KP230

CIRCLE: Development of 3D Protein Model Quality Assessment program using secondary structure prediction method and side-chain environment

Genki Terashi[1], Hiroko Sakai[1] , Kazuhiko Kanou[1], Tomoko Hirata[1], Mayuko Takeda-Shitaka[1], Hideaki Umeyama[1]

1) School of Pharmacy, Kitasato University, 5-9-1 Shirokane, Minato-ku, Tokyo 108-8641 JAPAN
terashig@pharm.kitasato-u.ac.jp

1. Introduction

The accurate prediction of protein structure is one of the major challenges in the field of bioinformatics. The Model Quality (MQ) assessment technique for distinguishing the near native models (high quality models) from decoys which are inferior models is one of the most important factors to achieve the accurate protein structure prediction. Many of the scoring functions for evaluating protein structures are founded on knowledge-based potentials, clustering methods, structural energies using molecular mechanics force fields, and the profile of sequence or structure (e.g. Verify3D, Inbgu, 3D-PSSM, ProQ). These scoring functions are used to assess the model quality and ultimately select the best model among a set of models. In this work, we have developed MQ Assessment Programs CIRCLE [1] and participated in Quality Assessment (QA) category of CASP8 [2] (The 8th Critical Assessment of Protein Structure Prediction, May-Aug 2008). CIRCLE aims at identifying the near native models and incorrect models without using consensus methods.

2. Method

CIRCLE considers two terms for the model quality: (1) model quality calculated from the side-chain environment of each residue (SideChainScore in equation(1)); and (2) similarity between the secondary structure propensities predicted for an amino acid sequence by PSI-PRED and the secondary structure of the three-dimensional model (SSscore in equation (1)). The side-chain environment for each residue is determined from the fraction of the molecular surface area of the side-chain covered by the polar atoms, the fraction of the side-chain area buried by any other atoms, and the secondary structure. According to the target difficulty, a total score is calculated as:

$$TotalScore = \begin{cases} \sum_{n}^{length} \left(0.35 \times SSscore + SideChainScore_{CM}\right)_n & CM \\ \sum_{n}^{length} \left(0.75 \times SSscore + SideChainScore_{FRNF}\right)_n & FRorNF \end{cases} \quad (1)$$

As shown in equation (1), the similarity score of the secondary structures (SSscore) is emphasized in difficult targets (FR: Fold Recognition, NF: New Fold) than easy targets (CM: Comparative Modeling).

In the QA category of CASP8, predictor groups provide quality estimates comprising scores between 0.0 and 1.0 for each protein structure model produced by server groups participating CASP8. Therefore, for each target, we convert estimated score of models into the values from 0.0 to 1.0 by scaling circle score of models which has minimum and maximum values.

3. Results and Discussion

The 103/128 (80%) native protein structures of CASP8 targets were published in CASP8 web site (Sep 2008). We calculated Pearson's correlation coefficient between converted CIRCLE score and the quality of models. We used the Global Distance Test Total Score (GDT_TS) as the quality of model compared to native.

The average of GDT_TS (x-axis) and correlation coefficient (y-axis) are shown in Fig.1. These results show that QA performance of CIRCLE depends on the quality of set of models which are evaluated (Table 1). The good correlation coefficients were obtained above 0.9 for the targets having the high average value of GDT_TS (above 50).

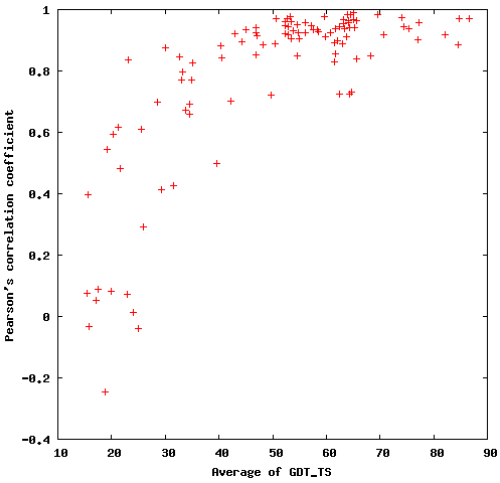

Fig. 1

Average of GDT_TS	Average of Pearson's correlation coefficient
0–25	0.24
25–50	0.75
50–75	0.92
75–100	0.93
ALL	0.78

Table 1

Additionally the best (T0423) and worst (T0460) examples of CIRCLE results are shown in Fig.2 and Fig.3. The x-axis and y-axis represents the circle score and GDT-TS of each model, respectively. In T0423 (Fig.2), CIRCLE score has high value of correlation coefficient (0.98), because high quality models (GDT_TS > 50) has high proportion of set of models. In contrast, in the case that no good models existed in the set of models (T0460 of Fig.3), CIECLE could not perform well (correlation coefficient = -0.24). These results indicate that CIRCLE still has a room to improve especially in difficult targets. We are planning to add other kind of scoring function calculated from evolutional information such as a sequence alignment score and consensus method.

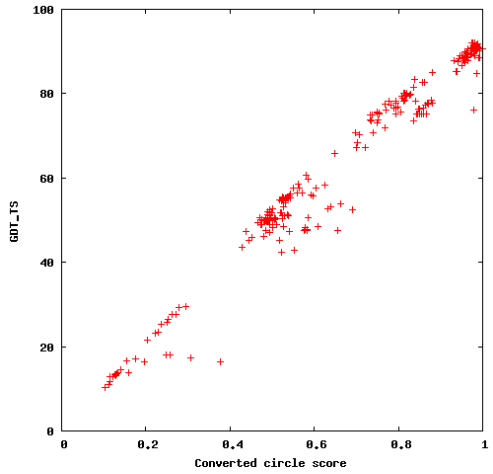

Fig. 2 T0423

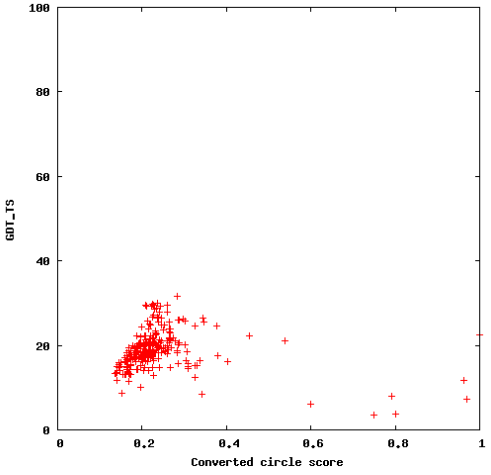

Fig. 3 T0460

References

1. http://predictioncenter.org/casp8/index.cgi
2. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H. Fams-ace: a combined method to select the best model after remodeling all server models. Proteins. 2007;69 Suppl 8:98-107.