CASP7 Predictions using the Wurst Server

Nasir Mahmood, Tina Stehr, Steve Hoffmann, Thomas Margraf,
Martin Mosisch, Gundolf Schenk, Paul Reuter, Thomas Huber[*], Andrew Torda
[*]Departments of Mathematics and Biology, University of Queensland, Australia
Center for Bioinformatics, Hamburg University, Germany
mahmood@zbh.uni-hamburg.de

Introduction

Wurst is a sequence to structure threading code, but is hopefully distinguished by:

- no assumption of Boltzmann statistics
- a sequence-structure term based on simultaneous sequence+structure classification
- sequence terms based on an optimised substitution matrix
- all other parameters from numerical optimisation

The server with all parameters as used for CASP is at http://www.zbh.uni-hamburg.de/wurst

Philosophy

We wanted to improve our ranking of guesses (templates), bring world peace and improve our sequence alignments. We may have partially succeeded in the last aim.

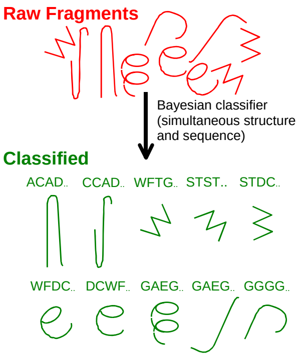
Methods

Alignments were generated using conventional Smith and Waterman approach (Gotoh algorithm), but the methodology has two interesting aspects:

1. The parameters (gap penalties, coefficients for contributions) came from numerical optimisation and
2. The main score function components were based on a maximally parsimonious Bayesian classification across both discrete (sequence) and continuous (structure) properties.

Score Function - Bayesian Classification of Fragments

The score function contained a sequence profile-profile term, but also the more interesting classification based term. A set of 5 to1.5×10^6 fragments (length 6 to 9) was viewed as a set of descriptors. The continuous properties (backbone angles) were modelled by Gaussian functions - the discrete (sequence) properties by multimodal Bernoulli distributions.



Raw Fragments

Bayesian classifier (simultaneous structure and sequence)

Classified

ACAD.. CCAD.. WFTG.. STST.. STDC..

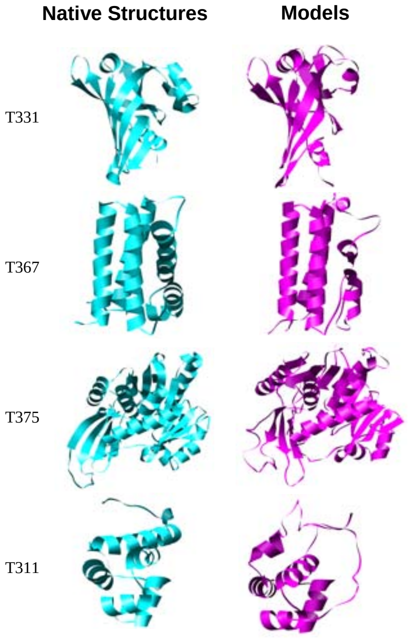
WFDC.. DCWF.. GAEG.. GAEG.. GGGG..

A classification was built using expectation maximisation. One uses an objective function which tries to optimise the probability that the statistical model agrees with the training data. It tends to minimise the number of classes - excess classes incur a penalty due to marginal probabilities.

Optimisation of Parameters

A set of about 2000 protein pairs was used with low sequence similarity, but some structural similarity. Within each pair, the sequence of "A" was aligned to the structure of "B" and the model quality assessed by comparison with the structure of "A" (a measure based on correct contacts).

CASP Predictions



Targets T331, T367, T375 and T311 have PSI-BLAST level homology

This was summed over all 2000 pairs and with alignments in both directions. This was used as the basis for a cost function, so parameters could be determined with a simplex optimiser. This had two benefits:

1. It really is possible to say that gap penalties or the coefficients weighting the terms were optimal in the context of our score functions.
2. The cost function provides a convenient way to test ideas in terms of their ability to produce good alignments.

Results

•**Good:**

We would make the claim that the alignment machinery in the server produces excellent sequence to structure alignments. All of the optimisation machinery has been geared to this goal.

•**Bad:**

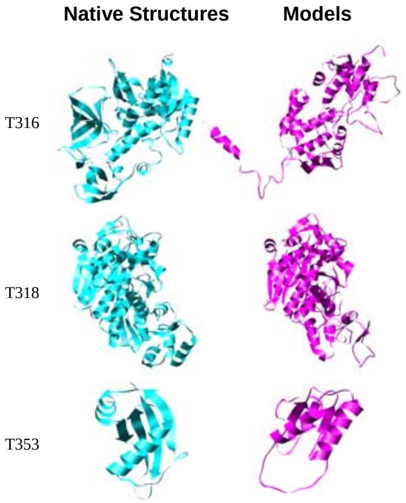
Our ranking of models was atrocious. Models based on completely wrong templates scored as well as the rather good alignments on better templates. For reasons of vanity, we have little interest in using the model assessment programs employed by other servers.

Results - CASP7

The wurst server was notable for its ability to make good models and rank them terribly.

The targets on the left all had only very remote sequence homology. Wurst generally did not use the best template, but produces excellent alignments. For this for structures, the rmsd is less than 2.6 Å with around 90% coverage.
Among the more difficult targets, T316 is a delight. The model for the second domain was amongst the very best predictions from any of the servers.

CASP Predictions



Targets T316, T318 and T353 had no reliable sequence homologues.

Targets	Seq. ID (%)	Template	rmsd (Å)	Fraction of target well predicted (%)
T331	16	2fhq_A	2.16	62
T367	19	1ufb_A	2.43	90
T375	24	1v19_A	2.59	87
T311	15	1adr_	1.65	62
T316	20	1k92_A	3.02	29
T318	26	1gyt_A	2.02	78
T353	19	1fe0_A	3.37	58
T327	17	2fbiA	4.99	76

Sequence identity of template to target and rmsd of model to correct structure and fraction of structure used to calculate rmsd.

Future

The weaknesses we are working on fixing are:

- a better solvation term
- ranking models / template selection

The classification methods used for this server have been used on pure structure (no sequence) problems and form the basis of a new server for rapid and surprisingly accurate structure alignment.