Poster I-67

3D Model quality evaluation using evolutionary information

Authors:
Nicolas Palopoli (*Centro de Estudios e Investigaciones, Universidad Nacional de Quilmes*)
Ezequiel Juritz (*Centro de Estudios e Investigaciones, Universidad Nacional de Quilmes*)
Diego Gomez-Casati (*IIB-INTECH, UNSAM-CONICET*)
Gustavo Parisi (*Centro de Estudios e Investigaciones, Universidad Nacional de Quilmes*)

Short Abstract: We use the structurally constrained model of protein evolution (SCPE) to assess 3D model quality. Our method uses site-specific substitution matrices generated with SCPE simulations for each model. We found the best structural model using maximum likelihood calculations using a reference alignment and the derived substitution matrices.

Long Abstract:

Knowledge of the three-dimensional structure of a protein can often be very useful for understanding its biological activity and function. Different approaches are being used to build structural models, which include homology modeling, threading and ab initio methods. The quality assessment of the models obtained using these methods is a difficult task, and the variety of methods to achieve it reflects the different aspects that can be evaluated in a prediction. In general, assessment methods search for structural similarity between the predicted model and the experimental structure in different ways. For example they could use the root mean square deviation (RMSD) or more complex measures as MaxSub[9] and GDT[10]. Other methods based on structural clustering or on energetic evaluations have also been proposed[3;4]. In this work we propose a new way to assess the quality of a model based on the use of the Structurally Constrained Protein Evolution model (SCPE [6]). The SCPE model simulates protein evolution by introducing random mutations into evolving sequences and selecting them against too much structural perturbation. Given a single protein structure, the SCPE model can be used to obtain a whole set of site-dependent amino acid substitution matrices[1]. As it is well established, protein structure constrains sequence divergence during evolution producing a fold-specific sequence pattern that could be detected in an alignment of homologous sequences. Then, using SCPE derived matrices for each generated model, the best structural model is the one producing the best fitness between this sequence pattern and the set of substitution matrices generated using the SCPE.

Structural models were generated for the D2 domain of starch-synthase III from Arabidopsis thaliana which we had previously identified as a member of the starch-binding domain (SBD) family [5]. The models were obtained using the program Nest and 11 SBD proteins as templates. For each structural model we run the SCPE as it was described previously[6] and a set of site-specific substitution matrices were obtained and evaluated using maximum likelihood calculations performed with HYPHY[8]. The reference alignment needed in the evaluation was obtained from CAMPASS database and consists of 11 structurally aligned SBD along with their close homologous proteins (89 sequences).

With these data we design three tests to evaluate a given model: a) the adequacy of SCPE for the different structural models, b) the specificity of the SCPE site-specific substitution matrices, c)the capacity of SCPE to assign a fold class. The SCPE model was compared with an unconstrained model of evolution (JTT model [2]). Comparison was performed using a likelihood ratio test. A shuffling protocol was performed for the site-specific substitution matrices for each model and 500 models were obtained of D2 using the program nest using randomly chosen templates. For each of these models we run SCPE and maximum likelihood calculations as explained above.

The combination of the tests mentioned allows the selection of the model that best reproduces the sequence pattern found in the SBD fold reference alignment. In our view, this model is the best one among the 11 built. It is important to note that it was not possible to select a given model using CAFASP MQAPs that mainly use both energetic and structural criteria for selection. We expect that this model better reflex the structural properties of the D2 domain to design future mutagenesis experiments to evaluate biochemical and functional characteristics.
We think that the use of evolutionary information is a promising method to assess the quality of structural models. Also, as substitution matrices contain more information than other sequence-based methods (for example profile-profile analysis), this methodology could also be a sensitive tool for fold assignment.

References

[1] Fornasari, M.S., Parisi, G., and Echave, J. 2002. Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. Mol Biol Evol. 19:352-6
[2] Jones, D.T., Taylor, W.R., and Thornton, J.M. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275-82
[3] Lackner, P., Koppensteiner, W.A., Domingues, F.S., and Sippl, M.J. 1999. Automated large scale evaluation of protein structure predictions. Proteins. Suppl 3:7-14
[4] Lee, M.R., Tsai, J., Baker, D., and Kollman, P.A. 2001. Molecular dynamics in the endgame of protein structure prediction. J Mol Biol. 313:417-30
[5] Palopoli, N., Busi, MV., Fornasari, MS., Gomez-Casati, D., Ugalde, R. and Parisi, G. 2006. Starch-synthase III family encodes a tamden of three starch-binding domains. In Press in Proteins, Structure, Function and Bioinformatics.
[6] Parisi, G. and Echave, J. 2005. Generality of the structurally constrained protein evolution model: assessment on representatives of the four main fold classes. Gene. 345:45-53
[7] Parisi, G. and Echave, J. 2001. Structural constraints and emergence of sequence patterns in protein evolution. Mol Biol Evol. 18:750-6
[8] Pond, S.L., Frost, S.D., and Muse, S.V. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 21:676-9
[9] Siew, N., Elofsson, A., Rychlewski, L., and Fischer, D. 2000. MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics. 16:776-85
[10] Zemla, A., Venclovas, C., Moult, J., and Fidelis, K. 1999. Processing and analysis of CASP3 protein structure predictions. Proteins. Suppl 3:22-9