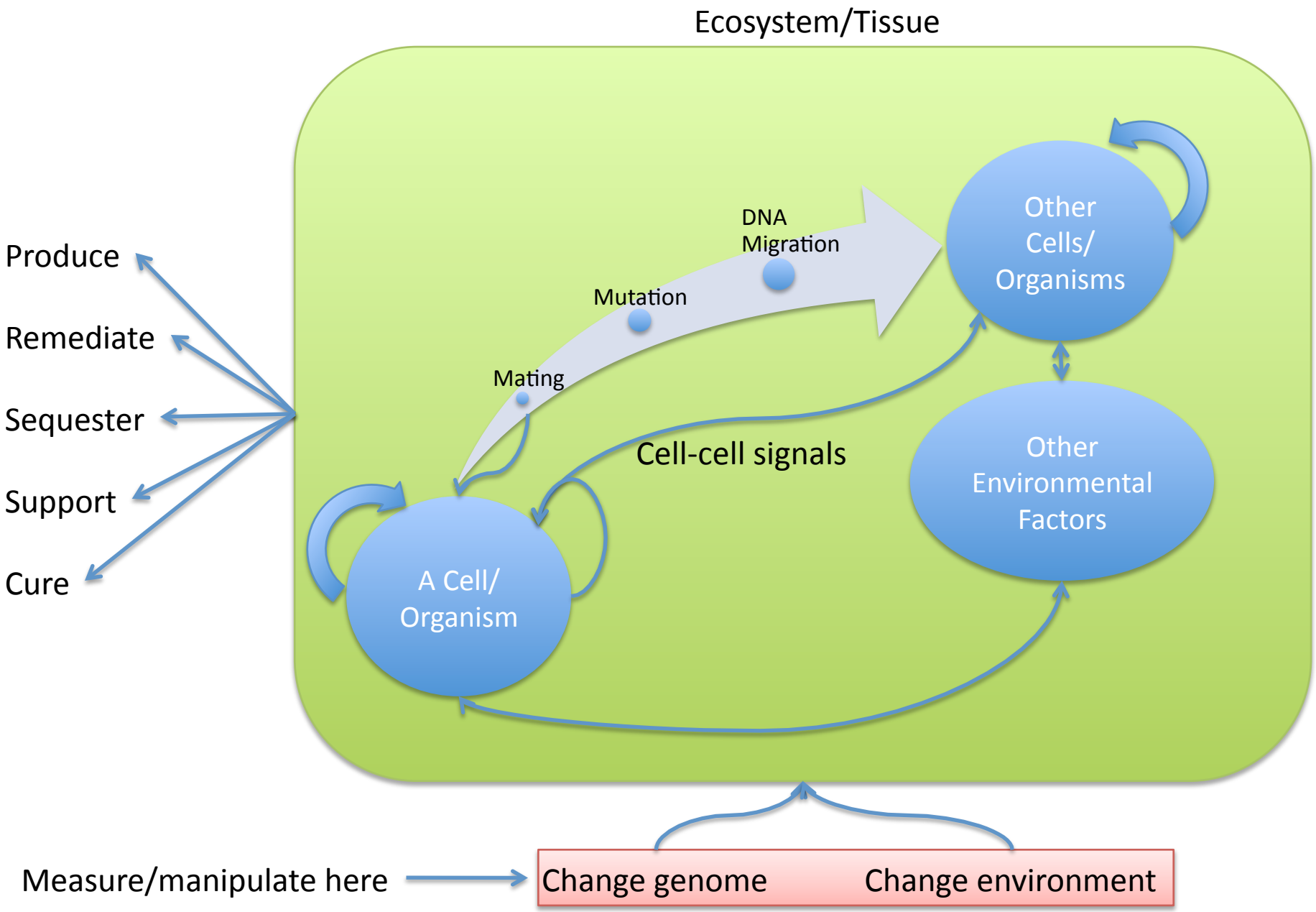
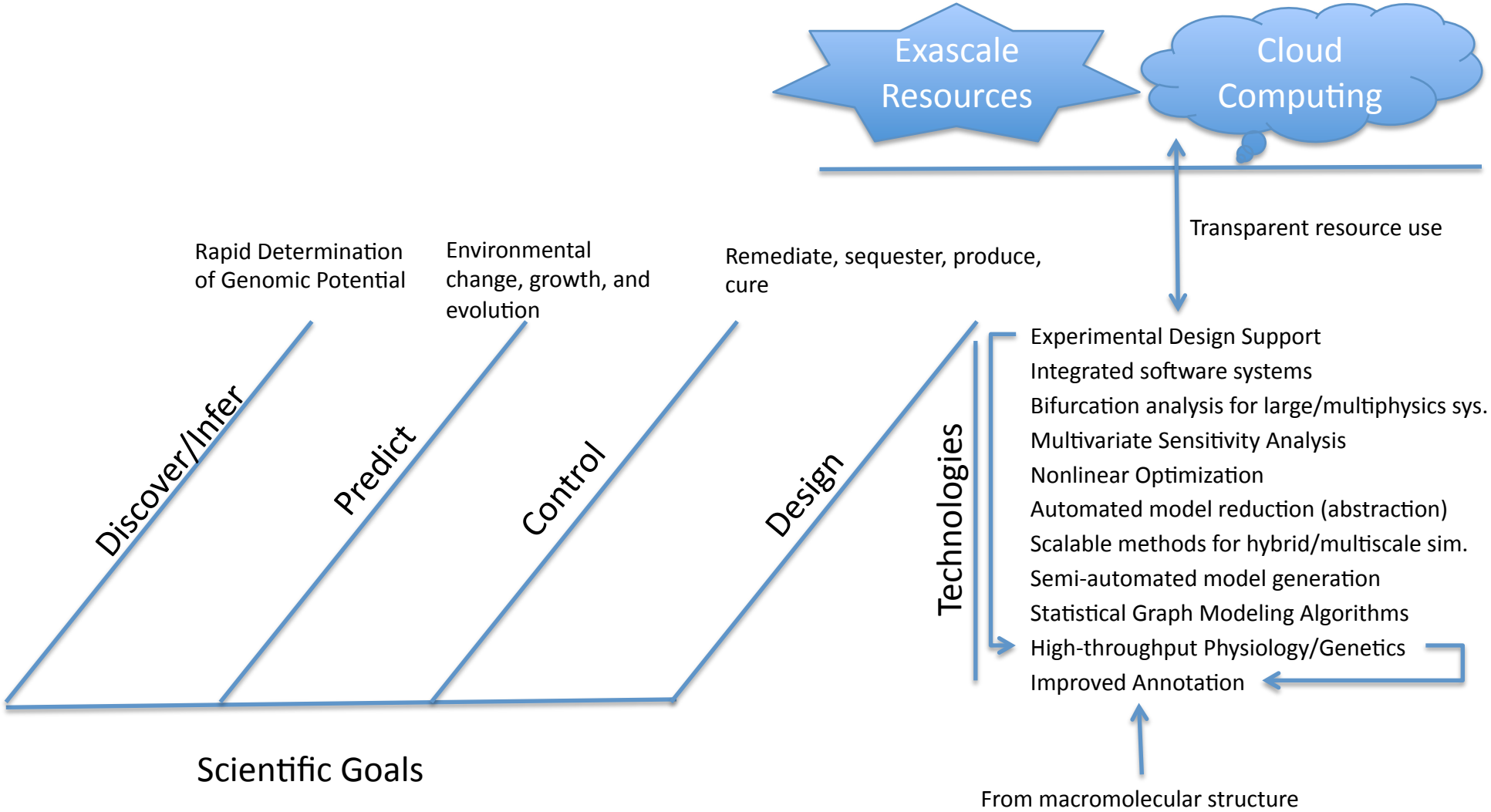

Ecosystem/Tissue
Produce
Remediate
Sequester
Support
Cure
DNA
Migration
Mutation
Mating
Other
Cells/
Organisms
Cell-cell signals
Other
Environmental
Factors
A Cell/
Organism
Measure/manipulate here
Change genome
Change environment

Goals and Outcomes



Grand Challenges

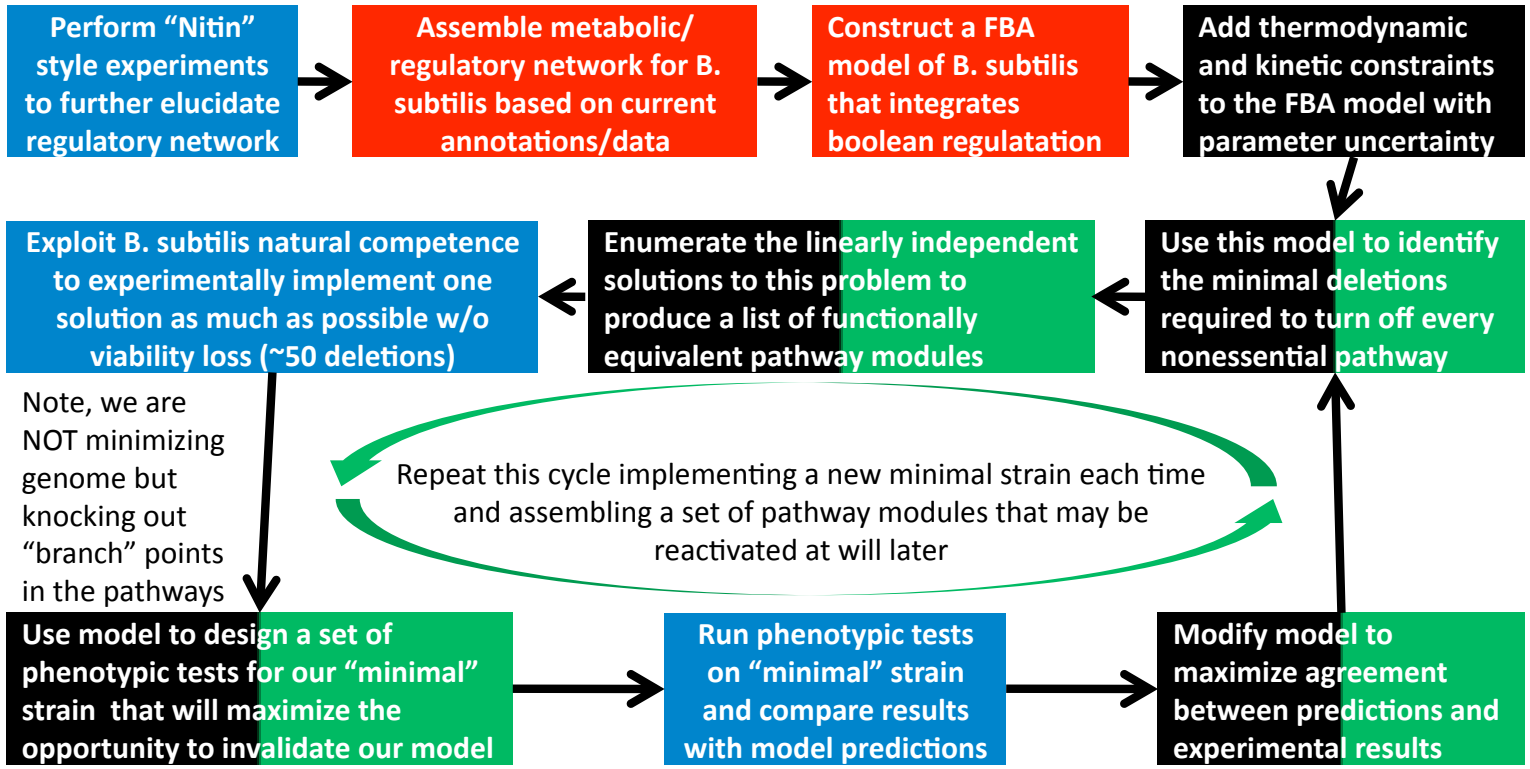
- Modeling Phenotype from Genotype Across the Tree of Life.
  - Rapid reconstruction of cellular networks across multiple time and space scales.
    - Predictive models of 10,000 organisms
    - Mechanistic understanding of single cell dynamics
  - Predict environmental "inputs" and outputs: "crack the signaling code"*
  - Predict optimal growth conditions from genotype
  - Optimize natural system activity of individuals and consortia
  - Individualized genomics

- Discover and Engineer Units of Function at all Network Scales
  - Discover the evolutionary principles of function generation and reuse
  - Design new function from composition of such modules

* Thanks, Mark

Example Stories: Chris Henry (and Costas Maranas)

Scientific objective: Develop a "complete" understanding of the metabolism of *B. subtilis* including identification of all transporters, the regulation of all metabolic pathways, and the pathway response to the environment.



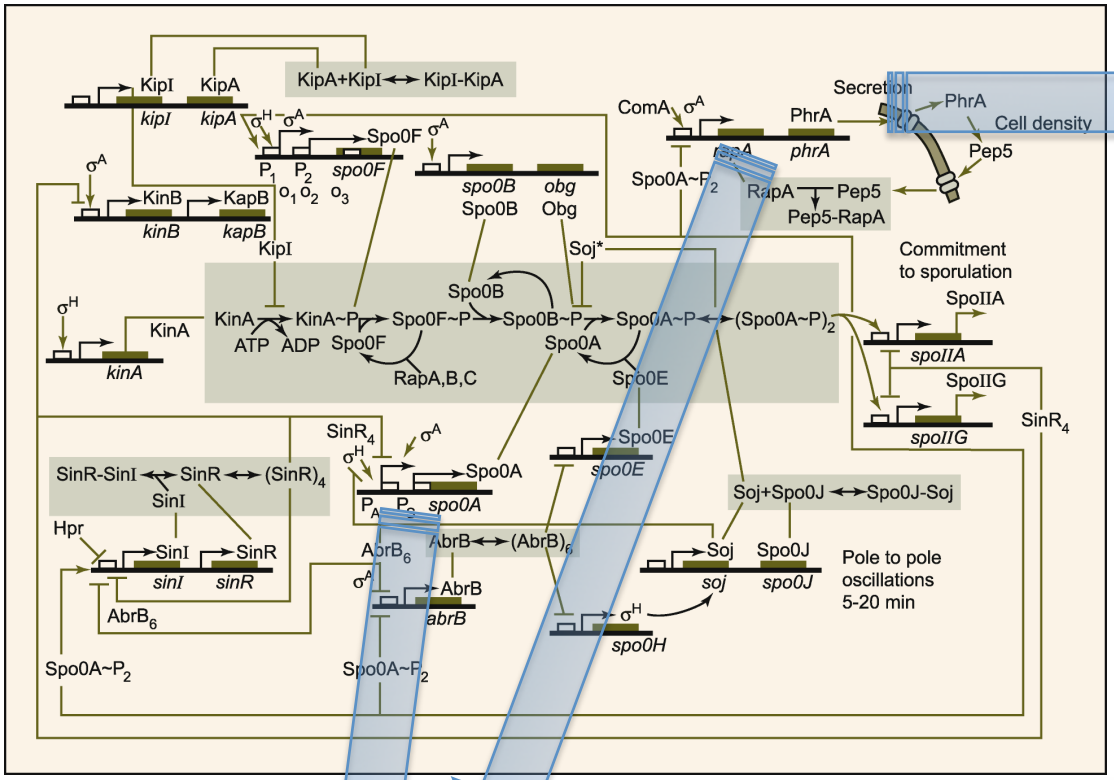
Desktop-scale computational steps in red

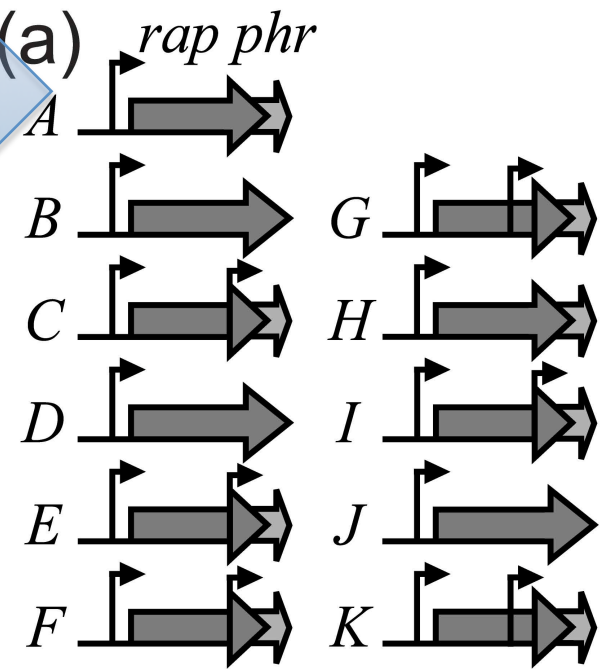
Massive scale computations in green

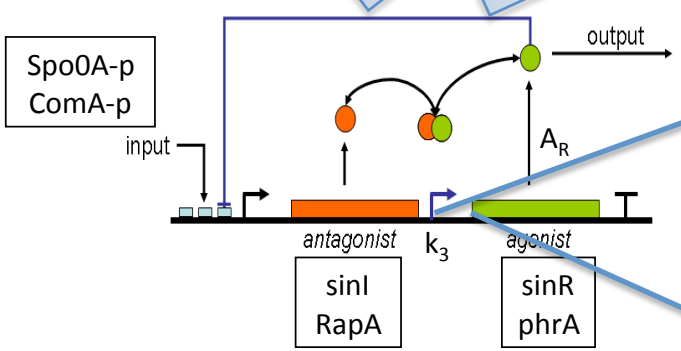
Experimental steps in blue

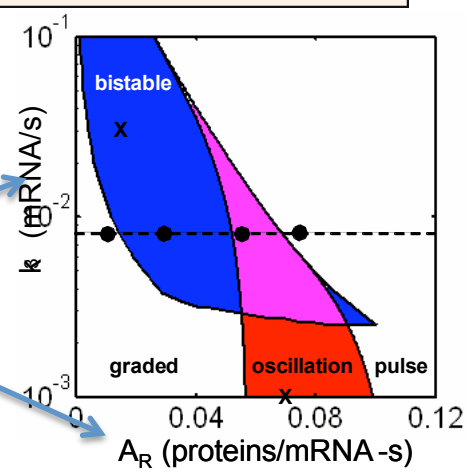
Steps requiring new algorithms in black

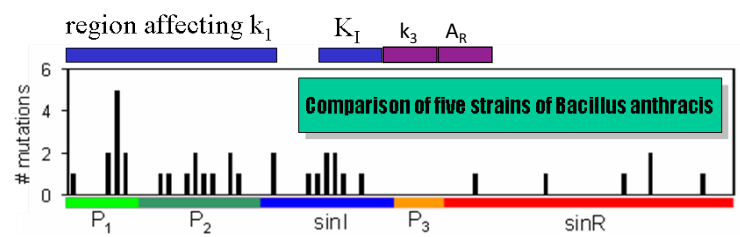
What switches on spores?







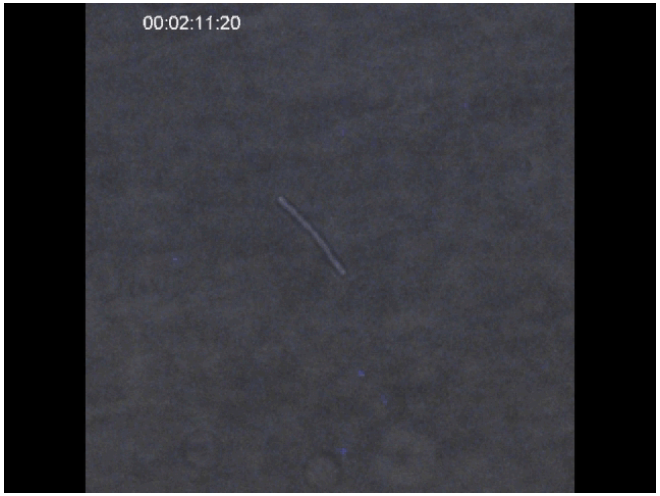




In *anthracis:*
Mutations mostly affect K_I and k_1
Threshold of the switch is most affected.

Across ALL sporulators
Very variable.

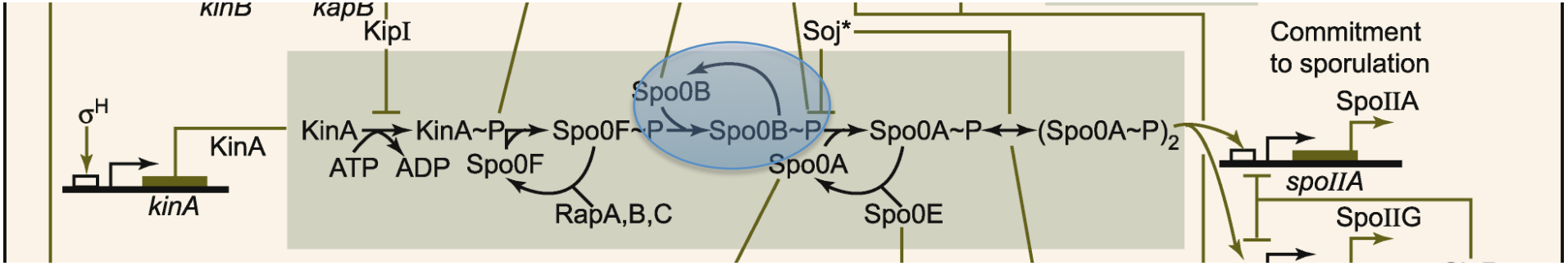
Induction of Spo0B



WT levels of Spo0B

Slight induction

Slightly larger induction



Impact

- We have an opportunity to truly understand a whole life form and its interaction with its environment.
- We can place it in evolutionary context and begin to understand how it and its parts arose from originally inorganic components?
- We can discover the motifs of function that we can exploit for human purposes

Overall Conclusions

- Data is still the main bottle neck but there are badly scaling computational problems in their analysis

- Behavior is responsive to complex combinations of inputs and combinatorially affected by large numbers of genes thus exploration of models or testing against data remains difficult.

- For this effort to have a profound impact on biology we need *high throughput*, *easy use, and fast concise presentation of data, model, and comparison.*

- While there is fundamental theory and algorithms that need to be developed– and is *critical*– the transparent access to large computation resources for a *large* number of users is the key. Most tools will be embarrassingly parallel or close but will still require outrageous numbers of processors for each project. And there will be many users.

- Sadly– experiments seem to scale along side computation– the larger and more complex the model the more experiments are needed to test.
  - The scaling may not have near the same exponents but it is far from trivial in time and cost
  - The data is "fragile" and needs to be tightly quality controlled, instantly accessible, and well annotated
  - Computation should be an everyday adjutant to experiment– but the computational throughput is currently too slow and complicated to be as generally useful as a knockout or microarray.

Annotation

- Sequencing is scaling such that 10k's of genomes and even more metagenomic sequence will be sequenced/unit time
- New families are plateauing, but still 30-50% unannotated function. Need structure prediction help. Experiments are key.
  - Investments
    - New algorithms for phylogenetic annotation N^2Log(N)
      - Deep theory of evolution for trees.
    - New algorithms for structural annotation (See macromolecular group)
    - New algorithms for "guilt-by-association" (M*N^2)
    - Experiments to broaden functional assignments (NOrganism*nCond*Nreplic)
    - Manual annotation/curation

High-throughput Physiology/Genetics

- Problems of scale
  - From single molecule to populations of cells
  - Multiplexes machines allow some whole genome assays– but biochemistry is still a problem, imaging is still a problem (~N^3) and tech like mass spec itself computational problems in spectra matching (~N^2)
  - Genetics is still case-by-case
  - Quality and reproducibility
  - Centralization/dissemination of data and standards
- Investments
  - Determining methods for quality control with data collection with technologies high distributed in individual laboratories (rather than a central resource).
  - Investment in creating higher throughput/high content resources for biochemistry, single cell and single molecule (and in situ) imaging/measurement
  - Investments in new culturing technologies in microfluidics to bioreactors
  - Massively parallel computation MUST be matched by massively parallel experiment.
  - Increasing computation in the loop design of experiments
    - Mathematics and CS for Design of experiments
  - Transparent integration of supercomputing and the HT experiment labs.
    - Embedded systems engineering for placing controls on the instruments
  - Visualization of complex data sets derived from these.
  - Mathematics for reduction of nonlinear multivariate datasets.
  - Image processing/segmentation-
  - Capture reagents (to target molecules for measurement)
  - Experiments that seek to discover the communication mechanisms and molecules between cells.

Statistical Graph Modeling Algorithms

- Theory for determining best array of experiments to input into the model
  - Which modalities
  - Which conditions and their combinations
  - Which timescales
  - How many replicates
- Currently– bicluster goes as N^2/N^3 and propagation and uncertain and branch/bound model reduction goes as Exp(cNedges) where c small constant.
- Investments
  - New theory for data integration and statistical modeling
  - Validated test sets for algorithm comparison
  - Mechanisms for data availability
  - Detection weak signals in high background noise (not subgraph isomorphism)– sequences, correlations, quantifications
  - Network motif detection
  - Automated text mining/natural language processing
  - Improve the feedback loop between statistical associations and annotation
    - Better use of molecular interaction data and sequence/TF information

Semi-automated model generation

- Data Driven to Dynamic Models
  - How do we use bioinformatic analysis of new data sets to build a model?
  - How do we force “annotation” of more biochemical features that would aid modelers.
  - Metabolic reconstruction through annotation
    - Validation
    - Feedback through inconsistency and holes
  - Rapid update of models when data updates
    - Link to papers
- A taxonomy of model types that would allow you traverse what kind of model to generate from data.
- Establishment of standards for different types of model
  - What is a “complete” model?

- How do you integrate models of different parts of a larger system
  - Tools for making this easier.
  - Controlled vocab, semantics and agreement with the bioinformatics folks with naming
- We’d like to– we have no clue?
- Maybe accessibility to the Statistical models

scalable methods for hybrid/multiscale sim.

- To understand how the 1D genome is transformed into 3D space.
- Current PDE simulations scale as O(N) but…
- Need for usable library and modeling frameworks that biologists can use easily.
- Do this with high heterogeneity in the physics (stochastic/deterministic) and space.
- Do this under uncertainty in data, parameters and (gasp) mechanism
- Languages for model representations: Formal languages? (Languages for simulation control).

- Investments
  - How do smoothly transition among levels of physical abstraction from fully rendered spatial/molecular level simulation, through mesoscale stochastics all way to smoothed ODE.
  - How do you compare spatial/temporal data and models in phylogenetic context.
  - How do you put together genetic/genomic data and spatial modeling.
  - How do you build complex heterogeneous models?
  - How do you build hybrid stochastic/deterministic models/statistical
  - New integrators (like Magnus Expansions, spectral) to support accurate
  - Measurement technology of biochemical activity (and forces, etc.) in live cells.

Automated model reduction

- Formal coarse graining, scale separation, and "balanced truncation"
  - Ronald Coifman?
- Automated nondimensionalization?
- Linear algebraic solutions for the above.
  - Pseudospectral methods?
  - Regularization
- Response surface methods and functional approximation.

Nonlinear Optimization

- Parameter estimation and Model Testing
  - Determination of the error bounds on the model
  - Parameter feasibility regions
    - Changes in regions upon compositing of submodels
      - Theory of module impendance and retroactivity
  - Model Invalidation
    - Inconsistency
  - Data invalidation
  - Algorithms for global optimization by clever parameter motion
    - Quasirandom search
      - How do we get to the tails of the distributions of our "stochastic" models.
- Measures of parameter conservation between organisms as a result and proxy for evolutionary pressure.
- Model Selection/Model Modification– minimal number of moves to explain the maximum amount of data. (Exp(N) but practically N^2 to N^3).
- Multiobjective optimization
- Integer Optimization

Multivariate Sensitivity Analysis

- Difference methods fail for stochastic systems
  - New algorithm
  - Petzold & Doyle?
- High dimensional sensitivity
  - Sethna and sloppiness.
    - Eigen directions.
  - David Rand (Warwick)
  - FAST (Fourier something or something) Saltelli
  - Denise Kirchner
- Understanding where to do your experiments
- How do we link sensitivity to evolvability?

Bifurcation analysis for large/ multiphysics sys.

- Auto, Oscill8, MatLab (CONTENT)
  - Crashes around 25 differential equations
- How does bifurcation change with composed models?
  - High dimensional (co-dim 3 and higher)
- Classification of dynamics of a model
  - Rapid inference of possibility for bifurcation from models.
  - Chemical Reaction Network Theory (Feinberg)
  - Different physical model classes.
    - Fully discrete and stochastic methods
      - Do deterministic bifurcations survive realistic noise
      - Do new bifurcations arise due noise.
- Map parameters to genotype determine which genotypes lead to bifurcation.

Experimental Design Support

- Optimize choice of data to answer question based on model
  - E.g. best data to estimate parameters
  - E.g. best data to discriminate between models
  - E. g. best data to discover missing pieces of model
  - E.g. best data to reduce the uncertainty in model prediction.

  - Tidor
  - Arkin/Flaherty

- The automated "Adam"– the geneticist.
- Issues of resolution
  - Time resolution
  - Relative vs. absolute measurement
  - Measures of constraint?
  - What data do you need to deal with multiscale models
- Closed loop control of Biology
  - Think Herschel Rabitz

It all depends on what the meaning of “Mission” is.

Yes, we can!