Robust Perron Cluster Analysis for Various Applications in Computational Life Science

Marcus Weber and Susanna Kube

Zuse Institute Berlin (ZIB), Germany

Abstract. In the present paper we explain the basic ideas of Robust Perron Cluster Analysis (PCCA+) and exemplify the different application areas of this new and powerful method. Recently, Deuflhard and Weber [5] proposed PCCA+ as a new cluster algorithm in conformation dynamics for computational drug design. This method was originally designed for the identification of almost invariant subsets of states in a Markov chain. As an advantage, PCCA+ provides an indicator for the number of clusters. It turned out that PCCA+ can also be applied to other problems in life science. We are going to show how it serves for the clustering of gene expression data stemming from breast cancer research [20]. We also demonstrate that PCCA+ can be used for the clustering of HIV protease inhibitors corresponding to their activity. In theoretical chemistry, PCCA+ is applied to the analysis of metastable ensembles in monomolecular kinetics, which is a tool for RNA folding [21].

1 Introduction

The application and improvement of cluster algorithms plays an important role in several areas of computational life science. Given a number of N objects $q \in \Omega$ with certain features, we are interested in identifying objects with similar behaviour in order to combine them into N_C clusters. For this purpose, we want to construct membership functions $y_i : \Omega \to [0,1], i = 1, \ldots, N_C$, $N_C \ll N$, which form a partition of unity. Then, each object in Ω can be assigned to the clusters with certain weights given by the values of the membership functions. A cluster can be considered as a vector that remains almost invariant under the action of a matrix T, i.e.

$$Ty_i \approx y_i. \tag{1}$$

In molecular dynamics, T is the discretised version of a spatial transition operator [14] and clusters are conformations for which the large scale geometric structure is conserved. In this case, the matrix T contains transition probabilities between different conformations. In general, T must be a row stochastic matrix. For example, it can result from the normalisation of a symmetric matrix whose entries represent some pairwise similarity measure, e.g. a covariance matrix.

Equation (1) is similar to an eigenvalue problem for an eigenvalue near $\lambda = 1$. A perturbation analysis shows that the space of eigenvectors of T corresponding to

eigenvalues near $\lambda = 1$ indicates a partition of Ω into the clusters we are looking for [4]. In Robust Perron Cluster Analysis, the space spanned by the membership functions y_i equals the space of the N_C first eigenvectors of T. In this case, the number N_C of clusters equals the number of discrete eigenvalues of T near $\lambda_1 = 1$. If each object is uniquely assigned to a cluster, then a rearranging of the rows and columns of T results in an almost block diagonal matrix. Therefore, the identification of clusters can also be seen as a detection of the almost block diagonal structure of T.
There are several other spectral methods which can be applied to reduce the dimensionality of given data. For example, Principle Component Analysis (PCA) and Independent Component Analysis (ICA) use the eigenvectors of a covariance matrix to compute a set of important directions within the data. However, they fail to separate non-overlapping data sets. An illustrative example can be found in [7]. PCCA+ was especially designed to identify spatially separated clusters and is close to Laplacian projection methods used in graph partitioning [20] [18], for example the relaxation of the *normalised cut* minimisation problem used by Shi and Malik [16] and the Multicut Algorithm by Meila and Shi [11]. The main differences between Robust Perron Cluster Analysis and these methods are:

- The results of Perron Cluster Analysis are given in terms of almost characteristic functions, i.e. fuzzy sets.
- These functions are a simple linear transformation of the eigenfunctions of the operator $\mathcal{T}$.
- There is a detailed perturbation analysis for the PCCA+ approach based on Markov chain theory, which provides robustness of this method.

2 Robust Perron Cluster Analysis Approach

The basis for Robust Perron Cluster Analysis is a stochastic matrix $T \in \mathbb{R}^{N \times N}$ with an eigenvalue cluster near 1. The clusters we are looking for are represented by vectors y_i, $i = 1, \ldots, N_C$, combined into a nonnegative matrix $Y \in \mathbb{R}^{N \times N_C}$. In order to meet the partition-of-unity constraint, Y has to be row stochastic, see also [3]. Since Y should fulfil

$$Ty_i \approx y_i,$$

the idea of PCCA+ is to construct Y as a linear transformation of the matrix $X \in \mathbb{R}^{N \times N_C}$, which contains the N_C first eigenvectors of T corresponding to eigenvalues near $\lambda_1 = 1$, see [5]. Therefore, the task for PCCA+ is to find a corresponding transformation matrix $\mathcal{A} \in \mathbb{R}^{N_C \times N_C}$, such that

$$Y = \mathcal{A}X$$

is a nonnegative, row stochastic matrix. Since there are many feasible solutions, one searches for a solution which maximises the functional

$$\sum_{i=1}^{n} \frac{\langle y_i, Ty_i \rangle_\pi}{\langle y_i, e \rangle_\pi} \to \max,$$

where $e = (1, \ldots, 1)$ is a constant vector, y_i is the ith column of Y and $\langle \cdot \rangle_\pi$ is a π-weighted inner product with the unique invariant row vector which meets $\pi = \pi T$. If the stochastic matrix T is the discretisation of a transition operator, then this is equivalent to the maximisation of metastability [5]. If the stochastic matrix is constructed based on a geometrical cluster problem (see below), then this optimisation problem minimises the overlap between different clusters. Instead of solving a constrained optimisation problem, another approach tries to find an optimal initial guess $\mathcal{A}$ wrt. the maximisation problem without regarding the non-negativity constraint for Y [20]. The smallest entry of Y, the so-called minChi-indicator, measures the feasibility of the initial guess as a solution of the clustering. This is also applied in order to determine N_C, i.e. the correct number of clusters. The minChi-indicator is used for the geometrical cluster problems shown in this paper.

For an application of Robust Perron Cluster Analysis in conformation dynamics see [5]. Now, we will give some other application examples for PCCA+.

3 Graph-based Spectral Clustering via PCCA+

Suppose we want to cluster $N_o \in \mathbb{N}$ objects, each of them described by $N_f \in \mathbb{N}$ features given by real numbers. That means we have to apply PCCA+ to an $N_o \times N_f$ real valued object-feature-matrix X. As input for PCCA+, we need an $N_o \times N_o$ diagonalisable stochastic matrix T which measures the similarity between objects in some sense. For this purpose, T is constructed out of a symmetric nonnegative matrix $W \in N_o \times N_o$ by scaling its rows to row sum 1, see [20]. The symmetric matrix W can be seen as weight matrix for an undirected graph where each object is represented by a vertex. The pairwise similarities between these vertices are expressed by weights of the corresponding edges. One example for computing this weight matrix can be taken from our analysis of gene expression data [20] in cooperation with the Max Planck Institute for Molecular Genetics. With some parameter $\beta > 0$, the weight $W(i, j)$ of the edge between object i and object j is defined as

$$W(i, j) = \exp(-\beta\, d^2(i, j)),$$

where $d(i, j)$ denotes the standard Euclidean distance between the ith and the jth row of X interpreted as vector in the N_f-dimensional space.

As an example, we examined the expression data of $N_f = 2000$ genes taken from $N_o = 50$ breast cancer patients. As preprocessing, we rescaled the features to zero mean and variance 1. After constructing T, we applied PCCA+ and got two clusters[1] y_1, y_2. Each patient $i \in \{1, \ldots, N_o\}$ was assigned to the cluster $k = 1, 2$, for which $y_k > 0.5$. In Figure 1 we compared the survival time of these two groups of patients and recognised a significant difference. The low p-value denotes the probability, that the difference of these two curves arises randomly. For a comparison with other clustering methods see [20].

[1] The minChi-indicator also allowed more than two clusters, $N_C = 2$ has been chosen in order to compare the results of PCCA+ with results from literature [20].

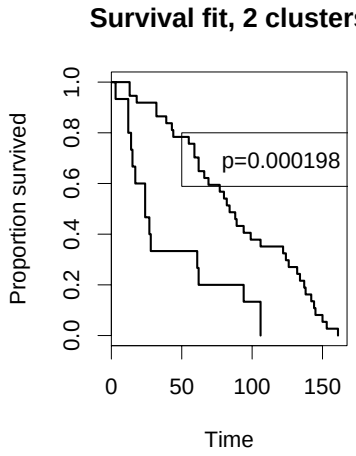

Fig. 1. Comparison of survival curves resulting from PCCA+ applied to gene expression data of breast cancer research.

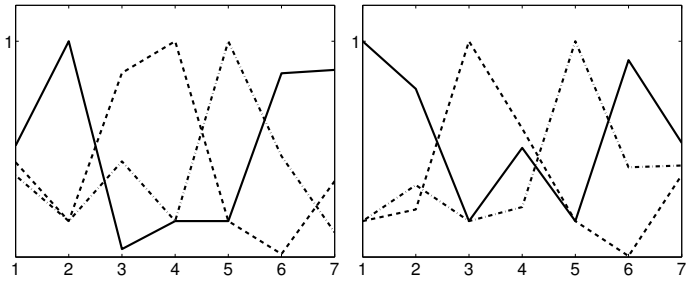

Fig. 2. Two clusterings of seven HIV protease inhibitors on the basis of 2311 HIV mutants. The computation of the activity coefficients differs between the two pictures. The three membership functions y_1, y_2 and y_3 are plotted as solid, dash and dash-dot line.

A second example for a graph based clustering turns up in the research of HIV protease inhibitors. We examined data kindly provided by Martin Däumer and Rolf Kaiser from the Institute of Virology, Cologne University, and Joachim Selbig from the Department of Biochemistry and Biology at the University of Potsdam [1]. The aim of this project is to find out if structural similarities between different inhibitors imply functional similarities. In a first step it was examined how good $N_o = 7$ different protease inhibitors bind to $N_f = 2311$ different mutants of HIV protease which are described by their genotype. This behaviour was measured by the activity coefficients. Our task was to identify those inhibitors with the same functional behaviour. For the computation of the 7×7-similarity matrix W, the pairwise correlation coefficients of the seven activity "vectors" have been shifted and normalised to the interval $[0, 1]$. Then the stochastic matrix T has been constructed and PCCA+ has been applied. The result was a 3-clustering of the protease inhibitors indicating different behaviour according to the HIV protease mutants. For a verification of the result, we used the fact that the activity coefficients can be computed in different ways. PCCA+ has been applied to these different activity coefficients and we always got similar results. In Figure 2 the results of two of the clusterings are shown. The x-axis shows the seven protease inhibitors. Their grades of membership, i.e. the curves for y_1, y_2 and y_3, are plotted in different line styles on the y-axis. Each HIV protease can be assigned to the cluster for which the corresponding grade of membership is maximal, i.e. for both experiments we get the result

$$\text{cluster}_1 = \{1, 2, 6, 7\}, \quad \text{cluster}_2 = \{3, 4\}, \quad \text{cluster}_3 = \{5\}.$$

Now it remains to examine the structural similarities between the different protease inhibitors which is still ongoing work. If it turned out that the structure of the inhibitors allows the same clustering, laboratory work could be done in a more tightly focused way.

4 Analysis of Metastable Ensembles in Monomolecular Kinetics

The understanding of transition pathways between different conformations of a molecule is an important issue in structural biology. Although the restriction of degrees of freedom to a few dihedral angles significantly reduces the complexity of the problem, this is still very difficult. Often, scientists are interested in single pathways, for example those over lowest energy barriers [2]. On the other hand, it is well known that molecular kinetics is not purely deterministic. All kinds of trajectories could appear, some with higher probability than others. Therefore, it seems natural to consider population probabilities. Starting with a given probability density in position space, we are interested in the evolution of the density to figure out intermediate states.
A description of molecular dynamics based on all conformations is unfeasible for large molecules. Therefore, we work with a set concept based on metastable

conformations as introduced in [14]. First, we reduce the position space to a number of N states represented by basis functions [19] or boxes [15]. Then, we cluster states into metastable conformations by applying PCCA+ to the transition rate matrix Q. The infinitesimal generator Q of T^τ provides important chemical information concerning transition pathways of single molecules. Given an initial weighting x_A of the states, one can compute the corresponding weights and the spatial configuration density at each time step $t \in [0, \infty)$ via

$$\dot{x} = Q^\top x, \qquad \text{with} \quad T^\tau = \exp(\tau Q). \tag{2}$$

This is the desired dynamic in configuration space, which is not based upon single molecules but upon ensembles.
It is easy to verify that the eigenvectors, which are essential for PCCA+, remain the same for the transition rate matrix Q. Assume, Q is diagonalisable by some nonsingular matrix X, i.e.

$$Q = X\Theta X^{-1} = X\mathrm{diag}(\theta_1, \ldots, \theta_p)X^{-1}.$$

Then

$$T^\tau = \exp(\tau Q) = X\exp(\tau\Theta)X^{-1} = X\mathrm{diag}(\exp(\tau\theta_1), \ldots, \exp(\tau\theta_p))X^{-1},$$

see [8]. Since $\exp(0) = 1$, an eigenvalue cluster of T^τ at 1 corresponds to an eigenvalue cluster of Q at 0. The number N_C of metastable sets is determined by this number of eigenvalues.
The entry $q_{ij}, i \neq j$, can be considered as the reaction rate of the monomolecular reaction

$$x_i \quad \rightharpoonup \quad x_j$$

where x_i stands representatively for the weight or "concentration" of state i. Equation (2) is not very interesting because the kinetics simply converges against the equilibrium distribution π. If one is interested in a simulation of a transition from metastable conformation A to a metastable conformation B and the corresponding transition behaviour, then (2) has to be solved as an initial value problem with initial distribution x_A and an absorbing end state given by the distribution x_B. Chemically, one would permanently eliminate conformation B out of the ensemble in order to push the reaction into the direction of this product. Mathematically this can be done by projection of x onto the orthogonal complement of the desired end point x_B before applying Q. Thus, the absorbing kinetics equation is:

$$\dot{x}(t) = Q^\top (x - \frac{\langle x, x_B\rangle}{\langle x_B, x_B\rangle} x_B), \quad x(0) = x_A. \tag{3}$$

The rate matrix Q can be obtained directly from the transition probability matrix T, but on the other hand, it offers a new approach to identify metastable conformations if the transition probability matrix is not available or difficult to compute. Furthermore, we are able to reduce our model not only to a set of basis functions whose number can be very large, but also to the few metastable sets which contain all important information about the system.

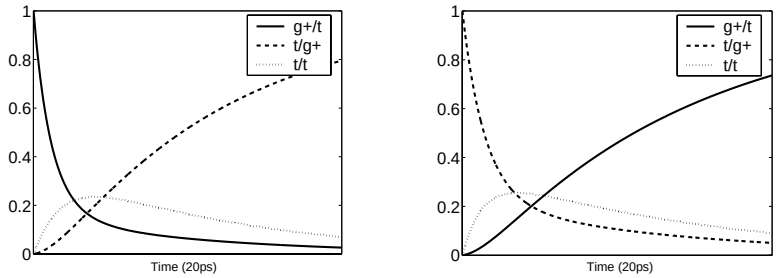

Fig. 3. Matlab [6] plot of a conformation kinetics simulation. *Left:* From g+/t conformation of pentane to the t/g+ conformation. *Right:* From t/g+ conformation of pentane to the g+/t conformation. Due to symmetry of pentane, both kinetics simulations should be equivalent. Differences result from unsymmetric approximations of transition probabilities.

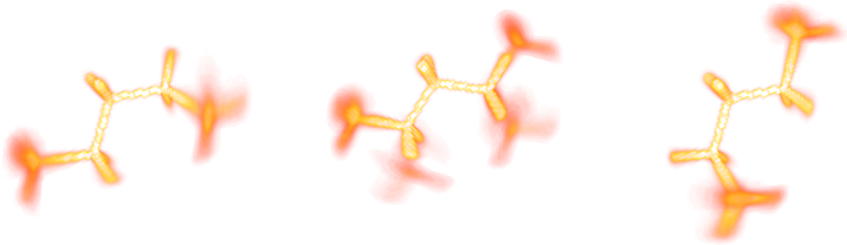

Fig. 4. Volume rendering of two conformations of pentane (left and right) and the corresponding transition macrostate (middle) in *amira/amiraMol* [17],[13].

Example: n-Pentane. We present the application to the n-pentane molecule $CH_3(CH_2)_3CH_3$ which was modelled with Merck Molecular Force Field [9][10] at a temperature of $300K$. The rate matrix Q was calculated directly from the transition probability matrix T. T itself resulted from a conformation dynamics simulation with ZIBgridfree, a program package based on meshfree methods which was developed at Zuse-Institute Berlin, see [19],[12].
We found 9 eigenvalues of Q close to 1,

$$\lambda = \{1.0000, 0.9988, 0.9985, 0.9978, 0.9976, 0.9967, 0.9947, 0.9601, 0.9589\},$$

followed by a gap to the 10th eigenvalue $\lambda_{10} = 0.8170$. This corresponds to 9 metastable conformations which can be distinguished according to the orientation of one of the two dihedral angles ($\pm$g and t denote the $\pm$ gauche and trans orientations):

$$\text{conformations} = \{-g/t, t/+g, -g/-g, t/t, t/-g, +g/t, +g/+g, -g/+g, +g/-g\}$$

The results for a $(g+/t) \rightharpoonup (t/g+)$ transition of pentane and the reverse experiment are shown in Figure 3. Only the concentrations of the conformations $(g+/t), (t/g+)$ and (t/t) are plotted. The corresponding Matlab algorithm needs less than 1 second CPU time for the computation of a 20ps reaction kinetics simulation with a 60×60-rate matrix Q, i.e. the numerical simulation of the "reduced model" is much faster than a full dynamics simulation of the same length. Figure 3 can be interpreted as follows. The conformational change from $(g+/t)$-pentane to $(t/g+)$ crosses the (t/t) conformation which can be seen as transition state. The transition from $(t/g+)$-pentane into $(g+/t)$-pentane is visualised in Figure 4. The left picture shows the start conformation $(t/g+)$, the right one the end conformation $(g+/t)$. At each step of the 20ps kinetics simulation, a similar density plot can be computed. The picture in the middle shows the transition state at 3.5ps simulation length. It can be considered as the mean conformation at this particular time.

Even though pentane is a very simple example, it illustrates very well the concept behind our method. From the chemical point of view, one could imagine that we start with a mixture of different molecules of the same chemical substance from which we know how the single molecules are distributed to the clusters. In this example, they all belong to the conformation $(g + /t)$. As time goes on, this distribution is driven towards equilibrium. Now, for example, suppose that molecules in a certain conformation are especially appropriate for a certain docking process, i.e. they do not contribute to the kinetics after this docking has taken place. This conformation is the target conformation of the reaction equation, here $(t/g+)$. The reaction kinetics calculation delivers information about the time scale of this process. Furthermore, it shows which other conformations are favoured in the meantime which can be of interest if several docking processes take place.

5 Conclusion

In the present paper, we have shown that Robust Perron Cluster Analysis (PCCA+) is a powerful tool for many cluster problems arising in computational life science. As input, PCCA+ expects a stochastic matrix T which can contain dynamics/kinetics information or similarity values from geometrical cluster problems. The aim of PCCA+ is to recover the almost block diagonal structure of T. The corresponding clustering is given in terms of a membership function for each of these "blocks". The number of almost-blocks in the matrix T need not to be known a priori. It is provided by the number of eigenvalues close to 1 or by the minChi-value. The property of the membership functions to be linear combinations of eigenfunctions allows their direct use in conformation kinetics. We prefer PCCA+ because it is easy to implement and has shown to be competitive with other clustering methods like Supervised Principal Component Analysis [20].

Acknowledgement. The authors especially want to thank Peter Deuflhard for various support of our work and for mathematical motivation. We also want to mention the cooperations with the group of M. Vingron at the Max Planck Institute for Molekular Genetics in Berlin, and the cooperation with J. Selbig at the University Potsdam and the Max Planck Institute for Molecular Plant Physiology. They provided us with application examples for PCCA+. Furthermore, we want to thank the group of P. Schuster from the Institute of Theoretical Chemistry at the University of Vienna for their support concerning conformation kinetics.

References

1. N. Beerenwinkel, B. Schmidt, H. Walter, R. Kaiser, T. Lengauer, D. Hoffmann, K. Korn, and J. Selbig. Diversity and complexity of HIV-1 drug resistance: A bioinformatics approach to predicting phenotype from genotype. *PNAS*, 99:8271–8276, 2002.
2. P. G. Bolhuis, C. Dellago, P. L.Geissler, and D. Chandler. Transition path sampling: throwing ropes over mountains in the dark. *Journal of Physics: Condensed Matter*, 12:A147–A152, 2000.
3. P. Deuflhard, W. Huisinga, A. Fischer, and Ch. Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Lin. Alg. Appl.*, 315:39–59, 2000.
4. P. Deuflhard and Ch. Schütte. Molecular conformation dynamics and computational drug design. In J.M. Hill and R. Moore, editors, *Applied Mathematics Entering the 21th Century*. ICIAM 2003, Sydney, Australia, 2004.
5. P. Deuflhard and M. Weber. Robust Perron Cluster Analysis in Conformation Dynamics. In M. Dellnitz, S. Kirkland, M. Neumann, and Ch. Schütte, editors, *Lin. Alg. App. – Special Issue on Matrices and Mathematical Biology*, volume 398C, pages 161–184. Elsevier Journals, 2005.
6. TheMathWorks Inc. Germany. Matlab(R) 6.5.0, 1994–2005.
7. D. Gleich and L. Zhukov. Soft clustering with projections: PCA, ICA, and Laplacian. Technical report, California Institute of Technology, Computer Graphics Research, 2004.

8. G.H. Golub and C.F. van Loan. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, 1996.
9. T.A. Halgren. *J. Am. Chem. Soc.*, 114:7827–7843, 1992.
10. T.A. Halgren. Merck molecular force field. *J. Comp. Chem.*, 17(I-V):490–641, 1996.
11. M. Meila and J. Shi. A random walks view of spectral segmentation. *AI and Statistics (AISTATS)*, 2001.
12. H. Meyer. Die Implementierung und Analyse von HuMfree–einer gitterfreien Methode zur Konformationsanalyse von Wirkstoffmolekülen. Master's thesis, Free University Berlin, February 2005.
13. J. Schmidt-Ehrenberg, D. Baum, and H.-C. Hege. Visualizing dynamic molecular conformations. In *IEEE Visualization 2002*, pages 235–242. IEEE Computer Society Press, 2002.
14. Ch. Schütte. *Conformational Dynamics: Modelling, Theory, Algorithm, and Application to Biomolecules*. Habilitation Thesis, Fachbereich Mathematik und Informatik, Freie Universität Berlin, 1999.
15. Ch. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys., Special Issue on Computational Biophysics*, 151:146–168, 1999.
16. J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
17. D. Stalling, M. Westerhoff, and H.-C. Hege. Amira - a highly interactive system for visual data analysis. In Christopher R. Johnson and Charles D. Hansen, editors, *Visualization Handbook*. Academic Press, November 2004.
18. D. Verma and M. Meila. A Comparison of Spectral Clustering Algorithms. Technical Report 03-05-01, University of Washington, 2003.
19. M. Weber. *Meshless Methods in Conformation Dynamics*. PhD thesis, Free University Berlin, 2005. In preparation.
20. M. Weber, W. Rungsarityotin, and A. Schliep. Perron cluster analysis and its connection to graph partitioning for noisy data. Technical Report ZR-04-39, Zuse Institute Berlin, 2004.
21. M. T. Wolfinger, W. A. Svrcek-Seiler, Ch. Flamm, I. L. Hofacker, and P. F. Stadler. Efficient computation of RNA folding dynamics. *J. Phys. A: Math. Gen.*, 37:4731–4741, 2004.