A mathematical approach for reference selection in high throughput proteomics

Brian L. LaMarche, Stephen J. Callister, Anuj R. Shah, Aaron T. Wright, Michael J. Wilkins, Mathew E. Monroe, Kevin Crowell, Gordon A. Anderson, and Richard D. Smith
Pacific Northwest National Laboratory, Richland, WA

Pacific Northwest NATIONAL LABORATORY

Overview

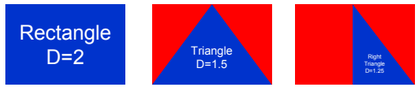
- Alignment of LC-MS features is used to correct for systematic variation; a critical step for comparative LC-MS data analysis.
- Alignment performed by selecting a single reference dataset; without *a priori* knowledge is non-trivial (may require a N-squared dataset comparison).
- Use of fractal geometry to assist selection of a reference dataset by computing a single global descriptor so N-x-N analysis is not required.
- Method can be extended to high throughput operations as it quantifies how a sample spans the mass and retention time dimension.
- This method may serve as a data quality metric.

Introduction

- The Minkowski-Bouligand (MB) dimension [1] is an approximation of fractal dimension. It can be used to describe the distribution of data in space.
- We calculate a MB dimension for a given dataset using the box-counting algorithm. This dimension is useful to alignment because MB can provide a single value describing how a dataset covers mass and LC retention time dimensions.
- To demonstrate this mathematical approach, LC-MS features were extracted from 566 LC-MS datasets of tryptically digested *Shewanella oneidensis* MR1 whole cell lysates. Each dataset is assigned a MB dimension.
- To show application to meta-proteomics analysis, 29 LC-MS datasets obtained from environmental biomass recovered from a U(VI) bioremediation project at the Rifle IFRC were aligned in an N-squared comparison to show the relationship between MB dimensions and cluster sizes.

Methods

- LC-MS features are extracted from raw mass spectra by first deisotoping using Decon2LS [2].
- These features are annotated with LC retention times and clustered using VIPER [3].

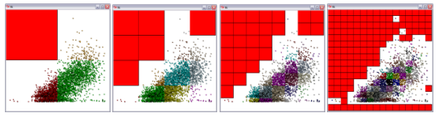


A fractal dimension is an abstract concept that an object's dimension is not a whole integer. Instead it has a real value dimension bounded between the two dimensions in a Euclidean space that it is modeled. For example, the rectangle shown covers the entire area so is said to have a dimension of D=2. The triangles do not quite span the entire space, but span more than the R=1 subspace; thus their dimension is real number bounded by R=[1,2].

- Box-Counting recursively divides LC-MS features into quadrants based on their position in the mass and normalized elution time (NET) plane.
- At each iteration the Minkowski-Bouligand Dimension (MB), *D*, is approximated as:

$$D = \frac{\log(N)}{\log(\frac{1}{r})}$$

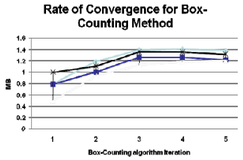
where *N* is the number of quadrants containing at least one LC-MS feature, and *r* is a scaling factor.



Iteration = 1 R = 2 Scale = ½ | Iteration = 2 R = 4 Scale = ¼ | Iteration = 3 R = 8 Scale = 1/8 | Iteration = 4 R = 16 Scale = 1/16

Iterations of the box-counting methods are shown. The red squares indicate boxes with no LC-MS features.

Since the Box-Counting algorithm is iterative, we show that the MB dimension can be computed in only a few iterations making it a tractable computation. It can also be extended to higher dimensions (e.g., R=3 using cubes).

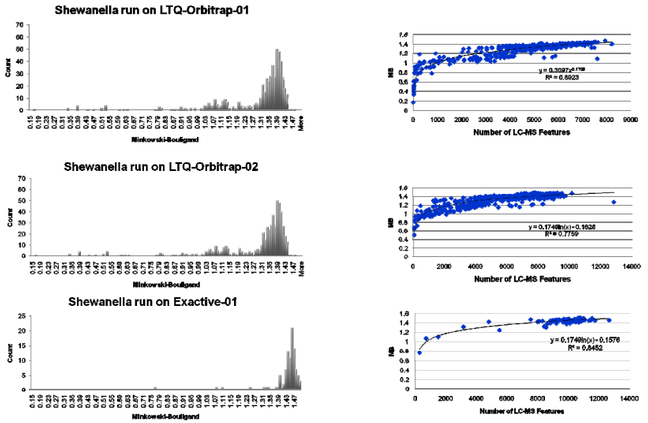


The Minkowski-Bouligand dimension was calculated for three *shewanella oneidensis* datasets of LC-MS features. The MB values are plotted with respect to each algorithm's iteration. The dimension converges within a few iterations.

Results

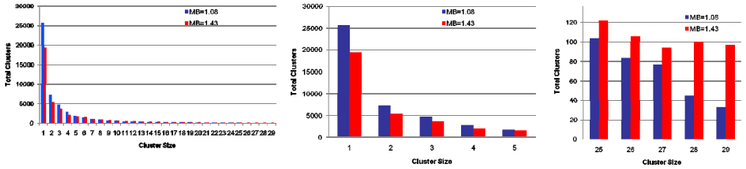
Minkowski-Bouligand as a global descriptor

- To show the MB dimension is not specific to a mass spectrometer or LC system, we analyzed feature maps from multiple instruments: two different LTQ-Orbitrap and one Thermo-Scientific Exactive mass spectrometers.
- LC was performed with uniform separation times using in-house custom LC platforms.
- Three histograms were constructed for each system showing similar MB distributions of the *Shewanella* datasets along with plots of MB values vs. the number of features identified.



Clustering

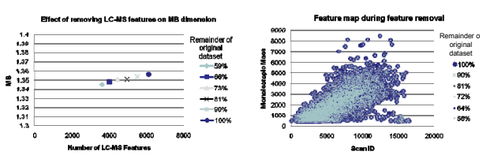
- Clustering of features is useful for identifying expression profiles across biological conditions. Profiles are constructed with two or more features, making smaller clusters less likely to provide meaninginful biological insight.
- Using the Rife IFRC datasets, a LC-MS feature map with low MB value produces more single member clusters than a higher MB value dataset when clustered using Single Linkage Clustering.



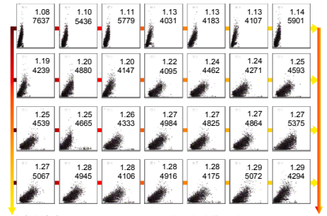
These histograms show the total number of clusters for each cluster size (member count). The far left histogram shows all cluster sizes. The middle histogram shows clusters with only a few members. The right histogram shows the larger clusters.

Space filling quantification

- To show the robustness of MB to the number of LC-MS features, we randomly removed LC-MS features from a *Shewanella* dataset up to 50%. Although 50% of the features are ultimately removed, the MB is only slightly affected.
- Far right we show how the shape of a LC-MS feature map changes with respect to the MB dimension. MB values increase starting upper left moving down and right.
- Note, in both plots MB is robust to the number of features present in the dataset.



% of original feature map	59%	66%	73%	81%	90%	100%
# points	3616	4018	4465	4962	5514	6127
MB dimension	1.345	1.348	1.350	1.350	1.353	1.356



LC-MS feature maps are visualized. MB values increase starting upper left heading right to bottom. This shows a correlation between shape quality and MB independent of feature count.

Peptide identification

- We matched two *Shewanella* datasets to a reference database containing 51,000 accurate mass and time tags filtered by a Peptide Prophet score of 0.99. The first data set had a MB dimension of 1.08 and 7637 features. The second data set had a MB dimension of 1.44 and 7640 LC-MS features.
- Matches were constrained by STAC (Statistical Tools for AMT tag Confidence*) scores to ensure confident matches to confident tags. Using a STAC cutoff score of .9 the dataset with a MB value of 1.44 had 1833 matches, while the dataset with a MB value of 1.08 matched to 1357 AMT tags.

*Manuscript in progress.

Conclusions

- The Minkowski-Bouligand metric is independent of instrumentation and thus useful as a global descriptor.
- The box-counting method converges after a few iterations; calculation of the Minkowski-Bouligand dimension scales linearly in time and space complexity.
- Since the box-counting method can be extended to higher dimensional data, it is useful for analyzing LC-MS measurements coupled with other separation types.
- Using a higher MB scored dataset as a reference for alignment improved clustering results for meta-proteomics samples.
- Our approach can assign a single value to a LC-MS feature set without requiring visual inspection of data, making it conducive for high-throughput operation.

Acknowledgements

The research described was conducted under the LDRD Program at the Pacific Northwest National Laboratory: a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RL01830.

Samples were analyzed using capabilities developed under the support of the NIH National Center for Research Resources (RR18522) and the U.S. Department of Energy Biological and Environmental Research (DOE/BER).

Significant portions of the work were performed in the Environmental Molecular Science Laboratory, a DOE/BER national scientific user facility at Pacific Northwest National Laboratory (PNNL) in Richland, Washington. PNNL is operated for the DOE by Battelle under contract DE-AC05-76RLO-1830.

References

1. Mandelbrot, B.B. (1982). *The Fractal Geometry of Nature*. W.H. Freeman and Company, 1982.
2. Jaitly N, Mayampurath A, Littlefield K, Adkins JN, Anderson GA, et al. Decon2LS: An open-source software package for automated processing and visualization of high resolution Mass Spectrometry Data. BMC Bioinformatics. 2009;10:87
3. Monroe ME, Tolić N, Jaitly N, Shaw JL, Adkins JN, Smith RD. VIPER: an advanced software package to support high-throughput LC-MS peptide identification. Bioinformatics. 2007 Aug 1;23(15):2021-3.

CONTACT: Brian LaMarche
Environmental Molecular Science Laboratory, K8-91
Pacific Northwest National Laboratory
P.O. Box 999, Richland, WA 99352
E-mail: brian.lamarche@pnl.gov