Prediction of Homology Model Quality with Multivariate Regression

Kristin Tøndel*

Department of Cancer Research and Molecular Medicine, Faculty of Medicine, Norwegian University of Science and Technology, MTFS, N-7489 Trondheim, Norway

Received February 25, 2004

A new method has been developed for prediction of homology model quality directly from the sequence alignment, using multivariate regression. Hence, the expected quality of future homology models can be estimated using only information about the primary structure. This method has been applied to protein kinases and can easily be extended to other protein families. Homology model quality for a reference set of homology models was verified by comparison to experimental structures, by calculation of root-mean-square deviations (RMSDs) and comparison of interresidue contact areas. The homology model quality measures were then used as dependent variables in a Partial Least Squares (PLS) regression, using a matrix of alignment score profiles found from the Point Accepted Mutation (PAM) 250 similarity matrix as independent variables. This resulted in a regression model that can be used to predict the accuracy of future homology models from the sequence alignment. Using this method, one can identify the target-template combinations that are most likely to give homology models of sufficient quality. Hence, this method can be used to effectively choose the optimal templates to use for the homology modeling. The method's ability to guide the choice of homology modeling templates was verified by comparison of success rates to those obtained using BLAST scores and target-template sequence identities, respectively. The results indicate that the method presented here performs best in choosing the optimal homology modeling templates. Using this method, the optimal template was chosen in 86% of the cases, as compared to 62% using BLAST scores, and 57% using sequence identities. The method presented here can also be used to identify regions of the protein structure that are difficult to model, as well as alignment errors. Hence, this method is a useful tool for ensuring that the best possible homology model is generated.

1. INTRODUCTION

During the past decade, homology modeling of protein structures has become a commonly used technique. Homology modeling is the procedure of generating a model of a protein using an experimental structure of a related protein as a template.[1−4] Many different programs are available for this purpose, and homology models of proteins are currently used in a wide variety of disciplines, ranging from drug design, to studies of mutations and protein engineering.[1] With the user-friendly modeling programs available, constructing a homology model of a protein is straightforward, but the quality of the results may vary a lot since automatic methods do not always find optimal alignments or loop predictions, especially when the sequence identity is below 40%.[1,5] An inaccurate homology model may be misleading, because relatively small structural errors may lead to large errors in e.g. binding energy calculations. Accuracies of the various homology model building methods are relatively similar when used optimally.[1,6] Other factors such as template selection and alignment accuracy usually have a larger impact on the model accuracy. Even homology models generated from very high quality sequence alignments might contain severe errors.[7] Hence, it is important to evaluate the quality of homology models made from high quality sequence alignments and to be able to predict the model quality for a given target-template pair. This is important both in order to select the correct template structures to use for the homology modeling and to evaluate whether useful information can be extracted from a future homology model. In this way, one can avoid spending time on modeling projects on target-template combinations that will not lead to models of sufficient quality.

Models of three-dimensional (3D) protein structures can be evaluated according to a variety of criteria, such as stereochemistry (bond lengths, bond angles, torsion angles, etc.), packing, formation of a hydrophobic core, residue and atomic solvent accessibilities, spatial distribution of charged groups, distribution of atom−atom distances, atomic volumes, and main-chain hydrogen bonding.[8−11] Large deviations from the most likely values have been interpreted as indicators of errors in the model structure. Methods based on 3D profiles and statistical potentials of mean force also exist, that take many of these criteria into account implicitly.[12−15] These methods evaluate the environment of each residue as seen in the model, compared to the expected environment as observed in experimental structures.

The accuracy of protein structure models can also be evaluated by comparison to experimental structures of the targets.[16−19] The most common method for comparison of two 3D structures is calculation of root-mean-square deviations (RMSDs) between corresponding atoms in the structures. However, the geometric measures only provide meaningful results when the entire extent of the proteins is comparable. For example, a set of partially correct structures cannot be ranked because the incorrect portions will dominate

* Corresponding author phone: +47 73 59 86 47; fax: +47 73 59 88 01; e-mail: kristin.tondel@ntnu.no.

10.1021/ci049924m CCC: $27.50
Published on Web 09/08/2004

an RMSD value. When restricting the comparison to certain parts of the structure, the choice of relevant parts may also be somewhat arbitrary. An alternative is to use the surface area of residue contacts, which does not require a superpositioning of the structures that are being compared. A new surface area based comparison method has been developed.[7] This method is similar to the Contact Area Difference (CAD) number[20] but differs in both calculational details and in the normalization of the CAD number. The surface areas are calculated using a Boolean logic based algorithm.[21] A two-dimensional matrix is constructed by calculating every pairwise contact surface area (A_{ij}) between all residues i and j in each protein structure. This matrix is referred to as the contact area matrix. When two protein structures are compared, the difference between the contact area matrices for the two structures is calculated. The elements in the resulting matrix are negative for incorrectly occurring and overestimated contacts, zero for correct contacts and non-contacting residue pairs, and positive for underestimated or missing contacts in the model structure. In the following, this matrix will be referred to as the interresidue contact area error matrix. Analysis of residue−residue contacts has been used to evaluate structure predictions[22] and the conservation of side-chain interactions in homologous proteins.[23] Contact-based measures can also be applied to simplified protein descriptions.[24]

The total, unnormalized CAD number for a reference structure R and a model structure M is given by eq 1.[20]

$$CAD = \sum_{i,j} |(A_{ij}^R - A_{ij}^M)| \quad (1)$$

The normalized CAD number is calculated as

$$CAD_{norm} = \frac{\sum_{i,j} |(A_{ij}^R - A_{ij}^M)|}{\frac{N}{2}(A_{ii}^R + A_{jj}^R)} \quad (2)$$

where N is the number of residues considered.[7]

All methods mentioned above for evaluation of the quality of protein structure models operate on the 3D models themselves. No methods exist that predict the model quality prior to the actual model building. Sequence identity between the target and template above 30% is a relatively good indicator of the expected homology model accuracy, but when the sequence identity is below 30% it becomes unreliable as a measure of the expected model quality.[1] Other measures of the similarity between the target and template amino acid sequences also exist, such as BLAST scores, which is a measure of the statistical significance of the alignment.[25] These measures are not family specific and provide no direct measure of the expected homology model quality, just an indication of the quality of the sequence alignment. Though a good sequence alignment is required to obtain a reliable homology model, a high statistical significance of the sequence alignment does not always imply high model accuracy. Predicting homology model quality is a difficult task and can only be done reliably within a specific protein family or for proteins related to the type of proteins for which the prediction method has been calibrated. The reason is that factors such as the effective mutation rate, the number and size of insertions and deletions, and the number of surface loops vary between protein families.[26] Even within a specific protein structure, some regions can be modeled with high accuracy, while others are more difficult to model. Loop modeling is known to be a difficult task, and much research is devoted to this part of the homology modeling procedure.[1,27,28] Loop modeling techniques range from searching databases of known protein structures for loops having similar end points, to molecular dynamics simulations.[1,29]

Prediction of the expected homology model quality, given a specific relationship between the primary structure of the target and template proteins, is useful for evaluating whether a homology model can be generated that suits the needs of the specific task. In some cases a model of very high accuracy is needed, while in other cases a model of lower quality can provide sufficient information. Misalignments are the largest source of errors in comparative modeling.[1] In this work, a new method for prediction of homology model accuracy has been developed, that operates only on the target-template sequence alignment. No information about the 3D structure is needed once the method has been calibrated, and the homology model quality can be predicted for a wide range of sequence identities. This method has been applied to the protein kinase family but can easily be extended to other protein families. RMSD values between the homology model structures and experimental structures of the same proteins, and differences in interresidue contact areas between the models and the target X-ray structures are used as measures of the model quality. The method presented here can be used to ensure that the correct templates and alignments are chosen, so that the best possible homology model is generated. It is also useful for identification of regions that are difficult to model, as well as errors in the alignment. Possibilities for improving the homology model quality by combination of several homology models are also discussed.

2. METHODOLOGY

A regression model has been developed for prediction of the accuracy of homology models of protein structures. This regression model was trained on 292 homology models of proteins for which experimental structures were available for comparison. Calculated RMSDs and differences in interresidue contact areas between the homology models and the target X-ray structures were used as measures of the accuracy of the homology models. The homology model quality data were used as dependent variables (**Y**) in the Partial Least Squares (PLS) regression analysis. The resulting regression model can be used to predict the accuracy of new homology models from the sequence alignment.

A matrix of alignment score profiles describing the similarity between the target and template amino acid sequences for each homology model was used as independent variables (**X**) in the regression analysis. Each element in this alignment score matrix contained the value of the Point Accepted Mutation (PAM) 250 similarity matrix[30] for a pair of amino acids at corresponding positions in the sequence alignment. Hence, for each homology model, a score value (corresponding to the PAM250 matrix value) for each pair of residues aligned in the sequence alignment used for the

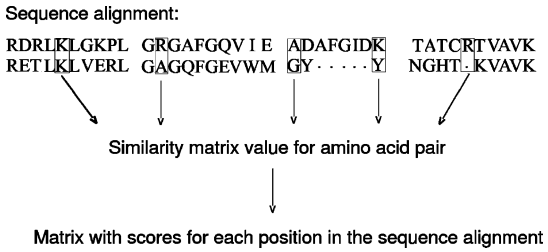

Figure 1. Generation of alignment score profiles.

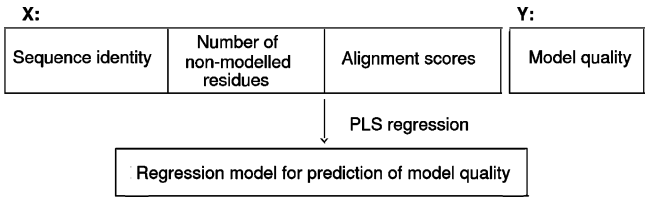

Figure 2. Multivariate analysis of the homology model quality data using alignment scores, sequence identity, and number of nonmodeled residues as independent variables.

modeling was found, resulting in a matrix of alignment scores, as illustrated in Figure 1. This describes how similar the target and template amino acid sequences are in each position in the alignment.

The PLS regression analysis is illustrated in Figure 2. In this model, the sequence identity between the target and the template and the number of nonmodeled residues (caused by gaps in the sequence alignment) were also added to the matrix of independent variables (**X**-matrix). A gap in the sequence alignment appears when an insertion or deletion has occurred during evolution, so that there is a region where the target and template structures differ in length. This often occurs in surface loops where the effective mutation rate is high. A high effective mutation rate means that, compared to other structural regions, a large fraction of the mutations that occur is leading to functional genes and results in changes in the amino acid sequence of the protein. Gaps in the sequence alignment lead to inaccuracies in the homology modeling and represent a great challenge when developing homology modeling methods. However, the effective mutation rate in active sites of protein structures is often relatively low,[26] so a homology model might be useful even though a part of the structure (e.g. a surface loop) is not modeled correctly.

Any other available information about the similarity between the target and template structures can also be added to the **X**-matrix to improve the predictive ability of the obtained regression model.

This regression model can be used to predict the accuracy of new homology models. Prior to constructing a new homology model, alignment score profiles can be generated from the sequence alignment between the target and the template. The regression model developed in this work can predict the homology model quality for new homology models from such alignment score profiles. This model can only be used within the protein family for which it has been trained, but similar regression models can be made in the same way for other protein families.

Outliers that should be kept out of the regression analysis can be identified by inspection of influence plots, that is, plots of the residual Y-variance against the leverage. Outliers that have a large effect on the results from the regression analysis will be placed in the upper, right-hand part of the influence plot. This can be used to identify members of a protein family that are difficult to model with homology modeling due to large deviations from the other proteins in the family.

The regression coefficients can be used to identify regions that are difficult to model, as well as alignment errors. Regions of the sequence alignment that contain many gaps (regions where the sequence alignment is of low quality) correspond to regions with large variations in the regression coefficients. Comparison of the residuals (for each alignment position) from prediction for a new homology model to the residuals for the homology models included in the regression analysis can also reveal errors in the sequence alignment. Such alignment errors will lead to deviations in the residual pattern.

2.1. Data Sets. *2.1.1. Protein Kinase Structures.* Two sets of protein kinase structures from the RCSB Protein Data Bank (PDB)[31] were selected for the homology modeling. One set (A) contains fourteen structures with pairwise sequence identities between 14% and 40% (Table 1). This set is a representative set of all protein kinase structures in the PDB. To maximize the structural diversity in the set, all pairs of structures in this set have sequence identity lower than 40%. Only structures with resolution better than 3 Å were selected.

The other set (B) consists of eleven protein kinase structures with pairwise sequence identities of 35−80% (Table 2). This is the sequence identity range where homology modeling is most frequently used. These eleven structures belong to the receptor tyrosine kinase (RTK) family. The RTK family was chosen because it is of great interest in e.g. drug design, and one of the families where experimental structures with the widest range of sequence identities are available in the PDB. When this work was carried out, the PDB contained 25 entries corresponding to protein kinases in the RTK family. To leave out multiple X-ray structures of the same proteins, only structures having lower sequence identities than 90% to each other were chosen. This resulted in 10 kinase structures, which provide a good representation of the structural diversity in the RTK family. In the interest of exploring the effects of conformational change, the apo-structure of the human insulin receptor protein kinase (HIRPK) was added to the set, so that the set now contains two copies of HIRPK.

For both sets of kinase structures, structures with a ligand in the adenosine triphosphate (ATP)-binding site and high-resolution structures were preferred when multiple structures of the same protein were present in the PDB.

The X-ray structures were superposed using the CE algorithm.[32] Pairwise sequence identities for the two sets of protein kinase structures are given in Tables 1 and 2. The corresponding Cα and Cβ RMSD values are given in the Supporting Information.

2.1.2. Homology Model Construction. A modeling pipeline has been developed for automatic all-against-all homology modeling from a multiple sequence alignment.[7] Two different homology modeling tools, WHAT IF (simple and advanced version)[33] and MODELLER,[1,27,34] can be used with this pipeline. This modeling pipeline has been used with the two sets of protein kinases described above. A multiple sequence alignment of the protein kinases in set A was created for another, separate research project.[35] This sequence alignment was based on a structural alignment made with the CE

Table 1. Pairwise Sequence Identities (%) for the Fourteen Protein Kinases in Set A[a]

PDB entry	1csn__	1b6c_b	1fgi_a	1ir3_a	2src__	1a6o__	1f3m_c	1hck__	1jnk__	1kob_a	1tki_a	1cdk_a	1phk__	1a06__
1csn__	100	21.4	17.6	19.1	19	16.2	18.3	19.7	18.5	17.8	14.1	20.7	17.6	15.9
1b6c_b	21.4	100	30.4	28.8	27	20.4	23	23.6	23.2	21.6	18.3	21.4	23	19.4
1fgi_a	17.6	30.4	100	37	39.8	18.5	23.8	25.5	26.7	22.4	20.1	21.9	19.6	25.4
1ir3_a	19.1	28.8	37	100	40.8	17.7	25	21.9	21.8	20.8	19.7	21.1	20.8	20
2src__	19	27	39.8	40.8	100	18.9	23.4	28.2	22.7	22.4	19.9	22	22.7	21.4
1a6o__	16.2	20.4	18.5	17.7	18.9	100	26.5	32.7	26.5	24	25.3	23.3	26.2	25.4
1f3m_c	18.3	23	23.8	25	23.4	26.5	100	32.2	27.9	29.8	26.5	28.1	30.4	31.2
1hck__	19.7	23.6	25.5	21.9	28.2	32.7	32.2	100	37.8	27.6	26.5	29.7	30.6	27.8
1jnk__	18.5	23.2	26.7	21.8	22.7	26.5	27.9	37.8	100	27.2	23.1	26.8	28.5	29.9
1kob_a	17.8	21.6	22.4	20.8	22.4	24	29.8	27.6	27.2	100	43.1	28.1	32.5	33.5
1tki_a	14.1	18.3	20.1	19.7	19.9	25.3	26.5	26.5	23.1	43.1	100	25.4	32.7	32.4
1cdk_a	20.7	21.4	21.9	21.1	22	23.3	28.1	29.7	26.8	28.1	25.4	100	34.6	32.4
1phk__	17.6	23	19.6	20.8	22.7	26.2	30.4	30.6	28.5	32.5	32.7	34.6	100	36.3
1a06__	15.9	19.4	25.4	20	21.4	25.4	31.2	27.8	29.9	33.5	32.4	32.4	36.3	100

[a] The entries are colored according to the similarity between the two proteins in each pair. Red: sequence identity < 30%, yellow: 30% ≤ sequence identity < 40%, green: 40% ≤ sequence identity < 50%, white: sequence identity ≥ 50%.

Table 2. Pairwise Sequence Identities (%) for the Eleven Protein Kinases in Set B[a]

PDB entry	1byg_a	1fgk_a	1fvr_a	1iep_a	1ir3_a	1irk__	1k3a_a	1qcf_a	1qpc_a	1vr2_a	2src__
1byg_a	100	38,7	35,9	45,2	37,4	37,4	36,2	42,2	44	39,4	43,6
1fgk_a	38,7	100	40,4	38,1	36,1	36,8	36,7	35,2	36,8	53	37,1
1fvr_a	35,9	40,4	100	41,2	34,9	34,9	35,8	35,8	36,4	38,6	35,7
1iep_a	45,2	38,1	41,2	100	39,8	40,2	43	48,1	48,1	38,5	48,3
1ir3_a	37,4	36,1	34,9	39,8	100	100	80,4	36,4	37,5	38,3	38,8
1irk__	37,4	36,8	34,9	40,2	100	100	79,4	36,7	37,1	38,3	39,1
1k3a_a	36,2	36,7	35,8	43	80,4	79,4	100	35,8	37	37,3	36,7
1qcf_a	42,2	35,2	35,8	48,1	36,4	36,7	35,8	100	75,7	41,5	66,4
1qpc_a	44	36,8	36,4	48,1	37,5	37,1	37	75,7	100	40,7	66,3
1vr2_a	39,4	53	38,6	38,5	38,3	38,3	37,3	41,5	40,7	100	35,5
2src__	43,6	37,1	35,7	48,3	38,8	39,1	36,7	66,4	66,3	35,5	100

[a] The entries are colored according to the similarity between the two proteins in each pair. Red: sequence identity < 30%, yellow: 30% ≤ sequence identity < 40%, green: 40% ≤ sequence identity < 50%, white: sequence identity ≥ 50%.

program[32] and manually edited based on prior knowledge about the functionality of different regions of the protein kinase structures. A multiple sequence alignment of the protein kinases in set B was first made using ClustalX.[36] This alignment was then aligned manually to the alignment of set A. For both sets of kinase structures, homology models were constructed for each sequence in the multiple sequence alignment of the set using, in turn, each of the other structures as template. This resulted in 292 homology models, made using templates having between 14 and 80% sequence identity to the target.

Phosphate groups were removed from phosphorylated tyrosine residues prior to the homology modeling, and crystallographic water molecules, ligands and ions were also purged from the template structures.

The advanced version of WHAT IF (WI-advanced)[33] was used for the homology modeling in this work. WHAT IF advanced maintains the backbone conformation of the template structure unchanged and models side-chains using a backbone-dependent rotamer library.[37] Insertions (gaps in the sequence alignment) are not modeled in this version of WHAT IF, and the resulting homology models thus frequently contain structural gaps. Since the method presented here for prediction of the homology model quality operates on the sequence alignment, loop modeling quality cannot be predicted by this method. However, the number of nonmodeled residues is taken into account in the prediction of the overall model quality.

2.1.3. Calculation of the Homology Model Accuracy. The homology models were verified by comparison to the experimental structures of the targets. Two different measures of the homology model quality were used: RMSD values (separate overall Cα, Cβ, and heavy atom (HA) RMSD) and difference in the interresidue contact areas between the target X-ray structure and the model structure.

The target and template X-ray structures were superposed using the CE algorithm.[32] Separate overall Cα, Cβ, and heavy atom RMSD values between targets and homology models were calculated using the rotation and translation matrices from the CE superpositioning of the target-template pair for the homology model.

Pairwise contact surface areas between the residues in each protein structure were calculated, using a new method that is similar to the Contact Area Difference (CAD) number,[20] but differs in both calculational details and in the normalization of the CAD number.[7] The surface areas are calculated using a Boolean logic based algorithm.[21] In this work, a 1.4 Å probe was used, along with a default set of van der Waals radii derived from the CHARMM27 force field,[38] to calculate the surface areas. Hydrogen atoms were ignored. The total, unnormalized CAD number was then used as a measure of the model accuracy.

2.1.4. Generation of Alignment Score Profiles. As a measure of the similarity between the target and template primary structures in each position in the sequence alignment, the value of the PAM250 similarity matrix[30] for that particular pair of amino acids was used. As mentioned above, separate multiple sequence alignments of each of the two sets of protein kinase structures were used for the homology modeling. To analyze the homology model quality for both sets of kinases simultaneously, the two sequence alignments used in the homology modeling were aligned to each other

as described above. The alignment scores were generated based on this common multiple sequence alignment (shown in Figure 4). To separate nonmodeled residues (caused by gaps in the sequence alignment) from modeled residues, the score value for a nonmodeled residue was set to −100.

2.2. Multivariate Regression Analysis of the Homology Model Quality Data. PLS regression in Unscrambler[39] was used to analyze the homology model quality data. Cα, C*β*, and heavy atom RMSD were analyzed together using PLS2, while a separate PLS1 model was made for the contact area error. The data set was centered prior to the regression analysis, and random leave-ten-out cross-validation was used. No variable selection was carried out. Outliers were detected by inspection of influence plots and removed from the analysis. The number of principal components (PCs) used was chosen by inspection of the explained Y-variation from the cross-validation.

Only cases where the target and template X-ray structures were in the same conformation (either active or inactive conformation) were considered, since a homology model made using a template structure in an active conformation cannot be compared to a target structure in an inactive conformation and vice versa.

2.3. Validation of the Method. The predictive ability of the regression model was validated by cross-validation, as described above. The ability of the regression coefficients to identify regions of the protein structures that are difficult to model was verified by comparison of the regression coefficient pattern with the multiple sequence alignment of the 23 kinases used to train the regression model.

To test whether this method can be used to detect alignment errors, an alternative alignment between two randomly chosen sequences from the multiple sequence alignment of protein structure set B, 1byg and 1fvr, was generated with ClustalX.[36] A new regression analysis of the contact area error similar to that described above was carried out, with 1byg and 1fvr kept out of the analysis. Using this alternative regression model, the contact area error for a homology model of 1fvr made using 1byg as template was predicted based on alignment scores generated from the alternative sequence alignment. The X-residuals from the prediction were calculated and compared to the mean residuals for all homology models included in the regression analysis (± two standard deviations).

As a verification of the method's ability to guide the choice of homology modeling templates, templates were chosen on the basis of predicted homology model quality from the cross-validation of the PLS regression model. The number of correct template choices made on the basis of the predicted homology model quality was compared to the success rate when BLAST scores[25] and the target-template sequence identity, respectively, were used to guide the template choice. Choices were made only between protein structures included in the PLS regression analysis. In each case, the optimal template choice was defined as the template structure corresponding to the highest calculated model quality obtained for the given target protein structure.

To get an idea of whether a combination of different homology models might improve the model quality, the average backbone Cα atom positions between all homology models of each protein was calculated. This resulted in an average backbone conformation for each protein, based on all homology models of that protein. For each homology model, the distance from this average model was calculated for all residue positions. To test whether the average model would perform better than the individual homology models, the average (over all residues in the model) distance of each model from the average model was correlated to the model quality.

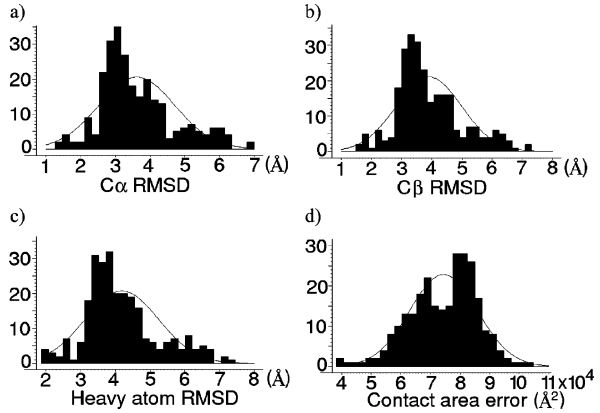

Figure 3. Histograms over (a) Cα, (b) C*β*, and (c) heavy atom RMSD values (Å) and (d) contact area differences (Å²) between the homology models obtained with WI-advanced and the target X-ray structures. Results for both kinase structure sets are shown in all histograms.

3. RESULTS

3.1. Data Sets. *3.1.1. Calculated Homology Model Accuracy.* Histograms over the obtained RMSD values and contact area differences between the homology models and the target X-ray structures are given in Figure 3.

The histogram in Figure 3(a) shows that most of the homology models have Cα RMSD values between 2.5 and 4 Å. The relatively high RMSD values are probably caused by the wide range of sequence identities between the targets and templates used for the homology modeling.

3.1.2. Alignment Score Profiles. Figure 4 shows the sequence alignment that was used to generate the alignment scores for the 23 proteins studied, together with the alignment score profiles for all homology models.

3.2. Regression Model for Prediction of Homology Model Quality. The homology model quality data set was analyzed with PLS regression. The predicted (from cross-validation) Cα, C*β*, and heavy atom RMSD for the homology models are shown in Figure 5. The predicted contact area error from the cross-validation is shown in Figure 6. The PLS score plots from the regression analysis show a separation of the samples according to sequence identity between the target and template. Statistical data (correlation coefficients (q) and explained Y-variation) from the regression analysis are given in Tables 3 and 4.

The results from this multivariate regression analysis show that the prediction of the contact area error is better than the RMSD value prediction. Interresidue contact area errors are not affected to that extent by one single loop conformation and are not dependent on structural superposition.[7]

The results presented in Figures 5 and 6 show that the homology model quality can be predicted with relatively high accuracy for the protein kinase family. Hence, the quality of future homology models can be predicted from alignment score profiles generated from substitution matrices. Similar regression models can be made for other protein families.

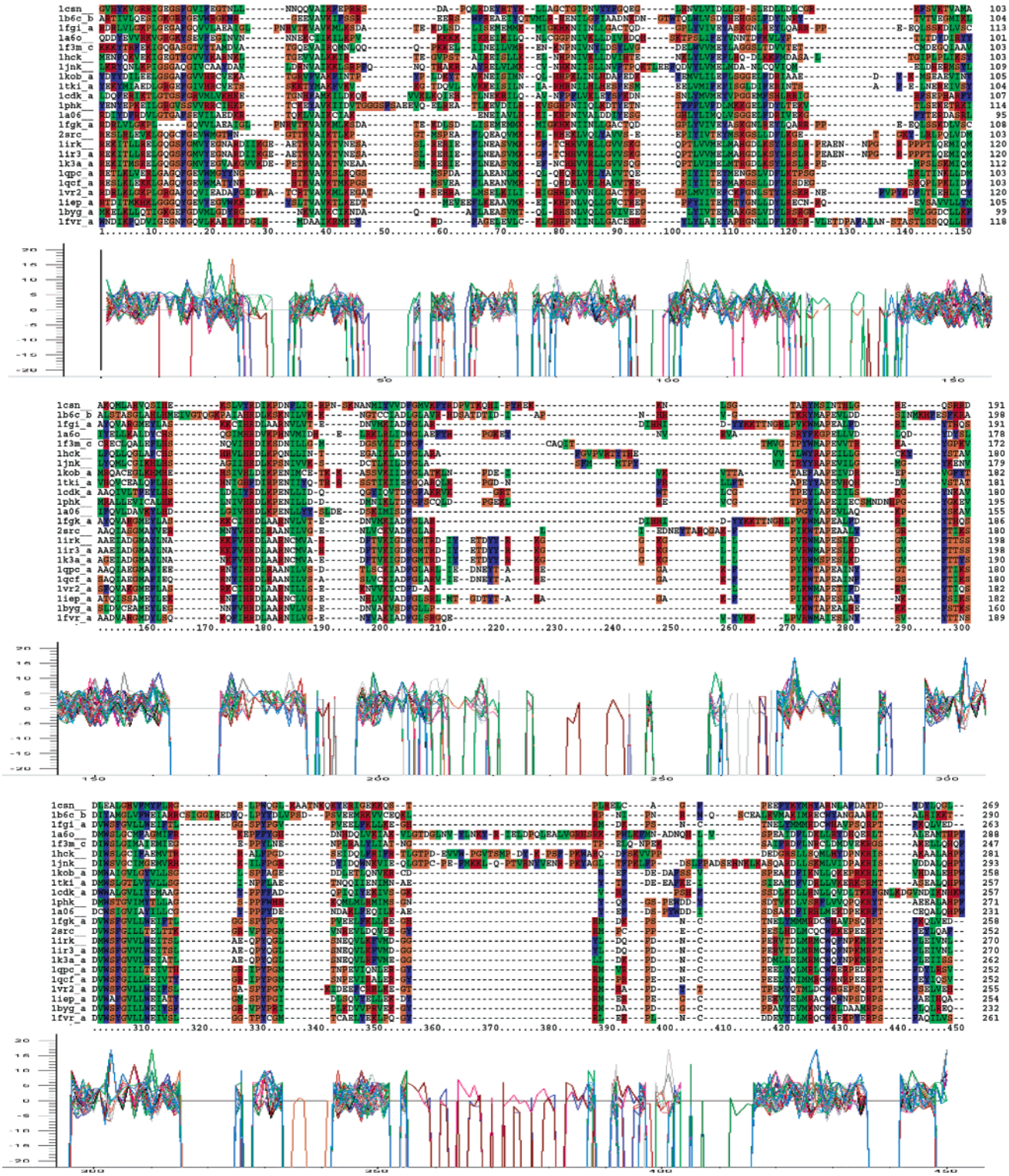

Figure 4. Multiple sequence alignment of the 23 protein kinases studied. This sequence alignment was used to generate the alignment score profiles shown below the alignment. The values on the horizontal axis correspond to the alignment positions. Score profiles for all homology models are shown.

3.3. Validation of the Method. Figure 7 shows the regression coefficients from the regression analysis. The regression coefficients from the regression model for the contact area error are shown together with the multiple sequence alignment used to generate the alignment score profiles in Figure 8. A comparison of the regression coefficients with the multiple sequence alignment shows that regions of the sequence alignment that contain many gaps (regions where the sequence alignment is of low quality) correspond to regions with large variations in the regression coefficients. Hence, the regression coefficients can be used to identify regions that are difficult to model, as well as alignment errors.

The residuals from prediction of the model quality for new homology models can be used to identify proteins that are difficult to model with homology modeling due to large deviations from the other members of the protein family. As

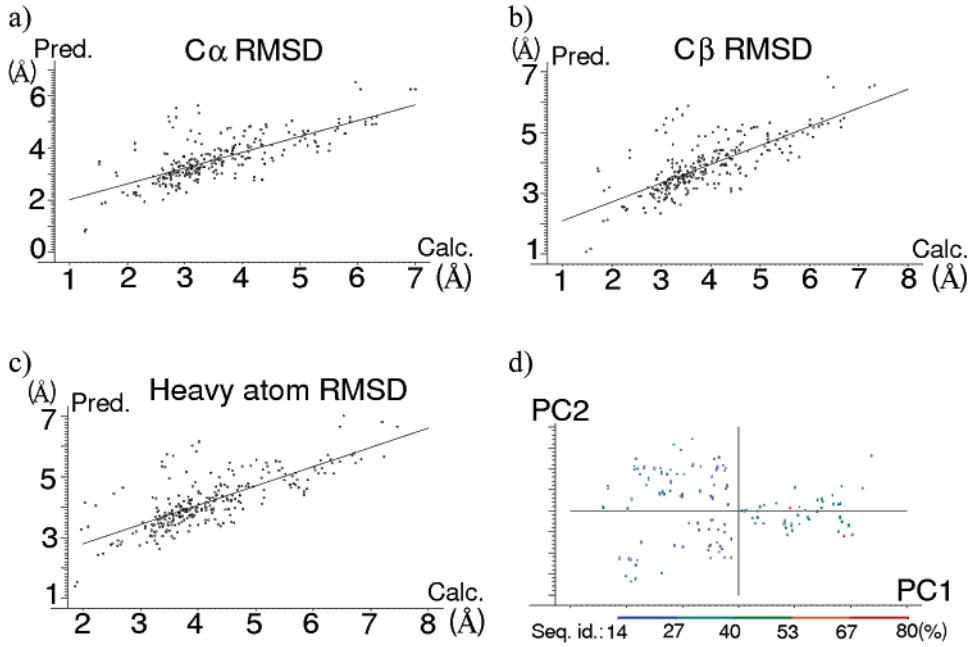

Figure 5. Predicted (from cross-validation) versus calculated values for (a) Cα RMSD (Å), (b) Cβ RMSD (Å), and (c) heavy atom RMSD (Å) for the homology models made using WHAT IF advanced. The PLS score plot (d) of PC2 vs PC1 is also shown. The samples are colored according to sequence identity between target and template.

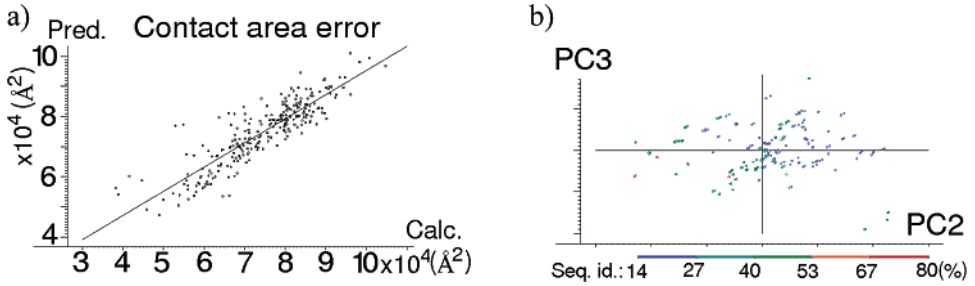

Figure 6. (a) Predicted (from cross-validation) versus calculated contact area error (Å^2) for the homology models made using WHAT IF advanced. (b) The PLS score plot of PC3 vs PC2. The samples are colored according to sequence identity between target and template.

Table 3. Statistical Data for the PLS2 Regression Model for Cα, Cβ, and Heavy Atom (HA) RMSD[a]

principal components	q (Cα RMSD)	q (Cβ RMSD)	q (HA RMSD)	explained Y-variation (%)
13	0.73	0.75	0.76	55.6

[a] The statistical data are from the cross-validation results.

Table 4. Statistical Data for the PLS1 Regression Model for the Contact Area Error[a]

principal components	q	explained Y-variation (%)
13	0.88	77.3

[a] The statistical data are from the cross-validation results.

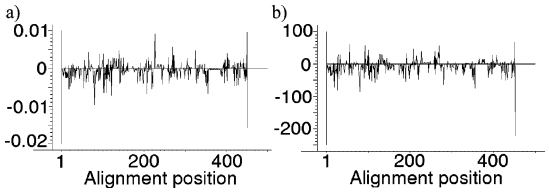

Figure 7. Regression coefficients from the regression analysis of (a) the heavy atom RMSD values and (b) the contact area error. The numbers on the horizontal axis correspond to the alignment positions.

explained earlier, such outliers can be identified by inspection of influence plots. The influence plots in Figure 9 show that no outliers that have a large effect on the results are present. The kinase structures with PDB entries 1f3m and 1b6c were previously removed from the PLS regression analysis because they were outliers.

Comparison of the residuals (for each alignment position) from prediction for a new homology model to the residuals for the homology models included in the regression analysis can e.g. reveal errors in the sequence alignment. Such alignment errors will lead to deviations in the residual pattern. To test this hypothesis, an alternative alignment between 1byg and 1fvr was generated. This alignment is shown together with the original alignment in Figure 10. Comparison of these two alignments reveals that the new alignment contains several deviations from the original alignment. Since the original alignment was corrected based on prior knowledge about the functionality of protein kinases, that alignment is more likely to be correct than the alternative one. Based on the alternative alignment, new alignment scores were generated, and the X-residuals from prediction of the contact area error were calculated using a regression model that had not been trained on 1byg and 1fvr. These residuals were

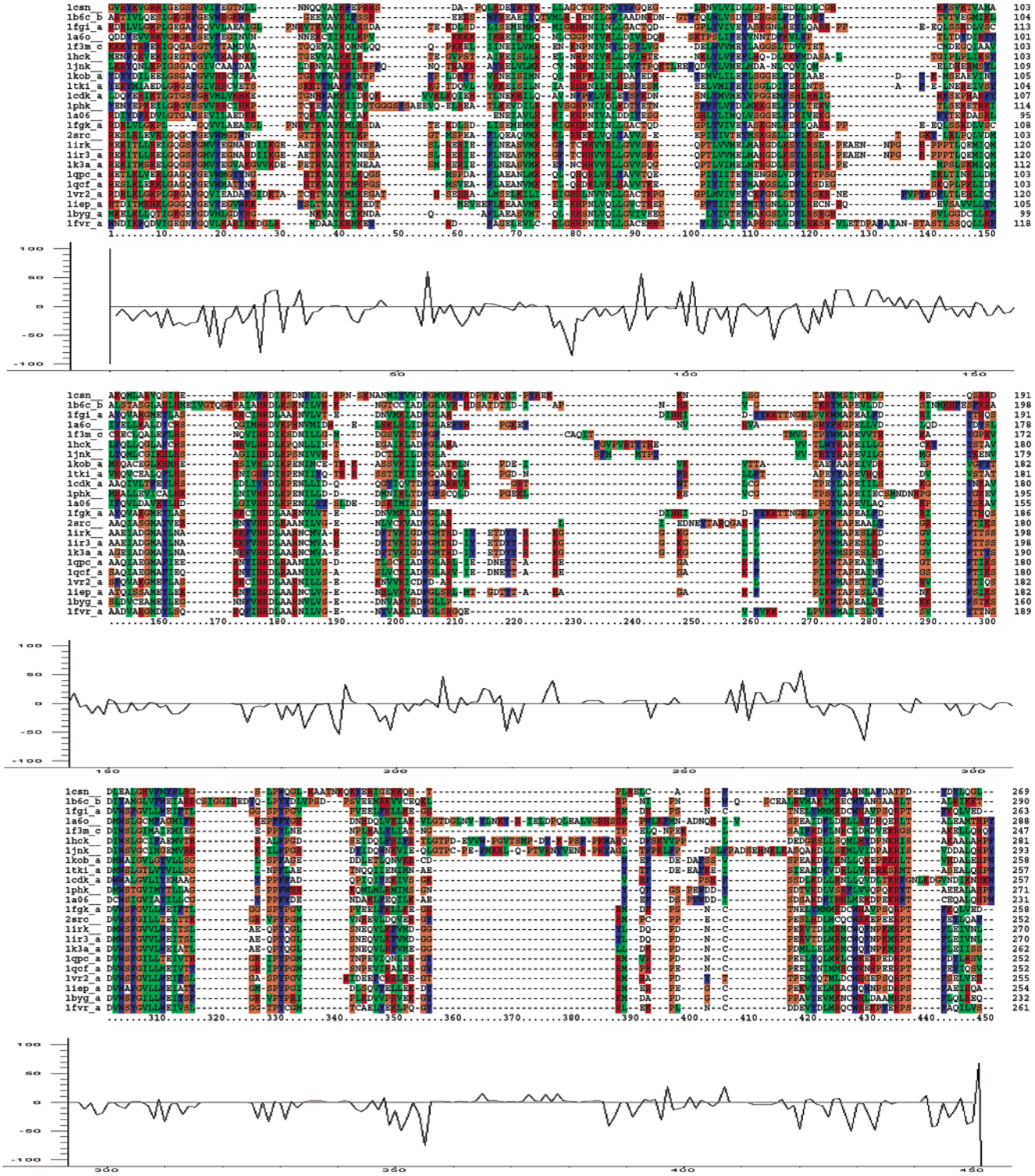

Figure 8. Regression coefficients from the regression analysis of the contact area error shown together with the multiple sequence alignment used to generate the alignment score profiles. The numbers on the horizontal axis correspond to the alignment positions.

compared to the residuals for all homology models for which the regression model was trained. The results are shown in Figure 10.

Comparison of the two alignments of 1byg and 1fvr and the curves in Figure 10 shows that in regions where the two alignments differ, the residuals for 1byg and 1fvr have large deviations from the mean residuals for the homology models included in the regression analysis. Hence, the X-residuals can provide useful information about alignment errors. As seen from Figure 10, there are a couple of regions where

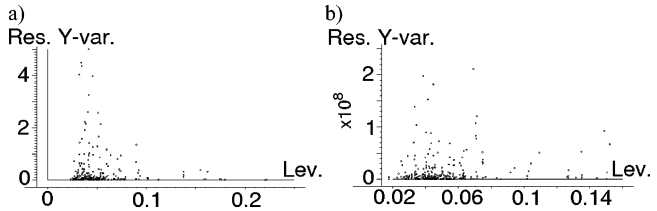

Figure 9. Influence plots (residual Y-variance versus leverage) from the regression analysis of (a) the RMSD values and (b) the contact area error.

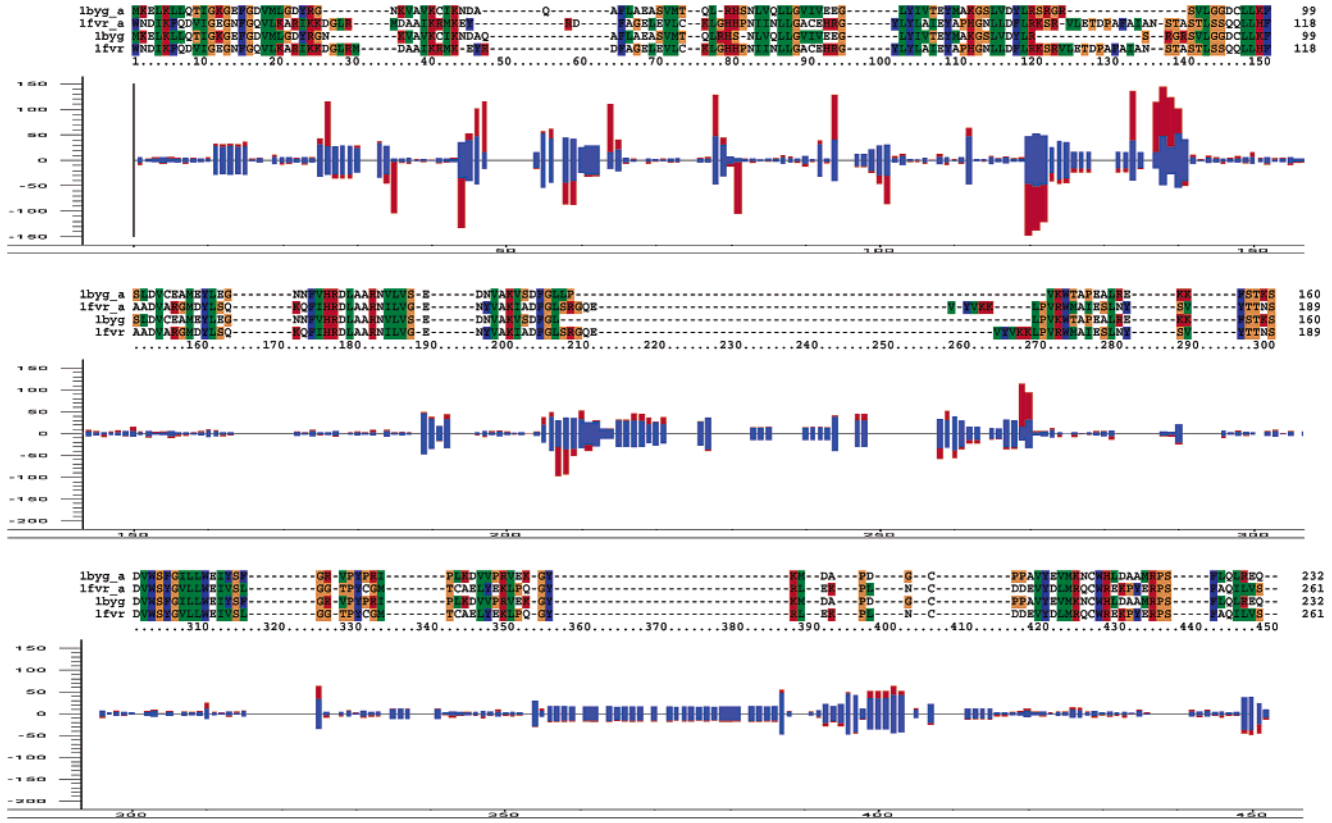

Figure 10. Alternative alignment between 1byg and 1fvr (last two sequences), aligned to the original alignment (first two sequences) from the multiple sequence alignment in Figure 4. The X-residuals from the prediction of the contact area error based on the alternative alignment of 1byg and 1fvr (red curve) are compared to the mean X-residuals for all homology models included in the regression analysis ± two standard deviations (blue curve). The numbers on the horizontal axis correspond to the alignment positions.

the residuals for 1byg and 1fvr deviate from the mean residuals even though the alignments are identical. This can e.g. be caused by large deviations between 1byg and 1fvr and the other structures included in the analysis in these parts of the sequence alignment. Hence, such deviations from the mean can also be caused by other factors than alignment errors and can only be used to identify regions where a closer look at the alignment might be necessary.

The ability of the PLS regression model to guide the choice of templates for the homology modeling was verified by comparison of success rates to those obtained using BLAST scores and target-template sequence identities, respectively, as indicators. The results given in Table 5 show that the method presented here performs significantly better than both BLAST scores and sequence identities in guiding the choice of homology modeling template. The presented method is especially useful in cases when only templates of low sequence identity to the target are available.

To test whether an average model would perform better than the individual homology models, the average (over all residues in the homology model) distance of each homology model from the average model was correlated to the model quality. Only the data for the proteins in kinase structure set B (sequence identities of 35−80%) gave a significant correlation. The results are shown in Figure 11.

The fact that the homology model quality is positively correlated with the distance from the average backbone conformation indicates that using a combination of several homology models might improve the model quality. Keeping the four marked outliers in Figure 11 out gives a correlation coefficient of 0.64 between the model error and the average distance from the mean model. Generation of plots such as the one shown in Figure 11 can be used to identify cases where a single template performs better than a combination of several templates. An example of a target-template pair where using a single template gives the best result is 1k3a and 1ir3. These two structures are so similar that using multiple templates in combination would probably introduce errors to the homology model. The sequence identity between 1k3a and 1ir3 is 80.4%. Such target-template pairs are placed in the lower, right-hand part of the plot in Figure 11, since the homology model quality will be high even though the homology model differs a lot from the average model. Hence, when the similarity between the target and template structures is high, using a single template is probably better than using an average model, since the template structure is more similar to the target than an average model will be. The other templates will make the average model differ more from the target.

4. DISCUSSION

The method presented here provides a new way to predict the quality of homology models directly from the sequence alignment between the target and template sequences. This method can be used prior to the actual homology model generation and allows the user to find out in advance whether it is possible to generate a model with the required accuracy. This is new, since existing methods for model quality

Table 5. Optimal Homology Modeling Templates (PDB Entries) and Templates Chosen on the Basis of Predicted Model Quality from the PLS Regression, BLAST Scores, and Target-Template Sequence Identity, Respectively[a]

		chosen template structure		
protein structure	optimal template	predicted model quality from the PLS regression	BLAST scores	target-template sequence identity
1csn__	1a06_	1a06_	*b*	1cdk_a (0.06)
1fgi_a	1ir3_a or 1vr2_a	1ir3_a	1vr2_a	2src__ (0.65)
1ir3_a	1k3a_a	1k3a_a	1k3a_a	1k3a_a
2src__	1qcf_a	1qpc_a (1.11)	1qcf_a	1qcf_a
1a6o_	1f3m_c or 1hck__	1f3m_c	1hck__	1jnk__ or 1f3m_c
1hck__	1a06_ or 1a6o_	1a06_	1jnk__ (0.74)	1jnk__ (0.74)
1jnk__	1a06_ or 1a6o_	1phk__ or 1a6o_	1hck__ (0.43)	1hck__ (0.43)
1kob_a	1tki_a	1tki_a	1tki_a	1tki_a
1tki_a	1kob_a	1kob_a	1kob_a	1kob_a
1cdk_a	1phk__	1phk__	1a06_ (1.54)	1phk__
1phk__	1cdk_a or 1a06_	1a06_	1a06_	1a06_
1a06_	1tki_a or 1kob_a	1kob_a	1phk__ (0.50)	1phk__ (0.50)
1byg_a	2src__, 1qpc_a or 1iep_a	1qpc_a or 1iep_a	1iep_a	1iep_a
1fgk_a	1ir3_a or 1vr2_a	1vr2_a	1vr2_a	1vr2_a
1fvr_a	1byg_a	1byg_a	1fgi_a (0.20)	1iep_a (1.45)
1iep_a	1byg_a	1byg_a	1qcf_a (1.22)	2src__ (1.18)
1irk__	1vr2_a	1vr2_a	1k3a_a (2.37)	1k3a_a (2.37)
1k3a_a	1ir3_a	1ir3_a	1ir3_a	1ir3_a
1qcf_a	2src__	1qpc_a (2.68)	2src__	1qpc_a (2.68)
1qpc_a	1qcf_a	2src__ or 1qcf_a	1qcf_a	1qcf_a
1vr2_a	1fgk_a	1k3a_a (0.82)	1fgi_a	1fgk_a
success rate (%)		85.7	61.9	57.1

[a] In cases where nonoptimal templates were chosen, the deviations from the lowest obtained Cα RMSD value (Å) are given in parentheses. The optimal template was defined as the template structure corresponding to the highest calculated model quality obtained for the given target protein structure. In some cases several templates correspond to homology models of approximately equal quality. Choices were made only between protein structures included in the PLS regression analysis. [b] No suitable template structure found among the structures included in the PLS regression analysis.

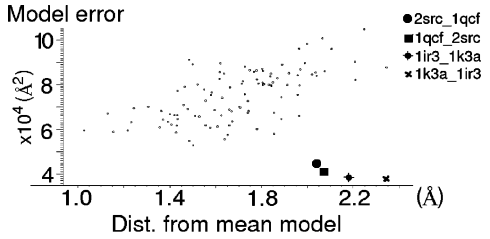

Figure 11. Contact area error (Å^2) for the homology models of the proteins in kinase structure set B versus the average (over all residues in the homology model) distance of each homology model from the average backbone conformation (Å) (2src_1qcf means the homology model of 2src made using 1qcf as template, and likewise for the other homology models).

prediction work on the protein structure models. However, a set of homology models of known accuracy is needed for calibration of the regression model that is used for prediction. To use this method, experimental structures for a representative set of the proteins on which the method is going to be used have to be available.

Time spent generating low quality homology models can be saved by using this method to rule out cases in which homology modeling is likely to fail and when it may succeed. The correct templates to use for the homology modeling can thereby more effectively be found. Using the method developed here, the model quality can be predicted based on a sequence alignment, possible problem regions can be identified using for example the residuals from the regression analysis, and corrections to the sequence alignment can be made. The impact of the alignment corrections on the expected homology model accuracy can then be tested by predicting the accuracy for the new alignment. Hence, this method is an effective tool for optimizing the sequence alignment prior to generating the homology model. Different homology modeling methods may also perform differently for a given sequence alignment. Using the approach developed in this work, regression models can be made that can predict the homology model quality resulting from several different homology modeling methods. This can guide the choice of modeling method and ensure that the best possible homology model is generated, given a certain sequence alignment. This method can also be used to find out whether an alignment corrected by an expert would do significantly better in the homology modeling than an automatically generated alignment, by comparison of the predicted homology model quality for the two types of alignments. In this work the alignments used for calibration were manually corrected, but this is not necessarily required for use of this method. Using an automatically generated alignment will give an estimate for the homology model accuracy one is likely to obtain with this type of alignment. This will be a measure of the lower limit for model accuracy, since a manually corrected sequence alignment is likely to give better results than an automatically generated one. However, the gain in obtained model quality is not necessarily large enough to be worth the extra work on the alignment. Hence, time can be saved by using this method to rule out in what cases manual alignment corrections give significantly better homology models, and in what cases one can equally well use an automatically generated alignment. A large limitation with this method is that the accuracy of the loop modeling cannot be predicted from the sequence alignment. However, the size and number of gaps can be taken into account in the regression modeling. This is relevant for the choice of

modeling templates.

The method presented here makes use of observed model qualities for already generated homology models for calibration and provides a protein family specific measure of the overall expected homology model quality not only the similarity between the target and template amino acid sequences. The methods used by alignment generation programs to identify the optimal alignment and templates are not based on observed homology model accuracy and are not protein family specific. A statistically significant sequence alignment does not necessarily mean that the accuracy of the generated homology model is high. Many other factors also affect the model accuracy, such as the choice of amino acid side-chain conformations. Since separate regression models can be made for different protein families and different homology modeling methods, homology model quality prediction can also guide the choice of modeling method. In some cases it is best to model different domains of the protein structure separately. Homology model quality prediction is useful for identifying what domains to model separately and what templates to use for the different domains. Plots of the regression coefficients from the regression analysis can be used to identify regions that are difficult to model, and X-residuals from the prediction can be used to detect alignment errors. Influence plots can be used to detect members of the protein family that will be difficult to model due to large deviations from the other members of the family.

In this work, the method has been used on a specific protein family, but the regression model can be made reliable for more than one protein family by including data for more proteins in the calibration. The number of structures that is needed for reliable predictions of the model quality depends on the diversity of the structures on which the model will be used. The method presented here can be used to get an idea of whether it is possible to generate reliable homology models for all proteins in a family using the experimental structures that are available and identify members of the family that are difficult to model with homology modeling. This gives useful information about what members of a protein family one needs experimental structures of, to adequately represent the structural diversity in the family.

Combination of several template structures in the homology modeling is widely used. The underlying idea is that multiple template structures provide more information than a single structure does. If the correct template structures are chosen, this is probably true. However, including structural information from template structures that do not have the required similarity to the target may introduce errors in the final homology model. In this case, using a single template with high similarity to the target is better than using this template in combination with other templates of lower similarity. The method presented here can be used to find the optimal combination of templates, and in which cases using a single template may give the best result. This can be done by generation of PLS regression models that use e.g. the average score value for the templates in each alignment position. In this way, the predicted homology model quality can be compared for models made using a single template and models based on multiple sequence alignments. The optimal number and combination of templates can thereby be found. Using an average score value for the templates in each alignment position will represent the lower limit for model accuracy that one is likely to obtain with the given templates, since a more optimal combination of the templates than just a simple averaging is likely to give better homology models.

One problem with most homology modeling methods that use a combination of multiple template structures is that a primary template (typically the one having the highest sequence identity to the target) is chosen, and information from the other template structures is often only used in gap regions. This makes the homology model very dependent on the primary template, and often this results in a model that is more similar to the primary template than to the target. This is, however, only the case when there are errors in the target-template alignment used for the homology modeling.[1] Hence, this technique is most useful in cases where a template structure of relatively high sequence identity is available. It is also difficult to obtain a correct sequence alignment in cases where the templates have low sequence identity to each other and to the target. In cases where only template structures of low sequence identity are available, including structural information from all templates along the entire sequence might be better than choosing one of them as a primary template. A reasonable question is therefore as follows: Is it possible to combine several homology models of low overall quality by using e.g. a weighted average of the backbone positions for each residue? The weights for each homology model should vary according to the similarity to the target sequence in that region. In this way, each homology model would contribute differently in different regions according to the local similarity to the target. One way to weight this average would be to use alignment score profiles such as those generated here, since they have been shown to be correlated with the obtained homology model quality. By using a weighted average, local similarities between the target and the different templates can be taken into account, and information from all templates can be effectively utilized. The target-template similarity might vary between the different regions of the protein structure. Different homology modeling methods might also perform differently in different regions of the protein. Hence, a combination of homology models generated using several modeling methods might improve the model quality. One problem with this procedure is that averaging the side-chain positions does not make sense. Hence, the side-chain conformations have to be determined after the backbone average is calculated. By generation of regression models calibrated for this type of homology models, the model quality can be predicted for different homology model combinations. A weighted average of the score values can then be used for each alignment position.

5. CONCLUSIONS

A new method for prediction of homology model quality has been presented, which is a useful tool for e.g. selection of template structures for the homology modeling and detection of alignment errors. This method can also be used to identify problem regions of a protein structure, as well as proteins that are difficult to model with homology modeling due to large deviations from the other members of the protein family. It will also be a useful tool for

improving the homology model quality by combination of several homology models. This method has been applied to protein kinases and can easily be extended to other protein families.

ACKNOWLEDGMENT

Thanks to Eric D. Scheeff and Philip E. Bourne at San Diego Supercomputer Center for providing the multiple sequence alignment of the protein kinases in set A. Michael Gribskov at San Diego Supercomputer Center is thanked for helping with the alignment of the kinases in set B. Thanks to Jens E. Nielsen for providing the homology modeling pipeline and to Stewart Adcock for helping with the calculations of the interresidue contact areas. Thanks also to Prof. J. Andrew McCammon and his research group at the Department of Chemistry and Biochemistry at University of California, San Diego, for allowing me to visit their group and carry out the homology modeling work there. Dr. Finn Drabløs at the Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology and Endre Anderssen at the Department of Chemistry, Norwegian University of Science and Technology are thanked for helpful discussions. The Norwegian Research Council is thanked for financial support.

Supporting Information Available: Pairwise Cα and Cβ RMSD values for the two sets of protein kinase structures used are given in Tables S1 and S2. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Marti-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291−325.
(2) Baker, D.; Sali, A. Protein structure prediction and structural genomics. *Science* **2001**, *294*, 93−96.
(3) Al Lazikani, B.; Jung, J.; Xiang, Z.; Honig, B. Protein structure prediction. *Curr. Opin. Chem. Biol.* **2001**, *5*, 51−56.
(4) Schonbrun, J.; Wedemeyer, W. J.; Baker, D. Protein structure prediction in 2002. *Curr. Opin. Struct. Biol.* **2002**, *12*, 348−354.
(5) Qian, B.; Goldstein, R. A. Optimization of a new score function for the generation of accurate alignments. *Proteins: Struct., Funct., Genet.* **2002**, *48*, 605−610.
(6) Koehl, P.; Levitt, M. A brighter future for protein structure prediction. *Nat. Struct. Biol.* **1999**, *6*, 108−111.
(7) Liu, J.; Tøndel, K.; Adcock, S.; Gribskov, M.; Niedner, H. R.; McCammon, J. A.; Nielsen, J. E. Homology modeling of protein kinases. Unpublished results.
(8) Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **1993**, *26*, 283−291.
(9) Laskowski, R. A.; Rullmann, J. A. C.; MacArthur, M. W.; Kaptein, R.; Thornton, J. M. AQUA and PROCHECK−NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **1996**, *8*, 477−486.
(10) Oldfield, T. J. Squid: a program for the analysis and display of data from crystallography and molecular dynamics. *J. Mol. Graphics* **1992**, *10*, 247−252.
(11) Hooft, R. W. W.; Sander, C.; Vriend, G. Verification of protein structures: side-chain planarity. *J. Appl. Crystallogr.* **1996**, *29*, 714−716.
(12) Lüthy, R.; Bowie, J. U.; Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* **1992**, *356*, 83−85.
(13) Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 355−362.
(14) Topham, C. M.; Srinivasan, N.; Thorpe, C. J.; Overington, J. P.; Kalsheker, N. A. Comparative modelling of major house dust mite allergen Der p I: structure validation using an extended environmental amino acid propensity table. *Protein Eng.* **1994**, *7*, 869−894.
(15) Melo, F.; Feytmans, E. Assessing protein structures with a nonlocal atomic interaction energy. *J. Mol. Biol.* **1998**, *277*, 1141−1152.
(16) Venclovas, C.; Zemla, A.; Fidelis, K.; Moult, J. Criteria for evaluating protein structures derived from comparative modeling. *Proteins: Struct., Funct., Genet.* **1997**, *Suppl. 1*, 7−13.
(17) Moult, J.; Fidelis, K.; Zemla, A.; Hubbard, T. Critical assessment of methods of protein structure prediction (CASP): Round IV. *Proteins: Struct., Funct., Genet.* **2001**, *Suppl. 5*, 2−7.
(18) Moult, J.; Fidelis, K.; Zemla, A.; Hubbard, T. Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins: Struct., Funct., Genet.* **2003**, *53*, *Suppl. 6*, 334−339.
(19) Cristobal, S.; Zemla, A.; Fischer, D.; Rychlewski, L.; Elofsson, A. A study of quality measures for protein threading models. *BMC Bioinformatics* **2001**, *2*, *5*.
(20) Abagyan, R. A.; Totrov, M. M. Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. *J. Mol. Biol.* **1997**, *268*, 678−685.
(21) le Grand, S. M.; Merz, K. M., Jr. Rapid approximation to molecular surface area via the use of Boolean logic and look-up tables. *J. Comput. Chem.* **1993**, *14*, 349−352.
(22) Göbel, U.; Sander, C.; Schneider, R.; Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins: Struct., Funct., Genet.* **1994**, *18*, 309−317.
(23) Russell, R. B.; Barton, G. J. Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J. Mol. Biol.* **1994**, *244*, 332−350.
(24) Gou, Z. Y.; Thirumalai, D. Kinetics of protein folding: Nucleation mechanism, time scales and pathways. *Biopolymers* **1995**, *36*, 83−102.
(25) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403−410.
(26) Lesk, A. M.; Chothia, C. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **1980**, *136*, 225−270.
(27) Fiser, A.; Do, R. K.; Sali, A. Modeling of loops in protein structures. *Protein Sci.* **2000**, *9*, 1753−1773.
(28) Mehler, E. L.; Periole, W.; Hassan, S. A.; Weinstein, H. Key issues in the computational simulation of GPCR function: representation of loop domains. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 841−853.
(29) Wieman, H.; Tøndel, K.; Anderssen, E.; Drabløs, F. Homology-based modelling of targets for rational drug design. *Mini-Reviews Med. Chem.* **2004**, *4*, 793−804.
(30) Schwartz, R. M.; Dayhoff, M. O. Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. A perspective is derived from protein and nucleic acid sequence data. *Science* **1978**, *199*, 395−403.
(31) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(32) Shindyalov, I. N.; Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **1998**, *11*, 739−747.
(33) Vriend, G. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **1990**, *8*, 52−56.
(34) Sali, A.; Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234*, 779−815.
(35) Scheeff, E. D.; Bourne, P. E. Evolution of the protein kinase-like superfamily, from a structural perspective. Unpublished results.
(36) Thompson, J. D.; Gibson, T. J.; Plewniak, F.; Jeanmougin, F.; Higgins, D. G. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **1997**, *25*, 4876−4882.
(37) Chinea, G.; Padron, G.; Hooft, R. W.; Sander, C.; Vriend, G. The use of position-specific rotamers in model building by homology. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 415−421.
(38) MacKerell, A. D.; Brooks, B.; Brook III, C. L.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. In *Encyclopedia of Computational Chemistry*; Schleyer, P. R., Ed.; John Wiley & Sons: New York, 1998; pp 271−277.
(39) The Unscrambler, Version 7.6 SR-1, CAMO ASA, 2000.

CI049924M