Molecular Systems Biology 7; Article number 537; doi:10.1038/msb.2011.70
Citation: ***Molecular Systems Biology* 7: 537**
© 2011 EMBO and Macmillan Publishers Limited All rights reserved 1744-4292/11
www.molecularsystemsbiology.com

CORRESPONDENCE

The self-assessment trap: can we all be better than average?

Molecular Systems Biology **7**: 537; published online 11 October 2011; doi:10.1038/msb.2011.70

Computational systems biology seems to be caught in what we call the 'self-assessment trap', in which researchers wishing to publish their analytical methods are required by referees or by editorial policy (e.g., Bioinformatics, BMC Bioinformatics, Nucleic Acids Research) to compare the performance of their own algorithms against other methodologies, thus being forced to be judge, jury and executioner. The result is that the authors' method tends to be the best in an unreasonable majority of cases (Table I). In many instances, this bias is the result of selective reporting of performance in the niche in which the method is superior. Evidence of that is that most papers reporting best performance choose only one or two metrics of performance, but when the number of performance metrics is larger than two, most methods fail to be the best in all categories assessed (Table I). Choosing many metrics can dramatically change the determination of best performance (Supplementary Table S1). Selective reporting can be inadvertent, but in some cases biases are more disingenuous, involving hiding information or quietly cutting corners in the performance evaluation (similar problems have been discussed in assessments of the performance of supercomputers, e.g., Bailey (1991)).

Even assuming that there is no selective reporting, we would like to argue that papers reporting good-yet-not-the-best methods (of which we found none in our literature survey of self-assessed papers listed in the Supplementary information) can still advance science. For example, a method that is not top ranked can still have value by unearthing biological results that are complementary to the results reported by other better performing methods. Furthermore, the effectiveness of a top-performing algorithm can be boosted when its results are aggregated with second and third best performers (Figure 1, and Supplementary Figures S1 and S2; Marbach *et al*, 2010; Prill *et al*, 2010). The discussion above suggests that self-evaluation is suspect and that insistence on publication of only best performing methods can suppress the reporting of good-yet-not-best performing methods that also have scientific value.

In biosciences, as well as in other natural sciences, we are often faced with situations that have been referred to as uncomfortable science, a term attributed to statistician John Tukey, in which the little available data are used both in the inference model and the confirmatory data analysis. The resulting overoptimistic 'confirmatory' results are often referred to as 'systematic bias'. Similarly, 'information leak' from data to methods can occur from improper and repeated cross-validation. In the general case, information leak results from developing or training an algorithm based on the entire available data set so that the test set is not independent. In some cases, the leak can occur subtly and inadvertently such as when a very similar sample is present both in training and test set. A better-known effect is 'overfitting', in which a model is developed with superior accuracy on its training data at the cost of reduced generalization of the model to new data sets. A notable example of this effect can be found in the

Table I Break out of 57 surveyed papers in which the authors assess their own methods

Number of performance metrics	Total number of studies surveyed	Authors' method is the best in all metrics and all data sets	Authors' method is the best in most metrics and most data sets
1	25	19	6
2	15	13	2
3	7	4	3
4	4	1	3
5	4	1	3
6	2	1	1

Note that we did not find any self-assessment paper where the presented method was not top ranked in at least one metric or data set. The survey was conducted over a large pool of scientific peer-reviewed papers selected as follows. First, a Google Scholar search using the keywords 'computational biology method assessment' was conducted. When papers with comparisons of methods were identified, we further examined (1) papers from the same journal issue and (2) downstream papers that cite the identified paper (as determined by Google Scholar). The 57 papers (see Supplementary information) resulting from the search span 22 journals. Most papers are in the categories of gene regulatory networks/reverse engineering (24/69), structure prediction/assessment (14/69) and DNA–protein interactions/regulatory element identification. An additional nine papers found in the same manner but not shown in the Table reported independently (not-self) assessed methods, of which only four were top performers, whereas five reported methods that ranked high but were not top performers.

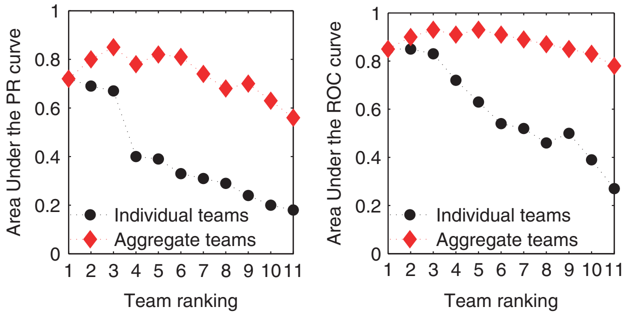

Figure 1 The performance metrics Area Under the Precision–Recall curve (AUPR, left panel) and Area Under the ROC curve (AUROC, right panel) for individual teams participating in a DREAM2 challenge. The challenge consisted of predicting transcriptional targets of the transcription factor BCL6. Even when as the performance of the individual teams decreases (black line and circles), the integrated prediction of the best performer and runner-up teams (red line and diamonds) outperformed the best individual team.

search for biomarker signatures in cancer. For about a decade, scientists have scoured high-throughput data to find collections of genes or proteins that can be used in diagnosis or prognosis of cancer. However, the tools used to find signatures in massive data sets can yield spurious associations with phenotype (Ioannidis, 2005), even when the results appear to be statistically sound in self-assessment. In most cases, unfortunately, these signatures do not generalize; taken to the task of showing the diagnostics or prognostics value of these signatures, the accuracy of the predictions is much poorer on impartial assessments on previously unseen patients than on the original data. This problem with cancer signatures is of sufficient general interest to be highlighted recently in the popular media (Kolata, 2011).

In order to alleviate the overestimation of accuracy from the many bias sources described above, we proposed a few guidelines:

(i) use third-party validation to test a model with previously unseen data
(ii) use more than one metric to evaluate the methods
(iii) report well-performing methods even if they are not the best performers on a particular data set
(iv) increase the awareness of editors and reviewers that superior performance in self-assessment is a biased demonstration of the method's value; instead, impartial assessment should be the preferred evaluation
(v) Establish a scientific culture that values timely, well-conducted follow-up studies that confirm or refute previous results

To a large extent, the remedies suggested above have been addressed in the context of genome-wide association studies (Chanock *et al*, 2007), and are embodied in existing independent assessments presented to the scientific community in efforts such as CASP (http://predictioncenter.org/), CAPRI (http://www.ebi.ac.uk/msd-srv/capri/) and DREAM (http://www.the-dream-project.org). In contrast to the usual practice of 'post-diction' (retrospective prediction) of known results as a way to test their methods, participants to these third-party collaborative competitions (alternatively known as challenges) submit predictions that are evaluated by impartial scorers against an independent data set that is hidden from the participants. The level of performance in these evaluations better tests the generalization ability of the methods, because the predictions are made based on unseen data, thus minimizing many of the above-discussed biases. We envision that a repository of blind challenges and data sets could be created (DREAM, for example, has 20 such data sets and challenges) with data produced on demand by third parties, especially funded to create verification data and challenges. This repository could be used to test the validity of many of the tasks that we deal with in Systems Biology, Bioinformatics and Computational Biology.

In summary, systematic bias, information leak and overfitting can all be considered facets of the same self-assessment trap. That is, by knowing too much about the desired results, the researcher gets snared into a trap of consciously or unconsciously overestimating performance. Moreover, the researcher is further lured to the trap by the common assumption that top performance is required for scientific value and publication. By exposing the self-assessment trap, we hope to lessen its effect with the ultimate goal of advancing predictive biology and improving human healthcare.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Conflict of interest

The authors declare that they have no conflict of interest.

Raquel Norel, John Jeremy Rice and Gustavo Stolovitzky
IBM Computational Biology Center, IBM T.J. Watson Research Center,
Yorktown Heights, NY, USA

References

Bailey DH (1991) Twelve ways to fool the masses when giving performance results on parallel computers. *Supercomputing Rev* **4:** 54–55

Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, Brooks LD, Cardon LR, Daly M, Donnelly P, Fraumeni Jr JF, Freimer NB, Gerhard DS, Gunter C, Guttmacher AE, Guyer MS *et al* (2007) Replicating genotype-phenotype associations. *Nature* **447:** 655–660

Ioannidis JP (2005) Microarrays and molecular research: noise discovery? *Lancet* **365:** 454–455

Kolata G. (2011) *Add Patience to a Leap of Faith to Discover Cancer Signatures*. The New York Times, New York (http://www.nytimes.com/2011/07/19/health/19gene.html?_r = 1&scp = 1&sq = Add%20Patience%20to%20a%20Leap%20of%20Faith%20to%20Discover%20Cancer%20Signatures%22&st=cse)

Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc Natl Acad Sci USA* **107:** 6286–6291

Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G, Stolovitzky G (2010) Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS One* **5:** e9202