IMPROVING HOMOLOGY MODELS FOR PROTEIN-LIGAND BINDING SITES

Chris Kauffman, Huzefa Rangwala, and George Karypis

Department of Computer Science, University of Minnesota
117 Pleasant St SE, Room 464
Minneapolis, MN 55455
E-mail: {kauffman,rangwala,karypis}@cs.umn.edu

In order to improve the prediction of protein-ligand binding sites through homology modeling, we incorporate knowledge of the binding residues into the modeling framework. Residues are identified as binding or nonbinding based on their true labels as well as labels predicted from structure and sequence. The sequence predictions were made using a support vector machine framework which employs a sophisticated window-based kernel. Binding labels are used with a very sensitive sequence alignment method to align the target and template. Relevant parameters governing the alignment process are searched for optimal values. Based on our results, homology models of the binding site can be improved if a priori knowledge of the binding residues is available. For target-template pairs with low sequence identity and high structural diversity our sequence-based prediction method provided sufficient information to realize this improvement.

1. INTRODUCTION

Accurate modeling of protein-ligand interactions is an important step to understanding many biological processes. For example, many drug discovery frameworks include steps where a small molecule is docked with a protein to measure binding affinity[1]. A frequent approximation is to keep the protein rigid, necessitating a high-quality model of the binding site. Such models can be onerous to obtain experimentally.

Computational techniques for protein structure prediction provide an attractive alternative for this modeling task[2]. Protein structure prediction accuracy is greatly improved when the task reduces to homology modeling[3]. These are cases in which the unknown structure, the target, has a strong sequence relationship to another protein of known structure, referred to as the template. Such a template can be located through structure database searches. Once obtained, the target sequence is mapped onto the template structure and then refined.

A number of authors have studied the use of homology modeling to predict the structure of clefts and pockets, the most common interaction site for ligand binding[4–6]. Their consensus observation is that modeling a target with a high sequence similarity template is ideal for model quality while a low sequence similarity template can produce a good model provided alignment is done correctly. This sensitivity calls for special treatment of the interaction site during sequence alignment assuming ligand-binding residues can be discerned a priori.

Identifying structural properties of proteins from sequence has become a routine task exemplified by secondary structure prediction. Recent work has explored predicting interaction sites from sequence[7]. As a measure of how well these methods perform, they may be compared to methods that identify interaction sites from structure[8]. We employ both structure and sequence-based schemes to predict interaction sites but, even given perfect knowledge of which residues are involved in binding, it is not clear how best to utilize this knowledge to improve homology models.

In this work we incorporate knowledge of the residues involved in ligand binding into homology modeling to improve the quality of the predicted interaction site. Our contribution is to show that this knowledge does help and can be predicted from sequence alone with enough reliability to improve model quality in cases where target and template have low sequence identity. To our knowledge, this is the first attempt to explore the use of predicted interaction residues in a downstream application such as homology modeling. We explore a variety of parameters that govern the incorporation of binding residue knowledge, assess how much the best performing parameter sets improve model quality, and whether these these parameters generalize.

2. RELATED WORK

2.1. Prediction of ligand-binding residues

Small molecules interact with proteins in regions that are accessible and that provide energetically favorable contacts. Geometrically, these binding sites are generally deep, concave shaped regions on the protein surface, referred to alternately as clefts or pockets. We will refer to residues in clefts as ligand-binding residues.

Predicting ligand-binding site residues from sequence information is similar to several site interaction prediction problems involving DNA[9–11], RNA[12, 13], and other proteins[14–16]. Specifically, Soga and coworkers studied the prediction of ligand-binding site residues using conservation information in the form of profiles and solvent accessible properties of potentially interacting residues[7].

Several methods have been developed to identify putative ligand-binding sites by determining pockets on the protein's surface using its structure. A popular approach for this task is to place grid points at a small separation throughout the space of the protein. Potential binding sites are defined by all grid points, atoms, or residues within a fixed radius of a central grid point. This point is typically assigned based on burial criteria. Software packages such as AutoLigand[17], Ligsite^{csc}[18], VisGrid[19], and PocketPicker[8] utilize this paradigm.

2.2. Homology modeling of binding site

The factors involved in modeling protein interaction sites have received attention from a number of authors. These studies tend to focus on showing relationships between target-template sequence identity and the model quality of surface clefts/pockets.

DeWeese-Scott and Moult made a detailed study of CASP targets[a] that bind to ligands[4]. Their primary interest was in atom contacts between the model protein and its ligand. They measured deviations from true contact distances in the crystal structures of the protein-ligand complexes. Though the number of complexes they examined was small, they found that errors in the alignment of the functional region between target and template created problems in models, especially for low sequence identity pairs.

Chakravarty, Wang, and Sanchez did a broad study of various structural properties in a large number of homology models including surface pockets[5]. They noted in the case of pockets, side-chain conformations had a high degree of variance between predicted and true structures. Due to this noise, we will measure binding-site similarity using the α-carbons of backbone residues. They also found that using structure-induced sequence alignments improved number of identical pockets between model and true structures over sequenced-only alignments. This point underscores the need for a good alignment which is sensitive to the functional region. It also suggests using structure alignments as the baseline to measure the limits of homology modeling.

Finally, Piedra, Lois, and Cruz executed an excellent large-scale study of protein clefts in homology models[6]. To assess the difficulty of targets, the true structure was used as the template in their homology models and performance using other templates was normalized against these baseline models. Though a good way to measure the individual target difficulty, this approach does not represent the best performance achievable for a given target-template pair which led us to take a different approach for normalization. We follow their convention of assessing binding site quality using only the binding site residues rather than all residues in the predicted structure. As their predecessors noted, Piedra et al. point to the need for very good alignments between target and template when sequence identity is low.

The suggestions from these studies, that quality sequence alignments are essential, led us to employ sensitive alignment methods discussed in Section 4.3.

3. DATA

3.1. Primary structure and sequence data

Primary data for our experiments was taken from the RCSB Protein Data Bank (PDB)[20] in January of 2008. Protein sequences were derived directly from the structures using in-house software (Section 7). When nonstandard amino acids appeared in the se-

[a] `http://predictioncenter.org`
[b] `http://astral.berkeley.edu/seq.cgi?get=release-notes;ver=1.55`

quence, the three-letter to one-letter conversion table from Astral[21] version 1.55 was used to generate the sequence[b]. When multiple chains occurred in a PDB file, the chains were treated separately from one another. Identical sequences are removed by sequence clustering methods in later steps. Profiles for each sequence were generated using PSI-BLAST[22] with default options and the NCBI NR database (version 2.2.12 with 2.87 million sequence, downloaded August 2005). PSI-BLAST produces a position specific scoring matrix (PSSM) and position specific frequency matrix (PSFM) for a query protein, both of which are employed for our sequenced-based prediction and alignment methods.

3.2. Definition of binding residues

We considered ligands to be small molecules with at least 8 heavy atoms. Specifying a minimum number of atoms avoids single atom ligands such as calcium ions which are not of interest for this study. Protein residues involved in the binding were those with a distance less than 5Å between heavy atoms in protein and ligand. In-house software was developed to filter ligands, compute distances, and report ligand-binding residues (Section 7).

3.3. Ligand-binding residue prediction

The PDBBind database[23] provided the initial set of data used to train a support vector machine (SVM) classifier (Section 4.1). To remove redundant entries, sequences were extracted from the 'refined' set of PDBBind structures, 1300 total structures and 2392 sequences, and clustered at 40% identity using the CD-HIT software package.[24] This resulted in 400 independent sequences for which profiles were generated. This set had sequence independence at 40% identity from the evaluation set, described later.

3.4. Homology modeling data

Homology modeling requires target-template pairs with some sequence or structure relation. To construct such pairs, we started with the Binding MOAD database[25] which collects a large number of PDB entries with associated ligands. The database gives a family representative for related proteins. For each representative with a ligand of 8 atoms or more, we searched the DBAli database of structure alignments[26] for significant structurally related proteins, (DBAli structural significance score of 20 or better). Since our aim is to study the alignment of ligand binding residues, we eliminated templates which did not contain a ligand of at least 8 atoms. Targets which had no hits in the database which satisfied these criteria were also eliminated. Finally, in order to evaluate the performance of the binding-residue prediction, we eliminated any target which had greater than 40% sequence similarity to the prediction training set from Section 3.3 according to CD-HIT.

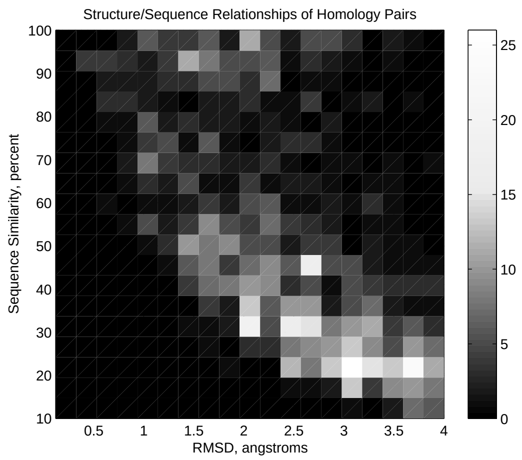


Fig. 1. The intensity of the heatmap indicates how many of the 1152 target-template pairs have the indicated RMSD-Sequence identity properties.

This left 409 unique targets, each having from one to twelve templates (average 2.8 templates per target) and 1,152 target-template pairs for the alignment. These pairs offer reasonable coverage of the sequence-structure relationship space according to their DBAli reports offering a range of easy (very similar sequences and structures) to hard homology modeling tasks (very different sequences and structures). DBAli is limited to structures related by less than a 4Å alignment and have at least 10% sequence identity which is reflected in our dataset. Figure 1 represents a distribution of the pairs over the RMSD-sequence identity landscape. The targets cumulatively represent 167,034 residues of which 9.1% are ligand-binding residues. This data was used for the evaluation of the ligand-binding residue prediction methods. An additional filtering step based on the generation of a quality baseline model was performed (see Section 5.2) which reduced the dataset to 1,000 target-template pairs for the statistical analysis of homology modeling results.

The identifiers for PDB entries used in our study may be obtained from the supplemental data (Section 7).

4. METHODS

The basis for most homology modeling approaches is to (1) obtain a structure template for a target sequence, (2) align the sequences of target and template, (3) let the target adopt the shape of the corresponding template residues, and finally (4) attempt some refinement of the shape. Our efforts center on step (2), properly aligning the binding residues of the target, assumed unknown, to those of the template, assumed known. Our hypothesis is that incorporating knowledge of these key residues will improve modeling of the binding site. In the following sections we describe how the binding residues of the target are predicted, how the target-template alignment is constructed, how baseline performance is generated from structure alignments, and the tools used to make a structure prediction.

4.1. Ligand residue prediction

4.1.1. *Structure-based prediction*

We chose to use PocketPicker for structure-based predictions of ligand-binding residues as it performed well in a recent benchmark by Weisel et al.[8]. It should be emphasized that in a true homology modeling situation, the target structure is unknown which precludes the use of structure-based predictors. They are employed here to benchmark whether binding residue prediction methods of any type are accurate enough to improve homology models.

PocketPicker reports the five largest pockets found in in the protein. Following the reasoning of Weisel et al., we defined binding residue prediction based on the single largest pocket (Pocket1) or on the largest three pockets (Pocket3) reported. These labels are evaluated for performance on the ligand-binding residue prediction task. For the homology modeling portion of the study, we used only the labels defined by the three largest pockets, Pocket3, to generate models.

4.1.2. *Sequence-based prediction*

Our predictions of ligand binding residues from sequence were made using a support vector machine (SVM) model[27]. In a previous work, we developed a generalized sequence annotation framework based on SVM learning which included prediction of ligand-binding residues[11,c]. In the present work we employed the same framework with a sliding window of size fifteen (seven to the left and right) around each residue to capture PSSM information on its neighbors. The framework is based off the SVM software package of Joachims[28] and eases the task of creating classification and regression models for sequence data.

A major advantage of SVM frameworks is their ability to exploit the so-called kernel trick which means roughly that similarity between data may be computed in a potentially high-dimensional, nonlinear space without greatly affecting efficiency. Thus, a kernel appropriate to a given type of data may be selected. In previous works, we have seen that the normalized second-order exponential kernel function ($nsoe$) is particularly useful in sequence prediction problems[11, 29, 30]. Details of the $nsoe$ kernel and framework may be found in the references.

4.2. Predicted secondary structure

Incorporating aspects of predicted structure into sequence alignment scoring has been shown to improve alignment quality[31]. In our preliminary studies, we found that alignments which did not utilize secondary structure produced far inferior homology models. To that end, we predicted secondary structure using YASSPP, a SVM-based predictor[29]. YASSPP produces a vector of three scores, one for each of the three types of secondary structure, with high positive scores indicating confidence in that class. We would like to use true secondary structure for the templates but must be careful to use a score calibrated to the YASSPP outputs. In order to create these scores, we used knowledge of the true structures of targets to calculate the average SVM prediction values for true helices, strands, and coils.

[c] Available as a tech. report at `http://www.cs.umn.edu/research/technical_reports.php?page=report&report_id=07-023`

Template residues in a helix used the average helix vector for their secondary structure and similarly for template strands and coils. This approach follows from the observation of Przybylski and Rost[32] that scoring the predicted secondary structure between two sequences improves their alignment. However, we avoid the need to make predictions for the templates by using the averaged feature vector of the appropriate type of secondary structure.

4.3. Sequence alignment

Previous analyses of homology models for clefts have used alignment methods that employ global scoring matrices, for example the ALIGN command that MODELLER provides[5, 6]. We improve on these methods by employing sensitive profile-to-profile scoring and also explore special terms related specifically to binding residues.

4.3.1. *Alignment scoring*

The basic alignment algorithm we use is derived from the work on PICASSO by Mittleman[33] which was shown to be very sensitive in subsequent studies by others[34, 30]. The details of our modification are found in a previous work[35] but are briefly described as computing an optimal local alignment using an affine gap model with matching residues i and j in sequences X and Y, respectively, scored as

$$\begin{aligned} S_{P2P}(X_i, Y_j) = &\sum_{k=1}^{20} PSSM_X(i,k) \times PSFM_Y(j,k) \\ &+ \sum_{k=1}^{20} PSSM_Y(j,k) \times PSFM_X(i,k), \end{aligned} \tag{1}$$

where $PSSM$ is the position specific scoring of a sequence and $PSFM$ is the position specific frequency matrix of a sequence. This is known as profile-to-profile scoring (P2P).

In addition to the P2P scores, we included scoring between secondary structure elements in the target and template. This was computed as a dot product of the YASSPP descriptor vectors (Section 4.2) and is referred to hereafter as SSE.

The P2P and SSE scores were combined linearly with half the matching score coming from each. We used a subset of 48 target-template pairs, picked for sequence/structure diversity, to optimize our gap opening and extension penalties for our affine gap model. After a grid search, these were set to 3.0 and 1.5 which produced the best homology models on standard alignments.

4.3.2. *Modified alignments: using binding labels*

As we sought to give special attention to the ligand binding residues, we incorporated one additional term into matching residues to reflect this goal. Each residue was labelled either as ligand-binding or not. In the case of the targets, these labels were either the true labels, as described (Section 3.2), the structure predicted labels, or a sequence-predicted labels, (both in Section 4.1). Templates always used true labels. The contribution of matching and mismatching binding and nonbinding residues was controlled using a matrix of the form

$$M_{lig} = \begin{pmatrix} 0 & m_{nb} \\ m_{bn} & m_{bb} \end{pmatrix}. \tag{2}$$

The parameters relate to a target-template nonbinding-binding mismatch (m_{nb}), target-template binding-nonbinding mismatch (m_{bn}), and target-template binding-binding match (m_{bb}). In all cases we considered, m_{bn} and m_{nb} were negative, penalizing a mismatch, while m_{bb} was positive, rewarding a match. The parameter to score a nonbinding-nonbinding match would appear in the upper left entry of M_{lig} but this match was considered neutral and thus set to zero throughout the study. The ligand modification was not weighted when combining it with P2P and SSE scores. The final form of scoring between residue X_i of target and Y_j of template is

$$\begin{aligned} S(X_i, Y_j) = &\tfrac{1}{2} S_{P2P}(X_i, Y_j) + \tfrac{1}{2} S_{SSE}(X_i, Y_j) \\ &+ M_{lig}(X_i, Y_j), \end{aligned} \tag{3}$$

where S_{P2P} is the profile-to-profile score, S_{SSE} is the dot product of the secondary structure vectors, and $M_{lig}(X_i, Y_j)$ is the modification matrix score based on the whether the residues are considered binding or not.

We refer to alignments formed from $m_{nb} = m_{bn} = m_{bb} = 0$ as *standard* alignments as they do not incorporate knowledge of the ligand-binding residues in anyway. Nonzero modification parameters are termed *modified* alignments. Our hypothesis is that for some set of parameters, the modified alignment will produce better homology models than the standard alignment.

4.4. Structure alignments

The sequence alignment of target and template is intended to approximate a map of structurally related portions. Accordingly, one could expect a sequence alignment derived from a structure alignment to give a very good starting point for the homology modeling process. This is, of course, impossible when the target is unknown. However, in a benchmark study such as ours the structure induced sequence alignment will give a reasonable baseline for the best performance that can be expected of sequence alignment.

MUSTANG is a software package which aligns structures and produces their induced sequence alignment[36]. We used MUSTANG (version 0.3) to produce a baseline alignment for each target-template pair. Homology models were produced for the MUSTANG alignments and used to normalize scores, described in Section 4.6. These structure-induced alignments are referred to as *baseline* alignments as they use a true structure relation between target and template giving the homology model the best chance for success.

4.5. Homology modeling

Once a sequence alignment has been determined between target and template, we used MODELLER to predict the target structure[37]. We employed version 9.2 of the MODELLER package which is freely available. As input, MODELLER takes a target-template sequence alignment and the structure of the template. An optimization process ensues in which the predicted coordinates of the target are adjusted to violate, as little as possible, spatial constraints derived from the template.

Details of our use of MODELLER are as follows. The 'automodel' mechanism was used which, given the sequence alignment, performs all necessary steps to produce a target structure prediction. We chose to generate a single model as a brief preliminary exploration indicated little changes when multiple models are generated (data not shown). As mentioned earlier, some template structures contained nonstandard amino acids for which MODELLER will fail. To that end, we used a modified table of amino acid code to type conversions, taken from ASTRAL as in Section 3.1, to model nonstandard residues as an analogous standard residue. The mechanism for defining such a table is described in the MODELLER manual[d] and the specific table we used is available with the other supplementary data (Section 7).

4.6. Evaluation

4.6.1. *Ligand-binding residue predictions quality*

We evaluated the sequence-based prediction of ligand-binding residues using the receiver operating characteristic (ROC) curve[38]. This is obtained by varying the threshold at which residues are considered ligand-binding or not according to the SVM output of the predictor. For any binary predictor, the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) determines standard classification statistics which we use for comparison between the structure-based and sequence-based predictors. These are

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (7)$$

4.6.2. *Homology modeling quality*

We chose to evaluate predicted structures (models) based on their RMSD from the true structure of the protein in question. A low RMSD indicates similarity between two structures. Calculations were done using in-house software which implements the quaternion method of computing RMSD[39]. Only the α-carbon coordinates are used for the RMSD computation. Following the convention of Piedra et al.[6], we computed the RMSD between only the ligand-binding residues in the model and those in the true structure as these residues are most important to models of the binding site. For brevity, this will

[d] `http://www.salilab.org/modeller/manual/node105.html`

be called the *lig*RMSD for ligand-binding residues RMSD.

Difficult modeling tasks are not expected to achieve a low RMSD: there is not enough information present in the template to deduce a high quality target model. Evaluating purely on the above RMSD criteria would not account for this factor. We chose to normalize the RMSD in the following way. Using the baseline sequence alignment (generated from structure, Section 4.4), we produced a model for the target. The *lig*RMSD was calculated for this model against the true structure and is denoted $lig\mathrm{RMSD}_{base}$. Sequence-only alignments were then used to generate homology models for the same target-template pairs. The *lig*RMSD for these models, denoted $lig\mathrm{RMSD}_{seq}$, was divided by the ligand RMSD of the corresponding $lig\mathrm{RMSD}_{base}$. The sequence alignments we produced were local while the baseline alignments were global. Using a local alignment means that some of the ligand-binding residues were potentially omitted from the alignment and subsequent model. For a given model, the total number of ligand binding residues is n_{tot} while the number of ligand-binding residues in the model is n_{mod}. We penalize the score of models by the ratio of total to missing residues. This gives a normalized homology score of

$$H = \frac{lig\mathrm{RMSD}_{seq}}{lig\mathrm{RMSD}_{base}} \times \frac{n_{tot}}{n_{mod}}. \qquad (8)$$

Due to the ratio that is taken here, the scores should follow a log-normal distribution. When doing our statistical analysis, we convert into log-space to calculate significance but report results in the usual space.

To test whether knowledge of the ligand-binding residues improved or degraded binding site models, we performed Student's t-Test on the normalized scores of the standard alignment predictions paired with the corresponding normalized scores for modified alignments. The null hypothesis is that the two have equal mean while the alternative hypothesis is that the modified alignments produce models with a lower mean (a one-tailed test). We report p-values for the comparisons noting that a p-value below 0.05 is typically considered statistically insignificant. We also report the mean improvement (gain) from using modified alignments. If the mean of all normalized homology scores for the standard alignments is $\bar{H}_{stand}$ and that of a modified alignment is $\bar{H}_{mod}$, the percent gain is

$$\%\mathrm{Gain} = \frac{\bar{H}_{stand} - \bar{H}_{mod}}{\bar{H}_{stand}}. \qquad (9)$$

A positive gain indicates improvement through the use of the ligand-binding residue labels while a negative gain indicates label use degrades the homology models.

5. RESULTS

5.1. Ligand-bind residue prediction from sequence and structure

Figure 2 illustrates the receiver operating characteristic (ROC) for the sequence-based predictor on the evaluation set. To produce binary labels, a threshold was chosen so that the number of predicted positives was approximately equal to the number of true positives. The threshold point is shown in Figure 2 and statistics of the labels it induces are shown in Table 1. Also in Table 1 we show the performance of the structure-based predictor on the targets based on binding-residue definitions from the largest single and largest three pockets, labeled Pocket1 and Pocket3 (Section 4.1).

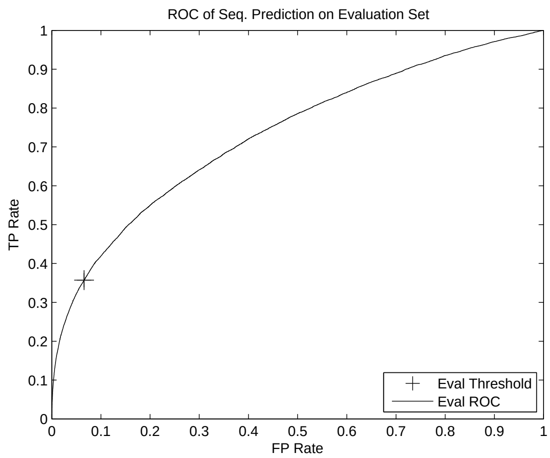


Fig. 2. ROC of sequence-only predictions of ligand-binding residues on evaluation set. The threshold position indicates the FPR and TPR for the predicted labels used in evaluation. The AUC is 0.7351 for the evaluation set.

In predicting ligand-binding residues, the sequence-only predictions are very comparable to those of the structure-based methods in terms of accuracy. As expected, the precision is worse than the best structure-derived labels method, but the two perform similarly when three of the largest pockets are used in the structure method.

Table 1. Performance statistics for predicting ligand-binding residues

Statistic	SeqPred	Pocket1	Pocket3
Accuracy	0.8813	0.8948	0.8302
Precision	0.3531	0.4430	0.3087
Recall	0.3572	0.5839	0.6907
Specificity	0.9341	0.9261	0.8443

A threshold of -0.91 was chosen for the sequence-based prediction as the cutoff for the positive class. Two variants of PocketPicker were used: positive residues generated from the single largest and three largest pockets, Pocket1 and Pocket3.

5.2. Homology modeling

Homology models were produced for the standard alignment procedure and for modified alignments that incorporated ligand labels derived from three sources: the true labels (Section 3.2), structure predicted labels, and sequence predicted labels (both in Section 4.1).

In some cases, the predicted structure that is produced by MODELLER is obviously wrong, for example when the model is in an extended rather than compact conformation. We removed structures for which the baseline alignment produced a model with greater than 10Å all-residue RMSD from the true structure. This left 1000 structures for the statistical analysis. Additional filtering was done on each target-template pair with failures being ignored for the analysis. Finally, we analyzed models in subgroups with specific sequence and structure properties and report the sample size of each group.

5.2.1. *Using true labels for binding residues*

The second section of Table 2 shows the improvement for alignments which used the true labels of ligand-binding residues. We found parameters $m_{bb} = 10$, $m_{nb} = m_{bn} = 0$ to provide the most improvement over standard alignments, though $m_{bb} \in \{7.5, 12.5\}$ with $m_{nb} = m_{bn} = 0$ produced only slightly inferior results. Also, $m_{bb} = 10, m_{nb} = -2.5, m_{bn} = 0$ performed well. We will discuss the issue of asymmetry in scoring later as it also pertains to the sequence and structure predicted labels.

The table shows sequence/structure subgroups along with the quality gained through the use of labels and whether the result is statistically significant (p-value ≤ 0.05). Improvement for the true labels occurs in low sequence identity groups with better gains in the higher structure diversity subgroup (2-4Å RMSD). At higher sequence identity, use of the labels improves performance only when the target and template are structurally diverse (0-50% identity and 2-4Å RMSD).

5.2.2. *Using structure-predicted labels*

We report the results of using structure predicted binding labels in the third section of Table 2. The best parameters we found in our grid search were $m_{bb} = 5, m_{nb} = 0$, and $m_{bn} = -2.5$, an assymetric scoring matrix. We see similar trends for the structure-predicted labels as were observed for the true labels with the largest gains appearing in the low sequence identity and high structural diversity areas of sequence-structure space. The magnitude of improvement for the structure-predicted labels appears greater in some cases than the true labels. We are still investigating the cause of this behavior.

5.2.3. *Using sequence-predicted labels*

The fourth section of Table 2 shows homology modeling results when sequence predicted labels are used. Again, asymmetric scoring parameters of $m_{bb} = 5, m_{nb} = 0, m_{bn} = -2.5$ provided the best performance. The significant gains are achieved only in the low sequence identity category and are greater in magnitude when the target-template structures are more diverse.

5.2.4. *Comparisons*

To compare the performance of true, structure-predicted, and sequence-predicted labels, we examine the first two rows of Table 2. These are the subgroups of pairs related by $\leq 30\%$ sequence identity and a DBAli structure alignment either between $0 \leq 4.0$Å or $2 \leq 4.0$Å. These two subgroups are where use of the ligand-binding labels appears to offer positive gains regardless of their source. The improvement given in these groups by the sequence-based labels are smaller than those for true and structure-

Table 2. Homology modeling results

		True Labels				Structure Labels				Sequence Labels			
SeqID	RMSD	N	$\frac{n_{mod}}{n_{tot}}$	%Gain	p-value	N	$\frac{n_{mod}}{n_{tot}}$	%Gain	p-value	N	$\frac{n_{mod}}{n_{tot}}$	%Gain	p-value
0≤ 30	2.0≤4.0	234	0.99	**3.51**	0.0009	234	0.98	**3.34**	0.0099	234	0.98	**2.03**	0.0276
0≤ 30	0.0≤4.0	254	0.99	**3.09**	0.0018	254	0.98	**3.95**	0.0037	254	0.98	**1.87**	0.0274
30≤ 60	0.0≤2.0	135	1.00	-0.02	0.5104	131	1.00	-0.93	0.7468	135	1.00	-0.50	0.7922
30≤ 60	2.0≤4.0	192	0.98	-1.40	0.9266	189	0.98	-1.10	0.7462	192	0.98	-2.33	0.9448
30≤ 60	0.0≤4.0	325	0.99	-0.83	0.9058	318	0.99	-1.04	0.8182	325	0.99	-1.58	0.9611
60≤100	0.0≤2.0	267	0.98	-0.34	0.9342	265	0.98	-0.72	0.8405	267	0.98	0.05	0.4334
60≤100	2.0≤4.0	121	0.99	-0.53	0.8451	120	0.99	-1.20	0.8492	121	0.99	-0.27	0.7274
60≤100	0.0≤4.0	388	0.98	-0.40	0.9626	385	0.98	-0.87	0.9217	388	0.98	-0.05	0.5838
0≤ 50	0.0≤2.0	116	1.00	-0.55	0.7718	114	1.00	1.28	0.2780	116	1.00	0.13	0.4109
0≤ 50	2.0≤4.0	395	0.98	1.73	0.0110	392	0.98	1.37	0.1204	395	0.98	0.03	0.4887
0≤ 50	0.0≤4.0	505	0.99	1.23	0.0230	500	0.99	1.38	0.0920	505	0.99	0.04	0.4769
50≤100	0.0≤2.0	312	0.98	-0.22	0.7796	308	0.98	-0.76	0.8812	312	0.98	-0.21	0.7647
50≤100	2.0≤4.0	152	0.99	-1.22	0.9072	151	0.99	-0.67	0.7519	152	0.99	-0.04	0.5167
50≤100	0.0≤4.0	464	0.98	-0.55	0.9374	459	0.98	-0.73	0.9123	464	0.98	-0.15	0.6701
0≤100	0.0≤2.0	426	0.99	-0.31	0.8587	420	0.99	-0.21	0.6091	426	0.99	-0.11	0.6688
0≤100	2.0≤4.0	546	0.99	0.92	0.0641	542	0.99	0.81	0.1817	546	0.99	0.01	0.4952
0≤100	0.0≤4.0	966	0.99	0.38	0.1469	956	0.99	0.37	0.2673	966	0.99	-0.05	0.5492

Columns one and two are the target-template sequence and RMSD ranges. The remaining columns relate specifically to each type of label. Columns three through six describe the sample size, ratio of modeled to total binding residues (Equation 8), percentage gain (Equation 9), and significance of results of models predicted using true labels. Columns six through eight describe the structure-predicted labels and columns nine through twelve the sequenced-predicted labels. The term $\frac{n_{ali}}{n_{tot}}$ is averaged over all models in the sample and, being close to one in all cases, indicates the majority of ligand binding residues are modeled.

based labels, but they are present and significant. It is also interesting to examine the last row of Table 2 and note that over the entire dataset, the true and structure-predicted labels offer positive though statistically insignificant gains while sequence-predicted labels slightly degrade model quality overall. This suggests use of labels only in the case when the only available templates are those with low sequence identity.

In many cases, the sequence-predicted labels did very well compared to the structure labels. An example of this is shown in Figure 3 for target 1h5q chain A produced by alignment to 1mxh chain D. In this case, the sequence-only method performs nearly identically to the structure-based method for deriving labels.

The magnitude of the ligand-ligand matching reward is different between the true and predicted label methods, 10 for true labels, 5.0 for the predicted labels. This is likely due to low precision for the predicted ligands.

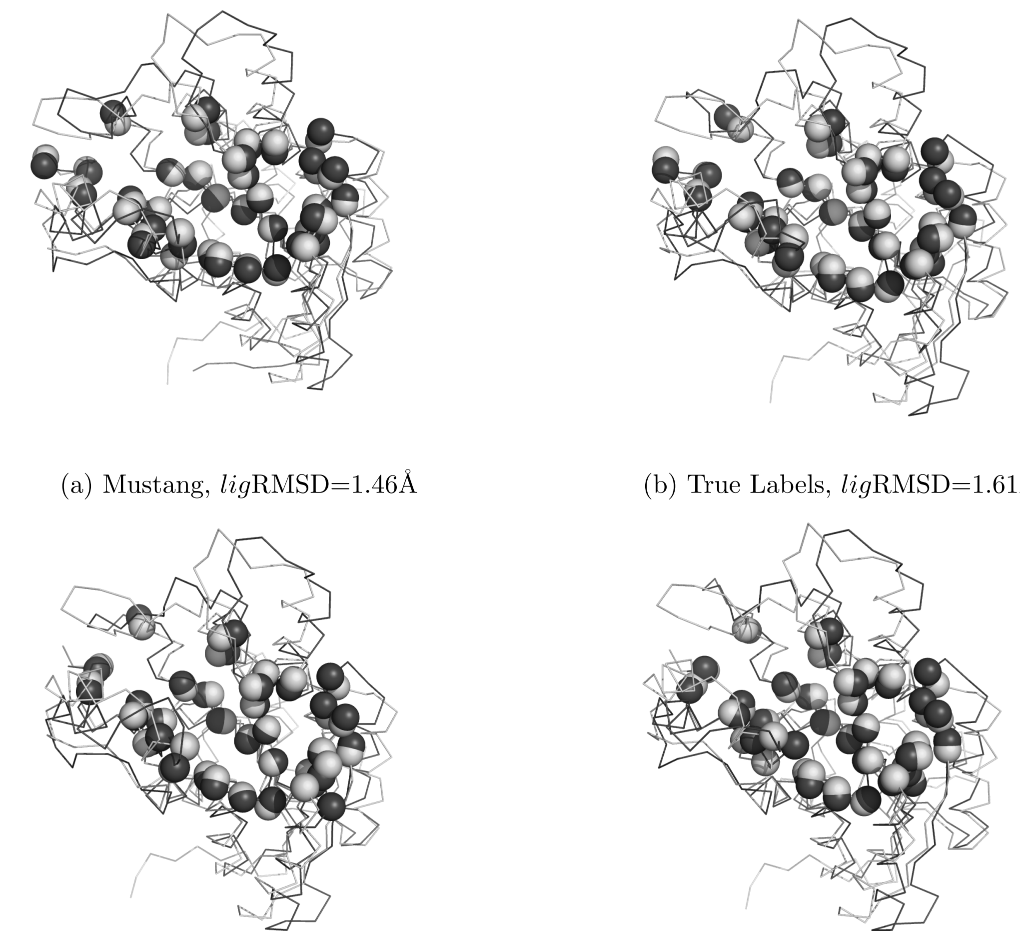
The success of asymmetric scoring parameters for predicted labels still requires further investigation. It was expected that the true signal from template ligands to govern the success of the scoring parameters. This would lead to a negative m_{nb} to penalize 'missing' known ligand binding residue in the template. This appears to be the case for true labels which had good performance for $m_{bb} = 10, m_{nb} = -2.5, m_{bn} = 0$. However, the opposite has shown to be true for both the sequence and structure-based alignments, that m_{nb} is neutral while m_{bn} is used to penalize the alignment of a predicted binding residue to a nonbinder in the template.

5.2.5. *Generalization of model parameters*

When proposing a parameterized model that shows prediction improvements, care is needed to ensure that the chosen parameters are not highly dependant upon the data used for measurement. Since our modified alignments depend on a small number of parameters that affect the scoring binding residue matches, we want to ensure that these parameters will reproduce the reported performance on future data. To that end, we performed a permutation test to determine the validate the modified alignments.

For the sequence/structure subgroups of interest, we took random subsets and performed paired Student's t-Test on the standard and modified alignment normalized scores. We took the average p-value over 1000 random subsets and used it as an indication of how well the parameters are expected to perform on future data.

Models generated using the true labels and the parameters $m_{bb} = 10, m_{nb} = 0, m_{bn} = 0$ had better



(a) Mustang, ligRMSD=1.46Å (b) True Labels, ligRMSD=1.61Å

(c) SeqPred, ligRMSD=1.74Å (d) PocketPicker, ligRMSD=1.75Å

Fig. 3. Homology models for target 1h5q chain A (template 1mxh chain D with 20% sequence identity and 2.48Å RMSD) produced by the 4 types of alignments. The protein has 260 residues with 35 ligand-binding residues. A backbone trace for the true model is shown in lightly colored, the predicted model in darkly colored, and the α-carbons of ligand-binding residues are shown as spheres. Images were produced with Pymol.

average p-values than other parameters in all the significant cases mentioned above indicating that they are likely to be applicable to future data.

Average p-values for the structure-based predicted labels and the parameters $m_{bb} = 5, m_{nb} = 0, m_{bn} = -2.5$ were better than other parameter sets. Again, significance was achieved in all the the cases above indicating good generalization.

Finally, the sequence predicted labels did not appear to have as good of generalization properties. At sequence identity $\leq$ 0-30% and RMSD $0 \leq 4$Å, the average p-values were between 0.08 and 0.11. An improved sequence predictions and a finer-grained grid-search will likely locate optimal parameters for the sequence-predicted labels generalize well.

6. CONCLUSIONS

We have explored the performance of a sequence-based and a structure-based ligand-binding residue predictor and have shown that making use of these predictions in a homology modeling framework can improve the overall quality of predicted structures. This effect is most pronounced when the sequence identity between the target and template is low.

Our prediction of ligand-binding residues from sequence was by no means perfect but the down-stream application shows that even noisy predictions can benefit homology models.

It is unclear at this point why using the structure-predicted labels from PocketPicker outperform the true labels but this may be a moot point as in real homology modeling the structure of the target is unknown. This result may suggest that an alternate definition for ligand-binding residues should be used, one which accounts for the location of a residue in a pocket as well as being within contact distance of the ligand.

There are several relevant directions to pursue in order to expand on the current work. Improving ligand-binding residue prediction from sequence will no doubt boost the performance of models generated via this mechanism. Though the set of parameters we explored for alignment modification was sufficient to indicate improvement, it was by no means exhaustive enough to claim that the optimal parameters were located. The particular values used for modifications are highly dependent on other aspects of the alignment process such as P2P scoring function. This remains a general problem worth studying: what is the best way to incorporate diverse information (profiles, SEE, ligand labels) into the scoring scheme for alignments? Extending the notion of a 'label' for a residue to a continuous value, indicative of confidence, will increase the flexibility of this part of the scoring scheme and remove the need to derive a threshold separating positive and negative classes.

7. ACKNOWLEDGMENTS, SUPPLEMENTS

The authors gratefully acknowledge support from the NIH Training for Future Biotechnology Development grant, NIH T32GM008347, and NIH RLM008713A, NSF IIS-0431135, and the U of MN Digital Technology Center.

Supplementary materials for this work are available online at `http://bioinfo.cs.umn.edu/supplements/ligand-modeling/csb2008`. These include the MODELLER modified residue table, the cross-validation results of section 5.2.5 and the binary programs for extraction, sequence alignment, and structure alignment.

References

1. N Moitessier, P Englebienne, D Lee, J Lawandi, and C R Corbeil. Towards the development of universal, fast and highly accurate docking//scoring methods: a long way to go. *Br J Pharmacol*, 153(S1):S7–S26, November 2007.
2. Philippe Ferrara and Edgar Jacoby. Evaluation of the utility of homology models in high throughput docking. *Journal of Molecular Modeling*, 13:897–905, Aug 2007. 10.1007/s00894-007-0207-6.
3. D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294(5540):93–96, Oct 2001.
4. Carol DeWeese-Scott and John Moult. Molecular modeling of protein function regions. *Proteins*, 55(4):942–961, Jun 2004.
5. Suvobrata Chakravarty, Lei Wang, and Roberto Sanchez. Accuracy of structure-derived properties in simple comparative models of protein structures. *Nucleic Acids Res*, 33(1):244–259, 2005.
6. David Piedra, Sergi Lois, and Xavier de la Cruz. Preservation of protein clefts in comparative models. *BMC Struct Biol*, 8(1):2, Jan 2008.
7. S. Soga, H. Shirai, M. Kobori, and N. Hirayama. Use of amino acid composition to predict ligand-binding sites. *Journal of Chemical Information and Modeling*, 47(2):400–406, 2007.
8. Martin Weisel, Ewgenij Proschak, and Gisbert Schneider. Pocketpicker: analysis of ligand binding-sites with shape descriptors. *Chemistry Central Journal*, 1(1):7, 2007.
9. Yanay Ofran, Venkatesh Mysore, and Burkhard Rost. Prediction of dna-binding residues from sequence. *Bioinformatics*, 23(13):i347–353, 2007.
10. Shandar Ahmad and Akinori Sarai. Pssm-based prediction of dna binding sites in proteins. *BMC Bioinformatics*, 6:33, 2005.
11. Huzefa Rangwala, Christopher Kauffman, and George Karypis. A generalized framework for protein sequence annotation. In *Proceedings of the NIPS Workshop on Machine Learning in Computational Biology*, 2007.
12. Michael Terribilini, Jae-Hyung Lee, Changhui Yan, Robert L. Jernigan, Vasant Honavar, and Drena Dobbs. Prediction of RNA binding sites in proteins from amino acid sequence. *RNA*, 12(8):1450–1462, 2006.
13. Manish Kumar, M. Michael Gromiha, and G. P S Raghava. Prediction of rna binding sites in a protein using svm and pssm profile. *Proteins*, 71(1):189–194, Apr 2008.
14. Yanay Ofran and Burkhard Rost. Predicted protein-protein interaction sites from local sequence information. *FEBS Lett*, 544(1-3):236–239, Jun 2003.
15. Ming-Hui Li, Lei Lin, Xiao-Long Wang, and Tao Liu. Protein protein interaction site prediction based on conditional random fields. *Bioinformatics*, 23(5):597–604, 2007.
16. Asako Koike and Toshihisa Takagi. Prediction of protein-protein interaction sites using support vector machines. *Protein Engineering, Design and Se-*

lection, 17(2):165–173, 2004.
17. Rodney Harris, Arthur J Olson, and David S Goodsell. Automated prediction of ligand-binding sites in proteins. *Proteins*, Oct 2007.
18. Bingding Huang and Michael Schroeder. Ligsitecsc: predicting ligand binding sites using the connolly surface and degree of conservation. *BMC Struct Biol*, 6:19, 2006.
19. Bin Li, Srinivasan Turuvekere, Manish Agrawal, David La, Karthik Ramani, and Daisuke Kihara. Characterization of local geometry of protein surfaces with the visibility criterion. *Proteins*, Nov 2007.
20. Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucl. Acids Res.*, 28(1):235–242, 2000.
21. John-Marc Chandonia, Nigel S Walker, Loredana Lo Conte, Patrice Koehl, Michael Levitt, and Steven E Brenner. Astral compendium enhancements. *Nucleic Acids Res*, 30(1):260–263, Jan 2002.
22. SF Altschul, TL Madden, AA Schaffer, J Zhang, Z Zhang, W Miller, and DJ Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucl. Acids Res.*, 25(17):3389–3402, 1997.
23. Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The pdbbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J Med Chem*, 47(12):2977–2980, Jun 2004.
24. Weizhong Li, Lukasz Jaroszewski, and Adam Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3):282–283, 2001.
25. Mark L Benson, Richard D Smith, Nickolay A Khazanov, Brandon Dimcheff, John Beaver, Peter Dresslar, Jason Nerothin, and Heather A Carlson. Binding moad, a high-quality protein-ligand database. *Nucleic Acids Res*, 36(Database issue):D674–D678, Jan 2008.
26. Marc A. Marti-Renom, Valentin A. Ilyin, and Andrej Sali. Dbali: a database of protein structure alignments. *Bioinformatics*, 17(8):746–747, 2001.
27. Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, 1995.
28. T. Joachims. *Advances in Kernel Methods: Support Vector Learning*, chapter Making large-Scale SVM Learning Practical. MIT-Press, 1999.
29. George Karypis. Yasspp: Better kernels and coding schemes lead to improvements in svm-based secondary structure prediction. *Proteins: Structure, Function and Bioinformatics*, 64(3):575–586, 2006.
30. Huzefa Rangwala and George Karypis. frmsdpred: predicting local rmsd between structural fragments using sequence information. *Comput Syst Bioinformatics Conf*, 6:311–322, 2007.
31. Jian Qiu and Ron Elber. Ssaln: an alignment algorithm using structure-dependent substitution matrices and gap penalties learned from structurally aligned protein pairs. *Proteins*, 62(4):881–891, Mar 2006.
32. Dariusz Przybylski and Burkhard Rost. Improving fold recognition without folds. *J Mol Biol*, 341(1):255–269, Jul 2004.
33. David Mittelman, Ruslan Sadreyev, and Nick Grishin. Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments. *Bioinformatics*, 19(12):1531–1539, Aug 2003.
34. A. Heger and L. Holm. Picasso: generating a covering set of protein family profiles. *Bioinformatics*, 17(3):272–279, Mar 2001.
35. Huzefa Rangwala and George Karypis. Incremental window-based protein sequence alignment algorithms. *Bioinformatics*, 23(2):e17–e23, Jan 2007.
36. Arun S Konagurthu, James C Whisstock, Peter J Stuckey, and Arthur M Lesk. Mustang: a multiple structural alignment algorithm. *Proteins*, 64(3):559–574, Aug 2006.
37. A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234(3):779–815, Dec 1993.
38. T. Fawcett. Roc graphs: Notes and practical considerations for researchers, 2004.
39. Chaok Seok Ken A. Dill Evangelos A. Coutsias. Using quaternions to calculate rmsd. *Journal of Computational Chemistry*, 25:1849–1857, 2004.