Modelling Principles in Computational Drug Design

Susanna Kube

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, Susanna Kube, Alexander Steidinger, Lionel Walter

Dept. Scientific Visualization

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

Cooperation:

FU Biocomputing:

Christof Schütte, Wilhelm Huisinga,Illja Horenko, Eike Meerbach

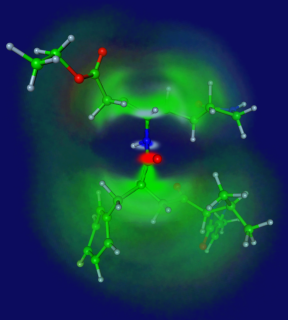
Berlin Center for Genome-based Bioinformatics (BCB)
DFG Research Center "Matheon"

Virtual lab

numerical ODEs:
molecular dynamics

nonlinear dynamics:
almost invariant sets

biochemistry:
RNA molecules,
prions,
viruses



statistics:
nearly uncoupled
Markov chains

computer science:
visualization

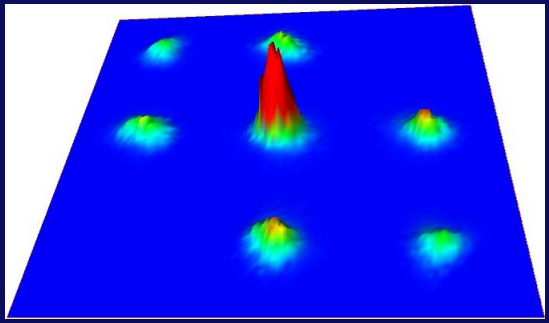
numerical linear algebra,
nonlinear optimization:
Perron cluster analysis

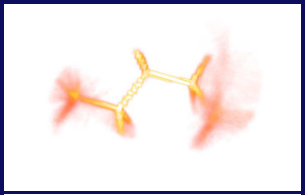
Outline

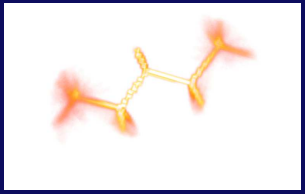
- Motivation
- Molecular Dynamics
- Transfer Operator Approach
- Molecular Kinetics
- Perron Cluster Analysis
- Eigenvector Computation
- Applications

Pentane

potential energy surface of pentane in the space spanned by the two essential dihedral angles at 300K







Outline

- Motivation
- Molecular Dynamics
- Transfer Operator Approach
- Molecular Kinetics
- Perron Cluster Analysis
- Eigenvector Computation
- Applications

Molecular Dynamics

- consider molecules in the (n, v, T)-ensemble (canonical ensemble)
- state=(momentum p, position q), $q \in \mathbb{R}^{3N},\ p \in \mathbb{R}^{3N}$
- total energy: $H(p,q) = K(p) + V(q) = \frac{1}{2} p^\top M^{-1} p + V(q)$
- Newton's equation of motion: $\dot{q} = M^{-1} p, \qquad \dot{p} = -\nabla V$
- formal solution: $(q(t+\tau), p(t+\tau)) = \Phi^\tau(q(t), p(t))$
- suitable integrator: e.g. Velocity Verlet
  - reversibel
  - symplectic (especially energy conserving)
  - momenta conserving

Molecular Dynamics

- in equilibrium: time-independent distribution of molecular states according to the Boltzmann-distribution

$$\begin{aligned}\pi(x) = \pi(p,q) &= \frac{1}{\mathcal{Z}} \exp(-\beta H(p,q)) \\ &= \underbrace{\frac{1}{\mathcal{Z}_p} \exp(-\beta K(p))}_{=\mathcal{P}(p)} \underbrace{\frac{1}{\mathcal{Z}_q} \exp(-\beta V(q))}_{=\mathcal{Q}(q)} \\ &= \pi_p(p)\, \pi_q(q)\end{aligned}$$

$$\mathcal{Z} = \int_{\mathbb{R}^{6N}} \exp(-\beta H(p,q)) dp\, dq, \quad \beta = \frac{1}{kT}$$

Concept change

molecular dynamics

point concept:

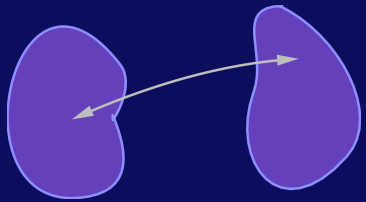
trajectory simulation

conformation dynamics

set concept:

metastable conformations

deterministic
mathematical model



stochastic
mathematical model

Outline

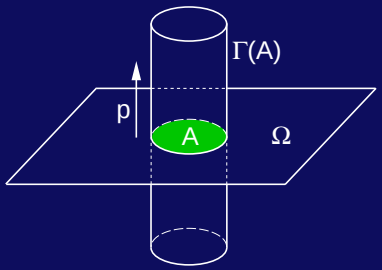
- Motivation
- Molecular Dynamics
- Transfer Operator Approach
- Molecular Kinetics
- Perron Cluster Analysis
- Eigenvector Computation
- Applications

Transfer operators

Position dependent Spatial Transition Operator

- Momenta averaging

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



Schütte, 1998

Schütte, Fischer, Huisinga, Deuflhard 1998

Conformations as Almost Invariant Sets

Consider subsets $A, B \subset \Omega$ and their characteristic functions χ_A and χ_B.

- Probability to be within A
  $w(A) = \int\limits_{\Gamma(A)} \pi(p,q)dq\ dp = \int\limits_{A} \mathcal{Q}(q)dq = \langle \chi_A, \chi_A \rangle_{\mathcal{Q}}$
- Probability to move from A $\rightarrow$ B during time τ
  $w(A, B, \tau) = \langle T^\tau \chi_A, \chi_B \rangle_{\mathcal{Q}} / w(A)$
- Probability to stay within A during time τ
  $w(A, A, \tau) = \langle T^\tau \chi_A, \chi_A \rangle_{\mathcal{Q}} / w(A)$

Conformations: almost invariant subsets of the position space Ω

$$w(A, A, \tau) \approx 1 \quad \leftrightarrow \quad T^\tau \chi_A \approx \chi_A$$

Essential degrees of freedom

Covariance analysis of torsion angles

Berendsen et al., 1993

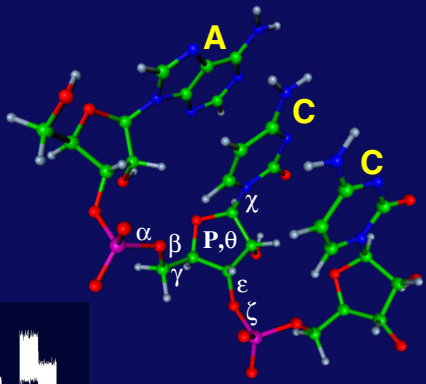
Schütte, Fischer, Huisinga, Deuflhard, 1998

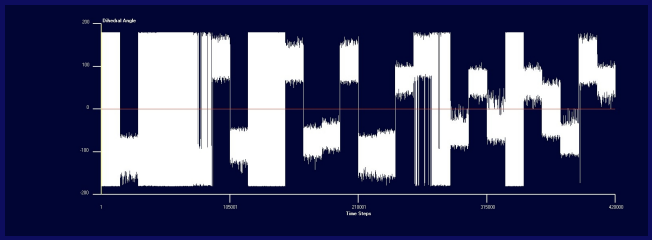
Example: trinucleotide ACC

$n = 37, \quad N > 10^{11}$

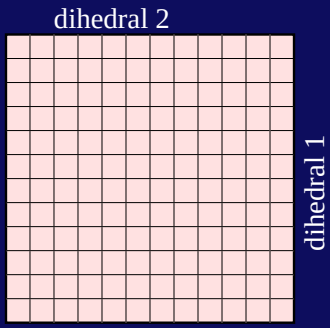
Essential degrees of freedom:

$n_{ess} = 4, \quad N_{ess} = 36$





Box discretization: curse of dimension



Number of spatial boxes: $N \approx s^n$

s : number of minima of torsion potential ($s = 2, 3$)

n : number of torsion angles of molecule ($n \approx 7$ per nucleotide)

Function based Discretization

A function $u(q)$ is called invariant unter the Transfer Operator T if

$$T^\tau u(q) \approx u(q)$$

Eigenvalue problem:

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

Galerkin approach:

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

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

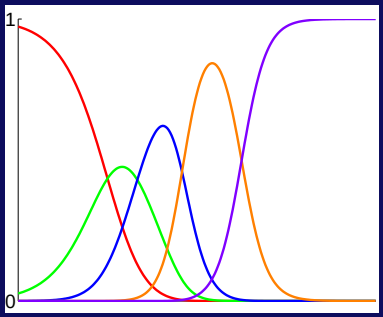
$$\boxed{P\alpha = S\alpha\lambda}$$

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

Radial Basis Functions

$$\varphi_i(\theta) = \frac{\exp(-\alpha \operatorname{dist}(\theta, \theta_i))}{\sum_{j=1}^{n} \exp(-\alpha \operatorname{dist}(\theta, \theta_j))}$$

- partition of unity
- strictly quasi-concave:
  $\varphi(\lambda x + (1-\lambda)y) > \min\{\varphi(x), \varphi(y)\}$
- unimodal
  (at most one maximizer per 2π-period)



Approximation of the Transition Operator

Approximation of matrices by Monte Carlo importance sampling:

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

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

Note:

- the points q_i must be distributed according to the Boltzmann density
- the transfered points $\overline{q}_i$ must be obtained from MD

Detailed Balance

Idea: generate a Markov chain $\{q_k\}_{k=1,2,...}$

$$q_1 \quad \rightarrow \quad q_2 \quad \rightarrow \quad q_3 \quad \rightarrow \quad q_4 \quad \rightarrow \quad \dots q_N$$

which satisfies the detailed balance condition

$$\pi(q_i)P(q_i \rightarrow q_j) = \pi(q_j)P(q_j \rightarrow q_i)$$

In equilibrium, the average number of moves leaving state i must be equal to the number of moves from all other states to state i:

$$\sum_j \pi(q_i)P(q_i \rightarrow q_j) = \sum_j \pi(q_j)P(q_j \rightarrow q_i)$$

Sampling Strategies

Generation of the set $\{q_i\}_{i=1}^N$

$$P(q_i \rightarrow q_{i+1}) = P_v(q_i \rightarrow q_{i+1}) P_a(q_i \rightarrow q_{i+1})$$

- Metropolis Monte Carlo: P_v is symmetric
  choose a new geometry randomly and accept with

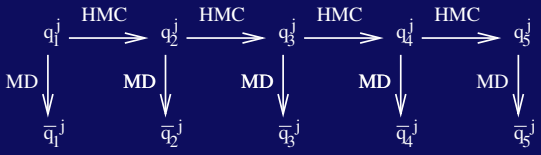
$$P_a(q_i \rightarrow q_{i+1}) = \begin{cases} \exp(-\beta(V(q_{i+1}) - V(q_i)), & V(q_{i+1}) > V(q_i) \\ 1, & \text{else} \end{cases}$$

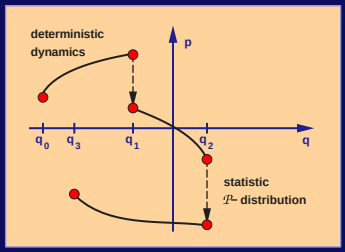
- Hybrid Monte Carlo: P_v is reversibel

  discrete stochastic dynamical system : $q_{i+1} = \xi_1 \Phi^\tau(q_i, p_i)$

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

$$P_a(q_i \rightarrow q_{i+1}) = \begin{cases} \exp(-\beta(H(q_{i+1}) - H(q_i)), & H(q_{i+1}) > H(q_i) \\ 1, & \text{else} \end{cases}$$


HMC
HMC
HMC
HMC
MD
MD
MD
MD
MD


deterministic
dynamics
p
q
statistic
distribution

Alternative Sampling Strategies

Problem: Slow convergence to the Boltzmann distribution because transitions between domains of low energy are rare events (trapping)

- replica exchange
  - start chains at different temperatures
  - high temperature chains are rapidly mixing
  - exchange points between different chains at certain time steps
- HMC with jumps
  - random jumps between domains of low energy
  - requires a list of minimized configurations (small cost compared to sampling costs)

Outline

- Motivation
- Molecular Dynamics
- Transfer Operator Approach
- Molecular Kinetics
- Perron Cluster Analysis
- Eigenvector Computation
- Applications

Transition Matrix

continuous-time:

$$Q = \lim_{t \mapsto 0+} \frac{P(t) - Id}{t} \quad \rightarrow \quad P(t) = \exp(tQ)$$

$P(t)$: transition semigroup in terms of infinitesimal changes of the transition probabilities
Q: infinitesimal generator

discrete-time: homogenous Markov jump process on a finite state space $S = \{0, 1, \ldots, N\}$, exponentially distributed jump times

$$P(t) = \exp(t\Lambda(K - Id))$$

K: transition matrix for the embedded Markov chain of the Markov jump process

$$Q = \Lambda(K - Id)$$

- $\lambda_i = -q_{ii}$: inverse average life time
- $k_{ij} = q_{ij}/\lambda_i$: probability of jumping to state j when leaving state i

Potential of Mean Force

- cartesian coordinates $(q_1, \ldots, q_N)$
  $\rightarrow$ internal coordinates $(\theta_1, \ldots, \theta_k, r_{k+1}, \ldots, r_n), \quad N = k + 3$
- separate relevant d. o. f. θ from irrelevant d. o. f. r
- Potential of Mean Force

$$V^{mf}(\theta) = -\frac{1}{\beta} \ln \left[\int \exp(-\beta V(\theta, r))\, d\, r \right]$$

problem: cannot be evaluated by simulations

$$\frac{\partial V^{mf}(\theta)}{\partial \theta_i} = \frac{\int \frac{\partial V(\theta,r)}{\partial \theta_i} \exp(-\beta V(\theta, r))\, d\, r}{\int \exp(-beta V(\theta, r))\, d\, r} \tag{0.1}$$

$$= \left\langle \frac{\partial V(\theta, r)}{\partial \theta_i} \right\rangle_r \tag{0.2}$$

derivative of potential of mean force is the ensemble averaged constraint force

Point Concept

- 1st order approximation of energy differences:

$$V^{mf}(\theta_i) - V^{mf}(\theta_j) \approx 0.5(\nabla V^{mf}(\theta_i) + \nabla V^{mf}(\theta_j))(\theta_i - \theta_j)$$

- approximation of transition probabilities:

$$\frac{\pi(\theta_j)}{\pi(\theta_i)} = \frac{P(\theta_i \to \theta_j)}{P(\theta_j \to \theta_i)} = \exp(-\beta(V^{mf}(\theta_j) - V^{mf}(\theta_i))) \quad \text{(Arrhenius)}$$

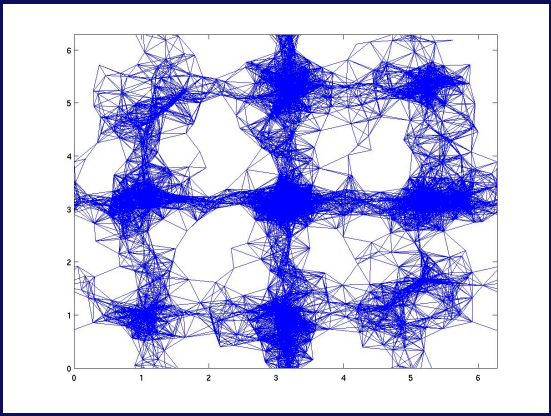
- set of possible transition ways:

$$\mathcal{W} = \{(i,j) \,|\, \theta_i, \theta_j \in \mathcal{S},\, d(\theta_i, \theta_j) \leq \varepsilon\}$$

d: distance measure according to the Euclidian distance in $\mathbb{R}^d$

Point Concept

We describe the dynamics of the molecule as a random walk on $\mathcal{S}$ along pathes in $\mathcal{W}$.



Master Equation

$x_i(t)$: probability of the molecule to be in state θ_i at time t
q_{ji} : transition rate from state θ_j to state θ_i

Master Equation

$$\dot{x}_i(t) = \sum_{j,\,(i,j)\in\mathcal{W}} q_{ji}\, x_j(t), \quad q_{ii} = -\sum_{j,\,(i,j)\in\mathcal{W}} q_{ij},$$

$$\dot{\mathbf{x}} = Q^\top \mathbf{x}.$$

Q: infinitesimal generator of an underlying continuous-time Markov jump process, given by the transition probability matrix P with $P(t) = \exp(tQ)$

Invariant Sets:
subset of states for which Q reduced to this subset represents a closed system (conservation of mass, row sum zero)

$$Q\chi = 0, \quad \chi_i \in \{0, 1\}$$

Transition Probabilities versus Transition Rates

Q can be reduced to a symmetric matrix Q^* by an orthogonal similarity transformation:

$$Q = D^{-1}S = D^{-1/2}(D^{-1/2}SD^{-1/2})D^{1/2} := D^{-1/2}Q^*D^{1/2}$$

$\longrightarrow$ Q is diagonalizable and has real eigenvalues

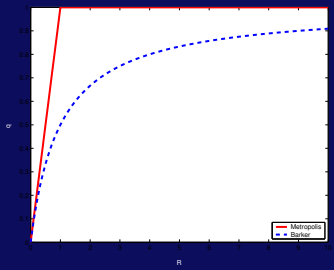
$$\lambda(Q) \subseteq \bigcup_{i=1}^{N}[-2|q_{ii}|, 0]$$

$$Q = X\Lambda X^{-1}$$

$$P(t) = \exp(tQ) = X\exp(t\Lambda)X^{-1} = X\,\mathsf{diag}(\exp(t\lambda_1), \ldots, \exp(t\lambda_n))X^{-1}$$

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

Approximation of Transition Rates



$$q_{ij} = \frac{1}{\pi_i} S(i,j), \quad j \in \mathcal{N}(i)$$

$$R_{ji} = \frac{\pi_i}{\pi_j}$$

Examples:

- $S(i,j) = \min(\frac{\pi_i}{N_i}, \frac{\pi_j}{N_j}), \quad q_{ij} = \frac{1}{N_i} \min(1, \frac{N_i}{N_j} R_{ij})$ Metropolis
- $S(i,j) = 1/(\frac{N_i}{\pi_i} + \frac{N_j}{\pi_j}), \quad q_{ij} = \frac{R_{ij}}{N_j + N_i R_{ij}}$ Barker

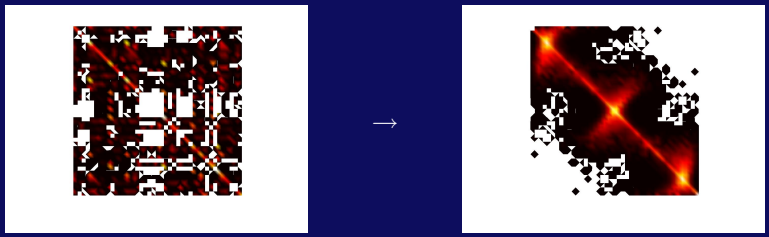
Outline

- Motivation
- Molecular Dynamics
- Transfer Operator Approach
- Molecular Kinetics
- Perron Cluster Analysis
- Eigenvector Computation
- Applications

What does PCCA+ do?

Row stochastic matrix T:

- $T(i,j) =$ Transition probability $i \to j$ in a detailed balanced Markov chain.
- $T(i,j) = \frac{S(i,j)}{\sum_l S(i,l)}$, where S is a symmetric measure of similarity.



Eigenvector Transformation

T_1	0	0
0	T_2	0
0	0	T_3

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

$\tilde{T}_1$	E_{12}	E_{13}
E_{32}	$\tilde{T}_2$	E_{23}
E_{31}	E_{32}	$\tilde{T}_3$

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

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

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,

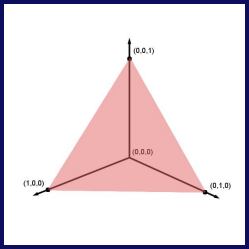
iv) for all $i = 1, \ldots, k$ there exists $l \in \{1, \ldots, N\}$ with $\widetilde{\chi}_i(l) = 1$.

Then

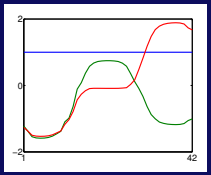
- 3 out of 4: easy to assure
- all 4: if solution exists, then unique

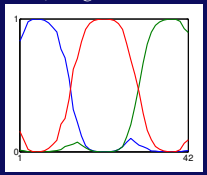
Inner Simplex Algorithm

- consider the n rows of $X \in \mathbb{R}^n$ as points in $\mathbb{R}^{N_C}$
- find those points which are the corners of a simplex
- these points define the transformation
- Deuflhard, Weber (2003)



Example n-butane: $N = 42$ states, $N_C = 3$ cluster





Quality Measurement

What is the appropriate number of clusters?

- minChi-indicator: the positivity requirement should be satisfied as good as possible

$$\text{minChi} = |\min_{i,j}(\chi(i,j))| \quad \rightarrow \quad \min!$$

- gap in the eigenvalue spectrum
- sharpness of the membership vectors χ

Outline

- Motivation
- Molecular Dynamics
- Transfer Operator Approach
- Molecular Kinetics
- Perron Cluster Analysis
- Eigenvector Computation
- Applications

Eigenvector Computation

$$Ax = \lambda x$$

- partial Schur Form: $AQ_k = Q_k R_k$ (orthonormal basis)
- deflation: avoid the multiple computation of eigenvalues

$$q_k \perp Q_{k-1}, \quad A[Q_{k-1}\, q_k] = [Q_{k-1}\, q_k] \begin{bmatrix} R_{k-1} & a \\ 0 & \lambda_k \end{bmatrix}$$

$\rightarrow$ (q_k, λ_k) is an eigenpair of the deflated matrix

$$A_d = (I - Q_{k-1}Q_{k-1}^*)A(I - Q_{k-1}Q_{k-1}^*)$$

- Jacobi-Davidson

  1. subspace method:
     - look for an approximate eigenvector q in a search space $V \perp Q_{k-1}$
     - compute the Schur form of the matrix

       $$M = V^* A_d V = V^* A V, \quad MS = ST,$$

       where S and T are ordered such that $\lambda = T(1,1)$ is closest to some target value τ
     - the pair $(q, \lambda) = (VS(:,1), T(1,1))$ is an approximation for a wanted eigenpair of A
  2. subspace expansion: expansion of V with the solution of the Jacobi correction equation

     $$(I - \tilde{Q}\tilde{Q}^*)(A - \lambda I)(I - \tilde{Q}\tilde{Q}^*)v = -r$$

     with $r = (A_d - \lambda I)q$ and $\tilde{Q} = [Q_{k-1}\, q]$

solution of the correction equation

- equivalent matrix system:

$$\begin{bmatrix} A - \lambda I & -\tilde{Q} \\ \tilde{Q}^* & 0 \end{bmatrix} \begin{bmatrix} v \\ \Delta\lambda \end{bmatrix} = \begin{bmatrix} -r \\ 0 \end{bmatrix}$$

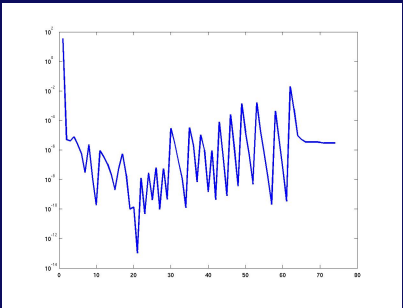
- note: we know $q_1 = (1, 1, \ldots, 1)$

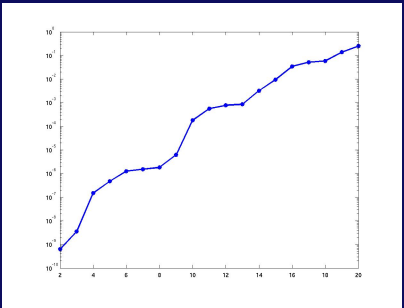
$$\begin{bmatrix} A - \lambda I & -q_1 & -\hat{Q} \\ q_1^* & 0 & 0 \\ \hat{Q}^* & 0 & 0 \end{bmatrix} = \begin{bmatrix} \tilde{M} & 0 \\ \hat{Q}^* & I \end{bmatrix} \begin{bmatrix} I & -\tilde{M}^{-1}\hat{Q} \\ 0 & Z \end{bmatrix}$$

  with $Z = \hat{Q}^* \tilde{M}^{-1} \hat{Q}$

- sufficient to solve it with a high quality preconditioner for $\tilde{M}$ (Sleijpen and Wubs 2003)
- choose $\lambda = \tau$ over all outer iterations and compute a sparse incomplete LU-decomposition with small drop tolerance (UMFPACK)

- epigallocatechine, $Q \subset \mathbb{R}^{1914 \times 1914}$
- the first 20 eigenvalues





Outline

- Motivation
- Molecular Dynamics
- Transfer Operator Approach
- Molecular Kinetics
- Perron Cluster Analysis
- Eigenvector Computation
- Applications

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 $\widetilde{W}$:

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

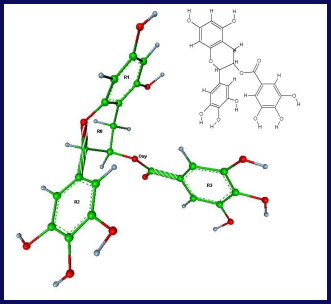
w_{ij} : probability to move from i to j

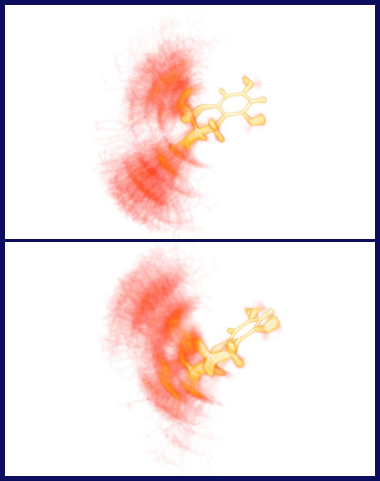
$w_{ii} \approx 1$: probability to stay within i

$w_{ii} > 0.5$

Epigallocatechine

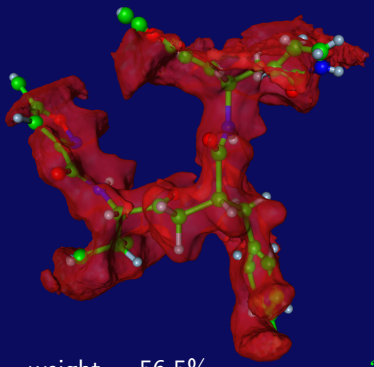
3d representation, line formula and density plot of conformations at 300K





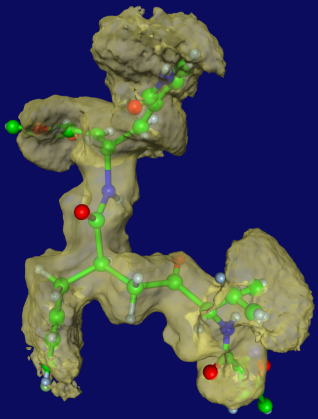
SARS protease inhibitor: conformations

Frank Cordes, Alexander Fischer, 2003



weight = 56.5%

300K



weight = 35.6%

Conclusions

- many challenging mathematical topics
- balance between modelling errors and numerical errors
- Visit our homepage: http://www.zib.de/MDGroup

Thank you for your attention!!!