Predicting causal effects in large-scale systems from observational data

To the Editor: Understanding cause-effect relationships between variables is of primary interest in many fields of science. The standard method for determining such relationships uses randomized controlled perturbation experiments. In many settings, however, such experiments are expensive and time consuming. Hence, it is desirable to obtain causal information from observational data, that is, from data obtained by observing the system of interest without subjecting it to interventions.

There are established methods to estimate causal effects from observational data when the possible causal relationships between the variables are known[1]. Many real-world problems, however, involve large-scale systems without such information. Although it is generally impossible to estimate causal effects in such systems, we recently proposed and mathematically justified[2] a statistical method to obtain bounds on total causal effects, under some assumptions (**Supplementary Methods**). We call this method intervention-calculus when the DAG is absent (IDA). IDA has not been experimentally validated until now, and there is a lack of experimental validation of causal inference methods in general.

We present here an experimental validation of IDA. As a first test, we used a compendium of gene expression profiles of *Saccharomyces cerevisiae*[3], containing 267 full-genome expression profiles of yeast deletion mutants (interventional data), together with 63 full-genome expression profiles of negative control experiments (observational data), all obtained under the same conditions. After initial data cleaning (**Supplementary Methods**), the interventional data contained expression measurements of 5,361 genes for 234 single-gene deletion mutant strains, and the observational data contained expression measurements of the same 5,361 genes for 63 wild-type cultures.

We used the interventional data as the gold standard for estimating the total causal effects of the 234 deleted genes on the remaining genes (that is, 234 × 5,360 effects; **Supplementary Methods**). We defined the top m percentage of these effects, where $m = 5$ or 10, as our target set and evaluated how well IDA could identify these effects from the observational data. We found that the q largest predicted effects from IDA, where $q = 50, 250, 1{,}000$ or 5,000 effects, corresponded significantly more often to effects in the target set than can be expected by random guessing, with $P < 0.001$ for all combinations of m and q (**Supplementary Table 1** and **Supplementary Methods**). For instance, for $m = 10$ and $q = 50$, IDA found 33 (66%) true positives, whereas random guessing yielded only 5 ± 2.1 true positives (10% ± 4.2%). Moreover, IDA improved substantially on Lasso[4] and Elastic-net[5], two state-of-the-art high-dimensional regression approaches commonly used to determine variable importance but not designed for causal inference (**Fig. 1a**, **Supplementary Table 1** and **Supplementary Methods**). For $m = 10$ and $q = 50$, these methods yielded 10 (20%) and 8 (16%) true positives, respectively. Finally, we found that the superior performance of IDA compared to that of the other methods was insensitive to the choice of m value for $m = 1, \ldots 50$ (**Fig. 1b**).

As a second test, we used data from the DREAM4 In Silico Network Challenge[6], a competition in reverse engineering of gene regulation networks. These data include several types of simulated mRNA expression levels, based on sophisticated biologically motivated simulation methods[6], for five networks of 10 genes and five networks of 100 genes. We used two types of observational data: (i) steady-state gene expression levels from unknown multifactorial perturbations of the networks and (ii) time series data on gene expression levels from the response and recovery of the networks to unknown external perturba-

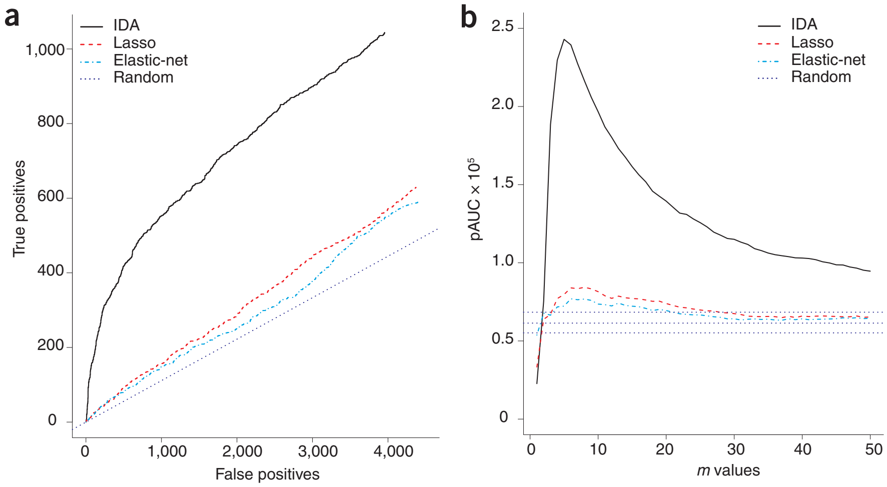

Figure 1 | Predicting causal effects from observational data (data are from ref. 3). (**a**) The number of true positives versus the number of false positives are plotted for the indicated methods, for the top 5,000 predicted effects from the observational data. The target set is the top 10% of the effects as computed from the interventional data. (**b**) The partial area under the receiver operating characteristic curve (pAUC) is plotted versus *m* values, when the target set is the top *m* percentage of the effects as computed from the interventional data. The pAUC was computed up to the false-positive rate determined by the top 5,000 effects from IDA for *m* = 10. The three horizontal lines for random guessing correspond to the 2.5th, 50th and 97.5th percentiles of a simulated distribution based on random orderings of effects.

tions (omitting the time stamps). We used interventional data on steady-state gene expression levels of known single-gene knockout experiments as the gold standard for determining the causal effects. We applied IDA, as well as Lasso and Elastic-net, to the observational datasets and evaluated how well the resulting top q predicted effects ($q = 10$ for the networks of size 10 and $q = 25$ for the networks of size 100) corresponded to the top m percentage ($m = 5$ or 10) of the effects as computed from the interventional data (**Supplementary Methods**). We counted the number of networks in which the partial area under the receiver operating characteristic curve (pAUC) was better than random guessing at significance level $\alpha = 0.01$ for both values of m (**Supplementary Methods**). By this measure, IDA was at least as good as Lasso and Elastic-net for all four possible combinations of the type of observational data (multifactorial or time series) and the size of the networks (10 or 100 genes). The difference was largest for the multifactorial data on the networks of size 10, where IDA was substantially better than Lasso and Elastic-net for three of the five networks (**Supplementary Fig. 1** and **Supplementary Table 2**). For instance, in this setting with $m = 10$ and $q = 10$, IDA found 4, 4, 5, 1 and 2 true positives for the five different networks, whereas Lasso found 1, 1, 0, 1 and 2 true positives and Elastic-net found 3, 1, 0, 1 and 1 true positives.

The results presented here on *S. cerevisiae* and the DREAM4 data are proof-of-concept results that IDA can predict the strongest causal effects in potentially large-scale biological systems by using only observational data. In particular, the results on *S. cerevisiae* demonstrate that we were able to do this in a challenging real-world setting where the number of variables (5,361) was much larger than the sample size (63) and the variables were substantially disturbed by noise. As IDA is supported by mathematical theory, we expect the results presented here to generalize to other problems.

Of course, statistical predictions based on observational data can never replace intervention experiments. In fact, whenever possible, IDA predictions should be followed up by intervention experiments. In this way, the predictions can serve as a new tool for the design of experiments, as they indicate which interventions are likely to show a large effect.

Software for IDA is available in the open source R-package pcalg (http://cran.r-project.org/web/packages/pcalg/index.html).

Note: Supplementary information is available on the Nature Methods website.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Marloes H Maathuis[1], Diego Colombo[1], Markus Kalisch[1] & Peter Bühlmann[1,2]

[1]Seminar for Statistics, Department of Mathematics, Eidgenössische Technische Hochschule (ETH) Zurich, Zurich, Switzerland. [2]Competence Center for Systems Physiology and Metabolic Diseases, Zurich, Switzerland.
e-mail: maathuis@stat.math.ethz.ch

1. Pearl, J. *Causality: Models, Reasoning, and Inference*. (Cambridge Univ. Press, Cambridge, UK, 2000).
2. Maathuis, M.H., Kalisch, M. & Bühlmann, P. *Ann. Stat.* **37**, 3133–3164 (2009).
3. Hughes, T.R. *et al. Cell* **102**, 109–126 (2000).
4. Tibshirani, R. *J. Roy. Statist. Soc. Ser. B* **58**, 267–288 (1996).
5. Zou, H. & Hastie, T. *J. Roy. Statist. Soc. Ser. B* **67**, 301–320 (2005).
6. Marbach, D., Schaffter, T., Mattiussi, C. & Floreano, D. *J. Comp. Biol.* **16** 229–239 (2009).

A method and server for predicting damaging missense mutations

To the Editor: Applications of rapidly advancing sequencing technology exacerbate the need to interpret individual sequence variants. Sequencing of phenotyped clinical subjects will soon become a method of choice in studies of the genetic causes of Mendelian and complex diseases. New exon-capture techniques will direct sequencing efforts to the most informative and easily interpretable protein-coding fraction of the genome. Thus, the demand for computational predictions of the impact of protein sequence variants will continue to grow.

Here we present a new method and the corresponding software tool, PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/, **Supplementary Software**), for predicting damaging effects of missense mutations. PolyPhen-2 is different from the earlier tool PolyPhen[1] in the set of predictive features, the alignment pipeline and the method of classification (**Fig. 1a**). PolyPhen-2 uses eight sequence-based and three structure-based predictive features (**Supplementary Table 1**), which were selected automatically by an iterative greedy algorithm (**Supplementary Methods**). The majority of these features involve comparison of a property of the wild-type (ancestral, normal) allele and the corresponding property of the mutant (derived, disease-causing) allele. The alignment pipeline selects a set of homologous sequences using a clustering algorithm and then constructs and refines its multiple alignment (**Supplementary Fig. 1**). The most informative predictive features characterize how likely the two human alleles are to occupy the site given the pattern of amino-acid replacements in the multiple-sequence alignment; how distant the protein harboring the first deviation from the human wild-type allele is from the human protein; and whether the mutant allele originated at a hypermutable site[2]. The functional importance of an allele replacement is predicted from its individual features (**Supplementary Figs. 2–4**) by a naive Bayes classifier (**Supplementary Methods**).

We used two pairs of datasets to train and test PolyPhen-2. We compiled the first pair, HumDiv, from all 3,155 damaging alleles annotated in the UniProt database as causing human Mendelian diseases and affecting protein stability or function, together with 6,321 differences between human proteins and their closely related mammalian homologs, assumed to be nondamaging (**Supplementary Methods**). The second pair, HumVar[3], consists of all the 13,032 human disease-causing mutations from UniProt and 8,946 human nonsynonymous single-nucleotide polymorphisms (nsSNPs) without annotated involvement in disease, which we treated as nondamaging.

We found that PolyPhen-2 performance, as presented by its receiver operating characteristic curves, was consistently superior compared to that of PolyPhen (**Fig. 1b**) and it also compared favorably with that of three other popular prediction tools[4–6] (**Fig. 1c**). For a false positive rate of 20%, PolyPhen-2 achieved true positive prediction rates of 92% and 73% on HumDiv and HumVar datasets, respectively (**Supplementary Table 2**).

One reason for the lower accuracy of predictions on HumVar is that nsSNPs assumed to be nondamaging in the HumVar dataset included a sizable fraction of mildly deleterious alleles. In contrast, most amino-acid replacements assumed nondamaging in