Evaluation of Hierarchies based on the Longest Common Prefix, or Baire, Metric

Pedro Contreras and Fionn Murtagh
{pedro, fionn}@ cs.rhul.ac.uk

Department of Computer Science
Royal Holloway, University of London

CSNA, June 9, 2007

Overview

- Introduction. The Baire metric concept
- Applying Baire metric to chemical data
- Clustering chemical data based on a Baire distance
- Comparison result
- Current work
- Conclusions
- References

Introduction

Baire space consists of countable infinite sequence with a metric defined in terms of the longest common prefix [A. Levi. Basic set theory, Dover, 1979 (reprinted 2002)] (The longer the common prefix, the closer a par of sequence)

Consider two floating point numbers with the first p digits identical. Then what we call their Baire distance is 2^{-p}. This distance is an ultrametric. [see http://www.cs.rhul.ac.uk/~fionn/papers]

Introduction

It follows that a hierarchy can be used to represent the relationships associated with this distance.

We address the issue of whether such a hierarchy is advantageous, computationally, for clustering large, high dimensional data sets

We seek to find inherent hierarchical structure in data, rather than fitting a hierarchy structure to data (as is traditionally used in multivariate data analysis).

Baire, or longest common prefix, ultrametric

Case of vectors x and y, with 1 attribute. Precision: digits 1, 2, ..., |K|

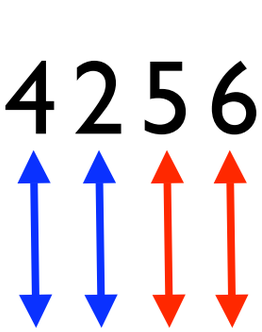
$$d_B(x_K, y_K) = \begin{cases} 1 & \text{if } x_1 \neq y_1 \\ \inf 2^{-n} & x_n = y_n \quad 1 \leq n \leq |K| \end{cases}$$

- Each coordinate is normalized, so is a floating point value.
- Then: we will define $d_B(x,y)$ based on sharing common prefix in all coordinates.

Baire, or longest common prefix

An example of Baire distance for two numbers (x and y) using a precision of 4

$x = 0.4256$



$y = 0.4278$

Baire distance between x and y:

$d_B(x_4, y_4) = 2^{-3} = |K| = 3$

That is:

k=1 -> $x_k = y_k$ -> 4

k=2 -> $x_k = y_k$ -> 2

k=3 -> $x_k \neq y_k$ -> 5≠7

Motivation: Matching of Chemical Structures

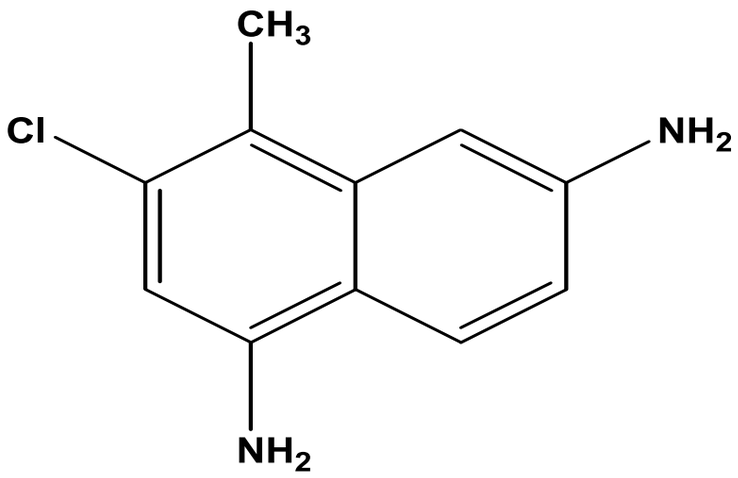
One of the most common problems in mining large chemical libraries is classifying the compounds into different classes.

Different classes could represent different levels of activity or could represent different types of compounds.

Motivation: Matching of Chemical Structures

- Clustering of compounds based on chemical descriptors or chemical representations, in the pharmaceutical industry.
- Used for screening large corporate databases.
- Chemical warehouses are expanding due to mergers, acquisitions, and the synthetic explosion brought about by combinatorial chemistry.
- We have started looking for local or global ultrametric characteristics on 1.2 million structures, with around 1500 descriptors. Later: larger sets.

Binary Fingerprints



Encode

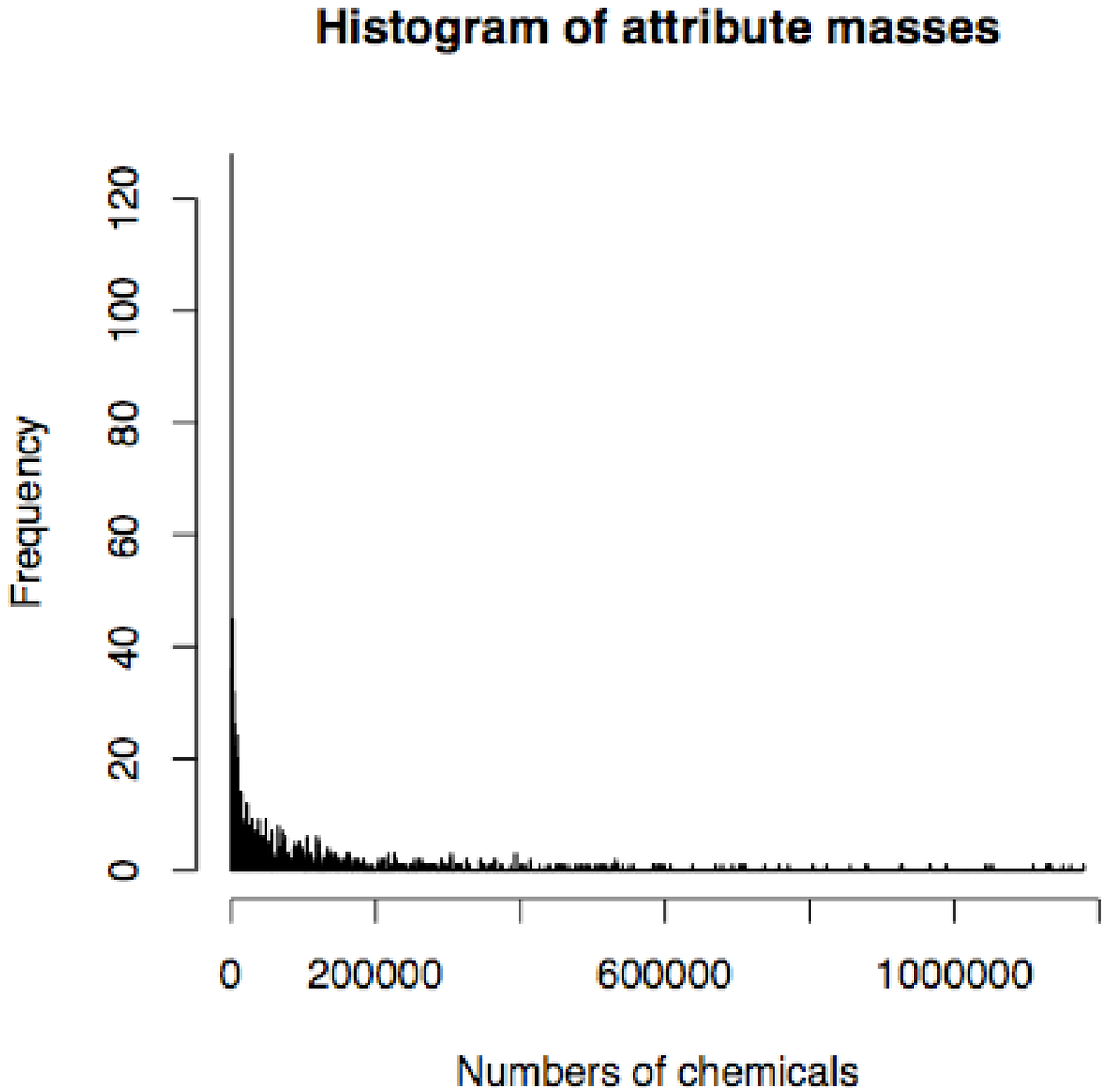
1 0 0 0 1 0 0 0 1 ...1

Fixed length bit strings such as:

- Daylight
- MDL
- BCI
- etc.

Data Characterization

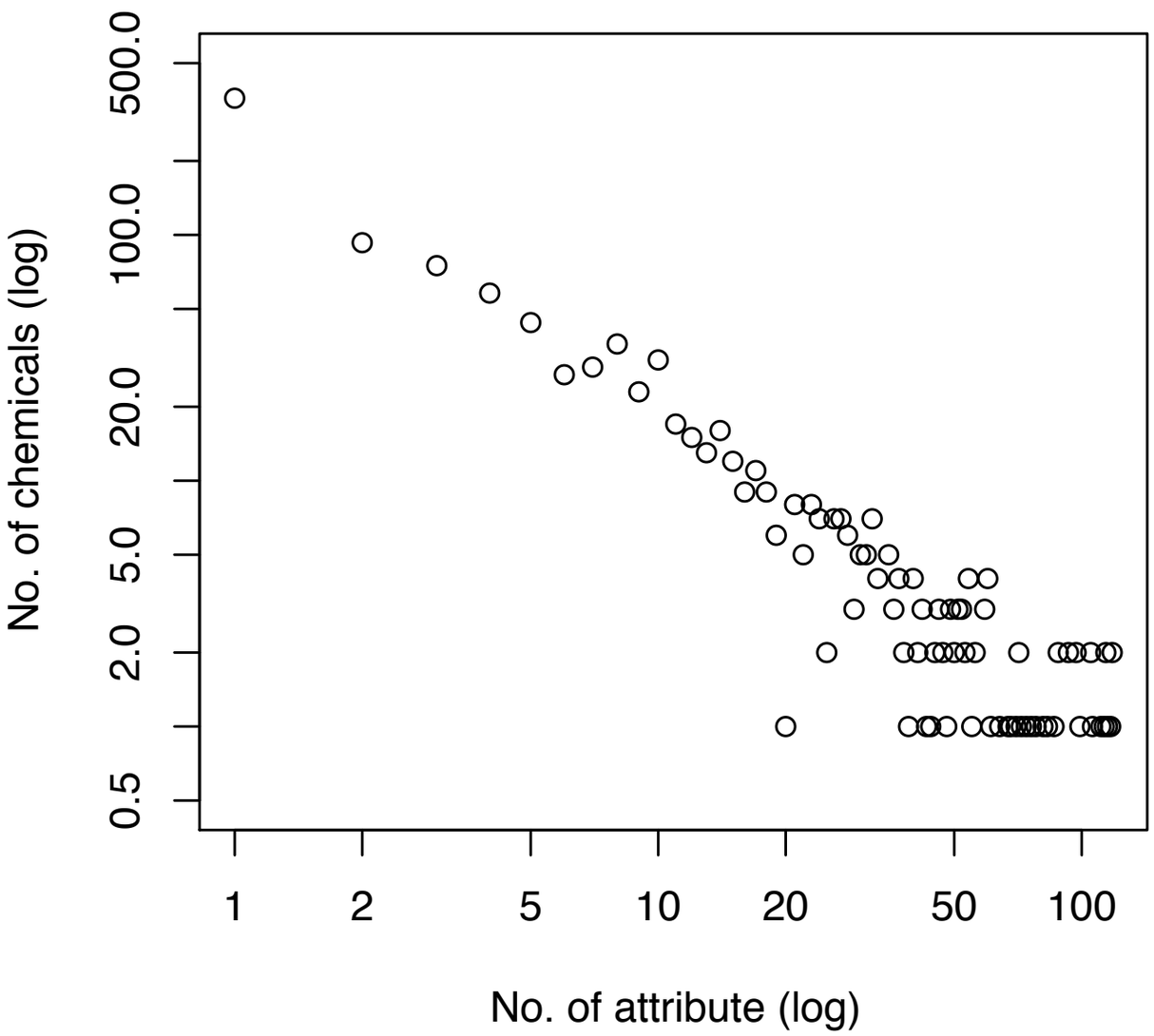
Data characteristics: 1.2M chemicals crossed by 1052 presence/absence attributes.



Histogram of column sums

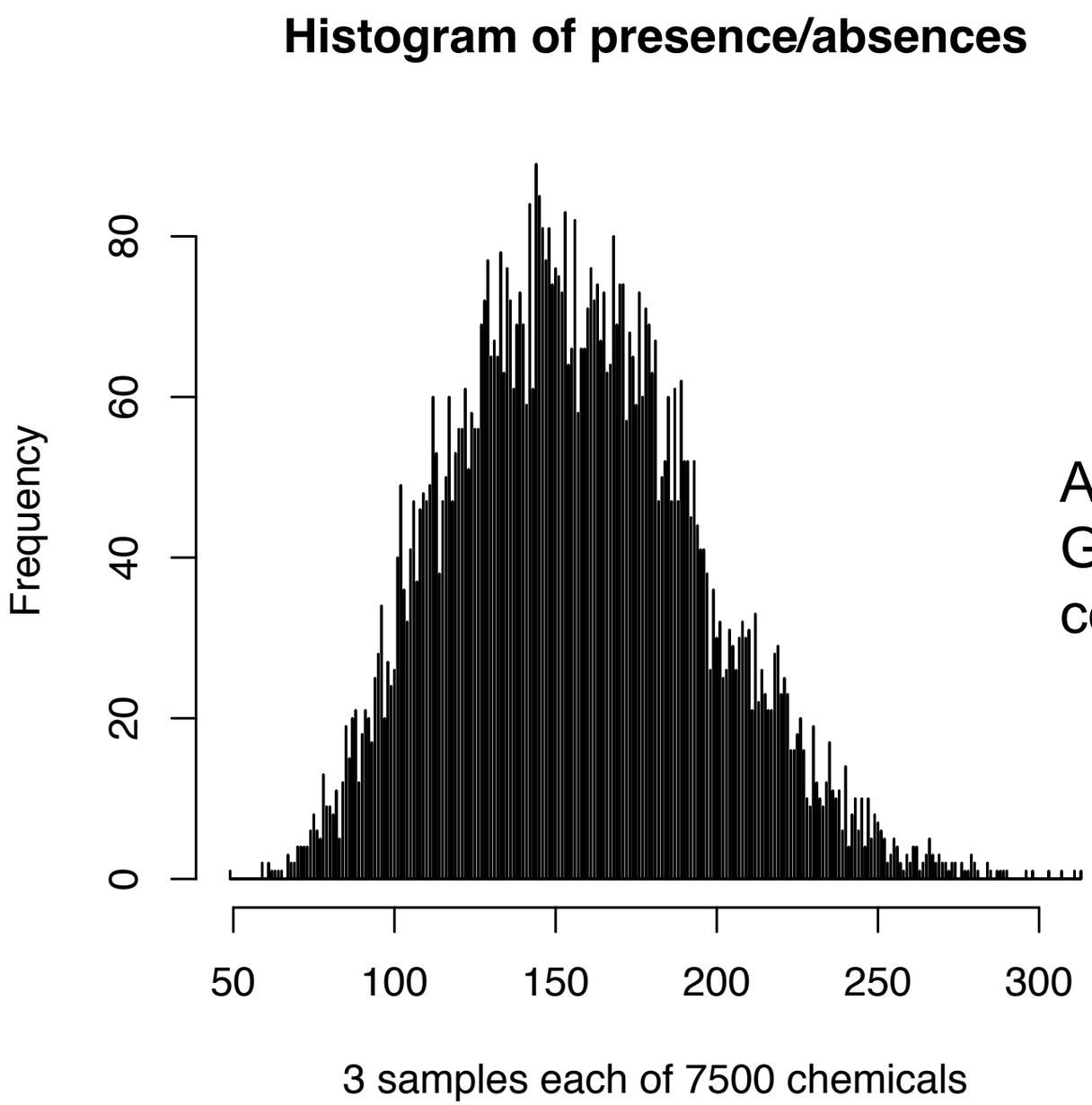
Data Characterization

Data characteristics: 1.2M chemicals crossed by 1052 presence/absence attributes.



Chemicals per attribute follow a power law. Find: probability of having more than *p* chemicals per attribute to be approximately $c / p^{1.23}$ for large *p* and for constant, c.

Data Characterization



Attributes per chemical are approximately Gaussian. Three different sub sets of chemicals contribute to this histogram.

Data Characterization

Then we have a data set that is:

- Highly sparse, occupancy is 8.6347%
- Attributes per chemical are ≈ Gaussian
- Chemical per attributes follows power law with exponent ≈ 1.23

Simple clustering hierarchy

- For precision k, k = 1, 2, ..., |K|
- For attribute set, J
- Determine random projections of all chemical vectors
- Sort projected values; determine identical values, to define cluster, implying chemicals that are projected into same value
- Find: for sets of 7500 chemicals, approx. 140 clusters at precision (number of digits) 1; approx. 2550 clusters at precision 2; approx. 6400 clusters at precision 3
- Appraisal of precision 1 case vis-à-vis k-means shows considerable similarity of results

Random projection schematically

A =

0	1	1	0
0	0	1	0
0	0	1	0
1	1	0	1

matrix Normalized by column sums

N =

0	0.5	0.33	0
0	0	0.33	0
0	0	0.33	0
1	0.5	0	1

Random vector; k = 2

R =

0.13	0.45	0.76	0.49

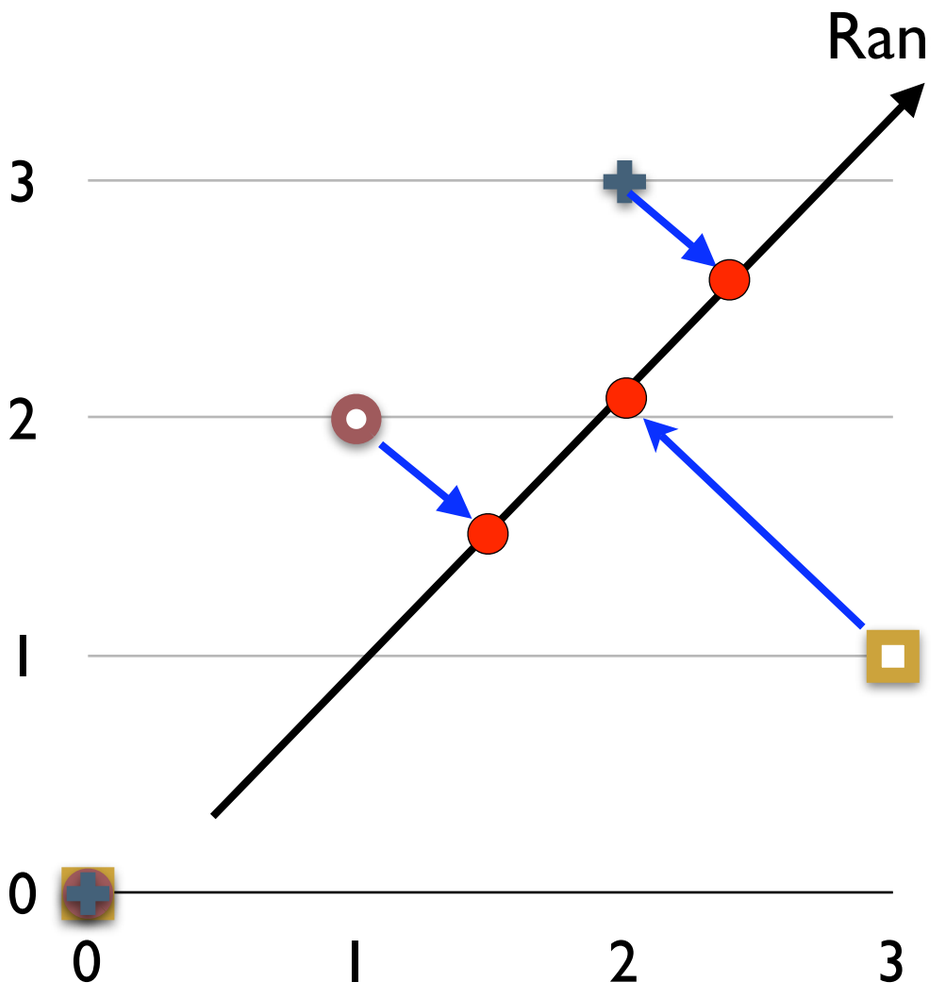
N x R = Random projected vector

0.47
0.25
0.25
0.84

Sorting ....

0.25	} 1st cluster
0.25	
0.47	} 2nd cluster
0.84	} 3rd cluster

Random projection and hashing



In fact random projection here works as a class of hashing function. Hashing is much faster than alternative methods because avoid the pair-wise comparisons required for partition and classification

If two points (p , q) are close, they will have a very small |p-q| (Euclidean metric) value; and they will hash to the same value with high probability; if they are distant, they should collide with small probability

Some results: Cluster example

Sig. dig. k	No clusters
4	6591
4	6507
4	5735
3	6481
3	6402
3	5360
2	2519
2	2576
2	2135
1	138
1	148
1	167

Results for the three different data sets, each consisting of 7500 chemicals, are shown in immediate succession. The number of significant decimal digits is 4 (more precise, and hence more different clusters found), 3, 2, and 1 (lowest precision in terms of significant digits).

Comparative evaluation

Sig. Dig.	No. Clusters	Largest cluster	No. discrep.	No. discrep. cl.
1	138	7037	3	3
1	148	7034	1	1
1	167	6923	9	7

Comparative evaluation: Results of k-means using as input the cluster centers provided by the 1 sig. dig. Baire approach relating to 7500 chemical structures, with 1052 descriptors.

Sig. dig. : number of significant digits used.
No. clusters: number of clusters in the data set of 7500 chemical structures, associated with the number of significant digits used in the Baire scheme.
Largest cluster : cardinality.
No. discrep. : number of discrepancies found in k-means clustering outcome.
No. discrep. cl. : number of clusters containing these discrepant assignments.

Current and future Work

– Textual information search in large backup or archived (document, email, etc.) repositories. Collaboration with ThinkingSAFE UK.

  – To support emergent compliance legislation

  – To be efficient and scalable to ~ 50 million documents

– Explore efficient hierarchy labeling for large repositories. e.g. BDB

Conclusion

- We find unusual symmetries in high dimensions (or low sample data) spaces
- Scalability is fundamental to handle requirements of massive data set analysis/processing
- Data coding is an essential part of data analysis

References

– F. Murtagh, G. Downs and P. Contreras, "Hierarchical Clustering of Massive, High Dimensional Data Sets by Exploiting Ultrametric Embedding". Submitted 2006. http://www.cs.rhul.ac.uk/~fionn/papers/ultrametrization-oct06.pdf

– F. Murtagh, "Hilbert Space Becomes Ultrametric in the High Dimensional Limit: Application to Very High Frequency Data Analysis", arXiv:physics/0702064v1, submitted, 2007.

– F. Murtagh, "On Ultrametricity, Data Coding, and Computation", Journal of Classification, 21, 167-184, 2004.

– F. Murtagh, "Thinking Ultrametrically", in D. Banks, L. House, F.R. McMorris, P. Arabie and W. Gaul, Eds., Classification, Clustering, and Data Mining Applications, Springer, 3-14, 2004.

– F. Murtagh, "Quantifying Ultrametricity", in J. Antoch, Ed., Compstat 2004: Proceedings in Computational Statistics, 1561-1568, Springer, 2004.

– F. Murtagh, "Identifying the Ultrametricity of Time Series", European Physical Journal B, 43, 573-579, 2005.