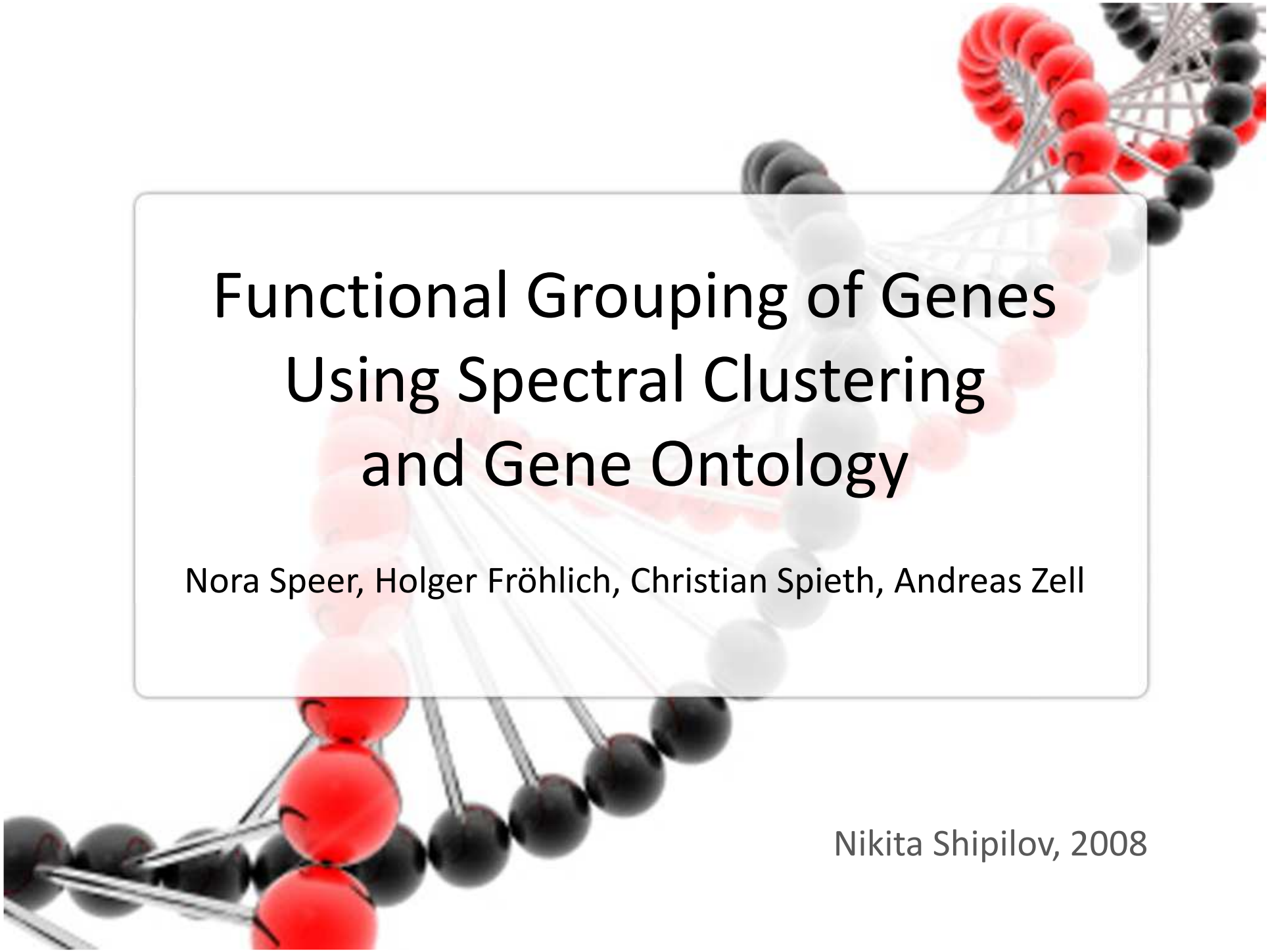


Functional Grouping of Genes Using Spectral Clustering and Gene Ontology

Nora Speer, Holger Fröhlich, Christian Spieth, Andreas Zell

Nikita Shipilov, 2008

Abstract

With the invention of high throughput methods, researchers are capable of producing large amounts of biological data. During the analysis of such data the need for a functional grouping of genes arises. In this paper, we propose a new method based on **spectral clustering** for the partitioning of genes according to their biological function. The functional information is based on **Gene Ontology** annotation, a mechanism to capture functional knowledge in a shareable and computer processable form. Our functional cluster method promises to automatize, speed up and therefore improve biological data analysis.

N. Speer, H. Fröhlich, C. Spieth, A. Zell

Gene Ontology

The Gene Ontology project, or **GO**, provides a controlled vocabulary to describe gene and gene product attributes in any organism.

[*http://www.geneontology.org/*]

The aim of the project is a collaborative effort to address the need for consistent descriptions of gene products in different databases.

Gene Ontology

Consists of two parts:

- Key concepts: the *molecular function* of gene products; their role in multi-step *biological processes*; and their localization to *cellular components*;
- the characterization of gene products using terms from the ontology.

A gene product might be associated with or located in one or more *cellular components*; it is active in one or more *biological processes*, during which it performs one or more *molecular functions*.

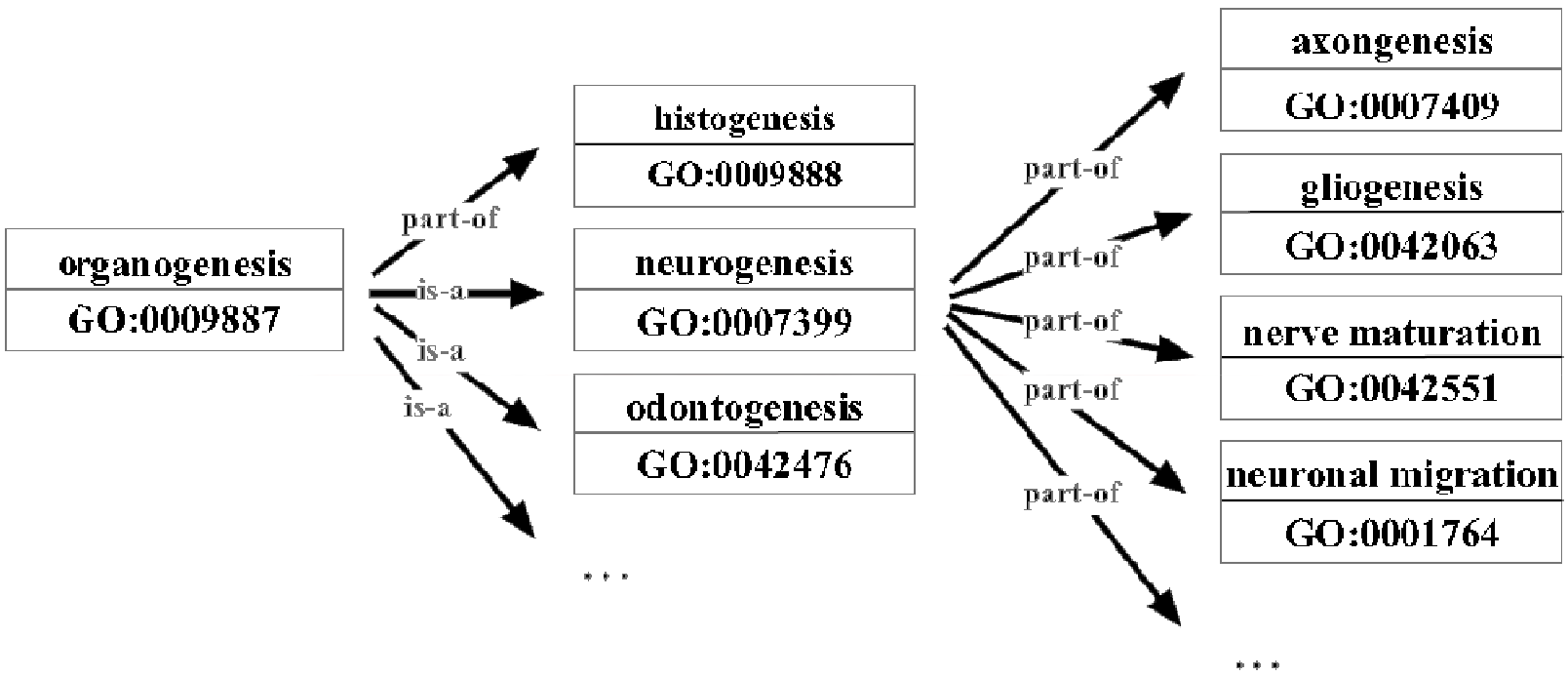
Gene Ontology

Each entry in GO has a unique numerical identifier of the form *GO:nnnnnnn*, and a **term** name, e.g. *cell*, *fibroblast growth factor receptor binding or signal transduction*.

Terms are classified into only one of the three ontologies.

The ontologies are structured as directed acyclic graphs (**DAG**), which are similar to hierarchies but differ in that a more specialized term (child) can be related to more than one less specialized term (parent).

Gene Ontology



Relations are: “is-a”, “part-of”

More about Gene Ontology

http://www.geneontology.org/

Methods

- Distances within the Gene Ontology
- Distances between Genes Using Feature Vectors
- Spectral Clustering using FV Representation
- Cluster Validity

Distances within the Gene Ontology

The **information content** of a term is defined as the probability of occurrence of this term or any child term in a dataset.

Following the notation in information theory, the IC of a term *c* can be quantified as follows:

$$\boldsymbol{IC}(c) = -ln\ \boldsymbol{P}(c)$$

$\boldsymbol{P}(c) = \frac{freq(c)}{N}$, where *N* is the total number of terms occurring in the dataset and *freq*(*c*) is the number of times term *c* or any child term of *c* occurs in the dataset.

Distances within the Gene Ontology

The similarity of two terms c_i, c_j can then be defined as followed:

$$\boldsymbol{sim}(c_i, c_j) = -ln \min_{c \in S(c_i, c_j)} P(c) = -lnP_{ms}(c_i, c_j)$$

where $S(c_i, c_j)$ is the set of parental terms shared by both c_i and c_j.

As the GO allows multiple parents for each term, two terms can share parents by multiple paths. We take the minimum $P(c)$, if there is more than one parent. This is called *the probability of the minimum subsumer*.

Distances within the Gene Ontology

The inverse of similarity is distance measure:

$$\boldsymbol{d}(c_i, c_j) = 2\, ln\, P_{ms}(c_i, c_j) - (ln\, P(c_i) + ln\, P(c_j))$$

Developed by J.J. Jiang and D.W. Conrath

Since genes are often annotated with more than one GO term, multiple functional distances can be computed between two genes. Therefore, we need to combine all or choose one of the calculated distances (best distance).

Distances within the Gene Ontology

Obviously, distance measurement causes a loss of information.

Additionally, the problem with using the best GO-distance between two genes *x* and *y* is that it can be 0, even if two genes are not identical, because they belong to the same functional class. This prevents from using distance directly as a metric for clustering.

Both problems can be solved by using a **feature vector representation** for each gene.

Methods

- Distances within the Gene Ontology
- Distances between Genes Using Feature Vectors
- Spectral Clustering using FV Representation
- Cluster Validity

Distances Using Feature Vectors

For each gene x we construct a feature vector $\phi_p(x)$ relative to prototypes $\mathrm{p} = (p_1, \ldots, p_N)^T$

$$\phi_p(x) = (d(x, p_1), \ldots, d(x, p_N))^T$$

Prototypes are just all genes from data set.

Now, the distance between two genes x and y is simply given by

$$d(x, y) = ||\phi(x) - \phi(y)||$$

Distances Using Feature Vectors

More specifically, we have the equality

$$d^2(x, y) = ||\phi(x) - \phi(y)||^2$$

$$= <\phi(x), \phi(x)> - 2<\phi(x), \phi(y)> + <\phi(y), \phi(y)>$$

$$= k(x, x) - 2k(x, y) + k(y, y)$$

That means by defining ϕ: X → H we map our data into some Hilbert space H. The scalar product in this space defines a kernel $k : X \times X \rightarrow R$ and hence a similarity measure between two genes x and y in our original input space X.

Methods

- Distances within the Gene Ontology
- Distances between Genes Using Feature Vectors
- Spectral Clustering using FV Representation
- Cluster Validity

Spectral Clustering

Given our representation of each gene as a feature vector, we can choose any clustering algorithm to group our data.

One of them is (Ng, Jordan, Weiss):

given the distance measure d on data $x_1, \ldots, x_n$; one computes the k largest eigenvalues and corresponding Eigenvectors of the graph Laplacian $L = D^{-1/2}KD^{-1/2}$ where $K = (\exp(d^2(x_i, x_j)/2\sigma^2))_{ij}$ and D is a diagonal matrix with $D_{jj} = \Sigma_i K_{ij}$.

Eigenvectors are then clustered e.g. by k-means.

Methods

- Distances within the Gene Ontology
- Distances between Genes Using Feature Vectors
- Spectral Clustering using FV Representation
- Cluster Validity

Clustering Validity

The **Silhouette value** for each point is a measure of how similar that point is to points in its own cluster vs. points in other clusters, and ranges from -1 to +1. It is defined as:

$$S(i) = \frac{\min(\overline{d_B}(i,j)) - \overline{d_W}(i)}{\max(\overline{d_W}(i),\ \min(\overline{d_B}(i,j)))}$$

where $\overline{d_W}(i)$ is the average distance from the *j*-th point to the other points in its own cluster, and $\overline{d_B}(i,j)$ is the average distance from the *i*-th point to points in another cluster *j*.

Experiments

Two datasets:

- The response of human fibroblasts to serum on cDNA microarrays. Only 238 genes out of 517 showed one or more GO mappings to *biological process* or a child term of *biological process*.
- The transcriptional profiling of human fibroblasts during cell cycle, 233 out of 388.

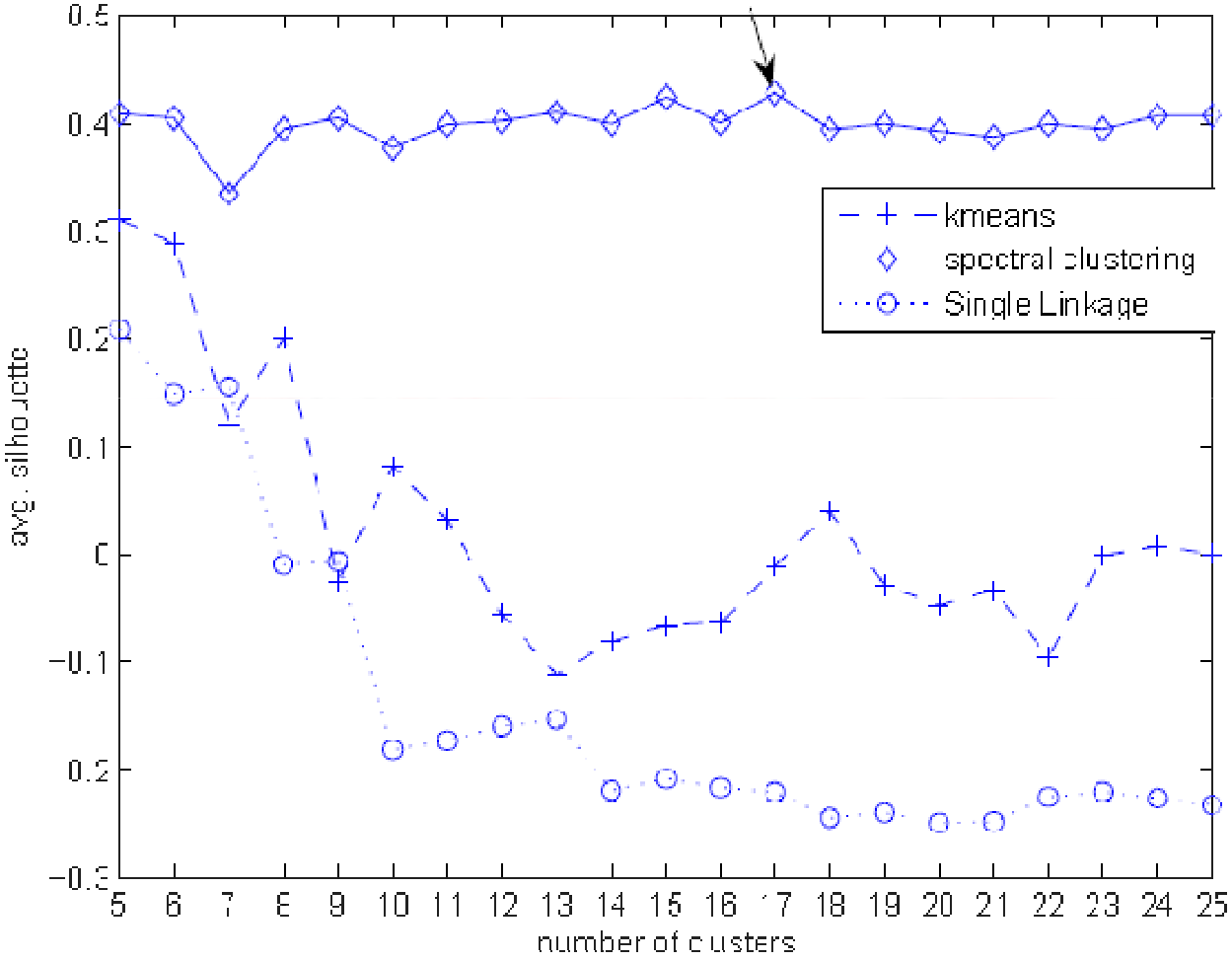
Duplicate experiments were carried out at 13 different time points ranging from 0 to 24 hours.

Results

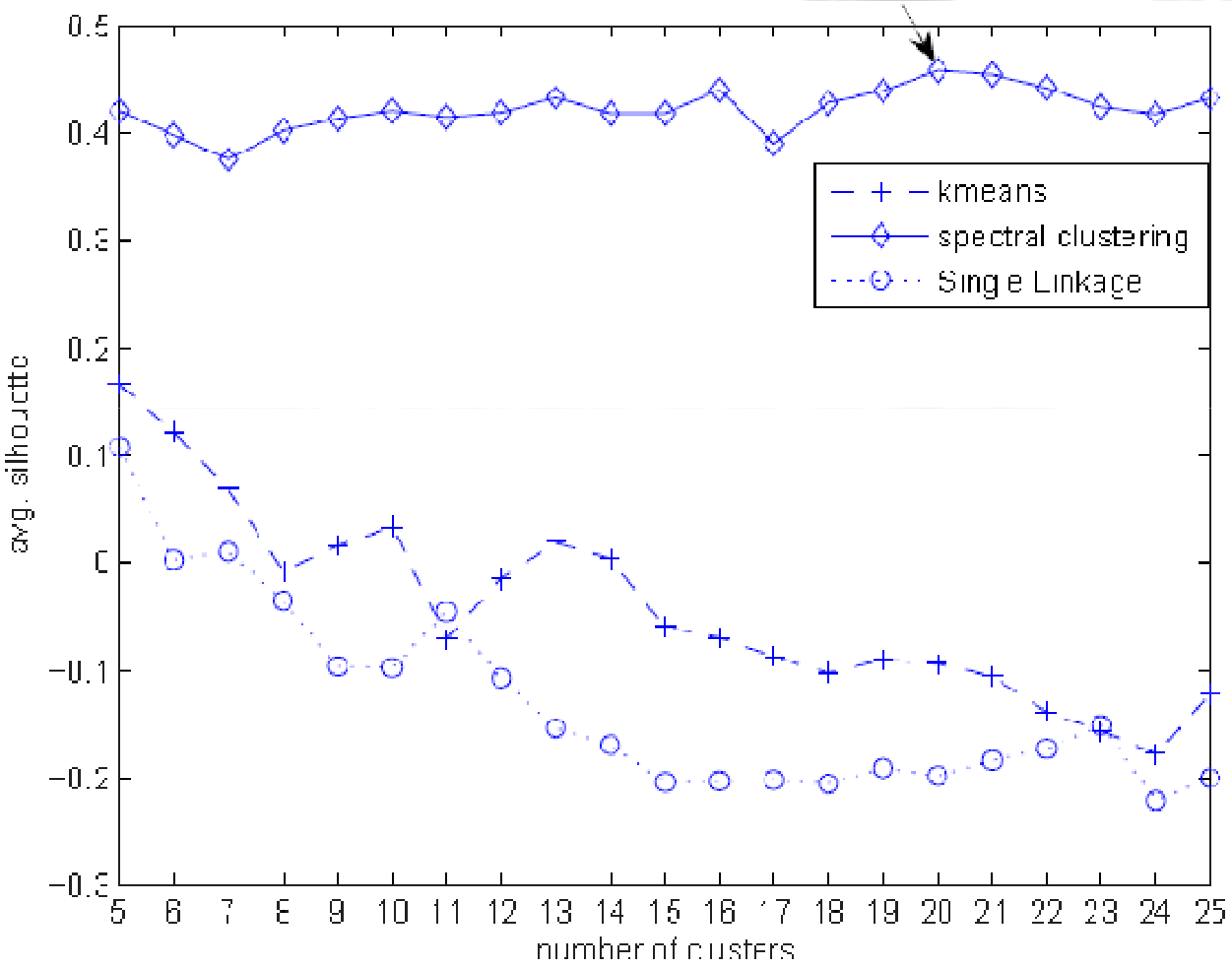
In the experiments, Spectral Clustering was compared to k-means and Single Linkage clustering which are also based on the proposed feature vector representation.

Validity was controlled by Silhouette index.

Results (dataset I)



Results (dataset I)



Conclusion

- Distances within the Gene Ontology
- Distances between Genes Using Feature Vectors
- Spectral Clustering using FV Representation
- Cluster Validity
- Test

Tutorial on Spectral Clustering

http://www.kyb.mpg.de/publications/attachments/Luxburg06_TR_%5B0%5D.pdf


Thank you!