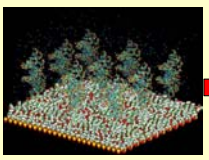
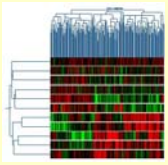
Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate

Sorin Draghici
Dept. of Computer Science
Wayne State University
Detroit, Michigan

WAYNE STATE UNIVERSITY

Onto-Express





AC2003
AG2038
GN3289
.
.
.

- Result of a microarray experiment is a list of differentially expressed genes.
- We would like to understand the underlying biological phenomenon affected by the set of genes.

WAYNE STATE UNIVERSITY

Biological processes

Gene 1
Apoptosis
Cell-cell signaling
Protein phosphorylation
Mitosis
…

Gene 2
Growth control
Mitosis
Oncogenesis
Protein phosphorylation
…

Gene 3
Gro
Mit
Onc
Pro
…

Gene 4
Nervous system
Pregnancy
Oncogenesis
Mitosis
…

• • •

Gene 100
Positive ctrl. of cell prolif
Mitosis
Oncogenesis
Glucose transport
…

WAYNE STATE UNIVERSITY

Biological processes

Apoptosis
Gene 1
Gene 53

Mitosis
Gene 2
Gene 5
Gene45
Gene 7
Gene 35
…

Glucose transport
Gene 7
Gene 3
Gene 6
…

Positive ctrl. of cell prolif.
Gene 7
Gene 3
Gene 12
…

• • •

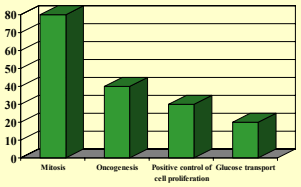
Oncogenesis
Gene 5
Gene 2
Gene 6
…

WAYNE STATE UNIVERSITY

Data mining results – first shot

Microarray (1000 genes) → Exp. → Data Analysis → Reg. genes (100 genes) → Onto-Express

mitosis – 80/100
oncogenesis – 40/100
p. ctrl. cell prol. – 30/100
glucose transp. – 20/100



Cancer?

WAYNE STATE UNIVERSITY

Are we jumping to conclusions?

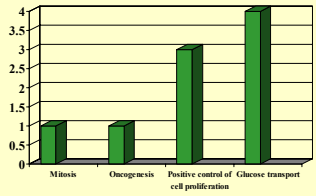
- What would happen if all genes on the array are involved in mitosis? Would then mitosis still be relevant?
- We need to see the results in the context of how many genes of each type there are on the array.

WAYNE STATE UNIVERSITY

Functional representation on the array

Function	Genes on array	# genes expected in 100 random genes	occurred
mitosis	800/1000	80	80
oncogenesis	400/1000	40	40
p. ctrl. cell prol.	100/1000	10	30
glucose transp.	50/1000	5	20

Occurred vs. expected



WAYNE STATE UNIVERSITY

The problem

- We expected 5 genes in glucose transport and we got 20. This is 4 times more than expected but it can still happen just by chance!!!
- What if we had 10 times more than expected. It can still happen just by chance but the probability of this happening is much lower.
- Question: what is the probability of obtaining the observed result just by chance?
- Can we calculate a confidence value associated with each category?

WAYNE STATE UNIVERSITY

Onto-Express Input

- Onto-Express accepts an input file containing
  - GenBank accession numbers
  - or UniGene cluster IDs
  - or Affymetrix probe IDs
  - or WormBase accession numbers (for C. elegans only).
- The file contains one identifier per line.
- The input file must be created in ASCII editor such as notepad.

WAYNE STATE UNIVERSITY

Biological processes affected in breast cancer

Veers, 415, Jan., 530—536, Nature, 2002

WAYNE STATE UNIVERSITY

Onto-Express features

- Organisms supported:
  - Human, Mouse, Rat, Drosophila, C. Elegans
  - Comparative genomics approach for others
- Probability distributions:
  - Hyper-geometric distribution
  - Binomial distribution
  - Chi-square distribution
  - Fisher's exact test
- Corrections for Multiple Experiment:
  - Bonferroni correction
  - Holmes correction
  - False Discovery Rate (FDR) correction
  - Sidak correction
- Integrated GO browser
- Interactive graphics
- 89,735 lines of code as of January 2004

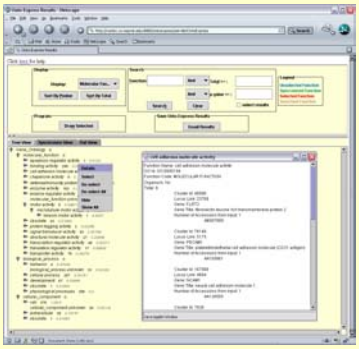
WAYNE STATE UNIVERSITY

OE output - flat view

- Flat view does not represent the hierarchical structure of GO.
- Clicking a left mouse button or selecting "details" from the pop-up menu shows the details for the corresponding function in another window.
- The graph can be sorted according to p-value, total or function name.
- It only displays results from one category at a time. In order to move to another category, select the desired category from drop-down list.

WAYNE STATE UNIVERSITY

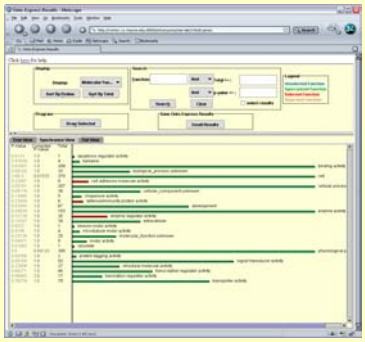
OE output – GO hierarchy



- Clicking a left mouse button on a node with non-zero total, displays details about the node, including the GO id, list of genes from the input file for the node, their accession numbers, UG cluster IDs, LocusLink IDs etc.
- Clicking a right mouse button on a node shows a pop-up menu that allows to select/deselect or show/hide a specific node. The results of these operations are shown in synchronized view.

WAYNE STATE UNIVERSITY

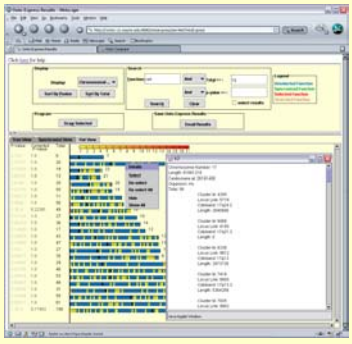
Synchronized View



- For each node with non-zero total in the tree view, there is a bar of proportional length.
- Only the nodes with non-zero total are displayed in the sync. view.
- The nodes selected in the tree view are highlighted in red color.
- The graph can be sorted by function name, total or p-value.
- Clicking the “draw selected” button in tree view will only show the bars in red.

WAYNE STATE UNIVERSITY

Chromosome view



- The chromosome view displays the number of UniGene clusters mapped on a particular cytogenetic location.
- The pop-up menu only shows if a right mouse button is clicked in the blue colored area.

WAYNE STATE UNIVERSITY

Onto-Compare

Choosing the best array for a given biological phenomenon

WAYNE STATE UNIVERSITY

Array bias

- Each array displays a different set of genes.
- Each set of gene represents different pathways to a different extent.
- Each array has a biological bias: some pathways/biological processes might be represented better, some might be represented more poorly.
- In a hypothesis driven scenario, there is a hypothesized biological mechanism in action. One should use the array or the combination of arrays that are most relevant for the phenomenon studied.

WAYNE STATE UNIVERSITY

Available choices

- ClonTech human apoptosis - 206 genes
- Perkin-Elmer apoptosis - 324 genes
- Sigma-Genosys human apoptosis– 198 genes

- ClonTech + Perkin-Elmer + Sigma-Genosys = 74
- ClonTech + Perkin-Elmer = 167
- Perkin-Elmer + Sigma-Genosys = 92
- Clontech + Sigma-Genosys = 92

WAYNE STATE UNIVERSITY

Apoptosis arrays comparison

Ontology Term	ClonTech	PE	Sig-Gen
Total genes on array	214	346	210
induction of apoptosis	16 [16]	27 [28]	23 [23]
anti-apoptosis	15 [15]	20[20]	21[21]
immune response	0 [0]	1 [1]	19 [19]
cell-cell signaling	9 [9]	9 [9]	18 [18]
cell surface receptor linked signal transduction	4 [4]	9 [9]	17 [17]
oncogenesis	22 [22]	28 [28]	16 [16]
regulation of cell cycle	30 [30]	30 [30]	12 [12]
positive regulation of cell proliferation	5 [5]	5 [5]	12 [12]
negative regulation of cell proliferation	16 [16]	20 [20]	10 [10]
induction of apoptosis by DNA damage	3 [3]	4 [4]	3 [3]
induction of apoptosis by extracellular signals	8 [8]	12 [12]	7 [7]
induction of apoptosis by hormones	1 [1]	1 [1]	1 [1]
induction of apoptosis by intracellular signals	2 [2]	2 [2]	2 [2]
induction of apoptosis by oxidative stress	0 [0]	0 [0]	1 [1]
induction of apoptosis via death domain recept	4 [4]	5 [5]	7 [7]
caspases	11[10]	14[14]	13[13]
tumor necrosis factor receptor	2 [2]	2 [2]	2 [2]
tumor necrosis factor receptor ligand	1 [1]	1 [1]	1 [1]
tumor necrosis factor receptor, type I	1 [1]	1 [1]	1 [1]
interleukins & interleukins receptors	0 [0]	0 [0]	16 [16]
Unique Sequences [Clusters]	99 [98]	133 [132]	129 [129]

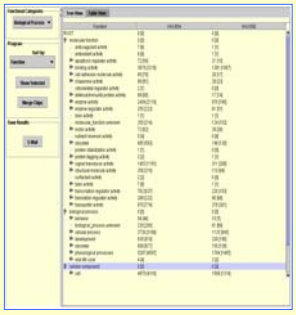
WAYNE STATE UNIVERSITY

Onto-Compare

- The arrays in the database are grouped by manufacturers and organism.
- A user can select any number of arrays by clicking the check box next to an array name.

WAYNE STATE UNIVERSITY

Onto-Compare Tree View



- Onto-Compare results are displayed in two different views: tree view and table view.
- Tree view reflects the tree structure of GO.
- The left-most column displays the name of an ontology term. The rest of the columns display total number of GenBank sequences found on a given array for the ontology term. The number in square brackets indicate the number of unique UniGene clusters for the term.

WAYNE STATE UNIVERSITY

Onto-Design

Choosing the best set of genes to construct a custom array for a given biological phenomenon

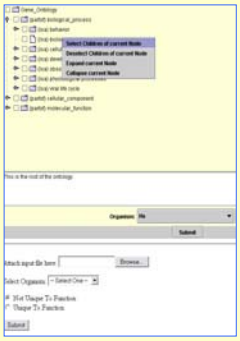
WAYNE STATE UNIVERSITY

Onto-Design

- Many focused commercial microarrays are available.
- However, given the complexity of the biological research, one may feel that none of the available microarrays represent the targeted biological processes and pathways to the extent needed.
- In such case, one may choose to design and print their own arrays.
- Printing custom array may also provide the ability to adapt the arrays to one's own experiment design and use of controls.

WAYNE STATE UNIVERSITY

Onto-Design Input Interface



- Input to OD is a list of ontology terms.
- A user can either select the terms by browsing the GO tree or can submit a file containing the terms one per line.
- A term can be selected by clicking the check box next to it.
- Clicking a right mouse button on a node displays a pop-up menu.
- "Select children of current node" expands the node recursively and selects all children of the node.
- "Expand current node" expands the current node without selecting any node.

WAYNE STATE UNIVERSITY

Onto-Design output

- The list of all known UG clusters for a given term is shown in the second column.
- A user can select check boxes next to the ontology terms and can carry out various set operations such as union, intersection and difference.
- Result of each operation is added as a new term in the table.
- After carrying out many set operations, the user can reset the view to original list by clicking "Reset Results".
- Clicking the "Unique to Function" button displays the list of UG clusters that are associated with one and only one term.

WAYNE STATE UNIVERSITY

Custom vs. commercial

Ontology Term	Sig-Gen	PE	ClonTech	Custom
Total genes on array	210 [206]	346 [324]	214 [198]	[229]/[250]
DNA fragmentation	4 [4]	3 [3]	1 [1]	[4]
DNA repair	4 [4]	9 [9]	6 [6]	[2]
I-kappaB phosphorylation	2 [2]	0 [0]	0 [0]	[0]/[2]
RAS protein signal transduction	1 [1]	3 [3]	3 [3]	[0]/[3]
anti-apoptosis	21 [21]	20 [20]	15 [15]	[53]/[58]
apoptosis	16 [16]	24 [24]	15 [15]	[79]/[85]
apoptotic program	7 [7]	7 [7]	8 [7]	[8]/[9]
caspase activation	1 [1]	2 [2]	1 [1]	[2]
cell death	0 [0]	1 [1]	0 [0]	[2]
cell motility	6 [6]	8 [7]	4 [4]	[3]/[4]
cell proliferation	20 [20]	19 [19]	21 [21]	[16]/[21]
cell-cell signaling	18 [18]	9 [9]	9 [9]	[24]/[25]
development	9 [9]	4 [4]	4 [4]	[11]/[13]
immune response	19 [19]	1 [1]	0 [0]	[9]/[10]
induction of apoptosis	23 [23]	27 [26]	16 [16]	[53]/[56]
induction of apoptosis by DNA damage	3 [3]	4 [4]	3 [3]	[5]/[6]
induction of apoptosis by extracellular signals	7 [7]	12 [12]	8 [8]	[18]
induction of apoptosis by hormones	1 [1]	1 [1]	1 [1]	[4]
induction of apoptosis by intracellular signals	2 [2]	2 [2]	2 [2]	[7]
induction of apoptosis by oxidative stress	0 [0]	0 [0]	1 [1]	[0]/[1]
induction of apoptosis via death domain recept	7 [7]	5 [5]	4 [4]	[8]
inflammatory response	8 [8]	4 [4]	2 [2]	[9]
killing transformed cells	0 [0]	1 [1]	0 [0]	[1]
killing virus-infected cells	0 [0]	1 [1]	0 [0]	[1]
negative regulation of survival gene products	1 [1]	2 [2]	2 [2]	[4]
neurogenesis	3 [3]	5 [5]	2 [2]	[8]
positive regulation of cell proliferation	12 [12]	5 [5]	5 [5]	[3]/[13]
proteolysis and peptidolysis	6 [6]	7 [7]	7 [6]	[7]/[8]
regulation of CDK activity	4 [4]	17 [17]	16 [16]	[2]/[9]
regulation of cell cycle	12 [12]	30 [30]	30 [30]	[16]/[32]
signal transduction	56 [56]	62 [60]	42 [42]	[55]/[57]

WAYNE STATE UNIVERSITY

Onto-Translate and Onto-Miner

WAYNE STATE UNIVERSITY

Question

- Paperino
- Aku Ankka
- Andres Ono
- Kale Ankka
- Donald Duck

- Different databases – different names
  - GeneBank - accessions
  - UniGene – cluster IDs
  - NetAffy – Affymetrix probe Ids

WAYNE STATE UNIVERSITY

Trivial? Not quite!!

- When annotating genomes same piece of information is stored and viewed differently across different databases.
- For example, more than one Affymetrix probe IDs can refer to the same GenBank sequence and more than one nucleotide sequence can be grouped in a single UG cluster.
  - How many different genes are represented on HGU133?
- The user has to be aware of these relationships between the different forms of the data in order to interpret the results correctly.
  - If the list of genes is submitted as accessions and the array contains 5 accession numbers corresponding to the same gene, the results will be skewed.
- Even if a user is aware of the relationships, the process of translating hundreds of genes one at a time is unfeasible.

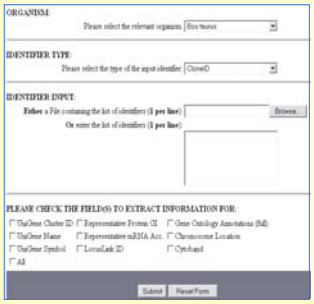
WAYNE STATE UNIVERSITY

Probes vs. UniGene

Probe IDs	On array	Selected	p value
apoptosis	170	18	0.105049
not apoptosis	830	82	
Total	1000	100	
UniGene clusters	On array	Selected	p value
apoptosis	110	15	0.044346
not apoptosis	785	72	
Total	895	87	

WAYNE STATE UNIVERSITY

Onto-Miner Input Interface



- Onto-Miner allows a user to query annotations for a set of genes.
- The user can either submit a list of genes in a file or can type the input in the text area.
- Input can be either clone ID, UG gene name, UG gene symbol, UG cluster ID, LocusLink ID or GenBank accession number.
- The results are returned as a tab delimited file.

WAYNE STATE UNIVERSITY

Conclusions

- Onto-Express can be used to interpret quickly the results of a microarray experiment.
- Onto-Compare can be used to select the best array or combination of arrays for a given biological hypothesis.
- Onto-Design can be used to design a custom array for a given set of biological hypotheses.
- Onto-Translate can be used to translate lists of genes between UniGene, GenBank accession and Affymetrix probe IDs.
- Onto-Miner can be used to gather quickly all annotations available about a known gene.

WAYNE STATE UNIVERSITY

Acknowledgements

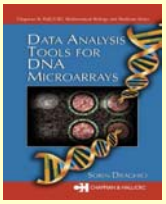
- National Science Foundation - DBI-0234806
- DoD / USAMRMC - DAMD 17-03-02-0035
- NIH / NCRR - 1 S10 RR017857-01
- NIH / NCI - 1 R21 CA10074001
- Sun Microsystems - EDU 7824-02344-US (equipment)
- Department of Health and Human Services - 1 R01 NS045207-01
- Michigan Life Science Corridor - MEDC GR-352
- NIH / NIBIB - 1 R21 EB00990-01
- Michigan Life Science Corridor - MEDC-538
- Michigan Life Science Corridor- GR-446

- Mott Center - Stephen Krawetz, Chuck Ostermeier, Rui Pires Martines
- Karmanos Cancer Institute – Michael Tainsky
- Dept. of Computer Science, WSU
  - Purvesh Khatri, Pratik Bhavsar, Valmik Desai

WAYNE STATE UNIVERSITY

More information

http://vortex.cs.wayne.edu
my lab web site
and access to all Onto-Tools
http://vortex.cs.wayne.edu/publications.htm
papers



•Onto-Tools, The toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. Nucleic Acids Research, 31(13):3775-81, July 2003.

•Sorin Draghici, Purvesh Khatri, Abhik Shah and Michael Tainsky. Assessing the functional bias of commercial microarrays using the Onto-Compare database. BioTechniques, Cancer and Microarrays, March 2003.

•Sorin Draghici, Purvesh Khatri, Rui P. Martins, G. Charles Ostermeier and Stephen A. Krawetz. Global functional profiling of gene expression. Genomics 81(2), Feb 2003.

•Khatri P., Draghici S., Ostermeier C., Krawetz S. - Profiling Gene Expression Utilizing Onto-Express. Genomics, 79(2), Feb 2002.

© Sorin Draghici (9/3/2001)

WAYNE STATE UNIVERSITY