

# APDB: a novel measure for benchmarking sequence alignment methods without reference alignments

Orla J. O’Sullivan<sup>1</sup>, Mark Zehnder<sup>3</sup> Des Higgins<sup>1</sup>, Philipp Bucher<sup>3</sup>, Aurelien Grosdider<sup>3</sup> and Cédric Notredame<sup>2,3</sup>.

<sup>1</sup>Conway Institute, University College Dublin Ireland

[ojos@student.ucc.ie](mailto:ojos@student.ucc.ie),

<sup>2</sup>Information Genetique et Structurale, CNRS, 31, Chemin Joseph Aiguier, 13 402 Marseille Cedex 20, France.

<sup>3</sup>Swiss Institute of Bioinformatics, Lausanne University, 155, Chemin des Boveresses, CH 1066, Epalinges.  
cnotred@igs.cnrs-mrs.fr

**Keywords.** Multiple Sequence alignments; protein structure; alignment evaluation.

## Introduction

APDB is a novel measure for evaluating the quality of a protein sequence alignment, given two or more PDB structures. We show how it is possible to avoid the use of reference alignments when PDB structures are available for at least two homologous sequences in a test alignment. Using this method it should become possible to systematically benchmark or train multiple sequence alignment methods using all known structures, in a completely automatic manner.

Benchmarking is usually accomplished by comparing test alignments to a set of reference alignments of the same sequences assembled by specialists with the help of structural information. Two such set of reference alignments, HOMSTRAD [1] and BALiBASE [2], were investigated in this study. One of the simplest ways of using reference alignments in benchmarking is to count the percentage of columns in the test alignment that are correctly aligned according to the reference alignment (column score) [3]. Although simple and convenient this method of benchmarking has one major drawback; it relies heavily on the reference alignment being correct. In APDB a test alignment is not evaluated against the reference alignment. Instead we measure the quality of structural superposition induced by the test alignment given any structures available for the sequences it contains. Using existing collections of reference multiple sequence alignments and existing sequence alignment methods, we show that APDB gives results that are consistent with those obtained using conventional evaluation methods (see Table 1).

## Table1: Correlation between APDB and CS on BALiBASE and HOMSTRAD

Test Set: indicates the test set being considered, either one of the BaliBase\_91 references or HOM\_43, a subset of HOMSTRAD. N indicates the number of test alignments in this category. ClustalW indicates a set of measures made on alignments generated with ClustalW. T-Coffee indicates similar measures made on T-Coffee generated alignments. Reference indicates measures made on the reference alignments as provided in BALiBASE or in HOMSTRAD. CS columns are the Column Score measures while APDB indicates similar measures made using APDB.

Test Set	N	ClustalW		T-Coffee		Reference	
		CS	APDB	CS	APDB	CS	APDB
Bali_91 Ref1	35	70.1	<b>59.9</b>	67.7	<b>58.3</b>	100	<b>64.7</b>
Bali_91 Ref2	23	32.7	<b>26.6</b>	33.9	<b>47.1</b>	100	<b>55.2</b>
Bali_91 Ref3	22	46.4	<b>38.5</b>	48.6	<b>46.9</b>	100	<b>53.2</b>
Bali_91 Ref4	11	52.0	<b>59.5</b>	52.5	<b>64.5</b>	100	<b>65.7</b>
HOM_43	43	35.4	<b>60.2</b>	38.9	<b>61.6</b>	100	<b>72.9</b>

## **Acknowledgements**

This work is funded by Enterprise Ireland.

## **References**

- [1] Mizuguchi K, Deane CM, Blundell TM, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci*; 7: 2469-71,1998
- [2] Thompson, J., Plewniak, F. & Poch, O. (1999). BaliBase: a benchmark alignment database for the evaluation of multiple sequence alignment programs. *Bioinformatics* **15(1)**
- [3] Karplus, K. & Hu, B. (2001). Evaluation of protein multiple alignments by SAM-T99 using the BALiBASE multiple alignment test set. *Bioinformatics* **17(8)**, 713-20.