ASSESSING PROTEIN MODELS: AN EVALUATION OF THE PERFORMANCE OF DIFFERENT SCORE TYPES

N. Benhabiles¹, P. Lackner¹, F.S. Domingues¹, S. Wegenkittl², A. Prlic¹ and M.J. Sippl¹,²

¹Center for Applied Molecular Engineering, University of Salzburg, Jakob Haringer Strasse 3, Salzburg, A–5020, Austria, ²Proceryon Biosciences GmbH, Jakob Haringer Strasse, 3, A–5020 Salzburg, Austria. Corresponding author: M.J. Sippl. E–mail: sippl@came.sbg.ac.at.

ABSTRACT

We examine the performance of different scoring schemes for the evaluation of structural models of proteins. The scores investigated include knowledge based potentials, secondary structure prediction and spatial compactness of the alignment. We find that the most discriminative scores are alignment score and secondary structure score. The alignment score is composed of knowledge based potentials and amino acid substitution scores. Using these scores, a coverage of approximately 30% is obtained for error rates below 1%.

INTRODUCTION

The prediction of proper structural models for amino acid sequences is required in many areas of molecular biology. For instance, an exciting new concept is the use of structural models in virtual screening. To be successful an obvious requirement is that the models generated reflect the structural features in sufficient detail and accuracy.

Fold recognition is frequently used to derive approximate models if no homologous structure is known. The quality of models is usually judged in terms of various scores. However, since fold recognition models are often incomplete, where side chains and loops are missing, discriminating between native like and missfolded models is quite demanding. To derive statistical sound measures of model quality it is necessary to study the scores using a data set reflecting the intended application and to determine the coverage and error rates that can be achieved.

In this work we systematically examine the performance of a set of scores based on amino acid substitution matrices, knowledge based potentials, predicted secondary
structures, measures of the spatial compactness, and their combinations. The data set consists of structurally related proteins that have no obvious sequence similarity which can be detected by FASTA (1). We find that the scores are pairwise uncorrelated in terms of regression analysis. A score combining knowledge based potentials with amino acid substitution scores, and a score based on predicted secondary structure are most discriminative. If length correction is used, these two scores reach coverage rates around 30% each.

METHODS

Database of structurally related proteins

The protein pairs used in this study were derived using CATH1.6 (2), a hierarchical domain classification of the protein structures in the PDB (3). The set consists of 516 related and 32141 unrelated protein pairs. The subset of related proteins consists of homologous and analogous pairs. Homologous pairs were derived using the same Class (C), Architecture (A), Topology (T) and Homology (H) for different Sequence (S) levels while analogous pairs were derived using the same C, A, T levels for different H and S levels. Cases with low structural similarity, having less than 35 structural equivalent residues as measured by ProSup (4), were excluded. Fold recognition is used to identify relationships when there is no clear sequence similarity. Therefore, pairs having clear sequence homology, i.e. E–values smaller than 0.01 according to FASTA (1), were excluded from the subset of related proteins. Unrelated pairs were derived using either different C, different A, or different T CATH levels. Pairs with more than 20 equivalent residues according to ProSup were removed to reduce the risk of having wrongly assigned unrelated pairs.

Generating threading models

The structural models were derived by a threading procedure (5) in which a Smith–Waterman algorithm (6) is used to align the target sequence to the template structure according to pair and surface mean force potentials and a BLOSUM40 amino acid substitution matrix (7).

Description of scores

We examined the following scores for model evaluation: The pair energy $E_p$ and surface energy $E_s$ are based on mean force potentials. These energies can be converted into $Z$–scores using a polyprotein (8, 9). Three $Z$–scores are tested: $Z_{pp}$, derived from pair energy, $Z_{ps}$, derived from surface energy, $Z_{pc}$, derived from a combination of $E_p$ and $E_s$ (9).

Another class of energy based scores are obtained by shuffling the sequence of the model and keeping the structure constant (10). We reverse the sequence, calculate the pair and surface energy and compute the difference to $E_p$ and $E_s$ respectively. This way we obtain $\Delta E_{rp}$ for pair energy and $\Delta E_{rs}$ for surface energy. In addition, we obtain a background distribution by repeatedly shuffling the sequence randomly and calculating the respective energies. From the mean and standard deviation the $Z$–score of $E_p$ and $E_s$ is calculated ($Z_{sp}$ for pair energy, $Z_{ss}$ for surface energy).

From the threading alignment of length L, we obtain the alignment score ($A_{SW}$), which results from applying a Smith–Waterman algorithm on a matrix $m(i,j)$. Each value in $m(i,j)$ contains a measure of the fitness of aligning the sequence residue i to the position
j in the template fold. Values of m(i,j) were obtained by combining information from pairs and surface mean force potentials (11) as well as from the BLOSUM40 substitution matrix.

A sequence similarity score is calculated, which is the sum of the amino acid substitution values of the aligned residues (A_{SEQ}). The percentage of identity (I_d) is also calculated.

The spatial compactness is measured by C_{α} – C_{α} contacts within a given distance. For each residue in the model we count the number of C_{α} atoms within a sphere of 9Å (C_9) and 15Å (C_15), respectively. The total sum is divided by the number of residues in the model.

Two scores were implemented that indicate the extent of the match between the target secondary structure prediction from DSC (12), and the template secondary structure assignment from DSSP (13). In order to generate the secondary structure predictions for each target, Gapped Blast (14) was used to search for homologues in the Swissprot database (15). Multiple sequence alignments of the target and resulting Blast hits were generated using Clustalw (16) and used as input for DSC. A matrix for the alignment of the secondary structure elements, helix, strand and coil has been derived from structurally equivalent residues in 3D structure superimpositions (17). The matrix is available at http://www.came.sbg.ac.at/Services/Services.html. Score S_1 is the sum of the secondary structure substitution values of the aligned residues. In addition we calculate a weighted sum using the probability of each predicted secondary state (score S_2).

Data were divided into three groups depending on the alignment length L. Group 1 includes the whole range of L. Group 2 contains alignments from L=80 to L=120, which corresponds to medium range domain size. Group 3 consists of alignments with L larger than 120.

A correlation analysis was performed for all the protein pairs of the testing set. The coverage (% of related pairs detected correctly) of each score was calculated by using a threshold admitting no more than 1% error (1% of the unrelated pairs over the threshold and therefore incorrectly assigned as related pairs).

RESULTS

Correlation analysis

The correlation analysis is performed on groups 1 and 2 in order to determine the degree of relationship between scores. As expected, correlation is found among scores of the same type, e.g. S_1/S_2, or C_9/C_15, or the group of solvation energy based scores (Z_{ps}, Z_{ss}, Z_{pc}, ΔE_{rs}), and the group of pair energy based scores (Z_{pp}, Z_{sp}, Z_{pc}, ΔE_{rp}). In addition, C_15 is found to be correlated to the group of solvation energy based scores, see the scatter plot in Figure 1. Finally, all scores are highly correlated to the alignment length (L). Except for C_15 and Z_{pp}, which increase proportionally to the square root of L, the relationship is a linear one. Using a regression model, this length dependence can be eliminated, but, as we will see below, this improves only the top performing scores (A_{sw}, S_1, and S_2) and lowers the discriminative power of the other scores.

Analysis of the coverage

The performance measurements would have been biased by the strong correlation between L and the single scores. To remove this bias, we normalized all scores by subtracting an appropriate (linear/non-linear) regression model. Results for the following three groups of data and for both, normalized and raw scores, are shown in Table 1.
Group 1 (arbitrary L)
It contains the complete testing set. The normalized versions of $A_{sw}$, $S_1$ and $S_2$ have coverage values ranging between 27 and 32%. The related raw scores show lower coverage. These are followed by energy scores, $E_p$ and $E_s$. Raw Z-scores show significant performance in comparison to the other scores but length correction significantly decreases their coverage.

Group 2 ($80 \leq L \leq 120$)
This subset of data is constituted of 151 related and 6289 unrelated pairs. Again three scores, $A_{sw}$, $S_1$ and $S_2$ give coverage around 20%. There is no considerable difference between normalized and raw scores. This is somewhat expected given the narrow range of alignment length in this group.

Group 3 ($L > 120$)
This group is made of 266 related and 4064 unrelated pairs. Raw Z-scores have lower values comparing to group 1. Normalization considerably increases the performance of $A_{sw}$ from 13 to 32%. The other scores do not seem to be affected.

Since the ratio between related and unrelated pairs is different in the 3 groups, one cannot compare the performance of one score between the different groups, but only the relative performances of the different scores within a group. Table 1 indicates that $A_{sw}$ and $S_2$ perform quite differently when the threshold is optimized for medium and long alignment lengths, with $S_2$ being slightly more reliable with medium length models and $A_{sw}$ clearly being optimal for long models. Therefore it would be desirable to take into account the alignment length in defining thresholds for different scores.

A significant difference is observed between raw and normalized scores. The coverage of raw scores is higher than that of the respective normalized ones, except for $A_{sw}$ in all groups and $S_2$ and $S_1$ in group 1.

$A_{sw}$ and $S_2$ give high coverage on all groups of data for both normalized and raw scores. $S_2$ gives slightly higher coverage than $S_1$ indicating that the weighting of substitution matrix entries according to the probability of the secondary structure states further improves this scoring scheme.

In addition, we performed the same kind of analysis restricting the test set to single domain protein pairs only. An overall increase of the score performance was observed (up to a coverage of 38% for $S_2$ in group 1). Nevertheless, the overall score distribution does not change significantly: $S_2$, $A_{sw}$ are the scores giving the best coverage rates.

CONCLUSION
The conclusion of this work is that a score combining knowledge based potentials with amino acid substitution scores, and a score based on predicted secondary structure are most discriminative. Applying length correction, they cover approximately 30% of the related pairs if an error rate of 1% is allowed. All other scores investigated have a coverage below 23%. Since the two best performing scores seem to be largely uncorrelated as judged by a regression analysis, it is most likely that a proper combination is even more discriminative. Our preliminary studies using a neural network indicate that coverage rates above 50% can be obtained. We are currently optimizing the architecture of the network. Details and final results will be published elsewhere.
Acknowledgements

Support for this work was provided by the Fonds zur Foerderung der wissenschaftlichen Forschung, Austria, Grant number P13710−MOB and Austrian Nationalbank, Jubilaeumsfonds, Austria, Grant number 7899.

REFERENCES

Figure 1. Scatter plot showing the correlation between $C_{15}$ and $Z_{ps}$. Only 10% of the related (boxes) and 1% (filled triangles) of the unrelated pairs are shown in the plot.
Table 1. The table shows the percentage of coverage reached with a threshold allowing at most 1% of error for the whole testing set (column 1), for medium alignment lengths L (column 2) and for large L (column 3). Numbers in brackets give the performance of raw scores, numbers without brackets give that of the alignment length corrected scores.

<table>
<thead>
<tr>
<th>Scores</th>
<th>Arbitrary L</th>
<th>80 ≤L≤120</th>
<th>L&gt;120</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>516 Rp / 32141 Up</td>
<td>151 Rp / 6289 Up</td>
<td>266 Rp / 4064 Up</td>
</tr>
<tr>
<td>A_{sw}</td>
<td>32 (29)</td>
<td>19 (15)</td>
<td>32 (13)</td>
</tr>
<tr>
<td>S_2</td>
<td>31 (23)</td>
<td>21 (21)</td>
<td>16 (15)</td>
</tr>
<tr>
<td>S_1</td>
<td>27 (22)</td>
<td>20 (21)</td>
<td>13 (13)</td>
</tr>
<tr>
<td>E_s</td>
<td>17 (22)</td>
<td>3 (3)</td>
<td>12 (13)</td>
</tr>
<tr>
<td>E_p</td>
<td>17 (20)</td>
<td>8 (8)</td>
<td>8 (6)</td>
</tr>
<tr>
<td>ΔE_{rs}</td>
<td>14 (18)</td>
<td>3 (3)</td>
<td>8 (9)</td>
</tr>
<tr>
<td>ΔE_{rp}</td>
<td>10 (18)</td>
<td>3 (4)</td>
<td>5 (5)</td>
</tr>
<tr>
<td>A_{seq}</td>
<td>11 (13)</td>
<td>8 (5)</td>
<td>8 (8)</td>
</tr>
<tr>
<td>Z_{ss}</td>
<td>5 (17)</td>
<td>3 (5)</td>
<td>3 (12)</td>
</tr>
<tr>
<td>Z_{sp}</td>
<td>3 (16)</td>
<td>2 (4)</td>
<td>4 (7)</td>
</tr>
<tr>
<td>Z_{pp}</td>
<td>4 (19)</td>
<td>3 (3)</td>
<td>5 (14)</td>
</tr>
<tr>
<td>Z_{ps}</td>
<td>8 (22)</td>
<td>8 (8)</td>
<td>7 (12)</td>
</tr>
<tr>
<td>Z_{pc}</td>
<td>7 (23)</td>
<td>6 (6)</td>
<td>6 (15)</td>
</tr>
<tr>
<td>C_{si}</td>
<td>1 (14)</td>
<td>1 (2)</td>
<td>1 (5)</td>
</tr>
<tr>
<td>C_q</td>
<td>1 (4)</td>
<td>0 (0)</td>
<td>1 (2)</td>
</tr>
<tr>
<td>L_1</td>
<td>0 (0)</td>
<td>6 (6)</td>
<td>6 (5)</td>
</tr>
</tbody>
</table>

Rp: number of related pairs. Up: Number of unrelated pairs. Abbreviations are described in the Methods section.