Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Biology and Chemistry
ELSEVIER
Computational Biology and Chemistry xxx (2006) xxx–xxx
www.elsevier.com/locate/compbiolchem

Technical comment

A comment on "Prediction of protein structural classes by a new measure of information discrepancy"

Kanaka Durga Kedarisetti, Lukasz Kurgan *, Scott Dick

Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alta., Canada T6G 2V4

Received 6 March 2006; received in revised form 15 June 2006; accepted 17 June 2006

Abstract

Protein structural class describes the overall folding type of a protein or its domain. A number of methods were developed to predict protein structural class based on its primary sequence. The homology of the predicted sequences with respect to the training sequences is a key attribute for the prediction performance. In this article we investigated the FDOD method developed by Jin et al. [Jin, L., Fang, W., Tang, H., 2003. Prediction of protein structural classes by a new measure of information discrepancy. Comput. Biol. Chem. 27, 373–380], which gave high prediction accuracy on a low homology dataset, and we empirically confirmed that the reported results were an artifact of improper implementation.
© 2006 Elsevier Ltd. All rights reserved.

Keywords: Protein structural class; FDOD measure; Sequence homology; Protein primary sequence

Dataset homology is a key characteristic in predicting protein structure classes; the accuracy of a classifier is usually poor on low-homology datasets. The one apparent exception is the FDOD measure developed by Jin et al. (2003), with a jackknife accuracy of 75.02% on a low-homology dataset. However, our attempts to duplicate this research have failed and we have determined that the high prediction accuracy is an artifact of improper implementation. We present analytical and empirical evidence to support this contention.

1. Investigation and experimental analysis of the FDOD measure

After careful analysis of the author's code, we found two subtle methodological errors:

Error 1—Inconsistent application of FDOD measure. While calculating FDOD measure, the probability of the query sequence p^l_{ik} is added to the pre-calculated average distribution. Hence, the Eq. (7) in Jin et al. (2003) was modified to

$$B_k(U^l_{S_1}, U^l_{S_2}, \ldots, U^l_{S_k}) = \sum_{i=1}^{m(l)} P^l_{ik} \log \frac{P^l_{ik}}{\sum_{j=1}^{s} P^l_{ij} + P^l_{ik}/(s+1)} \quad (1)$$

For the jackknife test, Eq. (1) was used for all structural classes except the class to which the query sequence belongs. In the latter case, the original Eq. (7) in Jin et al. (2003) was used. Since the FDOD algorithm predicts the structural class with the minimum B_k (Eq. (9) in Jin et al., 2003), using Eq. (7) in Jin et al. (2003) for the predicted class gives an unfair advantage (lower value) over other classes for which Eq. (1) is used.

Error 2—Pre-calculating the average distribution. The average distribution that is used in the denominator of the Eq. (7) in Jin et al. (2003) is pre-calculated using the entire dataset. Although this is correct for the resubstitution test, the query sequence should be withheld from this pre-computation for the jackknife test. Both errors undermine the jackknife test by making it a form of in-sample, rather than out-of-sample prediction, and causing an increase in the accuracy.

We have empirically confirmed the existence and effect of the above errors. We compare the author's original code, our implementation of FDOD, and our implementation with *precisely those errors* injected into our code; no other changes are made. We considered two datasets from Jin et al. (2003): the low-homology dataset (hereafter referred to as T30-1401), which was available to us, and a high-homology dataset (hereafter referred as 359), for which we reconstructed 332 out of the original 359 sequences; the remaining sequences became obsolete or have been updated. In addition, we considered another

* Corresponding author. Tel.: +1 780 492 5488; fax: +1 780 492 1811.
E-mail address: lkurgan@ece.ualberta.ca (L. Kurgan).

1476-9271/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2006.06.003

Table 1
Experimental results; *results from* column: 1, cited in Jin et al. (2003); 2, experiments with code provided by authors of Jin et al. (2003); 3, with our duplicated (errors injected) code; 4, with correct code (the two errors are excluded); *L*, subsequence length cited in Jin et al. (2003)

Dataset	*L*	Results from	Jackknife					Resubstitution				
			α	β	α/β	$\alpha+\beta$	Overall	α	β	α/β	$\alpha+\beta$	Overall
T30-1401	1	1	74.83%	67.45%	55.02%	35.49%	57.74%	75.84%	69.79%	55.94%	37.96%	59.39%
		2	75.17%	67.74%	55.02%	35.19%	58.28%	76.18%	70.09%	56.16%	36.73%	59.79%
		3	75.17%	67.74%	55.02%	35.19%	58.28%	76.18%	70.09%	56.16%	36.73%	59.79%
		4	**69.13%**	**61.58%**	**46.58%**	**23.77%**	**49.75%**	**75.84%**	**69.50%**	**56.85%**	**37.04%**	**59.39%**
	2	1	75.50%	72.14%	66.21%	40.74%	63.74%	83.89%	79.18%	72.37%	56.17%	72.73%
		2	75.50%	72.14%	65.07%	40.43%	63.29%	83.89%	78.59%	72.15%	57.41%	73.01%
		3	75.50%	72.14%	65.07%	40.43%	63.29%	83.89%	78.59%	72.15%	57.41%	73.01%
		4	**70.13%**	**70.97%**	**61.87%**	**35.80%**	**59.81%**	**80.87%**	**79.18%**	**73.97%**	**55.56%**	**72.45%**
	3	1	79.19%	75.07%	84.02%	58.95%	75.02%	98.66%	99.41%	99.77%	99.69%	99.43%
		2	79.19%	75.07%	84.02%	58.95%	75.02%	98.66%	99.71%	99.77%	99.69%	99.46%
		3	79.19%	75.07%	84.02%	58.95%	75.02%	98.66%	99.71%	99.77%	99.69%	99.46%
		4	**63.76%**	**68.33%**	**74.66%**	**44.75%**	**63.88%**	**95.97%**	**97.07%**	**96.80%**	**93.21%**	**95.86%**
359	1	1	67.10%	60.00%	42.40%	39.80%	51.50%	69.50%	63.50%	46.50%	41.90%	54.60%
332		2	69.74%	66.67%	43.62%	40.74%	54.52%	76.32%	66.67%	44.68%	45.68%	57.53%
		3	69.74%	66.67%	43.62%	40.74%	54.52%	76.32%	66.67%	44.68%	45.68%	57.53%
		4	**40.79%**	**44.44%**	**29.79%**	**17.28%**	**32.83%**	**76.32%**	**66.67%**	**45.75%**	**45.68%**	**57.83%**
359	2	1	67.10%	80.00%	86.90%	54.80%	72.40%	78.10%	92.90%	100.00%	78.50%	87.70%
332		2	69.74%	83.95%	86.17%	61.73%	75.90%	85.53%	93.83%	98.94%	77.78%	89.46%
		3	69.74%	83.95%	86.17%	61.73%	75.90%	85.53%	93.83%	98.94%	77.78%	89.46%
		4	**64.47%**	**77.78%**	**82.98%**	**49.38%**	**69.28%**	**82.90%**	**91.36%**	**100.00%**	**77.78%**	**88.55%**
359	3	1	91.50%	95.30%	100.00%	95.70%	95.80%	100.00%	100.00%	100.00%	100.00%	100.00%
332		2	90.79%	98.77%	100.00%	93.83%	96.08%	100.00%	100.00%	100.00%	100.00%	100.00%
		3	90.79%	98.77%	100.00%	93.83%	96.08%	100.00%	100.00%	100.00%	100.00%	100.00%
		4	**67.11%**	**64.20%**	**42.55%**	**58.03%**	**57.23%**	**81.58%**	**86.42%**	**58.51%**	**82.72%**	**76.51%**
25PDB	1	3	62.30%	57.79%	48.56%	26.19%	49.63%	63.43%	58.92%	48.56%	28.84%	50.87%
		4	58.24%	52.82%	40.17%	13.76%	42.42%	63.43%	58.92%	48.56%	28.84%	50.87%
	2	3	75.16%	67.74%	55.02%	35.19%	57.82%	76.18%	70.09%	56.16%	36.73%	59.32%
		4	59.60%	54.18%	47.11%	23.55%	46.96%	68.62%	65.91%	63.30%	44.71%	61.12%
	3	3	**31.38%**	**37.92%**	**81.50%**	**26.72%**	**42.86%**	**46.73%**	**49.89%**	**53.76%**	**33.60%**	**46.02%**
		4	**45.83%**	**48.53%**	**51.73%**	**32.54%**	**44.72%**	**92.33%**	**93.68%**	**96.53%**	**94.97%**	**94.22%**

low-homology dataset consisting of 1610 sequences selected from the publicly available 25%PDBSELECT list (Hobohm and Sander, 1994) (hereafter referred as 25PDB). Our empirical results with these datasets are presented in Table 1. For the T30-1401 and 359 datasets, the author's implementation and our implementation *with the methodological errors injected* achieved *identical* results, demonstrating that our analysis has correctly identified the errors. By comparison, resubstitution and jackknife accuracy in our clean implementation is substantially lower; see values in bold in Table 1. The reader will note a slight discrepancy between some of the results reported in Jin et al. (2003) and the author's own implementation; we are unable to provide an explanation for this discrepancy. Nonetheless, our empirical results clearly show that the high accuracy achieved in Jin et al. (2003) is a methodological artifact. Note that the results obtained for $L=3$ on 25PDB dataset with the correct implementation are superior to the faulty implementation; see values in bold in Table 1. The prediction accuracy for the 25PDB datasets is on average lower for both the faulty and the correct implementations when compared with the other two datasets. This is most likely due to lower homology of the 25PDB dataset. Finally, we observe that results obtained by the correct implementation of FDOD are more consistent than those of the faulty code.

2. Conclusions

We have examined the claims of high structural class prediction accuracy on a low-homology dataset reported in Jin et al. (2003). Code analysis of the authors' original implementation, followed by empirical testing, confirms that these claims are the result of methodological errors. Additional empirical testing on a separate low-homology dataset did not result in acceptable performance.

References

Hobohm, U., Sander, C., 1994. Enlarged representative set of protein structures. Protein Sci. 3, 522.
Jin, L., Fang, W., Tang, H., 2003. Prediction of protein structural classes by a new measure of information discrepancy. Comput. Biol. Chem. 27, 373–380.