Informatics in Drug Discovery

Workshop on a “Drug Discovery” Approach to Breakthroughs in Batteries

September 8-9, Cambridge

Ernst R. Dow, Ph.D.

dow@lilly.com

Group Leader / Senior Information Consultant, Eli Lilly and Company

Lilly

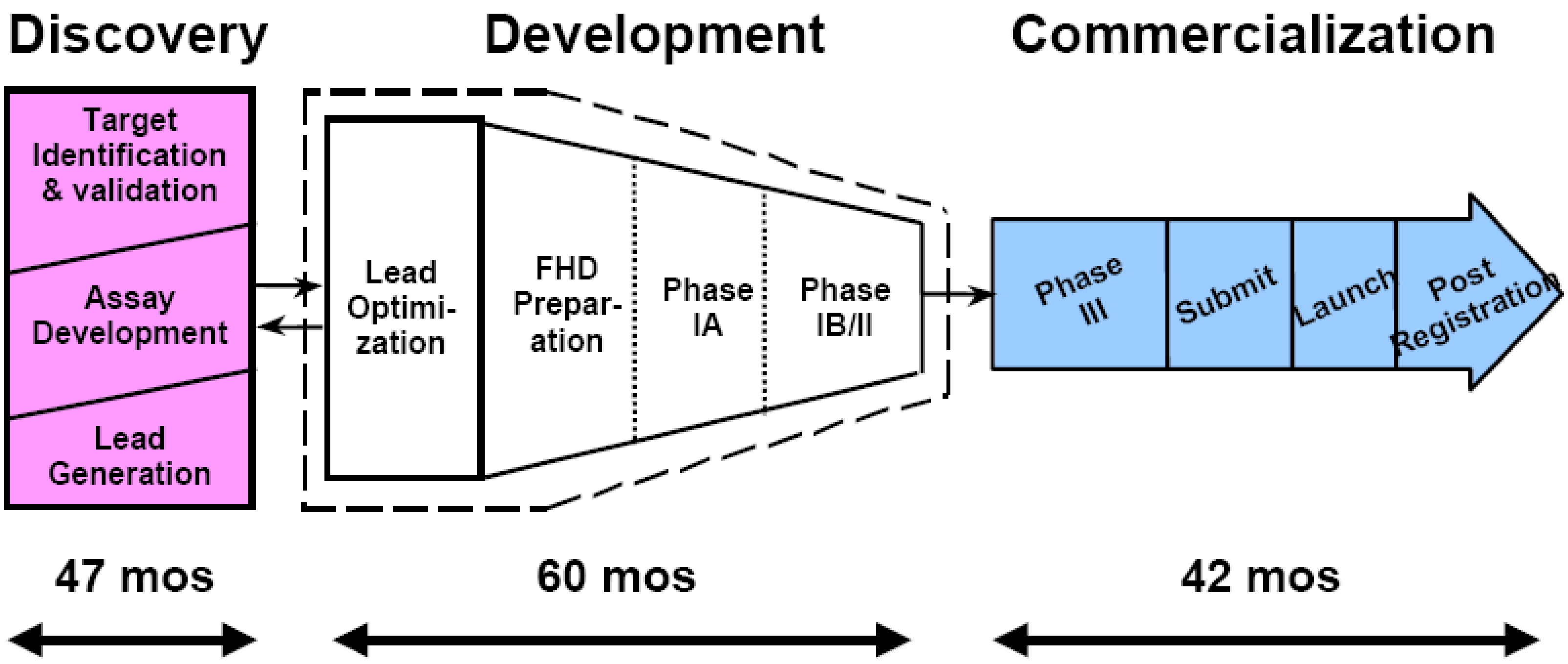
Answers That Matter.

Overview

- Brief overview of drug development
- QSAR – quantitative structure activity relationships
- Combinatorial synthesis
- Microarrays
- Data integration
- Phenotypic screening
- Managing research

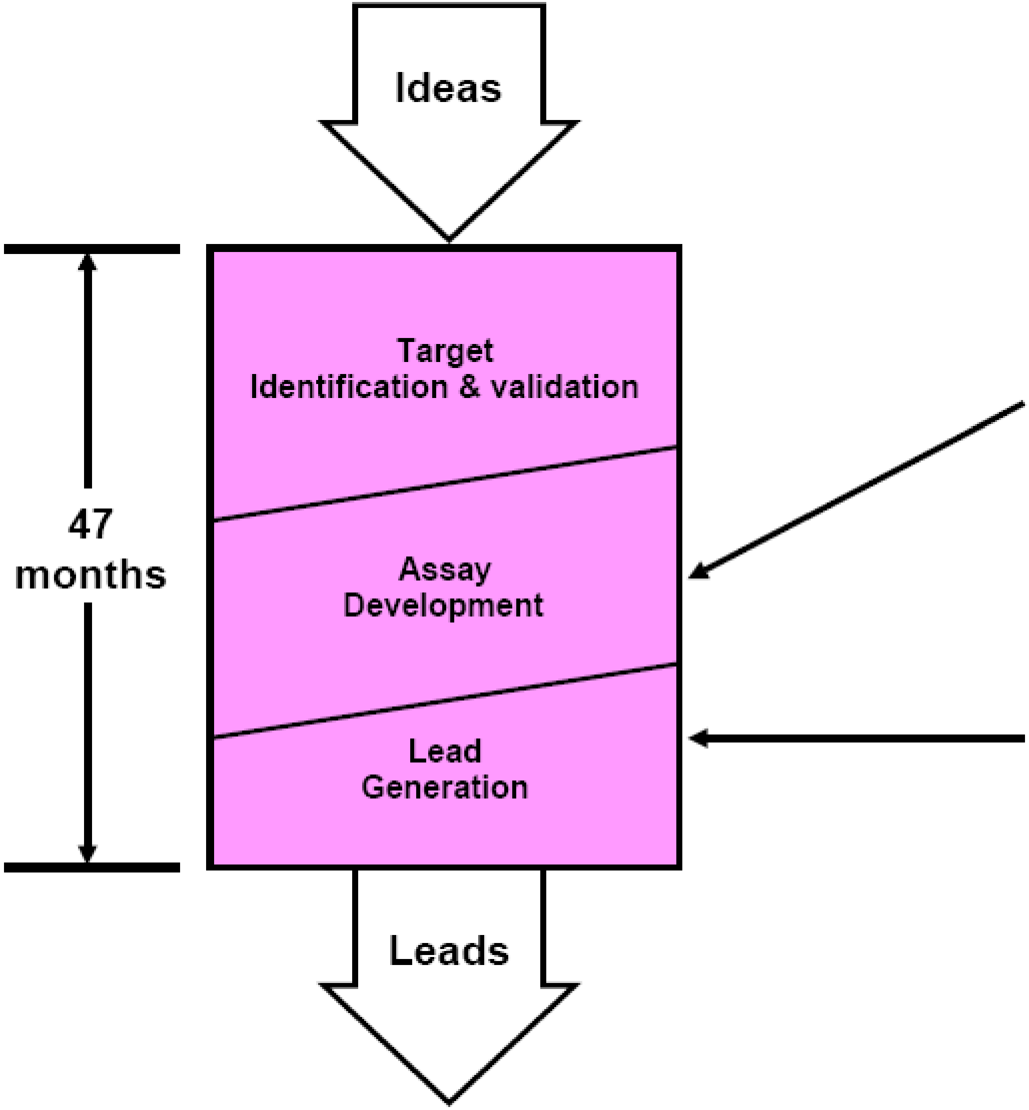
Our Development Process – The "Rocketship"

From Targets to Products



Our development process takes on average about 12 years to go from targets to products with an overall probability of success of about 2%

Discovery - Targets to Leads

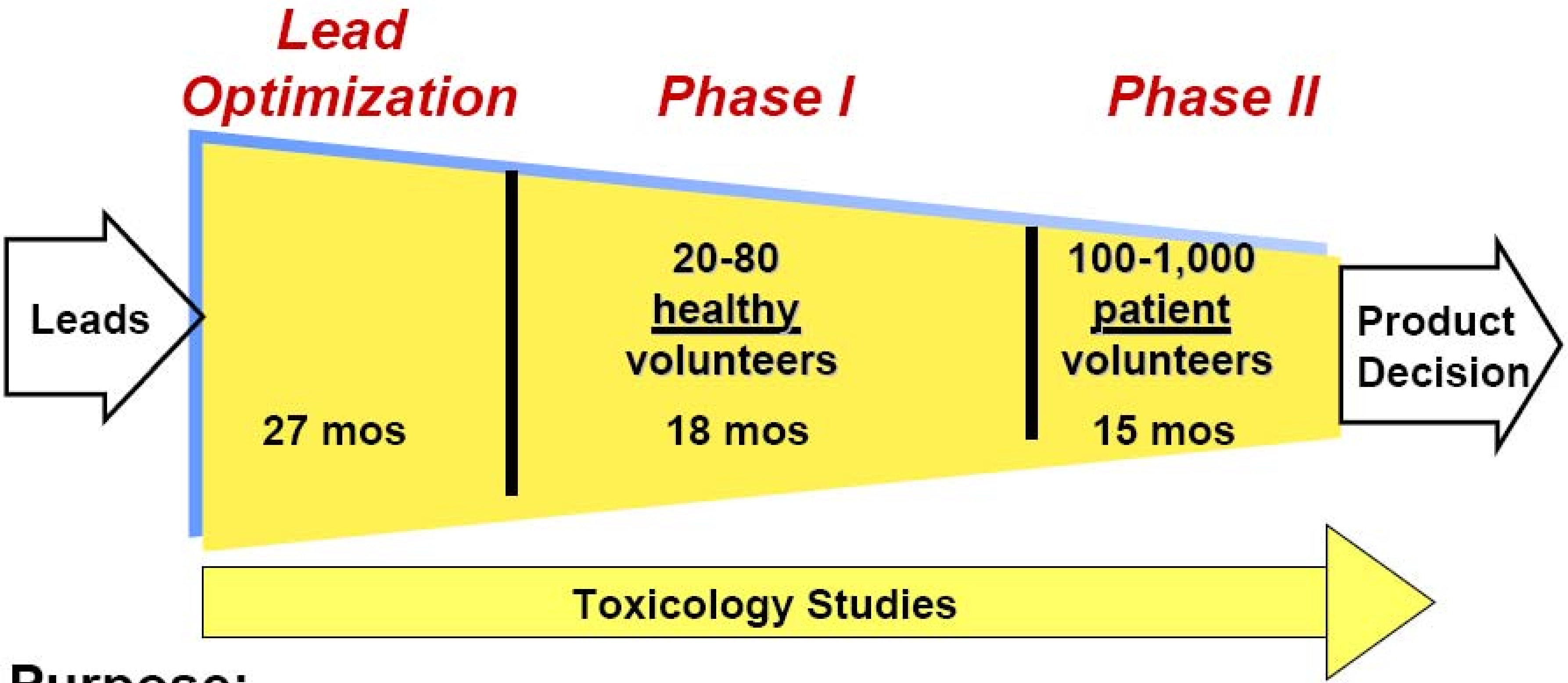


Purpose:

Develop a "screen" or test that helps us to determine if a chemical compound has the desired biological effect on a target

Testing or screening various chemical agents to determine if they have some level of the desired biological effect on a given target. Chemical libraries, combinatorial chemistry, and high throughput screening are employed here.

Development – Leads to Product Decision



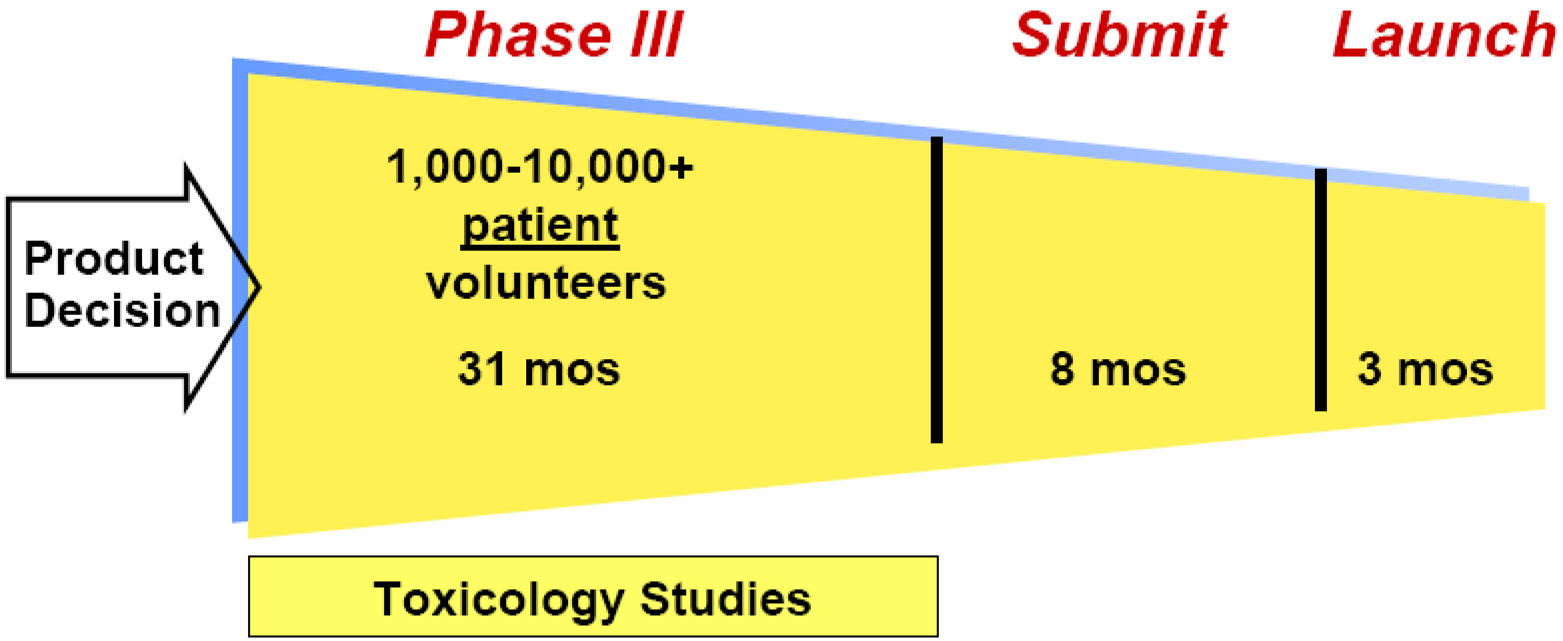
Purpose:

Lead compounds refined and optimized to give a **drug candidate** prior to going into humans. Tox, ADME, and PK studies begin.

First human dose is administered. Small doses given to healthy volunteers to test tolerance,safety, and dosing.

First efficacy dose given. Looking for therapeutic response and possible side effects. Most perilous time - approx. 75% will fail in Phase II.

Commercialization – Product Decision to Launch



Purpose:

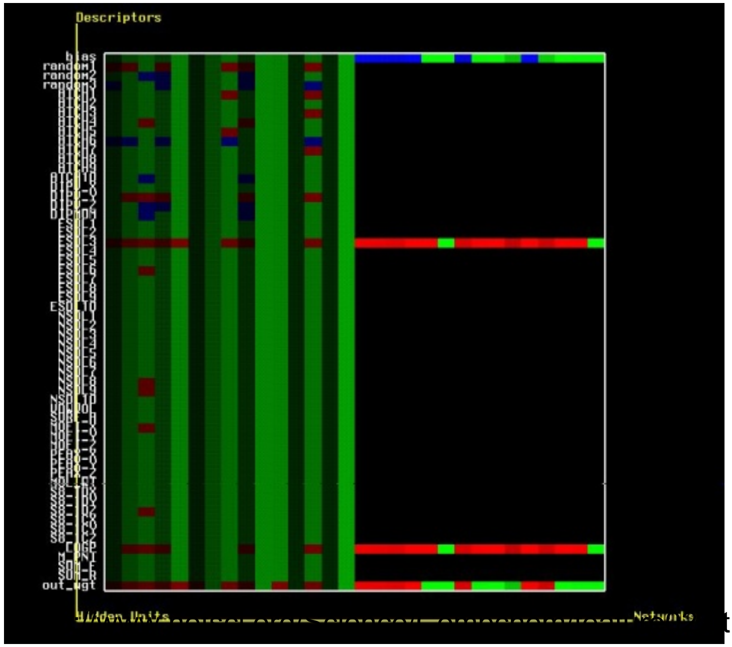
Biggest dollar spend as we study efficacy in thousands of patients worldwide to support a launch. Performance is often evaluated versus competitive drugs. Registration documents are prepared and submitted.

Dialogue with FDA during regulatory review process.

First approval and launch. Followed quickly by approval and launch in many other countries.

QSAR: Quantitative Structure Activity Relationship

- ~1990 – Have a set (10s) of molecules with an activity measure against an assay. Chemically intuitive descriptors are used to describe the molecules. Linear models used to find relationship between descriptors and activity.
- Could not realistically predict activity in new chemical spaces, but chemists could learn which descriptors would drive changes in activity and synthesize new
- Chemists would focus on synthesizing molecules that would vary those descriptors the most since they would presumably have the most effect on activity and the understanding of the chemical space.



tml

This figure shows using an artificial neural network for variable selection in QSAR. The weights of a different neural network are shown in each column, with the descriptors that were included in the reduced set on the right hand side. Green colors indicate weights near 0 and red or blue indicate positive or negative values. Note that several random inputs were used and these were then used to filter (prune) other inputs to build the smaller set of descriptors.

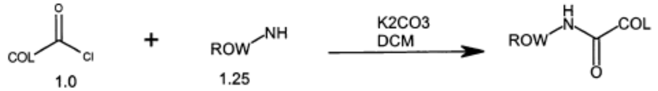
There are now companies that specialize in the model building, e.g. http://www.leadscope.com

High Throughput Screening

- Mid 90’s – large pharma gets involved in HTS
- Assumptions:
  - If we screen enough compounds, we will find new drugs.
  - *In vitro* assay is a good measure for affecting the target.
  - We understand biology enough to know that modifying the target will have the desired effect on human disease.
- Reality:
  - Too many “hits”.
  - Hits were often not drug-like molecules.
  - Too many of the “hits” were false positives.
  - Impurities could cause the activity.
- Currently:
  - Don’t screen blindly.
  - Save screening until once there is a starting point.
  - Informatics used to select a diverse library of compounds.

Combinatorial Synthesis: chemical reactions in plates

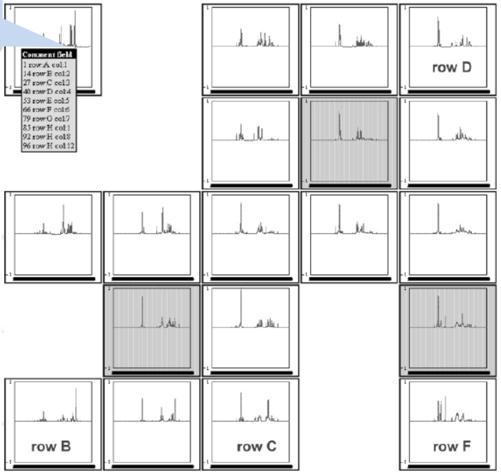
- Rapidly generate novel compounds with defined chemistry for screening
- Each row and each column has 1 compound
  - 8 + 12 starting compounds produce 96 new compounds
- Use flow NMR to verify structure in each well
  - Identifies outlier spectra to show undesired products, impurities, etc.
  - Many of these can be generated and it takes a trained NMR spectroscopist to interpret the spectra.
  - Tedious
- Informatics used to speed up and simplify the interpretation of NMR spectra by grouping similar spectra – outliers to go corners



Doped impurities

	1	2	3	4	5	6	7	8	9	10	11	12
A	▨											
B		▨										
C			▨									
D				▨								
E					▨							
F						▨						
G							▨					
H	▨							▨				▨

Spectra with impurities cluster together



Comment field
1 row:A col:1
14 row:B col:2
27 row:C col:3
40 row:D col:4
53 row:E col:5
66 row:F col:6
79 row:G col:7
85 row:H col:1
92 row:H col:8
96 row:H col:12

J. Comb. Chem., 4 (6), 622 -629, 2002

Microarrays / Gene Chips

What are Microarrays?

- Measure the expression level of essentially all the genes in a single sample
- Each chip has 30,000-50,000 probes: each can be a separate experiment
- Compare normal sample to treated sample
- Cannot simply use a pvalue for filtering: 10,000 experiments with a pvalue of 0.01 → 100 false positives

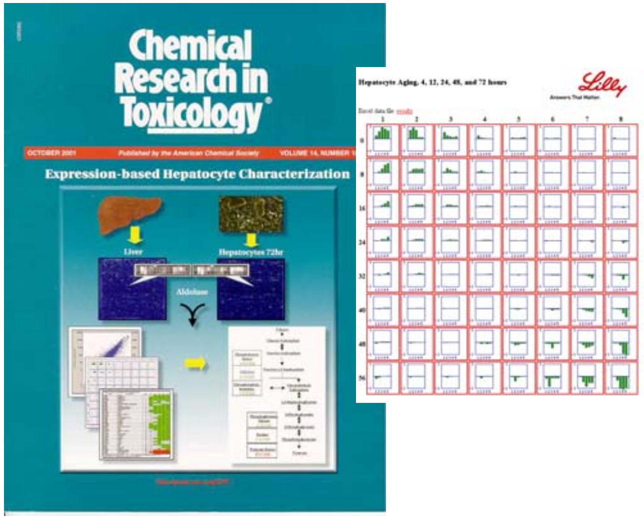
How to interpret so many results?

- Biologists are the experts in their therapeutic area – not informatics
- Often very familiar with a handful of genes and pathways
- 1000s of probesets changing
  - Easy to generate 1000s of hypotheses!
- Hypotheses can change based on arbitrary filtering criteria – much subjectivity
- Subjectivity makes it hard to know when one is done analyzing the data

Gene Analysis

1999 “List of Genes”

- 6,800 probesets on Affymetrix chip
- Clustering – HC, SOM, others
- Annotations ~ 30%
- Each chip tremendously expensive (few chips / study)
- Filter by fold change
- Pvalues
- “guilt by association”


Chemical Research in Toxicology
OCTOBER 2001
Published by the American Chemical Society
VOLUME 14, NUMBER 9
Expression-based Hepatocyte Characterization
Liver
Hepatocytes 72hr
Aldolase
Hepatocyte Aging, 4, 12, 24, 48, and 72 hours
Lilly
Answers That Matter

Chem. Res. Toxicol., **14** (9), 1218 -1231, 2001.

2008 “Biological Context”

56,000 probesets on Affymetrix chip

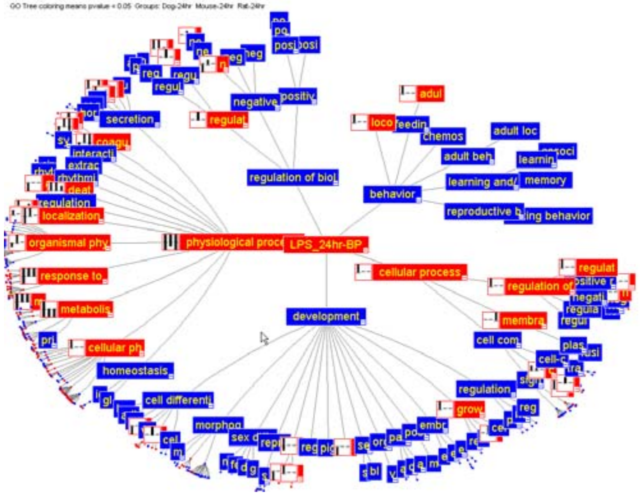
Clustering – HC, SOM, others

Annotations ~85%

Each chip less expensive
(many chips / study)

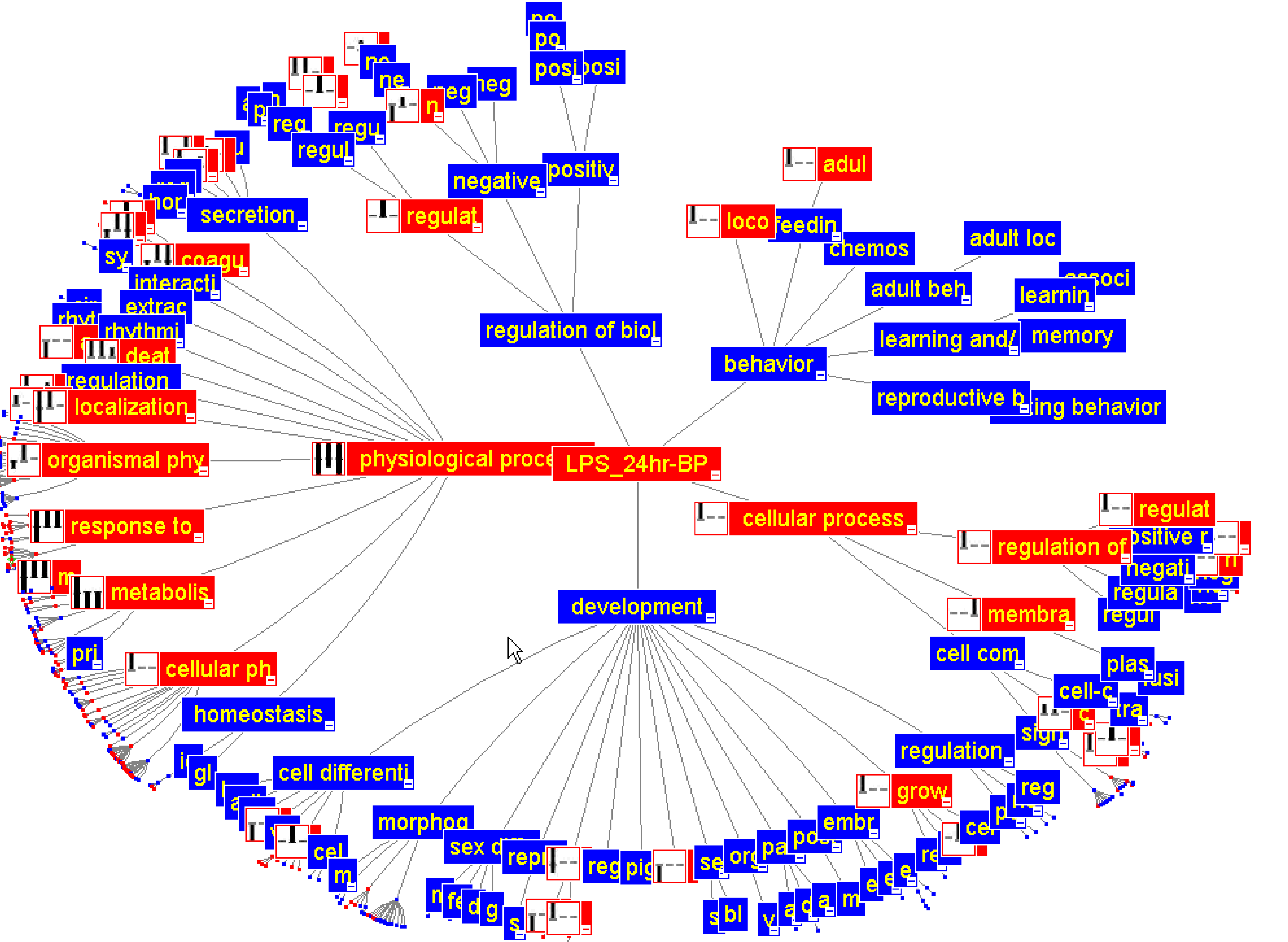
False discovery rates (multiple testing correction)

Gene Ontology



Systems Engineering. ICSEng 2005. 16-18 Aug. pp. 320- 325

GO Tree coloring means pvalue < 0.05 Groups: Dog-24hr Mouse-24hr Rat-24hr



Incorporating biology can change assumptions about filtering

We know biological changes are occurring, therefore, a good selection of genes should yield more significant biological groups.

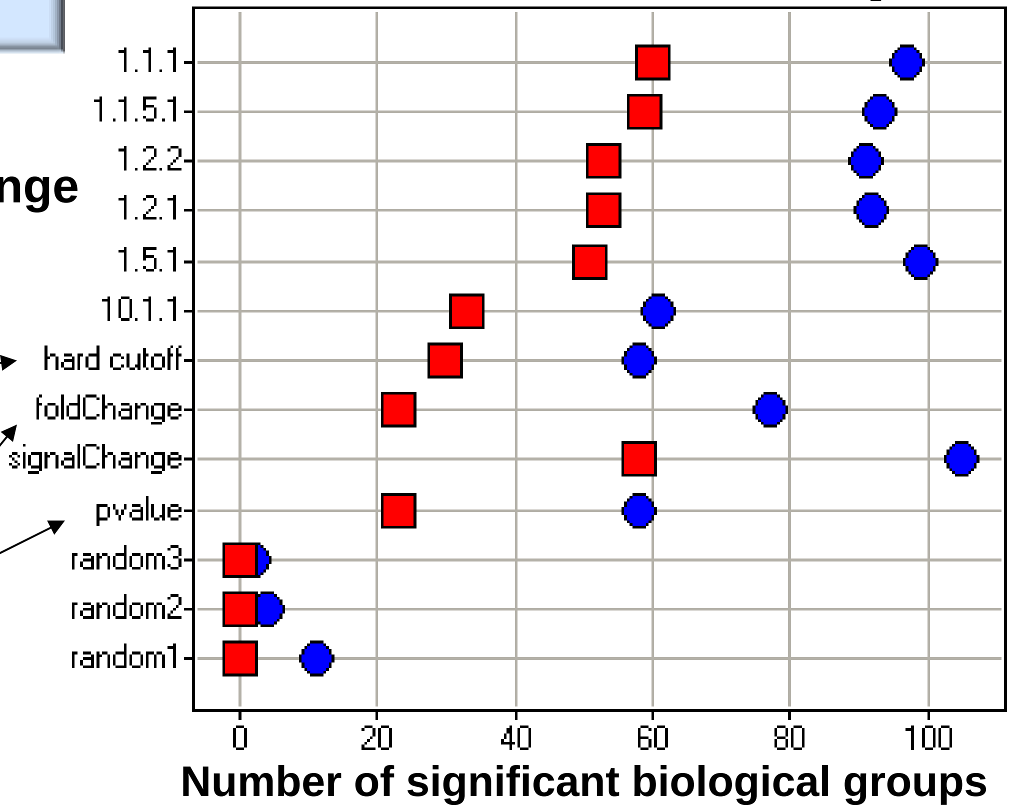
Ranking
pvalue : fold change : signal change

standard method (~2005):
pvalue (False discovery rate) < *x*
|fc|>1.2
|signal change| > 250

standard method (~2000):
pvalue < *~0.01*
|fc|> 2

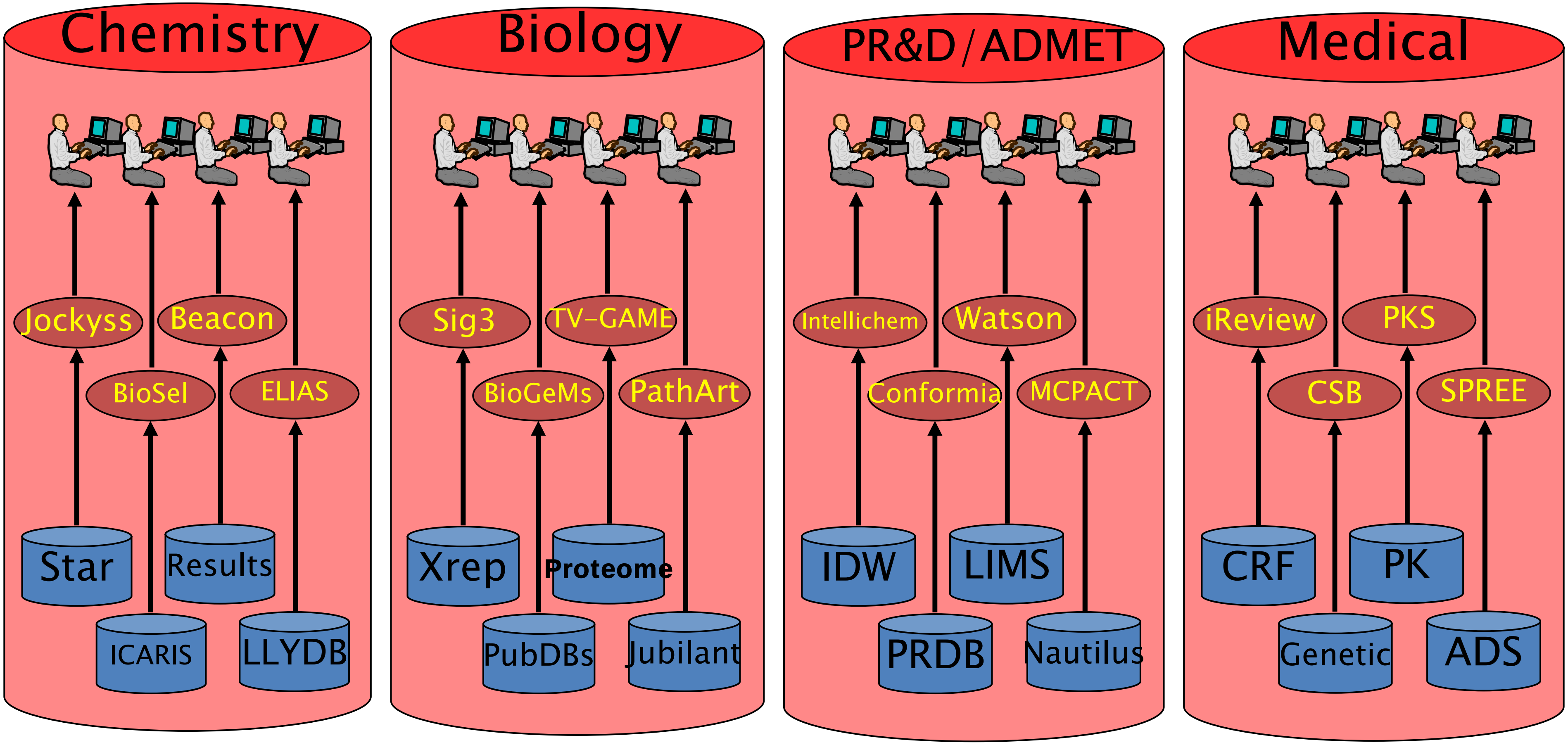
Family wise error < 0.05

fwe < 0.9



J. Cell Bio **102:6, pp. 1504 – 1518, 2007.**
Probeset list size of 1000 to 1010. Sham vs. Ovariectomy

Where are Internal Data? Silos of Silos



Chemistry

Jockyss
Beacon
BioSel
ELIAS
Star
Results
ICARIS
LLYDB

Biology

Sig3
TV-GAME
BioGeMs
PathArt
Xrep
Proteome
PubDBs
Jubilant

PR&D/ADMET

Intellichem
Watson
Conformia
MCPACT
IDW
LIMS
PRDB
Nautilus

Medical

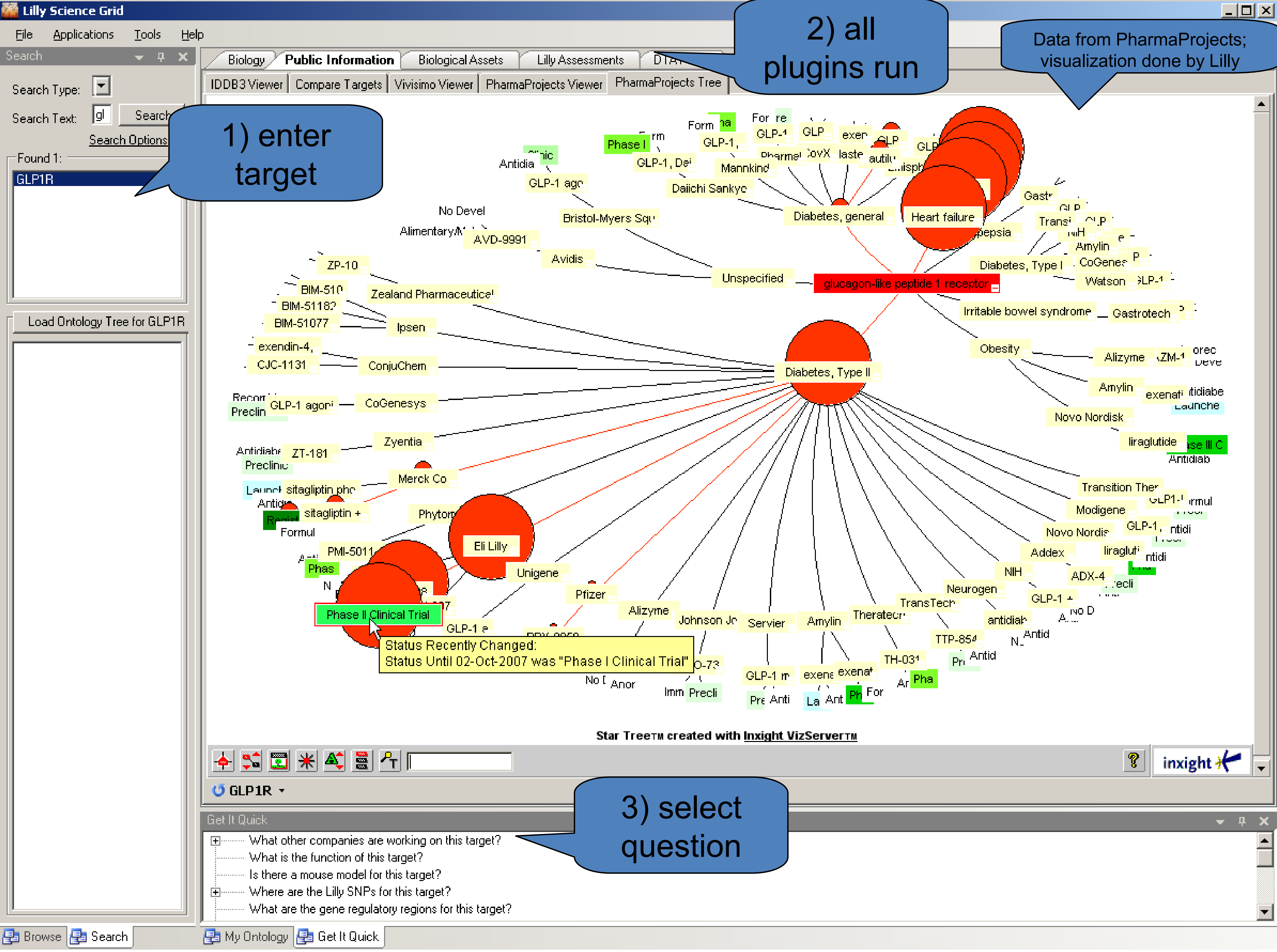
iReview
PKS
CSB
SPREE
CRF
PK
Genetic
ADS

- Tools, application, and data are standalone with limited interaction
- Scientists have great difficulty finding their data and associated tools
- Asking cross-domain questions (e.g. Discovery + Medical) very difficult
- Support becoming very impractical – estimated **400+** individual tools across silos
- Larger problem in older companies and regulated industries

How do we address?

- Use Discovery Target Assessment Tool (DTAT)
  - *DTAT allows scientists to evaluate drug targets. DTAT allows scientists to select the scientific question of interest and returns data that is in the appropriate context.*
- Built upon Life Science Grid: LSG available on http://www.sourceforge.net
- Uses RDF (resource description format) to store information about targets, pharmacology, internal development, disease
- Plugins use “listeners” to respond to appropriate data type and serve information
- Question framework allows scientists to learn how each data source provides relevant data
  - Questions stay relatively constant, data and sources change.
  - If informatics is doing proper job, we are providing the best answers for the questions.

Show DTAT



1) enter target

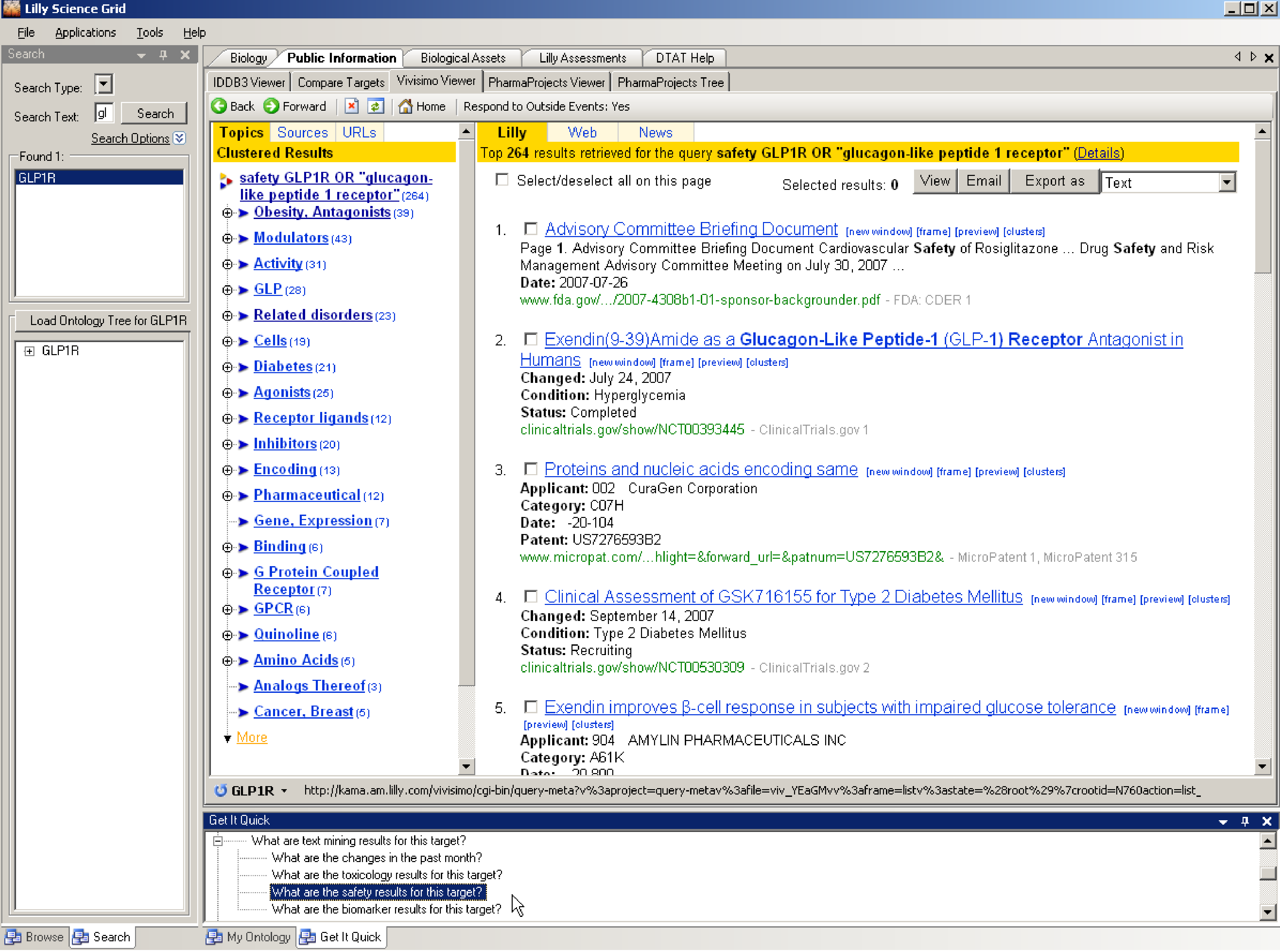
2) all plugins run

Data from PharmaProjects; visualization done by Lilly

3) select question

Status Recently Changed:
Status Until 02-Oct-2007 was "Phase I Clinical Trial"

- What other companies are working on this target?
- What is the function of this target?
- Is there a mouse model for this target?
- Where are the Lilly SNPs for this target?
- What are the gene regulatory regions for this target?



Search Type:

Search Text: gl Search

Search Options

Found 1:

GLP1R

Load Ontology Tree for GLP1R

⊞ GLP1R

Clustered Results

- safety GLP1R OR "glucagon-like peptide 1 receptor" (264)
  - ⊕ Obesity, Antagonists (39)
  - ⊕ Modulators (43)
  - ⊕ Activity (31)
  - ⊕ GLP (28)
  - ⊕ Related disorders (23)
  - ⊕ Cells (19)
  - ⊕ Diabetes (21)
  - ⊕ Agonists (25)
  - ⊕ Receptor ligands (12)
  - ⊕ Inhibitors (20)
  - ⊕ Encoding (13)
  - ⊕ Pharmaceutical (12)
  - Gene, Expression (7)
  - ⊕ Binding (6)
  - ⊕ G Protein Coupled Receptor (7)
  - ⊕ GPCR (6)
  - ⊕ Quinoline (6)
  - ⊕ Amino Acids (5)
  - Analogs Thereof (3)
  - Cancer, Breast (5)
  - More

Top **264** results retrieved for the query **safety GLP1R OR "glucagon-like peptide 1 receptor"** (Details)

□ Select/deselect all on this page   Selected results: **0**   View Email Export as Text

1. □ Advisory Committee Briefing Document [new window] [frame] [preview] [clusters]
   Page **1**. Advisory Committee Briefing Document Cardiovascular **Safety** of Rosiglitazone ... Drug **Safety** and Risk Management Advisory Committee Meeting on July 30, 2007 ...
   Date: 2007-07-26
   www.fda.gov/.../2007-4308b1-01-sponsor-backgrounder.pdf - FDA: CDER 1

2. □ Exendin(9-39)Amide as a **Glucagon-Like Peptide-1** (GLP-1) **Receptor** Antagonist in Humans [new window] [frame] [preview] [clusters]
   Changed: July 24, 2007
   Condition: Hyperglycemia
   Status: Completed
   clinicaltrials.gov/show/NCT00393445 - ClinicalTrials.gov 1

3. □ Proteins and nucleic acids encoding same [new window] [frame] [preview] [clusters]
   Applicant: 002 CuraGen Corporation
   Category: C07H
   Date: -20-104
   Patent: US7276593B2
   www.micropat.com/...hlight=&forward_url=&patnum=US7276593B2& - MicroPatent 1, MicroPatent 315

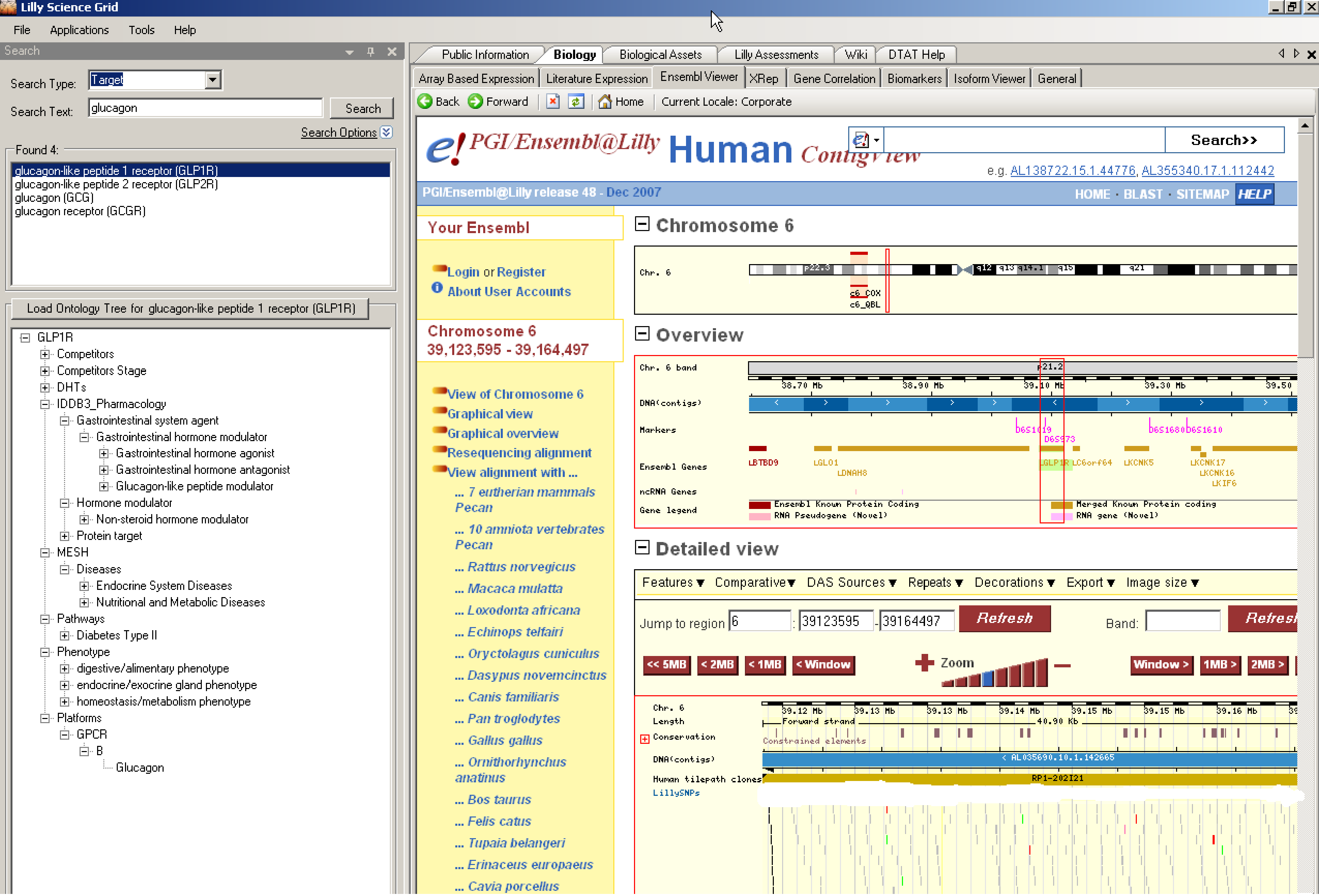
4. □ Clinical Assessment of GSK716155 for Type 2 Diabetes Mellitus [new window] [frame] [preview] [clusters]
   Changed: September 14, 2007
   Condition: Type 2 Diabetes Mellitus
   Status: Recruiting
   clinicaltrials.gov/show/NCT00530309 - ClinicalTrials.gov 2

5. □ Exendin improves β-cell response in subjects with impaired glucose tolerance [new window] [frame] [preview] [clusters]
   Applicant: 904 AMYLIN PHARMACEUTICALS INC
   Category: A61K
   Date: -20-800

GLP1R http://kama.am.lilly.com/vivisimo/cgi-bin/query-meta?v%3aproject=query-metav%3afile=viv_YEaGMvv%3aframe=listv%3astate=%28root%29%7crootid=N760action=list_

Get It Quick

- What are text mining results for this target?
  - What are the changes in the past month?
  - What are the toxicology results for this target?
  - What are the safety results for this target?
  - What are the biomarker results for this target?



Search Type: Target

Search Text: glucagon

Found 4:

- glucagon-like peptide 1 receptor (GLP1R)
- glucagon-like peptide 2 receptor (GLP2R)
- glucagon (GCG)
- glucagon receptor (GCGR)

Load Ontology Tree for glucagon-like peptide 1 receptor (GLP1R)

```
GLP1R
  Competitors
  Competitors Stage
  DHTs
  IDDB3_Pharmacology
    Gastrointestinal system agent
      Gastrointestinal hormone modulator
        Gastrointestinal hormone agonist
        Gastrointestinal hormone antagonist
        Glucagon-like peptide modulator
    Hormone modulator
      Non-steroid hormone modulator
    Protein target
  MESH
    Diseases
      Endocrine System Diseases
      Nutritional and Metabolic Diseases
  Pathways
    Diabetes Type II
  Phenotype
    digestive/alimentary phenotype
    endocrine/exocrine gland phenotype
    homeostasis/metabolism phenotype
  Platforms
    GPCR
      B
        Glucagon
```

Current Locale: Corporate

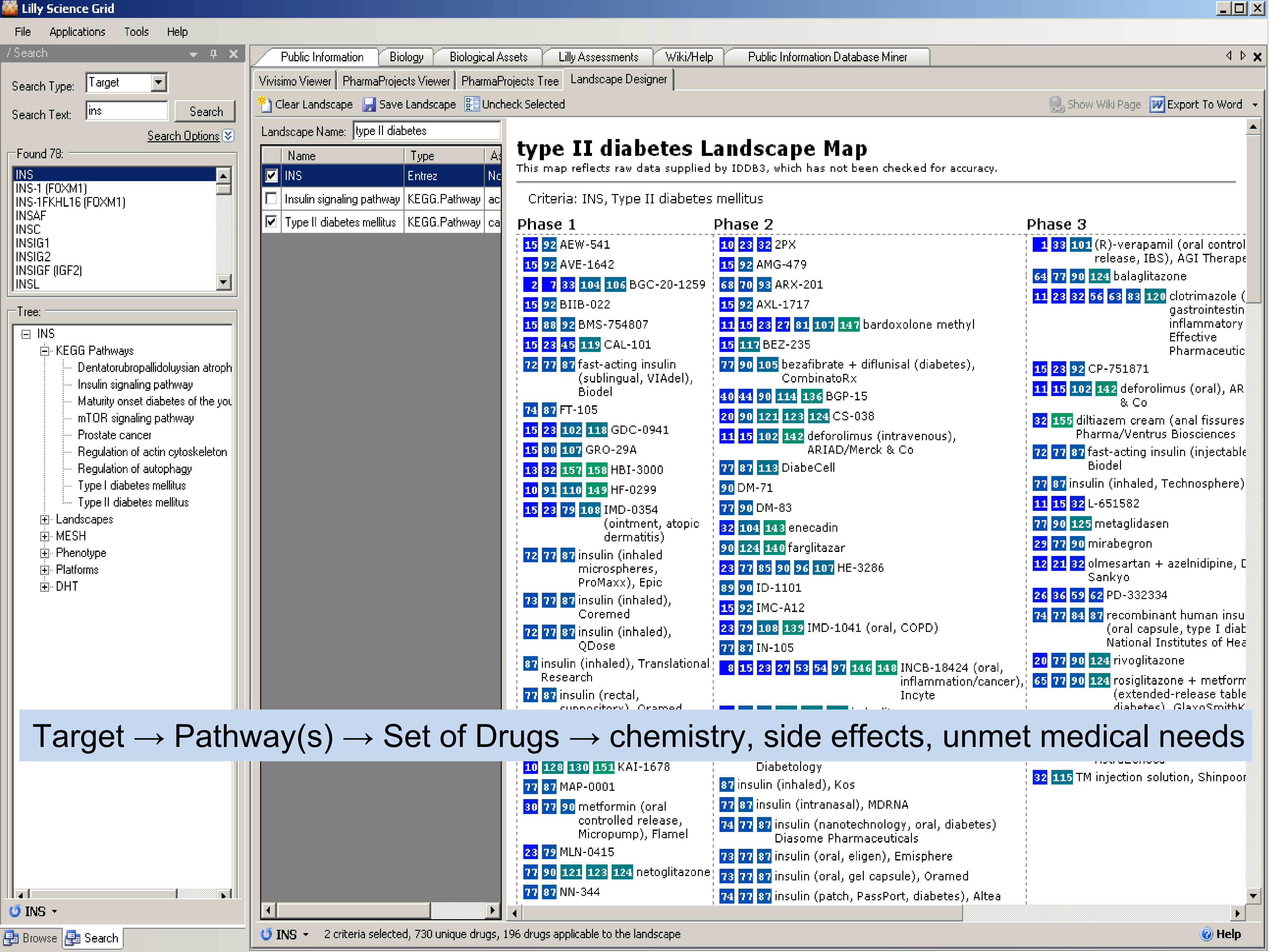
PGI/Ensembl@Lilly release 48 - Dec 2007

e.g. AL138722.15.1.44776, AL355340.17.1.112442

Chromosome 6
39,123,595 - 39,164,497

Jump to region 6 : 39123595 - 39164497

Band:



Lilly Science Grid

File Applications Tools Help

Search

Search Type: Target

Search Text: ins

Search

Search Options

Found 78:

- INS
- INS-1 (FOXM1)
- INS-1FKHL16 (FOXM1)
- INSAF
- INSC
- INSIG1
- INSIG2
- INSIGF (IGF2)
- INSL

Tree:

- INS
  - KEGG Pathways
    - Dentatorubropallidoluysian atroph
    - Insulin signaling pathway
    - Maturity onset diabetes of the you
    - mTOR signaling pathway
    - Prostate cancer
    - Regulation of actin cytoskeleton
    - Regulation of autophagy
    - Type I diabetes mellitus
    - Type II diabetes mellitus
  - Landscapes
  - MESH
  - Phenotype
  - Platforms
  - DHT

INS

Browse Search

Public Information | Biology | Biological Assets | Lilly Assessments | Wiki/Help | Public Information Database Miner

Vivisimo Viewer | PharmaProjects Viewer | PharmaProjects Tree | Landscape Designer

Clear Landscape | Save Landscape | Uncheck Selected | Show Wiki Page | Export To Word

Landscape Name: type II diabetes

	Name	Type	As
☑	INS	Entrez	Nc
□	Insulin signaling pathway	KEGG.Pathway	ac
☑	Type II diabetes mellitus	KEGG.Pathway	ca

type II diabetes Landscape Map

This map reflects raw data supplied by IDDB3, which has not been checked for accuracy.

Criteria: INS, Type II diabetes mellitus

Phase 1

15 92 AEW-541

15 92 AVE-1642

2 7 33 104 106 BGC-20-1259

15 92 BIIB-022

15 88 92 BMS-754807

15 23 45 119 CAL-101

72 77 87 fast-acting insulin (sublingual, VIAdel), Biodel

74 87 FT-105

15 23 102 118 GDC-0941

15 80 107 GRO-29A

13 32 157 158 HBI-3000

10 91 110 149 HF-0299

15 23 79 108 IMD-0354 (ointment, atopic dermatitis)

72 77 87 insulin (inhaled microspheres, ProMaxx), Epic

73 77 87 insulin (inhaled), Coremed

72 77 87 insulin (inhaled), QDose

87 insulin (inhaled), Translational Research

77 87 insulin (rectal, suppository), Oramed

10 128 130 151 KAI-1678

77 87 MAP-0001

30 77 90 metformin (oral controlled release, Micropump), Flamel

23 79 MLN-0415

77 90 121 123 124 netoglitazone

77 87 NN-344

Phase 2

10 23 32 2PX

15 92 AMG-479

68 70 93 ARX-201

15 92 AXL-1717

11 15 23 27 81 107 147 bardoxolone methyl

15 117 BEZ-235

77 90 105 bezafibrate + diflunisal (diabetes), CombinatoRx

40 44 90 114 136 BGP-15

20 90 121 123 124 CS-038

11 15 102 142 deforolimus (intravenous), ARIAD/Merck & Co

77 87 113 DiabeCell

90 DM-71

77 90 DM-83

32 104 143 enecadin

90 124 140 farglitazar

23 77 85 90 96 107 HE-3286

89 90 ID-1101

15 92 IMC-A12

23 79 108 139 IMD-1041 (oral, COPD)

77 87 IN-105

8 15 23 27 53 54 97 146 148 INCB-18424 (oral, inflammation/cancer), Incyte

Diabetology

87 insulin (inhaled), Kos

77 87 insulin (intranasal), MDRNA

74 77 87 insulin (nanotechnology, oral, diabetes) Diasome Pharmaceuticals

73 77 87 insulin (oral, eligen), Emisphere

73 77 87 insulin (oral, gel capsule), Oramed

74 77 87 insulin (patch, PassPort, diabetes), Altea

Phase 3

1 33 101 (R)-verapamil (oral control release, IBS), AGI Therape

64 77 90 124 balaglitazone

11 23 32 56 63 83 120 clotrimazole (gastrointestin inflammatory Effective Pharmaceutic

15 23 92 CP-751871

11 15 102 142 deforolimus (oral), AR & Co

32 155 diltiazem cream (anal fissures Pharma/Ventrus Biosciences

72 77 87 fast-acting insulin (injectable Biodel

77 87 insulin (inhaled, Technosphere)

11 15 32 L-651582

77 90 125 metaglidasen

29 77 90 mirabegron

12 21 32 olmesartan + azelnidipine, D Sankyo

26 36 59 62 PD-332334

74 77 84 87 recombinant human insu (oral capsule, type I diab National Institutes of Hea

20 77 90 124 rivoglitazone

65 77 90 124 rosiglitazone + metform (extended-release table diabetes), GlaxoSmithK

32 115 TM injection solution, Shinpoor

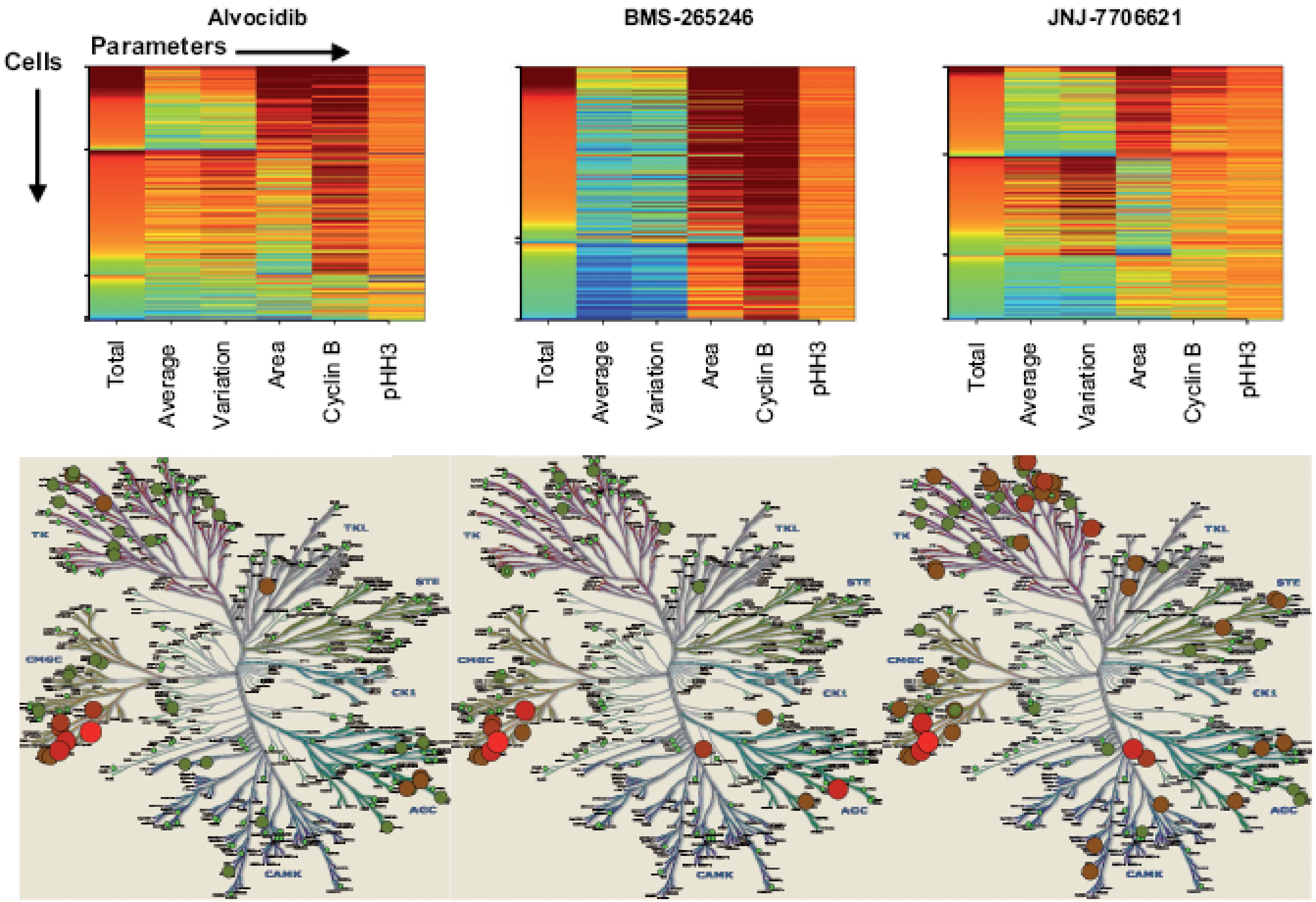
INS 2 criteria selected, 730 unique drugs, 196 drugs applicable to the landscape

Help

Target → Pathway(s) → Set of Drugs → chemistry, side effects, unmet medical needs

Phenotypic Drug Discovery

cyclin-dependent kinase 1 inhibitors



- In vivo (cell based assays), use imaging techniques to measure variety of biological parameters
- No need to choose a target - and possibly be wrong!
- **Current Opinion in Drug Discovery & Development 2008 11(3):338-345** Jonathan

Managing research

- Part of the challenge is how to manage the research, When development costs are high and failure is common, companies should structure research to seek truth first, success second.
- Project champions can often marshal resources to keep a project moving – may not be sufficiently motivated to do the experiment that could kill their idea
- Advocate early stages of research for "Truth Seekers". Evaluate many projects and rewarded for objectivity
- Since most molecules in the early stage fail, manage to assume failure of the asset instead of creating infrastructure to ramp up production early. This may delay a successful molecule, but otherwise there is a large opportunity cost as fewer early stage assets may be pursued.
- Clean up this page…
- "A More Rational Approach to New-Product Development", by Eric Bonabeau, Neil Bodick, and Robert W. Armstrong *Harv Bus Rev*. 2008 Mar;86(3):96-102

Summary

- Target focused research – assumes we know enough biology to optimize the right things
  - Initially optimized one parameter: activity (optimize only the cathode)
    - Must also optimize side effects, safety margin, population effects, dosing, etc.
  - Adjust design parameters to gain the most information
  - Help interpret the results
  - Adding background information can improve quality of results (optimize entire battery)
  - Integrating many data sources can improve the decision quality
- Phenotypic screening (measure performance of the car which is made up of a set of batteries with powertrain etc.)
  - Advances in technology allows higher throughput cell based assays that measure biology
  - Can skip the target stage
- How to reward scientists to remove molecules from the pipeline?

Backups

Life Science Grid

- LSG – an asynchronous web services (message oriented) “smart” client-side application deployed using Microsoft ClickOnce deployment strategy.
- Software Development: Microsoft Visual Studio 2005
- Client: Windows XP SP2, .NET Framework 2.0, WSE 3.0
- Server: Windows 2003 Enterprise Edition, SP1, .NET Framework 2.0 and IIS 6
- Databases Supported:
  - MySQL 5.0
  - Microsoft SQL Server 2005 Express Edition
  - Oracle Database 10g Express Edition

- Available on http://www.sourceforge.net . Search for LSG
- Framework will include sample public domain plugins
- Documentation “how to” for software developers

Data is being generated at an increasing rate – how to get relevant data?

- Difficult or impossible for any scientist to know all the sources – scientists asked to work more outside their own areas
- Nucleic Acids Research, DB issue
  - 1078 databases, 110 more than last year
  - links to more than 80 databases have been updated
  - only 25 obsolete databases have been removed
- Multiple ways of describing the same or similar data (same or similar depends on point of view)
  - MESH, PathArt disease, PharmaProjects indications, gene ontology, IDDB3 Pharmacology
  - Intelligent people can disagree, e.g., gene x causes cancer or gene x does not cancer. Both could have the same numerical results and have a different arbitrary cutoff.
  - How does one query across overlapping data?

Data are generated faster than they can be understood

- Must find data that are relevant
  - Tremendous duplication
  - What is the current answer?
  - wheat from chaff
- Find connections in data
  - visualization
  - words
  - Statistics
- Difficulty measuring value of data, e.g. compare to compute speed
  - database quality
    - database 1 vs. database 2
    - agreement
    - quality measure of each element
- Data curation is expensive
- More than just having the data: ability to retrieve relevant decision-making information must be part of the value metric

Informatics in Drug Discovery

This talk will begin with a brief overview of the various stages of drug development. Model building and chemical methods will initially be described from the early 90s. These will serve as a basis for comparison for later methods such as high throughput screening, medium throughput screening, and phenotypic drug discovery. Microarrays, with their ability to measure gene changes across the entire genome, will be described as a means of interrogating biological systems with the associated challenges of understanding the results. Recent work using the Life Science Grid will be covered as a means of integrating relevant information from many sources. Finally, other organizational shifts will be discussed that may facilitate more efficient breakthroughs.