

## **Receptor Tyrosine Kinase (RTK) Consortium Data and Modeling Workshop**

June 21, 2005

Environmental Molecular Sciences Laboratory

Pacific Northwest National Laboratory

### **RTK Members/Others in Attendance**

Boris Kholodenko, Consortium Chair, Thomas Jefferson University, USA  
Mariko Hatakeyama, Consortium Information, Data and Model Management, Secretariat Asia-Pacific, RIKEN Genomic Sciences Center, Japan  
Steven Wiley, Consortium Communication, Data and Model Management, Secretariat Americas, Pacific Northwest National Laboratory, USA  
Jan Hoek, Thomas Jefferson University, USA  
Herbert Sauro, Keck Graduate Institute, USA  
Vijay Chickarmane, Keck Graduate Institute, USA  
Sri Paladugu, Keck Graduate Institute, USA  
Haluk Resat, Pacific Northwest National Laboratory, USA  
David Tuck, Yale University School of Medicine, USA

### **Pacific Northwest National Laboratory Staff**

Mary Ace, Bill Cannon, Abbie Corrigan, Banu Gopalan, Julie Gephart, Kavita Patel-Stenoien, Karin Rodland, Eric Stephan, Ron Taylor, Brian Thrall, Bobbie-Jo Webb-Robertson, Katrina Waters

### **Workshop Purpose**

The purpose of this workshop was to define the first models that will serve as an initial focus of the four participated labs (PNNL, RIKEN, Keck Graduate Institute, Thomas Jefferson) of the RTK Consortium and to decide where data on these models will be stored, and to decide which types and formats of data to be used. These models will be preliminary and serve as a test bed for efforts to coordinate modeling and experimental work of the Consortium.

It will always be a challenge to combine modeling and experimental studies, especially for a distributed consortium. By choosing a focused set of experiments and data types, the expectation is to learn how to set up and use the infrastructure necessary to drive progress in the Consortium over the next decade.

### **Presentations**

Four presentations were made by Consortium members, describing areas of interest and capabilities at their respective labs. Each is summarized here.

### ***Data and Data Structures at PNNL – Steven Wiley, PNNL***

PNNL is studying the HMEC 184 system, a non-tumorigenic, relatively normal, immortal cell line. Basal phenotype. It can be grown in serum-free medium, but requires insulin/IGF1 and hydrocortisone for growth. Its growth requires activation of the EGFR. HMEC 184 is well characterized with respect to protein composition, gene expression profiles, and trafficking dynamics, and there are numerous genetically modified variants available.

Normal mammary epithelial cells express four different ligands that bind to the EGFR (TGF- $\alpha$ , EPR, AR, HB-EGF) and three different receptors (EGF-R, HER2, HER3) of the EGF family. Proteolytic processing is required for ligand activity. Removing the membrane-anchoring domain of EGF disrupts the ability of HMEC to organize in a 3D culture system.

PNNL is mapping inputs of growth factor dose, mutations, genetic alterations, and inhibitors to outputs of migration, microarrays, apoptosis, and proliferation using phosphorylation, enzyme assays, and substrate location. They are planning phosphoproteome analysis (non-directed MS-based, antibody-based), kinase assays (ERK, AKT, JNK1, IKK, MK2), and translocation assays (ERK, AKT).

PNNL's Biomolecular Systems Initiative (BSI) is assembling instruments and systems for high-throughput proteomics, imaging, and metabolite analysis; building the software infrastructure to capture and analyze high-throughput data; and creating a network-based modeling system to enable quantitative evaluation of complex systems.

PNNL has a data management system for proteomics data and is in the testing stage of an open-source system for collecting and managing microarray, Western blot, flow cytometry, and cell culture data. Plans are to add imaging and other high-throughput data in the near future.

**Suggested Projects.** Wiley had the following list of proposed projects for the Consortium:

- Transactivation: Modulate IGF1 and EGFR system (antibodies and ligands) and look at ERK activation, proliferation, or gene expression as output.
- Overexpression: Modulate HER2, HER3, and/or HER4 and examine the effect on EGF/hereregulin response.
- Oncogenes: Understand how vIII EGFR functions (it seems to act as dominant negative EGFR).

### ***RIKEN Data for Collaboration and Preliminary Data Analysis – Mariko Hatakeyama, RIKEN Genomic Sciences Center (GSC)***

RIKEN's ongoing projects:

Mathematical modeling of ErbB receptor signal transduction pathways

1. Kinetic model of intracellular signal transduction pathways
2. Prediction of ligand-induced gene regulatory network.

Ongoing cell lines are

- CHO: E1, E4, E1/4 (EGF, HRG-induced Western blot for 1)

- MCF-7 breast cancer (EGF, HRG-induced Western blot, microarray for 1 and 2).

Testing cell lines are PC-3 prostate cancer, LNCaP prostate cancer, and SK-OV-3 breast cancer. Microarray and kinetic modeling of overexpression, mutation and deletion of signaling-related genes in above cell lines are also in progress.

MCF-7 cells

Microarray time course

- Untreated, HRG (10 nM) EGF (10nM)—17 time points – good for identification of ligand-specific gene expression
- EGF, HRG (0.1, 0.5, 1, 10 nM)—early eight time points – good for identification of target genes and combining intracellular signaling and gene expression. Can estimate gene network.

Their gene network strategies include

Development of new algorithms for less computation time and larger network

- Combination with methods such as S-system, neural NW, Bayesian, GGM
- Incorporation of *a priori* knowledge (PPI information, transcription factor binding site)
- Resistance to noise.

They used the bootstrap method to compare EGF and HRG (heregulin) microarray data.

RIKEN has the following resources available.

Hardware

- Parallel computers for gene network prediction, parameter estimation.
- Molecular dynamics simulation via custom processors for protein-protein interaction analysis and *in silico* screening.

Tools

- ODE simulator YAGNS
- Parameter estimation algorithms
- Gene network estimation—S-system/neural network model
- PPI, TF databases
- Statistical analysis
- Protein 3D structure analysis of target molecules.

### **Suggested Projects**

- Kinetic modeling of EGF/HRG intracellular signaling in cancer cells – reasons of ligand specific kinetic patterns; what are key molecules/pathways/genes.
- Gene regulatory network comparison among cancer cells- for identification of tissue specific/cellular specific regulatory molecules.
- Integration/or overlay of intracellular signaling and gene networks
- Effect of mutation, deletion, over-expression of specific genes on the above to make a compatible model for cancer systems.

## *Network Models of RTK Signaling – Boris Kholodenko, Thomas Jefferson University*

Kholodenko reiterated the challenges of both bottom-up and top-down approaches to modeling RTK networks:

Bottom-up challenges:

- Incomplete knowledge of molecular mechanisms
- Lack of kinetic measurements of “isolated” signaling reactions
- Combinatorial explosion of the number of network states and spatial inhomogeneity

If you try to do a realistic mechanistic model of any pathway, the number of states of the network goes beyond the millions. It’s impossible to write that many differential equations.

Top-down challenges:

- Quantitative proteomics of protein phosphorylation
- Methods of computational inferring the architecture of cellular networks.

By maps of a network, he means the topology and strength of network connections.

In three graphs of the time course of EGF receptor signaling, he showed that temporal responses are controlled by 1) relative protein abundance, 2) phosphorylation-induced changes in the binding affinities, and 3) relative fractions of signaling proteins complexed with phosphatases.

Kholodenko gave an example of using a model to clarify the signaling puzzle. Data on the mismatch of SOS and Ras activation patterns suggest the strong control by p120 RasGAP. The dose-response patterns of SOS fraction in phosphotyrosine immunoprecipitates correlate with the kinetics of EGFR activation, whereas the Ras response to EGF is markedly more sustained.

**Modeling distinct mechanisms of the control of RasGAP activity.** Nobody has ever been able to show co-immunoprecipitation of RasGAP and EGFR—there’s nothing in the literature. The reason is that if you do computation, you see that the concentration of complex is <1 nanomole. In modeling the control of RasGAP activity, they found that data on isolated hepatocytes rule out some mechanisms, such as the change in RasGAP catalytic activity following the membrane recruitment by PIP3. There are no methods to measure RasGAP activation.

RasGAP is recruited to the plasma membrane not only by EGFR. RasGAP associates with p190 RhoGAP following EGF stimulation of isolated hepatocytes. Computational modeling explored feasible molecular scenarios, including the receptor- and PIP3-mediated recruitment of SOS and RasGAP to the plasma membrane, phosphorylation of RasGAP and p190 RhoGAP by soluble tyrosine kinases, and RasGAP interactions with phosphoinositides and p190 RhoGAP. Modeling suggests that a transient RasGAP association with EGFR followed by the capture of RasGAP through the formation of complexes with p190 RhoGAP can account for data on hepatocytes.

**Reducing combinatorial complexity of multidomain proteins.** Macro-state is a sum of microstates. Spatial separation of kinases and phosphatases in MAPK cascades can result in a dramatic decrease in the phosphorylation signal near the nucleus. If diffusion is not efficient,

how do signals propagate from membrane receptors to the nucleus? Cells may exploit additional mechanisms that involve phospho-protein trafficking within endocytic vesicles, scaffolding and active transport of signaling complexes by molecular motors and phospho-protein waves.

The interaction map of a cellular regulatory network is quantified by the local response matrix. The degree to which one protein affects another can be measured if you go to steady state of the system and increase each enzyme by 1%; fix all the directions. The response would be simply derived from Jacobian matrix/element.

**Untangling the wires: tracing functional interactions in signaling and gene networks.** The goal is to determine and quantify unknown network connections. The problem is that network (system) responses (R) can be measured in intact cells, whereas local response matrix, r (network interaction map), cannot be captured unless the entire system is reconstituted in vitro.

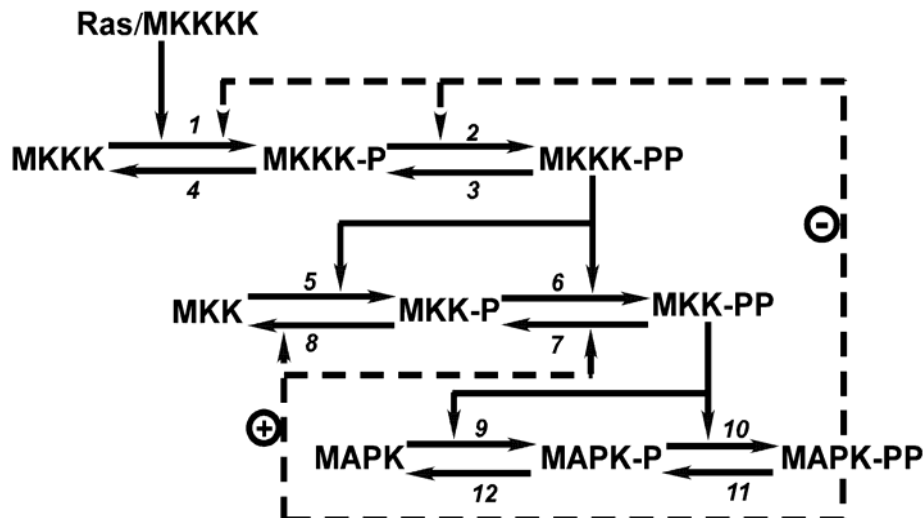
The steady-state solution is as follows: connection coefficients (Matrix r) are determined from the steady-state responses to successive perturbations to all modules (Matrix R):  $r = - (dg(R^{-1}))^{-1} \cdot R^{-1}$

Step 1: Determine global responses to three independent perturbations of the Ras/MAPK cascade, then measure the differences in steady-state variables following perturbations and generate the system response matrix.

Step 2: Calculate the Ras/MAPK cascade interaction map from the system responses. Two interaction maps (local response matrices) are retrieved from two different system response matrices.

Inferring dynamic connections in MAPK pathway is successful based on time dependence using orthogonality. First, one can deal only with the steady state. Another, more powerful method is to measure time course—Jacobian or time correlation—if possible.

The following figure is a kinetic diagram showing the transition of MAPK pathway from resting to stable activity state.



Oscillatory dynamics of the feedback connection strengths is successfully deduced.

***Challenges to Modeling Signaling Pathways. Yeast Sporulation Use Case – Vijay Chickarmane, Keck Graduate Institute***

The Keck Graduate Institute group is examining the IME1-IME2 interaction. The absence of nutrients induces IME1. They are doing automated reconstruction of networks with San Diego Supercomputing Center (SDSC) using Cytoscape. This software platform uses several databases to pull out all nearest neighbor interactions for a particular gene/protein. Hand curation is required to obtain more detailed interaction information and to convert the network into a stoichiometric network suitable for analysis. Stoichiometric models are stored in SBML format.

In modeling the gene-protein interactions, they find that rate laws are determined by the free energies of binding of the transcription factors and the RNAP (?). These include the interaction energies: cooperativity/inhibition.

Most kinetic data were derived from literature: kinetic constants, activities, absolute concentrations, and localization. They used Discovery Tool to search for possible oscillators and bistable switches in stoichiometric networks. The tool uses a search technique based on a genetic algorithm.

For parameter estimation and model validation, given a model and experimental data, an objective function is needed that tests the ability of the model to reproduce the experimental data, an optimization method that can be reliably used to generate the optimized model, and a means to determine whether a particular model is potentially correct. Fitting ODE-based models involves comparing a simulation with experimental times series data and adjusting parameter values in the model until the simulation matches the experiment. An objective function assumes that the experimental data are compatible with the data generated by the simulation, usually absolute data.

Fitting the model to the data is a problem. They find they need to do a lot of multiparameter simulation. They want to generate an ensemble of likely models. Monte Carlo runs using metropolis sampling to generate an ensemble. Each member represents a model with a reasonably good fit: 1) obtain confidence limits, 2) cluster plots in parameter space show correlations, 3) PCS give direction along which model is stiff/sloppy, and 4) stiff directions indicate important “control” points in the model.

Much current experimental data is, however, expressed in terms of relative changes. In such a case we can only hope to fit the qualitative changes. One approach is to fit correlations between paired data sets and corresponding pairs in the model.

The most difficult part of model fitting is discriminating between different model variants. We test models by their ability to fit a set of data. If a particular models fits better then it is deemed,

using appropriate statistical tests, that the model is the best model. How do we discriminate between several models, each of which fits the data reasonably well?

If the proposed model were in fact a genuine candidate, then they expect the parameter spaces to coincide. This is based on the assumption that rate constants do not change as a result of external or internal perturbations.

They are comparing two different probability distributions. The success for the methods requires assurance that the parameter space has been adequately sampled. They are investigating two approaches: 1) multiple GA (?) runs at different initial conditions, using bootstrap; and 2) running a few GAs, selecting the optimum from among these and then using Markov Chain Monte Carlo. Statistical tests need to be devised to determine when two parameter spaces are genuinely distinct.

**Summary.** KGI's goal is to build networks from literature and existing curated models and discover interesting dynamics. To fit models to data, it's essential to have absolute data. They will compare competing models by the invalidation approach.

### **Consortium Path Forward**

After the presentations, Steven Wiley led a discussion of the path forward for the RTK Consortium.

### **Data Needed**

First we need to define the system's topology, then go to the network, then to the simulation, then to the parameters. Need to determine which inputs and outputs to use. They must be varied widely, because we don't know what's upstream of this flow.

Input → signaling network → output

MIT is developing approaches to rapidly infer topology from large data sets—they are using HT-29 cell lines. Their concept is cue→signal→response. The relationship between signal and response appears to be constant, and so the real challenge is to model how cues (inputs) give rise to specific signals.

It's a matter of the whole constitution of things. If you look at the topology, it's built on cell stoichiometry. If there are different levels of components, how does that change output?

Microarrays have better time resolution than proliferation assays. These are global, but they feed back into the signaling layer. It can be difficult to separate out primary versus secondary responses but with care, it can be done. Ultimately, you need to find outputs that are fast enough, which is an experimental challenge. As a compromise, we could use microarrays to provide quantitative output data at 2 hours ± 30 minutes. Then most of the responses should reflect the initial state of the network and its topology.

Kholodenko's group has already investigated the kinetics of the EGFR response. In 0-10 minutes everything has peaked or passed. This response is almost certainly due to signals generated at the cell surface. In 20-30 minutes, you get into major changes; 30-60 minutes is never-never land where not much happens.

As a general rule, 0-10 minutes is a period of time within which many different signaling events happen. Therefore, it is a good starting point. This should simplify the modeling. It is not clear what to use as the output—one or multiple endpoints? It may well evolve into a longer time frame—however, the first need is to simplify the approach for the initial set of studies.

EGFR to ERK could be the extent of the initial model. We could use the level of ERK activity (or phosphorylation) as the output. That would be a very useful because it is well understood and somewhat constrained. For the cells used by PNNL (184A1 HMEC), you get a huge ERK response to EGF, but hardly any AKT. IGF1 gives you a great AKT response, but ERK response only by transactivating the EGFR

These profiles vary tremendously from one cell type to another, qualitatively. We don't know what regulates the intensity of signaling through all of the different pathways, and that's one of the most interesting questions. The literature shows that there are both qualitative and quantitative differences in how these things are operated, but these differences have to be linked to specific molecular mechanisms.

It would be great for us to work out a topology for one cell type, then immediately go to another one to see if we find another topology or whether we only find differences in the quantitative aspects of the components (levels of receptors, altered rate constants, etc.)

A possible long-term project might be finding if there are cell-specific maps or common maps. How much can differences between cells be ascribed to differences in the abundance of signaling proteins in the cell? These could have the exact same wiring diagram, but a different flow. This would be really interesting to explore.

There is general agreement to use ERK activity as the initial output because that leads to the next stage. We want to saturate the parameter space. Understand all of the factors that influence the generation of ERK activity. What do we want to assay? We can set up a large series of experiments to explore regulation of ERK activity. We will want to change time, dose, HER2, HER3, HER4 expression. We can get a dataset with lots of different inputs and measure the output. EGFR phosphorylation state would be easy to do, although it might not be very informative—we have assays of every phosphorylation site. 1173 is the most dominant site in HMEC. The number of heterodimers species is unknown.

The following are proteins nominated by the group to measure using EGFR phosphorylation. From these we will develop assays, get best antibodies from literature, etc.

**Top six candidates (ERK is a given)**

P85  
Raf1

B-Raf  
RasGTP  
SOS  
Src

**Others**

Shc	Jnk	EGF, then downregulates
AKT	Cbl	ERK1)
PLC $\delta$	DUSP6	MKP3
Grb2	DUSP3	PTEN
Grb7	MKP1 (totally	IRS2
STAT3	downstream, induced by	Gab1
P38 (within 5 minutes)		Gab2

The group will do enzyme assays on P32, fluorescent analogs, Western blot, and ELISAs. Though this may sound much like what the Alliance for Cell Signaling is doing, this scope is smaller and has a well-defined endpoint. It also restricts us. This is a prototype, hence it must be small in scope.

This is a foundation for us to start with, to learn how to work together. We don't need to set up central labs to do assays—this is unnecessary and very expensive.

**Joint Project**

Project number 1, signaling prototype project. Activation of 2 members of EGFR pathway, look at MAPK.

Question to answer is how to combine the data from the member labs. Once a central module is defined, others can be added on.

This project will be presented at the 2005 International Conference on Systems Biology in Boston October 19-22 (<http://csbi.mit.edu/icsb-2005>). Each lab (PNNL, RIKEN, Keck Graduate Institute, Thomas Jefferson) can do this. All labs should all use the same cell lines and same medium to have baseline for comparison. First step is to define the assays: which can currently be done and which will take some development work.

Goal: To show we can work together across a distance and get data that can be modeled. After establishing that, we can then move to a favorite cell line and answer specific questions.

We will need time series data on enzyme activity and a QA/QC protocol. Standardize methods for enzymatic assay are needed. We will develop a spreadsheet on a shared webpage in which data can be entered. Eric Stephan will check into creating a SharePoint site that all members can sign onto.

We will share PNNL proteomics, microarrays (cell atlas). These are good tools to see what's there—an encyclopedia with basic information about a project. Conditions, biological knowledge already known, information we find from these assays.

If you can use these data to try out different models, that would be good. Among the many questions to answer: Is there any feedback, and if so, how strong? Sensitivity? Negative feedback?

**Perturbation.** One possibility is to change EGF and also do it in cell type with HER2 and HER4 expressing. Or EGF and HRG. This complements what Mariko was showing. Low EGF, high EGF.

One idea was to talk to a company such as AstroZeneca or Genentech and allow them to use our data, if they give us small molecules and antibodies. We must do perturbations on all the doses (8): high EGF, low EGF, etc.

- Use PI3K INHIBITORS (LY294002 wortmannin)
- Use RAS (Farnesyl transferase) inhibitors
- Generate 5 time points
- Go to the literature; find best assays for each of these.

If you get these data, it will be enough for modelers so they can come up with something that is not trivial. Time series data. Break it down to absolute concentrations. Fractionate it later.

### **Timeline and Deliverables**

- Initial report on workshop posted on RTK website by **August 1**.
- Eric Stephan and Steve Wiley will set up a SharePoint drive or similar mechanism for posting data, exchanging it, getting modelers involved by **August 1**.
- After evaluation of what assays wanted (**no later than September 1**), will send information to others. Will write up general protocols and strategies. List the antibodies, time courses, time series, prices.
- We will be looking at perturbations using small inhibitors. Karin Rodland will do some test experiments with chemical inhibitors to determine which have an effect on ERK phosphorylation and are worth pursuing as perturbations in the larger experiments.
- We will present our results at ICSB 2005 at the one-day RTK Consortium workshop.