[111] R. Tibshirani, G. Walther and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[112] R. Agrawal, J. Gehrke, D. Gunopulos and P. Raghavan. *Automatic subspace clustering of high dimensional data for data mining applications*, volume 27. ACM, 1998.

[113] D. Barbará and P. Chen. Using the fractal dimension to cluster datasets. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 260–264. ACM, 2000.

[114] G. Sheikholeslami, S. Chatterjee and A. Zhang. Wavecluster: A multi-resolution clustering approach for very large spatial databases. In *VLDB*, volume 98, pages 428–439. 1998.

[115] R. Apweiler, A. Bairoch, C. H. Wu et al. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 32(suppl 1):D115–D119, 2004.

[116] P. E. Bourne, K. J. Addess, W. F. Bluhm et al. The distribution and query systems of the rcsb protein data bank. *Nucleic acids research*, 32(suppl 1):D223–D225, 2004.

[117] Y. Chen, K. D. Reilly, A. P. Sprague and Z. Guan. Seqoptics: a protein sequence clustering system. *BMC bioinformatics*, 7(4):1, 2006.

[118] M. Hauser, C. E. Mayer and J. Söding. kclust: fast and sensitive clustering of large protein sequence databases. *BMC bioinformatics*, 14(1):1, 2013.

[119] A. Krause, J. Stoye and M. Vingron. Large scale hierarchical clustering of protein sequences. *BMC bioinformatics*, 6(1):1, 2005.

[120] J. S. Bernardes, F. R. Vieira, L. M. Costa and G. Zaverucha. Evaluation and improvements of clustering algorithms for detecting remote homologous protein families. *BMC bioinformatics*, 16(1):1, 2015.