Robust Perron Cluster Analysis (PCCA+) in Conformation Dynamics

Peter Deuflhard, Marcus Weber*

Zuse Institute Berlin

Free University Berlin

Berlin Center for
Genom-based Bioinformatics

DFG Research Center
"Matheon"

ZIB Scientific Computing:

Dept. Numerical Analysis and Modelling

Peter Deuflhard, Frank Cordes, Marcus Weber,
Ulrich Nowak, Alexander Steidinger, Andreas May

Dept. Scientific Visualization

Hans-Christian Hege, Daniel Baum, Johannes Schmidt-Ehrenberg,
Timm Baumeister

Cooperation:

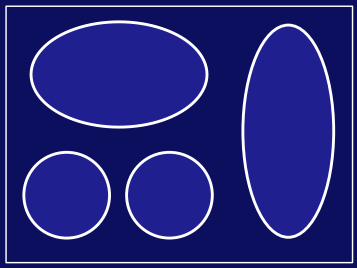
FU Biocomputing:

Christof Schütte, Wilhelm Huisinga, Alexander Fischer,
Illja Horenko, Carsten Hartmann, Phillip Metzner, Eike Meerbach

Berlin Center for Genome-based Bioinformatics (BCB)
DFG Research Center "Matheon"

Successive PCCA of dihedrals

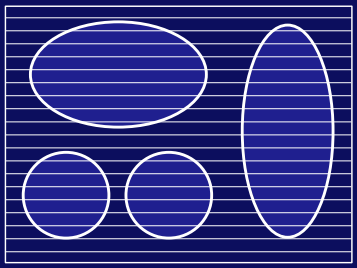
configuration space spanned by two dihedrals



Cordes, Weber, Schmidt-Ehrenberg, 2002

Successive PCCA of dihedrals

PCCA of first dihedral



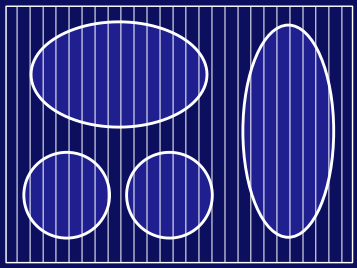
spectrum:

1.00 0.51 0.45 ...

Cordes, Weber, Schmidt-Ehrenberg, 2002

Successive PCCA of dihedrals

PCCA of second dihedral

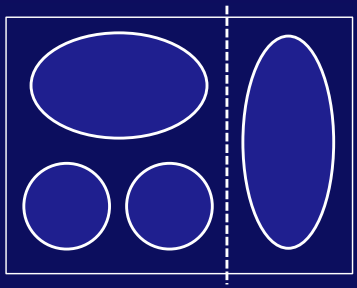


spectrum:

1.00 0.99 0.72 ...

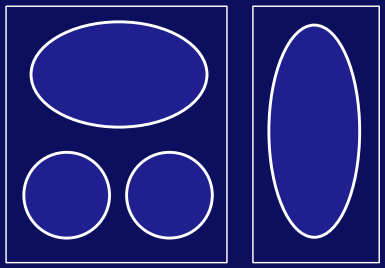
Cordes, Weber, Schmidt-Ehrenberg, 2002

Successive PCCA of dihedrals



Cordes, Weber, Schmidt-Ehrenberg, 2002

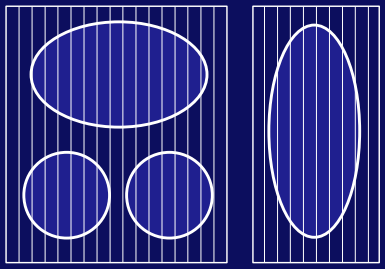
Successive PCCA of dihedrals



Cordes, Weber, Schmidt-Ehrenberg, 2002

Successive PCCA of dihedrals

PCCA of first dihedral



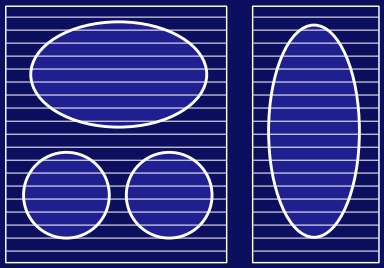
spectra:

1.00	0.71	0.63	...
1.00	0.76	0.61	...

Cordes, Weber, Schmidt-Ehrenberg, 2002

Successive PCCA of dihedrals

PCCA of second dihedral

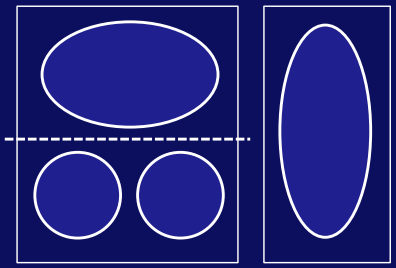


spectra:

1.00	0.99	0.73	...
1.00	0.75	0.64	...

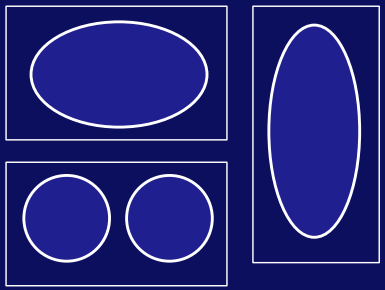
Cordes, Weber, Schmidt-Ehrenberg, 2002

Successive PCCA of dihedrals



Cordes, Weber, Schmidt-Ehrenberg, 2002

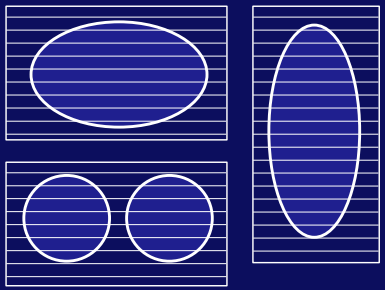
Successive PCCA of dihedrals



Cordes, Weber, Schmidt-Ehrenberg, 2002

Successive PCCA of dihedrals

PCCA of first dihedral



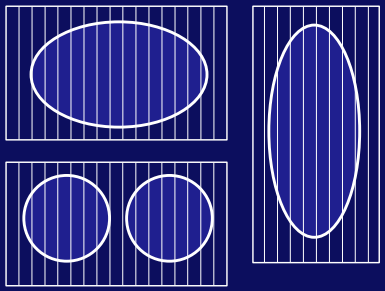
spectra:

1.00	0.71	0.68	...
1.00	0.75	0.64	...
1.00	0.69	0.61	...

Cordes, Weber, Schmidt-Ehrenberg, 2002

Successive PCCA of dihedrals

PCCA of second dihedral

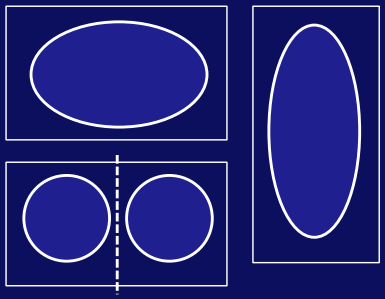


spectra:

1.00	0.71	0.68	...
1.00	0.76	0.61	...
1.00	0.99	0.90	...

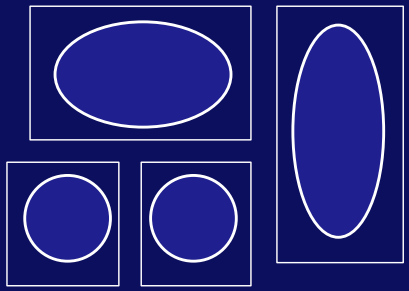
Cordes, Weber, Schmidt-Ehrenberg, 2002

Successive PCCA of dihedrals



Cordes, Weber, Schmidt-Ehrenberg, 2002

Successive PCCA of dihedrals



Cordes, Weber, Schmidt-Ehrenberg, 2002

Application Areas of PCCA+

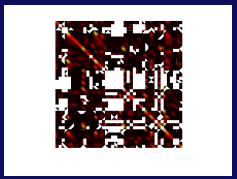
- Identification of conformations in drug design
- Identification of "connected conformations"
- Clustering of gene expression data

  (Weber, Rungsarityotin, Schliep, 2004)

What does PCCA+ do?

Input $\rightarrow$ PCCA+ $\rightarrow$ Output

(N, N)-transition matrix $\rightarrow$ classification of k almost invariant substructures



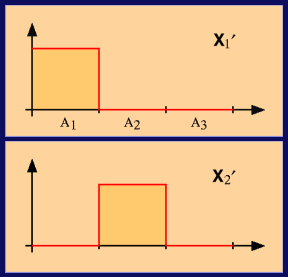
$\rightarrow$

Completely uncoupled Markov chains

$$T = \begin{pmatrix} T_1 & 0 & 0 \\ 0 & T_2 & 0 \\ 0 & 0 & T_3 \end{pmatrix}$$

$\lambda_1(T_{1,2,3}) = 1$

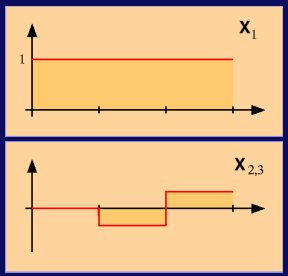
$X_i' = \chi_{A_i}$



$\lambda_{1,2,3}(T) = 1$

$X_1(T) = e = (1, \ldots, 1)$

$\chi = X\mathcal{A}$ linear combination



Linear Mapping

T: $(6, 6)$-transition matrix with 3 uncoupled blocks

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

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -2.02 & -0.55 \\ 1 & -2.02 & -0.55 \\ 1 & 0.48 & -0.91 \\ 1 & 0.48 & -0.91 \\ 1 & 0.50 & 1.24 \\ 1 & 0.50 & 1.24 \end{pmatrix} \cdot \begin{pmatrix} 0.20 & 0.41 & 0.39 \\ -0.40 & 0.33 & 0.07 \\ 0.00 & -0.47 & 0.47 \end{pmatrix}$$

“Nearly uncoupled” Markov chains

$$\widetilde{T} = \begin{pmatrix} \tilde{T}_1 & E_{12} & E_{13} \\ E_{32} & \tilde{T}_2 & E_{23} \\ E_{31} & E_{32} & \tilde{T}_3 \end{pmatrix}$$

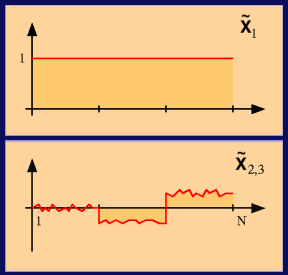
Deuflhard, Huisinga, Fischer, Schütte, 2000

“almost invariant” sets

$$\widetilde{\lambda}_1(T) = 1, \quad \widetilde{\lambda}_{2,3} = 1 - O(\epsilon)$$

$$\widetilde{X}_1(T) = e = (1, \ldots, 1)$$

PCCA: $\|\chi - \widetilde{X}\widetilde{\mathcal{A}}\|_\pi = \min$



Perturbation analysis: PCCA

Perturbation analysis: $\widetilde{T}(\epsilon) = T + \epsilon\, T^{(1)} + O(\epsilon^2)$

$\widetilde{X}(\epsilon) = X + \epsilon X^{(1)} + O(\epsilon^2)$

$\epsilon = 1 - \widetilde{\lambda}_2$

Lemma: $X^{(1)} = \chi B$

only level shifts

PCCA:

- k sign structures ($\epsilon = 0$) out of 2^{k-1} ones
- $k > 2$: "dirty zeroes" generic $O(\epsilon)$ effect!

Deuflhard, Weber 2003

Linear Mapping

T: $(6, 6)$-transition matrix with 3 almost uncoupled blocks

$$\widetilde{\chi} = \widetilde{X}\widetilde{\mathcal{A}}$$

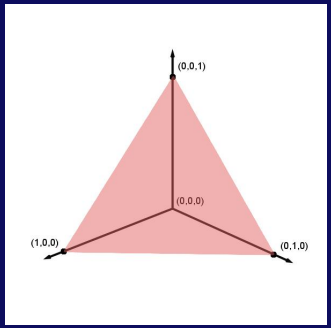
$$\begin{pmatrix} 1 & 0 & 0 \\ 0.8 & 0.2 & 0 \\ 0.1 & 0.8 & 0.1 \\ 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -2.02 & -0.55 \\ 1 & -1.52 & -0.62 \\ 1 & 0.23 & -0.66 \\ 1 & 0.48 & -0.91 \\ 1 & 0.50 & 1.03 \\ 1 & 0.50 & 1.24 \end{pmatrix} \cdot \begin{pmatrix} 0.20 & 0.41 & 0.39 \\ -0.40 & 0.33 & 0.07 \\ 0.00 & -0.47 & 0.47 \end{pmatrix}$$

Almost characteristic functions: PCCA+

$$\widetilde{\chi}(\epsilon) = \widetilde{X}(\epsilon)\widetilde{\mathcal{A}}(\epsilon), \qquad \widetilde{\mathcal{A}} = (\alpha_{ij})$$

Positivity: $\widetilde{\chi}_i(\epsilon) \geq 0$

Partition of unity: $\sum_{i=1}^{k} \widetilde{\chi}_i(\epsilon) = e$



$$(\widetilde{X}_2(l), \ldots, \widetilde{X}_k(l)) \in \widetilde{\sigma}_{k-1}$$
$$l = 1, \ldots, N$$

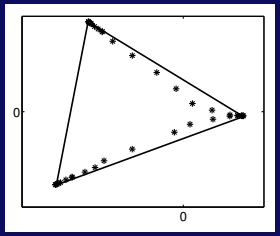
$$(\widetilde{\chi}_1(l), \ldots, \widetilde{\chi}_k(l)) \in \sigma_{k-1}$$
$$l = 1, \ldots, N$$

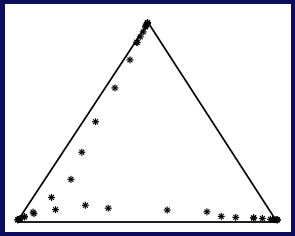
Deuflhard, Weber, 2003

Example n-butane: $\widetilde{X}$ versus $\widetilde{\chi}$

$k = 3,\ N = 42$

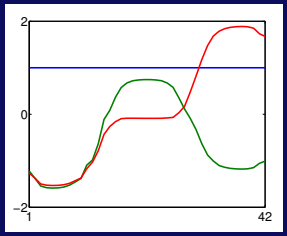
$\widetilde{\sigma}_2$

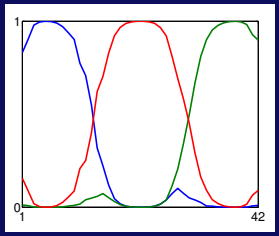




σ_2

$\widetilde{X}$





$\widetilde{\chi}$

Uniqueness of clustering

Theorem

Let

i) $\sum_{i=1}^{k} \widetilde{\chi}_i = e$,

ii) for all $i = 1, \ldots, k$ and $l = 1, \ldots, N$: $\widetilde{\chi}_i(l) \geq 0$,

iii) $\widetilde{\chi} = \widetilde{X}\widetilde{\mathcal{A}}$ with $\widetilde{\mathcal{A}}$ regular,

iv) for all $i = 1, \ldots, k$ there exists $l \in \{1, \ldots, N\}$ with $\widetilde{\chi}_i(l) = 1$.

Then

- 3 out of 4: easy to assure
- all 4: unique solution for almost characteristic functions

Constrained optimization problems

Scaling: $I_1[\alpha] = \sum_{i=1}^{k} \max_{l=1,\ldots,N} \widetilde{\chi}_i(l) \qquad \leq k$

Metastability: $I_2[\alpha] = \sum_{i=1}^{k} \frac{\langle \widetilde{\chi}_i, T\widetilde{\chi}_i \rangle_\pi}{\langle \widetilde{\chi}_i, e \rangle_\pi} \qquad < \sum_{i=1}^{k} \widetilde{\lambda}_i$

$I_{1,2}[\alpha] = \max$

subject to $\widetilde{\chi}(l) \in \sigma_{k-1}\,, \quad l = 1, \ldots, N$ and

$\widetilde{\chi} = \widetilde{X}\widetilde{\mathcal{A}}$

Application Areas of PCCA+

- Identification of conformations in drug design
- Identification of "connected conformations"
- Clustering of gene expression data

  (Weber, Rungsarityotin, Schliep, 2004)

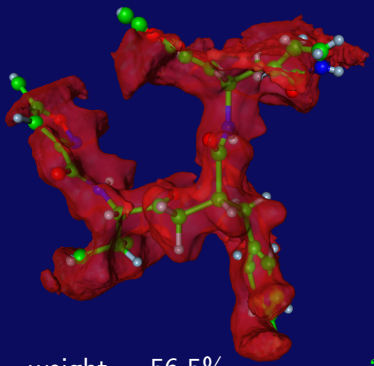
Example: Epigallocatechine

‘time interval’ = 5000 fs

‘time interval’ = 50 fs

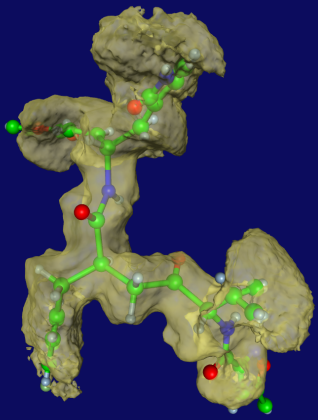
SARS protease inhibitor: conformations

Frank Cordes, Alexander Fischer, 2003



weight = 56.5%

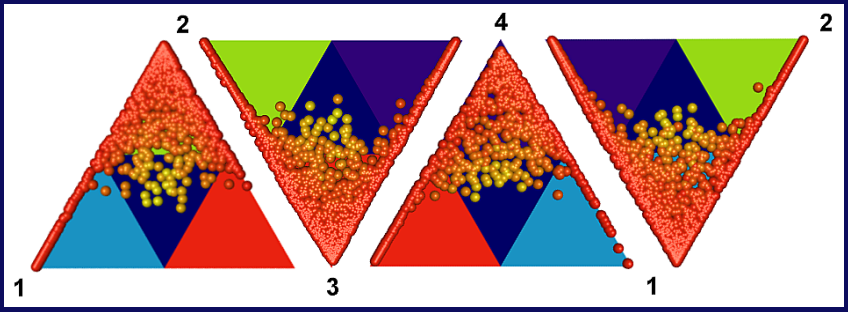
300K



weight = 35.6%

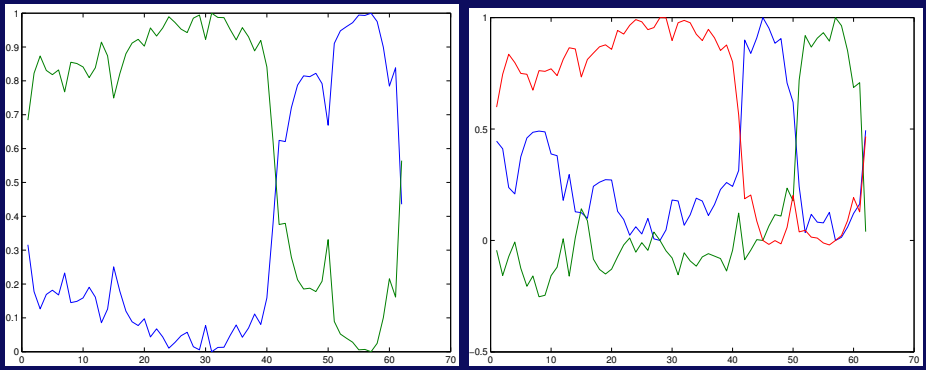
Example: Epigallocatechine

Johannes Schmidt-Ehrenberg, 2003



Gene Expression Data

Marcus Weber, Wasinee Rugsarityotin, Alexander Schliep, 2004



Cooperation with MPI for Molecular Genetics

Conclusions

- Robust cluster analysis via almost characteristic functions:

$$\chi - \widetilde{\chi} = O(\epsilon^2).$$

- Providing important informations: Identification of metastable sets, statistical weights, characterization of transition states...
- Geometrical clustering with PCCA+ is also possible.
- Visit our homepage: http://www.zib.de/MDGroup

Thank you for your attention!!!