J Comput Aided Mol Des (2012) 26:1097–1099
DOI 10.1007/s10822-012-9613-7

Call for Papers: GRC, CADD, and statistics, and all that

Anthony Nicholls

Received: 27 September 2012 / Accepted: 28 September 2012 / Published online: 9 October 2012

Abstract Molecular modeling and the art of computer-aided drug discovery seldom make much use of statistics, despite being fields that can not calculate important properties with great reliability. The 2013 CADD Gordon conference intends to examine what prevents a more effective use of statistics in routine modeling and to raise consciousness as to what is possible. Practical methods will be discussed, deeper issues in applying standard approaches addressed and research on successes and failures in other disciplines presented by invited experts.

Keywords Statistics · Molecular modeling · CADD

Introduction

The design of novel pharmaceuticals is one of the biggest, if not the biggest, industry on the planet and yet has been stubbornly resistant to industrialization. Other fields of endeavor, such as aeronautics, chemical manufacturing, even the perennially hopeful oil and gas industries, have all evolved from empiricism to highly refined and productive enterprises. Not so drug discovery. And while drug discovery is a heady brew of many components, the application of computers to the design of novel pharmaceuticals, or CADD, is ironically perhaps the least advanced in terms of its reliability. There are many good reasons for this, chief of which is that properties of importance are difficult, if not impossible, to calculate accurately, even given access to ever-growing computer power. We in CADD are still largely empirical and probably will be so for at least another decade or so. So if we must be empirical, we should at least attempt to be the best empiricists we can be. Which means statistics.

There seems to be few things more likely to make grown scientists cringe than the mentioning of statistics. Sometimes this is because there is a feeling that "Real Science" doesn't need statistics (ignoring that even the Higgs boson was not considered 'found' until the statistics were sufficient). Sometimes there is the "You can prove anything with statistics" view point, to which the rejoinder is that whether this is true or not it is even easier to prove anything without statistics. Finally, there is a self-consciousness that comes from having to admit one has forgotten most of what one knew, which wasn't very much to begin with—statistics not being a large part of our formal training.

It doesn't have to be so. Personally, after taking the time to both remind and teach myself some practical statistics I was surprised by how straightforward such methods (eventually) seemed and how indisputably useful they were. It is my belief that CADD would be well-served by becoming at least more aware of the possibilities statistics, both ancient and modern, offer to our daily work. It is with this in mind that I am making statistics an unusual centerpiece of the 2013 Gordon conference for CADD (Mount Snow, Vermont, July 21st–25th).

There are five aspects of this event that may interest members of the field: practical methods, profound issues, societal matters, outside perspectives and a critical poster session.

Practical methods

Perhaps one of the most glaring examples of statistically unaware molecular modeling has become is simply to look

A. Nicholls (✉)
OpenEye Scientific Software, Inc., Santa Fe, NM, USA
e-mail: anthony@eyesopen.com

for the prevalence of error bars in publications and presentations. Effects are presented unadorned with the very uncertainties necessary to judge a findings worth and validity. Alternatively, confidence limits are presented for different methods run over the same dataset to indicate equivalence of approaches—yet do not take account of covariance, i.e. error bars for the differences between tools. This is basic stuff. How do we, as a field, expect to make progress without mastering the tools that tell us how reliably method A is better than method B? This is but a simple, yet I think telling, example of the limitations the field labors under. There are many more.

There are also a whole host of useful methods that should be in a modelers toolbox: how to calculate means, variances, slopes and intercepts robustly, how to predict the best bin widths for histograms, how to estimate the expected degradation of performance from training to test set, when to classify and when to regress—or how to do both, averaging quantities of unequal certainty, distinguishing discrete distributions, knowing the significance of R**2 values, applying simple tests for model parsimony, using expected confidence limits from bootstrapping or other methods, calculating a reliable enrichment factor, adjusting for multiple hypothesis testing, optimizing the information from multiple time point measurement.. the list is a long one. These are all immensely practical and once explained cease to be "statistics" but instead become the way we do things.

Profound questions

There are some significant challenges that modeling poses to statistics—for instance, classical statistics is based on the concept of independence between samples, the so-called IID assumption (Independent and Identically Distributed). Typical datasets used to validate tools in modeling are neither independent nor identically distributed. Suppose, for example, a docking program ranks a thousand compounds. If the compounds were independent of each other then the probability any given compound is in fact active may depend on its rank (hopefully!) but should not depend on what compounds are ranked higher. However, suppose ten of the thousand compounds belong to a closely related series A, and nine of them are at the top of the list and prove to be active. No matter where the tenth compound falls in the list the likelihood of it being active is not independent of the fact that nine closely related molecules are active. Alternatively, suppose we are using a ligand-based method and a series is found to be active by similarity to a known ligand, but a second active series is missed entirely and, on investigation, it is found to act by a different mechanism—the underlying distribution (of properties, characteristics, variables) has changed for the second series. In our field we talk about this as the 'domain' of the problem and of the difficulties of making predictions 'out of the domain' of the training/knowledge set. These problems are difficult for statistics too.

Analysis is also complicated by the application of parameters to gain better agreement with observed data, at the risk of decreasing the effectiveness of future predictions (from the same distribution). CADD is far from unique in its poor appreciation of how to assess over-parameterization. Common methods, such as cross-validation, y-scrambling and even prospective prediction have fundamental problems, confounded by the non-IID behavior of the test data. For example, cross-validation can spot an over-parameterized model **asymptotically**,—i.e. in the limit of infinite data. With finite data there is a finite chance of mis-classification—a risk poorly understood in most applications. And non-ideality plays havoc with the reliability of all measures—is it significant to predict a compound very similar to the training set as active when this has a high likelihood anyway?

Which brings us to control studies. Far too many papers are published with little or no control experiments. The simplest, alluded to above, is to test how well 2D similarity methods work. Yes, your method may be getting great results but that needs to be assessed relative to the expected behavior of other simpler, faster, possibly more reliable approaches. NULL models are typically ones that do not attempt to use the results in-hand, such as model pKas, affinity by nearest neighbor comparison, docking by maximum common substructure overlap, but a more comprehensive discussion of what constitutes a reality check for new approaches is sorely needed.

Finally, how much should we trust the data we use to construct models? How much does it matter? How do we make robust models? What data do we need and how should we disseminate it? There are both statistical and society issues that ought to be addressed if modeling is to become a more mature discipline.

Societal issues

Just improving the field's awareness and comprehension of statistical practice will not help unless we understand the limits of experimental measurement, have access to relevant data, have journals that insist on proper standards and members of our field willing and able to reproduce the statistically extraordinary, and therefore interesting, result. Individual excellence in statistics will not alone help the field grow and prosper. As such, I hope to make this meeting a forum within which we can critically assess how the whole field may become better institutionally. We need

to grow beyond an ensemble of anecdotes and examples to robust rules of thumb and reliable practices. This is seldom discussed formally and it is about time we so did.

External opinions

Our field is but one of many trying to make sense of difficult to model data. An example of the consequence of thinking somehow our field is "different" is a common dalliance with strange new measures of performance rather adopting well-validated, widely-used approaches. Such behavior typically prevents rather than illuminates progress. Other fields have found their way forward by adopting better statistical standards, some are still inching their way there. It would be a mistake not to learn from their successes and their failures. As such, I am inviting leaders from other disciplines, such as economics, psychology, biology and machine learning, to share their experiences.

Methods posters

Typical CADD Gordon conferences consist of presentations on methods or applications of methods, occasionally with associated statistics although more often not. At the 2013 conferences such talks will be highlighted in the poster session where analysis of the statistical validity can be more easily be assessed. It is hoped that both by having experts present at the meeting and having multiple opportunities to learn during the sessions that poster sessions can represent how a more statistically adept field might appear.

Conclusions

Many have commented that this is a bold undertaking, and others that a conference on "statistics" will not fare well in this community. I agree with the former and not with the latter. This field has ignored the basics, the fundamentals by which we appreciate and value progress and as such does not progress, at least very rapidly. Furthermore, this is very far from a meeting about statistics, it is a meeting about how to use statistics to make our work better, more meaningful and more relevant. That is, perhaps, bold, but inevitably necessary. A meeting can occasionally make a difference to a field and I hope you will join me in making the CADD 2013 Gordon conference just such a occasion.