Report[1] on the first the DREAM-Project Steering Committee Workshop

On March 9 and 10, 2006 the steering committee for the DREAM project (DREAM stands for “Dialogue on Reverse Engineering Assessment and Methods”) convened to clarify the goals and procedures for the DREAM project. Held under the auspices of the NYAS Systems Biology Discussion Group, the Columbia University Center for the Multiscale Analysis of Genetic Networks (MAGNet), and the IBM Computational Biology Center, the event featured an evening symposium, on March 9, followed by an all-day, closed-door workshop on March 10, to discuss ideas. The conclusions of the closed door symposium, summarized below, will form the basis of the first open meeting for DREAM, scheduled for September 7 and 8, 2006, at the Wave Hill Convention Center, in New York.

In the evening discussion, Stolovitzky and Califano described their vision of a repository for data, algorithms, and literature that will help researchers evaluate and compare their reverse engineering methods. One inspiration for the project is the decade-old Critical Assessment of methods of protein Structure Prediction (CASP) project, which was described by John Moult of the University of Maryland. Unlike information about protein structure, however, data about network behavior take many different forms. Marc Vidal of Harvard University and the Dana-Farber Cancer Institute described one of those forms, the "interactome" of all direct physical interactions between proteins.

The next day's workshop featured various viewpoints about the types of networks that might be targeted for reverse engineering, the types of data that could be used for the task, and the appropriate metrics for evaluating success. The May 10 session featured a number of thought-provoking presentations. Presenters included Diego di Bernardo, Mark Gerstein, Andre Levchenko, Pedro Mendes, Mike Snyder and Mike Yaffe, who were joined by many other members of the DREAM steering committee.

At the time of this writing the steering committee for the DREAM project is composed of the following members (names in bold face attended the workshop and/or the evening discussion):

[1] This report is a synthesis of the full report produced by Don Monroe, which can be seen in the e-briefing on the DREAM workshop, in the web site of the NYAS: www.nyas.org.

Gary Bader (Univ of Toronto)
Joel Bader (Johns Hopkins Univ)
Diego Di Bernardo (TEGEM)
Hamid Bolouri (Inst for Sys. Biology)
Harmen Bussemaker (Columbia Univ)
Andrea Califano (Columbia Univ)
Jim Collins (Boston University)
Tim Gardner (Boston Univ)
Mark Gerstein (Yale Univ)
Alexander Hartemink (Duke Univ)
Trey Ideker (UCSD)
Andre Levchenko (Johns Hopkins)
Pedro Mendes (Virginia Tech)
John Moult (Univ of Maryland)
Andrey Rzhetsky (Columbia Univ)
Benno Schwikowski (Inst Pasteur)
Eran Segal (Weissmanm Inst of Science)
Ron Shamir (Tel Aviv Univ)
Mike Snyder (Yale Univ)
Gustavo Stolovitzky (IBM)
Marc Vidal (Harvard Univ)
Mike Yaffe (MIT)

During the May 10 session there was ample time for discussions, which were cordial but revealed significant differences among the diverse participants. An important focus of the session was how to compare the effectiveness of different methods to reverse engineer the network that produces a particular set of data. Evaluating this requires a "gold standard" network for which at least the true topology of connections is known. Many participants, especially the computational biologists, believe that synthetic networks, particularly if motivated by known biological pathways, are good candidates for this purpose because, at least for now, they are the only ones for which true and false interactions can be described with certainty. Experimental biologists, however, worry that unless the project addresses real biological networks, it could evolve into a mathematical exercise with little impact on biology.

In the following we make a synthesis of the topics discussed in the workshop and some preliminary conclusions. We end up with open questions, which we will continue to address in subsequent DREAM workshops.

Goals and Challenges of the DREAM project

What is DREAM. The DREAM project is intended to provide a forum for dialogue among researchers interested in the reverse engineering of biochemical interaction networks in the cell. The following are some of the challenges associated with the creation of such a project.

- *Gold Standard.* In the context of biological network reverse engineering, a Gold Standard is a reference network whose topology, interactions, and (possibly) kinetic parameters are exactly known and for which measurement data can be readily generated. Researchers trying to measure the performance of their methods could then quantitatively assess how much their reconstructed networks deviate from the reference one. Unfortunately no large-scale biological system is understood nearly well enough to serve this purpose. Artificial networks are well understood but may have limited relevance to biology. The absence of an agreed-upon Gold Standard is an important challenge for DREAM and represent an important difference with respect to the CASP model.

- *Comparison of Reverse Engineering methods.* The community can best understand the strength and weaknesses of different reverse engineering methods by testing them, on the same blind data generated from "Gold Standard" biological systems, where interactions are known with 100% certainty.
- *Limitations of In-silico Gold Standards.* Although researchers sometimes compartmentalize biological networks as "transcriptional," "signaling," or "metabolic," in reality none of these types of networks exists as an isolated entity, as each layer influences the others. The apparent interactions between events on a single layer, such as regulation of gene expression, must thus include events in other layers, albeit indirectly. Most synthetic networks address each layer in isolation. Additionally, such networks do not result from a evolutionary process but rather from random processes. As a result, some intrinsic network properties, such as homeostasis and fault tolerance, are not germane to synthetic networks. However, some recent models have emerged that are both multilayer and motivated by real, albeit partial, biological interaction knowledge.
- *Limitations of Experimental Gold Standards:* A key issue with experimental networks is that, in general, only a small fraction of the true biochemical interactions in a biological system are known. This poses a double challenge, making it very difficult to determine if an inferred interaction is indeed correct (true positive) or incorrect (false positive).
- *What do the network-edges represent?* Cellular networks are generally represented as graphs, where the edges represent interactions between different molecular species. Some high-throughput techniques hold the prospect of complete maps of one layer of interactions. For instance, protein-protein interactions have been exhaustively studied and compiled into "interactomes". In this case, an edge simply means that two proteins bind to each other, forming a stable or transient complex. As multiple interaction layers are considered, however, it becomes more and more complicated to understand what an edge truly represents. In some cases, for instance, a directed edge may indicate that the kinetics of one specie is a function (albeit an indeterminate one) of other species. A clear, although possibly context-specific definition of what edge represents should be a goal of the DREAM effort.
- *What do the network-nodes represent?* In some cases it is not even clear what the nodes of the network ought to represent. Frequently the nodes represent individual species such as a specific protein isoform in a given cellular compartment. At other times, the nodes collapse an entire array of gene byproducts, including mRNA, protein, phosphorylated proteins, etc, regardless of cellular compartment and sometimes across multiple tissues within an organism or across multiple species. Perhaps it would be better to represent the elementary "machines" that emerge from close coupling between different molecular species. Some researchers, for instance, have suggested the use of individual domains and binding sites as network-nodes.

- *Integration of different cellular networks.* Currently, most pathway inference approaches address a specific layer of biochemical interactions, such as transcriptional (Protein-DNA), signaling (Protein-Protein), protein complex (Protein-Protein), or metabolic (Protein-metabolite) interactions. The deepest biological insight will require combining different types of information, including data on gene expression, proteins and small molecules among others. Because of the wide variety of data types and network types, finding a common framework will be a continuing challenge for the DREAM project.
- *DREAM database.* The DREAM project will maintain a database that will serve as a clearinghouse for three important classes of data. One is a "database of high-probability predicted interactions" in search of validation by experimental biologists. Crisp predictions about the results of removing specific interactions in a real biological context should also get the attention of experimental biologists. The second is "a database of high-value cellular circuitry data that requires concerted attention by the reverse engineering community." The third one is a database of predictions of observable changes in cell behavior resulting from specific cellular perturbations. These resources along with the recurring workshops will help unite computational and experimental biologists as a community.
- *DREAM workshops.* The DREAM project will organize a yearly workshop. Once a suitable set of Gold Standards is defined, the workshop may adopt the double-blind competitive assessment style of the CASP meeting. I.e., the scientists will not know the structure of the network they are trying to predict, and the evaluators will not know who made the predictions. At least initially, however, this competitive structure will not be appropriate for the workshop, which may better serve the research community as a tool for self-evaluation. Whether the workshop will ever switch to a competitive framework will be determined by the Steering Committee in future editions of the workshop.
- *DREAM workshop challenges.* An additional challenge that DREAM will have with respect to CASP is that network experimental data and behavior tend to emerge in piecemeal fashion, unlike a protein structure that is unveiled as a whole. Even more challenging, however, is the much greater diversity of available data types and generative models as well as the lack of agreement on how to evaluate success in reverse engineering.
- *Different viewpoints and emphasis.* By highlighting the divergent viewpoints of its diverse participants, the March 10 meeting illuminated issues that must be addressed to pull together a true community of experimental and computational biologists to understand biological networks. It is clear that each of the many approaches to the problem offer some important insights into this challenging problem.

Where Should the Data Come From?

What model system? The DREAM project aims to compare reverse engineering methods by providing data from selected systems and by comparing the networks they infer using agreed-upon metrics. A critical issue, especially in this early phase, is what organisms and networks should be selected to generate the data to be used for method testing.

- *Two extremes for the model system.* As previously discussed, candidate networks for reverse engineering mostly fall into two extremes. On the one hand, in silico networks, whether inspired by biology or by mathematics. On the other hand, real biological networks for which there is consensus among biologists.
- *In silico models.* These networks let researchers rapidly explore what synthetic data types may be more convenient for network inference: is it mRNA expression profiles, proteomics data, measurements of phosphorylation states? They also can be used to answer questions about how sensitive their methods are to changes in topology, parameter values, kinetics, data availability, and noise. However, even biologically grounded mathematical models might be potentially far removed from their actual biological counterpart. For example, models may not include effects such as RNA interference, alternative splicing, and post-translational modifications.
- *In vivo models.* On the other hand, in vivo biological systems will contain a great many biochemical interactions that are uncharacterized/missing (false negatives) as well as interactions that have been incorrectly identified/characterized (false positives). The resulting simplified network models may therefore idealize and distort the true underlying interaction model. In other words, there is no firmly established experimental gold standard, and therefore, extracting data from these systems may not teach us if our methods are accurate because we don't know the extent to which they represent the biology.
- *A Middle ground: bioengineered systems.* Experimentalists were clear that they would be more interested in simple experimental systems rather than in complex synthetic model. To combine the strengths of the two types of systems, the tools of bioengineering can be used to insert relatively simple "designer networks" in an organism, such as yeast, or cell line. Unlike natural systems, such bioengineered systems would be almost perfectly known. Moreover, researchers who want to extend the results will not be limited to analyzing experimental data: they can be given the actual yeast strain and the corresponding perturbation plasmids. Results from such *in vivo* networks are more likely to be accepted as relevant by the experimental biology community than purely in silico models.
- *More discussions are needed.* No clear consensus emerged during the discussions about the best source of data for testing reverse engineering. The critical conflict between definitive evaluation and biological relevance is complicated by the very different metrics needed to evaluate success in models depending on whether the network is known. Probably each of the generative models has its own lessons to teach. As a result, the first workshop will explore all three alternatives and attempt to determine which ones are mature and acceptable to the community.

What Data Best Capture Network Behavior?

What kind of data should we gather? A critical issue both for the DREAM database and workshops is the choice of what data should be used for reverse engineering.

- *RNA, proteins or small molecules?* Biological networks involve interactions between several types of molecular species, including but not limited to DNA, RNA, proteins, and metabolites. Expression profile microarrays let researchers monitor the average mRNA concentration in a cell population on a genome-wide scale. Yet, other high-throughput techniques, such as transcription-factor binding microarrays (chip ChIP)), protein (MS-MS) and metabolite profiles, and genome-wide identification of protein-protein interactions and post-translational modifications, are less mature. Reverse engineers can also exploit reliable quantitative data from smaller-scale, more laborious techniques. It is likely that no technique provides all the necessary data for error-free network inference, and combining different types of data increases the chances of identifying biologically relevant relationships. The issue of appropriate datatypes will thus be a central one for the upcoming DREAM workshops.
- *Challenge for the DREAM database.* The diversity of data sources, and their uneven (albeit rapidly improving) quality, pose a challenge for DREAM in providing a repository for data useful for the task of inference of biological networks. We need to create a mechanism to determine which datasets will be initially made available to the community. The availability of predictions from a consistent set of data will make it somewhat easier to compare results in the absence of a perfect Gold Standard.
- *Quantitative data constrains models.* Even though some data types are rather qualitative in nature, quantitative data are crucial for specifying the more sophisticated network models. For example, time-dependent, quantitative data can provide tight constrains to model parameters.
- *Perturbing the networks.* Perturbations are a classic way to test network models of biological systems. Among these perturbations are the normal changes in internal or external cellular environment. Intentional changes such as heat shock or starvation modify the behavior of the systems, sometimes initiating large-scale changes in cellular behavior. In one sense, however, although these experiments reflect the response to a changing environment, they do not necessarily change the character of the network as such. Techniques such as gene knockout or RNA interference give experimental researchers the ability to manipulate specific aspects of the network itself. These tools essentially create new networks and are a powerful way to test reverse-engineered network models. Similarly, predictions of measurable biological response to perturbation, especially if biologically interesting, can be used to define assays that would at least indirectly validate the quality of the inferred models.
- *One more difference between CASP and DREAM.* In the protein-structure field, the CASP project benefits from the fact that the form of the starting sequence data is clearly defined. In contrast, the huge diversity of possible data for characterizing network behavior complicates the task of presenting it within a single framework to reverse-engineers. Such a diversity of input information makes it hard to ensure that different groups are using the same input for their task and complicates comparison of the methods. Disguising the identity of the network and its components would ensure a level playing field, but could degrade the value of the predictions. In addition

researchers may be reluctant to attack problems without knowing their biological relevance. The DREAM project will need to address these data challenges to make the evaluation of reverse engineering methods convincing.

- *Getting network data across species.* Even in protein structure prediction, however, researchers often exploit other knowledge, such as structures of analogous proteins, to improve their predictions. Similar information could make a big difference for reverse engineers as well. Using parallels with networks in other organisms, or evolutionary data, or various types of annotation or other information from the literature could improve the quality of networks.

Metrics for Evaluating Reverse Engineering Algorithms

As the DREAM project aims to compare reverse engineering methods, a critical issue is the choice of what the appropriate metrics for the comparison are. As there are many types of data that can be used for pathway inference and many ways to organize these data into networks and models, coming up with metrics that are good for all of them may be extremely hard.

- *Comparison with CASP.* Developing metrics even for the more straightforward task of protein structure prediction is not easy. For example, an algorithm might capture the essence of a structural motif, but if that motif were in the wrong position, its atoms would all deviate substantially from their correct coordinates. A metric that averaged displacements would show a poor match, although for many purposes the prediction would be good.
- *Don't forget the false negatives.* Comparing the results of a reverse engineering exercise seems easier for mathematically created networks or for the elusive "gold standard" biological system that is completely understood. In this case, the most basic comparison between a model network and its target is the topology: which nodes connect to which others? One popular test of a putative network is whether the inferred "edges" between nodes really exist. Although such positive tests are necessary, a realistic evaluation must also check the negative results: how often does the method fail to predict edges that really exist? In general, making a test more stringent eliminates false positives, but at the cost of increasing false negatives. There are typically many more negative connections than positive. Therefore appropriate metrics should be designed that test for both false positives and false negatives. "ROC" curves can be used to quantify the false positive/false negative tradeoff.
- *Indirect assessment of connections.* In the absence of a gold standard for real biological network, researchers could compare emergent behavior (which could be qualitative, such as over and under-expression, or quantitative, such as response to perturbations, time courses, etc.) between the model and data collected from its target. In general, the best way to evaluate a model network depends on both the nature of the model and the target and to the type of data supplied. The data used for evaluation may be the same as those used to infer the network, or some subset of the data may be held back for testing purposes.

- *What output of the model is more important?* The resulting metric will depend critically on what importance is ascribed to different features of agreement and disagreement. Emphasizing different features of the observed behavior can change the choice of network model, and/or the outcome of the comparison. The choice of what output to compare between model and target emphasizes the aspects of the system that researchers ultimately care about. This may be particularly appropriate because some parameters of the model network may be virtually impossible to determine from experiment.
- *Does the model represent the real system?* Matching only the output data, however, raises the question of how "real" the inferred networks are. Data may not constrain the model sufficiently, and more than one model could be consistent with the data. In addition to reporting the models, reverse engineers should propose new experimental perturbations that will illuminate the differences between model networks.
- *Parameter sensitivity in quantitative models.* In silico models of real biological processes have a behavior that is quite sensitive to the values of a handful of parameters. Most other parameters, however, could be varied over wide ranges without any obvious effect on the network behavior. It is important to recognize this limitation when evaluating reverse engineering algorithms, but it may be hard to know in advance which parameters will be poorly determined.
- *Validation of specific predictions.* One possible metric to be used is to ask for a concrete prediction to a given perturbation. So given the same data, and for a given perturbation, a metric could be devised to choose the model that best predicts the output behavior in the target system.

Open Questions

A number of questions remain that we hope will be the basis for fruitful discussions.

1. What is the best way to combine (at least) three levels of description: genes, proteins, metabolites?
2. Should DREAM be structured as a competition?
3. How can blinded network behavior data be obtained, represented, and publicized?
4. How can the modeling community best engage experimental biologists?
5. What kinds of data and networks should be represented in the DREAM database?
6. How can the project best balance the precision of artificial models with biological relevance?
7. What procedures can best ensure the quality of data in the DREAM database?