

# Cross Discipline Analysis made possible with Data Pipelining

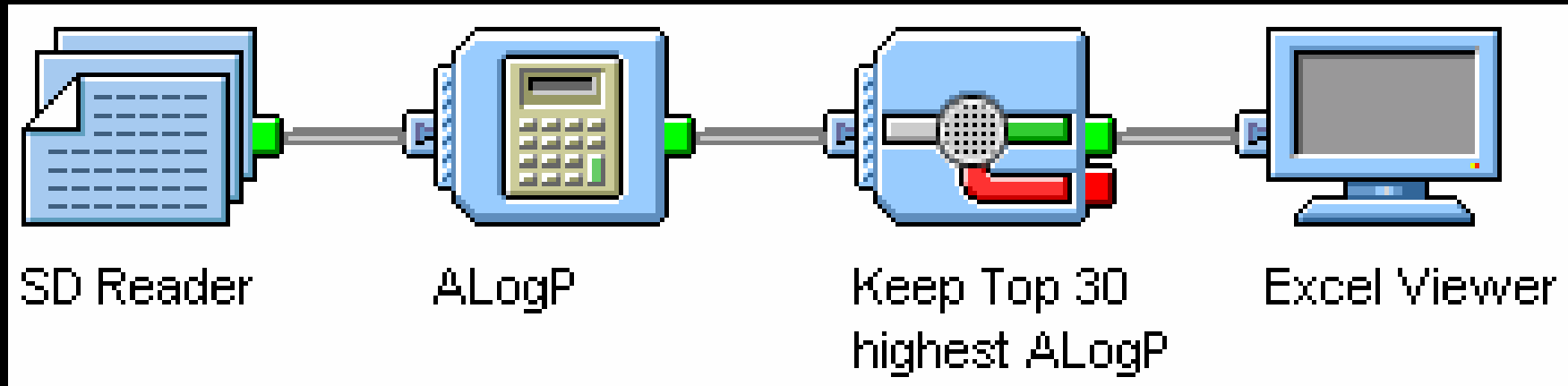
J.R. Tozer

SciTegic

## System Genesis

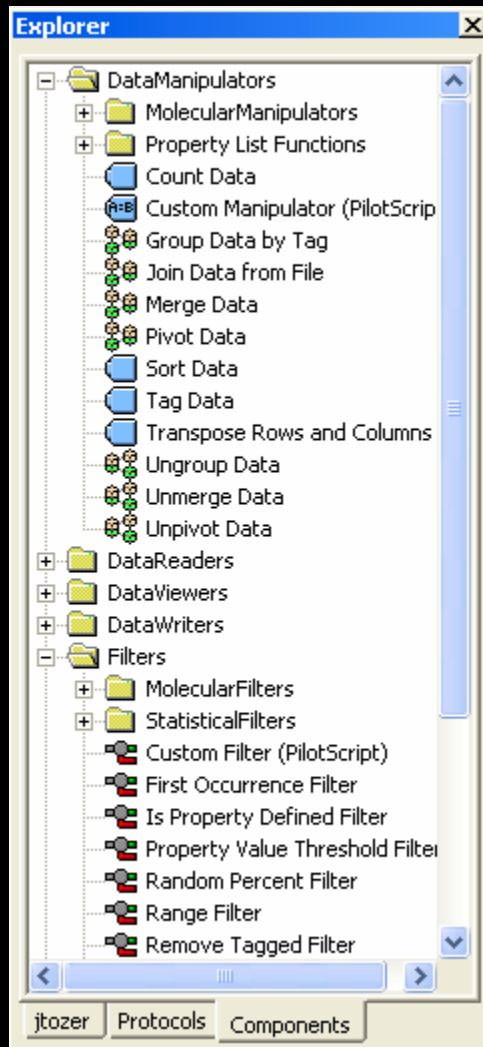
- Pipelining tool created to automate data processing in cheminformatics
- Modular system built with generic data types to accommodate diversity and change in the industry
- Validated in cheminformatics
  - Approach equally applicable to bioinformatics and beyond...

## Data Pipelining is...



- Graphical linking of independent modular steps
  - Readers
  - Calculators
  - Filters
  - Viewers
  - etc

## Data Pipelining is...



- Nearly 300 different “components” available
- Easily extensible through integration components
  - SOAP
  - ODBC
  - Perl
  - VBScript
  - Run Command-line program

## Chemo-Genomics

Ways of joining cheminformatics and bioinformatics:

### 1. Experimental data

- E.g., NCI cancer cell screen: Gene expression and compound activity data for the same set of cells

### 2. Pathways

- E.g., Kegg pathway database relates genes and compounds

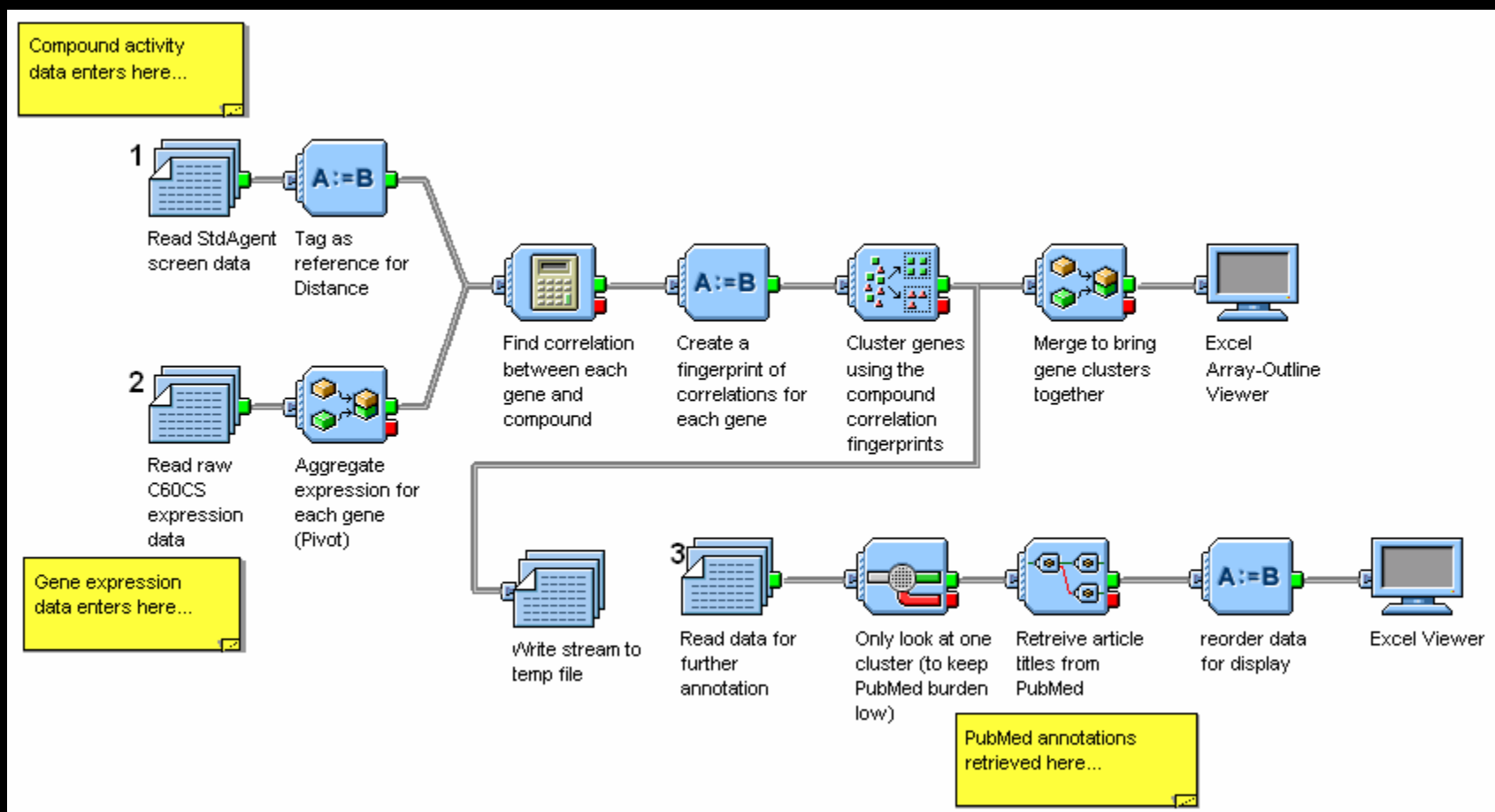
### 3. Protein-Ligand interactions

- E.g., Virtual docking technologies match genes with compounds that might bind

## Clustering Genes by Compound Activity

- NCI 60 Cancer Cell Screen
  - Expression levels for 10K genes
  - Growth inhibiting activity of 30K compounds
- Cluster genes according to the correlation of their expression level with the inhibiting activity of compounds.
  - i.e. If Genes A and B are always over-expressed in cells that are responsive to compounds X and Y, they would cluster together.
- Annotate cluster members with PubMed references

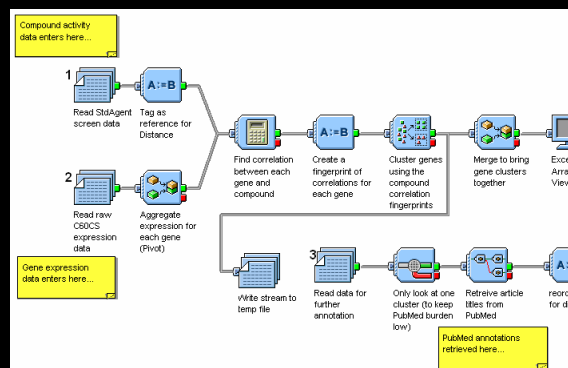
# Clustering Genes by Compound Activity



# Clustering Genes by Compound Activity

	1	2	3	4	5	6	7	8
	Cluster	GeneID	GeneName					
1	1	351 items	351 items					
2	2	553 items	553 items					
3	3	2 items	2 items					
4	4	37 items	37 items					
5	5	15 items	15 items					
950		GC11284	dual specificity phosphatase 3	The First Green Lineage cdc25 Dual-Specificity Phosphatase 3				
951		GC18321	dual specificity phosphatase 3	The First Green Lineage cdc25 Dual-Specificity Phosphatase 3				
952		GC16967	dual specificity phosphatase 3	The First Green Lineage cdc25 Dual-Specificity Phosphatase 3				
953		GC9724	zyxin	Zyxin interacts with the SH3 domains of the cytoskeletal proteins				
954		GC19002	phosphatidylinositol glycan, class F	A new aspect of the molecular pathogenesis of paroxysmal nocturnal hemoglobinuria				
955		GC13845	hemoglobin, zeta	Antisickling effects of an endogenous human alpha-like hemoglobin				
956		GC10289	annexin A5	Induction of cell death in activated hepatic stellate cells by annexin A5				
957		GC16514	integrin, alpha 5 (fibronectin receptor, alpha 5)	Extracellular fibrinogen binding protein, Efb, from Staphylococcus aureus				
958		GC10782	interferon gamma receptor 2	Interferon-gamma-mediated growth regulation of melanocytes				
959		GC14440	fragile X mental retardation 1	Metabotropic glutamate receptor activation regulates fragile X mental retardation 1				
960		GC10250	phosphoglucomutase 1	Increased fatty acid production in potato by engineering phosphoglucomutase 1				
961		GC18566	syndecan binding protein (syntenin)	PDZ tandem of human syntenin: crystal structure and function				
962		GC14515	protein tyrosine phosphatase, receptor type 11	Beyond the Metabolic Function of PTP1B				
963		GC16118	paired mesoderm homeo box 1	Multiple roles of Hoxa11 and Hoxd11 in the formation of the vertebrate tail				
964		GC17662	frizzled-related protein	Secreted Frizzled-related Protein 2 (SFRP2) is Highly Expressed in the Developing Brain				
965	6	3 items	3 items					
969	7	4 items	4 items					
974	8	9 items	9 items					

# Clustering Genes by Compound Activity



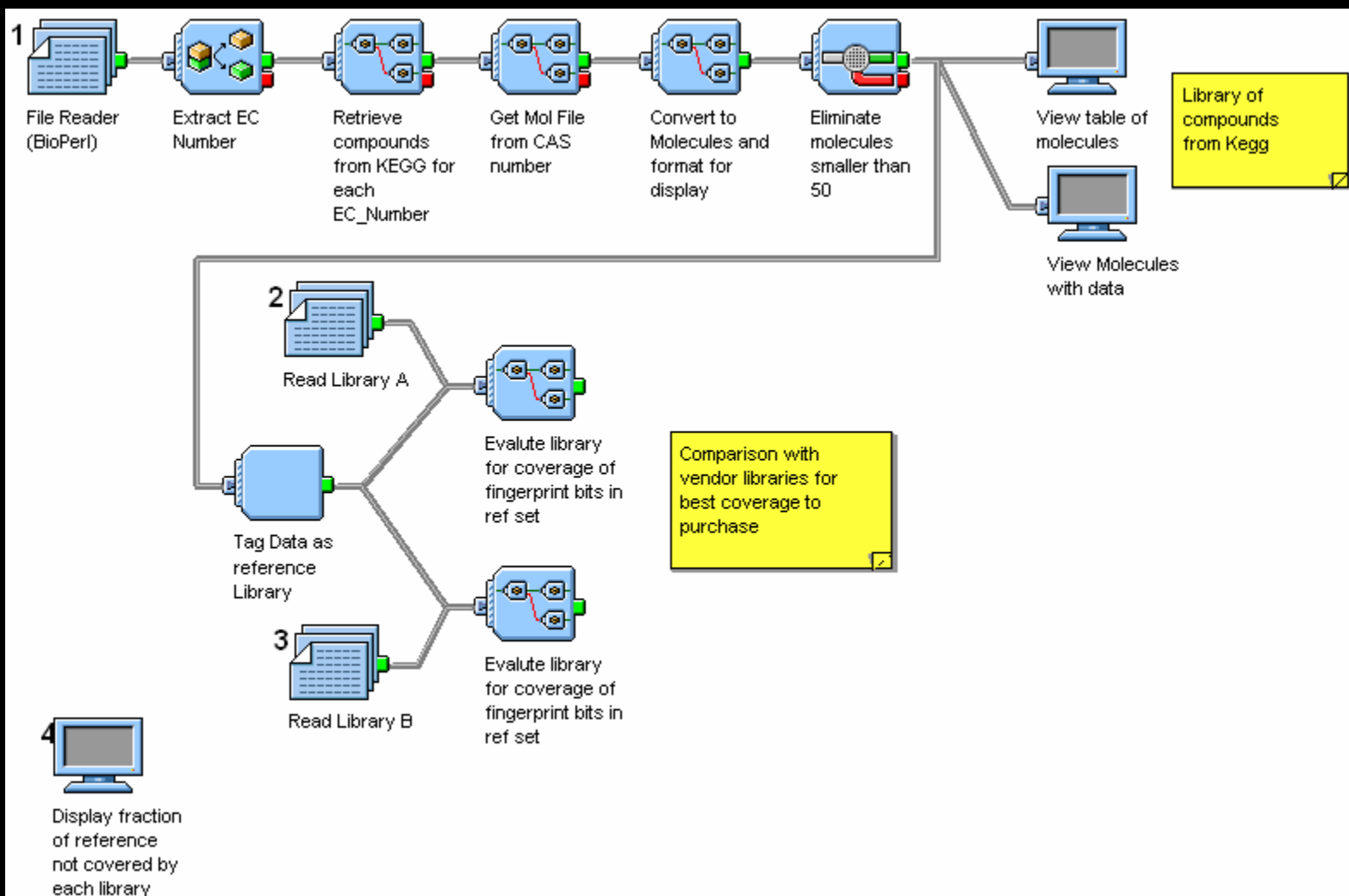
Cluster	GeneID	GeneName
1		
2	1	351 items
354	2	553 items
908	3	2 items
911	4	37 items
940	5	15 items
950	GC11284	dual specificity phosphatase 3 The First Green Lineage cdc25 Dual-Specificity Phos
951	GC18321	dual specificity phosphatase 3 The First Green Lineage cdc25 Dual-Specificity Phos
952	GC16967	dual specificity phosphatase 3 The First Green Lineage cdc25 Dual-Specificity Phos
953	DC9724	zysin Zysin interacts with the SHG domains of the cytoskel
954	GC19002	phosphatidylinositol glycan, class F A new aspect of the molecular pathogenesis of parr
955	GC13045	hemoglobin, zeta Antisickling effects of an endogenous human alpha-lik
956	GC10269	annexin A5 Induction of cell death in activated hepatic stellate cel
957	GC16514	integrin, alpha 5 (fibronectin receptor) Extracellular fibronectin binding protein, Efb, from Shp
958	GC10782	interferon gamma receptor 2 Interferon-gamma-mediated growth regulation of melar
959	GC14440	fragile X mental retardation 1 Metabotropic glutamate receptor activation regulates f
960	GC10250	phospholipase A2 Increased fatty acid production in potato by engineer
961	GC18956	syndecan binding protein (synterin) PDZ tandem of human synterin: crystal structure and
962	GC14515	protein tyrosine phosphatase, receptor type 2 Beyond the Metabolic Functions of PTP1B
963	GC16118	paired mesoderm homeobox 1 Multiple roles of Hoxa11 and Hoxd11 in the formation
964	GC17652	frizzled-related protein Secreted Frizzled-related Protein 2 (SFRP2) is Highly
965	6	3 items
969	7	4 items
974	8	9 items

- 10K genes x 125 compounds x 60 cells
- Protocol composed in an afternoon
- Executes in <5 minutes

## Assemble a library of compounds known to interact with genes of interest

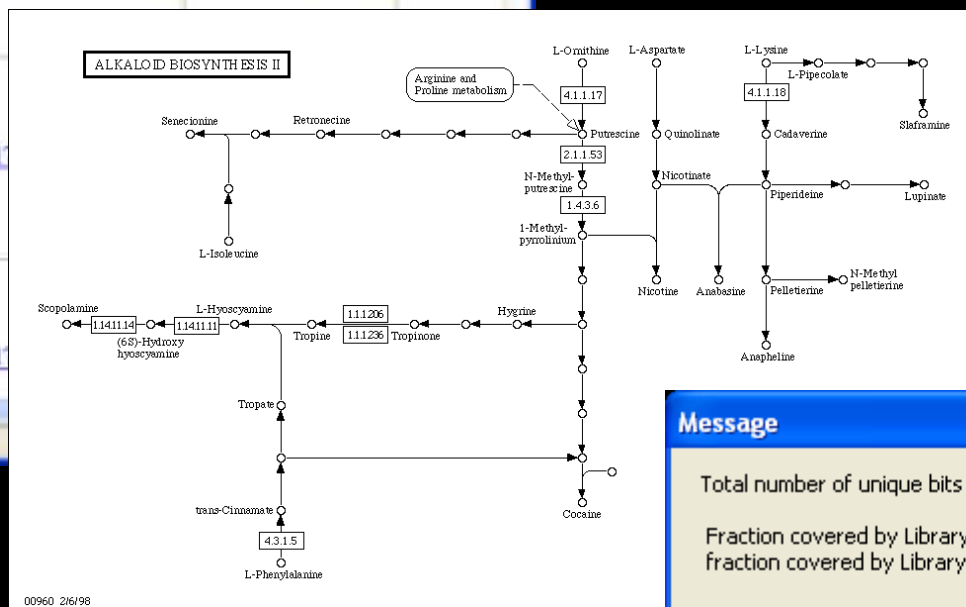
- Kegg database catalogs all the genes and small molecules involved in known pathways
  - Provides a bridge: Gene <-> Pathways <-> Compounds
- Given a list of interesting gene targets, assemble all the compounds that interact in the pathways they are involved in.
- Or, given a list of interesting compounds, determine what genes are involved in the pathways that contain them.

# Assemble a library of compounds known to interact with genes of interest



# Assemble a library of compounds known to interact with genes of interest

Molecule	Name	CAS number	KEGG Compound ID	KEGG Pathway ID	KEGG P
	DI-anabasine,	<a href="#">13078-04-1</a>	<a href="#">C06180</a>	<a href="#">MAP00960</a>	<a href="#">MAP00960</a>
	Hygrine,	<a href="#">496-49-1</a>	<a href="#">C06181</a>		
	Senecionine,	<a href="#">130-01-8</a>	<a href="#">C06182</a>		



**Message** ✕

Total number of unique bits in Reference library: 755

Fraction covered by Library A: 0.62  
fraction covered by Library B: 0.57

OK

## Find targets that might respond to compounds of interest

- Docking programs can evaluate the fit of compounds into binding pockets of proteins
- Given a list of interesting compounds, find which might bind to protein targets with known structures

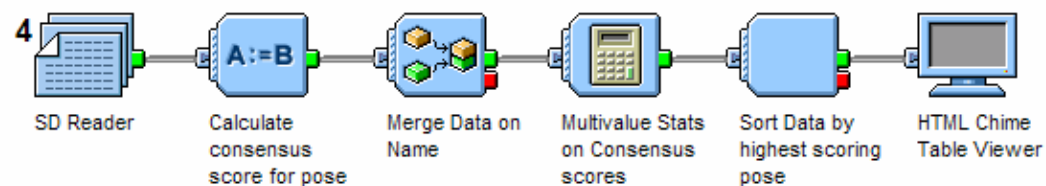
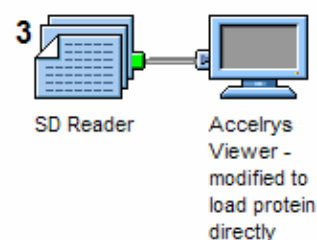
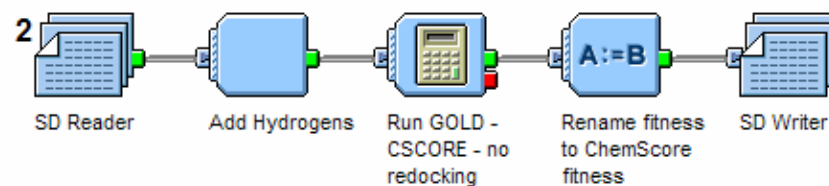
# Automated Protein-Ligand Docking

*Retrieve, prepare, dock & score ligands*

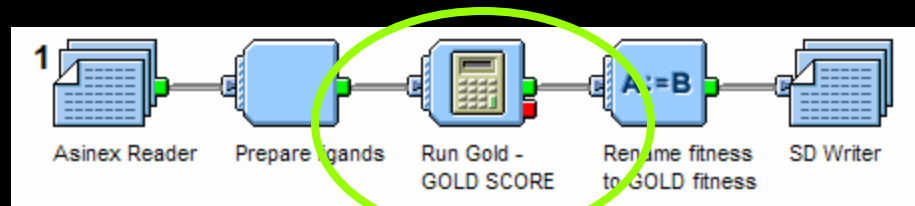
*Calculate second score*

*Display docked conformers in protein*

*Calculate consensus score for ligands and display statistics*



# Gold is integrated as a SOAP service



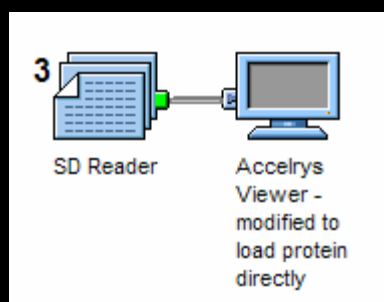
Parameters for: Run Gold - GOLD SCORE

Parameter Name	Parameter Value
SOAP Endpoint	http://localhost:8080/axis/services/GOLD
BatchSize	21
ProteinFile	C:\Documents and Settings\rbrown\Desktop\protein.mol2
NumberOfPoses	3
Settings	Library screening
UseChemScore	False
cavity_radius	14
cavity_origin_x	40.408
cavity_origin_y	8.195
cavity_origin_z	20.151

Parameters Error Handling Information

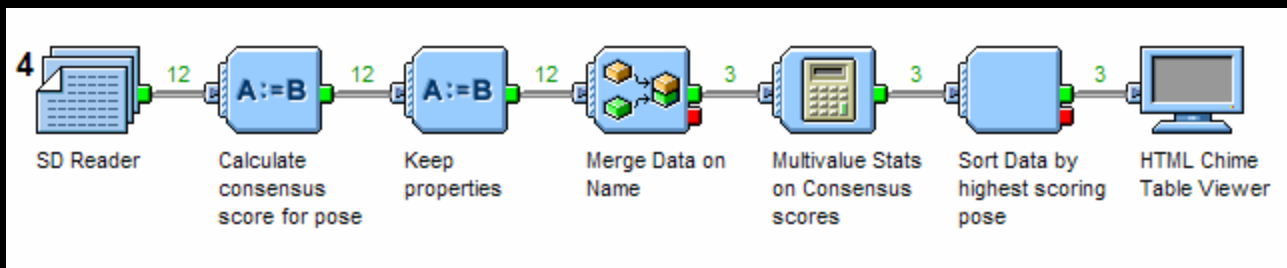
- Parameters exposed
- Server location
  - Batch size
  - Protein file
  - Mode
  - Scoring
  - Active site location

## Display resulting poses



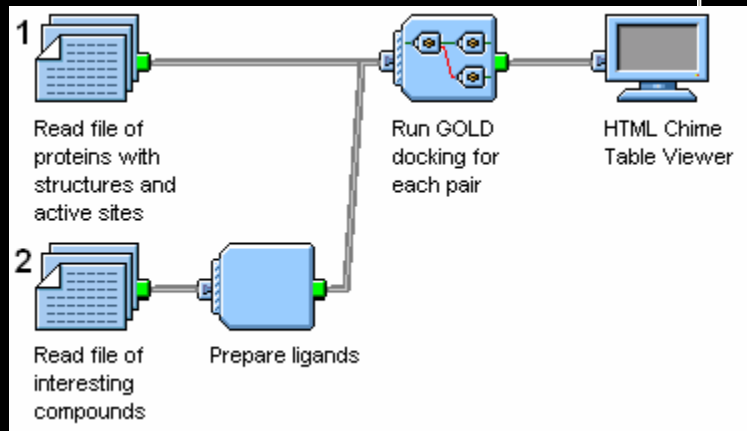
- DS.Viewer integration is an out-of-the-box component
- Uses VB/COM automation method

# Docking Results



The Structure	Name	Fitness_GOLD	Fitness_CScore	Consensus	Consensus_Mean	Consensus_StdDev	Consensus_N	Consensus_Max
	BAS 0000636	29.03 25.98 16.63 30.89 26.82 23.22	18.03 10.36 6.83 18.70 14.91 8.63	23.53 18.17 11.73 24.795 20.865 15.925	19.1691666666667	4.47971112969974	6	24.795
	BAS 0000525	26.96 22.63 22.15	16.96 14.24 21.22	21.96 18.435 21.685	20.6933333333333	1.60082444037095	3	21.96
	BAS 0000725	16.84 12.93 11.42	6.52 3.13 8.82	11.68 8.03 10.12	9.94333333333333	1.49533348194382	3	11.68

# High-Throughput Docking Results



The Structure	Name	Consensus_Max
	BAS 0000636	24.795
	BAS 0000525	21.96
	BAS 0000725	11.68

## Next Steps...

- Data Pipelining provides an ideal framework for linking the data and tools from different scientific domains.
- By eliminating the data and tool integration barriers, the challenge is to think of good scientific questions...
- Looking for collaborators
  - Unique data collections
  - Unique domain-bridging ideas

# Thank You

- Acknowledgements:
  - David Rogers (fingerprints and clustering)
  - Scott Markel (Kegg integration)
  - Rob Brown, Andrei Caracoti (Gold Integration)
- Questions?
  - (See more at SciTegic's Booth #127-129)