Data Mining and Validation of Associations between Drugs, Genes, Pathways and Diseases

Prof. Dr. Hasan Davulcu
School of Computing and Informatics, ASU

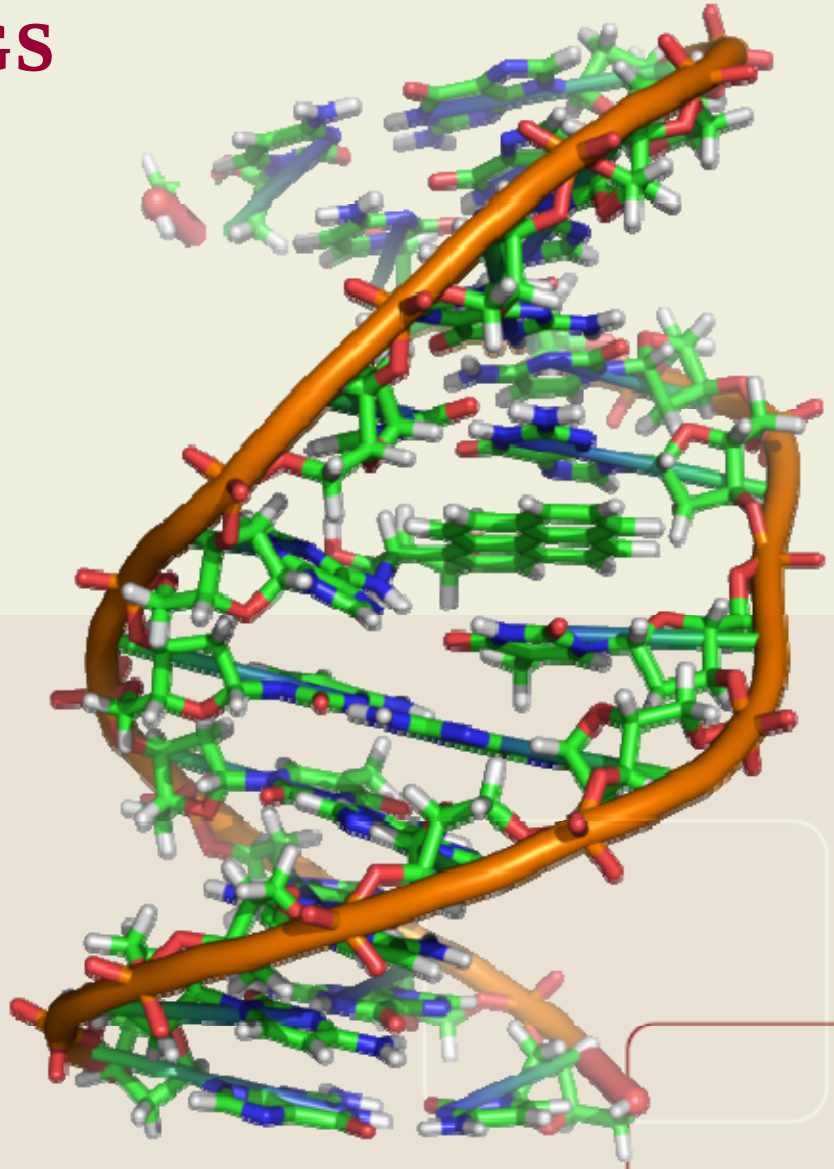
VEF Alumni Conference

Hanoi, Vietnam

December 12, 2008

DRUGS

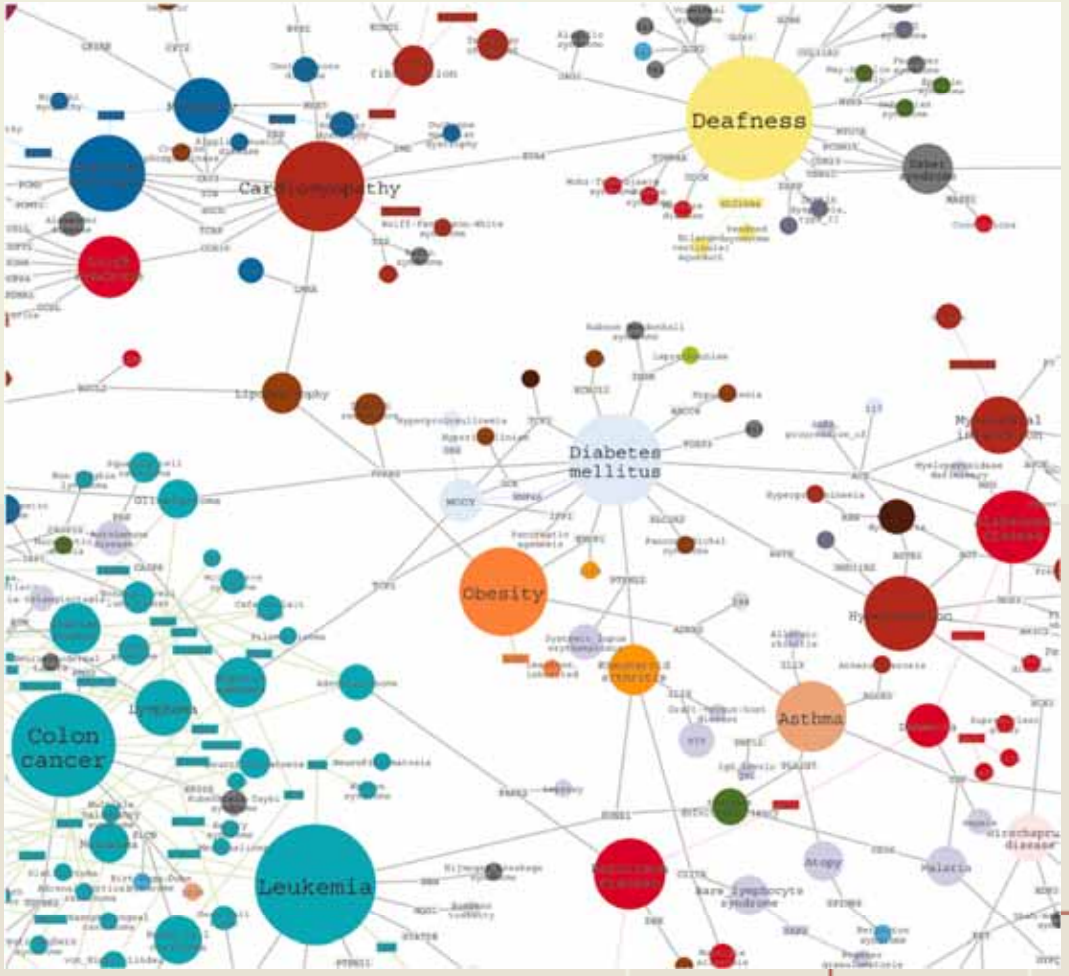
- **Developing new drugs is a lengthy and expensive process**
- Prestwick Chemical Library® containing **1120** FDA approved compounds
- Structurally diverse drug molecules of known safety and bioavailability in humans
- (e.g, statins, neuropsychiatry, cardiology, immunology, anti-inflammatory, and analgesia related drugs)



DISEASES

Wide spectrum of diseases:

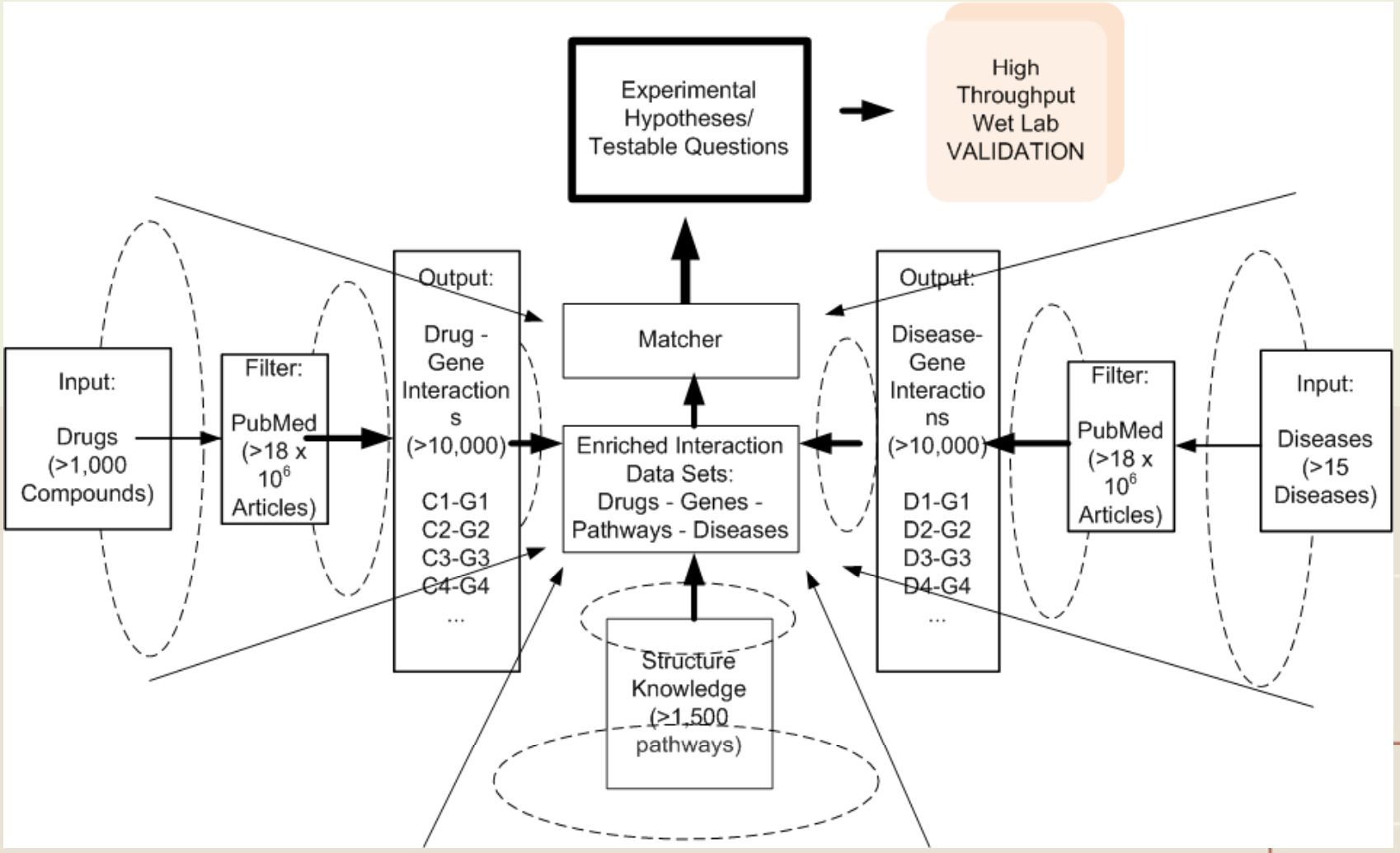
- Cancers
- Age-elated diseases Alzheimer's, Hypertension, age-related macular degeneration (AMD),
- diseases with strong genetic components such as sickle cell Anemia, Obesity, Diabetes,
- Endometriosis, Cholesterol-related defects, Multiple Sclerosis, Parkinson's, autism, and Inflammatory Bowel Disease



"Low-Hanging Fruit"

- Our approach is to identify new connections between diseases, molecular targets, and drugs that can be easily and cheaply translated into new treatments.
- The proposed approach will test the direct effect of hundreds of existing drugs on thousands of genes and, indirectly, on hundreds of pathways and dozens of diseases in which these genes are involved.
- We will employ an informatics step that will identify the best molecular target candidates

"Funnel" *in silico* informatics approach



Combine informatics and high throughput experiments:

a) the use of an information extraction system to extract protein, DNA, and RNA lists from 18 million PUBMED abstracts;

b) application of keyword search and summarization techniques on extracted data by utilizing graphical models and probabilistic inference techniques;

c) utilization of biological knowledge databases, such as pathway information, to identify lists of genes that are biologically related and are also associated from the literature with drugs and diseases;

d) high-throughput screening using the Prestwick Chemical Library® containing 1120 compounds that have structurally diverse drug molecules of known safety and bioavailability in humans, and corresponding to molecular scaffolds that likely hit multiple target classes;

e) validation of potential biological targets using quantitative real-time PCR (qPCR) [7] and Western blot analysis.

IntEx: Bio-Medical Interaction Extractor

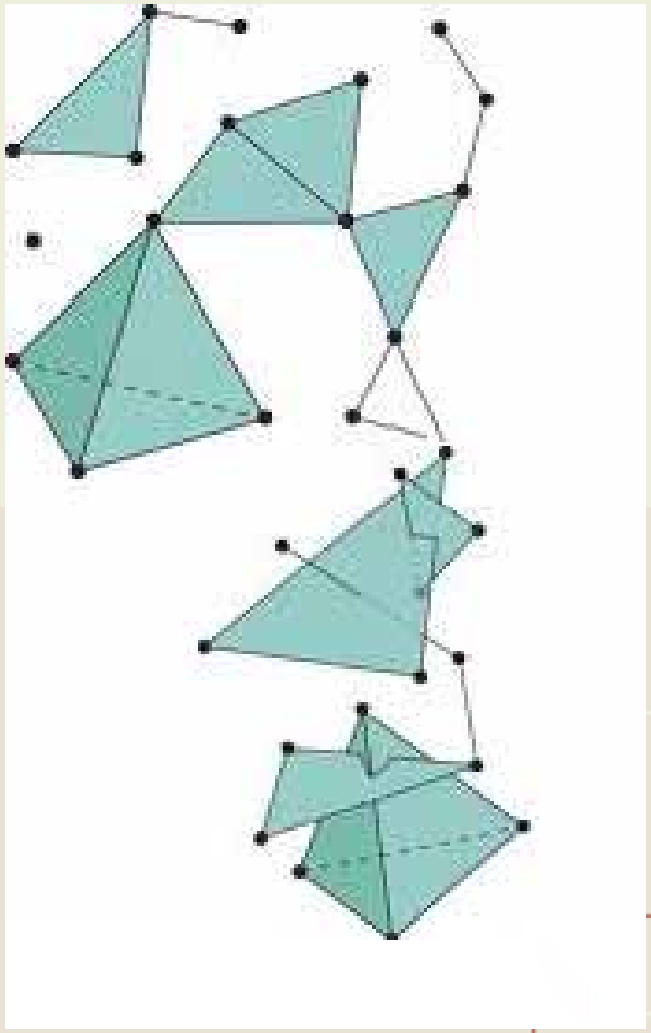
18 Million PUBMED abstracts

"HMBA could inhibit the MEC-1 cell proliferation by down-regulation of PCNA expression."

HMBA	down-regulates	PCNA expression
HMBA	inhibit	MEC-1 cell proliferation

Mining Synergies

- Recently, **Dr. Nguyen** developed an **undirected graphical model** that can be succinctly used to mine the **co-dependence patterns** among **groups of genes** from their pair-wise interactions.
- *We hypothesize that, these structures will capture biologically significant relationships among:*
  - *groups of common genes between diseases and drug compounds*
  - *while alternative models that examine pair-wise relations of genes alone will not be as expressive.*



Significant Subsets of Genes

Microsoft Excel - Triplets.xls

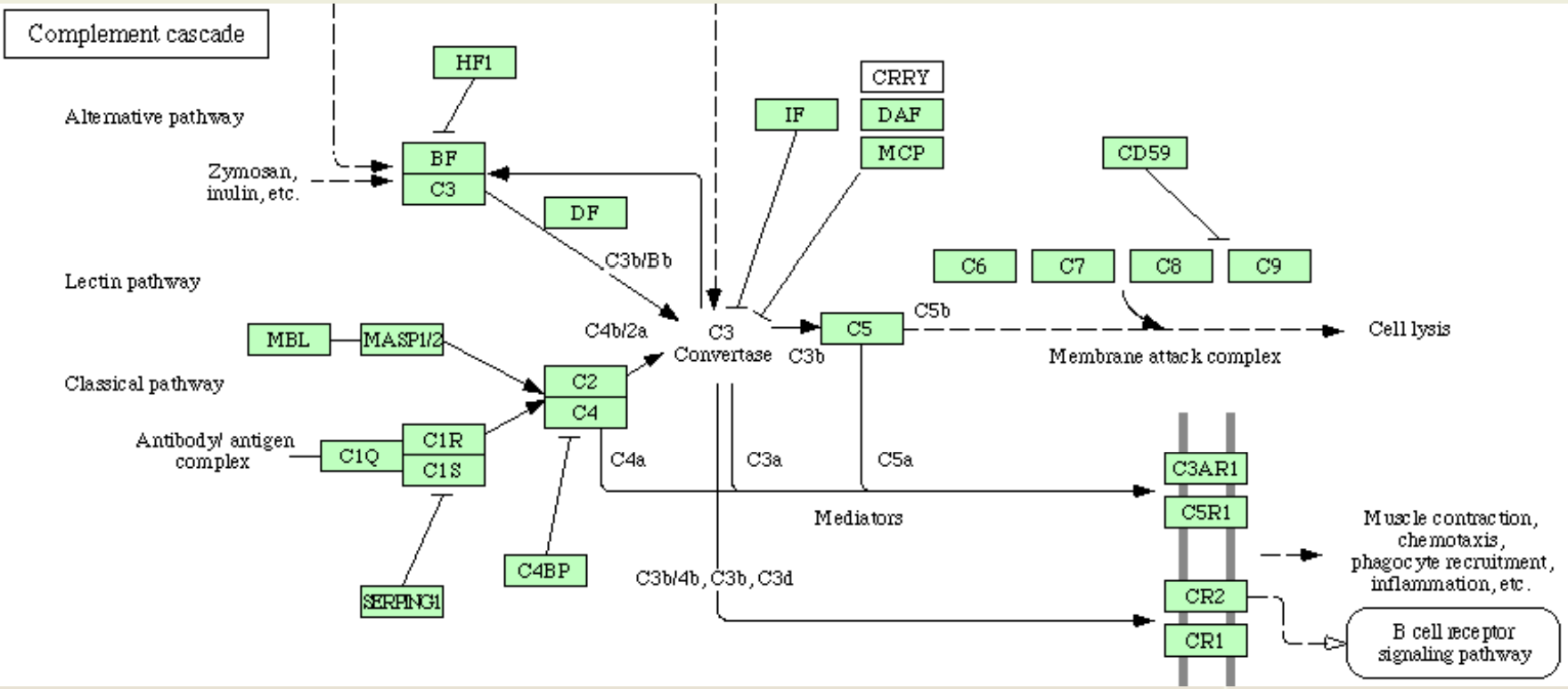
D1 Normalized Drug Count

	A	B	C	F	G
1	Drug Name	Disease Name	Common Gene Triplets	Combined Triplet Score	
2	lovastatin	breast cancer	{ statins , pravastatin , lovastatin }	15.736456	
3	lovastatin	breast cancer	{ fluvastatin , pravastatin , lovastatin }	13.006495	
4	lovastatin	breast cancer	{ atorvastatin , statins , lovastatin }	12.920878	
5	lovastatin	breast cancer	{ atorvastatin , lovastatin , pravastatin }	12.847203	
6	lovastatin	beta - thalassemia	{ gamma , alpha , beta }	12.715048	
7	lovastatin	pancreatic cancer	{ p53 , ras , p16 }	12.55407	
8	lovastatin	breast cancer	{ beta , alpha , gamma }	12.207747	
9	lovastatin	breast cancer	{ estrogen , mitogen , kinase }	11.54467	
10	lovastatin	diabetes	{ pravastatin , fluvastatin , lovastatin }	10.879509	
11	lovastatin	diabetes	{ pravastatin , fluvastatin , simvastatin }	10.781474	
12	lovastatin	breast cancer	{ beta , alpha , kinase }	10.50728	
13	lovastatin	breast cancer	{ fluvastatin , lovastatin , statins }	10.501539	
14	lovastatin	breast cancer	{ atorvastatin , statins , pravastatin }	10.051527	
15	lovastatin	diabetes	{ kinase , enzyme , insulin }	10.01778	
16	lovastatin	breast cancer	{ fluvastatin , lovastatin , atorvastatin }	9.973872	
17	lovastatin	breast cancer	{ fluvastatin , pravastatin , statins }	9.322737	
18	lovastatin	breast cancer	{ fluvastatin , pravastatin , atorvastatin }	9.016095	
19	lovastatin	breast cancer	{ mitogen , kinase , p38 }	8.872528	
20	lovastatin	pancreatic cancer	{ p53 , ras , oncogene }	8.8496	
21	lovastatin	breast cancer	{ alpha , kinase , mitogen }	8.263223	
22	lovastatin	breast cancer	{ atorvastatin , fluvastatin , statins }	7.988623	
23	lovastatin	diabetes	{ pravastatin , fluvastatin , statins }	7.8899	
24	lovastatin	diabetes	{ pravastatin , fluvastatin , atorvastatin }	7.827803	
25	lovastatin	diabetes	{ p38 , kinase , mitogen }	7.691954	
26	lovastatin	pancreatic cancer	{ simvastatin , pravastatin , lovastatin }	7.445238	
27	lovastatin	breast cancer	{ antibodies , beta , alpha }	7.269621	
28	lovastatin	breast cancer	{ ras , kinase , mitogen }	7.174084	
29	glimepiride	obesity	{ adipocyte , adipocytes , insulin }	6.709126	
30	lovastatin	pancreatic cancer	{ simvastatin , statins , lovastatin }	6.469714	
31	lovastatin	breast cancer	{ beta , alpha , integrin }	6.419386	

Sheet1

Ready Count=2692

Pathway topology



Complement pathway from the KEGG online database

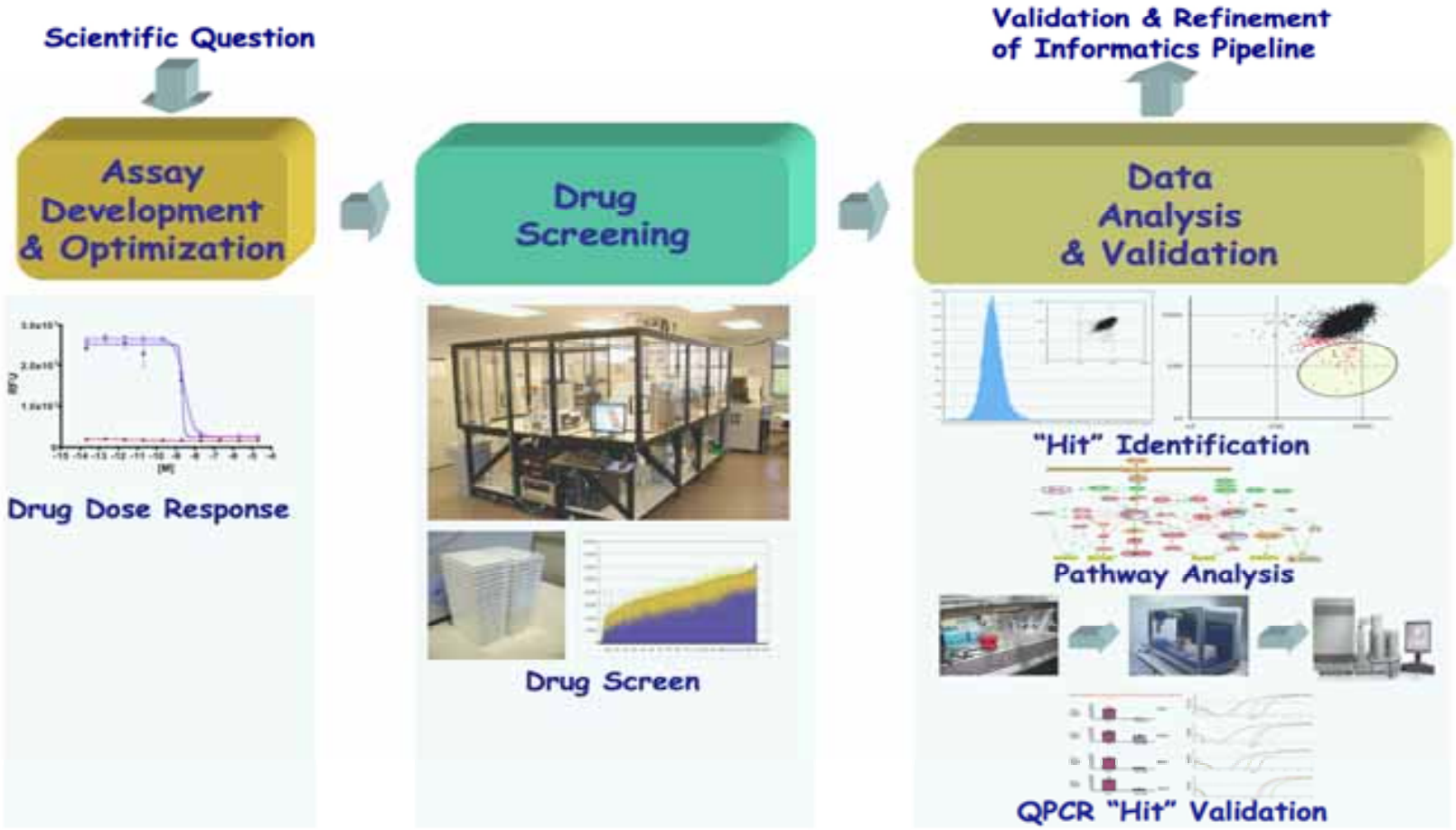
Sample Matches between DRUGS & PATHWAYS

Microsoft Excel - DiseaseDrug-Results092908.xlsx

	A	B	C	D	E
1	lovastatin	MAPK signaling pathway	hsa04010	16	ATF4,CD14,CDC42,CRKL,EGF,EGFR,FAS,FGF3,MYC,NF1,NGF,RAC1,RAP1A,RAP1B,SRF
2	lovastatin	MAPK signaling pathway	GM_Human_MAPK signaling pathway	15	ATF4,CD14,CDC42,CRKL,EGF,EGFR,FAS,MYC,NF1,NGF,RAC1,RAP1A,RAP1B,SRF,TNF
3	lovastatin	EGFR1 Signaling Pathway	GM_Human_EGFR1 Signaling Pathway	13	CBL,CDC42,CREB1,CRKL,EGF,EGFR,MYC,RAC1,RALB,SMAD3,SP1,SRC,STAT1
4	lovastatin	IL-3 Signaling Pathway	GM_Human_IL-3 Signaling Pathway	12	BAX,BCL2,CBL,CDC42,CREB1,CRKL,LYN,RAC1,RAP1A,SRC,STAT1,STAT6
5	lovastatin	Adipogenesis Human	GM_Human_Adipogenesis Human	12	BMP2,CREB1,FAS,GATA3,LIF,LPL,OSM,SMAD3,SP1,STAT1,STAT6,TNF
6	lovastatin	B Cell Receptor Signaling Pathway	GM_Human_B Cell Receptor Signaling Pathway	11	BAX,BCL2,CBL,CDK2,CDK4,CDK6,CREB1,CRKL,LYN,RHOA,STAT1
7	lovastatin	Androgen Receptor Signaling Pathway	GM_Human_Androgen Receptor Signaling Pathway	11	AES,AR,BRCA1,CDC2,EGFR,PTEN,RUNX2,SMAD3,SMAD4,SP1,SRC
8	lovastatin	TGF-beta Receptor Signaling Pathway	GM_Human_TGF-beta Receptor Signaling Pathway	11	AR,ATF3,CDC2,CDK2,CDK4,CDK6,MYC,RUNX2,SMAD3,SMAD4,SP1
9	lovastatin	Focal adhesion	hsa04510	11	BCL2,CDC42,CRKL,EGF,EGFR,PTEN,RAC1,RAP1A,RAP1B,RHOA,SRC
10	lovastatin	Cytokine-cytokine receptor interaction	hsa04060	10	CD40,EGF,EGFR,EPO,EPOR,FAS,LIF,OSM,PRL,TNF
11	lovastatin	gamma-Hexachlorocyclohexane degradation	hsa00361	10	CYP1A2,CYP2B6,CYP2C8,CYP2C9,CYP2D6,CYP3A4,CYP3A5,PON1,PON2,PON3
12	lovastatin	Pancreatic cancer	hsa05212	10	CDC42,CDK4,CDK6,EGF,EGFR,RAC1,RALA,RALB,SMAD3,SMAD4
13	lovastatin	T Cell Receptor Signaling Pathway	GM_Human_T Cell Receptor Signaling Pathway	10	CBL,CD2,CD4,CDC42,CREB1,CRKL,LYN,RAP1A,SRC,STAT1
14	lovastatin	Wnt Signaling Pathway	GM_Human_Wnt Signaling Pathway	9	CDC2,LDLR,LRP1,MYC,RAC1,RHOA,RUNX2,SMAD3,SMAD4
15	lovastatin	Genes specific to internal organs	GM_Human_Genes specific to internal organs	8	APOA1,APOB,EGFR,LPA,MST1,PON1,TF,TFPI
16	lovastatin	Arachidonic acid metabolism	hsa00590	8	CYP1A2,CYP2B6,CYP2C8,CYP2C9,CYP2D6,CYP3A4,CYP3A5,CYP4F2
17	lovastatin	p53 signaling pathway	hsa04115	8	BAX,BID,CDC2,CDK2,CDK4,CDK6,FAS,PTEN
18	lovastatin	Tryptophan metabolism	hsa00380	8	CAT,CYP1A2,CYP2B6,CYP2C8,CYP2C9,CYP2D6,CYP3A4,CYP3A5
19	lovastatin	Caffeine metabolism	hsa00232	7	CYP1A2,CYP2B6,CYP2C8,CYP2C9,CYP2D6,CYP3A4,CYP3A5
20	lovastatin	Androgen and estrogen metabolism	hsa00150	7	CYP1A2,CYP2B6,CYP2C8,CYP2C9,CYP2D6,CYP3A4,CYP3A5
21	lovastatin	Metabolism of xenobiotics by cytochrome P450	hsa00980	7	CYP1A2,CYP2B6,CYP2C8,CYP2C9,CYP2D6,CYP3A4,CYP3A5
22	lovastatin	Linoleic acid metabolism	hsa00591	7	CYP1A2,CYP2B6,CYP2C8,CYP2C9,CYP2D6,CYP3A4,CYP3A5
23	lovastatin	Cell cycle	hsa04110	7	CDC2,CDK2,CDK4,CDK6,PCNA,SMAD3,SMAD4
24	lovastatin	TGF-beta signaling pathway	hsa04350	7	BMP2,MYC,RHOA,SMAD3,SMAD4,SP1,TNF
25	lovastatin	Statin Pathway (PharmGKB)	GM_Human_Statin Pathway (PharmGKB)	7	ABCA1,APOA1,APOE,CETP,LDLR,LPL,LRP1
26	lovastatin	TGF Beta Signaling Pathway	GM_Human_TGF Beta Signaling Pathway	7	EGF,LIF,RUNX2,SMAD3,SMAD4,STAT1,TNF
27	lovastatin	Id Signaling Pathway	GM_Human_Id Signaling Pathway	7	ATF3,BMP2,CDK2,MYF5,NGF,SMAD3,SMAD4
28	lovastatin	Cell cycle	GM_Human_Cell cycle	7	CDC2,CDK2,CDK4,CDK6,PCNA,SMAD3,SMAD4
29	lovastatin	Apoptosis Mechanisms	GM_Human_Apoptosis Mechanisms	6	BAX,BCL2,BID,FAS,MYC,TNF
30	lovastatin	Mechanism of Gene Regulation by Peroxisome Prolifer	BC_Human_Mechanism of Gene Regulation by Peroxis	6	APOA1,CD36,LPL,MYC,SP1,TNF
31	lovastatin	IL-2 Signaling Pathway	GM_Human_IL-2 Signaling Pathway	6	BCL2,CBL,CREB1,CRKL,LYN,STAT1
32	lovastatin	IL-5 Signaling Pathway	GM_Human_IL-5 Signaling Pathway	6	BAX,CBL,CRKL,LYN,RAC1,STAT1
33	lovastatin	IL-7 Signaling Pathway	GM_Human_IL-7 Signaling Pathway	6	BAX,CBL,CDK2,CDK4,LYN,STAT1
34	lovastatin	G1 to S cell cycle control	GM_Human_G1 to S cell cycle control	6	CDK2,CDK4,CDK6,CREB1,MYC,PCNA
35	lovastatin	Delta-Notch Signaling Pathway	GM_Human_Delta-Notch Signaling Pathway	6	APP,CDK2,EGF,EGFR,SMAD3,SMAD4

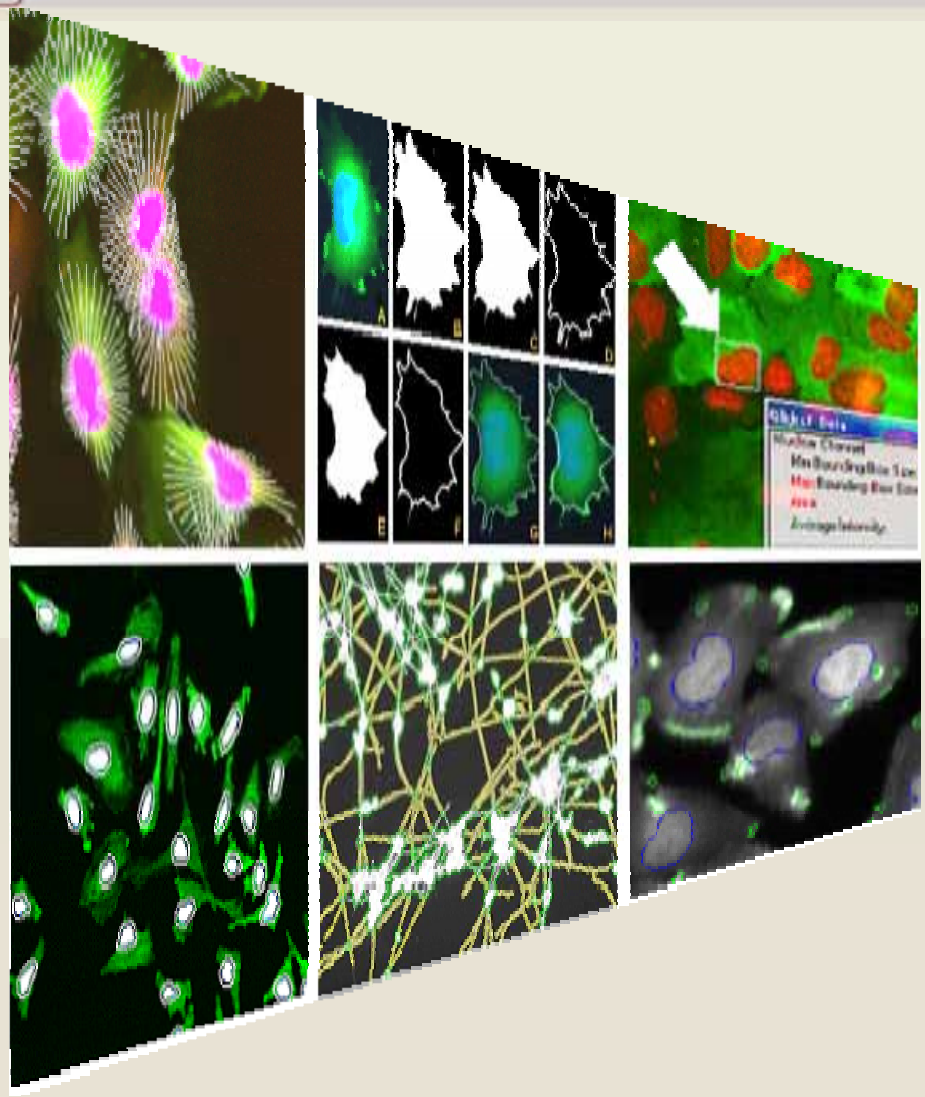
Pathway Counts / Pathway Genes

Flow of information in a Compound Library (Drug) Screening and qPCR Validation Study



Applied Biosystem's 7900HT Fast Real-Time PCR System

Image analysis for cell-based assays



- Object Intensity
- Nuclear Trafficking
- Plasma Membrane Spot
- Granularity
- Cell Cycle Trafficking
- Morphology
- Cell Viability
- Dual Object
- Neurite Outgrowth
- Micro-Nuclei

Courtesy of GE Healthcare

Team

- Dr. Hasan Davulcu Professor of Computer Science at ASU; CEO of uSuggest.com; 8 years of experience into e-commerce and data mining. Dr Davulcu was previously funded by US Army, DOD, NSF.

- Dr. Hung V. Nguyen Vice CEO PetroVietnam – PAIC and Senior Scientist at Information Technology Institute (ITI), Vietnam National University. Dr. Nguyen's expertise is in data mining which utilize probabilistic graphical models.

- Dr. Sukru Tuzmen is an Investigator and Head of the Molecular Genetics Laboratory in the Pharmaceutical Genomics Division at the Translational Genomics Research Institute (TGen).

Conclusion

- **The innovative, and unconventional nature of the research is that it lends itself to use in *existing drugs*, which can be easily and cheaply translated into improved treatments for a wide spectrum of diseases with enormous social impact.**