BIOINFORMATICS LETTER TO THE EDITOR Vol. 29 no. 3 2013, page 412
doi:10.1093/bioinformatics/bts687

Systems biology

Advance Access publication November 28, 2012

Concerning the accuracy of Fido and parameter choice

Oliver Serang

Department of Neurobiology, Harvard University and Proteomics Center, Boston Children's Hospital, 320 Longwood Avenue, Boston, MA 02115, USA

Associate Editor: Alex Bateman

Contact: Oliver.Serang@Childrens.Harvard.edu

Received and revised on September 17, 2012; accepted on November 22, 2012

A recent article by Huang and He (2012) illustrates a method for applying linear programming to protein inference and compares this method to a handful of established methods for protein inference: ProteinProphet (Nesvizhskii *et al*., 2003), MSBayes (Li *et al*., 2008) and Fido (Serang *et al*., 2010). Unfortunately, Huang and He (2012) gives the impression that protein probabilities computed by Fido are erratic, yielding either markedly superior performance (e.g. in the Sigma 49 dataset) or inferior performance (e.g. in the HumanMD dataset) compared with the rest of the field; however, this erratic performance is simply the effect of not choosing the parameters by grid search as specified in Serang *et al*. (2010). Performing a rough grid search ($\alpha \in \{0.01, 0.04, 0.09, 0.16, 0.25, 0.36\}$, $\beta \in \{0.0, 0.01, 0.025, 0.05\}$ and $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$) to jointly optimize discrimination and calibration substantially improves the quality of the probabilities computed (Fig. 1) and makes Fido competitive with ProteinProphet. Performing the grid search takes 4.42 s on a Core i3 laptop.

Not only are Fido's three parameters important for weighing many pieces of low-scoring peptide evidence against fewer pieces of high-scoring peptide evidence, but they also substantially determine the treatment of shared (i.e. 'degenerate') peptides. As a result, the optimal values for these parameters are influenced by many factors (e.g. the coverage of the experiment, the complexity of the sample and so forth), and these parameters are, understandably, dataset-specific; although using blind parameter estimates may give reasonable performance in some situations (particularly when the parameters come from a dataset with similar characteristics), Fido is *not* intended to be run without these parameters intelligently set (via the grid search noted earlier in the text or manually).

To be fair, there are situations where the means to estimate free parameters (i.e. true- and false-positive labels) are not available, and in these cases, it is not unreasonable to try parameters estimated from another similar data set; however, even in this case, the parameters in Fido are not intended to be chosen *arbitrarily* without regard for their performance. If no labels are available, manual parameter choice is an option (e.g. choosing parameters that rank a known highly abundant protein over others). In the rare case when *no* information for validation is available, Fido, like all models with free parameters, should not be used.

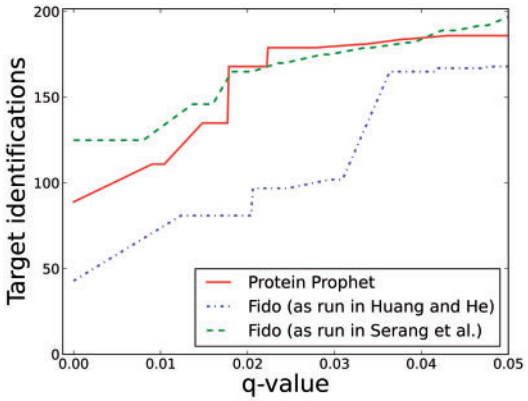

Fig. 1. The accuracy of Fido with poor and well-chosen parameters. Fido's performance on an example dataset (HumanMD) from Huang and He (2012) is significantly better when choosing parameters as specified in and Serang *et al*. (2010). ProteinProphet, a well-known method for protein inference, is shown for comparison

To better enable researchers to use Fido in the way intended and to aid in future comparisons, we have released an updated executable, which automatically performs the grid search as described in Serang *et al*. (2010) (without need of the previous shell script wrapper). It is available for free download from http://noble.gs.washington.edu/proj/fido/.

Funding: This work is funded by NIH grants NS007473 and NS066973.

Conflict of Interest: none declared.

REFERENCES

Huang,T. and He,Z. (2012) A linear programming model for protein inference problem in shotgun proteomics. *Bioinformatics*, **28**, 2956–2962.

Nesvizhskii,A. *et al*. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.*, **75**, 4646–4658.

Li,Y.F. *et al*. (2008) A Bayesian approach to protein inference problem in shotgun proteomics. *Lect. Notes Bioinform.*, **12**, 167–180.

Serang,O. *et al*. (2010) Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *J. Proteome Res.*, **9**, 5346–5357.