



Sixth Framework Programme for Quality of Life and  
Management of Living Resources

Project no. LSHG-CT-2006-037686

# EMERALD

Empowering the Microarray-Based  
European Research Area to Take a Lead  
in Development and Exploitation

EU Deliverable: 1.6

Due Date: October 2008  
Delivery Date: 8th October 2008

Version 1

Partner responsible: EBI

# Report on a Normalisation and Transformation Ontology (NTO) (deliverable D1.6)

Delivery Date: 08 Oct 2008

Version 1.0

Partner responsible: EBI

Author: James Malone, Helen Parkinson

## Contents

WP1 Quality Metric and Ontologies work package .....	3
1. D1.6 A normalization and transformation ontology as part of the MGED ontology .....	3
1.1 Development of the NTO .....	3
1.2 Scoping the ontology .....	3
1.3 Cooperation with external efforts .....	4
1.4 Reuse of existing ontologies .....	5
1.5 The Normalisation and Transformation Ontology .....	5
1.6 Evaluation .....	8
1.7 Dissemination of the NTO .....	9
1.8 Assistance with Usage of the NTO .....	10
1.9 Future Work .....	10
D3.6 Workshop III Ontology Workshop .....	11
D3.7 Workshop III report .....	11
Dissemination activities: .....	11
Publications: .....	12
References .....	12
Appendix A – GenePattern Use Case .....	13

## WP1 Quality Metric and Ontologies work package

The aim of the Quality Metric and Ontologies work package is to develop and disseminate quality metrics and tools for determining data quality and describing transformations. An important component of this work is the development of an ontology for describing normalization and transformation processes used in the microarray community. This report describes the development of this normalization and transformation ontology (NTO) in fulfillment of deliverable 1.6.

<i>Deliverable</i>	<i>Title</i>	<i>Due Date</i>	<i>Nature</i>	<i>Dissemination level</i>	<i>Status</i>
D1.6	A Normalisation and Transformation Ontology developed as part of the MGED ontology	24	R	PU	Complete
D3.6	Workshop III: Ontology Workshop	12	O	PU	Complete
D3.7	Report on workshop III	14	R	PU	Complete

### **1. D1.6 A normalization and transformation ontology as part of the MGED ontology**

The diversity in microarray experiment designs and applications requires that a large number of pre-processing approaches are available. In order to facilitate unambiguous and consistent descriptions of experimental data transformation the development of a 'normalisation and transformation ontology' (NTO) is required. This will provide a means to conceptualize and classify the approaches used, describe relationships between these concepts and store these in a machine readable form. Considerable work has been performed in creating the NTO involving collaborative community development and use case driven evaluation.

#### **1.1 Development of the NTO**

In order to progress with the development of the NTO, a design process was set out. This involved the following steps:

- *Scoping ontology* – The use of competency questions, use cases and scenarios for which the ontology will be used help define the scope of the development.
- *Cooperation with external efforts* – Close cooperation with the OBI project (the successor to MGED Ontology) constitutes an important component in order to collaborate with the various ongoing community developments including the emerging OBO foundry and the National Center for Bioontology in the US both of which are coordinating international ontology development.
- *Reuse of existing ontologies* – The reuse of ontologies currently in existence, either by incorporation into the NTO or by a mapping to the ontology as an external resource, reduces redundant development effort.
- *Evaluation* – The evaluation process consists of checking requirements through competency questions and testing the ontology in the target application environment to ensure it is fit for purpose.

#### **1.2 Scoping the ontology**

Standard ontology development methodology to determine the scope and requirements of an ontology is through the use of competency questions and use cases. After identifying the intuitive main scenarios envisaged for an ontology, a set of natural language questions, called

competency questions are used to extract the main concepts and their properties, relations and axioms of the ontology (Gómez-Pérez *et al*, 2004).

Competency questions and use cases for data transformation were recruited from members of the biomedical and data analysis communities. An example of the types of competency questions submitted is shown below. They are worded as questions which we would expect the ontology to be able to answer. Terms in scope for the NTO are shown in italics.

- Which genes have a 2 fold change in expression where *MAS5* has been applied as a *data transformation methodology*?
- Which *pre-processed microarray data* expresses values as *log ratios* (of two conditions) for a *specified logarithmic base*?

An example of a textual use case is shown below. These are generally more detailed and are given from an actor's point of view in terms of how they would interact with the ontology:

- An experimenter has conducted an expression microarray experiment involving two conditions with replicate assays per condition, where they have both biological and technical replicates. They are running two kinds of *differential expression analyses*: (a) one at the gene level and (b) one at the gene set level. In (a) the aim is to identify differential expressed genes (e.g. via *algorithms like PaGE and SAM*). In (b) the aim is to identify, from an a priori given collection of gene sets (e.g. user provided, or based upon GO annotation), which of these sets are *differentially expressed* as a whole (e.g. via *algorithms like GSEA or SAM-GSA*). Before running the analyses the data is preprocessed with the following data transformation series: (i) *filter out flagged reporters*, (ii) *normalize* the individual assays, (iii) *average* across technical replicates (but not across biological replicates). The above steps all require annotation using the ontology.

As well as simple competency questions and use cases, larger use cases were compiled following interactions with community representatives. An example of this is given at the following link

<http://obi.svn.sourceforge.net/viewvc/obi/trunk/docs/developer/DT/GenePatternUseCases.xls>

In this case, there are many examples of data transformations, often described in the context of the Gene Pattern software implementation used. This study demonstrates the variety of data transformation and normalizations that are used in the microarray community and the link between some of the transformations and software. Such use cases offer a rich mechanism by which the requirements can be captured and, crucially, later evaluated.

Use case recruitment is ongoing for future development of the ontology and updates are posted publicly at:

<https://wiki.cbil.upenn.edu/obiwiki/index.php/EvaluationPhase1Submissions>

### 1.3 Cooperation with external efforts

The initial strategy of this project was for the project consortium to work closely with the MGED consortium to 'develop a component of the MGED Ontology that can be used to describe data transformations.' However, the MGED Ontology has since been subsumed into the Ontology for Biomedical Investigations (OBI) (<http://obi.sourceforge.net/>) which has become the focus of efforts for many previously working on the MGED Ontology. The OBI project has the much wider scope of being able to represent any (not just microarray) biological and medical experiment and investigation. The ontology aims to model the design of an investigation, the protocols and instrumentation used, the material used, the data generated and the type analysis performed on it.

Although the focus of OBI is broader than the MGED Ontology, domain specific concepts still play a major role in the gathering of terms from the different biomedical communities, of which transcriptomics is an important part. For this reason, coordination and cooperation with OBI Consortium is inline with the NTO aims. OBI is an OBO foundry ontology and has data

transformation terms within its scope. This means that any separate effort would likely not be accepted as an OBO foundry ontology as efforts should be orthogonal. Moreover, the OBI Consortium benefits from a diversity of membership as well as an existing infrastructure framework which would be compatible with the NTO. This has allowed us to access use cases and expertise from many individuals and has resulted in an improved ontology.

#### 1.4 Reuse of existing ontologies

There are several ontology efforts currently in existence which have been investigated for incorporation into the NTO or by a mapping to the ontology as an external resource during development. One clear example of this is reusing components of the MGED ontology on which parts of the NTO were to be integrated. As OBI supersedes the MGED ontology, the relevant components were integrated as classes into the OBI effort.

#### 1.5 The Normalisation and Transformation Ontology

In order to understand the structure of the NTO, it is useful to introduce the relevant parts of the larger ontology, OBI, of which the NTO is an integral part. The normalisation and transformation parts of OBI, labeled simply 'data transformation' (as normalization is a specialized transformation) for the purposes of the OBI collaboration, form one of the key branches upon which the ontology was developed. This structure enabled the data transformation ontology to be developed concurrently alongside other areas of the ontology and, crucially, allowed for the interoperability with other areas of OBI which enable the important links between other ontological concepts. This includes areas such as the protocols of an experiment, the materials used and the entities representing information.

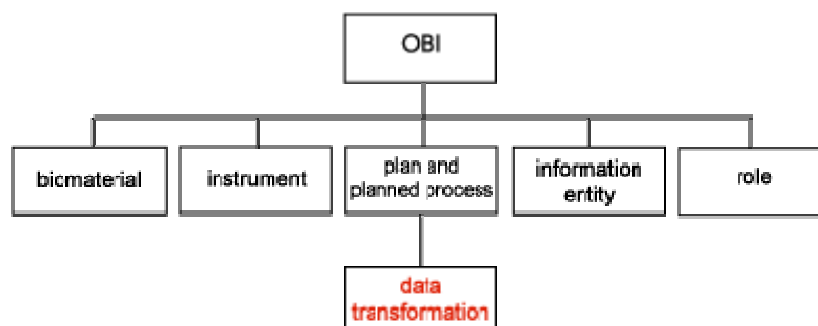


Figure 1. The Overall upper level OBI ontology structure with the NTO component data transformation branch highlighted in red.

An overview of the OBI structure can be seen in Figure 1. Here some of the key branch components which were used to devolve the work load are illustrated. Of interest here, data transformation (the NTO component) can be seen highlighted in red, which is a child of the branch of plan and planned process. This is reflective of the decision to group classes of 'processes' into a branch, of which data transformations are a child (i.e. a data transformation is a process). Importantly, the data transformation ontology, whilst a child of the process class, exists as a separate development branch so that development work can run un parallel within a branch and experts in the field that the branch describes.

The structure of the ontology aims to reflect the many uses that a particular data transformation concept can be put to by the inclusion of a 'data transformation objective'. This objective enables the ontology to represent concepts which can have several different uses. An example of some of these objectives from the ontology is shown in Figure 2.

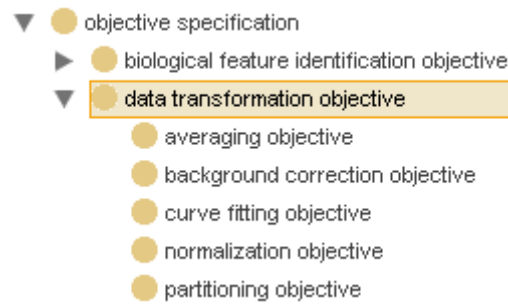


Figure 2. Data transformation objectives, used to describe the various aims a data transformation can be used for.

For example, considering a concept such as *k*-means clustering which can be used with the objective of class discovery or class assignment. In order to avoid multiple inheritance, the objective concept allows the concept *k*-means to be used in either context by assigning a particular objective given a particular usage. This makes the usage of the *k*-means concept in this example explicit, whereas giving *k*-means multiple parents of 'clustering' and 'classification' would introduce ambiguity and the complexities multiple inheritance can introduce when developing ontologies. These design principles have been previously discussed and reported in the deliverable D3.7.

There are currently 125 terms in the ontology file which are child classes of data transformation. These terms use many other classes from other parts of the ontology that are integrated with (OBI) and use relations also added for the benefit of the data transformation branch such as `has_objective` (to state when a transformation is being used for a specific goal) and `has_feature` (for instance to give a log transformation a base 10). Figure 3 illustrates the hierarchy implemented using the Web Ontology Language (OWL). The view is from the ontology editor Protégé.

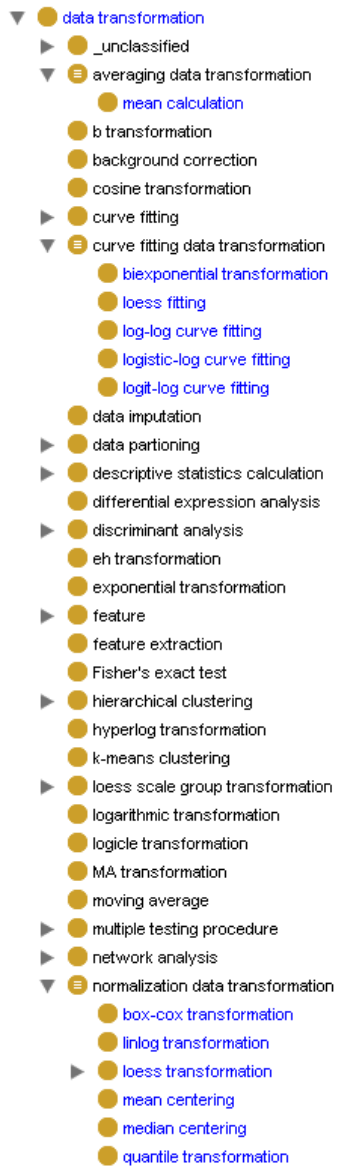


Figure 3. Data transformation ontology illustrated in the Protégé tool.

The hierarchy demonstrates a parent-child relationship, with the root here shown as data transformation (i.e. all of the classes shown are under the data transformation class and hence are types of data transformation). Of particular note are the classes which are displayed in blue. These classes have been classified (i.e. placed in the hierarchy by the reasoner) according to axioms describing them, rather than by the asserted parent. For example, considering the box-cox transformation, this class appears in blue under the normalization data transformation. A normalization data transformation is described as a class that has as an objective some normalization objective and this is asserted as a necessary and sufficient condition on the class. Used in this context, the necessary and sufficient condition is used to infer that any class where the rule 'has as an objective the normalization objective' is true then this is sufficient for the class to be placed under the normalization data transformation. Similarly, for the averaging data transformation the necessary and sufficient condition that it has as objective the averaging objective has meant that the mean calculation class has been placed under it as this class has the condition that it has as objective the averaging objective.

Using this approach, data transformations can be described as having multiple objectives and will appear in multiple places in the hierarchy, however this placement is inferred by the

reasoner and is not manually placed there. This eradicates some of the issues associated with multiple inheritance, such as the potential for inconsistencies and such automated generation of the hierarchy allows us to identify missing subsumption relationships and quickly add new classes

The current version of the NTO ontology, here shown as the data transformation ontology as part of OBI, is available from  
<http://obi.svn.sourceforge.net/viewvc/obi/trunk/src/ontology/branches/DataTransformation.owl>

## 1.6 Evaluation

As previously discussed, the requirements of the NTO were defined through a set of use cases and competency questions that were collected from the wider community. In order to validate that the ontology meets these requirements, a strategy was developed in order to evaluate the NTO which consisted of 3 components; competency questions, use cases and consistency checking. Competency questions and use case evaluation involves going back to the initial requirements and validating that the ontology meets the scope originally set out. This is akin to testing against a set of requirement in software engineering and is often used in ontology evaluation (Staab *et al*, 2001). Consistency checking concerns the use of a logical reasoner to computational evaluate the ontology for consistency and to infer hierarchies via the use of restrictions.

For the competency questions and use case evaluation, a repository was set up to keep a record of the documents containing the detailed information. This can be found at <http://obi.svn.sourceforge.net/viewvc/obi/trunk/docs/developer/DT/> The process undertaken was specific to the use case described in each instance. One of the usages described was the annotation of data, as can be seen in the GenePattern use case. For this use case, the terms provided were tabulated and for each and appropriate OBI class was assigned for annotation, where there was a successful match. Where there was not a successful match, (and the term was still considered in scope) the term was resubmitted to the tracker for inclusion in a future iteration of the ontology. The list of successfully matched terms with OBI classes was then returned to the submitter for their analysis and any further comments.

Table1. Example of an annotation use case submitted for inclusion in the ontology.

<b>NAME used by submitter</b>	<b>TASK TYPE (from use case submitter)</b>	<b>Comment (from NTO developers)</b>	<b>Closest current OBI Class</b>	<b>Action</b>
GSEA	Gene List Selection	One of the modules of the Gene Pattern software implementing the Gene Set Enrichment Analysis method by Mootha et al. This is a differential analysis of Gene Sets and a special type of differential expression analysis. Note that our current definition of differential expression analyses covers both analyses at the gene level and analyses at the gene set level. We might add subclasses corresponding to specific algorithms such as GSEA.	differential expression analysis	
KMeansClustering	Clustering		k-means clustering	
CART	Prediction	A type of decision tree learning. Also the name of one of the modules of the Gene Pattern software implementing this algorithm. We do plan to cover CART in DT, but are waiting on decision about proposed objectives ('prediction building' and 'class prediction').	None	Add decision trees (CART,...) to DT, direct child, assuming objectives are approved

An excerpt of one of the submitted use cases is illustrated in Table 1. This illustrates some of the information that is submitted, such as the name used by the submitter in their annotation work and the type of task the term relates to. Such information helps to add context to each of the terms submitted. Next to this there is a comment, if appropriate, from the developers of the NTO, often describing the interpretation taken when evaluating this particular term. Finally, the term is mapped to a class in the OBI ontology. This information would be used to determine if a class was present in the ontology, and therefore able to be successfully used to annotate this particular part of the use case, or whether further work was required. For instance, the CART term in this use case has resulted in an action for the inclusion of decision trees into the ontology. Table 1 illustrates an example of how the evaluation through use cases can both validate the terms that are contained by confirming they adequately cover the anticipated usage and, equally importantly, identify gaps in the ontology which can be corrected through this evaluation phase.

### 1.7 Dissemination of the NTO

Several activities have been planned to aid the dissemination of the NTO. The first of these was an ontology workshop, hosted at EBI in Cambridge in November 2007. This was organised largely to assist in the development of the ontology by inviting delegates with a

diverse range of backgrounds in order to gather a comprehensive range of perspectives and input from potential users of NTO, as well as a dissemination activity. This workshop was previously described in deliverable 3.7. This was followed by an invited talk at the EMERALD workshop at CAMDA 2007, aimed at the microarray community. Following this, the participation in an OBI workshop meeting in Vancouver, Canada, was also used to further extend the cooperation with the OBI project and assist development of the NTO. A poster was presented at ISMB 2008. A further 'ontology outreach workshop' is planned for November 2008 to disseminate to potential user communities and gather useful feedback and to evaluate the present ontology with a view to future enhancements and implementation of the ontology e.g. in text mining applications to aid data curators.

### **1.8 Assistance with Usage of the NTO**

Several steps have been taken to assist future communities who wish to use the ontology. Firstly, online documentation relating to the accessing and using of the ontology. This includes the following:

- Using OBI: [http://docs.google.com/Doc?docid=dzprnmw\\_10gk2mcpfb&hl=en](http://docs.google.com/Doc?docid=dzprnmw_10gk2mcpfb&hl=en)
- Summary of naming conventions and the formatting of the ontology: [http://docs.google.com/View?docid=dzprnmw\\_8fnr5nmd4](http://docs.google.com/View?docid=dzprnmw_8fnr5nmd4)
- Release notes for each version: [http://obi.svn.sourceforge.net/viewvc/obi/trunk/docs/user/release\\_notes.txt](http://obi.svn.sourceforge.net/viewvc/obi/trunk/docs/user/release_notes.txt)

Secondly, an OBI-user group and mailing list was established this year ([obi-users@googlegroups.com](mailto:obi-users@googlegroups.com)). The aim of this group is to provide expert advice to new and current users of the ontology from OBI developers and to provide a forum in which users can communicate with one another. Thirdly, the workshop to be held in 2008 will closely involve potential ontology users communities that are especially interested in the aspects of data transformation, and hence the NTO. This workshop will focus on describing the ontology to the users to help with their usage, gaining feedback given an examination of their submitted use cases and evaluating the existing structure to look for areas of future improvements. Furthermore, as part of a collaboration of the OBO Foundry, the ontology will undergo a series of expert reviews in order to help users and improve the ontology. This will result in an ontology that will be endorsed by the OBO Foundry to the bioinformatic community and provide a further mechanism to promote readout to the relevant audiences.

### **1.9 Future Work**

The ontology is currently being developed in an iterative fashion, with monthly releases of OBI (with NTO elements incorporated under the data transformation branch of work) being published as beta versions. The ontology files are available from the sourceforge svn site <http://obi.svn.sourceforge.net/viewvc/obi/>

The following have been tasks will be addressed between months 24-36:

1. Continue with adding concepts collected from communities into ontology
2. With reference to the above, in particular incorporation of some of the Software Ontology terms into the NTO since the SO is not sufficiently mature to provide coverage for the NTO use cases.
3. Journal publication on the ontology development to aid dissemination, - the manuscript is currently being written
4. Further evaluation using strategy detailed in this report and collect user feedback following the public 1.0 release
5. Further dissemination to community at outreach workshop in November 2008
6. Iteration using evaluation feedback
7. Assist developers to implement the ontology

### **D3.6 Workshop III Ontology Workshop**

An invitation only workshop was organized to facilitate the development of the ontology during a week-long face-to-face meeting with community experts and members of the Emerald consortium. Ongoing development subsequent to this workshop continued with collaboration with the Ontology for Biomedical Investigations (OBI) project, the successor to the MGED ontology. A workshop with the OBI consortium took place in early 2008 to facilitate the ongoing work and an agreed deadline for a first release of the ontology in autumn 2008. Dissemination has continued with a presentation at the EMERALD workshop of CAMDA 2007 and a poster has been accepted for presentation at ISMB 2008. A journal publication on the development work is planned for 2008 to aid dissemination. A subsequent ontology outreach workshop is planned for late 2008 to further disseminate the work to potential user communities. Complete proceedings of the workshop are available at: [https://wiki.cbil.upenn.edu/obiwiki/index.php/Data\\_transformation\\_workshop](https://wiki.cbil.upenn.edu/obiwiki/index.php/Data_transformation_workshop)

### **D3.7 Workshop III report**

The complete proceedings of the invitation only workshop are available online [https://wiki.cbil.upenn.edu/obiwiki/index.php/Data\\_transformation\\_workshop](https://wiki.cbil.upenn.edu/obiwiki/index.php/Data_transformation_workshop)

and a workshop report was produced in fulfillment of this deliverable. The workshop allowed us to access domain specialists in order to structure and extend the ontology. The subsequent teleconferences with these domain experts have occurred weekly and the ontology is now in the evaluation phase and will be part of a future OBI release.

### **Dissemination activities:**

The ontology and the wider EMERALD project has been disseminated at the following workshops and meetings:

Presenter	Meeting	Location	Audience	Date
James Malone	Manchester University Bioontologies Course	Manchester University, UK	Manchester University Ontology Group	Nov 2007
James Malone	CAMDA 07	Valencia, Spain	International microarray research community	December 13 & 14, 2007
James Malone Helen Parkinson	OBI Winter Workshop 2008	Vancouver, Canada	OBI development consortium	January 28- February 1, 2008
James Malone	ISMB 2008	Toronto, Canada	International bioinformatics research community	July 19-23, 2008
James Malone Helen Parkinson	Joint OBO Foundry and OBI workshop, 2008	Cambridge, UK	OBO Foundry members, OBI development consortium	July 7-11, 2008
James Malone	OWLED	Karlsruhe, Germany	Ontology developers community	Oct 26-27, 2008

## ***Publications:***

The OBI Consortium (2008) OBI: The Ontology for Biomedical Investigations. Poster in Bioontologies SIG, ISMB 2008.

The OBI Consortium (2008) OBI: The Ontology for Biomedical Investigations. Poster in ISMB 2008.

Melanie Courtot, William Bug, Frank Gibson, Allyson L. Lister, James Malone, Daniel Schober, Ryan Brinkman and Alan Ruttenberg. (2008) The OWL of Biomedical Investigations. OWLED 2008, Karlsruhe, Germany.

## ***References***

Bada MA and Altman RB (2000) Computational modeling of structural experimental data. *Meth. Enzymol.* 317, pp. 470–49.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, et al. (2001) Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nat. Genet.* 29, pp.365-371.

Gómez-Pérez A, Fernández-López M, Corcho O (2004) *Ontological Engineering*. Springer-Verlag, London.

Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P and Oinn T (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, 34, pp. W729–W732.

Rayner T, Rocca-Serra P, Spellman PT, Causton HC, Farne A, Holloway E, Liu J, Maier DS, Miller M, Petersen K, et al. (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics.*7(489).

Staab S, Schnurr HP, Studer R and Sure Y (2001) Knowledge processes and ontologies. *IEEE Intelligent Systems* 16(1), pp. 26-34.

Yeh I, Karp PD, Noy NF, Altman RB. (2003) Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). *Bioinformatics.* 19(2), pp. 241-8.

## Appendix A – GenePattern Use Case

NAME	TASK TYPE (from submitter)	Comment (EM)	Closest current OBI Terms (EM)	DT AI from conference call
ARACNE	Pathway Analysis	A network reconstruction algorithm ( <a href="http://amdec-bioinfo.cu-genome.org/html/ARACNE.htm">http://amdec-bioinfo.cu-genome.org/html/ARACNE.htm</a> ). Also the name of one of the modules of the Gene Pattern software implementing this algorithm. We need to cover network reconstruction (reverse engineering) in DT. Currently we have 'network analysis', but that assumes a network is already given and analyzes its properties. This is different from actually constructing a network from data.		Add DTs for network reconstruction-reverse engineering
AreaChange	Proteomics	One of the modules of the Gene Pattern software that calculates the fraction of the area under the spectrum attributable to signal. This should be cover once we expand subclasses of mass spectrometry analysis. One of the modules of the Broad Gene Pattern software. It exports experiment data from the caArray db and writes them into gct format. It's a utility and probably not in the scope of DT.	mass spectrometry analysis	Place generic algorithm for this under mass spectrometry analysis
caArrayImportViewer	Visualizer			Add DT 'format conversion'
CART	Prediction	A type of decision tree learning. Also the name of one of the modules of the Gene Pattern software implementing this algorithm. We do plan to cover CART in DT, but are waiting on decision about proposed objectives ('prediction building' and 'class prediction').		Add decision trees (CART,...) to DT, direct child, assuming objectives are approved
CARTXValidation	Prediction	One of the modules of the Broad Gene Pattern software. It implements CART with leave-one-out cross validation.	leave one out cross validation	

			method	
ClassNeighbors	Gene List Selection	One of the modules of the Gene Pattern software aimed at selecting genes that most closely resemble a profile, i.e. that are significantly correlated with a class template.		
ComparativeMarkerSelection	Gene List Selection	One of the modules of the Gene Pattern software. Similar to the above in aim.	differential expression analysis	Add differential analysis specific subclasses as needed
ComparativeMarkerSelectionViewer	Visualizer	One of the modules of the Gene Pattern software to work with and view the results of the above two. Implements various types of visualizations of the results. This is visualization and doesn't belong to DT.		
CompareSpectra	Proteomics	One of the modules of the Gene Pattern software to compare two spectra and determine similarity.	similarity calculation, mass spectrometry analysis	To be dealt with when expanding mass spectrometry analysis
ConsensusClustering	Clustering	One of the modules of the Gene Pattern software, implementing resampling-based clustering. Determines a consensus clustering among applications of a chosen clustering algorithm to perturbations of the original dataset. We do plan to cover clustering in OBI, but are waiting on decision about proposed objectives ('class discovery')	clustering method	Add consensus clustering to DT where appropriate in the hierarchy, once the clustering stuff is sorted out
ConvertLineEndings	Preprocess & Utilities	One of the modules of the Gene Pattern software. Converts line endings used into a file into line endings used by Perl in the host operating system. This is a file manipulation task, should not belong to the DT branch. Should we pass it to some other branch? Which?		

ConvertToMAGEML	Preprocess & Utilities	One of the modules of the Gene Pattern software which converts from Gene Pattern data formats to MAGE-ML. Maybe OBI should capture the more general concept of conversion from a given format (not necessarily from Gene Pattern) to MAGE-ML. But which branch does this belong to?	
CopyNumberDivideByNormals	SNP Analysis	One of the modules of the Gene Pattern software related to SNP analysis and normalization of SNP data. We should cover SNP analyses in DT.	DT needs to cover SNP analysis
CytoscapeViewer	Visualizer	One of the modules in the Gene Pattern software used to visualize the adjacency matrix output by the Gene Pattern ARACNE implementation. This is a visualization of a DT, not the DT itself and we had decided doesn't belong to our branch. The type of visualization should be covered by DENRIE.	Need a 'graph' concept in DENRIE, in the sense of (V=vertices, E=edges), not in the sense of graph of a function.
DownloadURL	Preprocess & Utilities	One of the modules of the Gene Pattern software. It's a utility aimed at downloading a file from a url and saving it into text. Probably not within the OBI scope.	
ExpressionFileCreator	Preprocess & Utilities	One of the modules of the Gene Pattern software. Creates an expression dataset (in Gene Pattern format) from Affy .CEL file. The summarization method to use (MAS 5, RMA, etc.) can be specified.	feature extraction Add specific feature extraction algorithms in DT, such as RMA, gcRMA, MAS5, GenePix. Sometimes algorithm name coincide with software name, must deal with this.
ExtractColumnNames	Preprocess & Utilities	One of the modules of the Gene Pattern software providing a utility to get the header from a .res file (this is a Gene Pattern format). Probably not in the scope of OBI.	
ExtractComparativeMarkerResults	Gene List Selection	One of the module of the Gene Pattern software. Creates a dataset and a feature list file from the results of ComparativeMarkerSelection.	projection and (row) submatrix extraction
ExtractRowNames	Preprocess & Utilities	Same comment as for ExtractColumnNames, except instead of the header of the file, the row names are extracted.	

FeatureSummaryViewer	Visualizer	One of the modules of the Gene Pattern software used to view a list of feature from a predictor. Probably not in our branch, but DENRIE. And OBI should cover the various visualization types used, not the specific Gene Pattern implementation.		
GeneCruiser	Annotation	One of the modules of the Gene Pattern software. Retrieves probe annotation (Locus Link, Unigene, etc.) for Affy arrays from the GeneCruiser web server. Not sure which branch this would belong to, if any.		
GeneListSignificanceViewer	Visualizer	One of the modules of the Gene Pattern software used to view the output of GeneNeighbors and ClassNeighbors. Same comment as for other viewers above.		
GeneNeighbors	Gene List Selection	One of the modules of the Gene Pattern software. Finds the N closest genes to a query gene according to a similarity calculation. Since we have similarity calculation in OBI, this is a result of performing several of these followed by ranking. Need to make sure our similarity calculation cover all possible types (currently we only cover euclidean and pearson), but when that is done, not sure if we need anything else.	similarity calculation	Add more similarity calculation methods Mahlanobis, etc.
GEOImporter	Preprocess & Utilities	One of the modules of the Broad Gene Pattern software. It export experiment data from the GEO db and writes them into gct format. It's a utility and probably not in the scope of DT.		
GLAD	SNP Analysis	One of the modules of the Gene Pattern software implementing Gain and Loss analysis of DNA. Detection methods for chromosomal aberrations need to be covered by DT.		DT needs to deal with methods to detect chromosomal aberrations
GlobalAlignment	Sequence Analysis	One of the modules of the Gene Pattern software implementing 3 dynamic programming sequence analysis algorithms. We need to cover alignment in DT (the 3 algorithms should be separate classes).	sequence analysis	

Golub.Slonim.1999.Science.all.aml.pipeline	pipeline	One of the modules of the Gene Pattern software. It reproduces several steps of analyses carried out in the Golub et al paper. We need to cover the various analysis steps (separate classes) in DT and whereas maybe we should have a concept of workflow/pipeline somewhere in OBI (not in DT probably), but we need not list all possible pipelines specific to papers.		Ask PLAN branch to capture the concept of pipeline or workflow
GSEA	Gene List Selection	One of the modules of the Gene Pattern software implementing the Gene Set Enrichment Analysis method by Mootha et al. This is a differential analysis of Gene Sets and a special type of differential expression analysis. Note that our current definition of differential expression analyses covers both analyses at the gene level and analyses at the gene set level. We might add subclasses corresponding to specific algorithms such as GSEA.	differential expression analysis	
GSEALeadingEdgeViewer	Visualizer	One of the modules of the Gene Pattern software that allows viewing of the Leading Edge analysis results from GSEA. Same comment as for other visualizations, it doesn't belong to DT.		
HeatMapImage	Image Creators	One of the modules of the Gene Pattern software to create Heat Maps. Doesn't belong to DT but to DENRIE.		
HeatMapView	Visualizer	One of the modules of the Gene Pattern software to view Heat Maps. Doesn't belong to DT.		Ask DENRIE about having a heat-map concept
HierarchicalClustering	Clustering		hierarchical clustering	
HierarchicalClusteringImage	Image Creators	One of the modules of the Gene Pattern software to create the image of the dendrogram from HierarchicalClustering. Doesn't belong to DT but to DENRIE.		
HierarchicalClusteringViewer	Visualizer	One of the modules of the Gene Pattern software to view the HierarchicalClustering dendrogram. Doesn't belong to DT.		

Hu68kHu35kAtoU95	Preprocess & Utilities	One of the modules of the Gene Pattern software providing a utility to map identifiers from one Affy chip to another. Probably doesn't belong to DT, does it belong to some other OBI branch?		
ImputeMissingValues.KNN	Missing Value Imputation	One of the modules of the Gene Pattern software implementing the KNN method of data imputation. We need to expand our 'data imputation' into subclasses, including that for KNN-imputation.	data imputation	Add KNN as tool for data imputation and class prediction
JavaTreeView	Visualizer	One of the modules of the Gene Pattern software to view Eisen's TreeView files. Same comment as for other viewers.		
KMeansClustering	Clustering		k-means clustering	
KNN	Prediction	A type of 'class prediction' methods. We are planning to cover this in DT but are waiting on decision about proposed objectives ('class prediction')		See above re KNN
KNNXValidation	Prediction	Similar to CARTXValidation but for KNN instead of CART		
KSscore	Statistical Methods	One of the modules of the Gene Pattern software to compute the KS score, which is for example used in GSEA to represent the positional distribution of a query gene sets within a ranked list of genes. What should be captured in OBI is probably the Kolmogorov-Smirnov test.		Need to deal with statistical tests. Proposal is to start by having a statistical testing class incorporating the main tests in flat fashion. We'll subsequently refine and revise whether or not we should be separating the test statistics out.
LandmarkMatch	Proteomics	GenePattern module implementing a method to increase the number of identified peptides in an LC-MS experiment. Belongs to mass spectrometry analysis.		To be dealt with when expanding mass spectrometry analysis
LocatePeaks	Proteomics	GenePattern module aimed at locating peaks in a spectrum. Belongs to mass spectrometry analysis.		To be dealt with when expanding mass spectrometry analysis
LOHPaired	SNP Analysis	GenePattern module aimed at detecting Loss Of Heterozygosity from paired normal-target samples. LOH algorithms should be encompassed in chromosomal aberration methods.		To be dealt with methods to detect chromosomal aberrations

Lu.Getz.Miska.Nature.June.2005.mouse.lung.pipeline	pipeline	GenePattern module implementing a specific pipeline using kNN class prediction. Comments similar to the Golub pipeline above. Method to capture in OBI is KNN class prediction.	
MAGEMLImportViewer	Visualizer	GenePattern module consisting of a visualizer to import MAGE-ML data into GenePattern. Doesn't belong to DT.	
MapChipFeaturesGeneral	Preprocess & Utilities	GenePattern module to replace gene accessions (features) in a dataset with another dataset as specified by a look-up table. Doesn't belong to DT.	
MergeColumns	Preprocess & Utilities	GenePattern module to merge files by columns retaining common rows. Probably not in the scope of OBI.	
MergeRows	Preprocess & Utilities	GenePattern module to merge files by rows retaining common columns. Probably not in scope of OBI.	
MINDY	Pathway Analysis	GenePattern module implementing the MINDY algorithm. This is an algorithm to infer genes that modulate the activity of a TF at post-transcriptional levels. This algorithm uses mutual information to measure the dependence between a TF and its target gene. The concept of mutual information should be captured in OBI, probably in DT. We should also capture pathway analysis and methods such as MINDY.	DT needs to capture mutual information and pathway analysis
MINDYViewer	Visualizer	GenePattern visualizer for MINDY utilizing heatmaps. As per above, the concept of heatmap should be captured by DENRIE.	
Multiplot	Visualizer	GenePattern module which creates customizable plots (scatter plots) of expression data-derived data.	dot plot
MultiplotExtractor	Visualizer	GenePattern user interface to save data created by Multiplot Preprocess. Should the general concept of interfact be captured in OBI, which branch?	

MultiplotPreprocess	Preprocess & Utilities	GenePattern module to create derived data from an expression dataset. DT should just worry about methods for 'deriving data', e.g. normalization, etc. We have several of these and more will be added.		
mzXMLToCsv	Proteomics	GenePattern module which converts an mzXML file to a zip of cvs file. Similar comment as for data format converters above.		
NMF	Projection	GenePattern module implementing NMF for class discovery. We have this in DT, under _unclassified. Note that this is also a dimensionality reduction.	non-negative matrix factorization	
NMFConsensus	Clustering	GenePattern module implementing consensus clustering using NMF. Should be sorted out together with NMF and consensus clustering.		Sort out after NMA and Consensus Clustering have been dealt with.
PCA	Projection	GenePattern module for principal component analysis implementing the MeV methods. We currently have PCA application to reduce the dimensionality of a dataset in OBI. We might also want to capture the principal component analysis method in its generality.	principal component analysis dimensionality reduction	Possibly add usage of PCA for clustering, besides its already existing usage in DT for dimensionality reduction.
PCAViewer	Visualizer	GenePattern visualizer for PCA. Outside of DT scope.		
PeakMatch	Proteomics	GenePattern module which performs peak clustering across multiple LC-MS sample runs. To be dealt with when expanding mass spectrometry analysis. Possibly a class discovery objective might be applicable for this.		To be dealt with when expanding mass spectrometry analysis
Peaks	Proteomics	GenePattern module which determines peaks in a spectrum using a series of digital filters. To be dealt with when expanding mass spectrometry analysis.		To be dealt with when expanding mass spectrometry analysis
PlotPeaks	Proteomics	GenePattern visualizer which plots the output of the Peaks module. Outside of DT scope.		
PredictionResultsViewer	Visualizer	GenePattern module to view the results of a predictor on a testing set. Not in DT scope.		

PreprocessDataset	Preprocess & Utilities	GenePattern module which performs various preprocessing operations: thresholding, filtering, discretization, normalization. DT should capture the individual operations. Filtering is captured, at least in generality, under subclasses of dimensionality reduction, and various normalizations are already captured. Might want to add terms for thresholding and discretization. Also need to capture filtering criteria somewhere (parameter/feature?).		To discuss how to best capture concepts such as 'criterion' for a projection (filtering), e.g. through parameter/feature maybe or in some other way. Also DT needs to add terms for binning and discretization.
ProteoArray	Proteomics	GenePattern module for feature detection and sample alignment for LC-MS proteomic data. To be dealt with when expanding mass spectrometry analysis.	feature extraction	To be dealt with when expanding mass spectrometry analysis
ProteomicsAnalysis	Proteomics	GenePattern module which run various steps in proteomics analysis on the input spectra. These include: quality assessment, normalization, peak detection, peak matching. The individual steps should be dealt with when expanding mass spectrometry analysis in DT.		Methods for quality assessment, normalization, peak detection, peak matching in mass spectrometry analysis need to be added. Data reordering might be needed, e.g. when one wants to produce heat-map reflecting sample similarities, they first reorder the input. Need to introduce a generic term for reordering (possibly with a 'criterion' parameter/feature).
ReorderByClass	Preprocess & Utilities	GenePattern module which reorders an expression file and a class file so that samples of the same class are together. Is this in DT/OBI scope?		
SelectFeaturesColumns	Gene List Selection	GenePattern module which generates a new file based on columns selected from an existing file.	column submatrix extraction	
SelectFeaturesRows	Gene List Selection	GenePattern module which generates a new file based on rows selected from an existing file.	row submatrix extraction	
SNPFileCreator	SNP Analysis	GenePattern module which creates a (GenePattern) .snp file from a set of Affymetrix SNP chip CEL files. Choice of 4 conversion algorithms is possible. This belongs to feature extraction. More specific conversion algorithms should be added under this class in DT.	feature extraction	Add specific feature extraction algorithms for SNP array data

SNPFileSorter	SNP Analysis	GenePattern module which sorts SNPs by chromosome and physical location. Should we include sorting/ranking terms in DT, with parameter/feature 'criterion'? Should we just add a specific term for this special case in SNP analysis?		See above, need reordering concept at a minimum
SNPMultipleSampleAnalysis	SNP Analysis	GenePattern module implementing the MSA algorithm developed at Upenn ( <a href="http://www.cbil.upenn.edu/MSA">http://www.cbil.upenn.edu/MSA</a> ). This identifies conodont aberrations across multiple samples. This algorithm should be dealt within methods for chromosomal aberration to be added to DT.		MSA should be captured in DT among methods to detect chromosomal aberrations
SnpViewer	Visualizer	GenePattern visualizer which displays SNP data (copy number or LOH) via a heat map. The concept of heat map should be captured by DENRIE.		
SOMClustering	Clustering	GenePattern module implementing the SOM clustering algorithm. SOM should be covered by DT (with class discovery objective).	clustering method	SOM should be added to DT (with objective class discovery).
SOMClusterViewer	Visualizer	GenePattern module to view SOM clustering results. Not in DT scope.		
SplitDatasetTrainTest	Preprocess & Utilities	GenePattern module which splits a dataset into a number of train and test subsets, either by percentage or by cross validation approaches. A 'partitioning' objective has been proposed by DT, pending approval. A parameter/feature 'criterion' might possibly be added to the latter.		Check the status of 'partitioning' objective. Consider adding a 'criterion' feature to this. Discuss with role branch the possibility of having a training and a testing role for datasets.
SubMap	Clustering	GenePattern module which takes as input a pair of independent microarray datasets, each with a sample subclass information, and which searches for matching pairs of subclasses between the 2 datasets. This uses GSEA for similarity measures between subclasses. Possibly a type of similarity calculation. Specific algorithm might be captured in DT.	similarity calculation	Possibly capture this specific algorithm in DT

SurvivalCurve	Survival Analysis	GenePattern module which draws a survival curve. DT should capture methods to identify genomics markers or predictive models for clinical outcomes. The specific type of plot (survival curve) might be captured by DENRIE.		Ask DENRIE about having a survival-curve term. DT should capture methods to identify genomics markers or predictive models for clinical outcome.
SurvivalDifference	Survival Analysis	GenePattern module to test if there is a difference between two or more survival curves based on sample classes defined by genomics data. It offers the choice of log-rank test and generalized Wilcoxon tests. These statistical tests should be covered by DT.		DT should cover statistical tests such as generalized Wilcoxon and log-rank tests.
SVM	Prediction	GenePattern implementation of the SVM algorithm. SVM should be captured in DT, with class prediction objective.		DT should capture SVM
TransposeDataset	Preprocess & Utilities	GenePattern module that transposes a dataset. Matrix transposition is a basic operation, not sure if it is necessary to cover it in DT.		DT should include a dataset transposition term.
UniquifyLabels	Preprocess & Utilities	GenePattern module which makes row or column labels unique. Probably not in OBI scope.		
VennDiagram	Visualizer	GenePattern module to display Venn Diagrams. Venn diagrams should be captured by DENRIE.		Propose DENRIE to include a term for Venn diagram
WeightedVoting	Prediction	GenePattern module which makes a weighted linear combination of relevant markers in a training set to perform class prediction. The specific algorithm might be captured in DT, with objective class prediction.		The specific algorithm might be captured in DT, with objective class prediction.
WeightedVotingXValidation	Prediction	GenePattern module which implements Weighted Voting with leave-one-out cross validation.	leave one out cross validation method	
XChromosomeCorrect	SNP Analysis	GenePattern module which doubles the intensity value for each SNP on the X chromosome from a male donor sample. A type of very specific preprocessing, possibly to be captured when adding specific terms for SNP analysis.		DT should probably capture this when dealing with SNP analysis terms