Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Structural Biology

Transition networks for modeling the kinetics of conformational change in macromolecules

Frank Noé[1] and Stefan Fischer[2]

The kinetics and thermodynamics of complex transitions in biomolecules can be modeled in terms of a network of transitions between the relevant conformational substates. Such a transition network, which overcomes the fundamental limitations of reaction-coordinate-based methods, can be constructed either based on the features of the energy landscape, or from molecular dynamics simulations. Energy-landscape-based networks are generated with the aid of automated path-optimization methods, and, using graph-theoretical adaptive methods, can now be constructed for large molecules such as proteins. Dynamics-based networks, also called Markov State Models, can be interpreted and adaptively improved using statistical concepts, such as the mean first passage time, reactive flux and sampling error analysis. This makes transition networks powerful tools for understanding large-scale conformational changes.

Addresses
[1] Research Center ''Matheon'', FU Berlin, Arnimallee 6, 14195 Berlin, Germany
[2] Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany

Corresponding author: Noé, Frank (noe@math.fu-berlin.de) and Fischer, Stefan (stefan.fischer@iwr.uni-heidelberg.de)

Current Opinion in Structural Biology 2008, **18**:154–162

This review comes from a themed issue on
Theory and simulation
Edited by Chandra Verma and Juan Fernandez Recio

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2008.01.008

Introduction

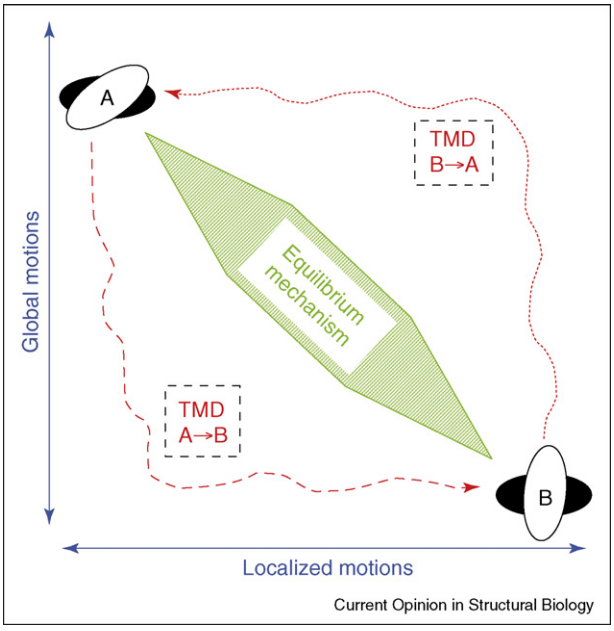
Conformational changes are crucial to the function of proteins and nucleic acids. A variety of processes exist, ranging from binding of macromolecules and their ligands [1], over complex conformational rearrangements switching between native protein substates [2,3••] to the folding of proteins and RNA [4,5]. Understanding the mechanisms of such transitions is challenging, as they involve many degrees of freedom and often occur *via* various pathways with many intermediates.

These processes are often simulated by driving the transition with a few (often one) pre-defined reaction coordinates or order parameters [6–8] while allowing the remaining degrees of freedom to relax. This assumes that the chosen reaction coordinates suffice to define all relevant states of the process, including the transition states, such that the slow transition events are entirely separated in the low-dimensional projection onto these coordinates. Such an approach may work for simple chemical systems, or for processes in which the driving coordinate corresponds to an experimentally applied external force, such as in force-clamp protein unfolding [9,10]. However, reaction-coordinate-based simulation methods tend to perturb the transition mechanism and yield wrong rates for complex equilibrium processes which, for example, involve domain rearrangements and folding. A striking example of this is the method of targeted (or steered) molecular dynamics (TMD), in which a constraint is put on the RMS-difference between the current and target coordinate-set in order to accelerate the transition towards the target coordinate during molecular dynamics simulation. When the equilibrium transition mechanism involves a concerted interplay between localized and large-scale motions, as it often does for macromolecules, TMD biases the order of these motions as illustrated in Figure 1[11].

Moreover, attempting to understand the thermodynamics and kinetics of complex systems in terms of the free-energy landscape in low-dimensional projections is, although appealing, often highly deceptive. This is due to the fact that low-dimensional projections shrink distances between points in space and produce overlaps between conformations that are separated in the full-dimensional state space. This makes free energy barriers disappear in the low-dimensional projection, often producing apparently smooth free energy surfaces with only one or two basins even for systems that contain many kinetically separated substates, as clearly shown in [12,13•].

In order to get an unbiased description of the intrinsically high-dimensional macromolecular dynamics, it is essential to abandon the attempt to control the molecular system in some pre-defined low-dimensional subspace and move instead towards a more comprehensive description of the transition process. The first step in this paradigm shift was the introduction of path-optimization methods [14,15], in which a curvilinear pathway describing the complete transition is treated as a continuous flexible chain of segments, starting from some initial guess (such as a linear interpolation between the two end conformers). This chain can be optimized without

Figure 1



Sequential bias during targeted molecular dynamics (TMD). Example of a transition between conformations A and B of a two-domain protein (shown in black and white). The equilibrium transition channel (green area) consists of paths in which soft domain motions (plotted vertically) alternate with stiff local motions (such as side-chain rearrangements necessary at the interfaces between domains, plotted horizontally). The RMSD-reaction coordinate of TMD responds most when the motion involves many atoms, thus, soft global motions react early to the pulling force, while localized changes occur late in a TMD pathway. Consequently, pulling $A \rightarrow B$ and $B \rightarrow A$ result in different pathways (red dotted lines), a clear indication that the transition mechanism has been biased.

application of external bias. An efficient method capable of doing this automatically for proteins is conjugate peak refinement (CPR) [14], which has allowed to determine several transition mechanisms in proteins [16,17,2,18]. Another popular method is the nudged elastic band (NEB) approach [15]. A typical application domain of path-optimization methods are allosteric processes, that is, conformational transitions where a small change in one region of the protein (such as the binding or the chemical modification of a ligand) triggers large changes in another region of the protein (such as tertiary or quarternary rearrangements). This often involves a well-defined transmission of structural information across the protein. For example, using CPR, it was recently possible to explain in detail the chemo-mechanical coupling between the catalytic ATPase site and the distant force-generating domain of myosin, an ATP-driven molecular motor [2,11].

The path-optimization approach is itself limited when a multitude of transition mechanisms co-exist and formulating many initial guess-paths becomes difficult. This could in principle be alleviated by path sampling methods [19,20] which sample from an ensemble of dynamic pathways connecting two end-states. However, this approach may fail to converge for complex transitions between well-defined end-states, or when the two end-states are separated by multiple transition channels. Thus, a further step towards obtaining a comprehensive description of the transition process and its kinetics was to partition the state space into discrete substates and cast the complex process into a network of simpler transitions between them (see Figure 2 for an illustration). In particular, biomolecular function often depends on the ability to undergo transitions between long-lived intermediates, that is 'metastable' states [21], which are well suited as substates in such a kinetic model. There are two approaches to building a transition network, the 'energy landscape' and the 'dynamical' approach, which are the focus of the present review. We start with a short overview of the theory underlying transition networks.

Modeling kinetics with discrete states

The transition process between conformational substates is often described with the memoryless Master equation:

$$\frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} = \mathbf{p}(t)\mathbf{K}. \tag{1}$$

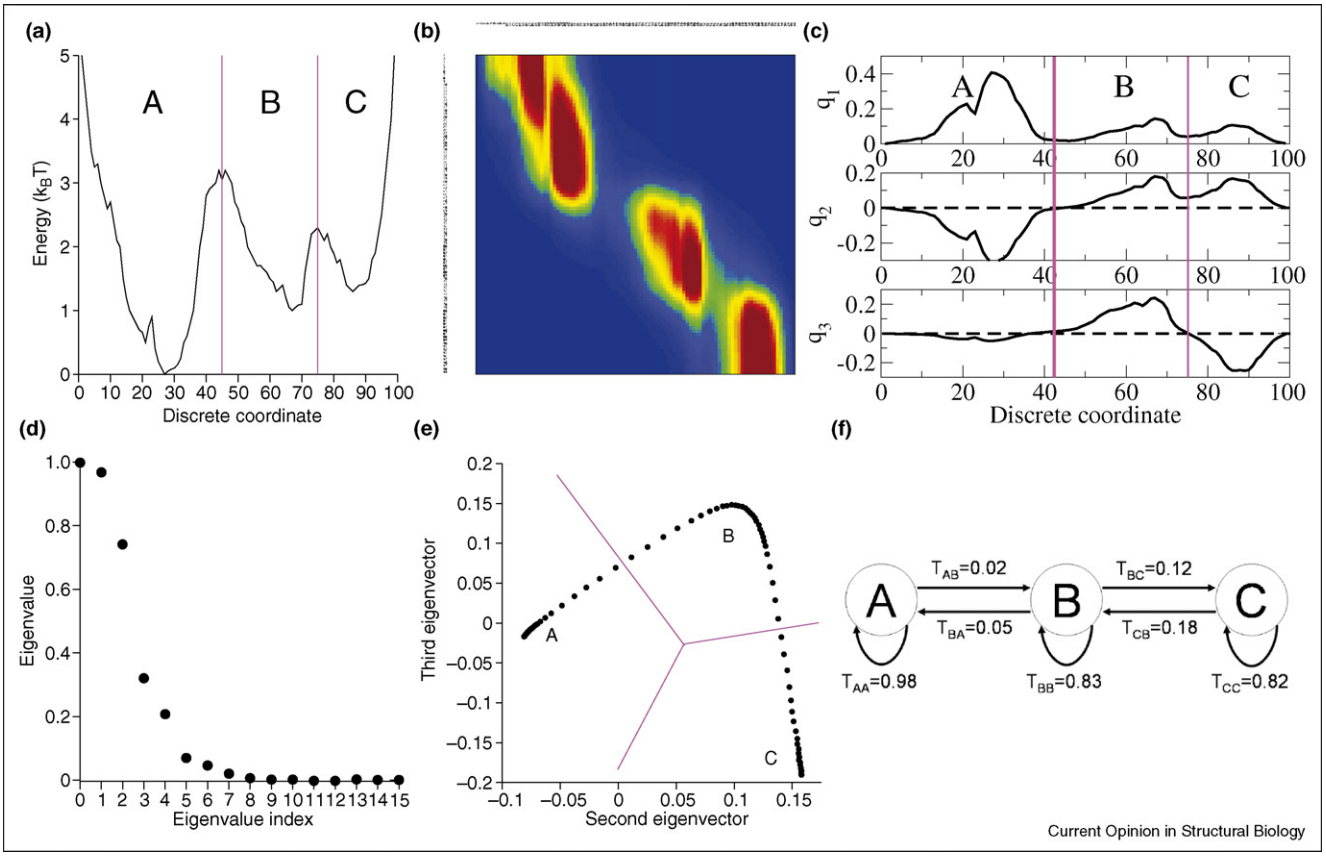
with $\mathbf{p}(t)$ being an m-dimensional column vector containing the probability to find the system in each of its m states at time t. $\mathbf{K}$ is a rate matrix with K_{ij} being the transition rate constant from state i to state j. The diagonal elements of $\mathbf{K}$ are $K_{ii} = -\sum_{j \neq i} K_{ij}$ to ensure mass conservation. Alternatively, the system dynamics can be described by a discrete-time Markov process using the transition matrix, $\mathbf{T}(\tau)$, whose entries T_{ij} provide the probability of the system to be found in state j at time $t + \tau$ given that it was in state i at time t (Figure 2b). The corresponding analog to Eq. (1) is

$$\mathbf{p}((k+1)\tau) = \mathbf{p}(k\tau)\boldsymbol{T}(\tau). \tag{2}$$

Eqs. (1) and (2) provide equivalent results at discrete times $t = k\tau$, $k \in \mathbb{N}_0$ and are related by $\mathbf{T}(\tau) = \exp(\tau\mathbf{K})$ [22]. Here, we will concentrate on $\mathbf{T}(\tau)$. Each left eigenvector of $\mathbf{T}$, $\mathbf{q}_i$, describes a particular 'transition mode' between substates, while the corresponding eigenvalue λ_i describes the fraction of molecules (i.e. $\lambda_i \leq 1$) that have not undergone the transition $\mathbf{q}_i$ after time τ (i.e., $\lambda_i \approx 0$ are fast modes, $\lambda_i \approx 1$ are slow modes, see Figure 2c and d). The first mode, $\mathbf{q}_1$, provides the stable equilibrium distribution of the system (Fig 2c, top), that is no transitions, and thus $\lambda_1 = 1$. In the example of Figure 2, the second transition mode, $\mathbf{q}_2$, corresponds to the slow ($\lambda_2 = 0.97$) exchange between basins A and basins B + C, as reflected by the opposite signs of the elements of $\mathbf{q}_2$ in these regions (Figure 2c, middle). The 'implied' timescale of a transition mode is given by

$$\tau_i^* = -\frac{\tau}{\ln \lambda_i} \tag{3}$$

Figure 2



The construction of a transition network. **(a)** Sample potential, defined over a one-dimensional coordinate that is discretized into 100 microstates. It has three metastable basins (A, B, and C). **(b)** Transition matrix $\mathbf{T}(\tau)$ for a Markov lagtime of $\tau = 200$ steps. The transition probability T_{ij} within time τ (blue: $T_{ij} = 0$, red: $T_{ij} \geq 0.1$) was obtained from a Metropolis Monte Carlo ($T = 1/k_B$), jumping each step only to the current or adjacent microstates. **T** exhibits three clusters corresponding to the metastable states. **(c)** Left eigenvectors of **T** indicating the transition modes among microstates. The first eigenvector gives the stationary distribution. The sign structure of the second eigenvector partitions the state space into two metastable states (thick magenta line), A and B + C. The sign structure of the third eigenvector further splits B and C (thin magenta line), obtaining three metastable states. **(d)** The eigenvalue spectrum of **T**. The clear gaps after 2 and 3 eigenvalues indicate how many states are metastable. **(e)** Coordinates of the 100 microstates projected onto the second and third right eigenvectors of **T**. Metastable states are identified by clustering the microstates in this eigenspace (magenta lines). **(f)** Transition network between the 3 metastable states A, B and C (the transition probabilities are obtained from Eq. (5)).

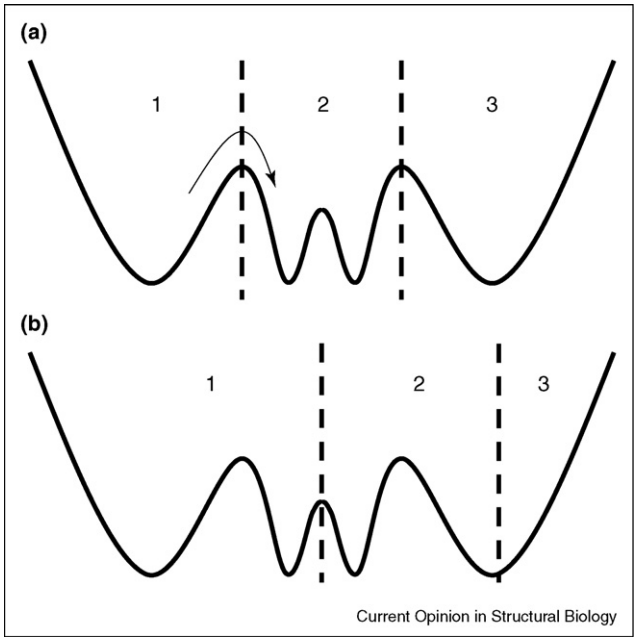
For equilibrium molecular dynamics, $\mathbf{T}(\tau)$ is positive definite ($\lambda_i > 0$ for all i) and there exists a unique equilibrium distribution, π, which fulfills detailed balance, $\pi_i T_{ij} = \pi_j T_{ji}$.

A system without memory is said to be Markovian. This means that, at any time t, the future of the system will depend only on its current state, $\mathbf{p}(t)$, and not on its past history. Besides inertial effects on short timescales, the most frequent cause of nonmarkovianity in macromolecules is the presence of state-internal barriers (see Figure 3). Any model will be Markovian for long enough lagtimes, τ, but in order to achieve a model with a short enough lagtime to be useful, the states should be defined such that large internal barriers are avoided. Thus, the model's substates should be metastable, that is, having minimal intra-state equilibration times and maximal interstate transition times. Most approaches first finely partition the configurational space of the molecule into a large set of states, called here *microstates*, which are then clustered together into fewer metastable states (Figure 2e). Figure 2f shows a coarse-grained transition network between metastable states.

Energy-landscape-based transition networks

In the 'energy landscape' approach, the microstates are often taken as the *attraction basins*, that is the set of configurations that minimize to the same local minimum [23], see Figure 4. Metastable substates are obtained by lumping groups of basins that are separated by low energy barriers [24,25]. The adjacent substates are connected to form a network, and the transition rates in the network are obtained from the energy of the saddle-points along the

Figure 3

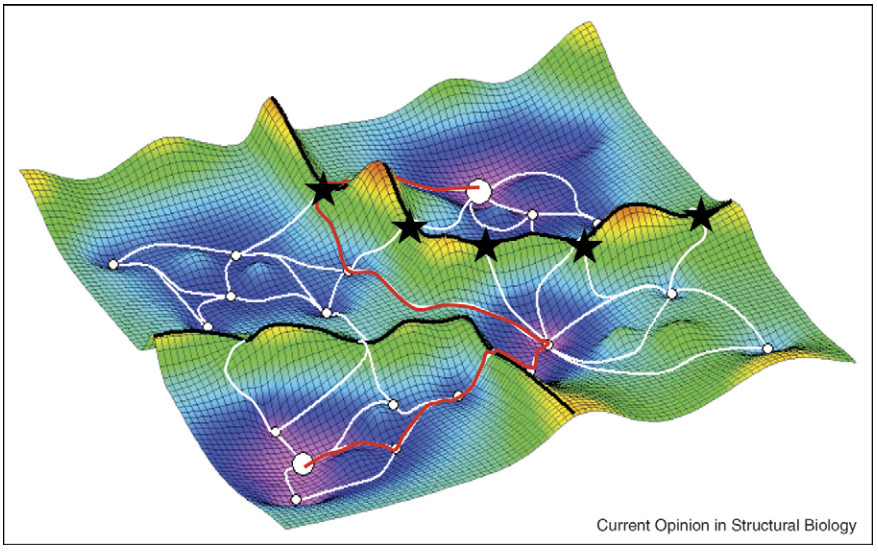


Intra-state barriers increase the Markov lagtime. **(a)** A trajectory passing from state 1 to state 2 at time t is, at a very short time $t + \Delta t$ later, more likely to go back to state 1 than to proceed to state 3 (which involves crossing the internal barrier in state 2). Thus, the system retains memory of its previous history as long as it did not equilibrate within state 2. After a sufficiently long lagtime τ, the system loses the memory from where it came into state 2. **(b)** This state definition has even higher state internal equilibration times, requiring longer lagtimes τ for a Markovian behavior, too long to be of practical use.

connections. The first-order saddle-points can be located directly with CPR [14], or with other path-optimization methods combined with transition state optimizers ([26] and [27•], pp. 284–287). For each connection in the network, the energy barrier of the highest (rate-limiting) saddle point is used in conjunction with a rate-law (usually transition state theory) to determine the transition rate, k_{ij}, from substate i to substate j. In the simplest case, the barrier is assumed to be purely enthalpic, that is, it is given by the potential energy difference between minima and saddle points. Entropic contributions can be included by, for example, using a harmonic approximation around the stationary points to estimate the vibrational free energies [28].

The main advantage of the energy-landscape approach is that it allows to explore transitions which involve high individual energy barriers that could not be crossed by unbiased molecular dynamics simulations. The main challenge of this approach is that the number of stationary points on the energy surface increases exponentially with the system size [29] and that computing first-order saddle points in large molecules is computationally expensive, rendering it infeasible to compute all barriers of the network. Therefore, energy-landscape based transition networks have been used for small systems, such as atom clusters and glasses (see e.g. [30,31]) or peptides [12,32,33]. Somewhat larger systems may be treated by employing methods that restrict the expensive saddle-point computations to a relevant subnetwork, for example by using discrete path sampling [20,34]. Recently, a

Figure 4



Transition network on an energy landscape. The substates are local energy minima (white bullets, the two end-states are shown as large bullets), connected by minimum energy subpaths (white lines) whose energy barriers can be calculated. Two essential properties of a transition network are the best path between the end-states (along which the reactive flux is maximal, shown as red line) and the energy ridge (i.e., the collection of rate-limiting saddle-points separating the end-states, shown as stars).

methodological breakthrough has allowed to comprehensively characterize the transition network of a complex protein transition [3••], the Ras p21 molecular switch. By using an adaptive approach based on graph theory [35], this method computes only those energy barriers which contribute to the global network properties of interest, such as the best path(s) connecting or the energy ridge(s) separating two transition end-states (Figure 4). Since this method reduces the number of energy barriers required by several orders of magnitude, it allows transition networks also to be constructed for complex transitions in large molecules [3••,35]. For example in the case of Ras p21, the hydrolysis of bound GTP induces a rearrangement of the fold of two regions (Switches I and II). It is conceivable that such a transition involves many pathways *via* various partially folded Switch regions. However, the computed TN showed that all energetically feasible pathways include a coupling between the motions of Switches I and II, inducing a preferred order of events. The suggested mechanism can be experimentally validated, for example, via point mutations in the key residues.

These works having spawned an energy-landscape theory [27•,36], which allow to interpret the behavior of molecular systems in terms of the features of the underlying energy landscape. In practice, the size of the molecules that can be meaningfully described by the energy-landscape approach is still limited because the fluctuations of the potential energy increases with the number of degrees of freedom in the system, thus giving rise to unrealistic energy differences in large systems. For this reason, explicit solvation is usually avoided and instead, implicit solvent methods (e.g. Generalized-Born) are used. Moreover, to compute kinetic properties with the energy-landscape approach, the transition rates between substates need to be modeled by some rate theory whose validity is often unclear. Finally, since the actual system dynamics is not available, the validity of the Markov property cannot be verified.

Dynamics-based transition networks

As a result of increased computational power and methodological advances, the construction of kinetic multi-state models directly from molecular dynamics (MD) data, often called Markov State Models (MSM), is becoming increasingly popular [37••–39–40]. In contrast to the energy-landscape approach, MSMs require no rate theory — rates are directly obtained from the observed dynamical transitions. Furthermore, it can be tested explicitly whether the MSM exhibits the Markov property and is consistent with the MD simulations. In contrast to extracting slowly converging properties from lengthy dynamical trajectories, MSMs can be constructed from simulations that need only to be long enough to be in local equilibrium: they must be long enough to equilibrate within individual substates and occasionally undergo transition to neighboring substates. Thus, processes that occur on very long timescales (such as milliseconds) can be correctly modeled using many short trajectories started from different conformations [42,43], which was already exploited in massively parallel simulation of peptide folding [43]. The main disadvantage of the dynamical approach is that it will fail if, for the process of interest to occur, individual barriers between microstates must be crossed that are too high to be sampled within the individual simulation times.

The microstates can be defined by geometrical proximity [37••,38••,41]. It is important to choose microstates such that they do not merge kinetically separated regions of state space, as this would prevent the model to be Markovian (see Figure 3 and [37••,42]). After assigning each structure along a given MD trajectory to one of the microstates, the transition matrix, $\mathbf{T}(\tau)$, is computed for each pair of microstates (i, j), as

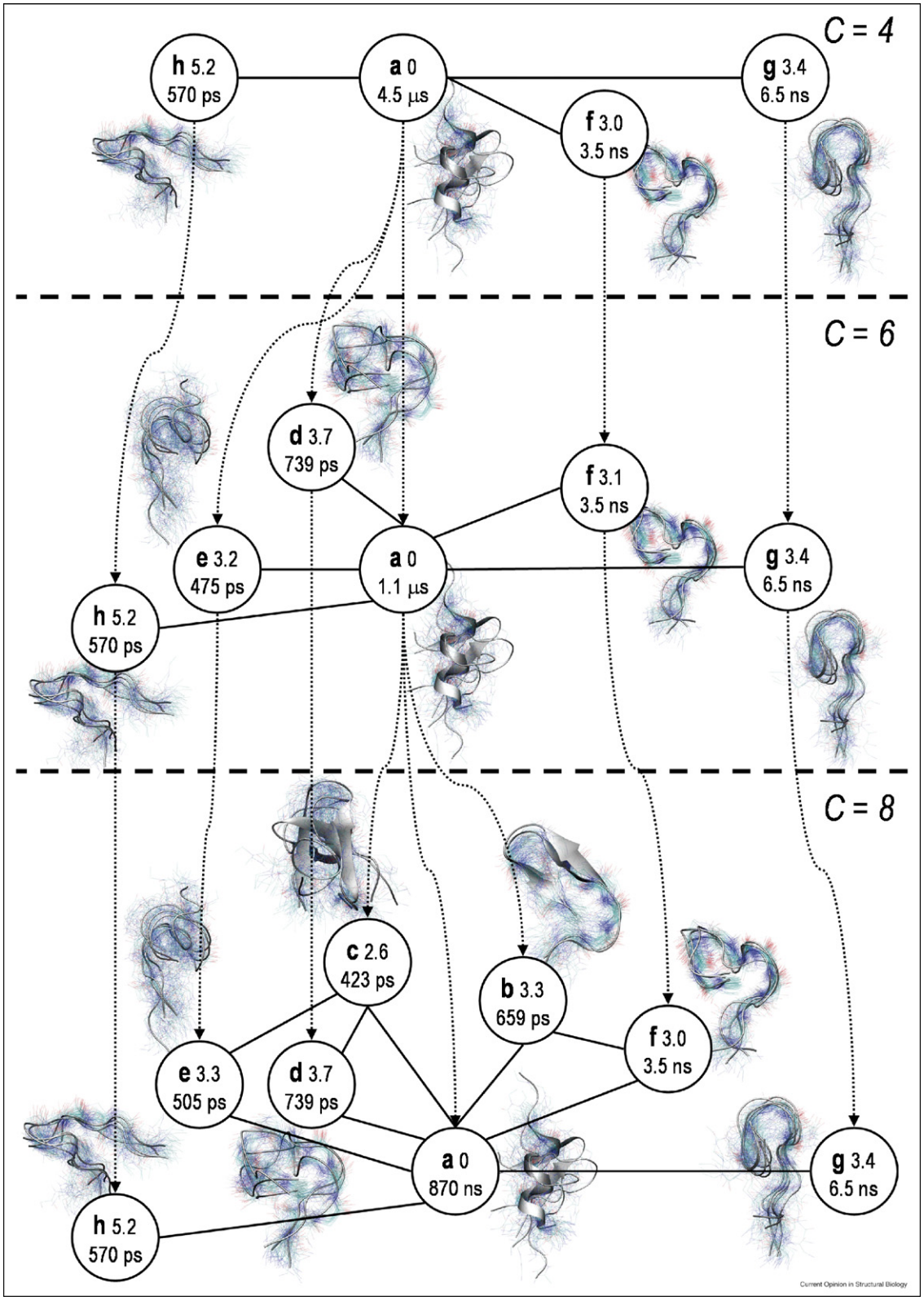
$$T_{ij} = \frac{\text{number of transitions } i \to j \text{ in time } \tau}{\text{number of starts in } i} \tag{4}$$

To maximize the time resolution of the model, the lagtime τ is chosen to be the minimal lagtime needed for the model to remain Markovian. This can be determined by exploiting the fact that, if the dynamics is Markovian at lagtime τ, it will also be Markovian at larger lagtimes $\tau' > \tau$, and any kinetic properties computed from the model should then be converged in τ. The method used in [37••,43,44] computes, for different lagtimes τ, $\mathbf{T}(\tau)$ and its set of implied timescales from Eq. (3). At lagtimes greater than or equal to the Markov lagtime, the implied timescales have converged.

A meaningful way of clustering the microstates into a set of C metastable states is to require that the transitions between microstates within each metastable state are much faster than the transitions between metastable states. This can be achieved by using an iterative splitting-and-lumping procedure [38•]. A similar partition can be obtained efficiently with the improved Perron cluster cluster analysis (PCCA) method [45,46], which exploits the fact that kinetically closely connected microstates have similar coordinates in the first C right eigenvectors (see Figure 2e). Note that simple geometric clustering is often unable to identify metastable states since large free energy barriers may lie between geometrically close conformations [37••]. A practical way to choose the number of clusters, C, is to define a timescale τ^*_{min} of interest. C is chosen equal to the number of implied timescales greater than τ^*_{min}. Note that C is a user-defined parameter and its choice is a compromise between improving conformational resolution and reducing statistical error. If there is a large gap between the C th and the $(C+1)$ th timescales, the total transition probability between metastable states I and J is approximately:

$$\bar{T}_{IJ}(\tau) \approx \frac{\sum_{i \in I, j \in J} \pi_i T_{ij}(\tau)}{\sum_{i \in I} \pi_i}. \tag{5}$$

Figure 5



Hierarchical transition network for the Ala_{12} peptide, computed for either 4, 6, and 8 metastable sets. Each circle corresponding to one metastable state contains its free energy relative to the most populated state, **a** (in kcal/mol, top), and its lifetime (bottom) at 300 K. Solid lines represent transitions that have occurred within a 4 μ s MD simulation. Dotted lines relate corresponding metastable states.

The resulting coarse-grained transition network captures the essential features of the transition process, as shown for example in Figure 2f.

The Markovianity of MSMs is often tested directly *via* convergence of the implied timescale test (see above) but this method is unreliable when statistics are poor. Other Markov tests have been proposed [44,47], but many of them suffer from being too tolerant, or from being ambiguous (John D Chodera, Stanford, personal communication). Ultimately, the best test for the model is to compare the molecular dynamics simulations to the predictions of Eq. (2)[37••,48]. A systematic test for Markovianity based on this criterion is still elusive.

So far, MSMs have been used mainly to model the dynamics of small polypeptides for which sufficient sampling could be achieved [43,40], some also testing the validity of the Markov assumption and checking for consistency with the dynamics [37••,38••,48]. An example application on Ala_{12} is shown in Figure 5.

Interpretation of transition networks

In principle, any property that can be calculated directly from simulation data, can also be obtained from the transition network. For example, the equilibrium distribution is obtained from the elements in the first left eigenvector of $\mathbf{T}(\tau)$ or the rate matrix $\mathbf{K}$, scaled such that they sum to 1 (see Figure 2c, top). This is how the relative free energies of the Ala_{12} macrostates in Figure 5 were calculated.

Another property of interest is the mean first passage time (mfpt), defined as the mean time f_i it takes to reach a given metastable state m for the first time when starting from another state i. All mfpt's (f_i, $i = 1, \ldots, m-1$) can be computed simultaneously from the transition matrix by solving the following linear system of equations [49•]:

$$\begin{bmatrix} T_{11}-1 & & \cdots & & T_{1m} \\ & T_{22}-1 & & & T_{2m} \\ \vdots & & \ddots & & \vdots \\ T_{m1} & \cdots & & T_{m-1,m-1}-1 & T_{m-1,m} \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \times \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{m-1} \\ f_m \end{bmatrix} = \begin{bmatrix} -\tau \\ -\tau \\ \vdots \\ -\tau \\ 0 \end{bmatrix}$$

A useful property is the committor probability c_i, defined as the probability to transition from a metastable state i to set B without hitting set A first, where A and B are given sets of metastable states. For example in protein folding, A and B can be chosen as the set of fully unfolded and folded conformers, respectively (c_i is then often referred to as p_{fold}, the probability of folding for conformer i). The set of states with $c_i \approx 0.5$ (i.e., having equal chance to fold or unfold) can be used as a definition for the transition state ensemble [50]. The committor can be calculated directly from the transition matrix by solving the following system of linear equations:

$$\begin{array}{ll} \sum_j (T_{ij} - \delta_{ij}) c_i = 0, & \forall\ i \notin (A \cup B) \\ c_i = 0, & \forall\ i \in A \\ c_i = 1, & \forall\ i \in B \end{array} , \qquad (6)$$

with $\delta_{ij} = 1$ for $i = j$ and 0 otherwise.

Finally, a statistically rigorous approach to computing the full set of individual transition pathways $A \to B$ along with their relative contributions to the overall $A \to B$ rate is given by the Transition Path Theory (P Metzner, C Schütte, E Vanden-Eijnden, Transition path theory for Markov jump processes, Annu Appl Prob, unpublished data). First, the flux from state i to state j that effectively contributes to the transition $A \to B$ is given by:

$$f_{ij} = \pi_i (1 - c_i) k_{ij} c_j,$$

where π_i is the probability to start in i, $(1 - c_i)$ is the probability to come from A rather than from B, k_{ij} is the transition rate from i to j, and c_j is the probability to transition to B rather than to A. The net flux is obtained from:

$$\bar{f_{ij}} = f_{ij} - f_{ji},$$

and provides a network of directed net fluxes. This network can be decomposed into individual $A \to B$ pathways, which can be sorted according to their contribution to the overall $A \to B$ flux, thus providing something similar to the best path, the next best path, etc.

The sampling problem

Even with the help of massively parallel computing platforms, such as folding@home, it is currently challenging to run enough MD sampling such as to obtain a well-converged kinetic model of complex conformational change in a protein. However, this may be achieved by using an adaptive approach, such that the sampling is limited to the statistically most rewarding parts of state space [3••,35,49•,51].

Since the transition matrix $\mathbf{T}(\tau)$ is estimated from finite MD trajectories, any property calculated from $\mathbf{T}(\tau)$ will have a degree of uncertainty. The first step is to estimate this uncertainty. Assuming an *a priori* uniform distribution of matrices, the likelihood that all the transitions

observed during the MD runs are consistent with a particular matrix $\mathbf{T}(\tau)$ is proportional to

$$p[\mathbf{T}(\tau)] \propto \prod_{i,j} T_{ij}^{c_{ij}}, \quad (7)$$

where c_{ij} denotes the actual number of transition events observed from state i to state j. The matrix that maximizes this likelihood turns out to be the transition matrix resulting from Eq. (4). The width of this likelihood density measures the statistical uncertainty of the matrix entries T_{ij} [49•]. Since generally one is more interested in the distribution (and corresponding uncertainty) of some target property $A(\mathbf{T}(\tau))$ that is computed from the transition matrix, this distribution can be computed by first generating transition matrices according to the density of Eq. (7), and then computing $A(\mathbf{T}(\tau))$ for each of the generated $\mathbf{T}(\tau)$ [49•]. Fast but approximate analytical methods have also been proposed [49•,51]. The development of efficient error analysis methods which evaluate the distribution of transition matrices that obey particular conditions (such as detailed balance or positive definiteness) is in progress.

Once the statistical uncertainty of A has been determined, the next step is to chose in which state i the next MD simulation should be started in order to give the largest decrease of this statistical uncertainty. This approach was considered in [49•] and it was shown that with adaptive sampling a given uncertainty could be achieved within a fraction of the computational effort compared to non-adaptive sampling.

Conclusion

Modeling the kinetics of macromolecules based on transition networks is a promising concept in current computational structural biology. This concept has been spawned by the energy-landscape approach, whose main limitation is that kinetics are only secondary to the model and must be recovered *via* a rate theory. Markov state models are computationally more demanding, but can be tested and analyzed using standard tools from statistics and algebra. Major open problems in the field include: First, the development of a reliable and unsupervised test for Markovianity; second, the development of efficient methods to generate transition matrices that exhibit a number of properties such as detailed balance or positive definiteness; thus allowing for a better estimation of the statistical uncertainty; third, the development of a numerically stable and algorithmically efficient adaptive sampling strategy, which includes an adaptive re-definition of microstates and metastable states, thus alleviating the sampling problem. Upon solving these problems, transition networks are set to become the standard tool for studying and understanding complex transitions in macromolecules.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Ostermann A, Waschipky R, Parak FG, Nienhaus GU: **Ligand binding and conformational motions in myoglobin**. *Nature* 2000, **404**:205-208.

2. Fischer S, Windshuegel B, Horak D, Holmes KC, Smith JC: **Structural mechanism of the recovery stroke in the myosin molecular motor**. *Proc Natl Acad Sci U S A* 2005, **102**:6873-6878.

3. •• Noé F, Krachtus D, Smith JC, Fischer S: **Transition networks for the comprehensive characterization of complex conformational change in proteins**. *J Chem Theory Comput* 2006, **2**:840-857.
This study presents adaptive methods on how to compute only a small part of a transition network while still finding the globally best path and the global energy ridge for a given transition of interest. These methods have opened the way for the application of transition networks to the comprehensive characterization of complex change in proteins.

4. Jäger M, Zhang Y, Bieschke J, Nguyen H, Dendle M, Bowman ME, Noel J, Gruebele M, Kelly J: **Structure-function-folding relationship in a ww domain**. *Proc Natl Acad Sci U S A* 2006, **103**:10648-10653.

5. Kobitski AY, Nierth A, Helm M, Jäschke A, Nienhaus GU: **Mg^{2+}-dependent folding of a Diels–Alderase ribozyme probed by single-molecule FRET analysis**. *Nucleic Acids Res* 2007, **35**:2047-2059.

6. Schlitter J, Engels M, Krüger P: **Targeted molecular dynamics: A new approach for searching pathways of conformational transitions**. *J Mol Graphics* 1994, **12**:84-89.

7. Laio A, Parrinello M: **Escaping free energy minima**. *Proc Natl Acad Sci USA* 2002, **99**:12562.

8. Shean J-E, Brooks CL III: **From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding**. *Annu Rev Phys Chem* 2001, **52**:499-535.

9. Lu H, Isralewitz B, Krammer A, Vogel V, Schulten K: **Unfolding of titin immunoglobulin domains by steered molecular dynamics**. *Biophys J* 1998, **75**:662-671.

10. Gräter F, Shen J, Jiang H, Gautel M, Grubmüller H: **Mechanically induced titin kinase activation studied by force-probe molecular dynamics simulations**. *Biophys J* 2005, **88**: 790-804.

11. Koppole S, Smith JC, Fischer S: **The Structural Coupling between ATPase Activation and Recovery Stroke in the Myosin II Motor**. *Structure* 2007, **15**:825-837.

12. Krivov SV, Karplus M: **Hidden complexity of free energy surfaces for peptide (protein) folding**. *Proc Natl Acad Sci U S A* 2004, **101**:14766-14770.

13. • Muff S, Caflisch A: **Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a -sheet miniprotein**. *Proteins* 2007.
This study demonstrates how state-based kinetic models may reveal features that would be hidden in projections on few reaction coordinates.

14. Fischer S, Karplus M: **Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom**. *Chem Phys Lett* 1992, **194**:252-261.

15. Jónsson H, Mills G, Jacobsen KW: **Classical and Quantum Dynamics in Condensed Phase Simulations, ch. Nudged Elastic Band Method for Finding Minimum Energy Paths of Transitions**. World Scientific; 1998:385-404.

16. Fischer S, Michnick S, Karplus M: **A mechanism for rotamase catalysis by the fk506 binding protein (fkbp)**. *Biochemistry* 1993, **32**:13830-13837.

17. Noé F, Ille F, Smith JC, Fischer S: **Automated computation of low-energy pathways for complex rearrangements in proteins: application to the conformational switch of ras p21**. *Proteins* 2005, **59**:534-544.

18. Cui Q, Karplus M: **Triosephosphate isomerase: A theoretical comparison of alternative pathways**. *J Am Chem Soc* 2007, **123**:2284-2290.

19. Dellago C, Bolhuis P, Chandler D: **Efficient transition path sampling: application to Lennard–Jones cluster rearrangements**. *J Chem Phys* 1998, **108**:9236-9245.

20. Wales DJ: **Discrete path sampling**. *Mol Phys* 2002, **100**:3285-3305.

21. Frauenfelder H, Sligar SG, Wolynes PG: **The energy landscapes and motions of proteins**. *Science* 1991, **254**:1598-1603.

22. van Kampen NG: *Stochastic Processes in Physics and Chemistry*. 4th ed.. Amsterdam: Elsevier; 2006.

23. Stillinger FH, Weber TA: **Packing structures and transitions in liquids and solids**. *Science* 1984, **228**:983-989.

24. Becker OM: **Geometric versus topological clustering: An insight into conformation mapping**. *Proteins* 1997, **27**:213-226.

25. Evans DA, Wales DJ: **The free energy landscape and dynamics of met-enkephalin**. *J Chem Phys* 2003, **119**:9947-9955.

26. Henkelman G, Jóhannesson G, Jónsson H: **Progress on Theoretical Chemistry and Physics, ch. Methods for Finding Saddle Points and Minimum Energy Paths**. Kluwer Academic Publishers; 2000:269-300.

27. • Wales D: *Energy Landscapes*. Cambridge: Cambridge University Press; 2003.
The most comprehensive review of the energy landscape approach.

28. Wales DJ: **Energy landscapes: calculating pathways and rates**. *Int Rev Phys Chem* 2006, **25**:237-282.

29. Stillinger FA: **Exponential multiplicity of inherent structures**. *Phys Rev E* 1999, **59**:48.

30. Wales DJ: **Structure, dynamics, and thermodynamics of clusters: tales from topographic potential surfaces**. *Science* 1996, **271**:925-933.

31. Calvo F, Bogdan TV, de Souza VK, Wales DJ: **Equilibrium density of states and thermodynamics properties of a model glass former**. *J Chem Phys* 2007, **127**:044508.

32. Becker OM, Karplus M: **The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics**. *Journal of Chemical Physics* 1996, **106**:1495-1517.

33. Levy Y, Becker OM: **Effect of conformational constraints on the topography of complex potential energy surfaces**. *Phys Rev Lett* 1998, **81**:1126-1132.

34. Evans DA, Wales DJ: **Folding of the GB1 hairpin peptide from discrete path sampling**. *J Chem Phys* 2004, **121**:1080-1090.

35. Noé F, Oswald M, Reinelt G, Fischer S, Smith JC: **Computing best transition pathways in high-dimensional dynamical systems: application to the $\alpha_L \rightleftharpoons \beta \rightleftharpoons \alpha_R$ transitions in octaalanine**. *Multiscale Model Sim* 2006, **5**:393-419.

36. Brooks CL III, Onuchic JN, Wales DJ: **Taking a walk on a landscape**. *Science* 2001, **293**:612-613.

37. •• Noé F, Horenko I, Schütte C, Smith JC: **Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states**. *J Chem Phys* 2007, **126**:155102.
Analysis how to obtain and test kinetic models that hierarchically decompose state space using molecular dynamics of polyalanines. The PCCA method, the usefulness of metastable states versus geometric clustering and possible sources for nonmarkovianity are illustrated.

38. •• Chodera JD, Dill KA, Singhal N, Pande VS, Swope WC, Pitera JW: **Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics**. *J Chem Phys* 2007, **126**:155101.
This study presents an adaptive method for identification of metastable states from simulation data. The method is applied to three examples from massively parallel simulations.

39. Horenko I, Dittmer E, Fischer A, Schütte C: **Automated model reduction for complex systems exhibiting metastability**. *Multiscale Model Sim* 2006, **5**:802-827.

40. Schultheis V, Hirschberger T, Carstens H, Tavan P: **Extracting Markov models of peptide conformational dynamics from simulation data**. *J Chem Theory Comp* 2005, **1**:515-526.

41. Rao F, Caflisch A: **The protein folding network**. *J Mol Biol* 2004, **342**:299-306.

42. Swope WC, Pitera JW, Suits F, Pitman M, Eleftheriou M: **Describing protein folding kinetics by molecular dynamics simulations. 2. Example applications to alanine dipeptide and beta-hairpin peptide**. *J Phys Chem B* 2004, **108**: 6582-6594.

43. Elmer SP, Park S, Pande VS: **Foldamer dynamics expressed via Markov state models. i. explicit solvent molecular-dynamics simulations in acetonitrile, chloroform, methanol, and water**. *J Chem Phys* 2005, **123**:114902.

44. Swope WC, Pitera JW, Suits F: **Describing protein folding kinetics by molecular dynamics simulations. 1. Theory**. *J Phys Chem B* 2004, **108**:6571-6581.

45. Schütte C, Fischer A, Huisinga W, Deuflhard P: **A direct approach to conformational dynamics based on hybrid Monte Carlo**. *J Comp Phys* 1999, **151**:146-168.

46. Deuflhard P, Weber M: Robust perron cluster analysis in conformation dynamics, ZIB Report, vol. 03–09, 2003.

47. Park S, Pande VS: **Validation of Markov state models using Shannon's entropy**. *J Chem Phys* 2006, **124**:054118.

48. Chodera JD, Swope WC, Pitera JW, Dill KA: **Long-time protein folding dynamics from short-time molecular dynamics simulations**. *Multiscale Model Sim* 2006, **5**:1214-1226.

49. • Singhal N, Pande VS: **Error analysis and efficient sampling in Markovian state models for molecular dynamics**. *J Chem Phys* 2005, **123**:204909.
This study provides an approach to estimate the errors of a Markov State Model and shows that such an error analysis can be used in an adaptive sampling method.

50. Lenz P, Zagrovic B, Shapiro J, Pande VS: **Folding probabilities: A novel approach to folding transitions and the two-dimensionalising-model**. *J Chem Phys* 2004, **120**:6769-6778.

51. Hinrichs NS, Pande VS: **Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics**. *J Chem Phys* 2007, **126**:244101.