Empirical limits for template-based protein structure prediction: the CASP5 example

B. Contreras-Moreira*,[1], I. Ezkurdia, M.L. Tress, A. Valencia*

Protein Design Group, Centro Nacional de Biotecnología, Madrid, Spain

Received 29 November 2004; revised 17 December 2004; accepted 5 January 2005

Available online 19 January 2005

Edited by Robert B. Russell

Abstract Most protein structure prediction methods use templates to assist in the construction of protein models. In this paper, we analyse the current state of template-based modelling approaches and reach an estimate of the empirical limits of these methods. Our analysis show that current prediction methods are already reaching these empirical accuracy limits in the easier cases, where finding a close homologue to the native target structure is not a problem. However, we find that even in the absence of alignment errors and using optimal templates, template-based methods have intrinsic limitations, suggesting that other methodologies, such as ab initio procedures, must be used if accuracy is ultimately to be improved.

Keywords: Template-based protein structure prediction; Comparative modelling; Fold-recognition; Fragment reconstruction; Accuracy limit

1. Introduction

Methods for protein structure prediction can be classified into two basic classes: those which use physical principles to fold a protein and those which use experimentally determined structures to help reconstruct the protein of interest. The first class is usually known as ab initio approaches [1]; the second includes related techniques such as comparative modelling, fold recognition and threading [2–7]. These generally use sequence alignments to map the sequence to be modelled onto protein templates of known structure and are guided by criteria such as sequence similarity or secondary-structure compatibility.

This paper deals mainly with the second class of methods, template-based methods. The empirical basis for these approaches comes from the observation by Chothia and Lesk [8] that protein sequence identity and structural similarity are correlated. According to their original results there are clear empirical limits for protein structure predictions based on single templates: for proteins sequences around 95% identical backbone deviations are expected to be under 1 Å RMS; when the sequence identity drops to 30%, deviations grow to around 4 Å RMS. These limits broadly agree with the observed performance of comparative modelling servers as measured by continuous benchmarks such as EVA [9] (see [10] for a review), and ultimately affect the quality and therefore the applicability of template-based predictions [11].

In addition to these natural restrictions, methods for template-based prediction of protein structure must solve two technical problems: the choice of the template closer to the target structure, and the derivation of the sequence alignment between the query and template protein closer to the optimal structural alignment. The lack of satisfactory solutions for these two problems has been identified as negatively affecting the performance of fold recognition and comparative modelling methods in previous "Critical Assessment of Techniques for Protein Structure Prediction" experiments (CASP [12]) [13,14].

However, choosing the correct template and alignment are not the only problems facing predictors. Even those models built from the correct template and alignment often require substantial refinement in order to be sufficiently close to the native target structure. This paper seeks to estimate the limits of current template-based structure prediction techniques under ideal conditions, that is building a model a posteriori using multiple optimal templates and in the absence of alignment errors.

We do that by allowing models to be built by combining aligned fragments from several templates, selected by structural similarity. We then measure, using the CASP GDT_TS score [15], how the best fragment-based predictions compare to the native target structure.

Additionally, we ask how far the predictions are from these best possible models. This gives us a better idea of how successful the current modelling methods are, how good they could be in the absence of the sequence alignment problem, and can implicitly tell us to what extent ab initio methods would be needed to improve the current performance of template-based methods.

2. Datasets, methods and algorithms

A collection of 68 targets, as split in domains by the CASP5 organisers, was taken as our test set (see http://predictioncenter.llnl.gov/casp5). These targets are proteins whose experimental structures were about to be released at the time CASP5 started (May, 2002). To model

*Corresponding authors. Fax: +34 91 585 45 06.
E-mail addresses: contrera@ccg.unam.mx (B. Contreras-Moreira), valencia@cnb.unam.es (A. Valencia).

[1] Present address: Programa de Genómica Computacional, Centro de Ciencias Genómicas, Av. Universidad s/n, Colonia Chamilpa, 62210 Cuernavaca, Morelos, México.

0014-5793/$30.00
doi:10.1016/j.febslet.2005.01.014

these targets we needed a library of PDB templates and for that we used a 90% non-redundant set of PDB chains from April 2002, obtained from PDB-SELECT [16]. Since CASP5 started in May 2002 we were sure that we did not have access to templates not available to the predictors at that time. This library included 6182 PDB chains.

Then, we designed our procedure with five steps to be applied to every CASP5 target:

(1) Search the template library for a list of significantly similar PDB structures.
(2) Superimpose all found templates in the target's frame of reference.
(3) Calculate a large number of fragment-based models from the ensemble of templates and evaluate them using typical CASP evaluation parameters.
(4) Compare the best fragment-based model to the target structures.
(5) Compare the best fragment-based model to the best model produced by CASP5 predictors.

Note that no fragment readjustment is performed, since we felt this was an ab initio technique.

Each of these steps was implemented as follows:

1. To search the template library we used the program MAMMOTH [17] and took only those templates that yielded a $-\ln E$ score over 4.5, to avoid using marginally similar structures. Only the top 40 hits were considered.
2. To superimpose the selected templates in the frame of reference of the target we used two programs, MAMMOTH and LGA [15], to generate alternative sequence-independent superpositions. Other superimposition protocols could be added in this step, but for demonstration purposes we felt that two were enough. In the case of MAMMOTH, the coordinates of the superimposed template needed to be transformed using the rotation matrices and translations provided in the output.
3. This was the most important step in our protocol, the generation of a collection of models for our target by fragment reconstruction from the superimposed templates. This step included several substeps, as shown in Fig. 1 and was based on a previous work [18]:

3.1. Construct a multiple alignment from the pairwise structural alignments between target and templates, with the target as the frame of reference. This multiple alignment can be regarded as a matrix with each row in the matrix corresponding to a template, each column to an aligned set of residues and their backbone coordinates.
3.2. In this sub-step, we define "fragment" as a contiguous set of template residues that have been aligned without gaps by either MAMMOTH or LGA. As suggested by related work [19,20], fragment length is an important parameter and here we tried values of 5 and 9 residues. To score fragments we used a score similar to GDT-TS [15,21], the main evaluator used in CASP experiments. GDT-TS score measures similarity between two structures based on a combination of the fractions of matching residues within distance cutoffs of 1, 2, 4, and 8 Å. MyGDT is similar to GDT-TS but calculated on just one superimposition. It is calculated as P1 + P2 + P4 + P8/4, where P.n. is the % of residues in the template closer than n Å to the corresponding residues in the target. In this sub-step, fragments of the chosen length in the matrix are labelled and their myGDT scores are calculated and stored.
3.3. For each labelled fragment a new fragment-based model by growing it towards both N and C termini within the matrix applying iteratively these greedy rules:
  3.3.1. If one or more fragments are available in the matrix to grow the model, choose the best one and add it. Fragments are scored according to their local myGDT score with respect to the target's coordinates.
  3.3.2. Otherwise, if possible, extend the solution model by one residue.
3.4. Rank all obtained fragment-based solutions in terms of global myGDT scores with respect to the target length and select the best ones within a given tolerance (set to 1 myGDT unit in this experiment).

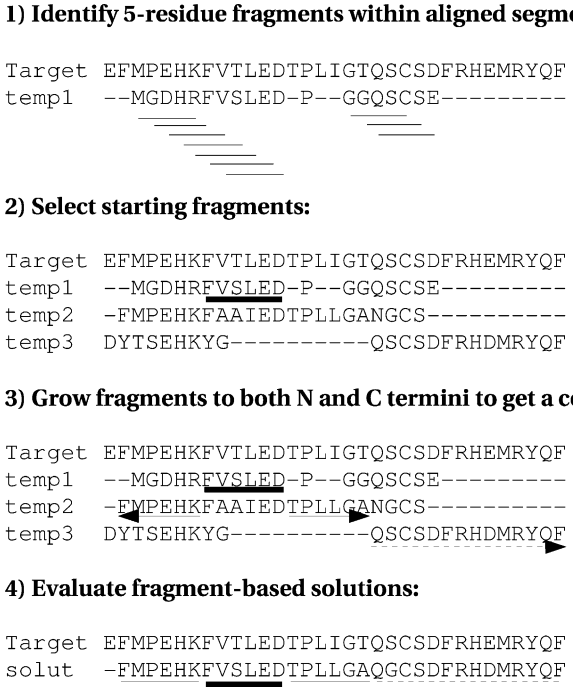

Fig. 1. Graphical example of the construction of fragment-based models from a set of superimposed templates. A possible fragment-based solution is built starting from a fragment of 5 residues in template 1. This fragment is then grown in both left and right (N and C-terminal) directions. Going left there are three options, three possible fragments and the one from template 2 was taken for having the best local myGDT score with respect to the target's coordinates. Towards the right of the starting fragment, initially only the fragment from template 2 was available. Then another fragment was extracted from template 3 and this was actually extended since no other fragments were available (see rules 3.3.1 and 3.3.2).