NAVAL SURFACE WARFARE CENTER
DAHLGREN DIVISION

Exploratory Data Analysis of Large Document Collections

Jeff Solka Ph.D. and Avory Bryant

Presented by:
Jeff Solka Ph.D.
Principal Scientist
Computational Mathematics & Statistics Branch
dlgr_nswc_q21@navy.mil

NAVSEA
WARFARE CENTERS
DAHLGREN

Agenda

- The problem manifests itself computationally
- The role of clustering
- The problem manifests itself visually
- The solution
  - Hierarchical aggregation
  - Multi-feature clustering
- Acnecdotal results on PubMed
  - Clustering on keywords
  - Clustering on chemicals
- Wrap-up

Acknowledgements

- Office of Naval Research In-house Laboratory Independent Research (ILIR) Program

- Avory Bryant
  - Pursuit of excellence in the face of every increasing demands.

The Problem Manifests Itself Computationally

- Given a large collection of data, n > 10^6, and associated meta-tag structure how does one understand the structure within the data?
  - 3,129,445 PubMed abstracts published between 2000 and 2005

- Associated meta-tag structure includes
  - Titles
  - Authors
  - Institutions
  - …

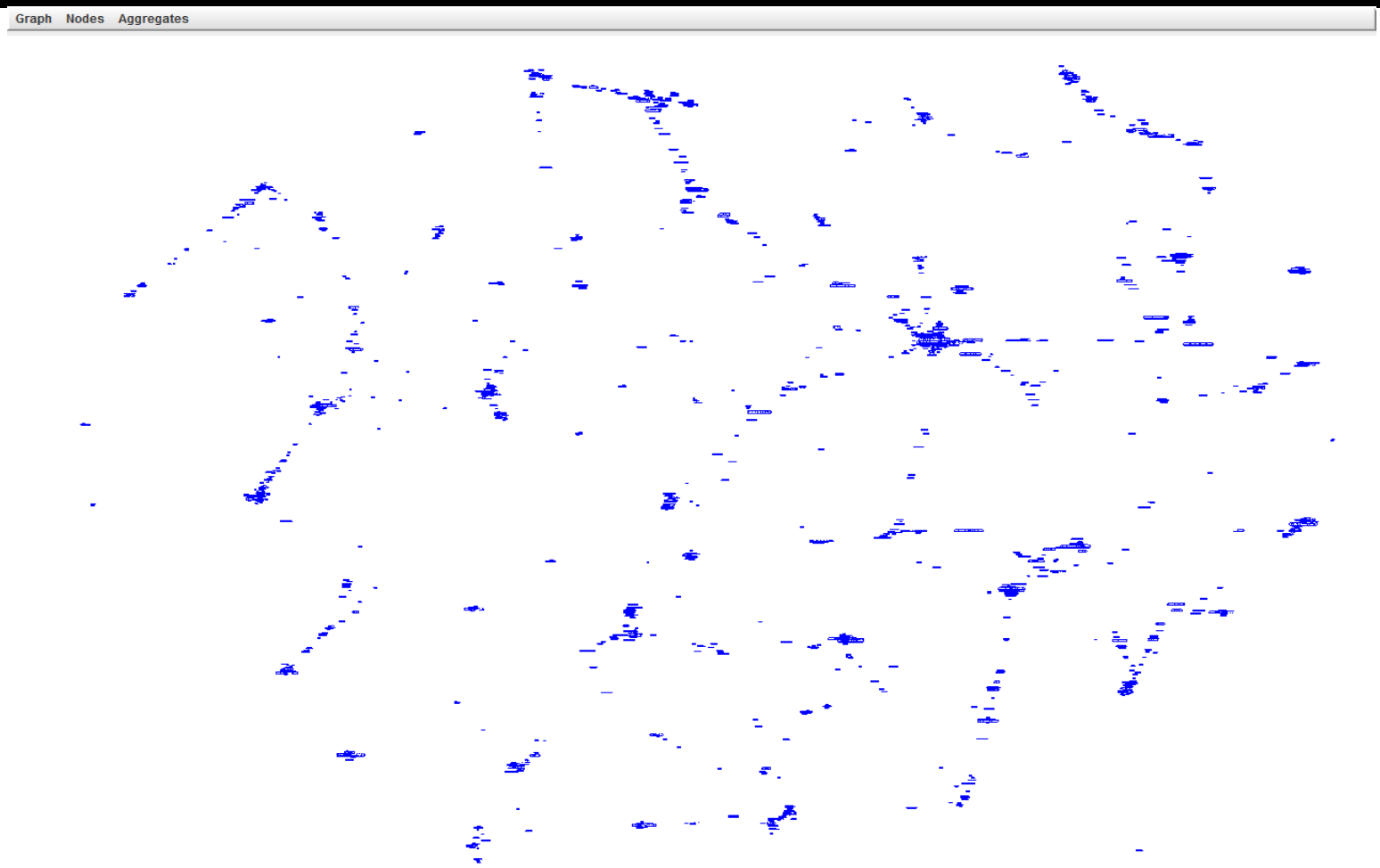
NAVSEA
WARFARE CENTERS
DAHLGREN

The Role of Clustering

- We will use clustering as a means to understand the data
- How shall we encode the data?
- How shall we visualize the data?

NAVSEA
WARFARE CENTERS
DAHLGREN

The Problem Manifests Itself Visually



1500 node cluster solution calculated on the PubMed data

NAVSEA WARFARE CENTERS DAHLGREN

Multi-feature Document Clustering

- Treating each entity as a single sample (i.e. keyword)
  - Compute distance measures between samples using various feature sets.
    - Keywords, Subjects, Text, References, Sources, …
  - Combine distance measures using a scaled mean method to create a single distance measure between samples.
  - Cluster samples.

Dissimilarity Measures Between Topic Areas

- Combination of several vector space models
  - Topic-Term, Topic-Keyword, Topic-Reference, Topic-Source, …

	Term 0	Term 1	…	Term M
Topic 0	F_{00}	F_{01}	…	F_{0M}
Topic 1	F_{10}	F_{11}		F_{1M}
…	…	…	…	…
Topic N	F_{N0}	F_{N1}	…	F_{NM}

F_{ij} – Frequency of term j in topic i

TF-IDF weighting, Cosine Similarity

  - Mean of the normalized dissimilarities

$$d(i,j) = \frac{\sum_{f=1}^{p} \frac{d_{i,j}^{f} - d_{\min}^{f}}{d_{\max}^{f} - d_{\min}^{f}}}{p}$$

d(i,j) – dissimilarity topics i and j

p – number of dissimilarity types (term, keywords, …)

d^f – type f dissimilarity matrix

NAVSEA
WARFARE CENTERS
DAHLGREN

Our Focus in the PubMed Data?

- Keywords
  - These include keywords and medical subject heading (MESH) terms
    - lymphocyte activation, cell cycle proteins, middle aged, …
- Chemicals
  - Iron, soil pollutants, anticoagulants, …

NAVSEA WARFARE CENTERS DAHLGREN

Sample Sorted Distance Results (Cluster PubMed 2000-2005 Chemicals)

Chemical Keyword Vectors

cdc2-cdc28 kinases	cyclin dependent kinase 2	7.13E-04
atorvastatin	heptanoic acids	0.001762
mycophenolate mofetil	mycophenolic acid	0.003505
cyclooxygenase 2	prostaglandin endoperoxide synthases	0.005044
clopidogrel	ticlopidine	0.006101
endothelial growth factors	vascular endothelial growth factors	0.006908
factor v	factor v leiden	0.011455
cftr protein human	cystic fibrosis transmembrane conductance regulator	0.011744
ghrelin	peptide hormones	0.013936
mitogen activated protein kinase 1	mitogen activated protein kinase 3	0.014455

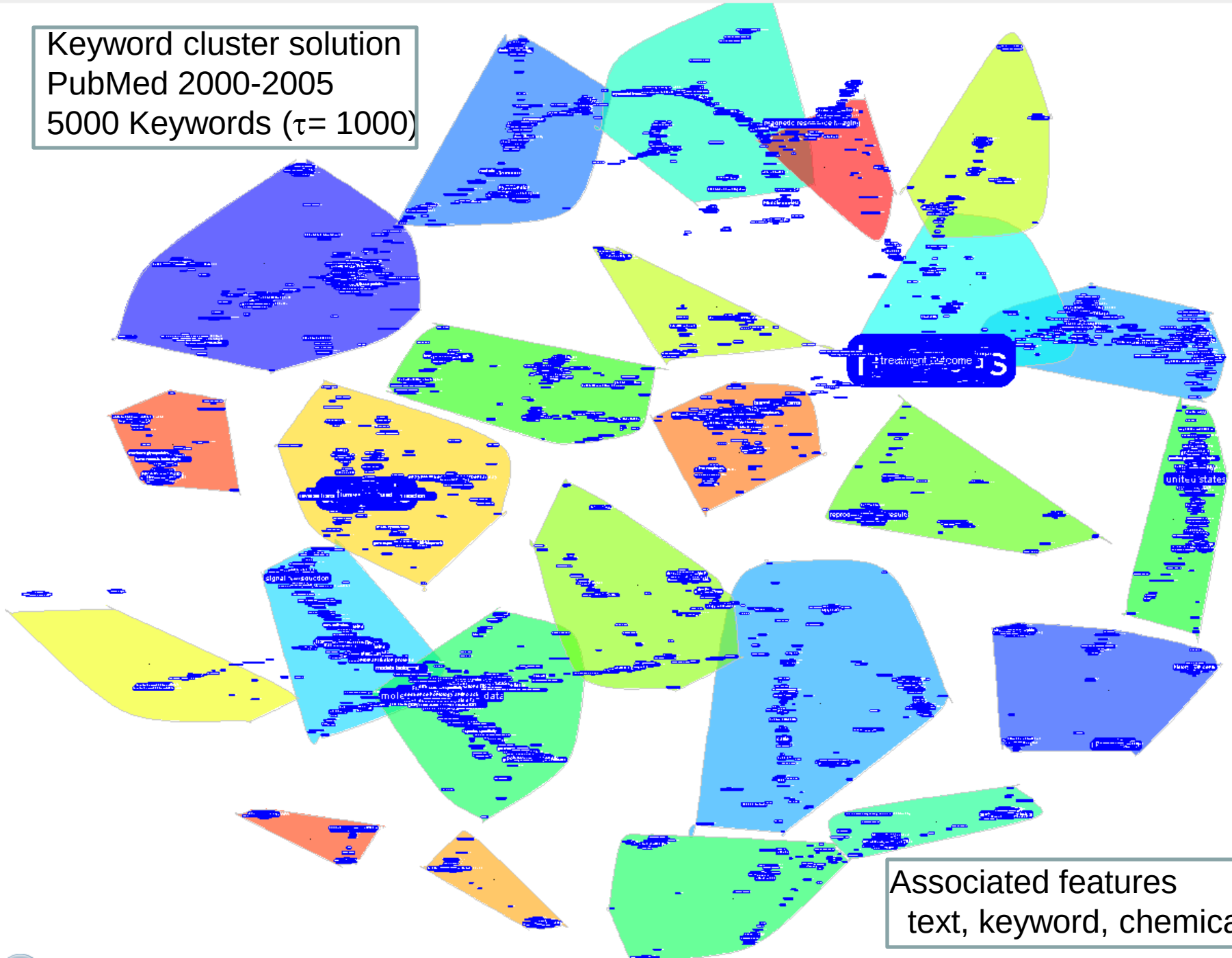
Chemical Term Vectors

atorvastatin	heptanoic acids	0.001716614
ghrelin	peptide hormones	0.002913594
hif1a protein human	hypoxia inducible factor 1 alpha subunit	0.004588425
cdc2-cdc28 kinases	cyclin dependent kinase 2	0.005242527
cftr protein human	cystic fibrosis transmembrane conductance regulator	0.005525887
endothelial growth factors	vascular endothelial growth factors	0.005641282
factor v	factor v leiden	0.00591284
mycophenolate mofetil	mycophenolic acid	0.006154239
beta catenin	ctnnb1 protein human	0.008612394
cyclooxygenase 2	prostaglandin endoperoxide synthases	0.00951004

Chemical Chemical Vectors

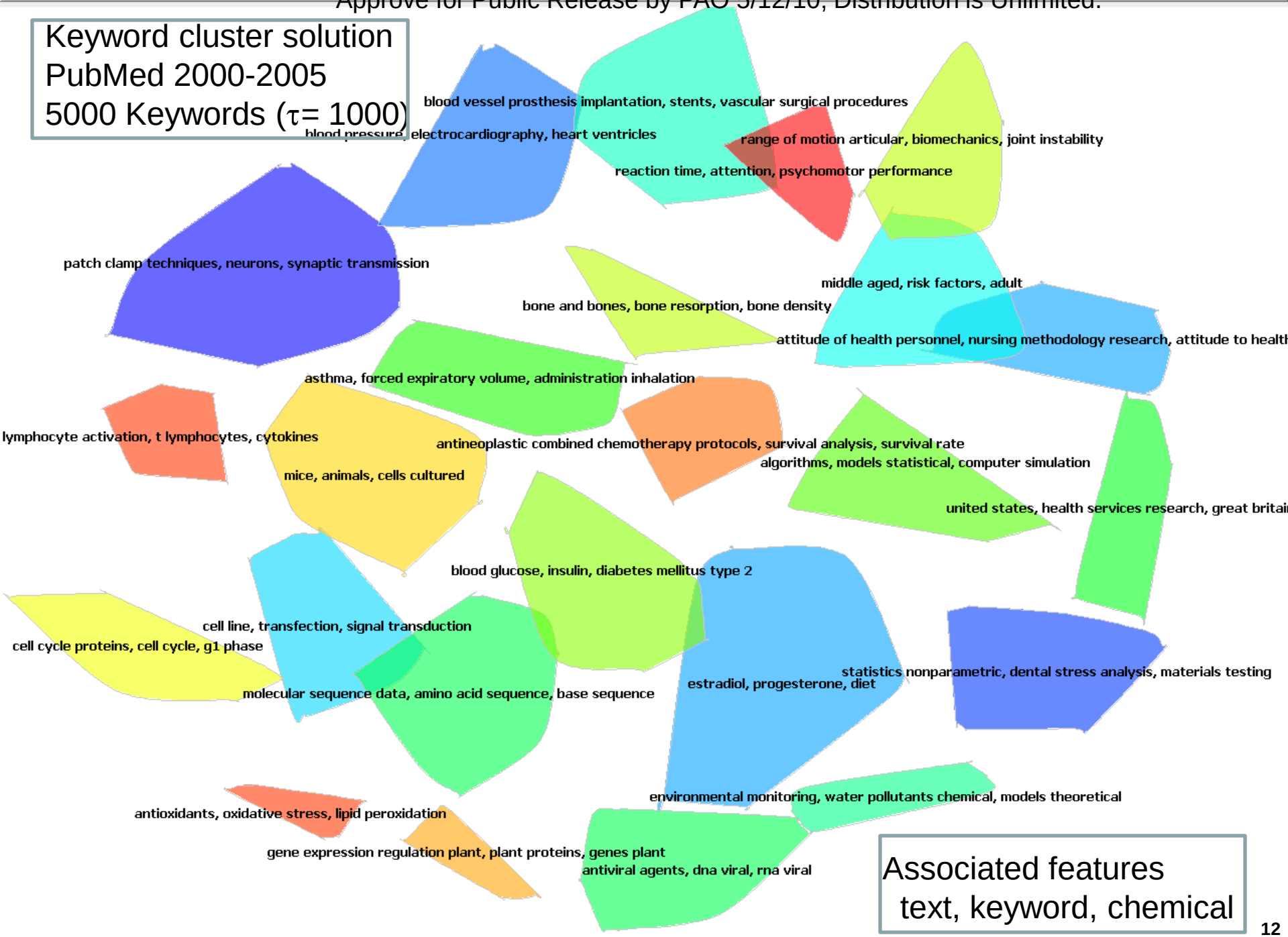
cdc2-cdc28 kinases	cyclin dependent kinase 2	7.13E-04
atorvastatin	heptanoic acids	0.001762271
mycophenolate mofetil	mycophenolic acid	0.003504753
cyclooxygenase 2	prostaglandin endoperoxide synthases	0.005044222
clopidogrel	ticlopidine	0.006100714
endothelial growth factors	vascular endothelial growth factors	0.006907523
factor v	factor v leiden	0.011455357
cftr protein human	cystic fibrosis transmembrane conductance regulator	0.01174432
ghrelin	peptide hormones	0.013935566
mitogen activated protein kinase 1	mitogen activated protein kinase 3	0.014455438

Keyword cluster solution
PubMed 2000-2005
5000 Keywords (τ= 1000)



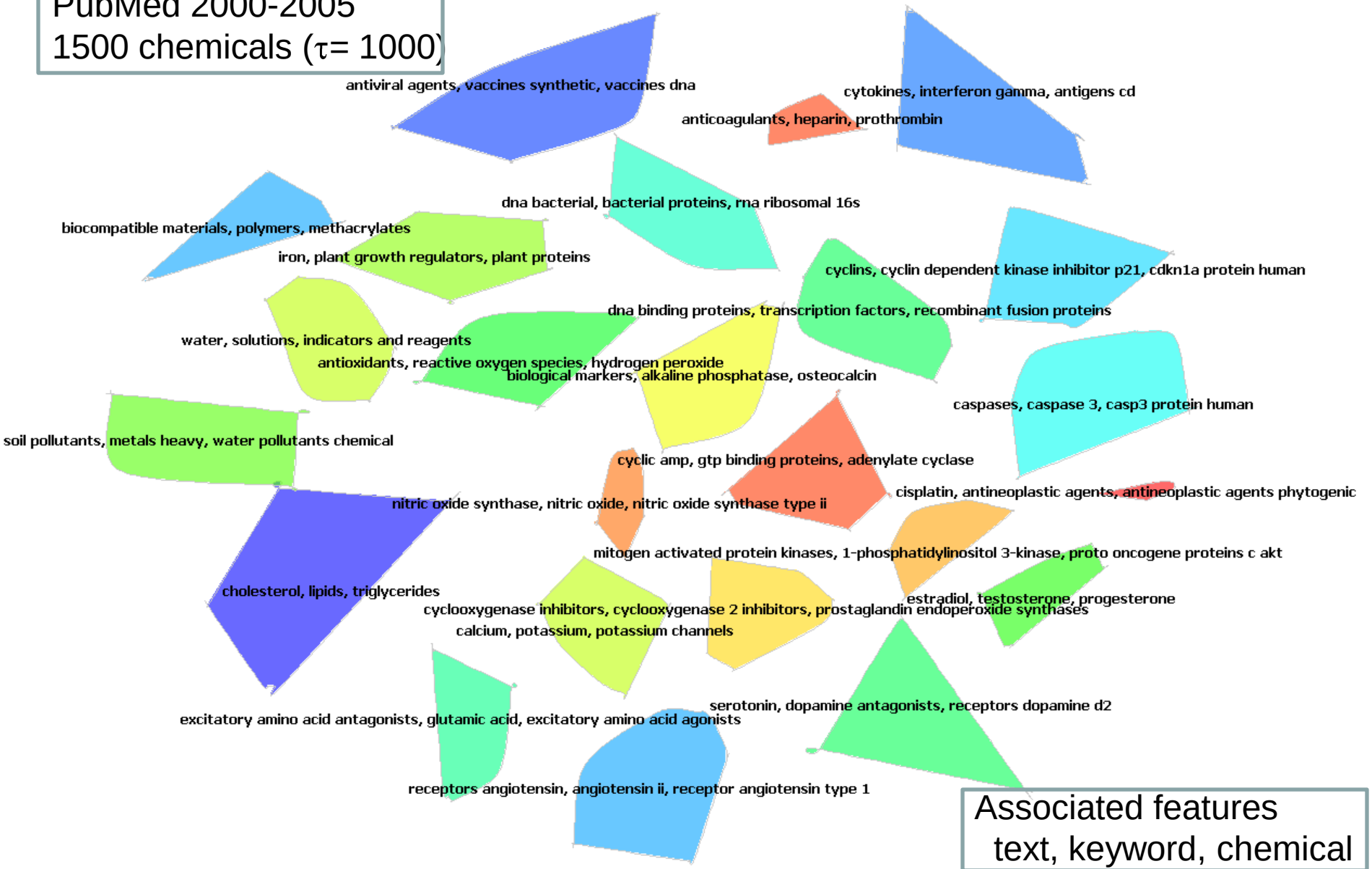
Associated features
text, keyword, chemical

Keyword cluster solution
PubMed 2000-2005
5000 Keywords (τ= 1000)

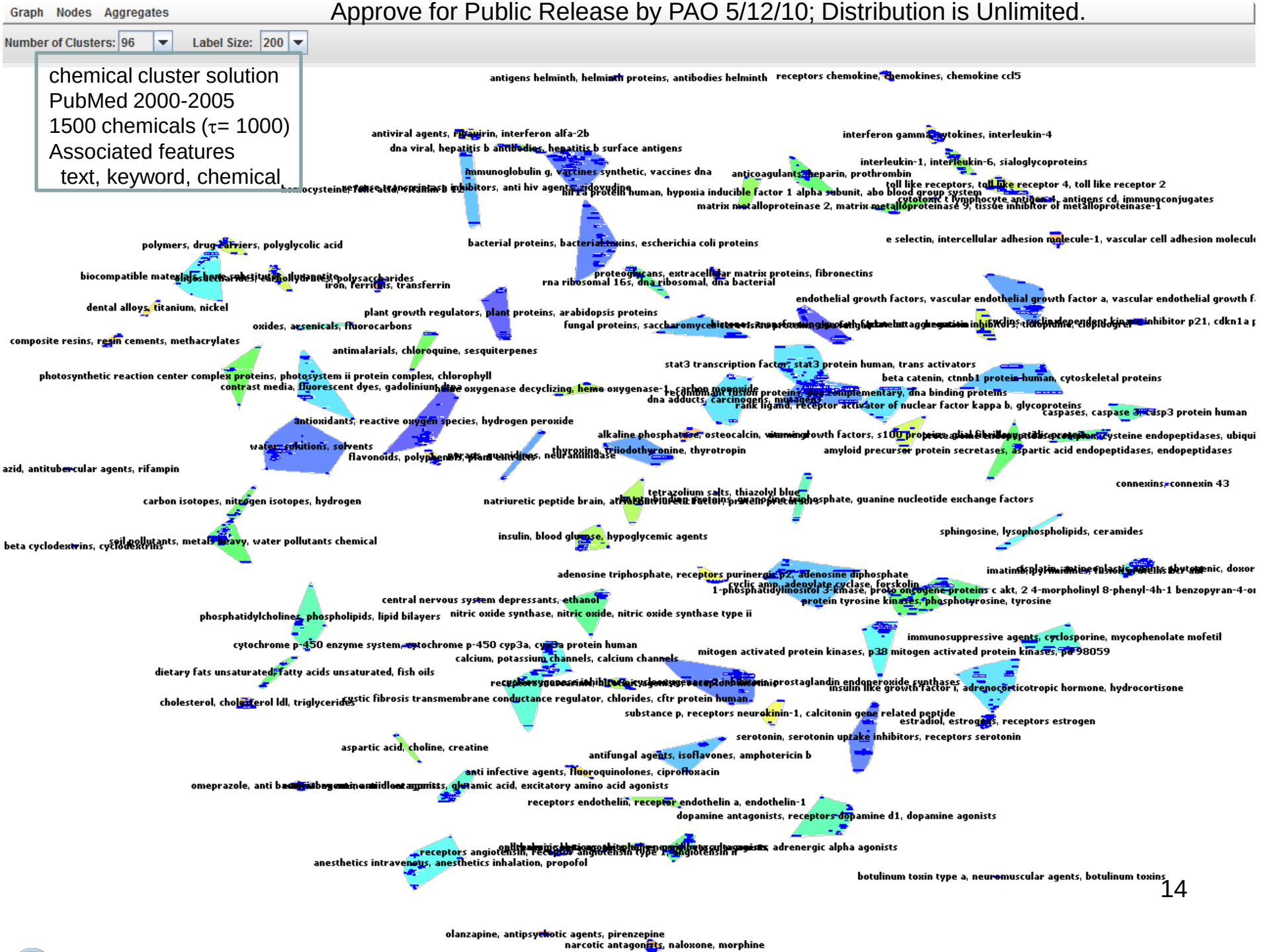


Associated features
text, keyword, chemical

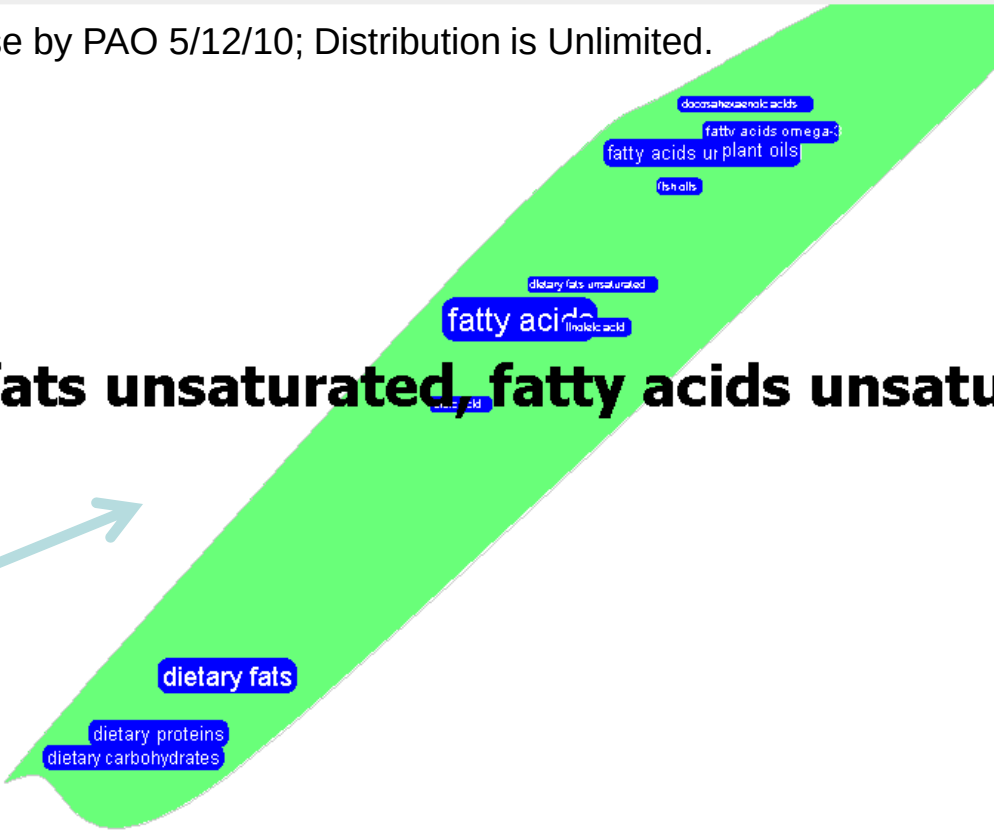
chemical cluster solution
PubMed 2000-2005
1500 chemicals (τ= 1000)



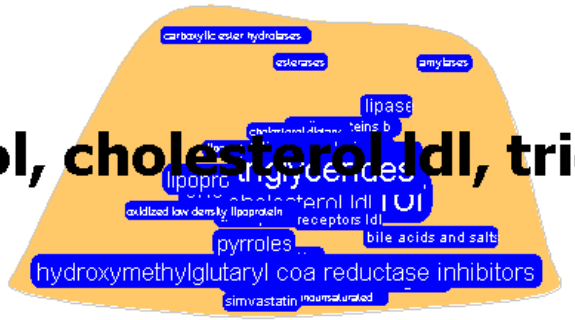
Associated features
text, keyword, chemical

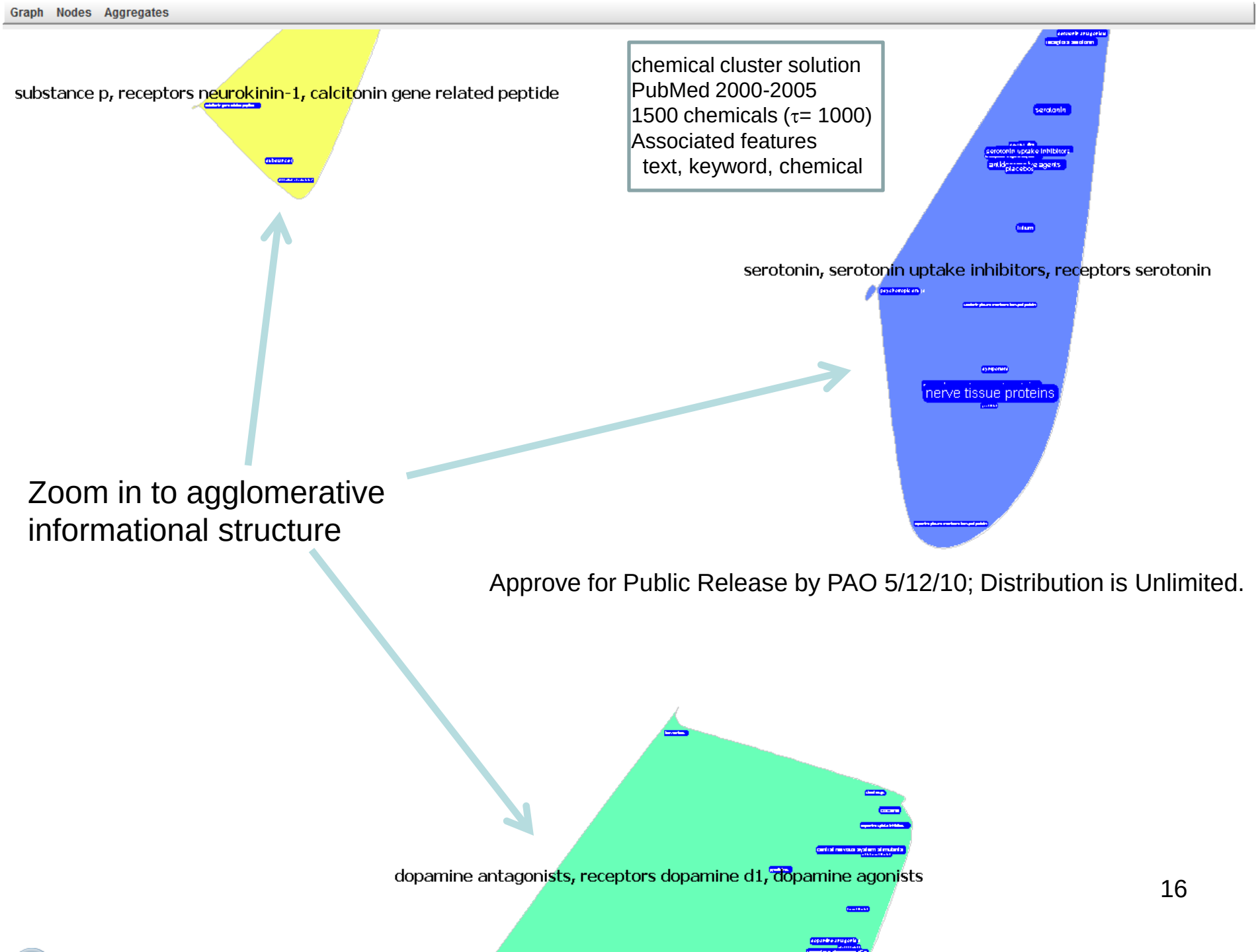


chemical cluster solution
PubMed 2000-2005
1500 chemicals (τ= 1000)
Associated features
text, keyword, chemical

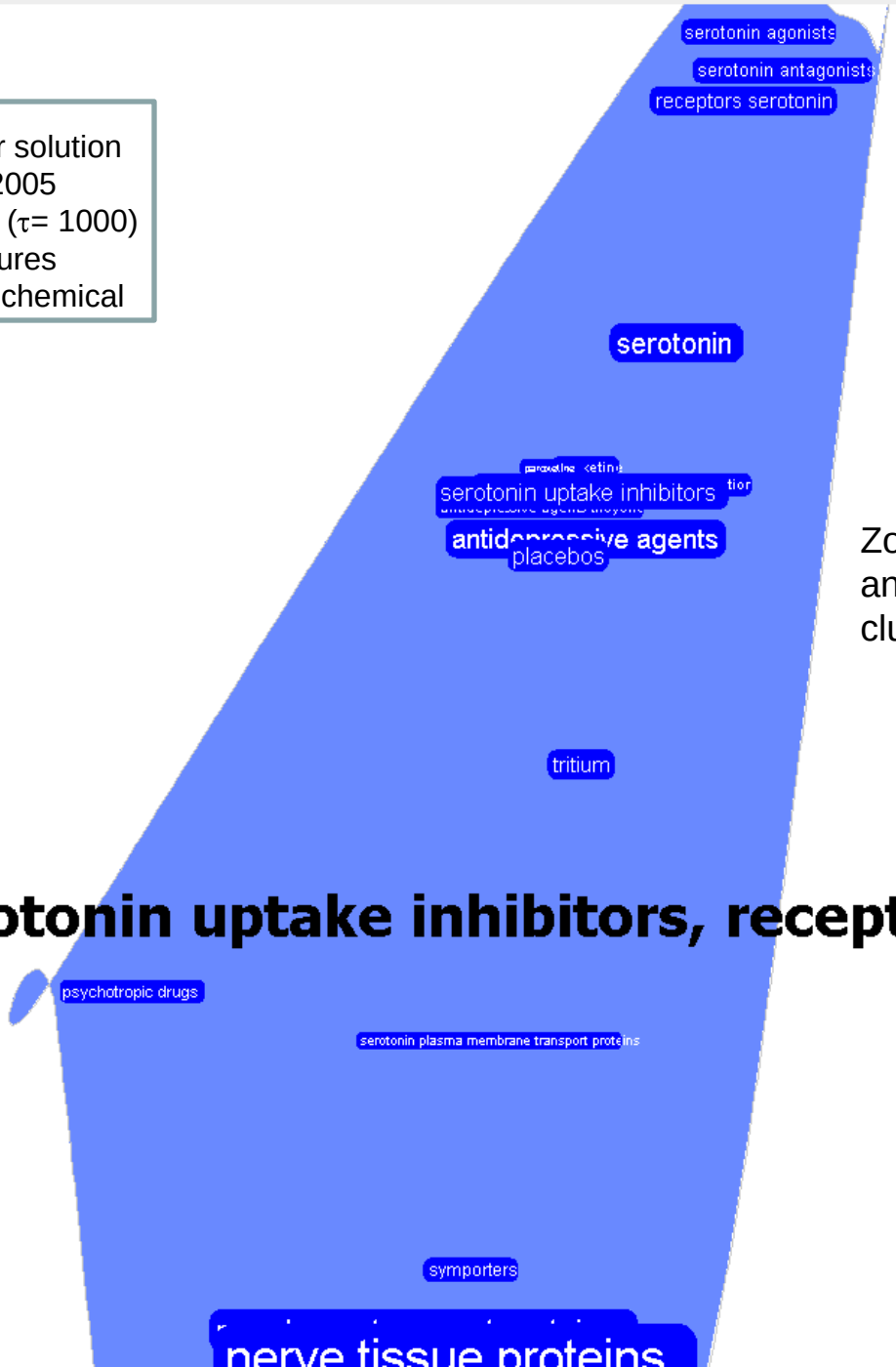


Zoom in to reveal agglomerative informational structure




Graph Nodes Aggregates
substance p, receptors neurokinin-1, calcitonin gene related peptide
chemical cluster solution
PubMed 2000-2005
1500 chemicals (τ= 1000)
Associated features
text, keyword, chemical
serotonin, serotonin uptake inhibitors, receptors serotonin
nerve tissue proteins
Zoom in to agglomerative informational structure
Approve for Public Release by PAO 5/12/10; Distribution is Unlimited.
dopamine antagonists, receptors dopamine d1, dopamine agonists

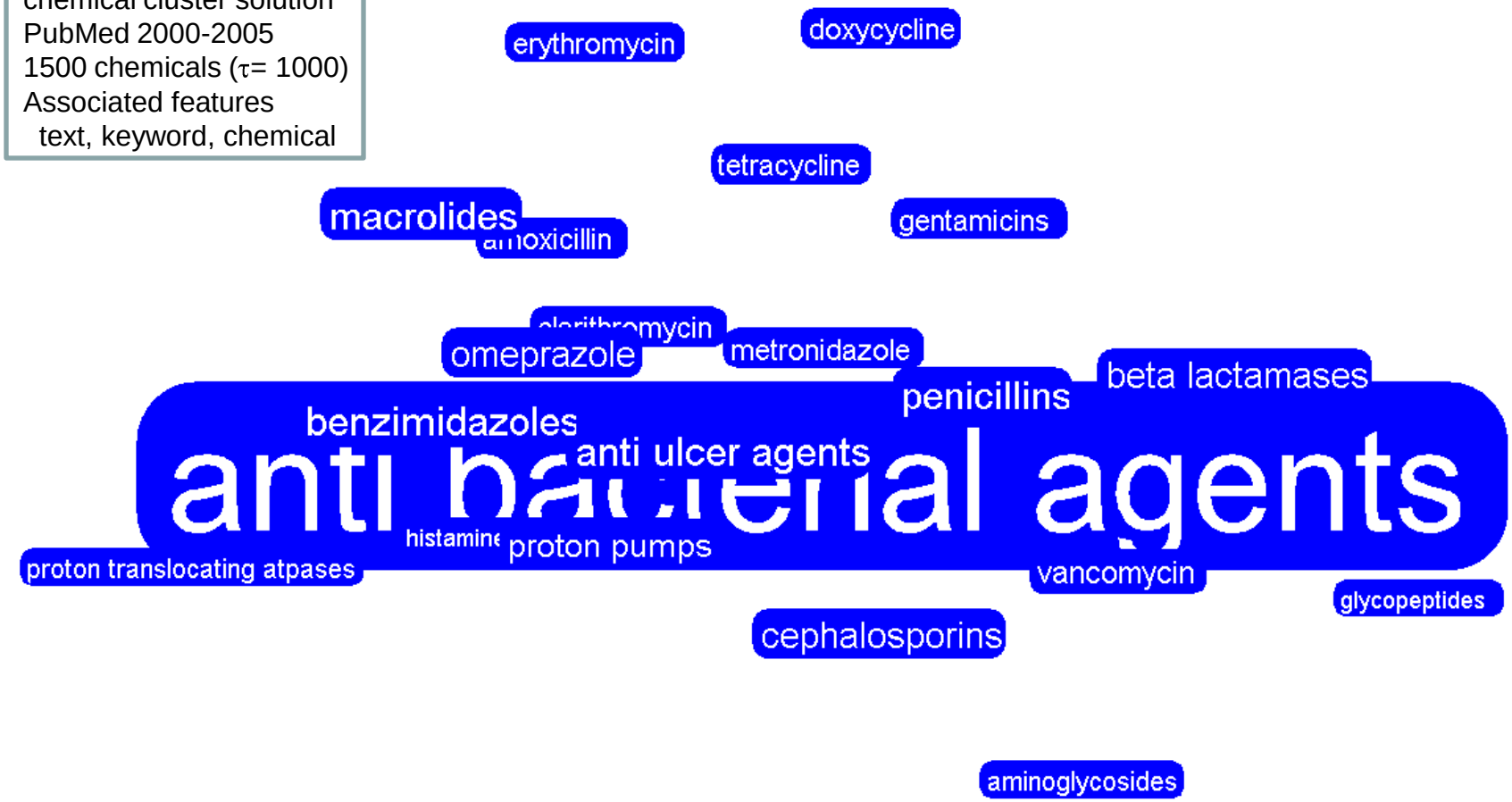
chemical cluster solution
PubMed 2000-2005
1500 chemicals (τ= 1000)
Associated features
text, keyword, chemical



Zoom in to reveal individual chemicals and their relationships in the serotonin cluster

rotonin, serotonin uptake inhibitors, receptors serotonin

chemical cluster solution
PubMed 2000-2005
1500 chemicals (τ= 1000)
Associated features
text, keyword, chemical



Zoom in to reveal individual chemicals and their relationships in the antibacterial and ulcer cluster

NAVSEA
WARFARE CENTERS
DAHLGREN

Wrap-up

- The future
  - Integration of this prototype software into our other analytical environments
  - Condition the plot based on ancillary data
  - Network visualization
  - Full content/multi-lingual exploitation