *et al.*, 2005) decoy sets. Selection criteria were the popularity of the individual datasets in the literature and diversity of the data regarding their authors and the method of decoy generation used. Two further decoy sets obtained by the Rosetta method [Rosetta All (Simons *et al.*, 1997) and Rosetta Tsai (Tsai *et al.*, 2003)] were included to reflect their increasing usage in the literature, and we added decoy sets obtained by comparative modelling [MOULDER decoy set (Eramian *et al.*, 2006)] and loop modelling (Jacobson *et al.*, 2004). Finally, we used collections of servers models for 10 targets from the recent CASP8 competition. The diverse set (see Supplementary Material) included targets of different difficulty, different sizes and different secondary structure types.

Our aim here was to assess the quality of the original decoy sets, without the introduction of further degrees of freedom based on the particular preprocessing steps used. Therefore, processing of the structures was kept at a minimum. All structures with missing non-hydrogen atoms were discarded from the analysis, and TINKER was used to add hydrogens to the experimental structures and the decoys, where necessary. Unless explicitly stated, the structures were not further modified, i.e. neither the experimental structure nor the decoys were subjected to relaxation or energy minimization. Evidently, the minimization of conformations can have a significant impact on some of our observations, and, where relevant, we therefore also discuss the effects of local minimization. Unconstrained local minimization (where applied) was performed in TINKER using limited memory L-BFGS minimization to an RMS gradient of 0.1 kcal/mole/Å, under the Amber99 force field with the GBSA solvation model.

Section 3.4 additionally employs decoy sets that were custom-designed for our study and were obtained using the low-resolution mode of Rosetta (version 2.3) (Simons *et al.*, 1997). Rosetta is a Monte Carlo technique for *de novo* structure prediction that generates candidate conformations using fragment assembly and a statistics-based scoring function.

2.2 Potential energy function

The Amber99 physics-based potential energy function [as implemented in the TINKER molecular modelling software (Ponder, 2004)] was used in all experiments reported in this article. However, the core observations made in this article have been verified using TINKER's implementations of the Charmm27 and OPLSaa force fields, too, and, importantly, it turns out that these observations are general to all three energies. Amber99 can be written as a linear combination of six terms $E_s = E_{bs} + E_{ab} + E_{it} + E_{ta} + E_{vdw} + E_{cc}$, where E_{bs}, E_{ab}, E_{it} and E_{ta} are the bonded terms constraining bond lengths, bond angles, improper torsion angles and torsion angles, respectively. E_{vdw} and E_{cc} are the non-bonded forces, which arise from van der Waals attractive and repulsive forces and electrostatic interactions, respectively. In our experiments, all of these terms were analysed individually, thus treating the Amber99 potential as a 6D vector score $E_v = (E_{bs}, E_{ab}, E_{it}, E_{ta}, E_{vdw}, E_{cc})^T$, rather than a single energy value. A seventh dimension can be added through the inclusion of an implicit solvation model, and the GSBA model was used for this purpose.

2.3 Correlation analysis

The Kendall's Tau rank correlation is a non-parametric test of correlation, which considers the difference between the number of concordant and discordant pairs in a sample (normalized by the number of possible pairs). The primary use of correlations in this article was in establishing the rank correlation between a given energy term and RMSD to the native to provide an indication of the ranking performance of the energy term on a given decoy set. The statistical software R was used to compute all correlations and statistical significance tests reported in this article.

3 RESULTS

3.1 Trivial discrimination of the native

We first investigated the degree to which obvious differences between the native and the decoys are present in a representative collection of decoy sets, and whether these differences vary for

Table 1. Success rate of the Amber99 energy function with (Amber99 + GBSA) and without solvation model (Amber99) and its individual energy components (Bond stretching = BS, Angle Bending = AB, Improper Torsion = IT, Torsion Angle = TA, Van der Waals = VDW, Charge-charge = CC, Implicit Solvation = GBSA) at identifying the native structure and also shown is the success rate obtained when selecting the (at most 14) extrema with respect to each term (Extr.). The results show that the native can be discriminated from the decoys along specific individual energy terms more frequently than along the overall energy term

Dataset	Amber99	BS	AB	IT	TA	VDW	CC	GBSA	Amber99+GBSA	Extr.
RosettaAll	33/42	12/42	13/42	41/42	22/42	34/42	18/42	16/42	32/42	42/42
RosettaTsai	18/30	12/30	9/30	26/30	16/30	18/30	5/30	0/30	17/30	30/30
4state	4/7	1/7	2/7	6/7	1/7	5/7	5/7	1/7	4/7	7/7
lmds	9/10	10/10	8/10	10/10	4/10	9/10	6/10	6/10	8/10	10/10
lattice	6/8	2/8	2/8	2/8	3/8	7/8	3/8	3/8	5/8	8/8
fisa	1/4	4/4	4/4	3/4	2/4	2/4	1/4	1/4	1/4	4/4
MOULDER	15/18	7/18	6/18	10/18	8/18	15/18	17/18	10/18	15/18	18/18
vhp_mcdm	5/5	5/5	5/5	5/5	5/5	5/5	3/5	3/5	5/5	5/5
Loop	1/15	7/15	4/15	3/15	1/15	1/15	1/15	1/15	2/15	10/15
CASP8 targets	0/10	0/10	0/10	1/10	0/10	1/10	2/10	0/10	0/10	4/10
All decoys	92/149	60/149	53/149	107/149	62/149	97/149	62/149	41/149	89/149	138/149

the different types of decoys. For this purpose, the native and the decoys were evaluated using the Amber99 all-atom force field. We then checked whether the native structure takes a value outside of the range of that taken by the decoys (smaller or larger) for any of Amber99's constituent energy components. For 139 out of 149 of the decoy sets considered, such trivial discrimination is indeed possible, as the native corresponds to an extremum under at least one Amber99 term (Table 1). This implies that, for many popular decoy sets, the set of candidate solutions can be narrowed down to a set of (at most) 14 solutions by an entirely naive selection scheme. The decoy sets taken from the recent CASP competition prove to be the most difficult out of the data considered: in these data, trivial identification of the native is only possible for four out of the 10 proteins considered. When considering alternative evaluation measures, it is evident that the seemingly impressive performance of individual Amber99 terms at identifying the native structure does not carry over. For example, when considering the identification of the best decoy structure within the set (which is a scenario much more representative of a real prediction scenario), this structure is a part of the set of Amber99 extrema in only six out of the 149 decoy sets considered (see Supplementary Material). Furthermore, the correlations between individual Amber99 terms and RMSD vary for different decoy sets, but, on average, are poor (see Supplementary Material). The strong contrast between these results and those presented in Table 1 underlines that an evaluation based on the identification of the native provides an unrealistic assessment of scoring function performance in most prediction scenarios.

In many cases, the discriminatory ability of certain bonded energy terms on specific decoy sets can be linked directly to sampling artefacts of the decoys. For example, the good performance of the Improper Torsion term on the Rosetta All and Rosetta Tsai decoy sets can be fully explained by a conformational artefact of all decoys in a specific improper torsion angle of the Asparagine and Glutamine residues, and, the discriminatory ability of the same term on the 4state and lmds dataset arises due to systematic differences with different backbone improper torsions. On the decoy sets containing such artefacts, the discriminatory performance of the corresponding terms is significantly reduced after local minimization of the conformations (see Supplementary Material). This indicates that these artefacts are less likely to influence the evaluation of physics-based energy functions (assuming that these have been evaluated on sufficiently minimized/relaxed conformations), but they may have an impact in studies focused on knowledge-based potentials, where, typically, minimization of the decoys has not been employed.

For the non-bonded terms, much of their discriminatory performance is retained after local minimization of the decoys (see Supplementary Material), indicating that there is a genuine undersampling with respect to these terms by specific methods of decoy generation. Note, for example, the strong performance of the charge—charge term on the comparative modelling (MOULDER) set, where 17 out of 18 natives can be identified based on electrostatic interactions alone. After local minimization, the charge—charge terms continues to identify the native in the majority of these decoy sets.

Similar results can be observed for the van der Waals term: when this term discriminates the native from the decoys in the original dataset, it largely retains this ability on the minimized data. For example, the van der Waals term alone identifies 16 out of 21 structures in the 4state, lmds and fisa decoy sets, and it continues to do so for 14 out of 21 structures on the minimized data. The observed undersampling of the van der Waals term is not entirely surprising given the fact that, for many of these decoy sets, full atom models were not employed at the early stages of the decoy generation. As a result, steric clashes were not considered sufficiently during optimization and the decoys score badly (compared with the native) when evaluated as full-atom models. Our observations indicate that heavier use of all-atom scoring functions is necessary to derive decoy sets in which selection of the native is genuinely challenging for an all-atom scoring function. The loop decoy set and the CASP8 targets are examples of decoy sets for which such a more global optimization of full-atom models appears to have been achieved, and, consequently, they appear more difficult both with respect to the van der Waals term and the overall all-atom scoring function.

Overall, our results underline that evaluation of scoring functions based on the identification or the rank of the native is a weak

test of scoring function performance and may have little bearing regarding a scoring function's performance at selecting low-RMSD decoys. This problem arises both due to the presence of artefacts and due to the lack of optimization of decoys with respect to certain properties. While the presence of artefacts may be reduced through a preprocessing step involving the minimization of decoys, this does not sufficiently address the undersampling of decoys w.r.t. certain energy terms. This observation is consistent with the recent work by Wroblewska and Skolnick (2007), which shows that extensive relaxation of decoys (rather than just local minimization) is required for closing the energy gap between the decoys and the experimental structure. Finally, our analysis shows that certain problems are shared across different types of decoy sets and that evaluation of scoring functions on a range of sets is therefore not necessarily a foolproof means of overcoming the weaknesses of these evaluation criteria.

3.2 Non-independence of sampling points

If the ability to identify the native is an unrealiable indicator of general scoring function performance (see previous section), alternative methods should be used. One such method is the computation of a correlation coefficient between the scores assigned to the decoys and their distance to the experimental structure. The key aim behind this type of analysis is to check whether a scoring function is able to reliably rank decoys and thus provide an accurate guidance throughout the search space.

A number of different correlation coefficients exist in the literature. While the Spearman rank correlation and Pearson correlation are the methods most commonly used in this context, both of these tend to overestimate the correlations present in a dataset. Consequently, Kendall's Tau has been recently suggested as a more reliable and interpretable statistic (Paluszewski and Karplus, 2008) and has been employed in this article. Nevertheless, the general trends in our results carry over to other correlation coefficients and related measures such as the enrichment score. Furthermore, our experiments use the RMSD as the measure of distance to the experimental structure, but it is worth noting that alternative measures exist and that, as shown in Pettitt *et al.* (2005), the choice of distance measure itself may have an effect on the degree of correlation observable on a given decoy set. This variability is thought to be caused by inadequacies of the individual distance measures (Pettitt *et al.*, 2005), but it may also reflect subtle dependencies between a scoring function and the particular distance measure used (such as similar penalization factors for incomplete models).

When analysing correlation coefficients and drawing conclusions on the general performance of a scoring function, there is an implicit assumption that the set of decoys considered provides a representative (independent and identically distributed, i.i.d.) sample of the conformational space. Our aim in the following is to investigate how specific violations of this assumption impact on the correlations observable on a given decoy set, and to show that biases result that may be a general problem in decoy-based scoring function evaluation.

The first type of decoy set we consider in this context are those obtained from MD simulations. There has been some discussion in the literature regarding the reliability of events observed in individual MD simulations and it has been suggested that (i) repeated short MD runs are more effective at sampling conformational space than a single long run (Grossfield *et al.*, 2007; Hess, 2002), and that (ii) several MD simulations may be required to obtain an objective picture of the likelihood of certain events, as the results observed in an individual trajectory may be caused by random fluctuations (Likic *et al.*, 2005). It is evident that consecutive snapshots in a given MD trajectory are highly dependent (non-ergodic) and it is known that even long MD trajectories only yield a limited number of samples that can be considered to be i.i.d. (Grossfield *et al.*, 2007; Lyman and Zuckerman, 2007), thus providing an unbiased sample of the conformational space.

The vhp_mcdm decoy set is the only decoy set in the Decoys 'R' Us database that has been obtained from MD simulations. It is constituted of energy minimized snapshots from the production runs of five independent MD trajectories of 100 ns, consisting of 1251 snapshots each. One of the simulations uses the native as the starting structure, whereas the other four start from conformations previously obtained by Monte Carlo simulation. Fogolari *et al.* (2005) compare the correlations of the molecular mechanics (MM) energy force field, both with and without the use of a solvation model, across the dataset obtained through the union of all five trajectories. They report correlations with RMSD of 0.66 for MM/GBSA compared with a correlation of 0.21 for MM without a solvation model, and conclude that this demonstrates the importance of solvation effects (Fogolari *et al.*, 2005).

Here, we re-examine these correlations (using the Amber99 force field and Kendall's Tau rank correlation coefficient) to provide an example of the vagaries of non-independent samples and their effects on measured correlations. As illustrated in Table 2, the correlations observed for the full dataset are in rough agreement with the results in the original paper (Fogolari *et al.*, 2005), showing that a higher correlation of 0.38 as compared with 0.11 is obtained when an implicit solvation model is included. However, separate analyses of the trajectories show that the same performance difference cannot be consistently observed (Table 2) and that, in general, the correlations observed for all Amber99 energy terms vary significantly across the individual trajectories (see Supplementary Material). This gives an

Table 2. Correlations (Kendall's Tau) with RMSD for the Amber99 energy function (with and without GBSA) on the five individual trajectories composing the MD dataset (F1, F3, F4, F7 and NATIVE) and the combined data (All decoys)

Decoy set	Amber99 + GBSA	Amber99
F1	0.26	0.22
F3	0.06	−0.28
F4	0.29	0.38
F7	−0.14	−0.028
NATIVE	0.50	0.11
All decoys	0.38	0.11
Normal distributions	0.30	0.078
Cluster means	0.6	0.0

Correlations are also computed across the mean points of the five trajectories (Cluster means) and across a regenerated version of the combined data, where the points in each trajectory are replaced by points sampled from a normal distribution placed around the mean point of the trajectory (normal distributions). The correlations observed vary significantly between trajectories. The correlations observed on the regenerated data are comparable to those observed on the original combined data, suggesting that the correlations are influenced significantly by the mean position of the trajectories.

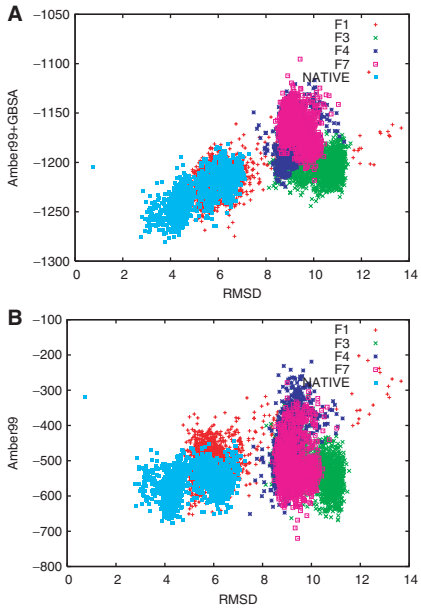

Fig. 1. Scatterplot of all conformations in the five independent trajectories of the vhp_mcdm decoy set: (**A**) Amber99 energy + GBSA versus RMSD; and (**B**) Amber99 versus RMSD. The plot shows that the correlation of the energy scores with the RMSD scores on this dataset is influenced significantly by the dynamics in a single trajectory (NATIVE), as well as the relative positions of the five 'clusters' of solutions, corresponding to individual trajectories.

indication that these trajectories reveal information about limited (and different) parts of conformational space only.

The observations above raise the question of where the correlations on the full dataset arise from. A scatterplot of RMSD versus energy (Fig. 1) indicates that, due to the similarity of decoys originating from the same trajectory, correlations observed on the full decoy set arise primarily as a consequence of differences *between* the five trajectories. To illustrate this further, Table 2 shows the correlation computed over the mean values of the five trajectories only. A distinct difference between Amber99 and Amber99+GBSA (in terms of their correlation with RMSD) can be observed, but the result has no statistical significance, as this correlation is computed over a set of five independent points only. In a follow-on experiment, the original data were replaced by five normal distributions that were generated using the means and standard deviations of the energy terms and the RMSD values of each trajectory. Analysis of the correlations for these data shows that the 'performance advantage' of Amber99+GBSA largely remains. While the number of samples in the data now *appears* to be sufficiently high to infer statistical significance, no such inference is warranted due to the small number of samples that are genuinely independent. Due to the high similarity of consecutive snapshots, similar limitations hold for the original decoy set and the statistical significance of correlations on these data is more reliably assessed by establishing the genuine number of i.i.d. samples in each trajectory.

In general, we suggest that correlation analyses on MD decoy sets may be problematic if the potential non-independence of samples is not properly accounted for. The risk of incorrectly estimating correlations can be mitigated by (i) using larger numbers of independent trajectories, or (ii) using methods to estimate the real sample size correctly, following approaches previously described in the MD literature (Grossfield *et al.*, 2007; Lyman and Zuckerman, 2007).

3.3 Low RMSD bias

A primary difficulty in the design of good decoy sets is the conflict between the first and the fifth of the five design objectives identified previously (the four of Park and Levitt and the supplementary one of Tsai and coauthors): on one hand, decoy sets are required to include structures that are close to the native structure (first criterion), but, on the other hand, knowledge of the native should not significantly bias the actual sampling process (fifth criterion). Early decoy sets were generated by using the native structure as the starting conformation for some kind of perturbation procedure and were thus clearly in violation of the fifth condition. A slightly more indirect approach was taken in the lmds dataset, but it was obtained using a scoring function optimized to generate near-native structures for this particular set of proteins, so again, knowledge of the native played a fundamental role during conformational sampling. Recent decoy sets have tackled the dilemma by avoiding the use of the native during conformational sampling, but employing knowledge of the native at a later stage to reduce the final decoy set and enrich it with near-native solutions. In this section, the effects of this last approach are examined experimentally. Density-based methods (Bonneau *et al.*, 2001; Jiang *et al.*, 2003; Shortle *et al.*, 1998; Wang *et al.*, 2004) of decoy selection are taken as an archetypal method, and we consider how enrichment with natives impacts on them. More generally, we can also expect issues with other scoring function types when used on 'enriched' (and therefore non-i.i.d.) samples, though possibly for more intricate reasons, which may be less predictable and more difficult to analyse.

In a recent paper (Wang *et al.*, 2004), a decoy-dependent discriminatory function was presented that ranks all decoys based on the computation of all-against-all RMSDs. The evaluation (Wang *et al.*, 2004) was performed using a range of decoy sets from the Decoys 'R' Us database. Correlations of this density score with RMSD of as high as 0.9 were reported, and it is notable that the strongest correlations for this method were observed for those data sets that make significant use of the knowledge of the native structure. For example, the best results were obtained on the 4state dataset, where decoys are obtained through enumeration of conformations surrounding the native structure and further selection of low-RMSD decoys.

High (but varying) correlations observed on the Rosetta Tsai decoy sets (Wang *et al.*, 2004) lend themselves to our own analysis here, as each of the 30 decoy sets within the Rosetta Tsai collection contains two distinct subsets: a 'default' set of 1000 decoys, and an 'enriched' set of 400 decoys that has been enriched with near-native structures (these have been filtered from a large number of runs). In the following, we refer to the combination of these two subsets as the 'combined' decoy set (a small fraction of the original datasets is composed of a third type of decoy, but their generation is not described in the original paper, so they are excluded from the analysis).

In our experiments, we analysed the correlation (based on Kendall's Tau) between the density score and RMSD to the native. Figure 2 shows how this correlation changes for the 30 decoy sets when considering the 'default' decoys rather than the 'combined' decoy set. From these data it is clear that, for the majority of decoy

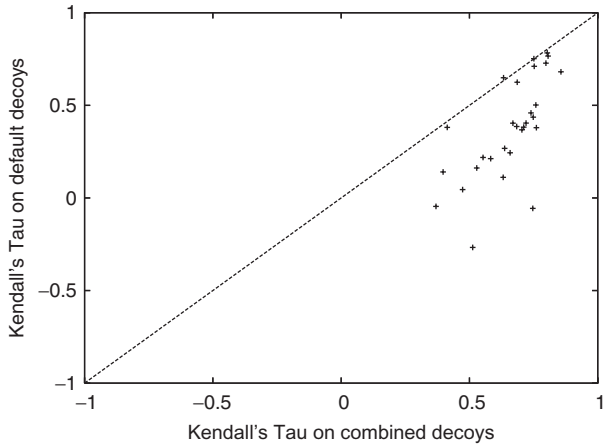

Fig. 2. Analysis of the correlation (Kendall's Tau) between RMSD (to the native) and the density score on the 30 Rosetta Tsai datasets. The correlation observed on the 'combined' decoy set is plotted versus the correlation observed on the 'default' decoy set. Each point corresponds to the correlations observed on a set of 1000 ('default') or 1400 ('combined') decoy structures for a given protein. If the enrichment with near-native decoys had no significant effect on the correlation, all points would be expected to be scattered closely around the line (no significant change in correlation). In contrast, for the majority of decoy sets, the correlation on the 'default' decoy set is significantly reduced compared with the correlations observed on the 'combined' data. The correlation is now below 0.5 for 22 out of 30 decoy sets, compared with just three out of 30 on the 'combined' data. The correlations observed on the 'default' decoy set are those that we can expect in a realistic prediction scenario, where artificial enrichment with near-native structures cannot be achieved.

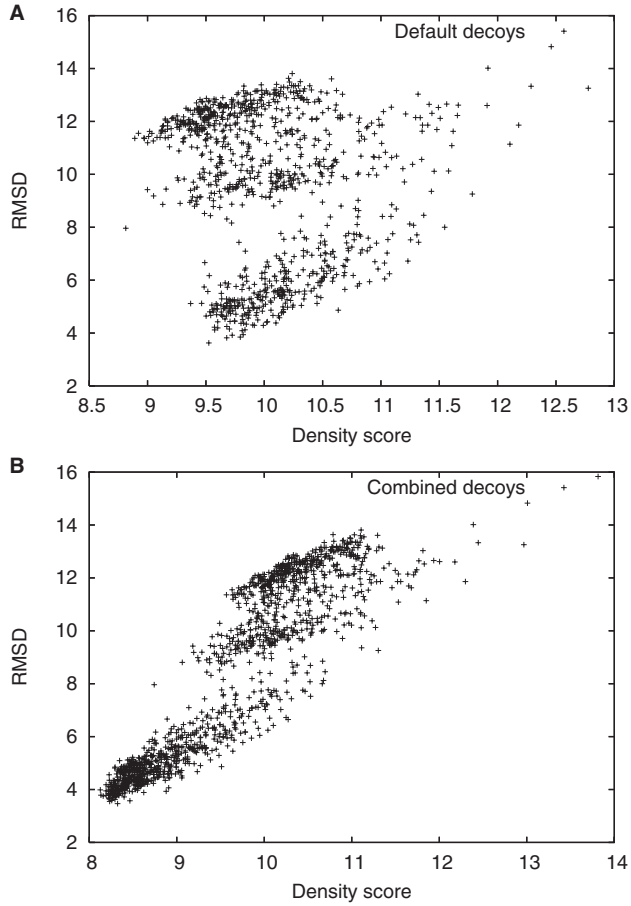

Fig. 3. Scatterplot of RMSD versus density score for the 2ptl decoy set and its 'default' subset. (**A**) Default data; (**B**) combined data. The plot of the default decoys appears to indicate that at least three energy basins have been sampled and that the density score provides some useful gradient within basins, while failing to discriminate correctly between them (a negative overall correlation between density score and RMSD is observed). The plot of the combined decoys shows that the enrichment with near-native structures leads to an ability to discriminate between the basins and thereby induces a strong positive correlation.

sets in the Rosetta Tsai collection, the enrichment with near-native structures significantly increases the correlation observed between the density score and the distance to the native. This means that, as a result of the artificial enrichment with near-native structures, the performance of the density score on the RosettaTsai decoy set is significantly overestimated compared with what can be expected in a real prediction scenario (also see the caption of Fig. 2). Figure 3 underlines this by visualizing the dramatic change in correlation (as perceived in a scatterplot) for one of the 30 Rosetta Tsai decoy sets (2ptl).

3.4 Systematic sampling biases

In this final results section, we discuss why scoring function evaluation on a given decoy set needs to take into account any relationships between the scoring function under evaluation and the scoring function (if any) used in the decoys' generation. The following thought experiment illustrates how such a relationship may directly affect the performance of a scoring function.

Assume we are given the problem of optimizing the number and positions of the vertices of a Hamiltonian cycle. The desired solution is a perfect square, but we have no direct access to this information and have knowledge only about the desired properties of the structure: (i) it shall have four vertices, (ii) it shall have edges of equal length and (iii) the smaller of the two angles at each vertex shall be exactly 90°. Each of these three properties can be formulated as an individual objective, and the set of perfect squares will be Pareto optimal (Steuer, 1986; Supplementary Material) with respect to these objectives. Now assume we want to identify the importance/optimal weighting between the three objectives, using a decoy dataset generated through optimization of the first objective only. This will consist of various quadrilaterals, and the first objective will show no discriminatory ability on this decoy set. In contrast, the second and third objective will provide good rankings of the structures in terms of their quality. If the second or third objective were used during decoy generation, roles would be reversed.

In the thought experiment, all Pareto optima are also optima on each of the three objectives individually. But in real optimization problems, such a single optimal point in the objective space does not usually exist and we have to find solutions corresponding to optimal trade-offs between the objectives. We will then be able to 'over-optimize' a single objective at the cost of another one, which may lead to a negative (rather than just an absence of) correlation.

Equivalent effects can be reproduced and observed for protein decoy sets. To demonstrate this, Rosetta's low resolution mode

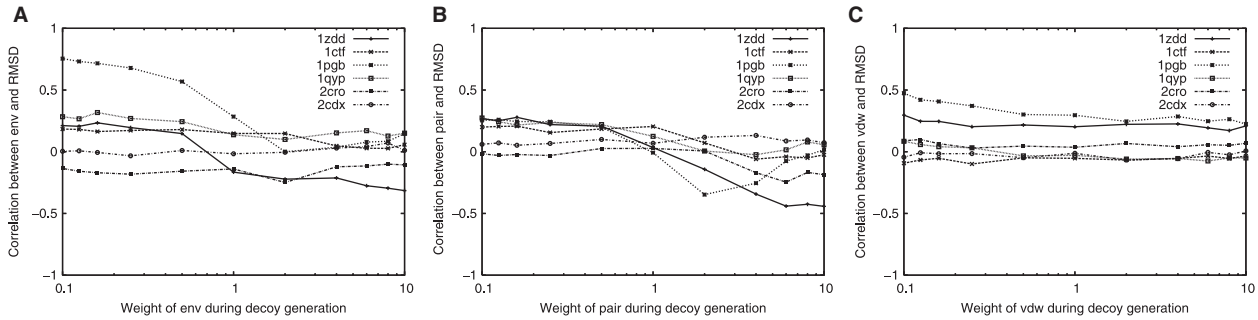

Fig. 4. Illustration of the dependency between the scoring function used during decoy generation and scoring function performance on the decoy set. The low resolution mode of Rosetta version 2.3 was used to generate decoy sets consisting of 1000 conformations each for six small proteins. During the generation of each decoy set, the weight of one of the three core terms (A: the env term; B: the pair term; C: the vdw term) of Rosetta's low resolution scoring function was readjusted by a factor of $1/10, 1/8, 1/6, 1/4, 1/2, 1, 2, 4, 6, 8, 10$, resulting in 30 different decoy sets of size 1000, overall. The correlation of the three terms with RMSD on the resulting decoy sets was then analysed. The full set of graphs is available in the Supplementary Material. For several proteins, the results show a clear anti-correlation between the weight of a given term during decoy generation and its correlation with RMSD on the resulting decoy set.

was used to generate decoy sets for six small proteins. Rosetta's low resolution energy function uses a knowledge-based scoring function, which, in its standard form, consists of nine individual terms. Three of its core terms are the pair, env and vdw term, and the weights of these three terms were varied systematically (using a set of 10 different weights $\{1/10, 1/8, 1/6, 1/4, 1/2, 1, 2, 4, 6, 8, 10\}$) in our experiments. For each setting, a separate decoy set consisting of 1000 conformations was generated. Overall, this resulted in the generation of 6×31 decoy sets of size 1000 each.

The effect of the variation in weight on the correlations with RMSD observable on the resulting decoy sets is shown in Figure 4. The graphs are largely in agreement with the effects anticipated from our thought experiment: for several of the proteins, increases in the weight of a given term cause a significant decrease in the correlation observable for this term, resulting in small or even negative correlations. One of the reasons why distinct negative correlations are not observed for Rosetta's vdw term is that it has a minimum value of 0 that can be obtained quite easily: if the weight of the vdw term is increased sufficiently, the large majority of decoys achieve this value, resulting in the absence of correlation.

Evidence of sampling-based biases can also be found in popular decoy sets from the literature. Decoy sets for the same protein show different patterns of correlations dependent on the sampling methods and scoring functions used during their generation. More specifically, distinct signature patterns can be observed when considering the correlations of the individual Amber99 terms with RMSD for all decoy sets employed in this study, and plotting the distribution of these correlations for every method of decoy generation (see Supplementary Material). For example, we observe a consistent presence of positive and (often strong) correlations of the bond stretching term with RMSD on the MOULDER decoy set, which is not observed for any of the other methods of decoy generation. Note that correlation patterns will change when subjecting decoys to local minimization.

The results above indicate that the use/non-use of a given property of native protein structures during decoy generation introduces a bias against/towards the same (or a closely related) property during scoring function evaluation. If this effect is not taken into account, experimental results may be misinterpreted. An example of this from the literature is a study concerned with the evaluation of eight empirical energy functions on a decoy set generated by MD simulations (Wang *et al.*, 1995). The scoring function (Wang *et al.*, 1995) used during decoy generation was the Amber4 potential and this was also one of the energy functions later evaluated on the decoy set. Amber4 turned out to have the least discriminatory power on this decoy set, which may be explained, at least partially, by the existence of the biases discussed in this section.

4 DISCUSSION

The experiments in this article were aimed at demonstrating some fundamental pitfalls in the decoy-based evaluation of scoring functions.

The first part of the article considered the evaluation of scoring functions based on the identification of the experimental structure. For a large number of publicly available (and commonly used) decoy sets, it was shown that the identification of the native is trivial and does not provide an adequate picture of scoring function performance. While the underlying problems of artefacts and undersampling appeared to be reduced for the more recent decoy sets in our analysis (in particular the loop decoys and the CASP data), we argue that problems not trivially identified by the Amber99 energy terms may continue to be present in these data and that evaluations based on the rank or z-score of the native should always be complemented by more powerful measures of scoring function performance.

The second part of the article was concerned with biases that may affect correlation analyses on protein decoy sets. The existence of these biases can be understood in terms of the nature of the sample presented by a given decoy set. Due to the huge number of possible protein conformations, decoy sets that provide a representative sample of the entire conformational space are difficult if not impossible to obtain and the field currently relies on the use of specific prediction techniques to obtain candidate conformations. In this context, a decoy set generated by random restarts of a given prediction methodology can be interpreted as an i.i.d. sample drawn from a probability distribution specific to this particular prediction technique. Consequently, correlation analyses on such a sample can be used for statistically sound inferences about the complementarity between a prediction method and a

scoring function (i.e. the performance of a scoring at ranking the structures returned by a particular prediction method). Importantly, however, the results of such analyses may not be transferable to other types of data (i.e. predictions generated by a different type of prediction technique) and will therefore not usually support general statements about the performance of a scoring function (as seen in Section 3.4). Furthermore, obvious violations of the i.i.d. assumption during sampling (as seen in Sections 3.2 and 3.3.) may flaw analyses or reduce the effective sample size.

From the above view of a decoy set as a sample drawn from a specific probability distribution it follows that the merging of decoy sets generated by different techniques can be problematic in its own right. In particular, the results obtained on a merged decoy set will be specific to data following the same underlying distribution and will not necessarily transfer to the individual constituent decoy sets (or unseen data). Hence, the pooling of structures obtained by different predictors is primarily useful if the merged decoy set accurately reflects the collection of structures expected to be encountered in a real prediction scenario.

Evidently, this latter observation is of relevance regarding decoy-based analyses on CASP decoy sets, as well as the interpretation of the results in CASP's MQAP category. If the distribution of server models observed during CASP cannot be accurately reproduced in practical applications (e.g. due to limitations regarding the number of input models or the lack of availability of some CASP servers), it is unclear whether the ranking of methods and the promising results observed during CASP7 and CASP8 can carry over to real applications. Regarding the ranking of methods, there already is some evidence that suggests not (McGuffin, 2007). We therefore believe that CASP's MQAP exercise may be more meaningful if the pool of models was restricted to include only those generated by a handful of publicly available servers.

5 CONCLUSION

We find that existing decoy sets used commonly for scoring function assessment and design remain problematic on a number of fronts, despite the field's appreciation of certain issues (Park and Levitt, 1996; Tsai *et al.*, 2003) and the continued, indeed increasing reliance, on these data.

For many established decoy sets, discrimination of the native is possible based on the individual terms of a standard physics-based energy function only, indicating problems with the optimization of particular properties and pointing to the presence of artefacts. This result underlines the importance of discouraging the practice of evaluating scoring functions by the rank of the native alone. If evaluation based on the rank of the native is to be used, we suggest that some obvious artefacts could be reduced through the minimization and/or relaxation of the decoys and the native prior to any other analyses. Furthermore, the quality of new decoy sets could be controlled by analyses similar to ours, where discrimination is checked against individual energy terms.

The article has further discussed that decoy sets may be interpreted as samples from a probability distribution, and has outlined a number of important conclusions that immediately follow from this view. First, violations of the i.i.d. assumption during decoy generation may significantly reduce the effective sample size and, generally, flaw inferences on the resulting data. Second, decoy-based analyses do not usually allow one to arrive at claims about the general performance of a scoring method, as current decoy sets do not adequately characterize the entirety of a protein's conformational search space.

On the positive side, decoy sets can be used for valid inferences about the performance of scoring functions on a particular type of data. Above all, our experiments therefore underline the importance of evaluating scoring methods on decoy sets that are genuinely representative of the application scenario at which the methods are targeted.

Funding: Special Training Fellowship in Bioinformatics from the Medical Research Council, UK (to J.H.); David Phillips Fellowship from the Biotechnology (to J.K.); Biological Sciences Research Council, UK (to J.K.).

Conflict of Interest: none declared.

REFERENCES

Bonneau,R. *et al*. (2001) Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins*, **S5**, 119–126.

Eramian,D. *et al*. (2006) A composite score for predicting errors in protein structure models. *Protein Sci.*, **15**, 1653–1666.

Fogolari,F. *et al*. (2005) A decoy set for the thermostable subdomain from chicken villin headpiece. Comparison of different free energy estimators. *BMC Bioinformatics*, **6**, 301.

Fujitsuka,Y. *et al*. (2004) Optimizing physical energy functions for protein folding. *Proteins*, **55**, 88–103.

Ginalski,K. *et al*. (2003) 3D-Jury: a simple approach to improve protein structure prediction. *Bioinformatics*, **19**, 1015–1018.

Grossfield,A. *et al*. (2007) Convergence of molecular dynamics simulations of membrane proteins. *Proteins*, **67**, 31–40.

Hess,B. (2002) Convergence and sampling in protein simulations. *Phys. Rev. E*, **65**, 031910.

Hsieh,M.J. and Luo,R. (2004) Physical scoring function based on AMBER force field and Poisson-Boltzmann implicit solvent for protein structure prediction. *Proteins*, **56**, 475–486.

Hu,C. *et al*. (2004) Developing optimal non-linear scoring function for protein design. *Bioinformatics*, **20**.

Jacobson,M.P. *et al*. (2004) A hierarchical approach to all-atom protein loop prediction. *Proteins*, **55**, 351–367.

Jiang,Z. *et al*. (2003) How well can we predict native contacts in proteins based on decoy structures and their energies? *Proteins*, **52**, 598–608.

Keasar,C. and Levitt,M. (2003) A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. *J. Mol. Biol.*, **329**, 159–174.

Krishnamoorthy,B. and Tropsha,A. (2003) Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics*, **19**, 1540–1548.

Laskowski,R.A. *et al*. (1993) Procheck: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, **26**, 283–291.

Lazaridis,T. and Karplus,M. (1999) Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.*, **288**, 477–487.

Lee,M.C. and Duan,Y. (2004) Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized born solvent model. *Proteins*, **55**, 620–634.

Li,X. and Liang,J. (2007) Knowledge-based energy functions for computational studies of proteins. In Xu,Y. *et al*. (eds), *Computational Methods for Protein Structure Prediction and Modeling, Volume 1: Basic Characterization*. Biological and Medical Physics, Biomedical Engineering. Springer, New York, NY.

Likic,V.A. *et al*. (2005) A statistical approach to the interpretation of molecular dynamics simulations of calmodulin equilibrium dynamics. *Protein Sci.*, **14**, 2955–2963.

Lu,H. and Skolnick,J. (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, **44**, 223–232.

Lundstrom,J. *et al*. (2001) Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci.*, **10**, 2354–2362.

Luthy,R. *et al*. (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.

Lyman,E. and Zuckerman,D.M. (2007) On the structural convergence of biomolecular simulations by determination of effective sample size. *J. Phys. Chem. B*, **111**, 12876–12882.

McConkey,B.J. *et al*. (2003) Discrimination of native protein structures using atom-atom contact scoring. *Proc. Natl Acad. Sci. USA*, **100**, 3215–3220.

McGuffin,L.J. (2007) Benchmarking consensus model quality assessment for protein fold recognition. *BMC Bioinformatics*, **8**, 345.

Moult,J. *et al*. (2007) Critical assessment of methods of protein structure prediction — round VII. *Proteins*, **69**, 3–9.

Paluszewski,M. and Karplus,K. (2008) Model quality assessment using distance constraints from alignments. *Proteins*. **75**, 540–549.

Park,B. and Levitt,M. (1996) Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.*, **258**, 367–392.

Pettitt,C. *et al*. (2005) Improving sequence-based fold recognition by using 3D model quality assessment. *Bioinformatics*, **21**, 3509–3515.

Ponder,J.W. (2004) TINKER: Software tools for molecular design 4.2. Available at http://dasher.wustl.edu/tinker/ (last accessed date October 15, 2008).

Samudrala,R. and Levitt,M. (2000) Decoys 'R' Us: a database of incorrect protein conformations to improve protein structure prediction. *Protein Sci.*, **9**, 1399–1401.

Shortle,D. *et al*. (1998) Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl Acad. Sci. USA*, **95**, 11158–11162.

Simons,K. *et al*. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **268**, 209–225.

Steuer,R. (1986) *Multiple Criteria Optimization. Theory, Computation, and Application*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley, New York.

Tress,M.L. *et al*. (2003) Predicting reliable regions in protein alignments from sequence profiles. *J. Mol. Biol.*, **330**, 705–718.

Tsai,J. *et al*. (2003) An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins*, **53**, 76–87.

Verma,A. and Wenzel,W. (2007) Protein structure prediction by all-atom free-energy refinement. *BMC Struct. Biol.*, **7**, 12.

Wang,K. *et al*. (2004) Improved protein structure selection using decoy-dependent discriminatory functions. *BMC Struct. Biol.*, **4**, 8.

Wang,Y. *et al*. (1995) Discriminating compact nonnative structures from the native structure of globular proteins. *Proc. Natl Acad. Sci. USA*, **92**, 709–713.

Wroblewska,L. and Skolnick,J. (2007) Can a physics-based, all-atom potential find a protein's native structure among misfolded structures? - large scale AMBER benchmarking. *J. Comp. Chem.*, **28**, 2059–2066.

Yang,J.M. and Chen,C.C. (2004) GEMDOCK: a generic evolutionary method for molecular docking. *Proteins*, **55**, 288–304.

Zemla,A. (2003) LGA: a method for finding 3D similarities in protein structure prediction. *Nucleic Acids Res.*, **31**, 3370–3374.

Zhang,C. *et al*. (2004) An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Sci.*, **13**, 400–411.