

Representing Biological Knowledge to facilitate reasoning and drug discovery research

Atul Butte, MD, MS

Pediatric Endocrinologist and Bioinformatician,
Children's Hospital, and Harvard Medical School
Founding Scientist and Scientific Advisor,
Genstruct Inc.

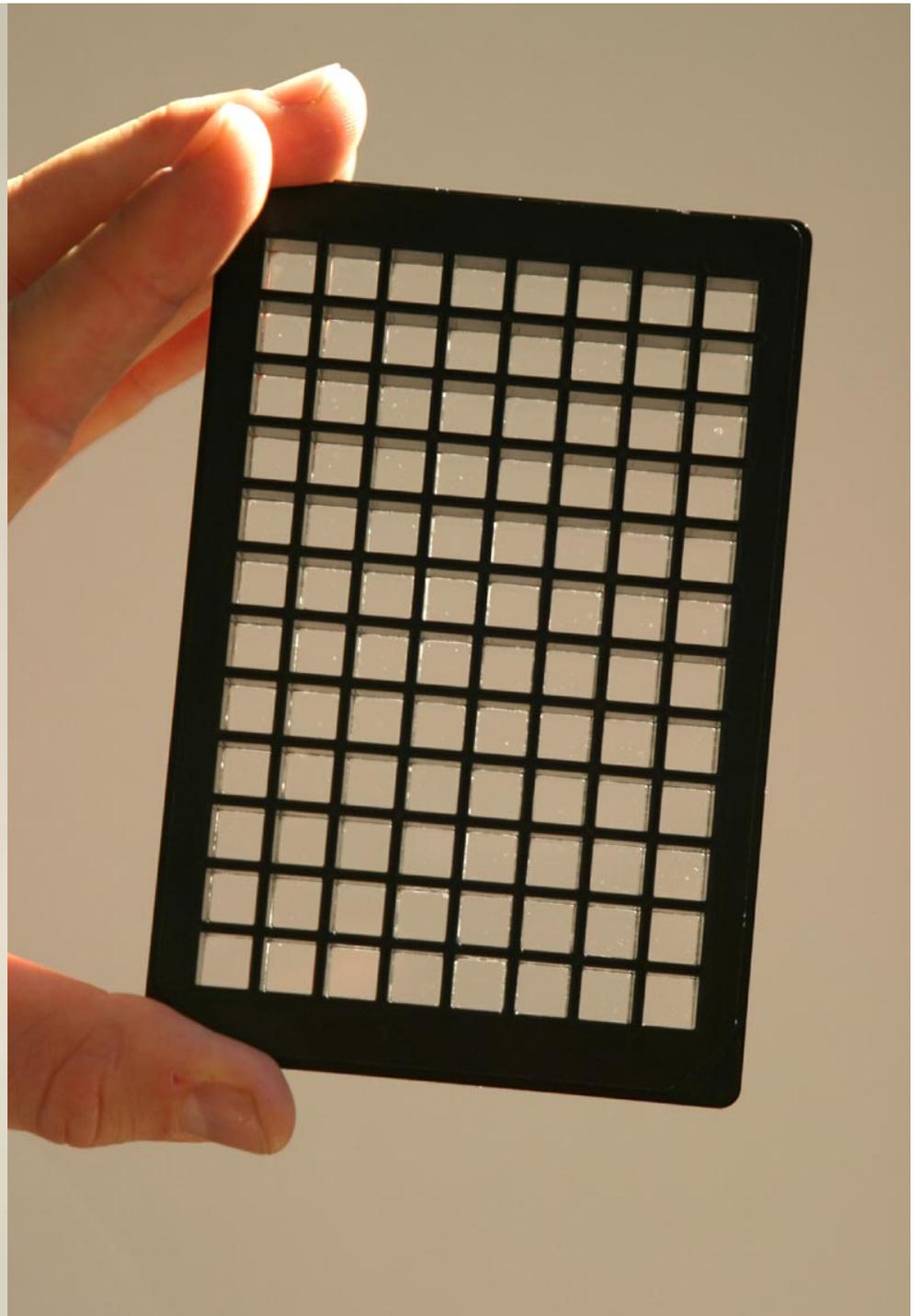
November 5, 2003

Bioinformatics at the Children's Hospital Informatics Program

- **Funded by 14 NIH grants across 7 institutes**
- NIDDK "Career Development in for Pediatric Endocrinologists" DK063696
- **NIDDK "Diabetes Genome Anatomy Project" DK60837: bioinformatics core**
- NIDDK "Gene Expression in Prediabetes": bioinformatics support
- NIDDK "Biotechnology Center" processing 50 microarrays/week U24 DK058739: bioinformatics core
- NIDDK "Surrogate Markers for Early Stage Diabetic Retinopathy": bioinformatics support
- NHGRI PhD Training Program "Bioinformatics and Integrative Genomics"
- NLM funded "Biomedical applications of the Next Generation Internet" N01 LM093536
- NCI funded "Improved diagnosis in ALL" R21 CA95618: bioinformatics support
- **NHLBI Program in Genomic Applications "Genomics of Cardiovascular Development, Adaptation, and Remodeling" U01 HL066582: bioinformatics core**
- **NHLBI Program in Genomic Applications "Innate Immunity In Heart, Lung, And Blood Disease" U01HL066800: bioinformatics core**
- NHLBI funded "AT2 receptor-mediated gene programming of smooth muscle" R01 HL58516: bioinformatics support
- NINDS program project grant "Gene expression in normal and diseased muscle during development" P01 NS040828: bioinformatics core
- NINDS funded "Functional Genomic Analysis of the Developing Cerebellum" R21 NS041764: bioinformatics core
- NIAID funded "Novel Approaches to Achieve Allograft Tolerance" R01 AI050987: bioinformatics core

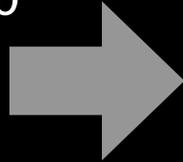
Genstruct Inc.

- Genstruct is a biotechnology company pioneering the development of new therapies through integrative biology
 - independently and through pharmaceutical partnerships
- Genstruct builds and exploits large conceptual models of biology to break cognitive barriers in discovery research and medicine
- Genstruct's models are generating and testing hypotheses for:
 - Disease Pathophysiology
 - Compound Mechanisms of Action
 - Predictions for Animal Models
 - Target Prioritization

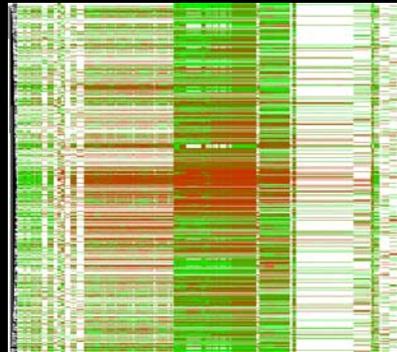


Microarrays
3-10 MB

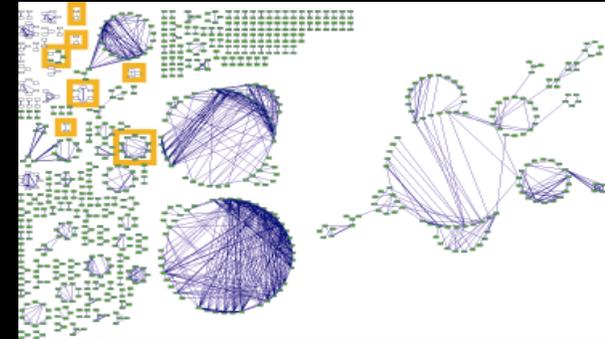
At least 8000
publicly
available
microarrays



Microarray Analyses



Bioinformatics



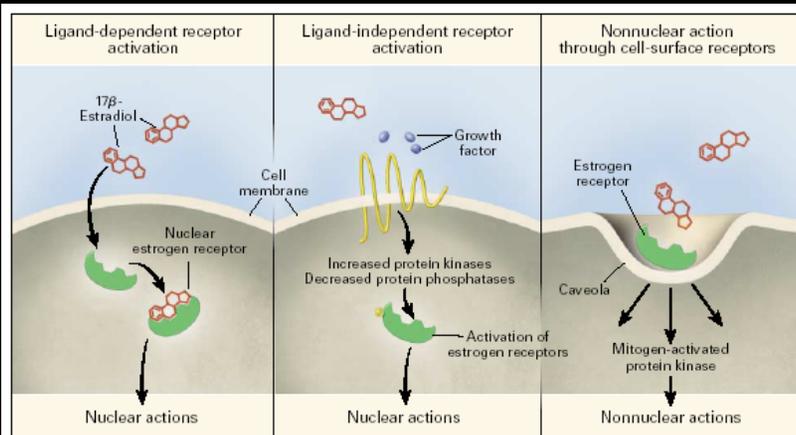
Wafers
5-10 TB

10K and 100K
SNP chips

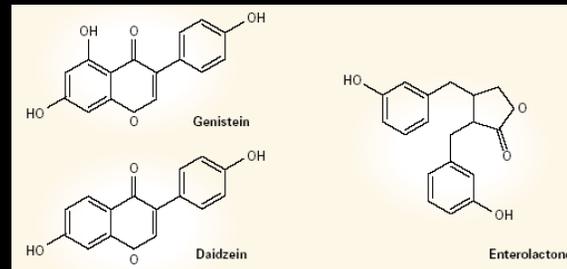


The Next Impediment to Discovery

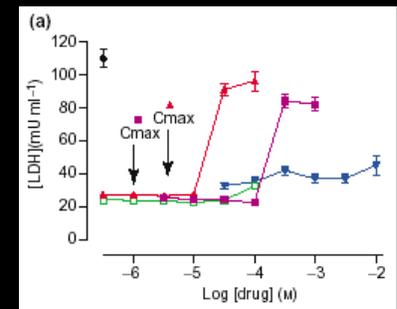
The Causative Target



The Best Agent



The Least Toxic



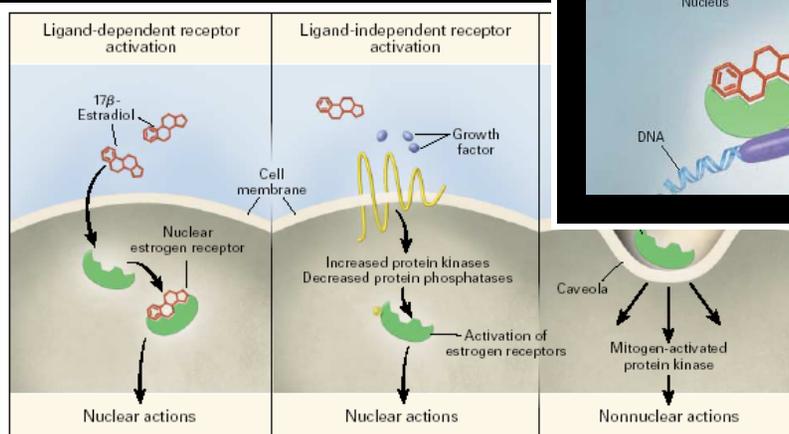
Microarrays
3-10 MB

At least 8000
publicly
available
microarrays

Wafers
5-10 TB

10K and 100K
SNP chips

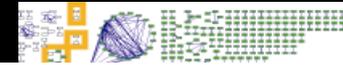
The Causative Target



Microarray Analyses

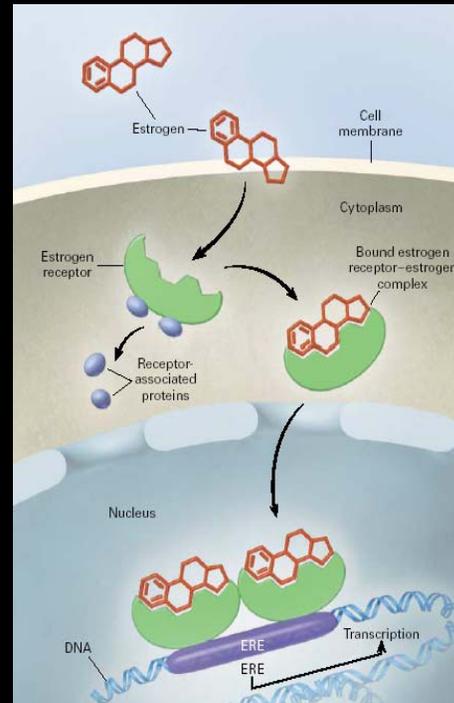


Bioinformatics

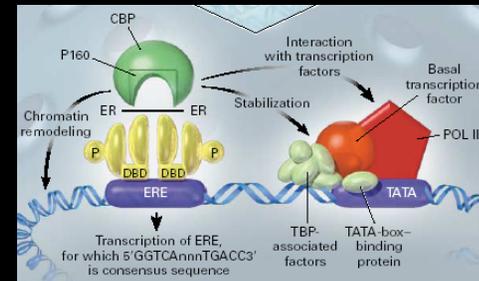


Knowledge Assembly

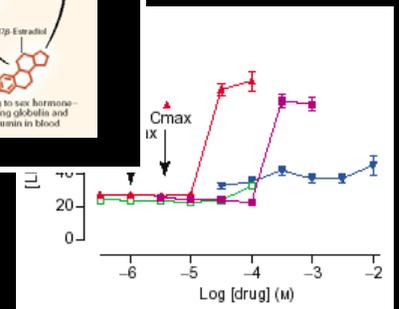
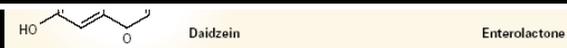
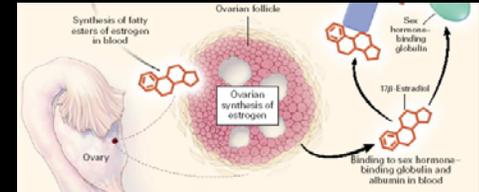
Known Signal Transduction



Known Transcriptional Control

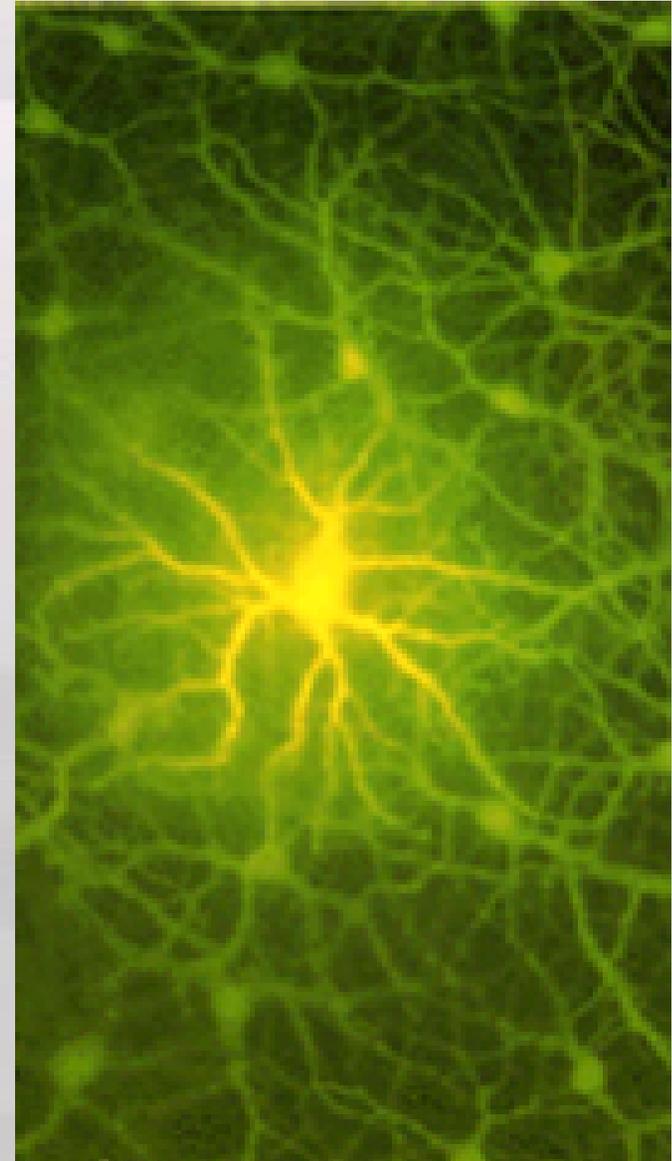


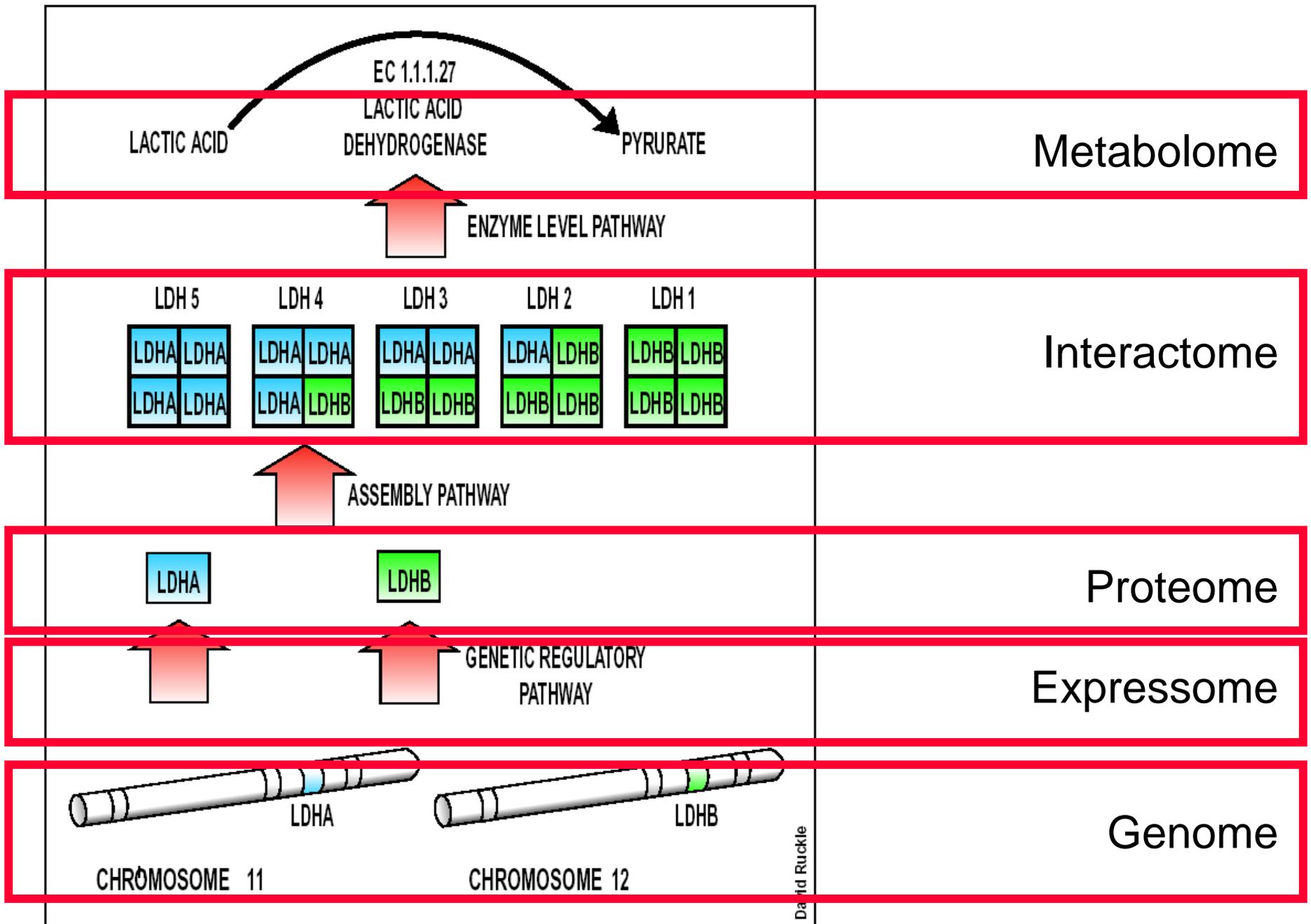
Known Pathophysiology



Background

- Organisms are networks of Cells
- Cells are networks of proteins
 - Which regulate genes
 - Which make networks of proteins
 - And so on and so forth
- We currently understand very little about the functioning of the simplest regulatory networks
- A conceptual framework for dealing with large biological networks does not exist



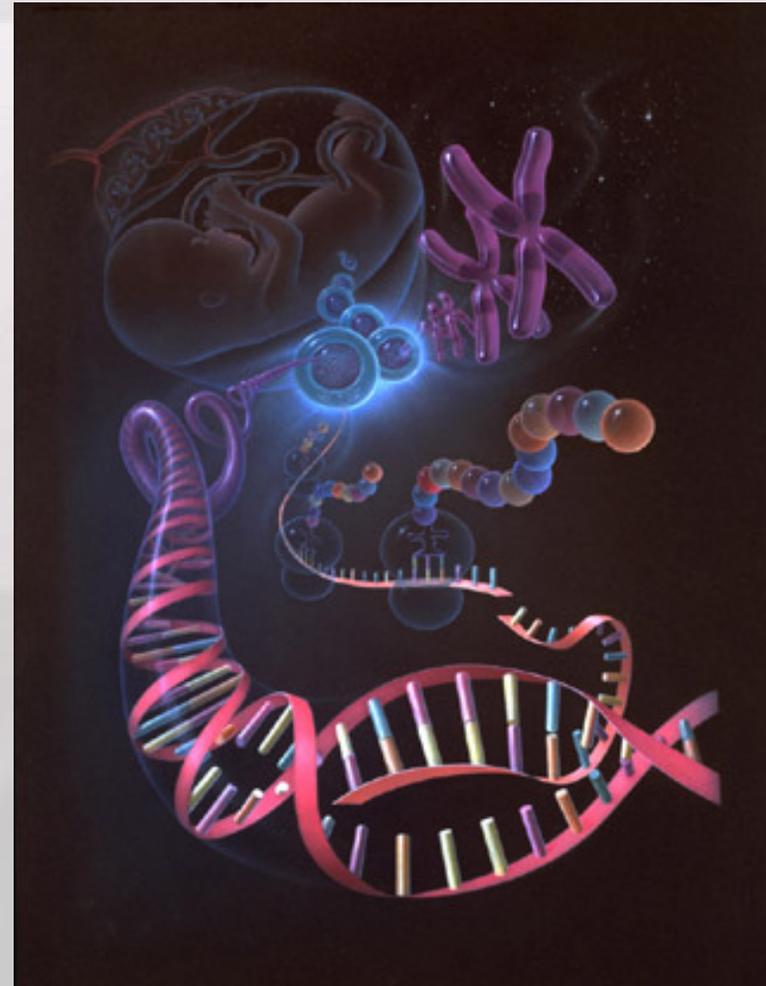


The Vocabulary

- Microarray testing: 500K identifiers
- LocusLink: 200k genes, 7.5 pieces of information on each
- dbSNP: 4m human, 200k others
- GeneOntology: 15k terms
- Laboratory tests: 20k
- Taxonomy: 300k species
- Prosite: 8k domains
- DNA: 200k exons human, 200k mouse
- At least 6 million objects/concepts in biology today
- Interconnections between them: n^2

The Challenge

- New technologies enable the gathering of enormous amounts of information about biology
- The need is to develop an understanding of how biological systems organize and integrate complex processes



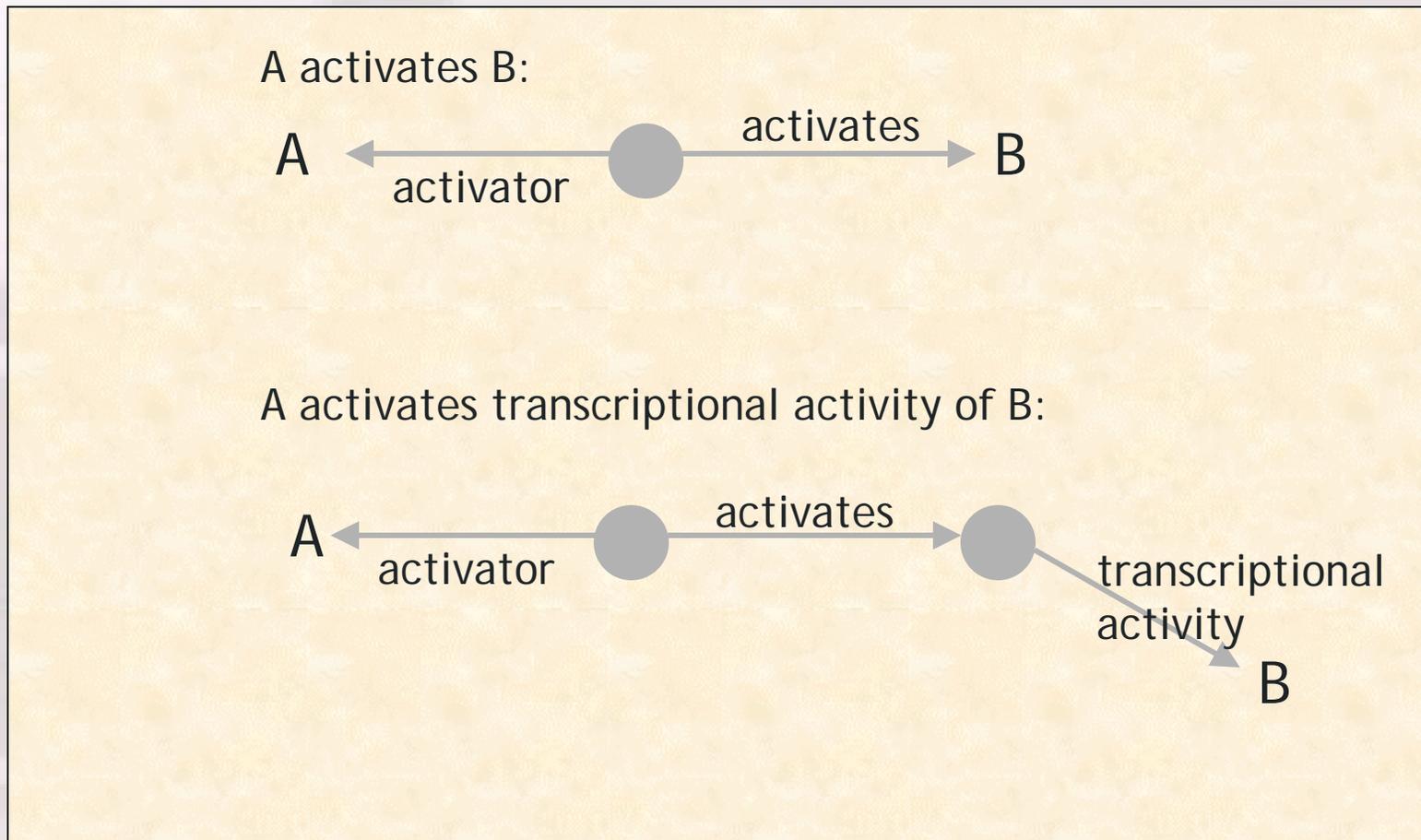
The Genstruct Vision

- Explore biology with an *integrative* rather than *reductionist* approach
 - Apply this new knowledge to the development of groundbreaking new therapies
- Achieve this through logical *reasoning* on the implications of *qualitative* biological networks
 - Networks of biochemical processes regulate the state of the cell and its response to stimuli and environment
- Goal: understand how biological networks function and use these findings to benefit human health

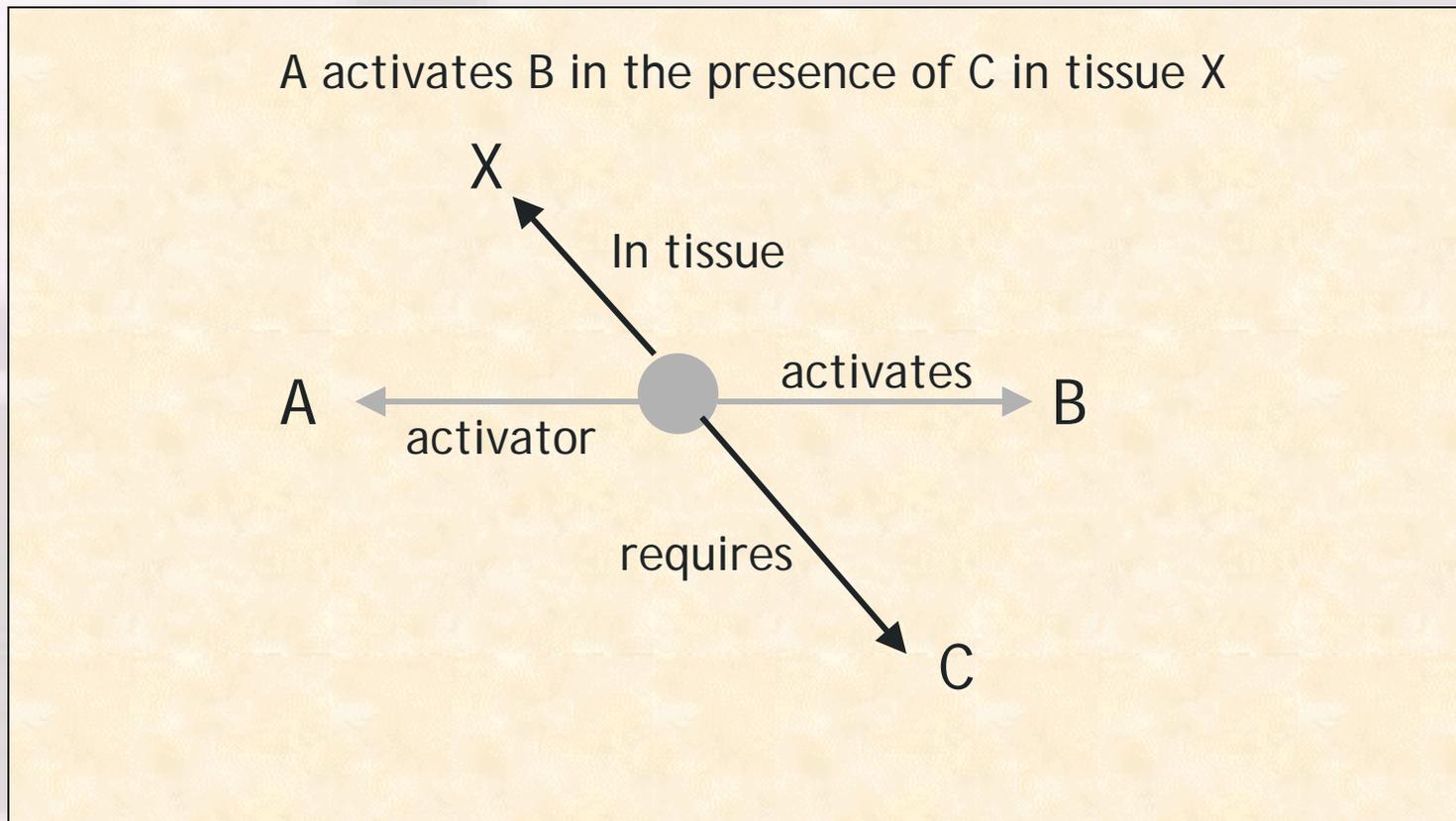
Our Approach

- Assemble models that integrate fundamental knowledge about large biological networks
 - The networks that underlie all biological systems
- Interrogate these models using computable logic to understand and predict biological responses
 - What was the response to drug?
 - How can we obtain an optimal response?
- Genstruct developed a framework for assembling and interrogating large biological networks
 - Defining key steps in biological pathways associated with disease states
 - Integrating complex genetic and environmental interactions leading to disease

Biological Case Frames: a framework for capturing knowledge in biology



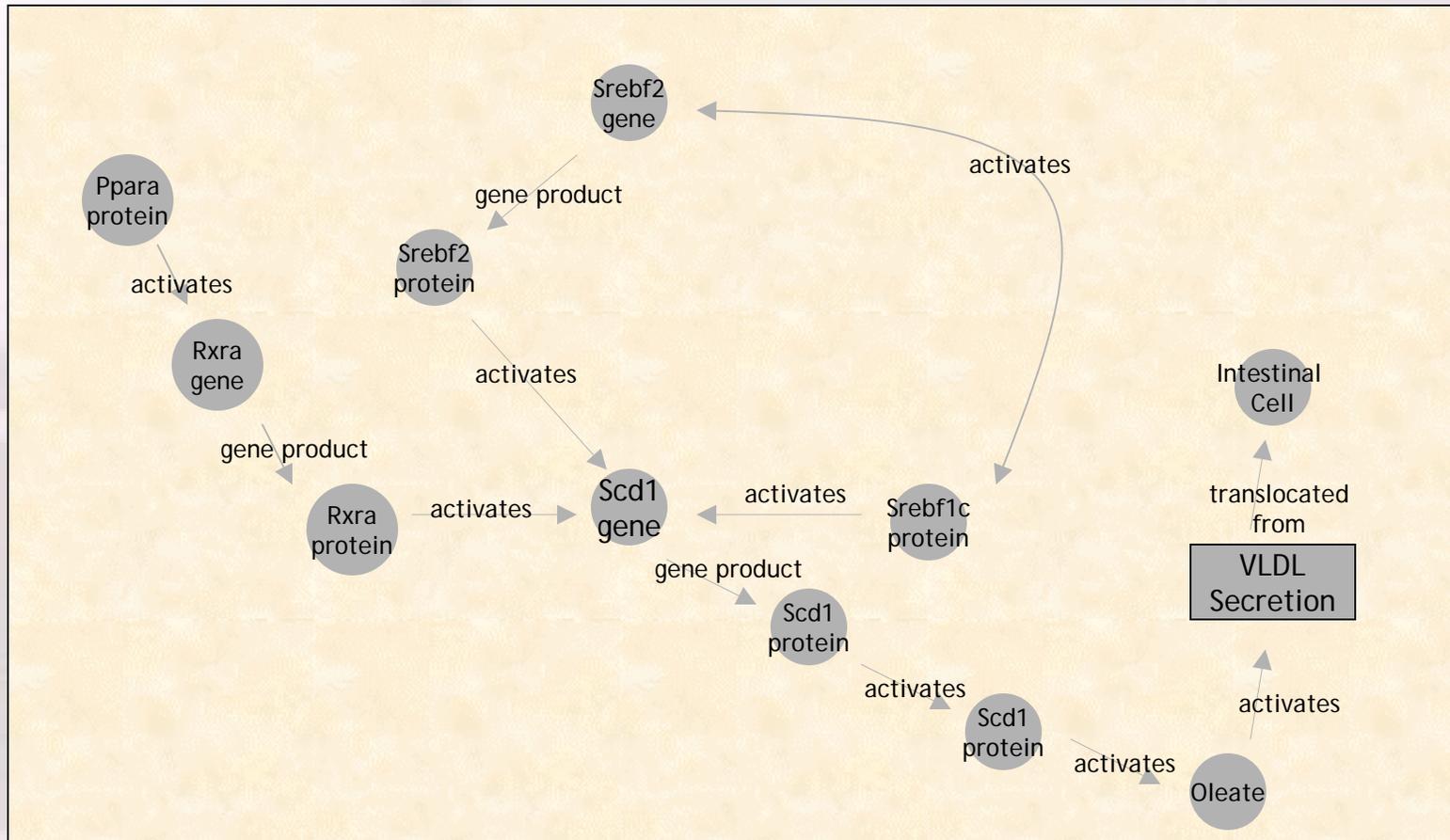
Biological Case Frames capture complex biological functions



Properties of biological case frames

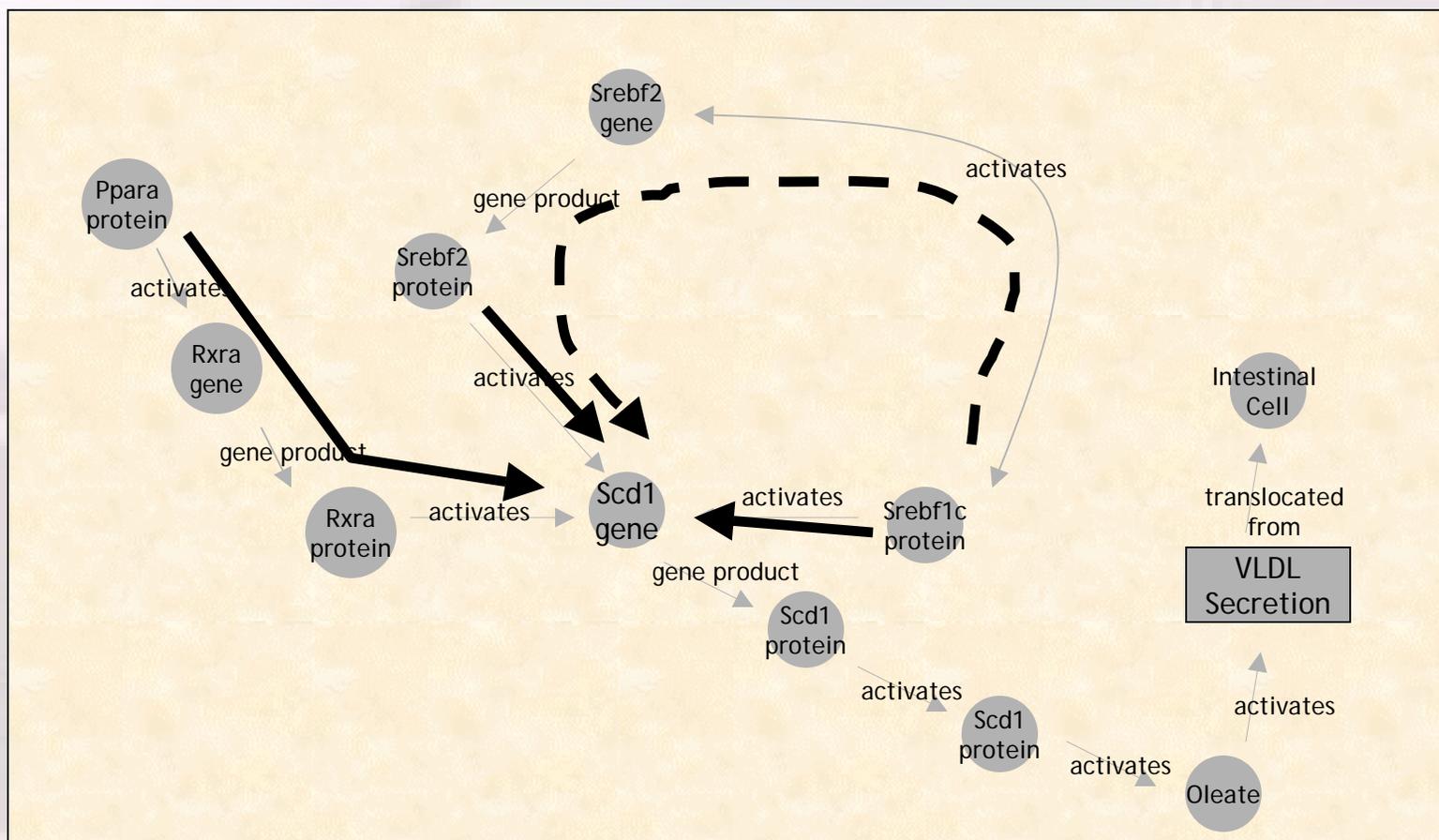
- Complete: Each case frame represents an *encapsulated* concept in biology.
- Context Free: Form and meaning do not change across biological contexts
 - species / tissue / compartment / stage
- Globally Consistent: Any case frame can be placed in any context.
- Scalable: Case frames can be connected to form graphs of indefinite size and complexity.
- Conflicting knowledge can co-exist.

Knowledge Assembly - Building Conceptual Models of Biology

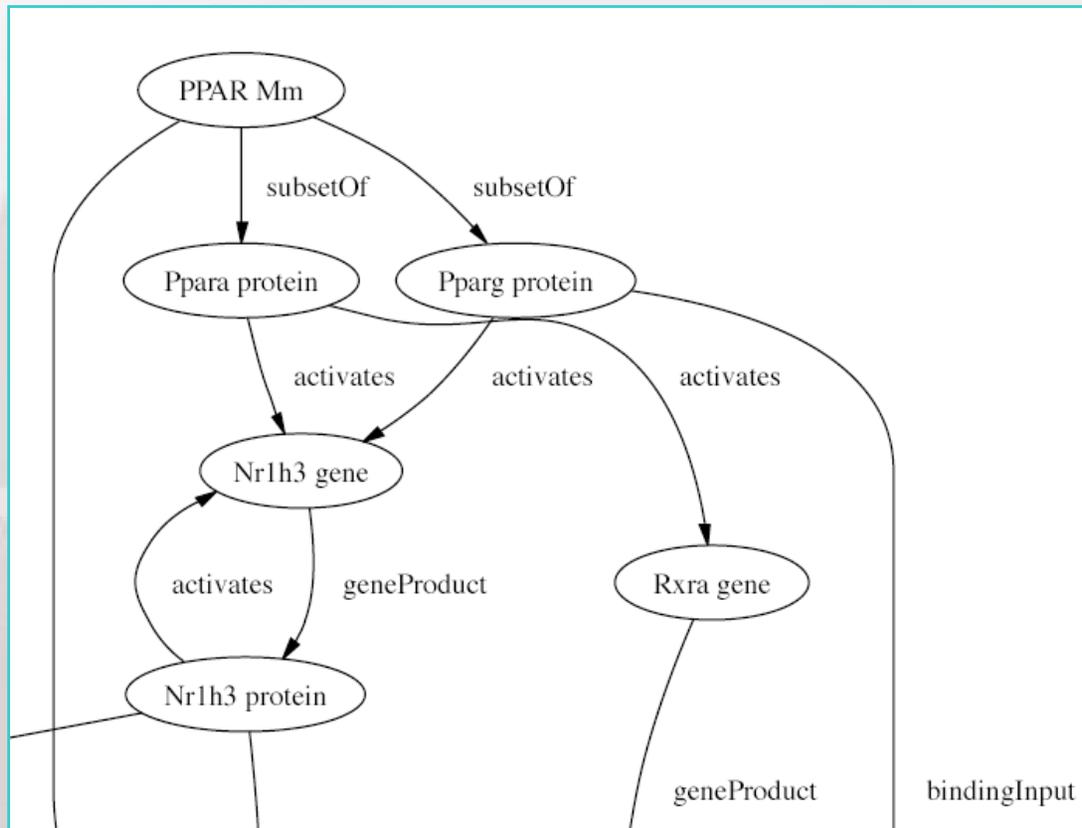


KA's capture and represent knowledge in computable networks

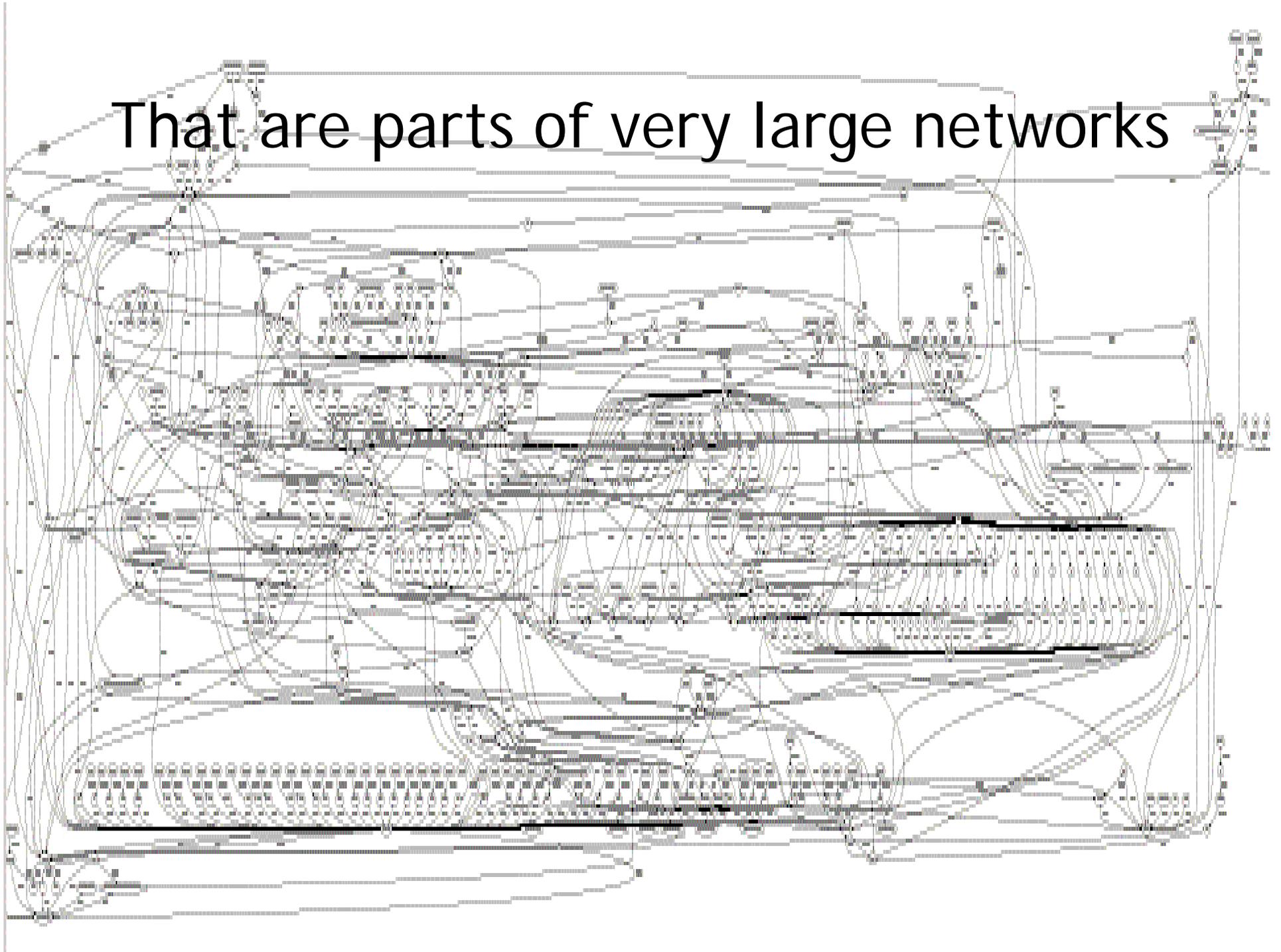
Reasoning through Biological Case Frames



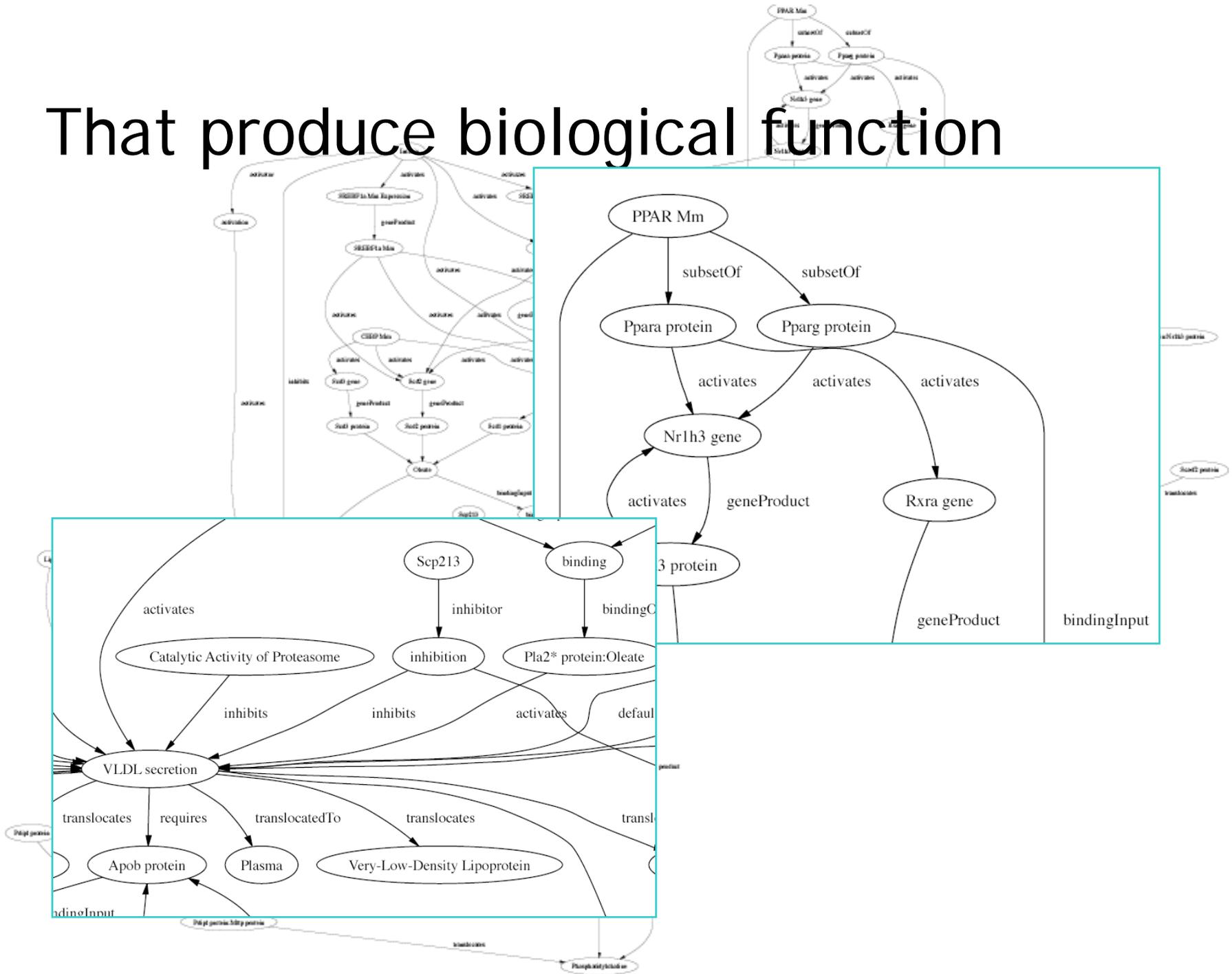
Small Networks...



That are parts of very large networks

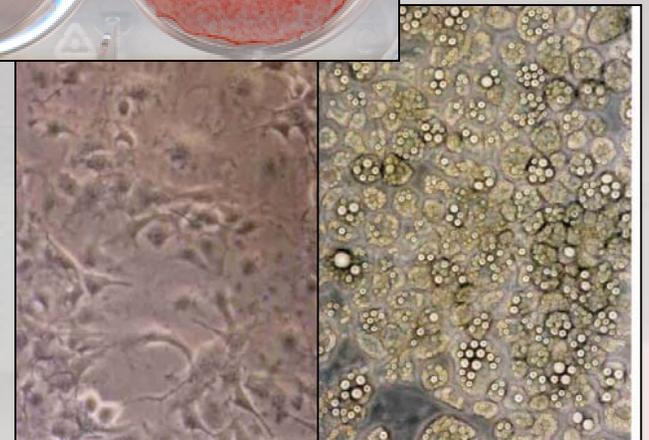
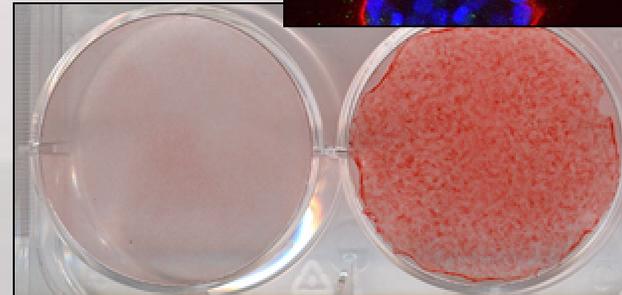
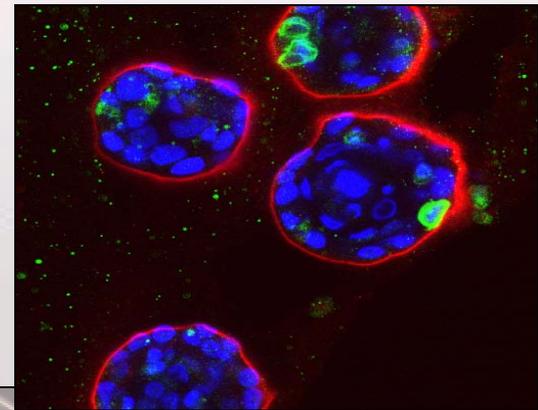


That produce biological function



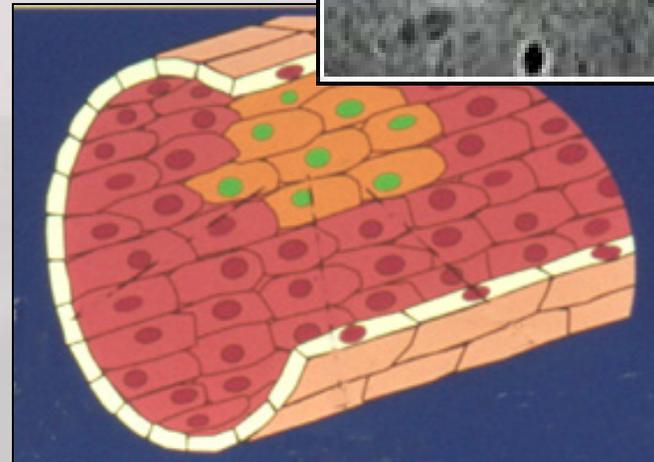
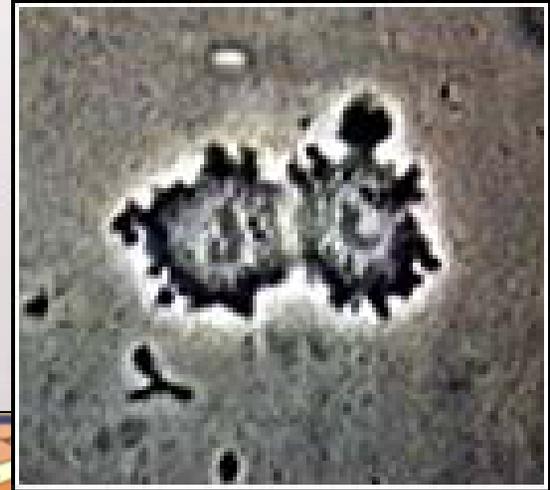
Genstruct's Drug Discovery Programs

- **Prostate Cancer**
Identification and validation of novel control points (targets) for androgen-resistant prostate carcinoma
- **Diabetes**
Identification and validation of novel control points (targets) for type 2 diabetes



Drug Development Partnerships with Pharma

- **Solid Tumors:**
Identified mechanisms of resistance and putative biomarkers for pre-clinical oncology compound (1Q03)
- **Cardiovascular:**
Identified compound MoA for phase II dyslipidemia compound (3Q03)



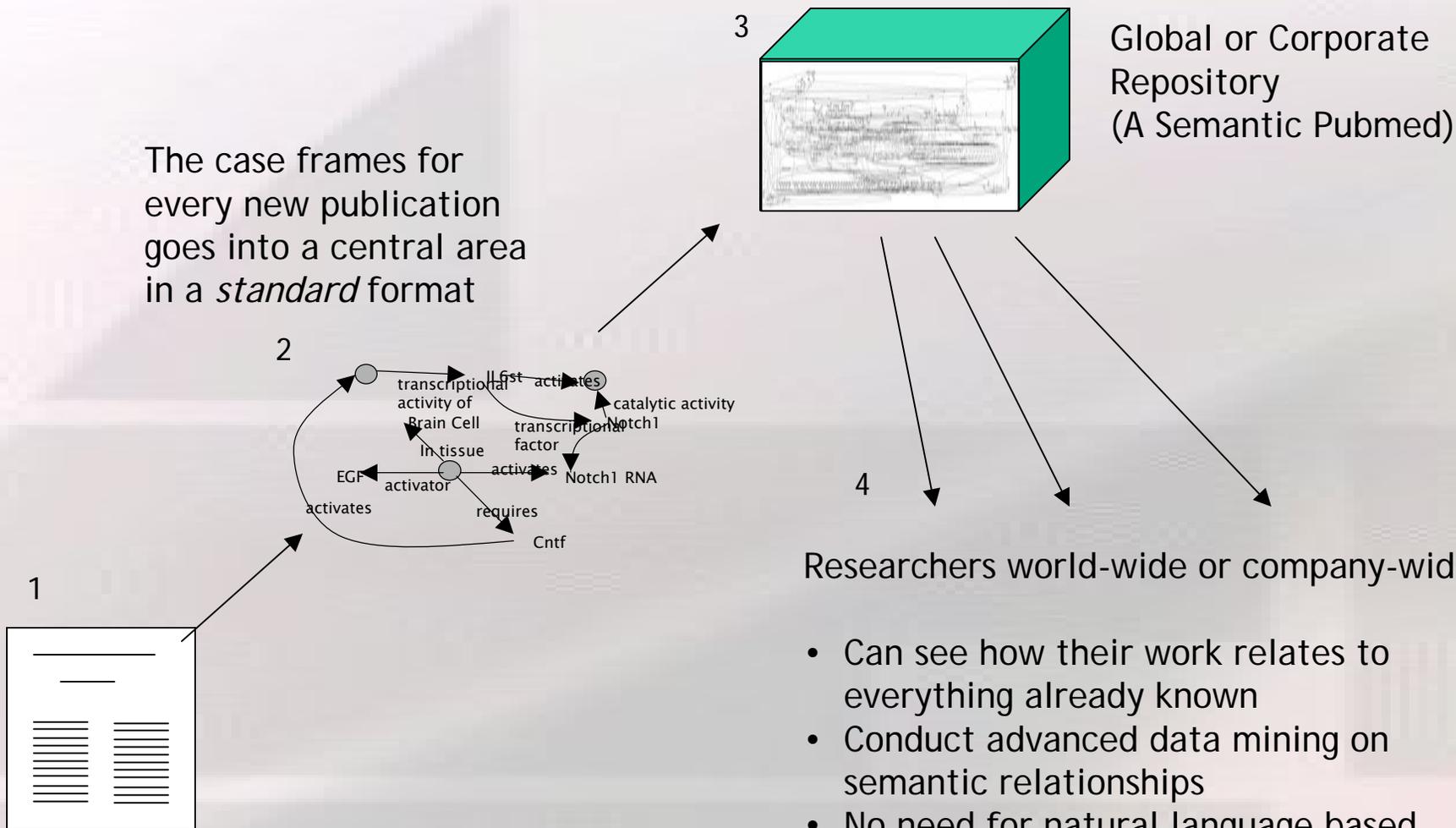
*Both with top 5 pharma

A look towards standards

- The case frame representation allows us to represent any concept in biology.
- Case frames can be connected to one another to construct complex *stories* (e.g disease process).
- The semantic essence of any paper or abstract can be represented as a collection of case frame statements.
- What if every publication came with an abstract and a set of case frames...

Looking to the future

The case frames for every new publication goes into a central area in a *standard* format



Researchers world-wide or company-wide:

- Can see how their work relates to everything already known
- Conduct advanced data mining on semantic relationships
- No need for natural language based literature mining

Questions?

Please contact Genstruct at 617.547.5421 or:

Atul Butte, Scientific Advisor
Keith Elliston, CEO

atul_butte@harvard.edu
kelliston@genstruct.com