Clustering of Metastable Sets given by Transition Rates

Marcus Weber, Susanna Kube

Zuse Institute Berlin

Free University Berlin

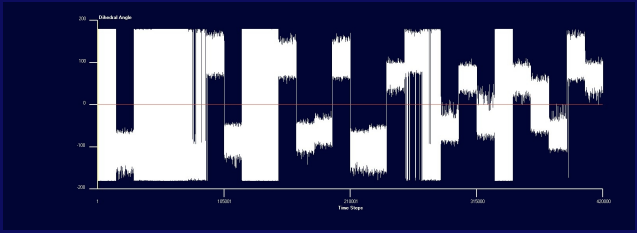
Berlin Center for
Genom-based Bioinformatics

DFG Research Center
"Matheon"

First Paper

Deuflhard, Huisinga, Fischer, Schütte, 2000:

“Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains”



Outline

- From deterministic to stochastic models
- Perron Cluster Analysis

Concept change

molecular dynamics

point concept:

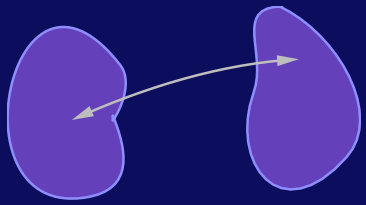
trajectory simulation

conformation dynamics

set concept:

metastable conformations

deterministic
mathematical model



stochastic
mathematical model

Molecular Dynamics

Modelling of molecules in classical MD:

$$H(q,p) = \frac{1}{2} p^\top M^{-1} p + V(q)$$

q: positions of the atoms, p: momenta

Corresponding canonical equations of motion: $\dot{q} = M^{-1}p, \qquad \dot{p} = -\nabla V$

Formal solution: $(q(t+\tau), p(t+\tau)) = \Phi^\tau(q(t), p(t), \tau)$

Stationary canonical density:

$$f_0(x) = \underbrace{\frac{1}{Z_p} \exp\left(-\frac{\beta}{2} p^\top M^{-1} p\right)}_{=\mathcal{P}(p)} \underbrace{\frac{1}{Z_q} \exp(-\beta V(q))}_{=\mathcal{Q}(q)}, \qquad \beta = 1/k_B\mathcal{T}$$

Conformations as Almost Invariant Sets

$S \subset \Gamma$ is called invariant under the flow Φ^τ iff

$$\Phi^\tau(S) = S \quad \forall\, \tau > 0$$

Conformations: almost invariant subsets of the position space Ω

Probability to stay within $S \subset \Omega$: $\delta(S,\tau) = w(S,S,\tau) \approx 1$ with

$$w(B,C,\tau) = \frac{1}{\int_B \mathcal{Q}(q)dq} \int_B \left\{ \int_{\mathbb{R}^d} \chi_C(\xi_1 \Phi^\tau(q,p)) \mathcal{P}(p)dp \right\} \mathcal{Q}(q)dq$$

Spatial Transition Operator:

$$T^\tau u(q) = \int u(\xi_1 \Phi^\tau(q,p)) \mathcal{P}(p)dp$$

$$w(S,S,\tau) \approx 1 \quad \leftrightarrow \quad T^\tau \chi_S \approx \chi_S$$

Spatial Discretization

Eigenvalue problem:

$$Tu = \lambda u, \qquad \lambda \approx 1$$

Galerkin approach:

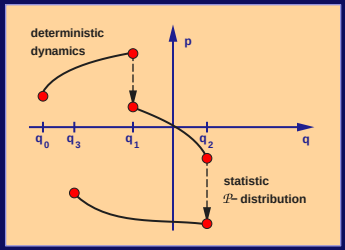
$$u(q) = \sum_{i=1}^{n} \alpha_i \varphi_i(q)$$

$$\sum_{i=1}^{n} \alpha_i \langle T\varphi_i, \varphi_j \rangle_{\mathcal{Q}} = \lambda \sum_{i=1}^{n} \alpha_i \langle \varphi_i, \varphi_j \rangle_{\mathcal{Q}}, \quad \forall\, j$$

$$\boxed{P\alpha = S\alpha\lambda}$$

T selfadjoint in $L^2_{\mathcal{Q}}$ $\quad\rightarrow\quad$ P symmetric (as well as S)

Approximation of the Transition Operator

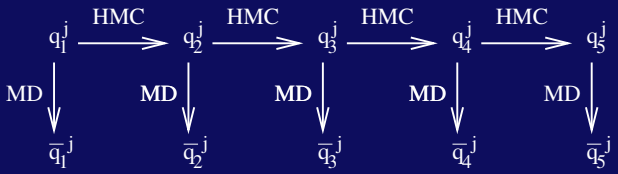


T: Markov operator in $L^1_{\mathcal{Q}}(\Omega)$ $\quad\rightarrow\quad$ generates a Markov chain $\{q_k\}_{k=1,2,\ldots}$ via the discrete stochastic dynamical system

$$q_{k+1} = \xi_1 \Phi^\tau(q_k, p_k)$$

p_k: randomly chosen from the momentum distribution $\mathcal{P}$

Approximation of the Transition Operator



Approximation of matrices by Monte Carlo importance sampling:

$$S(i,k) = \langle \varphi_i, \varphi_k \rangle_{\mathcal{Q}} \approx \frac{1}{N} \sum_{l=1}^{N} \varphi_i(q_l)\varphi_k(q_l),$$

$$P(i,k) = \langle \varphi_i, T\varphi_k \rangle_{\mathcal{Q}} \approx \frac{1}{N} \sum_{l=1}^{N} \varphi_i(q_l)\varphi_k(\overline{q}_l)$$

Transition Probabilities versus Transition Rates

$$T^\tau = \exp(\tau Q)$$

Assume $Q = X^{-1}\theta X$.

$$\exp(\tau Q) = X^{-1}\exp(\tau\theta)X$$

$$P = X^{-1}\Theta X$$

with $\lambda_i = \exp(\tau\theta_i)$.

Eigenvalue cluster of P at 1 corresponds to an eigenvalue cluster of Q around 0.

Stochastic Matrices

At this stage:

- Metastable in terms of an eigenvalue problem.
- Numerical routine for the computation of P and S.
- Possibly reformulation in terms of transition rates.

For Perron Cluster Analysis we need transition rate matrices or stochastic matrices.
Matrix T as discretization of T^τ:

$$T = D^{-1}P$$

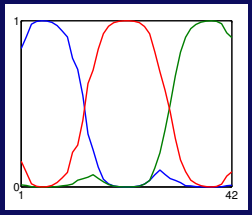
D is diagonal matrix, where d_{ii} is the ith-row sum of P. If all φ_i are characteristic, then $D = S$.

What does our algorithm do?

Input $\longrightarrow$ mapping $\longrightarrow$ Output

(N, N)-stochastic matrix $\longrightarrow$ (fuzzy) cluster analysis with k membership functions

$\longrightarrow$

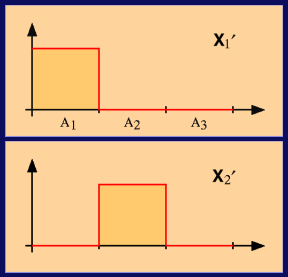


Completely uncoupled Markov chains

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

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

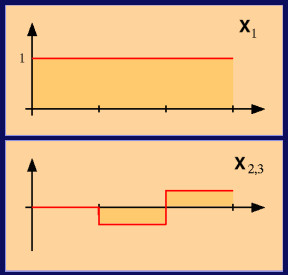
$X_i' = \chi_{A_i}$



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

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

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



Linear Mapping

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

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

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

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

linear combination: $\chi - \widetilde{\chi} = O(\epsilon^2)$

sign structure:

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

Deuflhard, Weber, 2003 Meila, Shi, 2001

“Nearly uncoupled” Markov chains

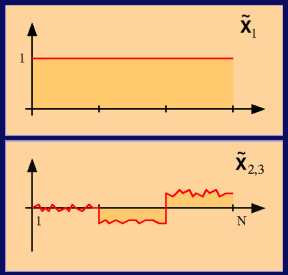
$$\widetilde{T} = \begin{pmatrix} \tilde{T}_1 & E_{12} & E_{13} \\ E_{32} & \tilde{T}_2 & E_{23} \\ E_{31} & E_{32} & \tilde{T}_3 \end{pmatrix}$$

Deuflhard, Huisinga, Fischer, Schütte, 2000

“almost invariant” sets

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

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



Linear Mapping

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

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

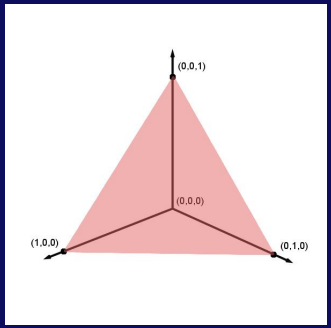
$$\begin{pmatrix} 1 & 0 & 0 \\ 0.8 & 0.2 & 0 \\ 0.1 & 0.8 & 0.1 \\ 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -2.02 & -0.55 \\ 1 & -1.52 & -0.62 \\ 1 & 0.23 & -0.66 \\ 1 & 0.48 & -0.91 \\ 1 & 0.50 & 1.03 \\ 1 & 0.50 & 1.24 \end{pmatrix} \cdot \begin{pmatrix} 0.20 & 0.41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

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



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

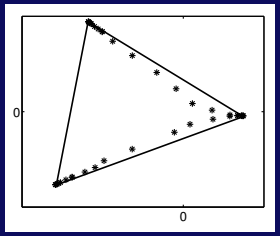
$$(\widetilde{\chi}_1(l), \ldots, \widetilde{\chi}_k(l)) \in \sigma_{k-1} \qquad l = 1, \ldots, N$$

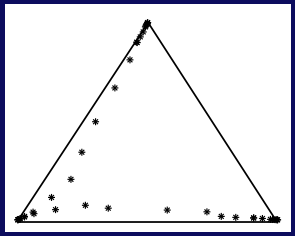
Deuflhard, Weber, 2003

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

$k = 3,\ N = 42$

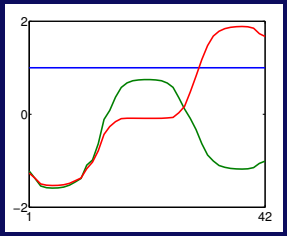
$\widetilde{\sigma}_2$

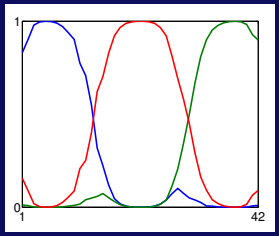




σ_2

$\widetilde{X}$





$\widetilde{\chi}$

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
- Discretization (via "tensor product")

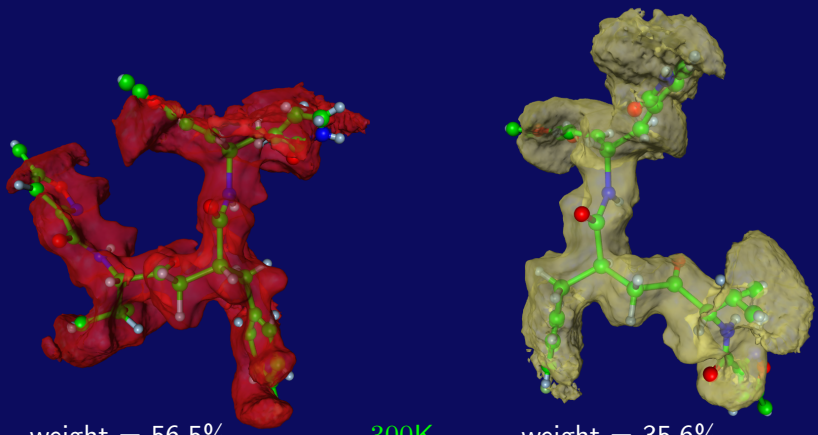
Example: Epigallocatechine

‘time interval’ = 5000 fs

‘time interval’ = 50 fs

SARS protease inhibitor: conformations

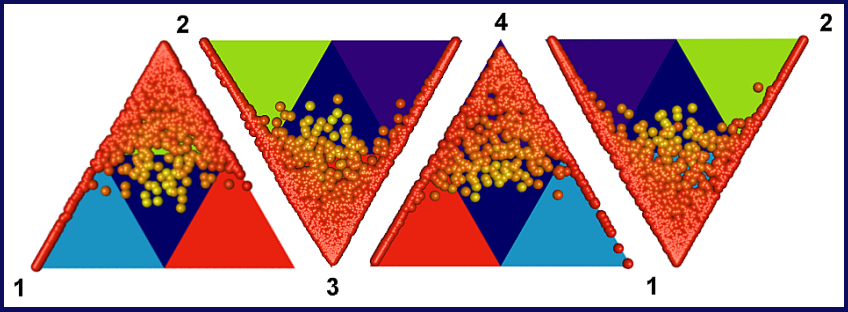
Frank Cordes, Alexander Fischer, 2003



weight = 56.5% 300K weight = 35.6%

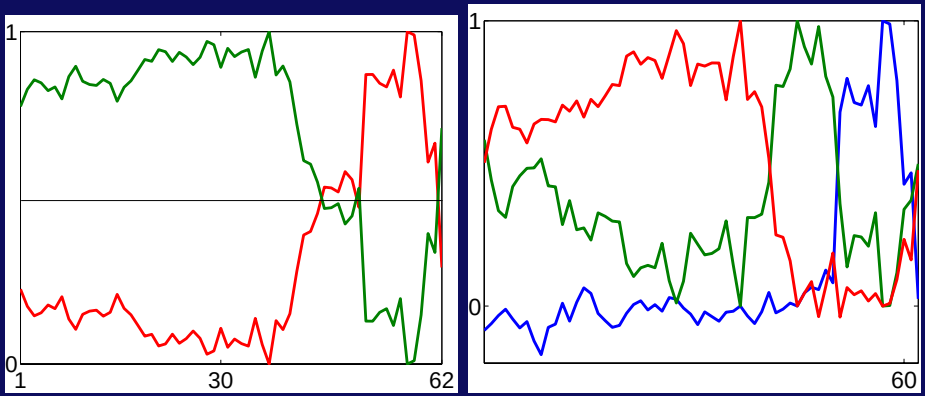
Example: Epigallocatechine

Johannes Schmidt-Ehrenberg, 2003



Gene Expression Data

Marcus Weber, Wasinee Rugsarityotin, Alexander Schliep, 2004



Cooperation with MPI for Molecular Genetics

Information for drug design

Perron index k

3D-Visualization:

metastable conformations: $i = 1, \ldots, k$

probabilities to be within i: $\widetilde{\pi}_i$

significant dihedrals

(k,k)-coupling matrix $\widetilde{W}$:

$$\begin{pmatrix} w_{11} & \cdots & w_{1k} \\ \vdots & \ddots & \vdots \\ w_{k1} & \cdots & w_{kk} \end{pmatrix}$$

w_{ij} : probability to move from i to j

$w_{ii} \approx 1$: probability to stay within i

$w_{ii} > 0.5$

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!!!