

## The-more-the-better and the-less-the-better

To many biologists, geneticists and bioinformaticians, the excitement of genomics comes from systematic analyses of large amounts of information, such as complete end-to-end DNA sequences, densely packed genetic markers on chromosomes, and sometimes, comprehensive population genetics history in places like Iceland and Finland where extensive genealogical data are available. Also, realizing the importance of sample sizes in mapping susceptibility genes in complex and common diseases, various national and international consortiums were established, and meta analyses were frequently in use on pooled data. Almost everybody agrees that ‘the more (information), the better’.

There are two senses in the word ‘more’ used here. One concerns the search space, and another concerns the sample size. It is easy to understand why one would like to see both or either one of them to be large. The reason for demanding a complete search space is that we do not want to miss anything. If we fail to detect a genetic linkage or association signal for a human disease, could it be that we have not covered all genomic regions with enough markers, or is it because we have not compiled a complete list of all coding genes? Having a complete search space will remove these doubts.

The reason for larger sample size is well-known in statistics: if the statistical signal is weak, only a larger dataset has a chance to uncover it with confidence. Also, if the sample size is much smaller than the number of variables, such as the ‘large  $p$  (number of genes), small  $n$  (sample size)’ situation in microarray data, the variable space is not adequately explored, and degenerate fitting models are possible. The ‘large  $p$ , even larger  $n$ ’ situation is preferable.

‘The more, the better’ trend is a natural byproduct of the genomic era, and will undoubtedly continue as ever more advanced biotechnology produces bioinformation faster and cheaper. However, for a specific biology project or a particular human disease study, not all genes are involved and not all chromosomal regions are relevant. An equally important process of removing the irrelevant information allows us to focus on the key areas. A cartoon in Weiss and Terwilliger (2000) compared the search of human disease genes with finding a needle in a haystack. In this example, reducing the haystack size instead of increasing it helps the chance of finding the needle. This principle might be called ‘the less, the better’.

‘The less, the better’ is a recurring theme in many other fields: Occam’s razor and the principle of parsimony in philosophy, signal-to-noise ratio and the curse of dimensionality in engineering, simplicity of fundamental laws in physics, minimum length description and lossy data compression in computer science, Morgan’s Canon in animal psychology, and so on. The topic of model and variable selection (Burnham and Anderson, 2002) in statistical learning is aimed at simplifying description of data without sacrificing the quality of description.

In genomics, genetics and bioinformatics, the parsimony principle has been actually practiced often. The whole point of filtering out low-density and non-differentially-expressed genes in a microarray discriminant analysis is to remove genes that are unlikely to contribute to the phenotype difference. For some microarray

datasets, classifiers with extremely small number of genes are able to distinguish normal and diseased samples or samples of different types.

In the leukemia data from Golub *et al.* (1999), for example, the expressions of *zyxin*, a component of cell adhesion plaques which may play a role in signal transduction, are thoroughly higher in all 11 acute myeloid leukemia (AML) samples than in all 27 acute lymphoblastic leukemia (ALL) samples (Li and Yang, 2002). Even if such a perfect classification rate is not achieved in a larger dataset, this single-gene classifier should still perform well in distinguishing AML and ALL. Similar observation on classification successes with smaller classifiers was also made in Xiong *et al.* (2001).

Rare Mendelian diseases provide another, somewhat trivial, example for requiring fewer genes in describing the data. In most cases, only one gene is behind a rare Mendelian disease. The genetic heterogeneity is probably the only source of complication. As for complex diseases, the number of genes involved in the disease surely increases. But the increase is not without a limit and only a very small proportion of the total  $\sim 25\,000$  human genes is expected to be the susceptibility genes. The idea of exclusion mapping (Edwards, 1987) is to delete signal-less chromosome region from further investigation. One may also apply the variable selection technique to narrow down the marker list as well as the chromosome regions (Li and Nyholt, 2001).

Interestingly, there is a similar debate in statistics on whether the best statistical model should contain fewer variables and parameters. According to the late professor Leo Breiman (Breiman, 2001), the majority of traditional statisticians prefer simple models that have simple explanation and perhaps tractable mathematics; whereas machine-learning-inspired statisticians prefer ‘black box’ and complicated algorithm with better prediction performance.

As for the latter category, a new breed of predictors has appeared under the names such as ensembles methods, model aggregation and model averaging. These approaches can achieve excellent prediction rates even when no particular care is taken to reduce the variable list to a minimum. In the example of ‘random forest’ (Amit and Geman, 1997), when it was applied to microarray classification datasets, it performed comparatively well as with other predictors without tree pruning (Díaz-Uriarte and Alvarez de Andrés, 2006). Does using multiple models violate the principle of ‘the less, the better’?

In machine learning and statistics literature, there were discussions on two different interpretations of Occam’s razor: one being ‘simplicity is a goal in itself’, and the other is ‘simplicity leads to greater accuracy’. Some people expressed reservation on the second interpretation (Domingos, 1999). Another argument for ensemble methods is that owing to bias-variance trade-off (Hastie *et al.*, 2001), a single model either overfits the data (small bias and large variance) or underfits the data (large bias and small variance). Having many models somehow reduce the variance, reminiscent of the variance reduction by having many samples.

We can view this issue more intuitively. First, are the irrelevant variables and parameters kept in an aggregation model really contributing to the prediction? In Bayesian subset selection (George and McCulloch, 1993) for example, all variables are kept but some of them have zero weight and others only contribute to the data

marginally. This may lead us to the concept of ‘effective number of variables/parameters’. Second, what is the nature of variation among samples that cause a prediction failure? Is it because of some true biological and genetic heterogeneity? If so, then using multiple models indeed makes sense.

Suppose a human disease is caused by (say) three genes and one environmental factor, a model or a classifier with these four variables should be the optimal model. Rather than ‘the more, the better’ or ‘the less, the better’, the answer is ‘number four is the best’. Unfortunately, this number and the nature of the disease etiology is often unknown to us. If the disease is heterogeneous, caused by these four factors in some patients but by different genes in others, then an extra layer of complexity is introduced to the data. Perhaps the new principle should be stated as follows: keep the appropriate level of model complexity that matches that of the data and at least throw away the irrelevant information.

Wentian Li

*The Robert S Boas Center for Genomics and Human Genetics,  
Feinstein Institute of Medical Research,  
North Shore LIJ Health System,  
Manhasset, NY 11030, USA*

## REFERENCES

- Amit, Y. and Geman, D. (1997) Shape quantization and recognition with randomized trees. *Neural Comput.*, **9**, 1545–1588.
- Breiman, L. (2001) Statistical modeling: the two cultures. *Stat. Sci.*, **16**, 199–231.
- Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multi-Model Inference*. 2nd edn. Springer-Verlag, New York, NY.
- Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.
- Domingos, P. (1999) The role of Occam’s razor in knowledge discovery. *Data Mining and Knowl. Disc.*, **3**, 409–425.
- Edwards, J.H. (1987) Exclusion mapping. *J. Med. Genet.*, **24**, 539–543.
- George, E.I. and McCulloch, R.E. (1993) Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.*, **88**, 881–889.
- Golub, T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
- Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning*. Springer-Verlag, New York, NY.
- Li, W. and Nyholt, D.R. (2001) Marker selection by Akaike information criterion and Bayesian information criterion. *Genet. Epidemiol.*, **21** (Suppl. 1), S272–S277.
- Li, W. and Yang, Y. (2002) How many genes are needed for a discriminant microarray data analysis. In Lin, S.M. and Johnson, K.F. (eds), *Methods of Microarray Data Analysis*. Kluwer Academic, pp. 137–150.
- Weiss, K.M. and Terwilliger, J.D. (2000) How many diseases does it take to map a gene with SNPs? *Nature Genet.*, **26**, 151–157.
- Xiong, M. *et al.* (2001) Feature (gene) selection in gene expression-based tumor classification. *Mol. Genet. Metabolism.*, **73**, 239–247.