
ProtCV

Protein Clustering & Visualization of mass spectra peak lists

Stavroula Ventoura
Eugenia G. Giannopoulou, Elias Manolakos

Department of Informatics and Telecommunications
University of Athens

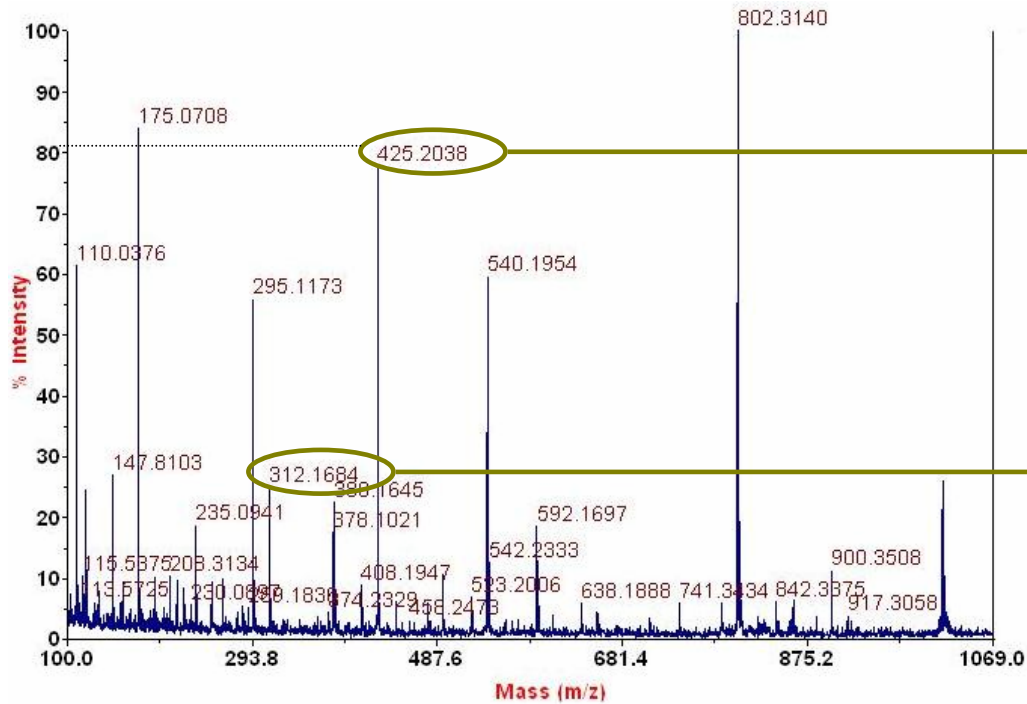
Objectives

- Identify new proteins previously unidentified
 - Increase confidence of identification
 - Discover biomarkers in clinical proteomics applications
 - Explore the spectral relationship between protein spots in a 2D-gel
 - Possibly assign function to proteins using “guilt by association”
 - Discover protein networks
-

ProtCV Steps-Tabs

- Load Data Set
 - the peak list files coming from MALDI spectrometer
 - Peak lists Preprocessing
 - Filtering – Binning
 - MS-Screener – Protein Vector file creation
 - Vector Pre-processing
 - Normalization / Scaling of vectors
 - Scaling of each vector
 - Normalization of each bin of every vector
 - Clustering (HC, k-means)
 - Visualization (HeatMap, Dendrogram, Cluster Set)
 - Clustering Validation (Silhouette, Dunn, Davies-Bouldin)
-

Vector File Creation



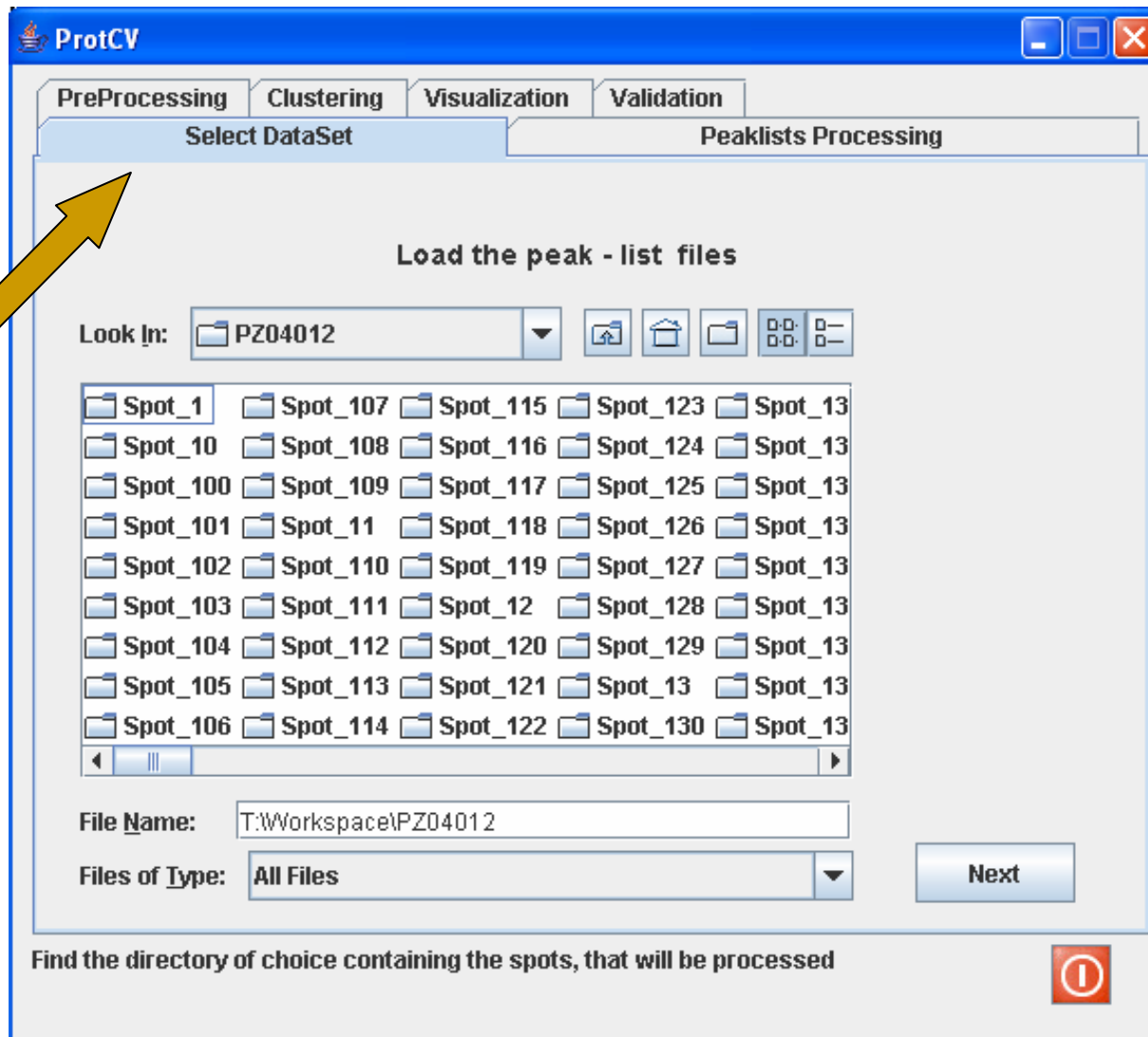
Concatenation of all vectors into a vector file

ms_name	...	295.1173	312.1684	378.1021	425.2038	540.1954	...
Spot_1	...	1250	456	254	1800	658	...

ms_name	...	295.1173	312.1684	368.6369	378.1021	385.6369	480.256	503.0515	540.1954	558.128	588.128	623.2441	...
Spot_1	...	1250	456	0	254	235.5	1803.85	0	124	658	2254.4	0	653.98
Spot_2	...	0	0	235.5	0	0	1803.85	124	0	0	2254.4	653.98	...

Load Data Set

Choose a dataset, containing peaklists from MALDI



Filtering – Binning

ProtCV

PreProcessing Clustering Visualization Validation

Select DataSet Peaklists Processing

Show ToolTips

Peaklist Processing

Filtering

MS-Screener

Call MS-Screener

Binning

Load peaklists

Create vectors

Original or Filtered?

Choose whether you want to load the original or the modified spots

Original Spots Modified Spots

XML

Export peaklists

Back Next

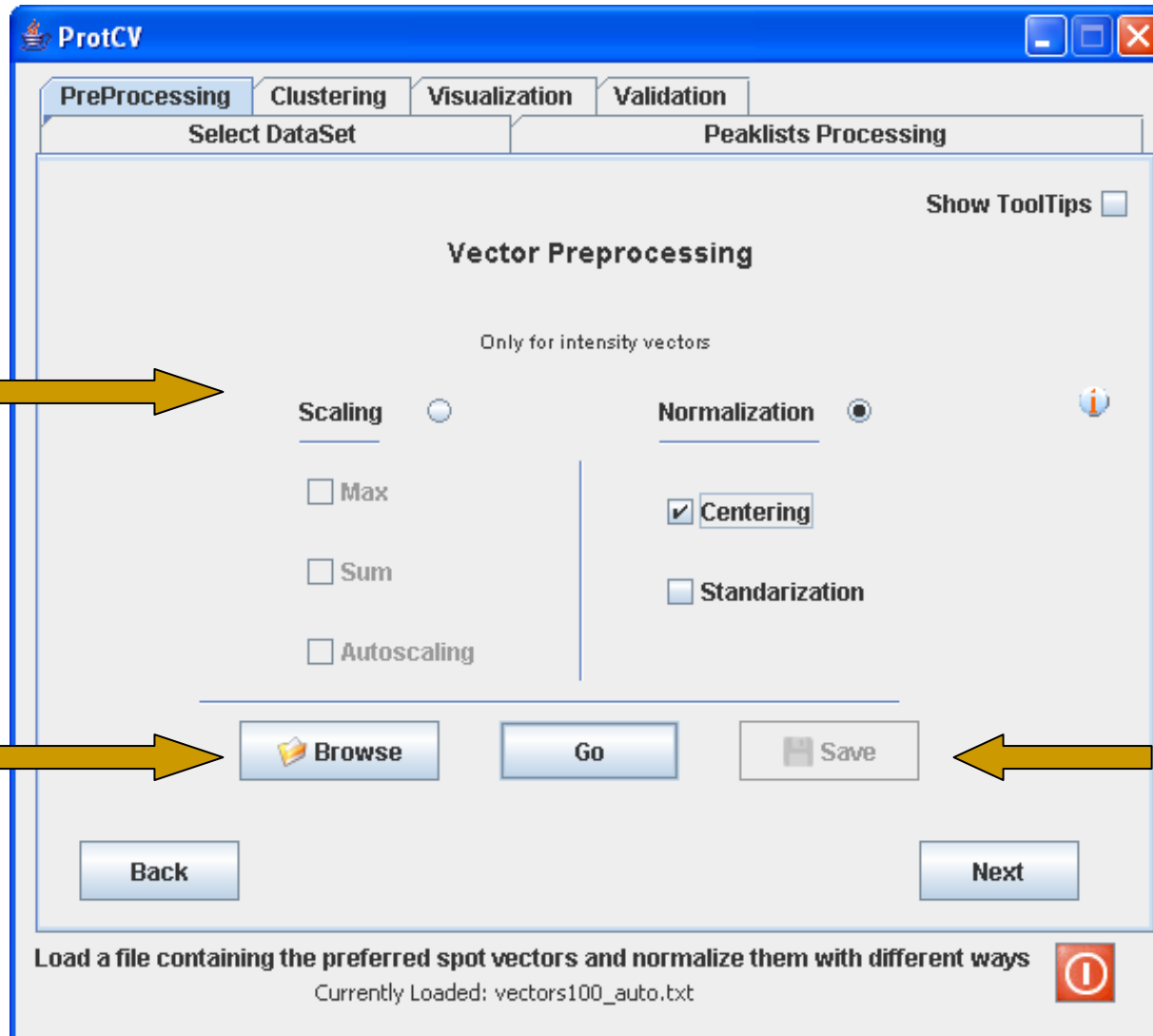
After choosing our working set, MS-Screener can be called for further processing

Remove
Contaminant
Masses

Create peaklist
vector file

Export modified
peaklists to
rerun MASCOT

Preprocessing

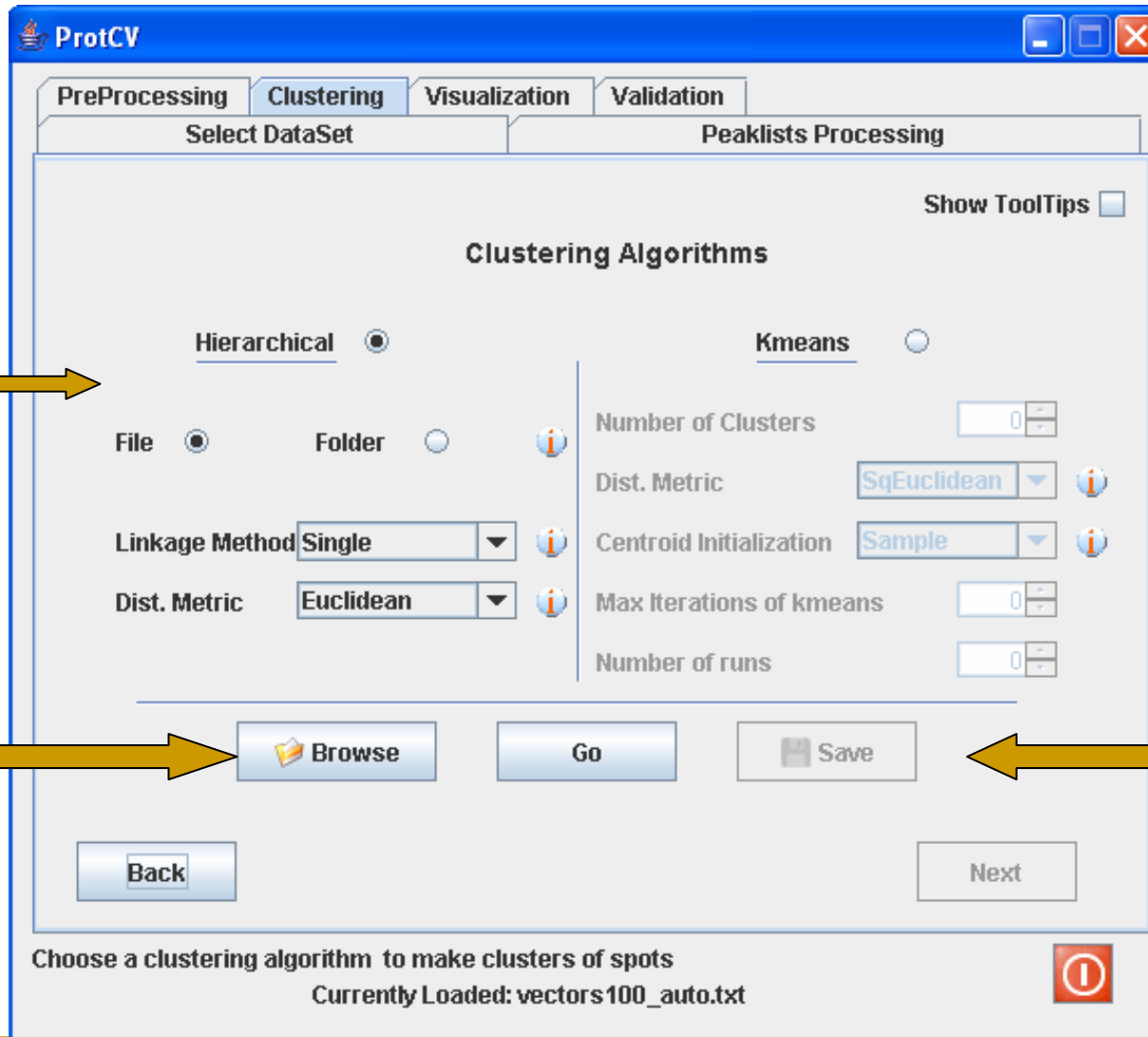


Choose a preprocessing method

Load a vector file

Save results

Clustering



Choose a clustering algorithm



Load the vector file



Save clustering results



Cluster Visualization Step

ProtCV

PreProcessing Clustering Visualization Validation

Select DataSet Peaklists Processing

Show ToolTips

Cluster Visualization

Dendrogram Heat Map Cluster Set

Threshold: Default [0] [0]

or [0] [0]

Cut dendrogram in clusters at threshold

Linkage Meth. Dist. Metric

Threshold: Default [0] [0]

or [0] [0]

Dimension: 1 2

No choices

File Browse View View Details

Folder

Back Next

Visualize the clustering results in 3 ways, created from the clustering algorithms

Define the parameters of the visualization method



Load the vector file or a folder of vector files



Choose a visualization method



Visualization – Dendrogram

Info about Clustered Spots

Clustered Spots with their Accession Number

Spots	Accessions	Description
Spot 167	HSB7_HUMAN	(Q9UBY9) Heat-shock protein beta-7 (HspB7) (Cardi...
Spot 107	TTHY_HUMAN	(P02766) Transthyretin precursor (Prealbumin) (TBP...
Spot 53	CCA4_HUMAN	(Q9BXL8) Cell division cycle associated protein 4 (H...
Spot 52	HXB4_HUMAN	(P17483) Homeobox protein Hox-B4 (Hox-2F) (Hox-...
Spot 196	TTHY_HUMAN	(P02766) Transthyretin precursor (Prealbumin) (TBP...
Spot 354	K2CE_HUMAN	(P48668) Keratin, type II cytoskeletal 6E (Cytokera...
Spot 352	WWP2_HUMAN	(O00308) Nedd-4-like E3 ubiquitin-protein ligase W...
Spot 149	TWF1_HUMAN	(Q12792) Twinfilin 1 (A6 protein) (Protein tyrosine k...
Spot 135	TTHY_HUMAN	(P02766) Transthyretin precursor (Prealbumin) (TBP...
Spot 92	AB3G_HUMAN	(Q9HC16) DNA dC->dU editing enzyme APOBEC-3G...
Spot 106	UBIQ_HUMAN	(P62988) Ubiquitin
Spot 59	ATNG_HUMAN	(P54710) Sodium/potassium-transporting ATPase g...
Spot 361	GBB3_HUMAN	(P16520) Guanine nucleotide-binding protein G(I)/G...

Buttons: To TAGGO, Save, Exit

Go to Spot: Go

Spots vs distance

Export to TAGGO, tool that connects the accession numbers with their GO information

Save selected spots for further processing

Merge of clustering results with protein identification results

Find a certain spot

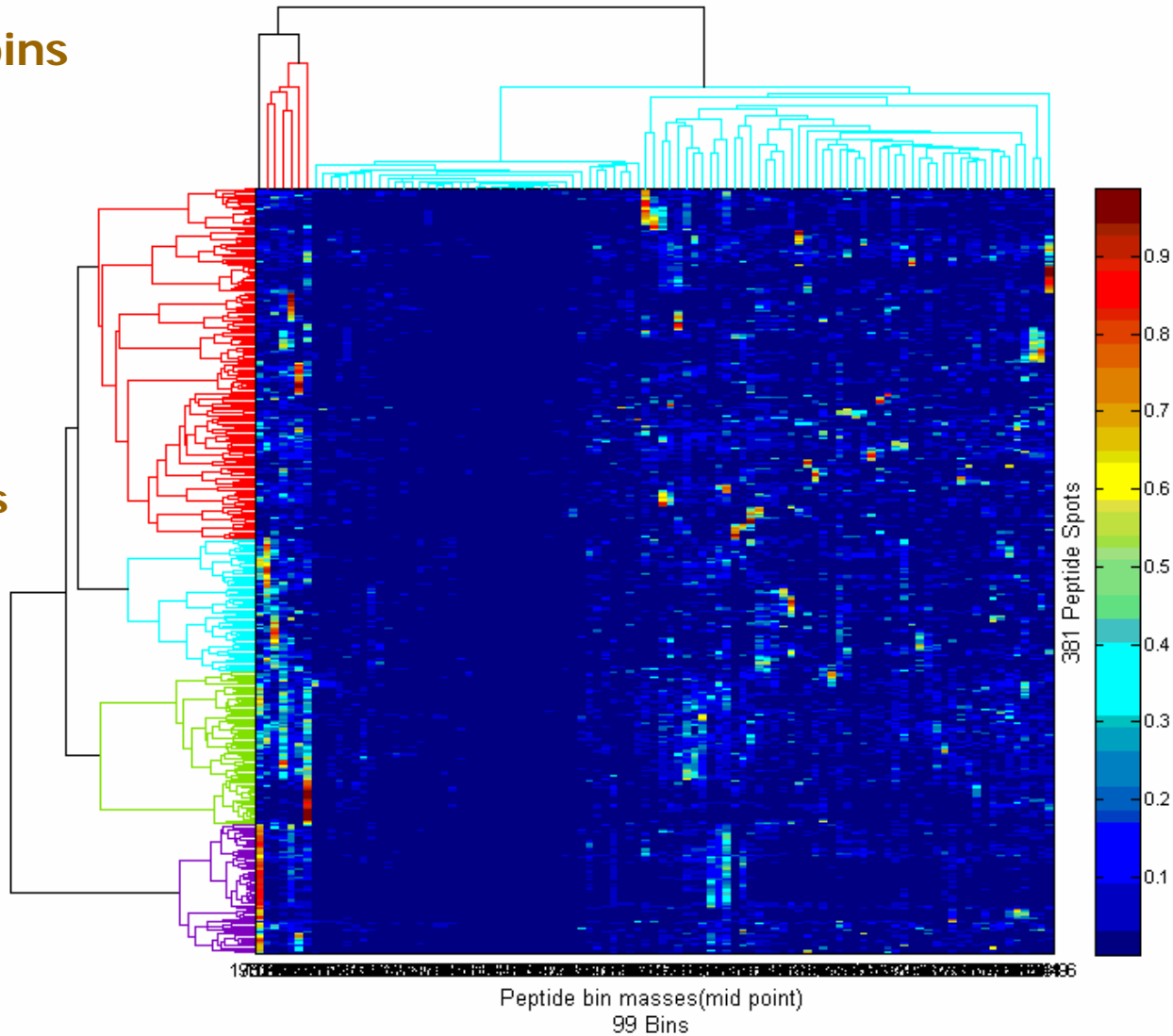
Visualization – HeatMap

Spots vs bins

Bin Clusters

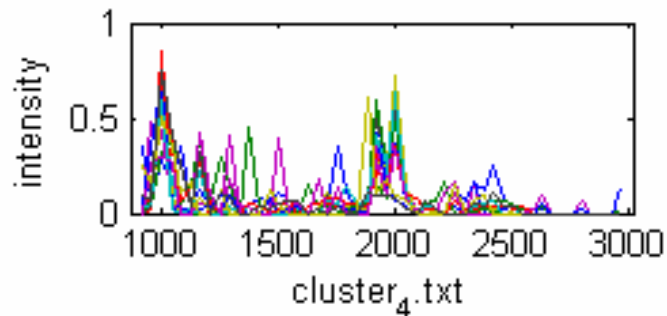
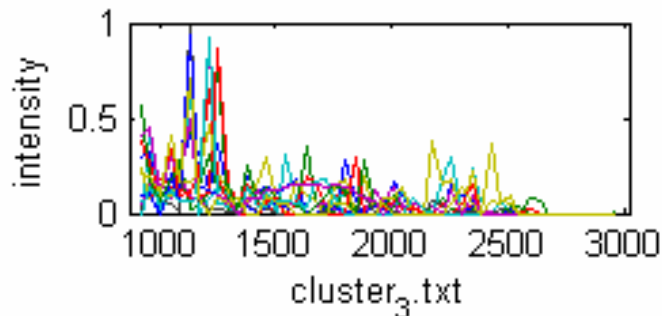
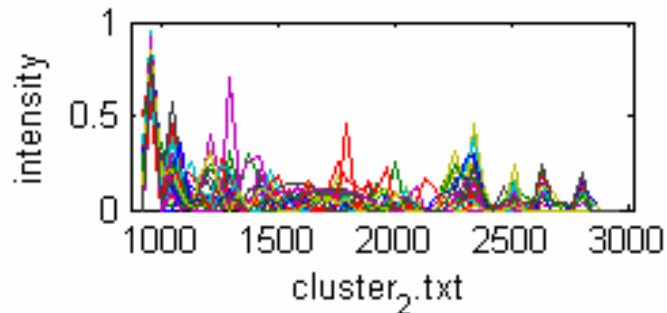
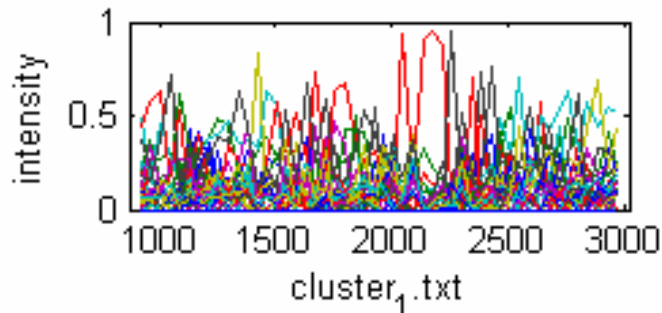
4 major Spots
Clusters
created

Distance scale



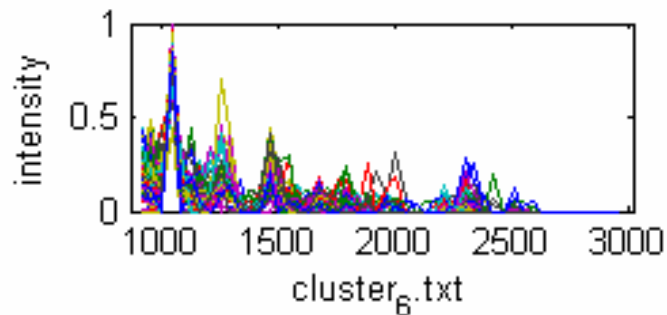
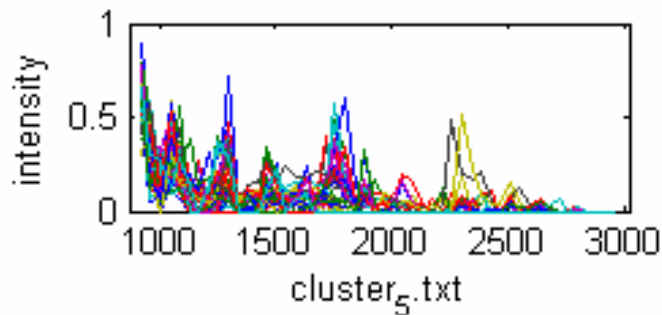
Visualization – ClusterSet

Mass vs
intensity

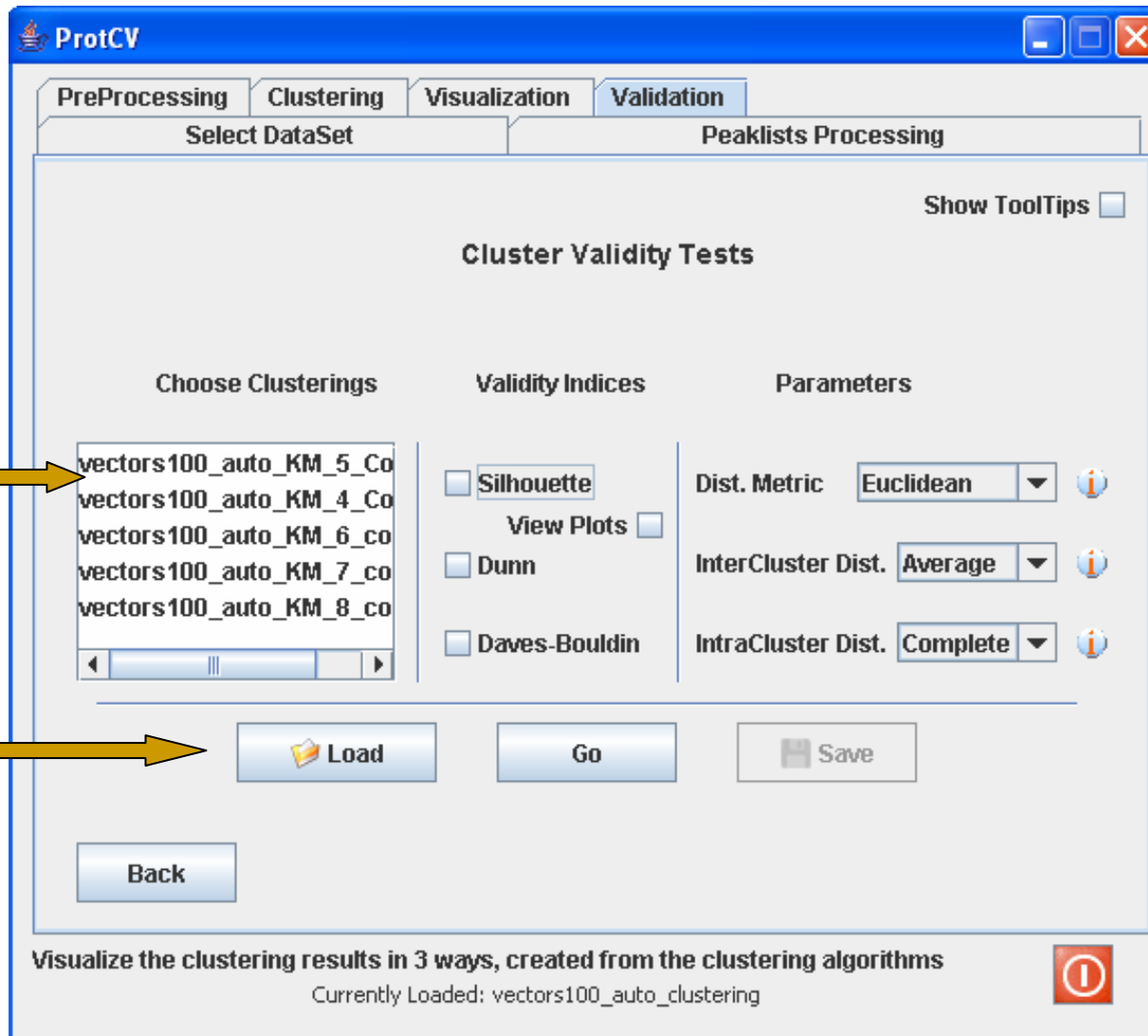


6 Clusters

Check how
dense a
cluster is,
from the
variation of
its cluster
members



Cluster Validation



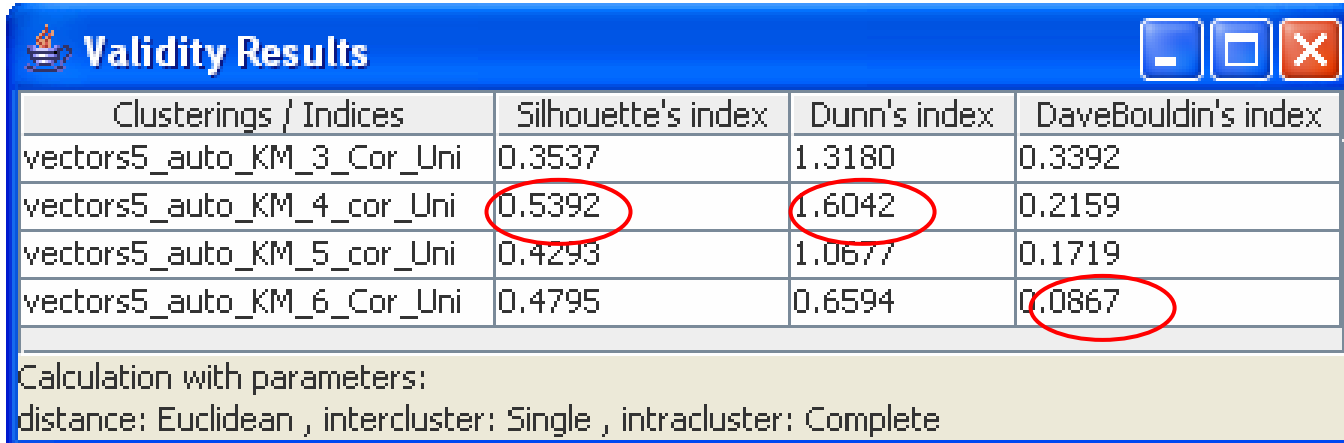
Select clustering partitions

Load the a folder containing many different clusterings of the same vector file

Choose validity index and define their parameters

Cluster Validation – Results

Validity Tests for finding the suitable number of clusters

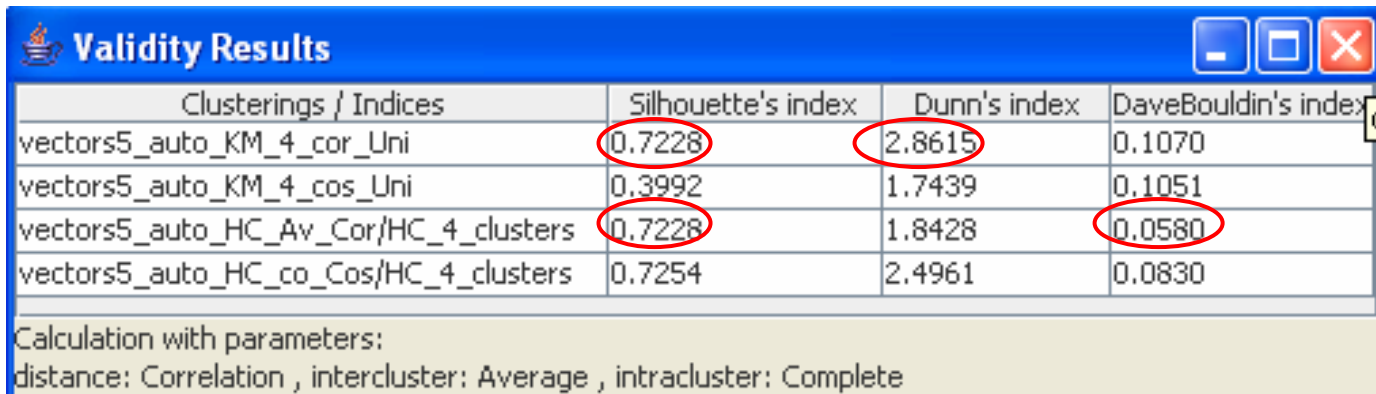


Validity Results

Clusterings / Indices	Silhouette's index	Dunn's index	DaveBouldin's index
vectors5_auto_KM_3_Cor_Uni	0.3537	1.3180	0.3392
vectors5_auto_KM_4_cor_Uni	0.5392	1.6042	0.2159
vectors5_auto_KM_5_cor_Uni	0.4293	1.0677	0.1719
vectors5_auto_KM_6_Cor_Uni	0.4795	0.6594	0.0867

Calculation with parameters:
distance: Euclidean , intercluster: Single , intracluster: Complete

Validity Tests for finding the suitable clustering



Validity Results

Clusterings / Indices	Silhouette's index	Dunn's index	DaveBouldin's index
vectors5_auto_KM_4_cor_Uni	0.7228	2.8615	0.1070
vectors5_auto_KM_4_cos_Uni	0.3992	1.7439	0.1051
vectors5_auto_HC_Av_Cor/HC_4_clusters	0.7228	1.8428	0.0580
vectors5_auto_HC_co_Cos/HC_4_clusters	0.7254	2.4961	0.0830

Calculation with parameters:
distance: Correlation , intercluster: Average , intracluster: Complete

Related work

- MS-Analyzer (<http://dns2.icar.cnr.it/proteus/>)
 - Preprocessing
 - Supports only **binning/normalization** (not scaling)
 - Data Mining
 - Supports all data mining methods in WEKA (including **classification**)
 - Visualization
 - Simple mass spectra **plot**
 - No cluster validation performed
-

Conclusions

■ Cluster Analysis

- Define the relationship between protein spots in a 2D-gel
- Prediction of function of unknown proteins through association (clustering) with known ones

■ Clustering Validation

- protein clusters can be validated using multiple methods in order to find the clustering that best captures the underlying structure of a peak lists dataset.
-

Thank You!
