ToolShop: prerelease inspections for protein structure prediction servers

Leszek Rychlewski

Bioinformatics Laboratory, BioInfoBank Institute, ul. Limanowskiego 24A,
60-744 Poznan, Poland

Received on May 8, 2001; revised on June 20, 2001; accepted on July 2, 2001

ABSTRACT

Summary: The ToolShop server offers a possibility to compare a protein tertiary structure prediction server with other popular servers before releasing it to the public. The comparison is conducted on a set of 203 proteins and the collected models are compared with over 20 other programs using various assessment procedures. The evaluation lasts ca one week.

Availability: The ToolShop server is available at http://BioInfo.PL/ToolShop/. The administrator should be contacted to couple the tested server to the evaluation suite.

Contact: leszek@bioinfo.pl

Supplementary information: The evaluation procedures are similar to those implemented in the continuous online server evaluation program, LiveBench. Additional information is available from its homepage (http://BioInfo.PL/LiveBench/).

Protein structure prediction servers potentially provide a very useful service to the biological community (reviewed by Moult, 1999). The utilization of such services without mandatory caution can result in false conclusions and potentially wasted experimental resources used to validate the predicted structural or functional features. A small number of misleading predictions or misunderstanding of the outcome of the algorithms can quickly result in loss of consumer confidence, with potentially long lasting reluctance to use the service again. Clever marketing strategies and carefully planned market introduction plans are designed and followed when a company launches a major new product. This is obviously seldom the case when introducing a new protein fold recognition program by an academic research team. Most of the new methods are accompanied by results obtained on in-house benchmarks, which offer little possibility of comparison with other popular services offered to the community. This can be the source of an important problem. The community of users is not able to assess a priori the real value of the new algorithm and a new high-quality service has a much tougher struggle to gain justified recognition in the field. The two main options, which could be followed by the developers, are to sign up the new server to a continuous structure prediction evaluation program, such as LiveBench (Bujnicki et al., 2001a) (http://BioInfo.PL/LiveBench/) or Eva (http://maple.bioc.columbia.edu/eva/) or to wait for the next CAFASP experiment (Fischer et al., 1999), which is traditionally conducted every two years. The first option offers the advantage of a much larger sample size. The second option has the important advantage of being a blind test. Both have however the disadvantage of the relative long latency between the start of the evaluation and the availability of the official results.

The ToolShop program has been launched to speed up this process, to assist in the introduction of new services and to offer a final check before the new protein structure prediction site is released. The service is based on the experience from previous community wide evaluation experiments, including CAFASP (Fischer et al., 1999) and LiveBench (Bujnicki et al., 2001a). The same evaluation methods are used and the current set of query proteins is identical to the set selected during LiveBench-2 that includes 203 proteins (J.M.Bujnicki, A.Elofsson, D.Fischer and L.Rychlewski, manuscript submitted). Models for all query proteins obtained from over 20 popular structure prediction methods are used for the comparison of performance. The evaluation includes the following characteristics (checklist).

1 Sensitivity

The sensitivity analysis is conducted the same way as in the LiveBench program (Bujnicki et al., 2001a). Various model quality assessment methods are used to verify the correctness of the results of the servers. The list of methods includes standard rigid body superposition approaches (MaxSub; Siew et al., 2000, and LGscore, described in LiveBench-1; Bujnicki et al., 2001a) as well as contact map overlap evaluation (Touch, described in LiveBench-1; Bujnicki et al., 2001a). Each method produces its own ranking and four such rankings are summarized to produce a total ranking. The main difference between blind prediction and benchmarks is that the structures of the tar-
gets are known before. Most of the new prediction services will have the ToolShop targets in their template databases. Thus the ToolShop server removes all hits reported by the servers, which point to PDB proteins, which are ‘younger’ than the PDB files of the target. The dates of files available from the RCSB ftp server are used for this filtering procedure. This has two contrary implications for the evaluation of performance. If the prediction server uses only representative templates and a group of correct (similar) proteins is represented by a too young protein, then the whole group is effectively removed. This reduces the size of the effective databases. The size of the template databases is known to influence the observed ranking. On the other hand many servers use sequence databases to produce profiles for the targets or templates. The size of the sequence databases grows rapidly. New sequences should result in more specific profiles and should positively influence the prediction accuracy. Another factor is the rising accuracy of components of threading methods, i.e. secondary structure prediction programs. If such components are used, the prediction could be contaminated by the knowledge of the correct secondary structure. The net outcome of both opposing effects is usually positive. (All servers, which were evaluated twice at two different time points obtained better results in the second, later evaluation.) Thus, very good sensitivity results obtained during the ToolShop test must be confirmed by other evaluation procedures like the then current LiveBench, EVA or the next CAFASP.

2 Specificity

The specificity of the prediction is evaluated by counting the number of correct hits with a more significant score than the first 1–10 false predictions. The goal of this analysis is to assess how reliably the reported confidence score can be used to estimate the accuracy of the prediction. The number of correct hits with higher confidence score depends also on the sensitivity of the prediction. Nevertheless the analysis offers a reasonable comparison of the methods. The fact that some servers of lower sensitivity exhibit high specificity and vice versa (Bujnicki et al., 2001a) confirms the utility of this evaluation strategy. Higher sensitivity of the service improves the specificity because more correct templates can be found above a given score cutoff. Because the sensitivity is affected by the time the evaluation is conducted (see Section 1), the specificity will be affected slightly as well. However, according to our experience, this effect is marginal.

3 Added value

If the sensitivity or the specificity of the new algorithm is not ranked first, it is of course not a reason to postpone or drop the release of the new service. In most of the cases the new prediction method provides a valuable service creating an added value to the ‘community’ of services. This can be demonstrated by the number of times the new service returned the best model (model with strongest similarity to the correct structure). The majority of the coupled services have such predictions (Bujnicki et al., 2001a). The most important added value comes however from correct predictions where other services have failed. This analysis is conducted on a pair-wise basis separately for the easy and hard targets. The results clearly demonstrate that all common fold recognition servers provide complementary predictions to other services. On average 10% more correct hits could be produced if two services of similar sensitivity would be combined perfectly.

The main advantage of the ToolShop service is the speed of the analysis. The evaluation can be completed in few days. An additional advantage of the ToolShop service is its close relation to the LiveBench program (Bujnicki et al., 2001a) and the Structure Prediction Meta Server (Bujnicki et al., 2001b). If, after the analysis of the evaluation results the developer decides that the service is ready to go public, the more official LiveBench evaluation could begin instantly. The advantage of being part of the Meta Server on the other hand is the immediate access to the ‘protein structure prediction market’. The Meta Server also provides many useful format translation features and links to other services. This could reduce the effort needed to create convenient interfaces to new valuable prediction methods and help the developers of algorithms to focus on their main research subject.

ACKNOWLEDGEMENTS

I would like to thank Janusz M. Bujnicki, Arne Elofsson and Daniel Fischer for very useful discussions and for the model assessment software (Arne and Daniel).

REFERENCES


