Chapter 5

Applications

5.1 Conformational analysis of biomolecules

5.1.1 Introduction

The analysis of biomolecular structure and function is one of the real challenges of scientific computing nowadays. Advances in this area will have tremendous impact on the design and identification process of new pharmaceutical drugs. The enrichment of chemical databases with structural and functional information will allow the use of *virtual screening* procedures, reducing time and costs of the pharmaceutical research decidedly.

The key concept to characterize *structure* has become the characterization in terms of *geometric conformations*, often just called conformations in the literature. In contradiction to structure, *function*, seems to depend on the dynamic properties of the molecule and therefore should be rather characterized by what has been called *metastable conformations*. Any type of conformations consists of sets of possible molecular states. In geometric conformations such sets are defined via the geometric similarity of different states. In metastable conformations such sets are defined via the high probability of the molecule to stay in such a set, once it is in such a set.

In classical molecular dynamics [2] a molecule is modeled by a Hamiltonian function

$$H(q,p) \;=\; \tfrac{1}{2}\, p^T M^{-1} p \;+\; V(q),$$

where q and p are the corresponding positions and momenta of the atoms, M denotes the diagonal mass matrix, and V is a differentiable potential. The Hamiltonian function H is defined on the phase space. The corresponding canonical

equations of motion

$$\dot{q} = M^{-1}p, \qquad \dot{p} = -\text{grad}\, V \tag{5.1}$$

describe the dynamics of the molecule. The formal solution of (5.1) with initial state $x_0 = (q(0), p(0))$ is given by $x_t = (q(t), p(t)) = \Phi_V^\tau x_0$, where Φ_V^τ denotes the flow.

In [14] a first attempt had been made to identify metastable conformations on the basis of the so-called Perron-Frobenius operator. That approach, though principally opening the door to the new concept of conformation dynamics, had been more or less restricted to toy molecules. In a further step, performing some momenta averaging based on the Boltzmann distribution f_0 for given heat bath temperature, the Perron-Frobenius operator in phase space has been replaced by a different Markov operator in position space [58, 59]. This new operator has much nicer theoretical properties and it may be interpreted as the transfer operator of an underlying Markov chain $X(t)$. This Markov chain can be realized via Hybrid Monte-Carlo (HMC) methods [22]:

- random choice of momenta from a Gaussian distribution,
- deterministic propagation of the molecular system by the flow Φ_V^τ with potential V and over short time τ,
- acceptance or rejection of new configurations by an appropriate transition kernel K of the underlying Markov process, e.g., Metropolis-Hastings.

Like classical Monte-Carlo, HMC also suffers from possible *trapping* in local potential wells. In order to overcome this unwanted effect, an adaptive temperature version has been worked out [22] that embeds the given problem into a family of problems with flow $\Phi_V^{\tau,s}$ in terms of an embedding parameter $s \in [0, 1]$. At $s = 0$, only a few metastable subsets need to be identified, whereas at $s = 1$ a rich structure of conformations might arise. Two types of embedding are in quite common use: *temperature embedding* and *potential embedding*. Upon examining the equations of motion, one immediately sees that, in the context of HMC, temperature embedding can be realized by the following flow:

$$\Phi_V^{\tau,s} = \Phi_{sV}^{s^{-2}\tau}, \tag{5.2}$$

which requires a scaling of the potential and the time step of propagation [58].

Any kind of embedding stimulates the idea of a hierarchical algorithm consisting of the following steps:

1. Simulate the molecular system for a specific parameter (say, high temperature), which causes the flow to overcome specific energy barriers.

2. Identify metastable subsets.

3. Increase the parameter (say, lower the temperature), but restrict the simulation to one of the metastable subsets. Go to step 1.

This algorithm will generate a hierarchy of subsets that can be sampled independently at each level. The restriction of an HMC-simulation to a given metastable subset C_s requires only a slight modification of the Markov kernel K to K_s [23]. The additional rule is that any configuration outside the subset C_s will be rejected. Detailed balance still holds for this modified Markov kernel so that K_s is still reversible. Since C_s is metastable, only a few rejections will be expected with respect to the new rule. Moreover, trapping should thus be avoided, since energy barriers towards all other metastable subsets can be ignored. A further exploitation of this embedding structure is given in [23], where an uncoupling/coupling technique has been suggested and worked out.

A schematic diagram of such a hierarchy is given in Figure 5.1. As can been seen there, each cluster needs to be described by appropriate boundaries. To save computer time over the whole simulation, one is interested in efficient descriptions of the identified metastable subsets (see section 1.3).

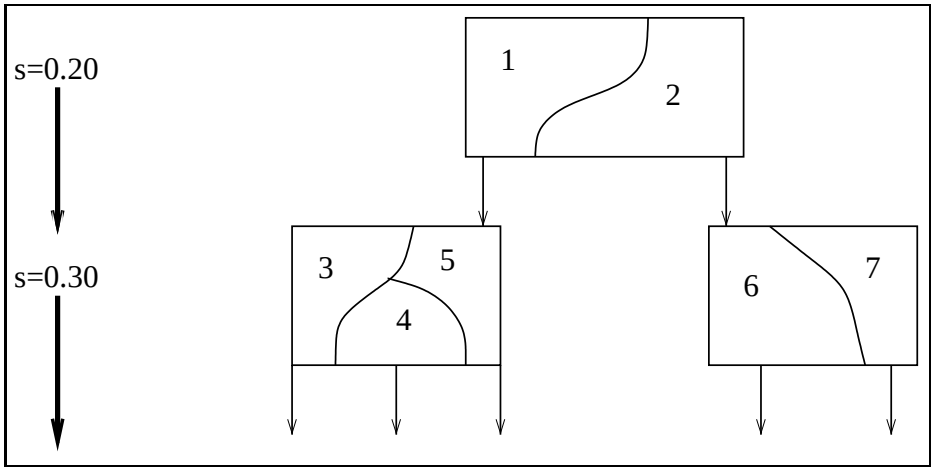

Figure 5.1: **Hierarchical scheme of clustering combined with parameter embedding.** The numbers denote metastable conformations at different levels of the hierarchical embedding scheme.

As described in section 1.1, the problem of finding metastable conformations can be transformed into a cluster problem, if we use a sufficiently long Markov chain $X(t)$ as a representative trajectory. Since $X(t)$ is reversible (see [59]), we can use Perron Cluster analysis to determine an optimal number k of metastable conformations (see section 4.3).

Based on an uniform box decomposition, the conformations of small molecules like *n-pentane* have been recently analyzed successfully [58]. For larger molecules such a simple decomposition is not possible, because the number of boxes explode (see section 2.2). Therefore the use of approximate box decompositions, computed via the SOBM algorithm, allows for the first time the conformational analysis of molecules of practically relevant size.

5.1.2 Adaptation of SOM and SOBM to cyclic data

One easily checks that the computing time of the SOM and the SOBM algorithm strongly depends on the dimension of Ω. The dimension of the position space of molecules is three times the number of atoms and therefore it is very large even for small molecules. The following observation leads to a reduction of the dimension: For each molecule there exists a set of so called *torsion angles*, which are sufficient for a rough reconstruction of the spatial position of each atom of the molecule together with the corresponding equilibrium bonds and angles [39]. Without loss of generality we assume each torsion angle within $[-\pi, \pi]$. Then we define Ω as the space spanned by the torsion angles of the molecule. Since the analysis of cyclic data is different from non-cyclic data (see [24] for a comprehensive introduction), it is not surprising that we have to adapt the SOM and the SOBM algorithm to cyclic data.

First one has to choose a suitable distance measure. We suggest to use the distance on the q-dimensional unit circle, i.e. we define $\mathrm{dist} : \Omega \times \Omega \to \mathbf{R}_0^+$ via

$$\begin{aligned} \mathrm{dist}(x,y) &:= F(d_1(x_1,y_1),\ldots,d_q(x_q,y_q)) := (\sum_{i=1}^{q} d_i(x_i,y_i))^{1/2} \\ \text{with} \quad d_i(x_i,y_i) &:= (\sin(x_i)-\sin(y_i))^2 + (\cos(x_i)-\cos(y_i))^2 \end{aligned}$$

for $x, y \in \Omega$, where x_i and y_i denote the values of the ith torsion angle.

Next we have to assure that the codebook is adapted in the right direction (see Figure 5.2). For the SOM algorithm this requires that the input vector $x(t)$ or the old codebook vector $w_s(t)$, respectively, may need to be transformed first, before the new codebook vector $w_s(t+1)$ can be computed according to Eq. (3.5):

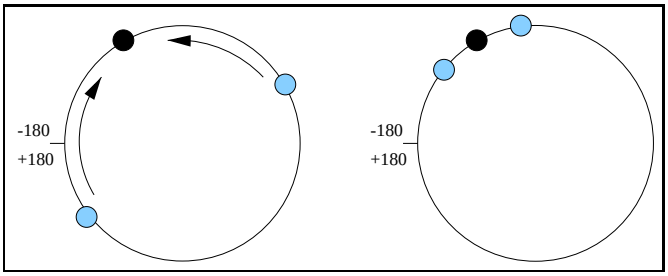

Figure 5.2: **Example: Adaptation of the codebook vector (grey) in direction of the input vector (black) on the shortest way.**

Cyclic Transformation Rules (SOM)

1. IF $w_{s_i}(t) \geq 0$ AND $x_i(t) < 0$ AND $\text{abs}(w_{s_i}(t)) + \text{abs}(x_i(t)) > \pi$
   THEN $x_i(t) := x_i(t) + 2\pi$

2. IF $w_{s_i}(t) < 0$ AND $x_i(t) \geq 0$ AND $\text{abs}(w_{s_i}(t)) + \text{abs}(x_i(t)) > \pi$
   THEN $w_{s_i}(t) := w_{s_i}(t) + 2\pi$

   Note that we have $\text{abs}(x) := \sqrt{x^2}$ for $x \in \mathbf{R}$.

After the new codebook vector has been computed, eventually it must also be transformed so that each component $W_{s_i}(t+1)$ is inside the interval $[-\pi, \pi]$. Figure 5.3 shows an one-dimensional example for the first case.

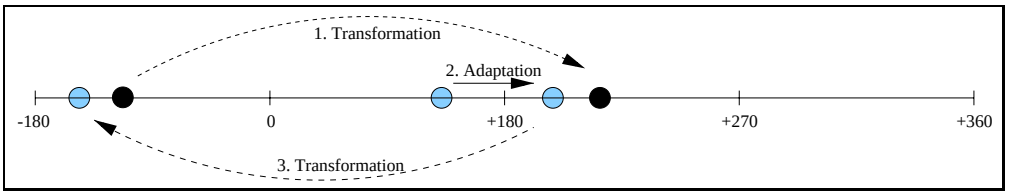

Figure 5.3: **Example: Transformations of the codebook vector (grey) and the input vector (black) to guarantee correct adaptation.**

To use cyclic data within the SOBM algorithm, we need more sophisticated rules, because we have to distinguish between normal and complementary intervals:

If $l_{s_i} < r_{s_i}$, we call $\hat{W}_{s_i} := [l_{s_i}, r_{s_i}]$ a normal interval. But we allow also the case $l_{s_i} > r_{s_i}$. In this case we have $\hat{W}_{s_i} := [-\pi, \pi] \setminus [r_{s_i}, l_{s_i}]$, i.e. $\hat{W}_{s_i}$ is the complementary interval of $[r_{s_i}, l_{s_i}]$.

First we have to refine function $g : [-\pi, \pi]^3 \to [0, 1]$ used within the codebook adaptation rules (see Eq. (3.7)):

Case 1: $a < b$. Set

$$g(a,b,x) := \begin{cases} 1 & \text{if } x \notin [a,b] \wedge d_i(x,a) \leq d_i(x,b) \\ 0 & \text{if } x \notin [a,b] \wedge d_i(x,a) > d_i(x,b) \\ \frac{b-x}{\iota([a,b])} & \text{else} \end{cases}$$

with $\iota([a,b]) := (b-a)$.

Case 2: $a > b$. Set

$$g(a,b,x) := \begin{cases} 1 & \text{if } x \in [b,a] \wedge d_i(x,a) \leq d_i(x,b) \\ 0 & \text{if } x \in [b,a] \wedge d_i(x,a) > d_i(x,b) \\ \frac{2\pi+(b-x)}{\iota([a,b])} & \text{if } x \notin [b,a] \wedge x \geq a \\ \frac{b-x}{\iota([a,b])} & \text{else} \end{cases}$$

with $\iota([a,b]) := 2\pi + (b-a)$.

Next we have to specify the necessary transformations to guarantee a correct adaptation of the codebook boxes:

Cyclic Transformation Rules (SOBM)

If $\hat{W}_{s_i}(t) := [l_{s_i}(t), r_{s_i}(t)]$ with $l_{s_i}(t) > r_{s_i}(t)$ or if $x_i(t)$ is not inside the complementary interval $\hat{W}_{s_i}(t)$, i.e. $x_i(t) \in [r_{s_i}(t), l_{s_i}(t)]$, then we have to consider the earlier defined cyclic transformation rules for the SOM algorithm, with $l_{s_i}(t)$ and $r_{s_i}(t)$ instead of $W_s(t)$. But if $x_i(t)$ is inside the complementary interval $\hat{W}_{s_i}(t)$, i.e. $x_i(t) \notin [r_{s_i}(t), l_{s_i}(t)]$, one has to consider slightly different transformation rules to assure that the boundaries are adapted towards the correct direction:

IF $g(l_{s_i}(t), r_{s_i}(t), x_i(t)) > g(-r_{s_i}(t), -l_{s_i}(t), -x_i(t))$ THEN
  Use the cyclic transformation rules (SOM) for the adaptation of $l_{s_i}(t)$.
  IF $x_i(t) > r_{s_i}(t)$ THEN
    First set $x_i(t) := x_i(t) - 2\pi$, afterwards adapt $r_{s_i}(t)$ directly
    (i.e. without further transformation).
  ELSE
    Adapt $r_{s_i}(t)$ directly.
ELSE
  Use the cyclic transformation rules (SOM) for the adaptation of $r_{s_i}(t)$.
  IF $x_i(t) < l_{s_i}(t)$ THEN
    First set $x_i(t) := x_i(t) + 2\pi$, afterwards adapt $l_{s_i}(t)$ directly.
  ELSE
    Adapt $l_{s_i}(t)$ directly.

If the width of the interval $[l_{s_i}(t), r_{s_i}(t)]$ is nearly 2π, then one observes sometimes the artifact that left and right boundaries interchange so that the interval becomes "too small". In this case the adaptation step has to be skipped and the interval $[-2\pi + \epsilon, 2\pi - \epsilon]$ has to be fixed as the new value of $\hat{W}_{s_i}(t+1)$.

5.1.3 Numerical results: HIV protease inhibitor

The fact that the cleavage of the HIV polyprotein by HIV protease is essential for viral propagation, has made the HIV protease a key target for the design of drugs against AIDS. The recent development of HIV protease inhibitors has dramatically improved the therapeutic outcome for many AIDS patients. Unfortunately, these inhibitors are very expensive and the effectiveness of therapy can encounter problems with drug-resistant viral strains. So there is further strong interest in the development of other classes of HIV protease inhibitors [10]. It is obvious that with a deeper understanding — including knowledge about the dynamic behavior — of the existing inhibitor molecules, it becomes much easier and cheaper to find and to design new inhibitor classes. In the following we present the numerical results of the conformational analysis of the HIV-protease inhibitor VX-478.

The inhibitor VX-478 of the enzyme HIV protease consists of 70 atoms. The molecule was parameterized by the Merck molecular force field (MMFF) [37]. Figure 5.4 shows one possible state (configuration) of the molecule.

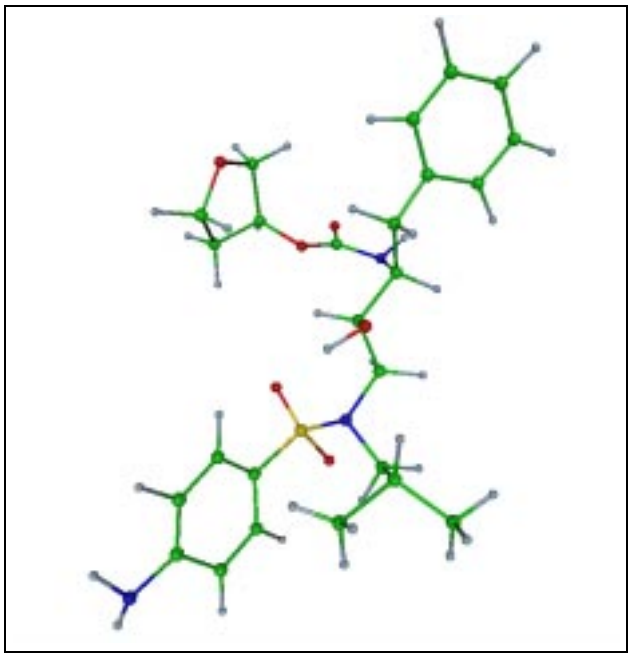

Figure 5.4: **Possible configuration of the HIV-protease inhibitor.**

As noted in Eq. (5.2), the sampling of a thermodynamic distribution at various temperatures within a temperature embedding can be realized by a correlated scaling of time steps and potential [58].

The Hybrid Monte Carlo (HMC) simulations are performed with temperature dependent time steps (fs = femtoseconds)

$$\tau = \frac{1.4}{\sqrt{\frac{300}{T[K]}}} \text{ fs}.$$

Each new configuration is generated by a propagation of the system over a random length between 40 and 80 time steps and each simulation consists of 5 independent Markov chains. For every configuration 34 torsion angles are stored which are sufficient for a rough reconstruction together with the corresponding equilibrium bonds and angles. Convergence of the HMC-simulation is reached, as soon as the Gelman and Rubin quotient R [34, 9] is sufficiently close to the value 1. Note that the choice of what is "sufficiently close to 1" is rather critical, because on the one hand one is interested in fast simulations, but on the other hand a worse convergence bears the risk of sampling not the whole configurational space. In [30] the focus was definitely on fast simulations, leading to a sampling of only parts of the configurational space. Together with a slight different choice of parameters[1] this has led to a detection of conformations even at rather high temperatures. In the following the results of simulations with much better convergence properties are presented, where the Gelman and Rubin quotient accomplishes the rigorous condition $\|1 - R\| \leq 0.05$.

Based on the five Markov chains we have constructed the data set V, the frequency function f and the homogeneity function h_S as described in section 1.1. The computation of the approximate box decomposition of V with respect to f was done automatically via a combination of the SOM and the SOBM algorithm with pruning and early stopping (see section 3.3+3.5). Note that the chosen parameters are comparable with the suggestions in the SOM literature [48]:

1. As an upper bound for the number of partitions Θ_s we have chosen an upper bound $\Bbbk := 600$, what is large enough to guarantee robust results, i.e. nearly equal results, if $\Bbbk$ is changed slightly.

2. The computation of a 25×24 SOM was done by performing $u \cdot \Bbbk$ ordering steps (with $\alpha(0) = 1.0$, $\eta := \eta_{gaussian}$ and $\gamma(0) = 12$) and $u \cdot \Bbbk$ convergence steps (with $\alpha(0) := 0.1$, $\eta = \eta_{bubble}$ and $\gamma(0) = 1$), where $u := 50$ denotes the average number of codebook updates.

[1] In [30] shorter time steps and a propagation of fixed length were used. This has reduced the flexibility of the molecular system.

3. We have initialized the SOBM codebook by using only the codebook vectors w_p with $f(\Theta_{w_p}(V)) \geq 2u$. Then we have performed convergence steps (with $\alpha(0) := 0.005$, $\eta := \eta_{bubble}$ and $\gamma(0) := 1$), until the overlap between the codebook boxes has exceeded 0.1%. We have used the final codebook to derive an approximate box decomposition of V according to Lemma 3.2.2.

Cluster identification

For the cluster identification, we have used our extended multilevel approach. First we look at the results, without decomposition refinement (see Table 5.1):

$\mathcal{T}$[K]	N	k	spectrum	coupling matrix	overlay [%]
900	60000	53	$\underline{1.000}$ 0.830 0.805 0.791	1.000	26.5
700	31000	72	1.000 $\underline{0.930}$ 0.885 0.876 0.860 0.795 0.790	0.924 0.076 0.018 0.982	40.5
700-C_0 RS	60000	65	1.000 0.890 0.820 0.798 0.768	1.000	35.5
700-C_1 RS	42000	92	1.000 0.896 0.875 0.824 0.820	1.000	36.4

Table 5.1: **Hierarchical temperature embedding for HIV protease inhibitor with resimulation at level** $\mathcal{T} = 700K$ (N = number of configurations per Markov chain, k = final number of codebook boxes).

While for $\mathcal{T} \geq 900K$ the Perron cluster analysis only identifies one conformation, one observes a large spectral gap between the second (0.930) and the third

(0.885) eigenvalue of the transition matrix $\mathcal{S}$ at level $\mathcal{T} = 700\,K$. To prove the metastability of the identified clusters C_0 and C_1, a resimulation at the same level was performed. As expected the gap between the 1 and the second eigenvalue grows for both clusters, but there are also large gaps between the second (0.890) and the third eigenvalue (0.820) for the first cluster and between the third (0.875) and the fourth (0.824) eigenvalue for the second cluster. But if one looks again at the original spectrum at level $\mathcal{T} = 700\,K$, one finds another large gap between the fifth (0.860) and the sixth (0.795) eigenvalue. Obviously the configurational space at level $\mathcal{T} = 700\,K$ decomposes into two strongly metastable clusters, but also into five weaker metastable subsets (see Table 5.2).

$\mathcal{T}$[K]	spectrum	coupling matrix					overlay [%]
700	1.000						40.5
	0.930	0.908	0.021	0.024	0.031	0.018	
	0.885	0.014	0.874	0.022	0.001	0.090	
	0.876	0.013	0.018	0.879	0.006	0.085	
	0.860	0.044	0.002	0.015	0.896	0.043	
	0.795	0.004	0.029	0.033	0.006	0.928	
	0.790						

Table 5.2: **Weaker metastability:** Five conformations for HIV protease inhibitor at level $\mathcal{T} = 700K$ (31000 configurations, 72 final codebook boxes).

Next we have refined the decomposition after step (2) and performed step (3), until the decomposition was fine enough. At level $\mathcal{T} = 700\,K$, we have achieved the results presented in Table 5.3.

The number of final codebook boxes has increased, leading to a larger second eigenvalue (0.952), a larger gap size and a better coupling matrix. Additionally the overlay has increased (47.7% in comparison with 40.5%), while the overlap still has remained near zero.

For a temperature embedded simulation at level $\mathcal{T} = 500\,K$ inside the both metastable clusters C_0 and C_1, our cluster method computes 4 $(C_{00}, C_{01}, C_{02}, C_{03})$ and 3 conformations (C_{10}, C_{11}, C_{12}) respectively (see Table 5.3). The seven identified conformations have weights $f(C_i)$ according to Table 5.4.

Figure 5.5 and Figure 5.6 show average configurations for always two out of the seven conformations at $\mathcal{T} = 500\,K$. To allow a better comparison the two average configurations are aligned in a plane defined by three common atoms.

$\mathcal{T}$[K]	N	k	spectrum	coupling matrix	overlay [%]
900	60000	53	$\underline{1.000}$ 0.830 0.805 0.791	1.000	26.5
700	31000	113	1.000 $\underline{0.952}$ 0.898 0.889 0.886 0.830 0.802 0.794	0.934 0.066 0.015 0.985	47.7
500-C_0	60000	101	1.000 0.962 0.949 $\underline{0.945}$ 0.917 0.903 0.896	0.921 0.015 0.040 0.023 0.012 0.920 0.023 0.044 0.034 0.024 0.919 0.023 0.010 0.024 0.012 0.954	51.8
500-C_1	60000	72	1.000 0.952 $\underline{0.942}$ 0.920 0.908 0.891	0.961 0.029 0.010 0.025 0.964 0.012 0.044 0.062 0.894	47.3

Table 5.3: **Hierarchical temperature embedding for HIV protease inhibitor with decomposition refinement** (N = number of configurations per Markov chain, k = final number of codebook boxes).

C_{00}	C_{01}	C_{02}	C_{03}	C_{10}	C_{11}	C_{12}
3.3%	4.1%	3.9%	7.4%	33.8%	40.1%	7.5%

Table 5.4: **Weights of the seven conformations for HIV protease inhibitor at level $\mathcal{T}$ = 500 K**

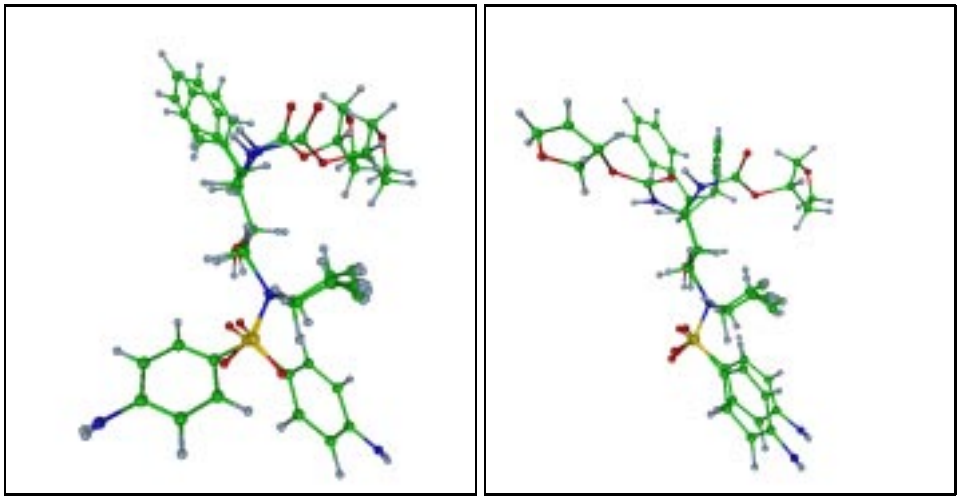

Figure 5.5: **Visualization of conformations of HIV protease inhibitor:** Average configurations for two metastable conformations at temperature level $\mathcal{T} = 500\,K$ (left: C_{00} and C_{02}, right: C_{02} and C_{11}).

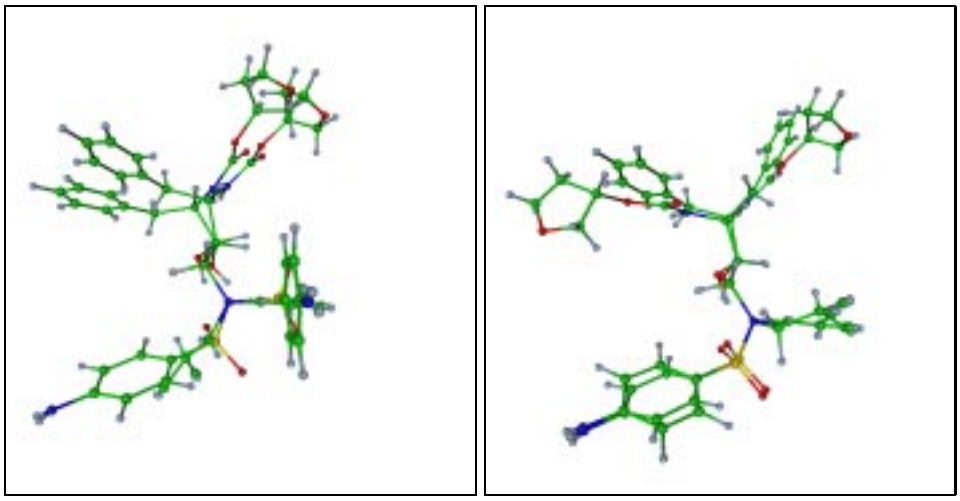

Figure 5.6: **Visualization of conformations of HIV protease inhibitor:** Average configurations for two metastable conformations at temperature level $\mathcal{T} = 500\,K$ (left: C_{01} and C_{03}, right: C_{01} and C_{10}).

For comparison purposes, we have also used mere VQ instead of SOM. In this case Perron Cluster analysis leads to four metastable clusters instead of the three conformations C_{10}, C_{11}, C_{12} at $T = 500K$. Upon careful examination of the results, however, one observes that one of the four clusters is nearly empty — this is the kind of pseudo-clusters already mentioned in chapter 3.

Cluster description

Using the corresponding approximate box decomposition (see Figure 5.7 for a projection of codebook boxes computed by the SOBM algorithm on two out of the 34 torsion angles), we have identified 17 discriminating torsion angles for the clustering $\mathcal{C} := \{C_0, C_1\}$ at $T = 700K$. Further we have used the corresponding 113 codebook boxes to determine reduced membership rules of C_0 and C_1. Here is one of these membership rules for cluster C_1:

IF $v_{*,3} \notin [18.9, 151.8]$ AND $v_{*,4} \notin [-169.4, -29.2]$ AND $v_{*,5} \notin [-82.3, 58.5]$ AND $v_{*,6} \notin [29.3, 168.0]$ AND $v_{*,7} \notin [-45.4, 94.6]$ AND $v_{*,8} \notin [-103.7, 29.0]$ AND $v_{*,15} \notin [-36.4, 99.9]$ AND $v_{*,16} \notin [-160.0, -22.5]$ AND $v_{*,17} \notin [-138.5, -10.1]$ AND $v_{*,18} \notin [-52.1, 67.7]$ AND $v_{*,19} \in [-61.8, 177.7]$ AND $v_{*,26} \in [-148.1, 77.0]$ AND $v_{*,27} \in [-158.1, 68.4]$ AND $v_{*,29} \in [-144.4, 89.2]$ AND $v_{*,30} \in [-110.5, 107.4]$ AND $v_{*,31} \in [-152.7, 76.5]$ AND $v_{*,32} \in [-99.3, 121.3]$ THEN $v = (v_{*,1}, \ldots, v_{*,34}) \in C_1$

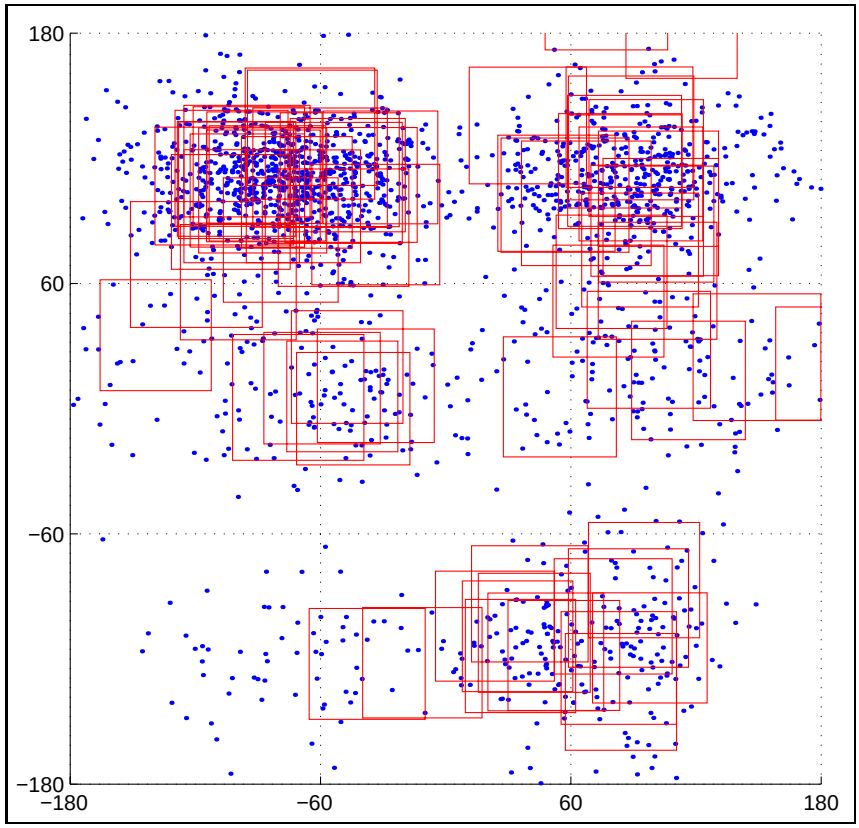

Figure 5.7: **Example: Adaptive box decomposition for HIV protease inhibitor.** Visualized projection of codebook boxes on two out of 34 torsion angles.

5.1.4 Prospect: Virtual screening

Clustering techniques and especially self-organized neural networks have been already used for the analysis of molecular dynamics [43, 41]. But all suggested algorithms have the deficit that they use a geometric cluster model: They try to group geometric conformations to metastable conformations by an investigation of a suitable visualization of the transition probabilities between the geometric conformations. Obviously such a procedure is only possible if the number of geometric conformations is very small, as it is only the case for simple molecules. In contradiction, the method described in the previous subsections is able to compute metastable conformations also for large and complex molecules. Therefore it can be used for a virtual screening of chemical databases.

Example: Virtual screening of CDK inhibitor

Virtual screening of chemical databases is a powerful tool for the identification of derivatives of already known molecules with a function of pharmaceutical interest. Figure 5.8 shows a virtual screening process for the *CDK inhibitor indirubin* in principle: First we have to perform a conformational analysis of indirubin and also of all molecules inside the database, to generate knowledge about their function. Then we have to use suitable matching algorithms (see [52]) to identify molecules inside the database with a similar structure and similar metastable conformations as the indirubin molecule. For a first application of conformational analysis within a virtual screening project see [30].

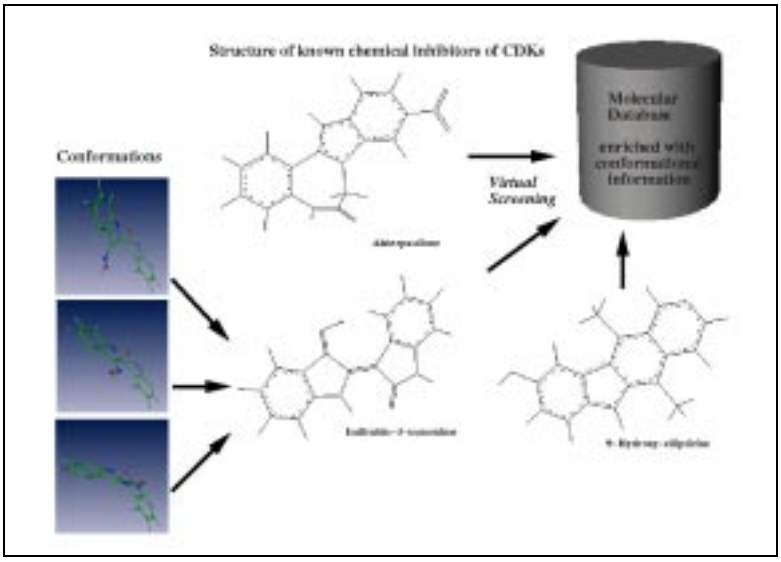

Figure 5.8: **Virtual Screening of CDK inhibitor indirubin.**

5.2 Cluster analysis of insurance customers

Cluster analysis is a powerful tool for insurance companies to get a better understanding of their customer structure, e.g., to design new tariffs or services. In the following we will present a successful applications of our new cluster approach for the analysis of insurance data that has been done in cooperation with RISK-CONSULTING, KÖLN. For a description of a further application see [31].

5.2.1 Modeling

Suppose that each insurance customer can be described by a set of q attributes, e.g., age, sex, occupation. As described in the appendix, we can easily transform the corresponding Ω to a normalized metric space and therefore the customers can be interpreted as points in a set $V \subset \Omega$. Since we want to identify groups of customers, who have similar properties with regard to the different attributes, we have to solve a geometric cluster problem. If the data quality is good, i.e. if we have for each customer valid values for nearly all attributes, we can use a homogeneity measure h_d based on the Euclidean distance function $d = d_{euclid}$. Otherwise we have to use more sophisticated distance measures as, e.g., the Tanimoto measure [48] or measures that use information levels [28]. Since each customer is unique, we use a frequency function f with $f(v) = 1$ for all $v \in V$. If the number of clusters is unknown a priori, we transform h_d into a stochastic homogeneity function $\widetilde{h}_d$ as described in Lemma 4.3.11 so that we can use our extended multilevel approach based on Perron Cluster analysis. Since we cannot be sure that the homogeneity function $\widetilde{h}_d$ corresponds to the same optimal clusters as the original homogeneity function h (see the earlier discussion in connection with Lemma 4.3.11), we have to validate the identified clusters carefully. This is especially necessary, if the artificial construction of $\widetilde{h}_d$ leads to a spectrum with much noise, i.e. a spectrum where the separation between the Perron Cluster and the remaining part is difficult. Obviously an efficient cluster description based on an approximate box decomposition is a helpful tool for cluster validation.

5.2.2 Numerical results: Whiplash Injury Patients

Within our application we have clustered 2153 customers of a German health insurance company with a diagnosis of *whiplash*[2] during the observation years 1996 and 1997. The number of attributes after transformation of Ω into a normalized metric space was 185.

[2]Whiplash (German: Schleudertrauma) is an injury to the cervical spine and its soft tissues caused by forceful flection of the neck, especially that occurring during an automobile accident.

The computation of an approximate box decomposition of V was done with a combination of the SOM and the SOBM algorithm as described in section 3.3. We have used early stopping, but we have not pruned neurons to allow a visual comparison with the results generated by using only the SOM algorithm.

1. As an upper bound for the number of partitions Θ_s we have chosen $\Bbbk := 99$, what is large enough to guarantee robust results, i.e. nearly equal results, if $\Bbbk$ is changed slightly.

2. The computation of a 11×9 SOM was done by performing $100\Bbbk$ ordering steps (with $\alpha(0) = 0.9$, $\eta := \eta_{gaussian}$ and $\gamma(0) = 5$) and $300\Bbbk$ convergence steps (with $\alpha(0) := 0.1$, $\eta = \eta_{bubble}$ and $\gamma(0) = 1$).

3. Using the codebook vectors w_p, we have initialized the SOBM codebook boxes. Then we have performed convergence steps (with $\alpha(0) := 0.005$, $\eta := \eta_{bubble}$ and $\gamma(0) := 1$), until the overlap between the codebook boxes has exceeded the value 0.1%.

In a first trial, we have stopped after step (2). We have used the codebook vectors w_p to determine a decomposition of V and performed a Perron Cluster analysis (see Table 5.5):

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	$\Gamma_{f,h_d}(k=3)$	$\Gamma_{f,h_d}(k=5)$
1.00	0.81	0.72	0.60	0.51	0.38	0.34	0.71	0.60

Table 5.5: **Whiplash Patients:** Perron Cluster analysis using 9×11 SOM.

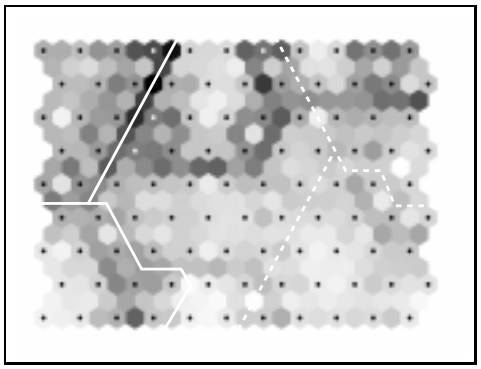

Figure 5.9: **Whiplash Patients:** SOM gray-level visualization including cluster borders computed via Perron cluster analysis (solid border: clusters for $k = 3$, dashed border: two additional clusters for $k = 5$).

The two largest gaps are between λ_3 and λ_4 and between λ_5 and λ_6 respectively. Figure 5.9 shows the borders of the computed clusters within a SOM gray-level visualization[3].

Next we have performed additionally step (3). We have computed an approximate box decomposition of V based on the final codebook boxes and we have used Perron Cluster analysis to determine an optimal clustering:

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	$\Gamma_{f,h_d}(k=3)$	$\Gamma_{f,h_d}(k=5)$
1.00	0.81	0.73	0.62	0.54	0.43	0.35	0.69	0.62

Table 5.6: **Whiplash Patients:** Perron Cluster analysis using 9×11 SOBM.

The algorithm suggests 3 or 5 clusters. Since we have not pruned neurons after step (2), we can visualize the SOBM with gray-levels (see Figure 5.10). The borders computed via Perron Cluster analysis corresponds to the the borders indicated by the dark-shades. Especially the right upper cluster is clearly identified. This cluster contains customers that has been taken over by the insurance company from another company many years ago. It is very interesting that these customers have been grouped together, because we have not used the corresponding attribute within our analysis, i.e. the information "customer has been overtaken" was not given explicitly. Nevertheless there exists a strong relationship between these customers, hidden inside the used attributes. Our cluster algorithm was able to detect these relationship and therefore has generated knowledge.

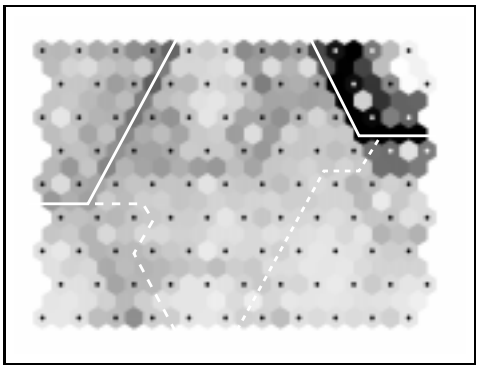

Figure 5.10: **Whiplash Patients:** SOBM visualization including cluster borders computed via Perron cluster analysis (solid border: clusters for $k = 3$, dashed border: two additional clusters for $k = 5$).

[3]SOM gray-level visualization is used to determine the clusters by visual investigation (see [48]). Dark shades represent low homogeneity between the codebook vectors, while light shades represent a high homogeneity . Other techniques for cluster visualization are presented in [61]