

SEPARATING THE WHEAT FROM THE CHAFF

Tips on how to identify and characterize essential movements in frantically shaking proteins

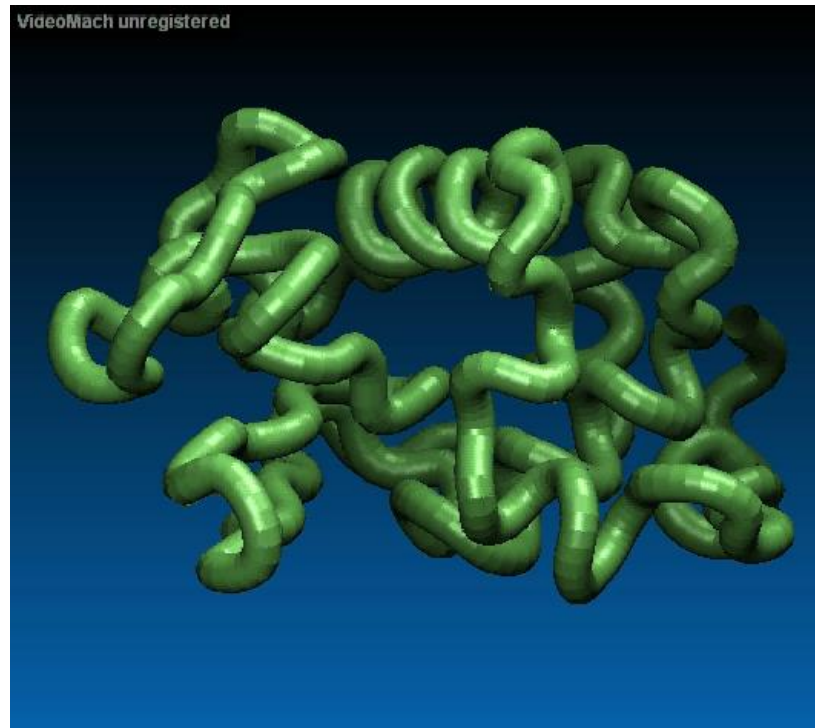
Why do we do MD?

- Originally: to collect data for statistical mechanics
 - Based on the ergodic hypothesis.
 - Calculate energies, free energies, diffusion coefficients, etc.
- To see the movements of macromolecules
 - The problem: “Imagine living in a world where a Richter 9 earthquake raged continuously...at the scale of proteins Brownian motions are even more furious than that.”
(G. Oster and H. Wang, Molecular motors, Chapter 8. DOI: 10.1002/3527601503.ch8)



Why is that a problem?

- Interesting movements, relevant for protein functioning, are mixed with the noisy irrelevant movements



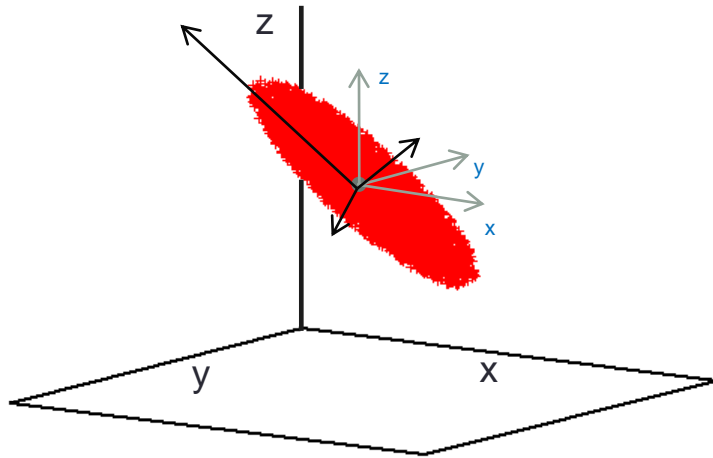
Principal component analysis

- Procedure taken from multivariate statistical analysis.
- Introduced in MD by Karplus and Berendsen.
- Aims to identify a reduced set of coordinates able to describe the relevant movements.
- Does it (always) fulfil its aim?
- Can we improve it?

Outlook of the presentation

- PCA:
 - Fundamentals.
 - Utility / Limitations.
- Consistent PCA.
- Concatenated PCA.
- PCA of inter/intra subunit movements.
 - P2X4 as example.
- Conclusions.

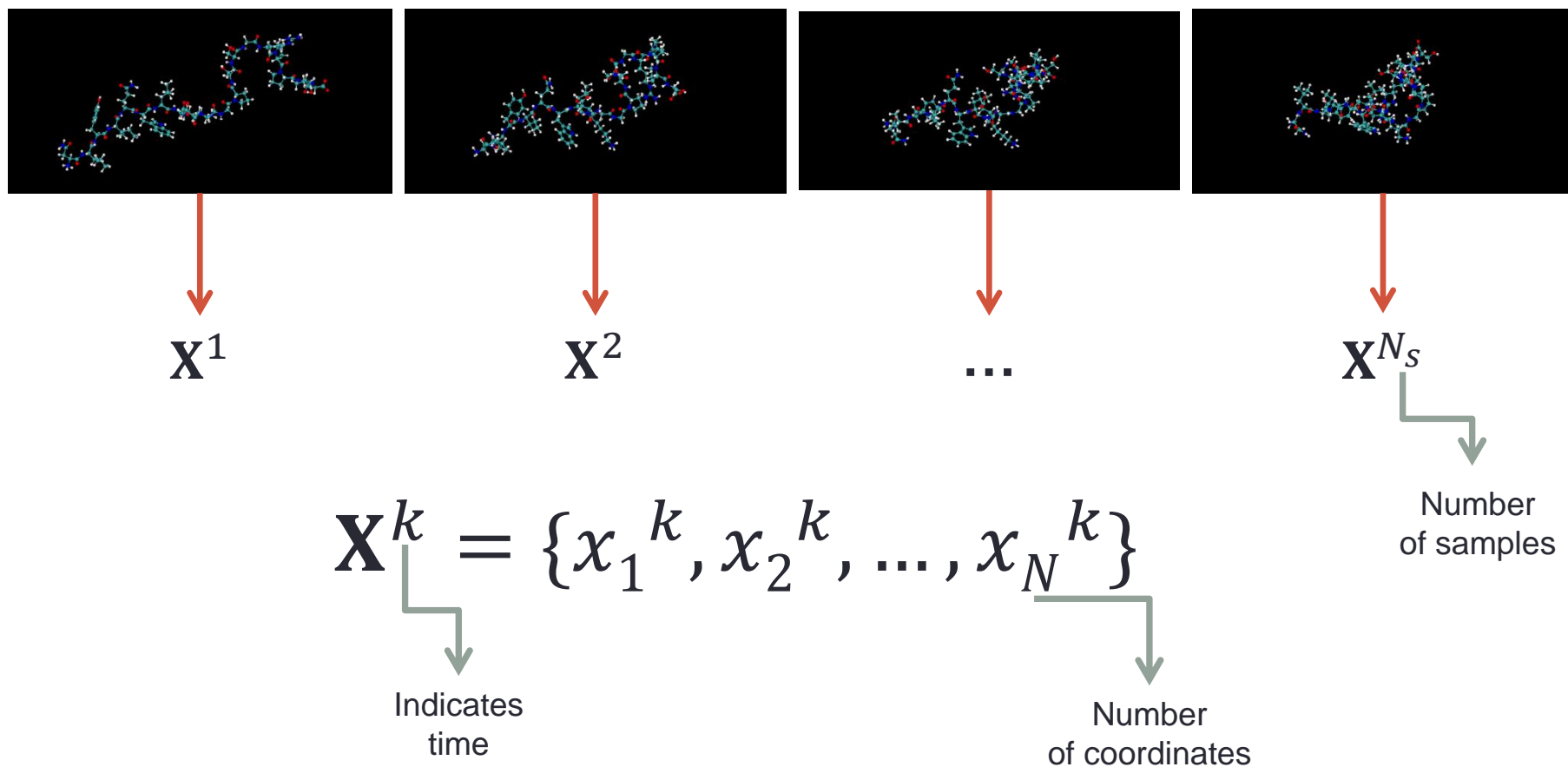
What does PCA do? (basically)



- Transform local coordinates to collective coordinates.
- Just a few collective coordinates explain most of protein fluctuations.
- Allows a reduction of the dimensionality.

How does it do that?

- Collect coordinates from a MD



How does it do that

- Compute the correlation matrix

(covariance matrix too)

$$\mathbf{C} = \begin{pmatrix} C_{11} & \cdots & C_{1N} \\ \vdots & \ddots & \vdots \\ C_{N1} & \cdots & C_{NN} \end{pmatrix} \quad C_{ij} = \frac{1}{N} \sum_{k=1}^{N_s} (x_i^k - \bar{x}_i) \cdot (x_j^k - \bar{x}_j)$$

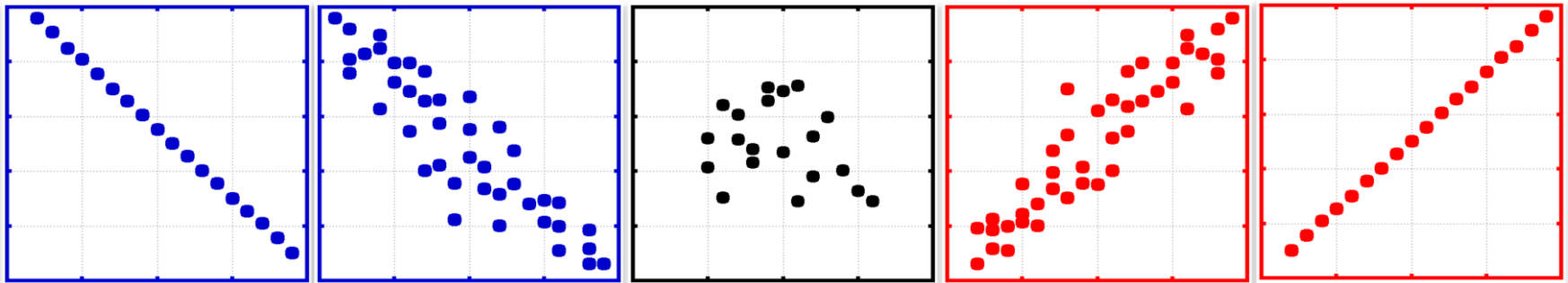
Linear dependence

Anti-correlated

Uncorrelated

Correlated

Linear dependence



$$C_{ij} = -1$$

$$-1 \leq C_{ij} \leq -0.7$$

$$|C_{ij}| \approx 0$$

$$0.7 \leq C_{ij} \leq 1$$

$$C_{ij} = 1$$

How does it do that?

- Diagonalize the correlation matrix

$$\mathbf{R}^T \mathbf{C} \mathbf{R} = \Lambda$$

Diagonal matrix

$$R = \begin{pmatrix} R_{11} & \cdots & R_{1N} \\ \vdots & \ddots & \vdots \\ R_{N1} & \cdots & R_{NN} \end{pmatrix}$$

\mathbf{V}_1 \mathbf{V}_N

$$\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_N \end{pmatrix}$$

Eigenvalue of \mathbf{V}_1 Eigenvalue of \mathbf{V}_N

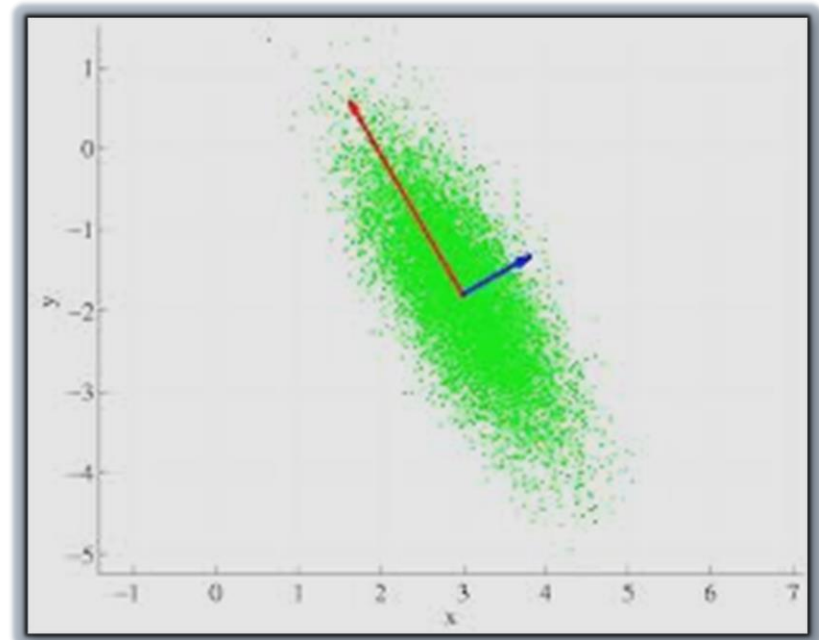
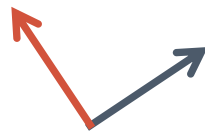
Eigenvectors of matrix \mathbf{C} \Rightarrow Orthonormal \Rightarrow Constitute a basis set

Example in 2D

$$\mathbf{C} = \begin{pmatrix} \frac{1}{N_s} \sum_{k=1}^{N_s} (x^k - \bar{x})^2 & \frac{1}{N_s} \sum_{k=1}^{N_s} (x^k - \bar{x})(y^k - \bar{y}) \\ \frac{1}{N_s} \sum_{k=1}^{N_s} (x^k - \bar{x})(y^k - \bar{y}) & \frac{1}{N_s} \sum_{k=1}^{N_s} (y^k - \bar{y})^2 \end{pmatrix}$$

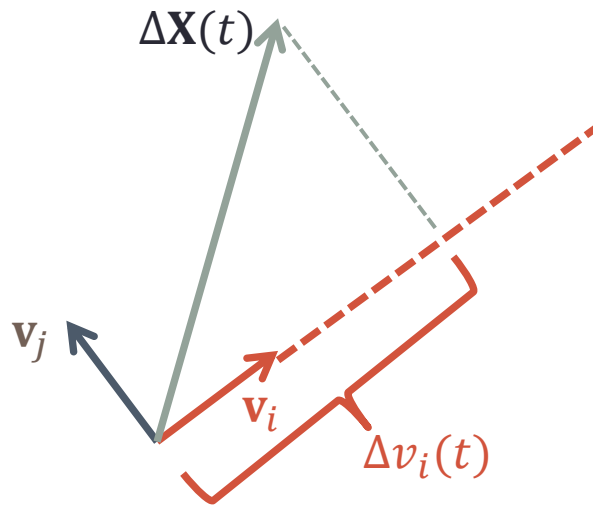
$$\mathbf{V}_1 = \begin{pmatrix} R_{11} \\ R_{21} \end{pmatrix}$$

$$\mathbf{V}_2 = \begin{pmatrix} R_{21} \\ R_{22} \end{pmatrix}$$



Meaning of eigenvalues and eigenvectors

- The i -eigenvalue measures the squared displacement on the direction of eigenvector \mathbf{v}_i



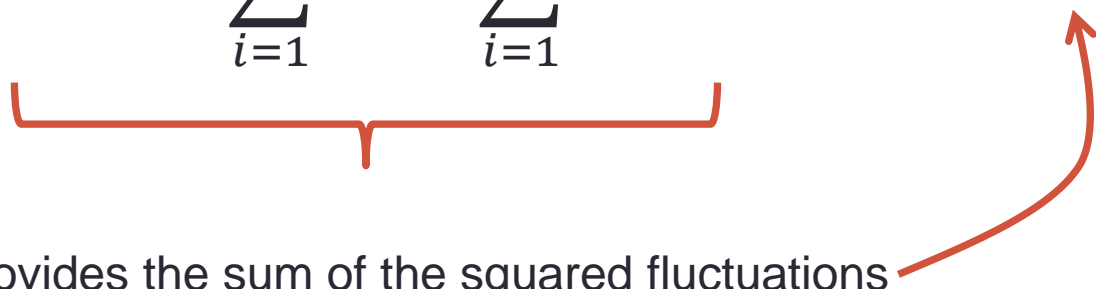
$$\Delta v_i(t_k) = \mathbf{v}_i \cdot \Delta\mathbf{X}(t_k)$$

$$\lambda_i = \frac{1}{N_s} \sum_{k=1}^{N_s} (\Delta v_i(t_k))^2$$

The importance of the eigenvalues

$$\mathbf{C} = \begin{pmatrix} C_{11} & \cdots & C_{1N} \\ \vdots & \ddots & \vdots \\ C_{N1} & \cdots & C_{NN} \end{pmatrix}$$

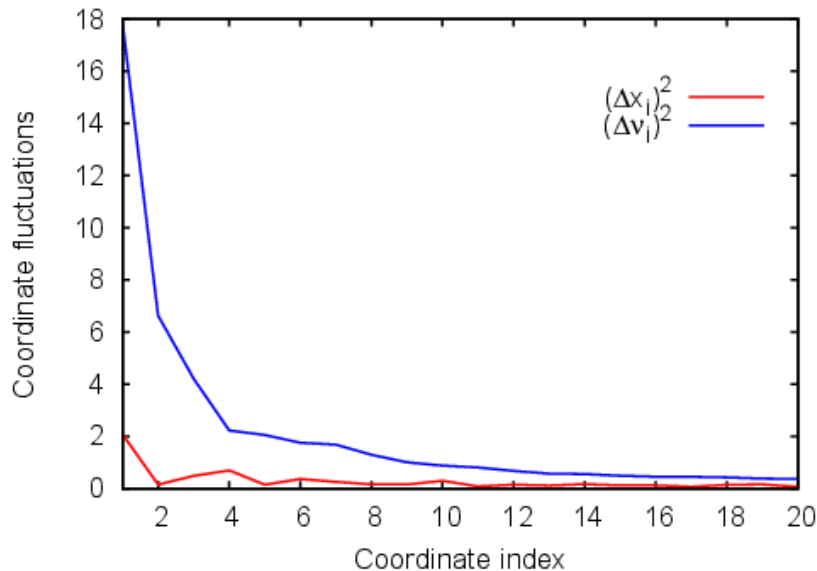
$$\Delta = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_N \end{pmatrix}$$

$$\text{Tr}[\mathbf{C}] = \sum_{i=1}^N C_{ii} = \sum_{i=1}^N (\Delta x_i)^2 \quad \equiv \quad \text{Tr}[\Delta] = \sum_{i=1}^N \lambda_i = \sum_{i=1}^N (\Delta v_i)^2$$


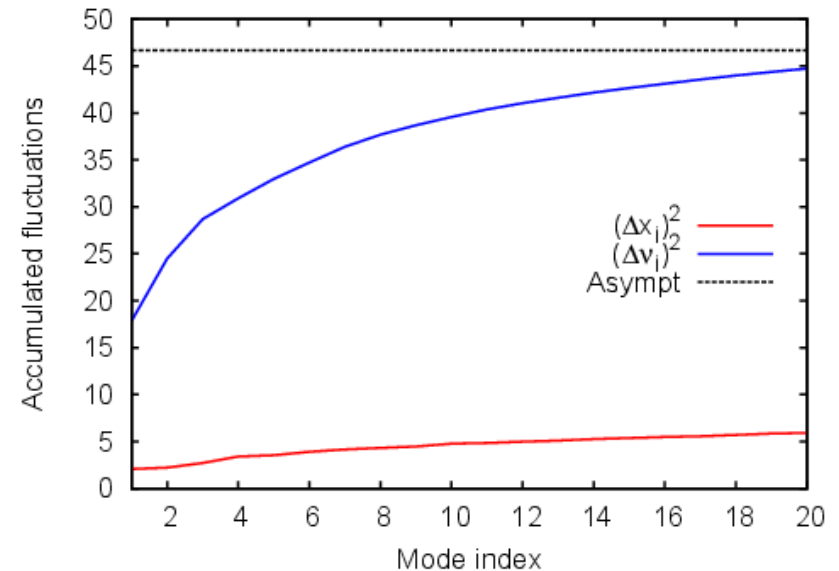
Provides the sum of the squared fluctuations

Cartesian coordinates vs. Principal components

Individual squared fluctuations



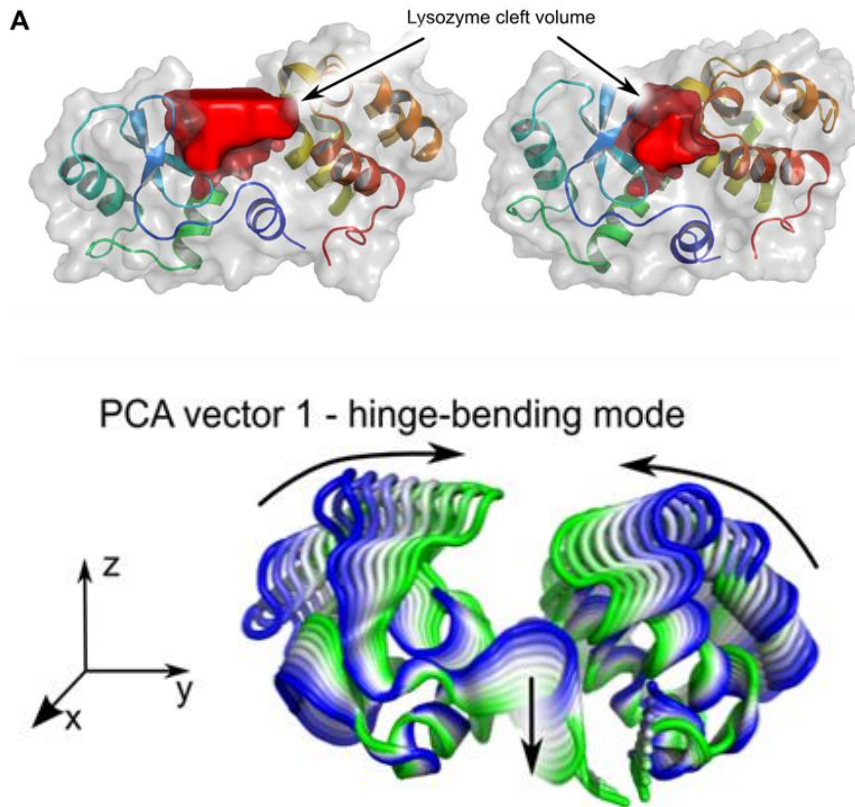
Accumulated squared fluctuations



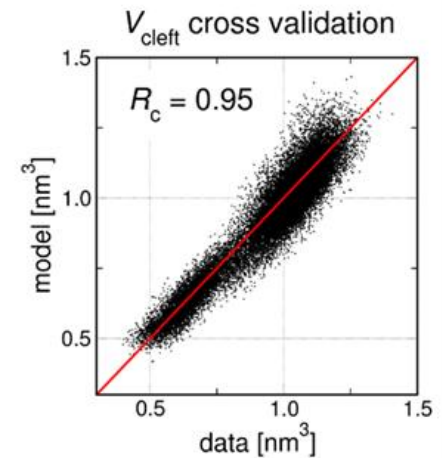
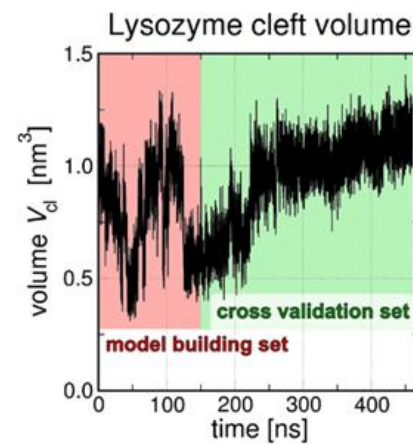
- Total fluctuations are concentrated in a few PC-modes (< 20).
- Total fluctuations are equally distributed among all Cartesian coordinates (714).

Vectors of the essential space are able to describe important movements

- There are plenty of examples.

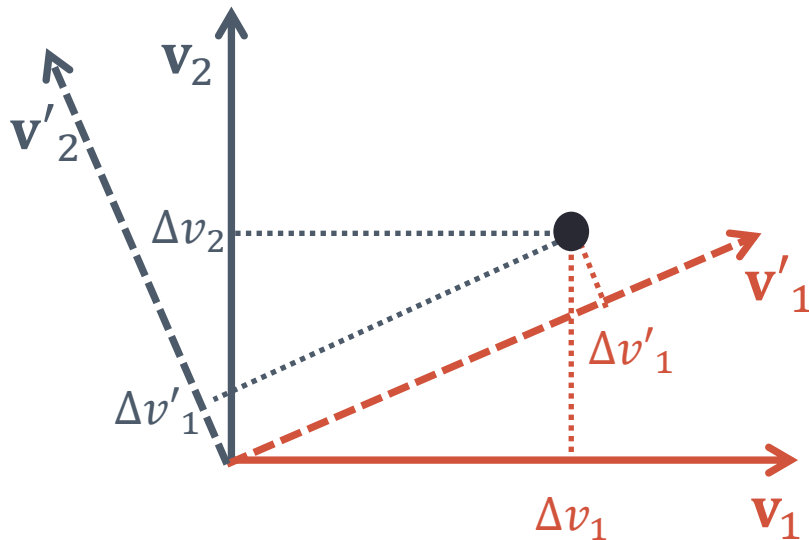


J. S. Hub and B. L. de Groot,
 Plos Comput. Biol. 5(8): e10004802009.

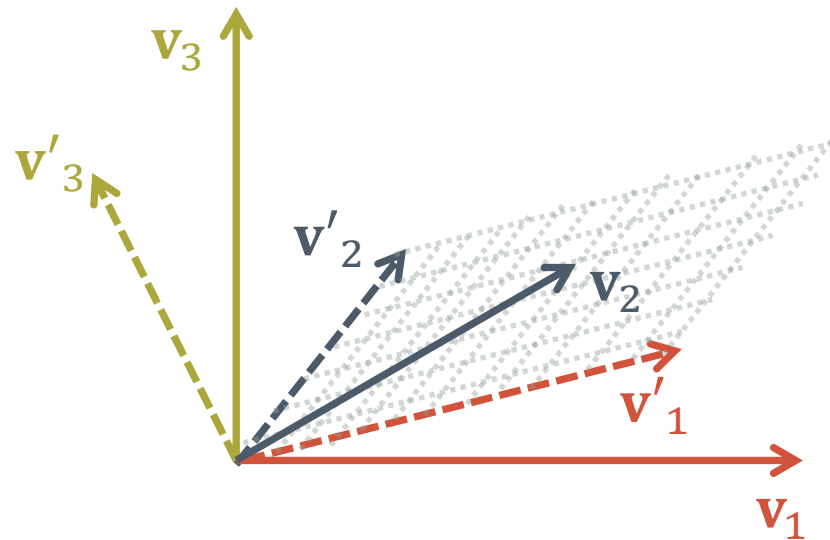


The essential space (subspace)

- Contains the most important eigenvectors
 - How many are truly “essential”?
 - The problem with defining a subspace.



$\{\Delta v_1, \Delta v_2\}$ and $\{\Delta v'_1, \Delta v'_2\}$ span the same subspace

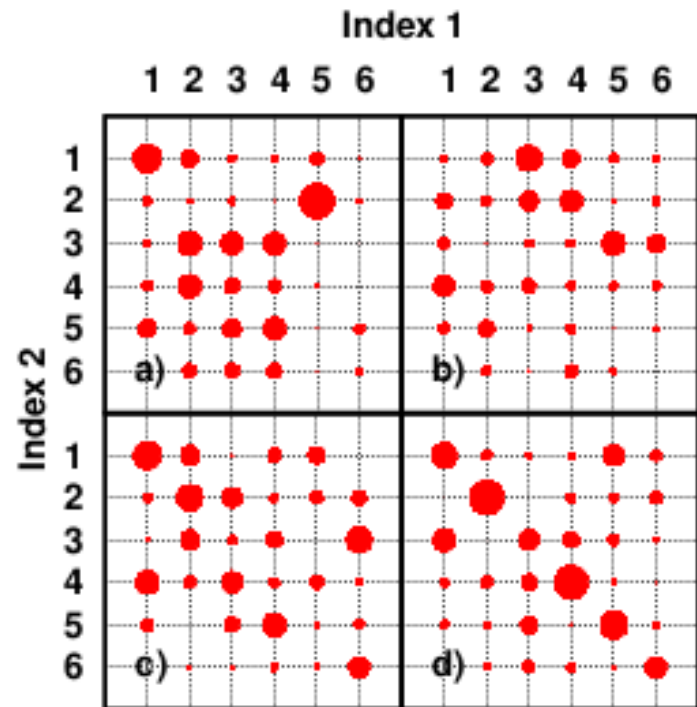


$\{\Delta v_1, \Delta v_2\}$ and $\{\Delta v'_1, \Delta v'_2\}$ do not span the same subspace

Are reproducible the main PC-modes?

- Run equivalent trajectories.
- Compute the PC-modes for each of them.
- Compute the scalar product for the PC-modes of 2 alternative runs.

$$\mathbf{V}_i \cdot \mathbf{V}'_j = \left. \begin{array}{l} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{array} \right\} \text{ Ideally!}$$



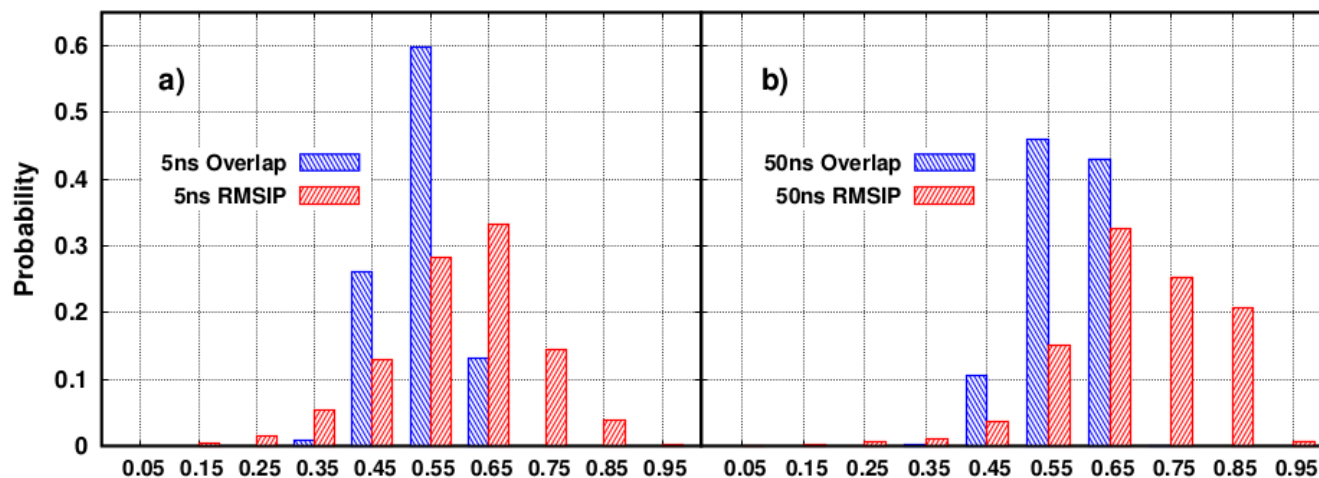
Four independent comparisons.
Each of 50 ns. System: BPTI.

Are reproducible the essential spaces?

- Run equivalent trajectories.
- Compute the PC-modes.
- Compute the RMSIP for the ES of alternative runs.

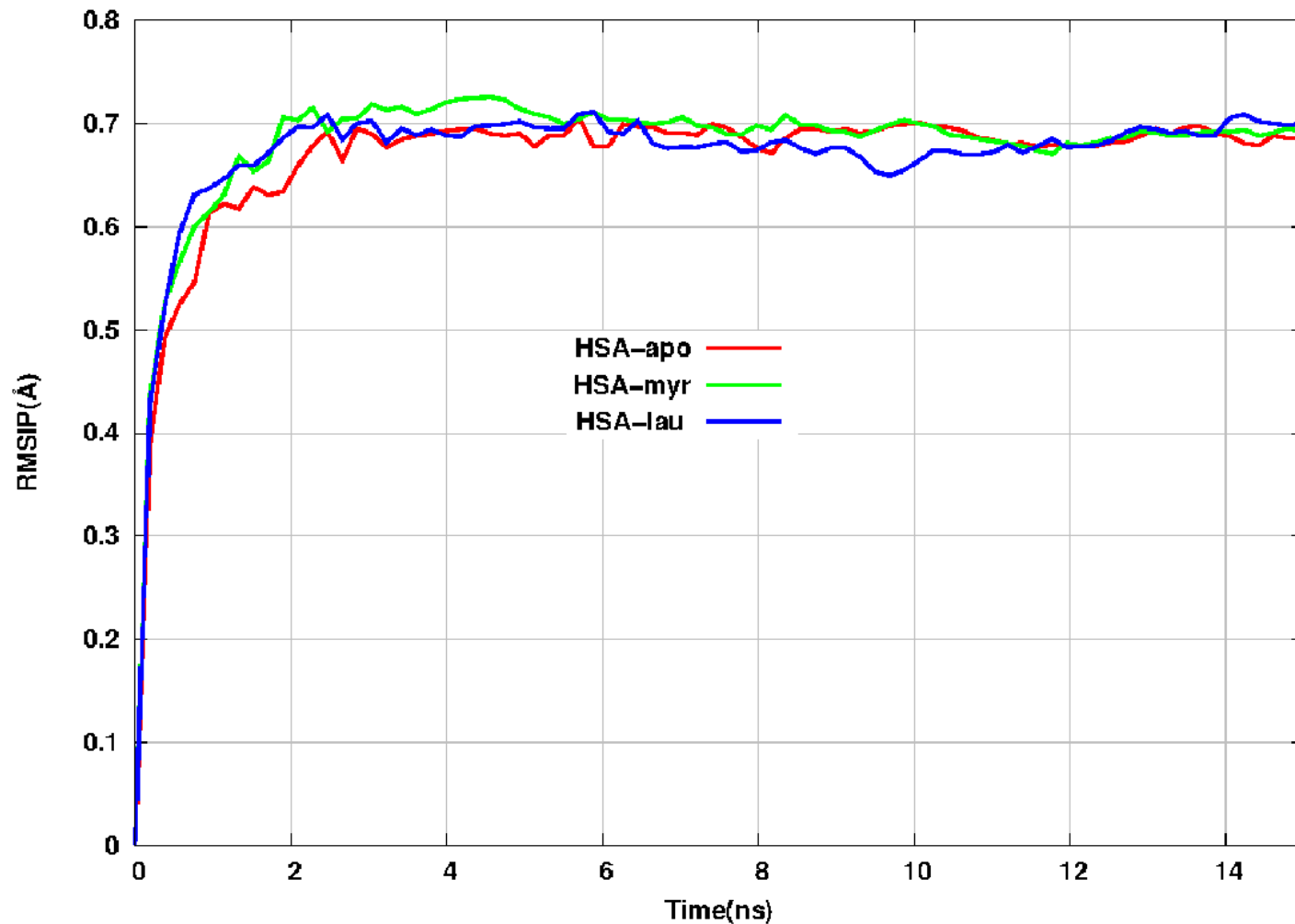
$$RMSIP = \frac{1}{M} \sum_{i=1}^M \sum_{j=1}^M \mathbf{v}_i \cdot \mathbf{v}'_j$$

$$RMSIP = \begin{cases} 1 & \text{if they span the same subspace} \\ 0 & \text{if subspaces are orthogonal} \end{cases}$$

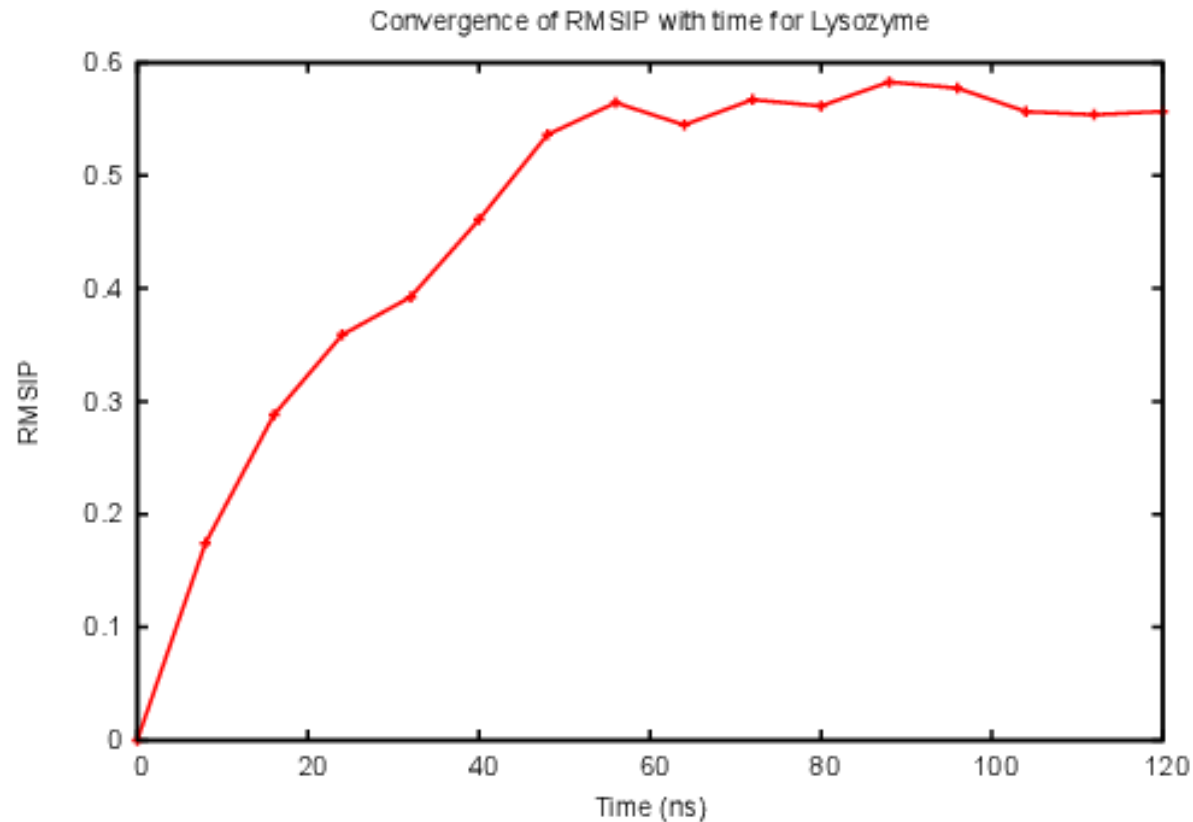


Huge # of trajectories
System: BPTI

Increasing time does not solve the problem

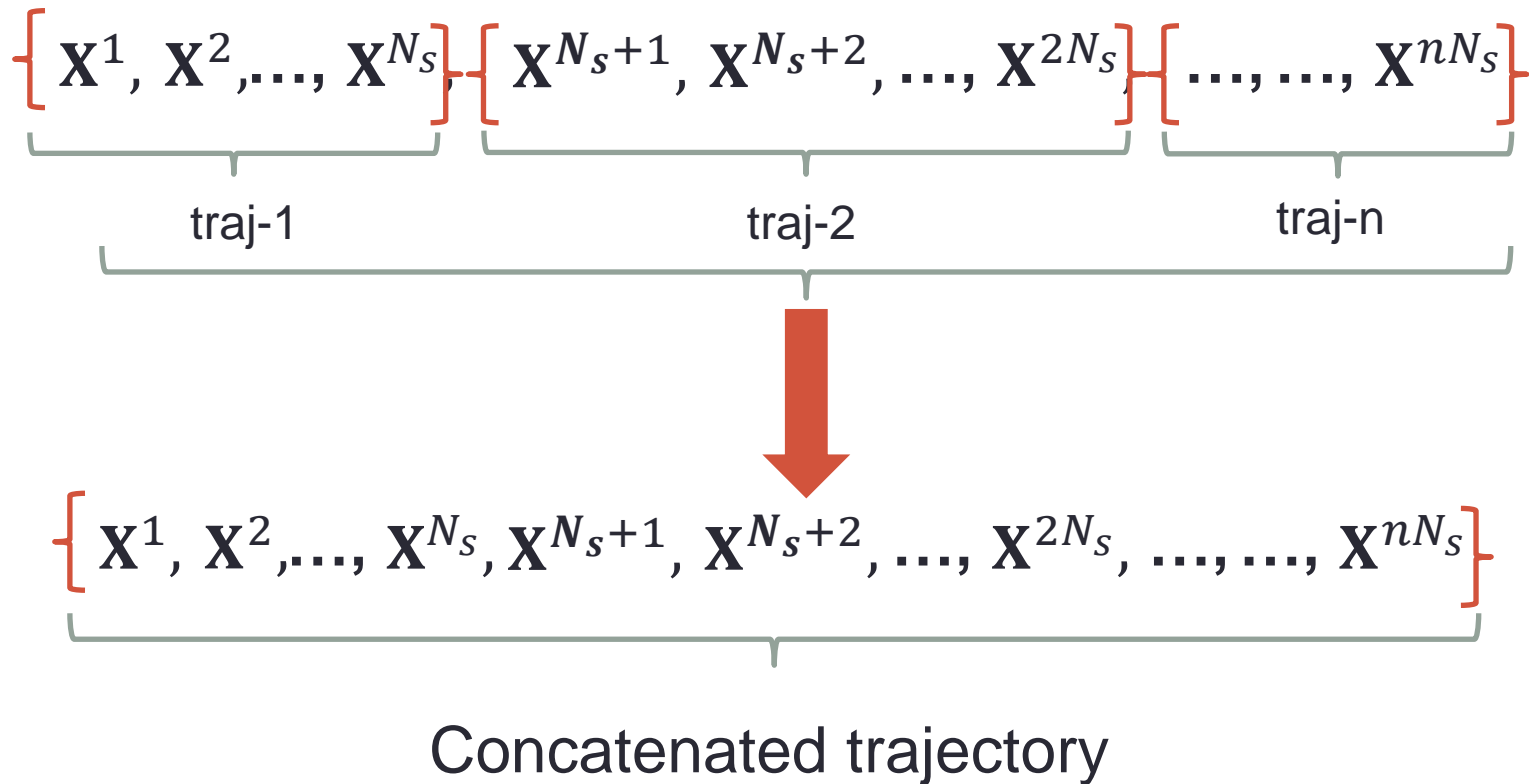


Increasing time does not solve the problem



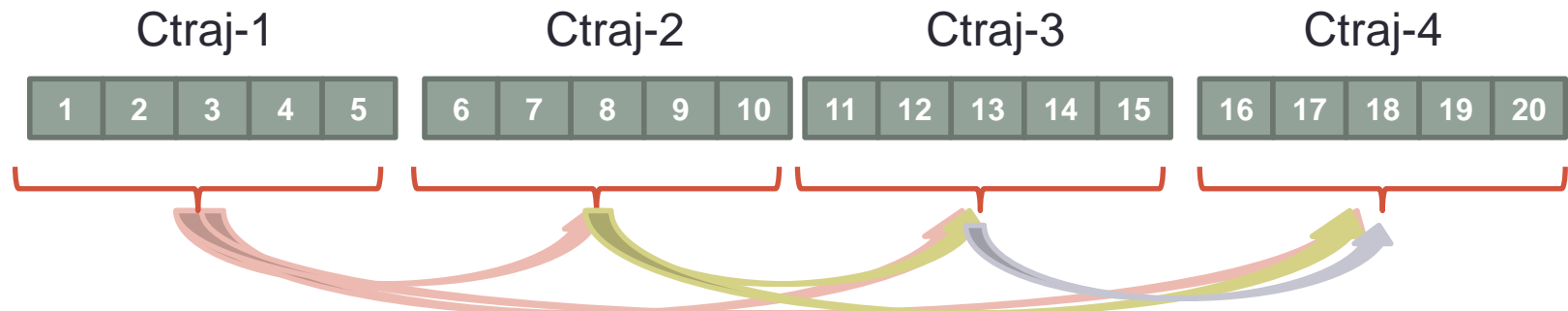
A simple way to improve the consistency of the PC-modes

- Concatenate equivalent trajectories!

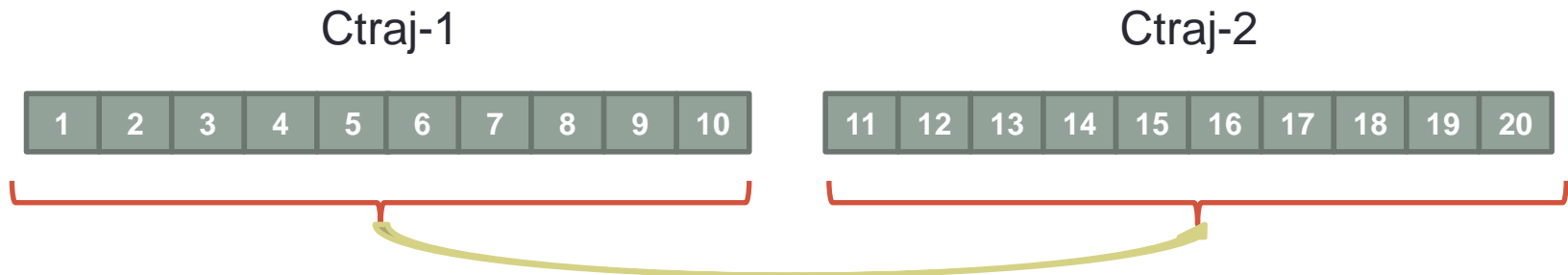


How to check that it works?

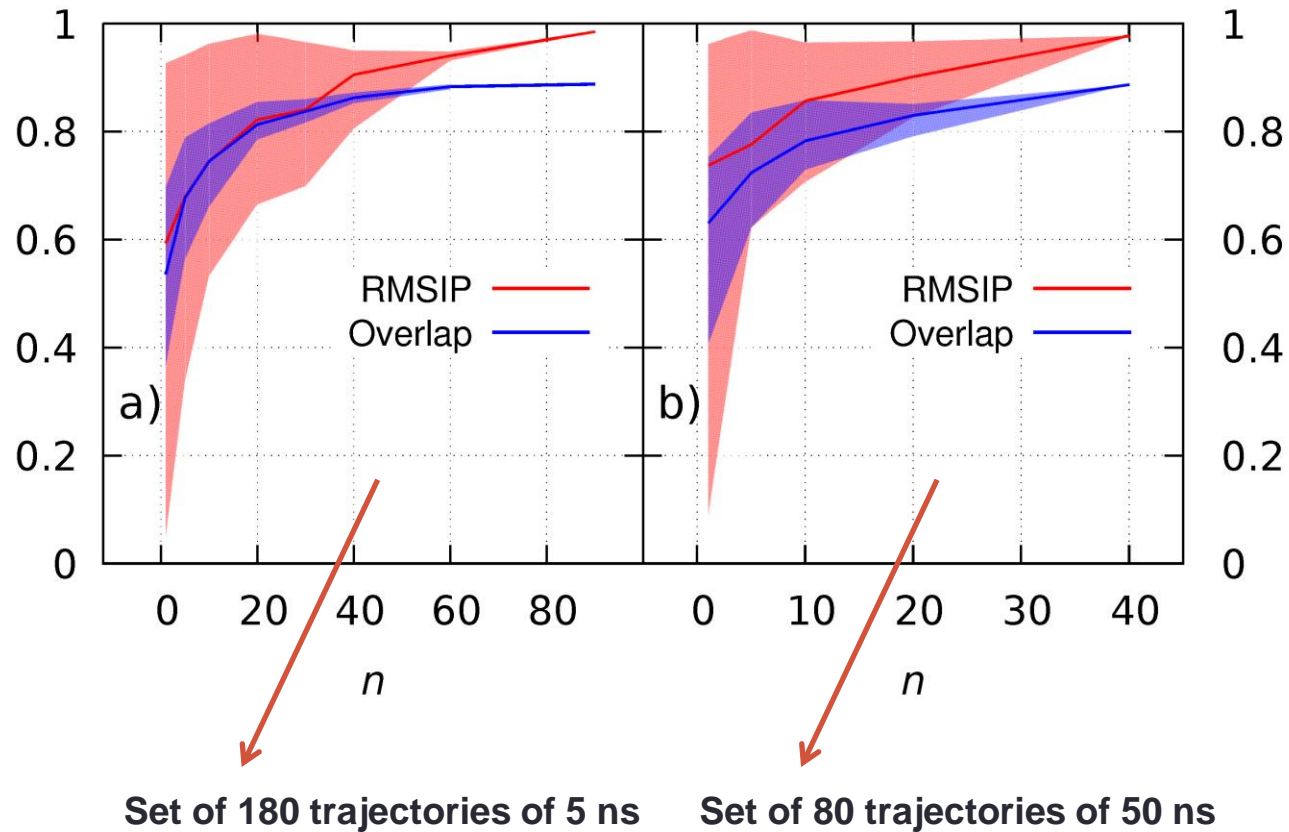
- Estimate the RMSIP values that can be obtained using different number of concatenated trajectories



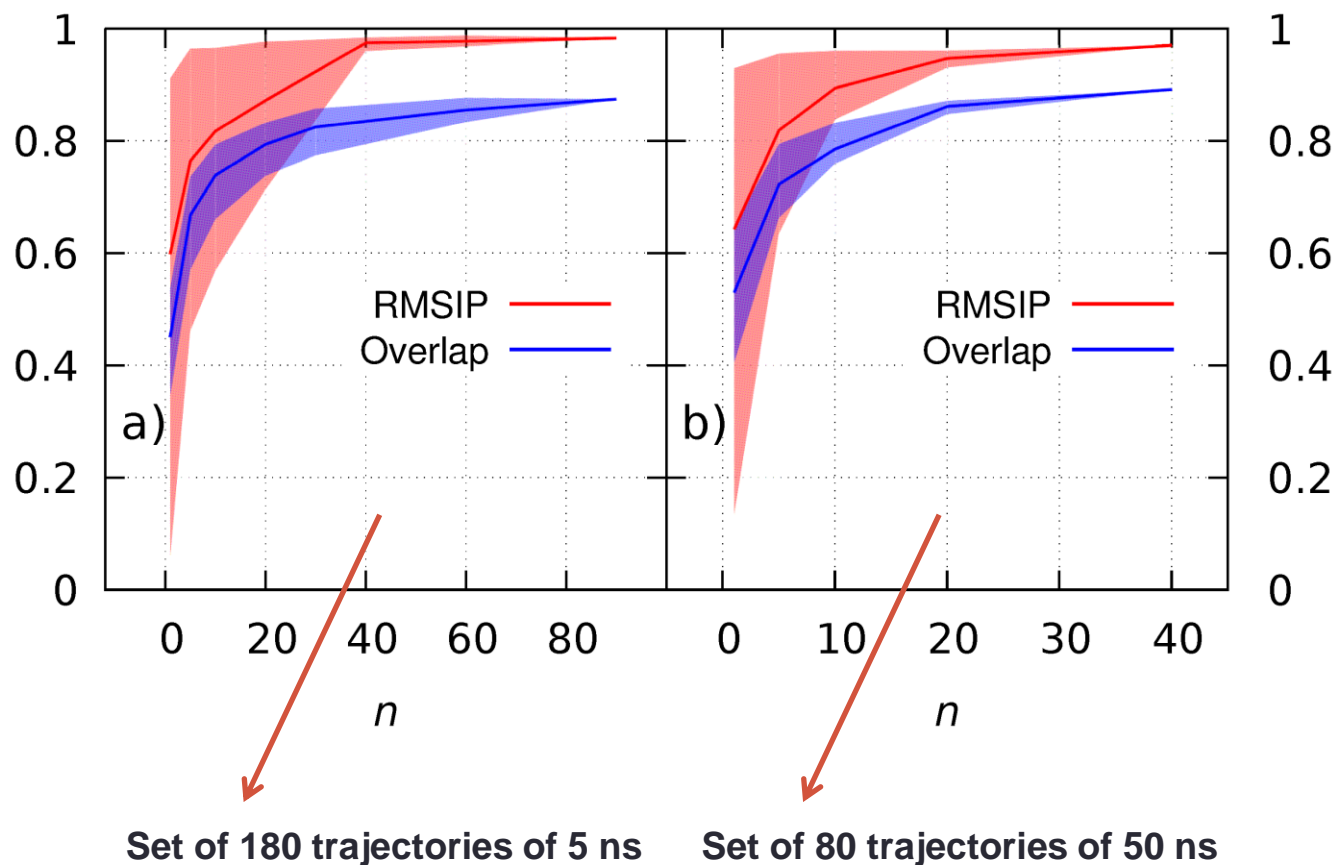
$$\text{Number of independent values of RMSIP} = \frac{N_{ctrhaj}(N_{ctrhaj} - 1)}{2} = 12$$



Results for BPTI

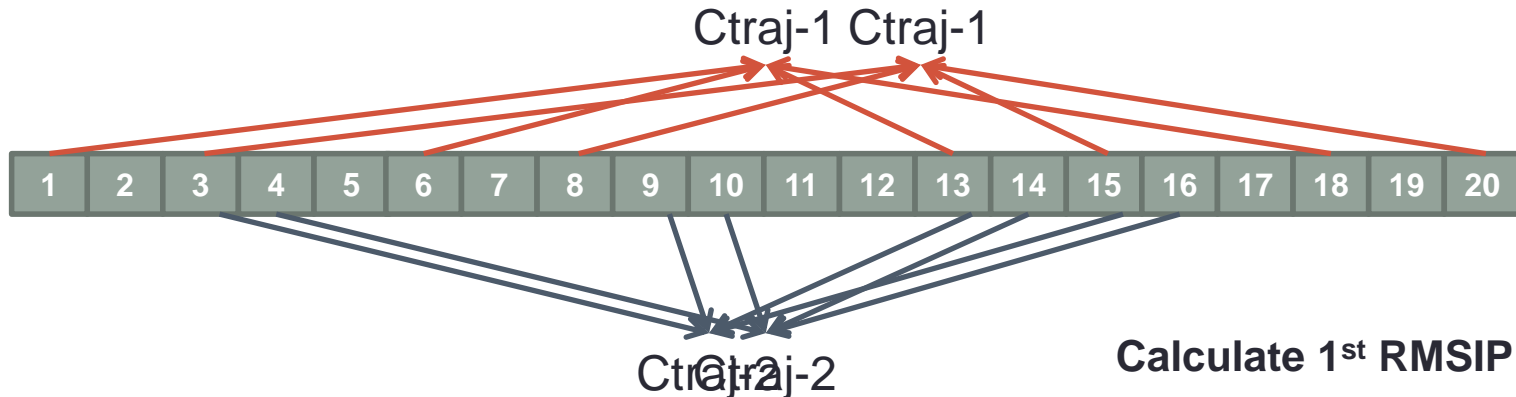


Results for lysozyme



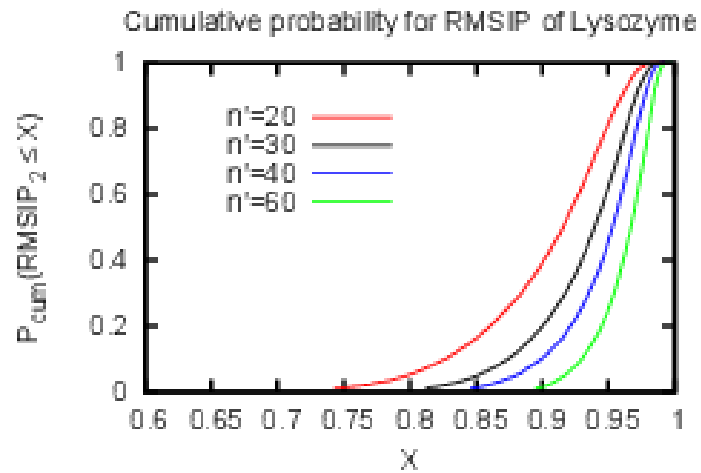
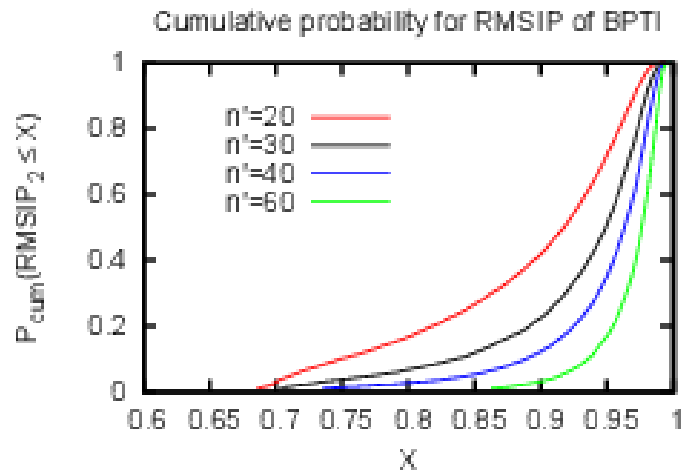
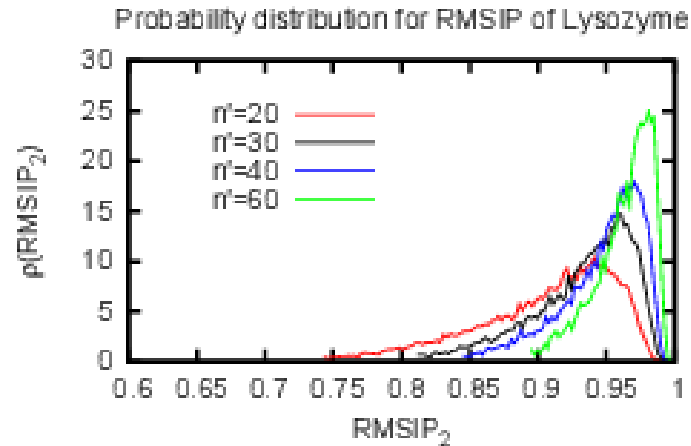
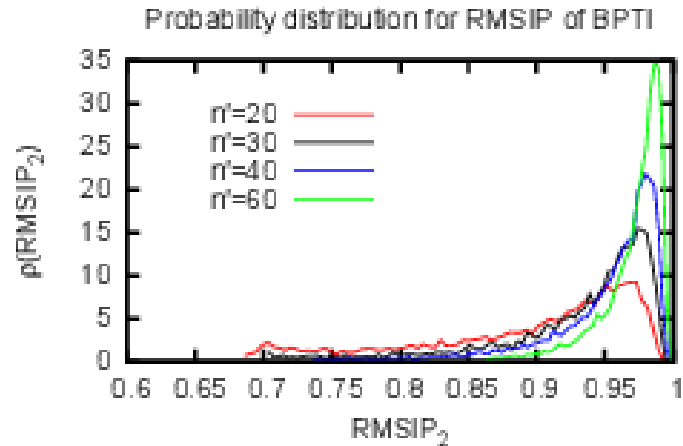
RMSIP distributions

- Previous procedure affords statistically-independent RMSIP values.
 - But for large n we obtain too few values.
 - Too low variability.
- To get more variability
 - Compute an even larger number of trajectories.
 - Form alternative pairs of concatenated trajectories by selecting at random from this set.



Calculate 1st RMSIP value
Calculate 2nd RMSIP value

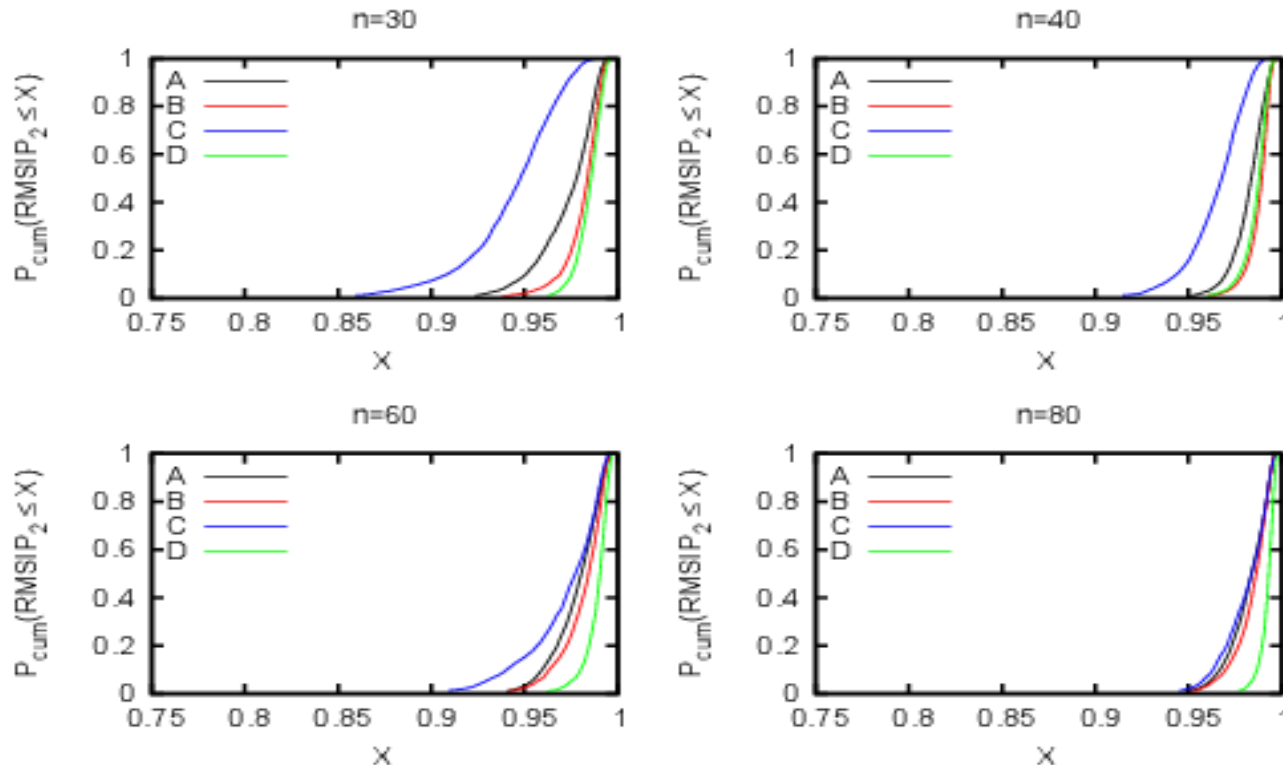
RMSIP distributions



How to assess the convergence?

- If $n/2$ trajectories provide good convergence, n trajectories provide good convergence, too.

Cumulative probabilities for RMSIPs obtained with n and $n/2$ trajectories



Why does it work?

- We need to understand what can be expected from the PC-modes of a concatenated trajectory.
 - *“The essential dynamic analysis can be performed on a combined trajectory (constructed by concatenating the trajectories). This is a powerful tool to evaluate similarities and differences between the essential motions in different trajectories of the same protein. If the motions are similar, then the eigenvalues (and eigenvectors) coming from separate trajectories and from the combined trajectory should be similar.”*

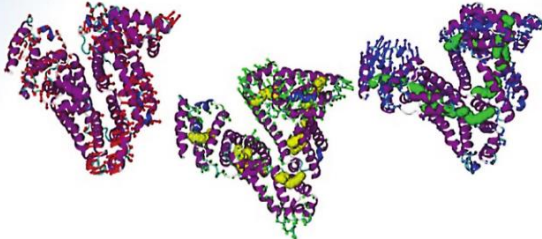
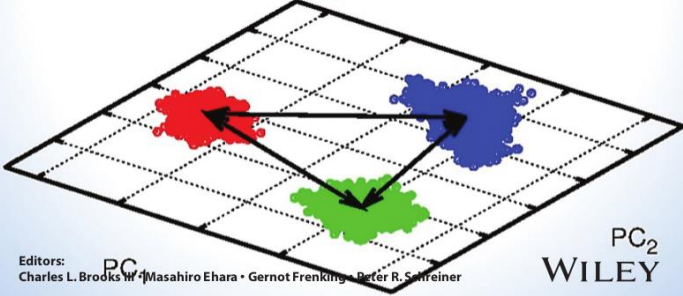
Van Aalten et. al. Proteins: Structure, Function and Genetics, 22, 45-54, 1995.

The correlation matrix of concatenated trajectories

Volume 36 | Issues 7-8 | 2015
Included in this print edition:
Issue 7 (March 15, 2015)
Issue 8 (March 30, 2015)

Journal of COMPUTATIONAL CHEMISTRY
Organic • Inorganic • Physical
Biological • Materials

www.c-chem.org

$$C^{ABC} = \frac{C^A + C^B + C^C}{3} + S^{ABC}$$



Editors: Charles L. Brooks III • Masahiro Ehara • Gernot Frenking • Peter R. Schreiner

WILEY

Journal of
**COMPUTATIONAL
CHEMISTRY**

WWW.C-CHEM.ORG

FULL PAPER

New Insights into the Meaning and Usefulness of Principal Component Analysis of Concatenated Trajectories

Gustavo Pierdominici-Sottile and Juliana Palma*

A comparison between different conformations of a given protein, relating both structure and dynamics, can be performed in terms of combined principal component analysis (combined-PCA). To that end, a trajectory is obtained by concatenating molecular dynamics trajectories of the individual conformations under comparison. Then, the principal components are calculated by diagonalizing the correlation matrix of the concatenated trajectory. Since the introduction of this approach in 1995 it has had a large number of applications. However, the interpretation of the eigenvectors and eigenvalues so obtained is based on intuitive foundations, because analytical expressions relating the concatenated correlation matrix with those of the individual trajectories under consideration have not been provided yet. In this article, we present such expressions for the

cases of two, three, and an arbitrary number of concatenated trajectories. The formulas are simple and show what is to be expected and what is not to be expected from a combined-PCA. Their correctness and usefulness is demonstrated by discussing some representative examples. The results can be summarized in a simple sentence: the correlation matrix of a concatenated trajectory is given by the average of the individual correlation matrices plus the correlation matrix of the individual averages. From this it follows that the combined-PCA of trajectories belonging to different free energy basins provides information that could also be obtained by alternative and more straightforward means. © 2014 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23811

Introduction

Principal component analysis (PCA) has been widely used to characterize the dynamics of proteins since it allows to detect important directions in their multidimensional configurational space.^[1] These directions are obtained from molecular dynamics (MD) simulations by diagonalizing the correlation matrix.^[2] Usually, a few eigenvectors stand out for having eigenvalues far larger than the rest. Movements along these directions account for the largest structural variations of the peptidic chain, describing the so-called essential dynamics (ED) of the protein. Motions along the remaining eigenvectors just correspond to trivial, nearly Gaussian fluctuations. There have been many discussions on the reliability, usefulness, and meaning of the vectors identified by PCA.^[3-7] Besides, several tools have been provided to assess their stability and convergence.^[11] The main hypothesis of the approach is that the ED of a protein, determined with PCA, contains the motions relevant to its function.^[2] This hypothesis has gained support from the build-up of MD studies that describe a close relationship between the first eigenvectors of the correlation matrix and the functional motions of several proteins.^[8-14] The PCA method is closely related to quasiharmonic analysis, a method that provides an affordable approach to compute configurational entropies.^[15-17] Finally, we should note that PCA only identifies linear correlations between atomic fluctuations. More sophisticated procedures have to be used to detect correlations beyond linearity.^[18,19]

An extension of the PCA method consists of diagonalizing the correlation matrix obtained by concatenating two or more independent trajectories, each corresponding to an alternative conformation of the same protein.^[20] They could be, for exam-

ple, the trajectories for the holo and apo forms of a protein, the active and inactive forms of an enzyme, the open and closed structures of a channel protein or distinct oligomerization states of a given protein subunit. It is known that the main eigenvectors of these combined correlation matrices (CCM) no longer describe the largest deformations of the conformations involved. Instead, it has been asserted that they highlight differences in the structure and dynamics of the proteins under comparison. The occurrence of static modes among the eigenvectors of the CCM has frequently been reported (see for example Refs. [20-24]). They are identified as eigenvectors of the CCM for which the projections of the individual trajectories differ significantly.

To the best of our knowledge, an analytical expression relating the CCM to the structures and correlation matrices of the individual trajectories involved has not been provided yet. In this article, we present such formulas for the cases of two, three, and an arbitrary number of concatenated trajectories. We believe that they will be useful to enlighten the interpretation of the results of combined-ED analysis and to guide its discussion. Among other things, these expressions allow to predict the number of static modes to be expected and afford a precise and clear meaning for the eigenvalues and directions of these eigenvectors.

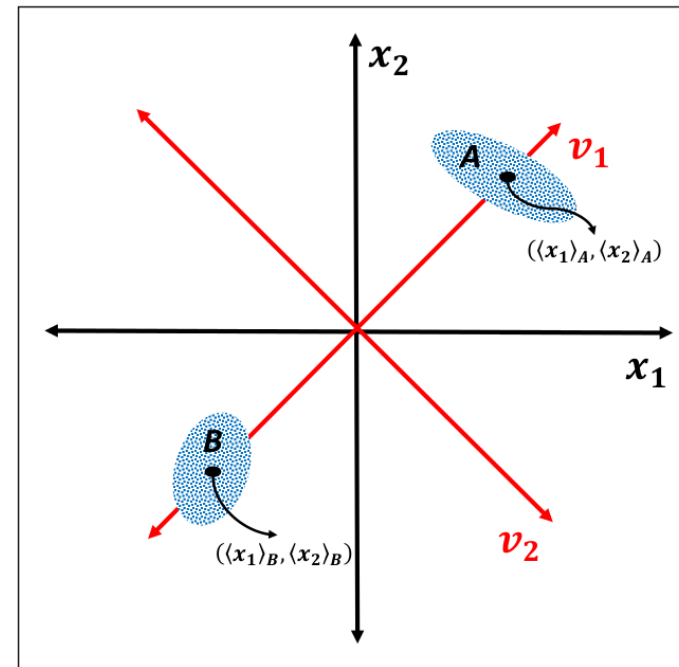
G. Pierdominici-Sottile, J. Palma
Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes,
Sierra Peña 352, Bernal, B1876BND, Argentina
E-mail: juliana@unq.edu.ar
Contract grant sponsor: CONICET and Universidad Nacional de Quilmes
© 2014 Wiley Periodicals, Inc.

The correlation matrix of concatenated trajectories

- Is the average of the individual correlation matrices plus the correlation matrix of the individual average structures.

$$\mathbf{C}^{(2)} = \frac{\mathbf{C}^A + \mathbf{C}^B}{2} + \mathbf{S}^{(2)}$$

$\mathbf{C}^{(2)}$ → Corr matrix of concat traj
 $\frac{\mathbf{C}^A + \mathbf{C}^B}{2}$ → Individual corr matrices
 $\mathbf{S}^{(2)}$ → Corr matrix of average structures



The correlation matrix of concatenated trajectories

$$\mathbf{C}^{(2)} = \frac{\mathbf{C}^A + \mathbf{C}^B}{2} + \mathbf{S}^{(2)}$$

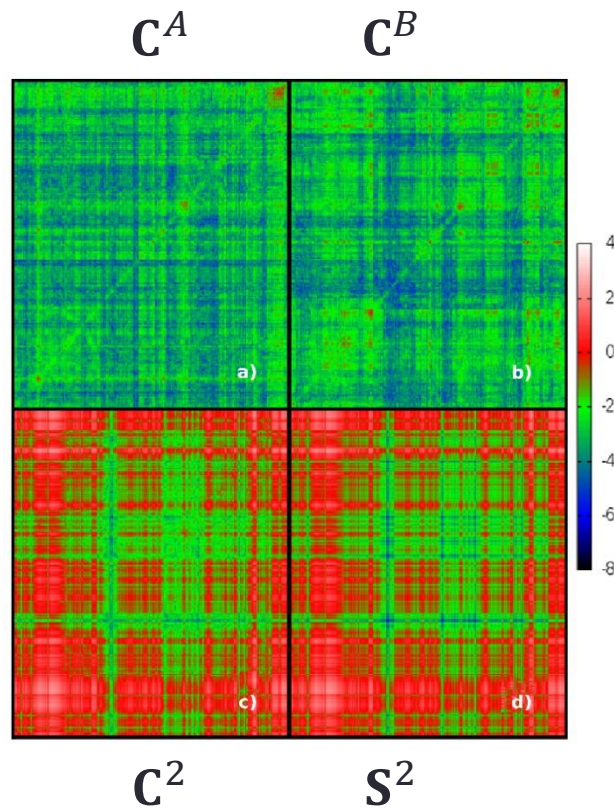
Information about fluctuations
observed in individual trajectories

Dynamic contribution

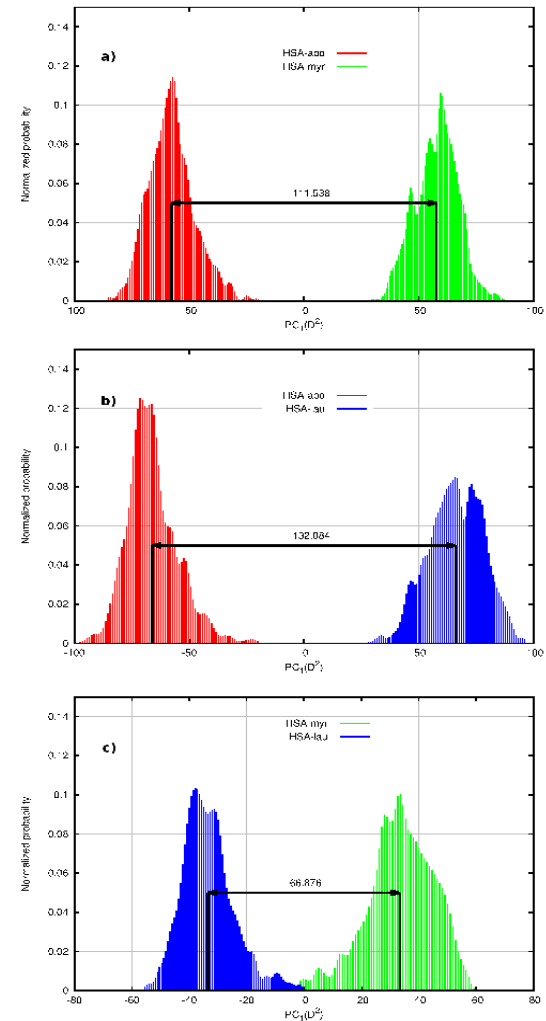
Information about differences in
average structures.

Static contribution

If the static contribution dominates



For $n=2$ the \mathbf{S} matrix has a single eigenvector

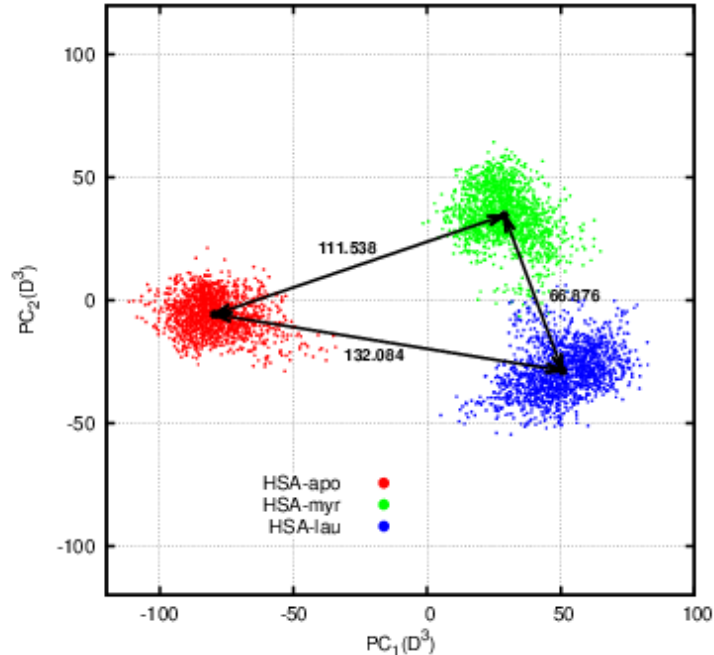


If the static contribution dominates

$$\mathbf{C}^{(3)} = \frac{\mathbf{C}^A + \mathbf{C}^B + \mathbf{C}^C}{3} + \mathbf{S}^{(3)}$$

The $\mathbf{S}^{(3)}$ matrix has two eigenvectors.

They span the plane that contains the three average structures.



If the static contribution is negligible

$$\mathbf{C}^{(n)} = \frac{\sum_{i=1}^n \mathbf{C}^{(i)}}{n} + \mathbf{S}^{(n)}$$

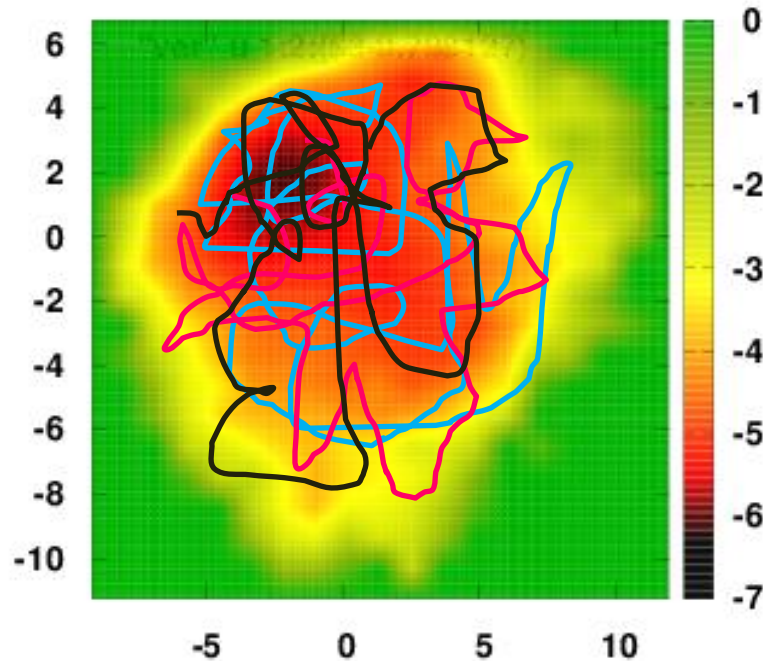


$$\mathbf{C}^{(n)} \approx \frac{\sum_{i=1}^n \mathbf{C}^{(i)}}{n}$$

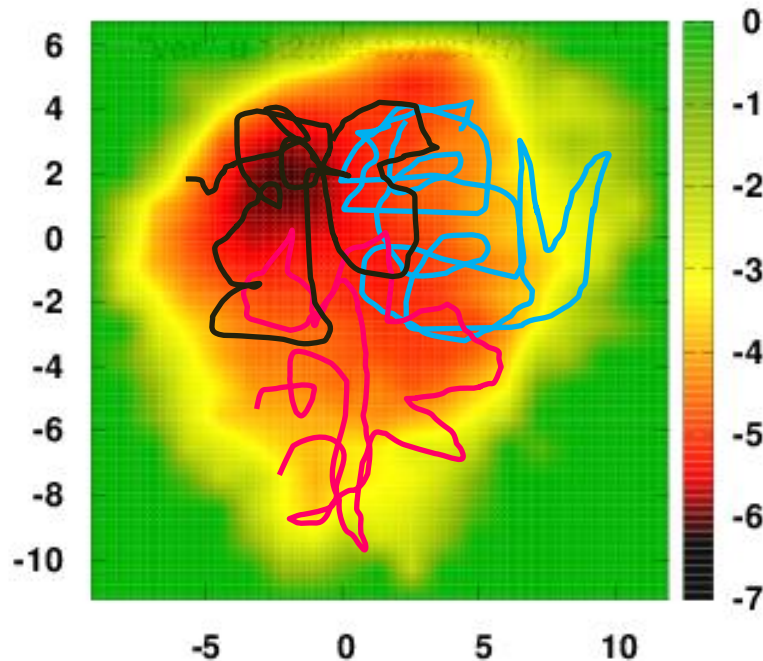
The statistical error in the elements of $\mathbf{C}^{(n)}$ is that of the individual the $\mathbf{C}^{(i)}$ divided by $n^{1/2}$.

When is negligible $\mathbf{S}^{(n)}$?

- When the fluctuations of individual trajectories are much larger than differences between the average structures.



What happens if the trajectories are biased?



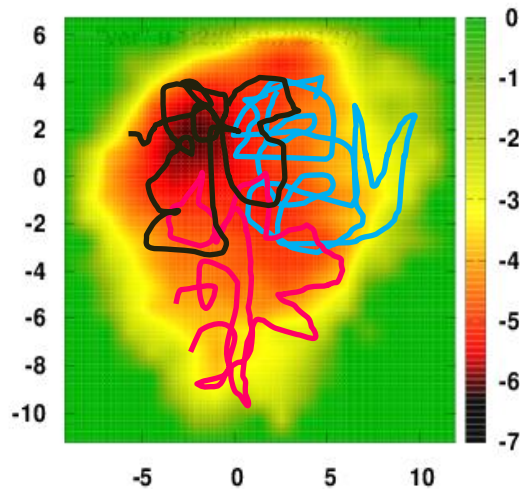
We can still have:

$$\mathbf{C}^{(n)} \approx \frac{\sum_{i=1}^n \mathbf{C}^{(i)}}{n}$$

How?

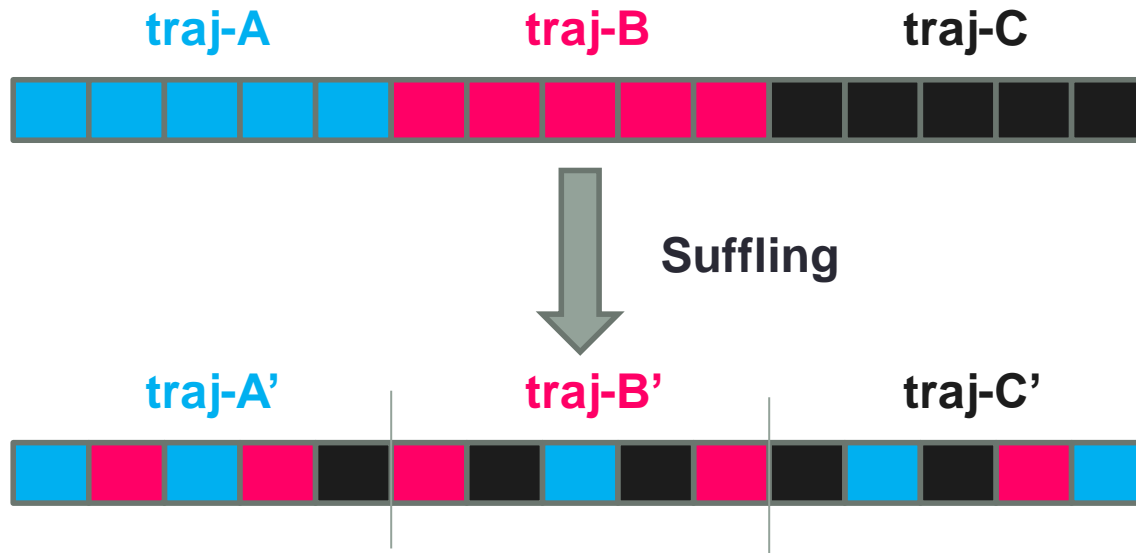
Why?

- Because correlation matrices are independent of the order of the samples



$$\mathbf{C}^{(3)} = \frac{\mathbf{C}^A + \mathbf{C}^B + \mathbf{C}^C}{3} + \mathbf{S}^{(3)}$$

Non negligible



$$C^{(3)} = \frac{C^{A'} + C^{B'} + C^{C'}}{3} + S^{(3)'}$$

Neglibigle

Correlation matrices for concatenated trajectories

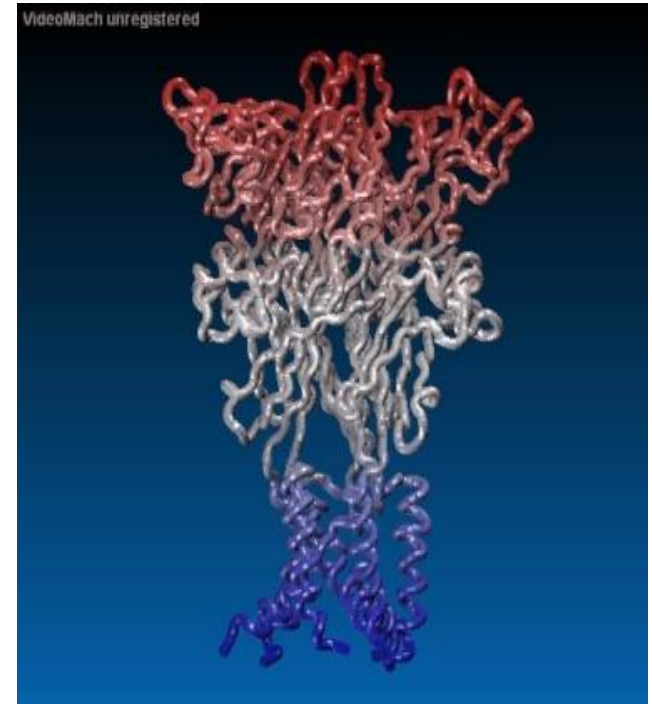
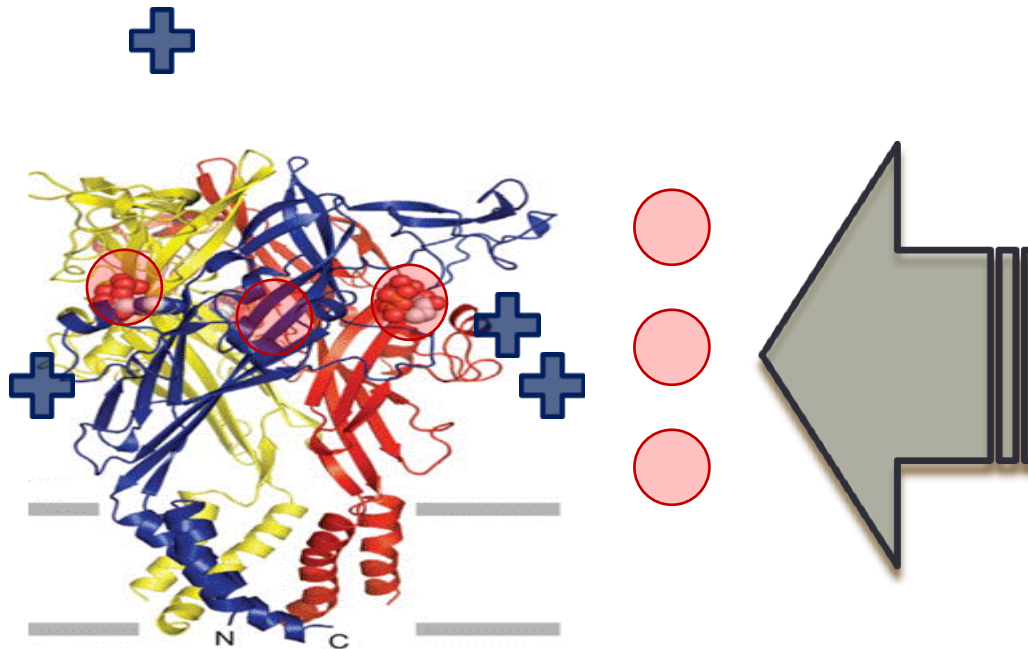
- For two or more separated free energy minima
 - Are dominated by the static contributions.
 - Little interest.
- For a single free energy minima
 - Have reduced statistical uncertainty.
 - Can be used to define consistent/reproducible PC-modes.
- For two or more connected free energy minima
 - To be studied...

PC-MODES OF INTER / INTRA MOVEMENTS

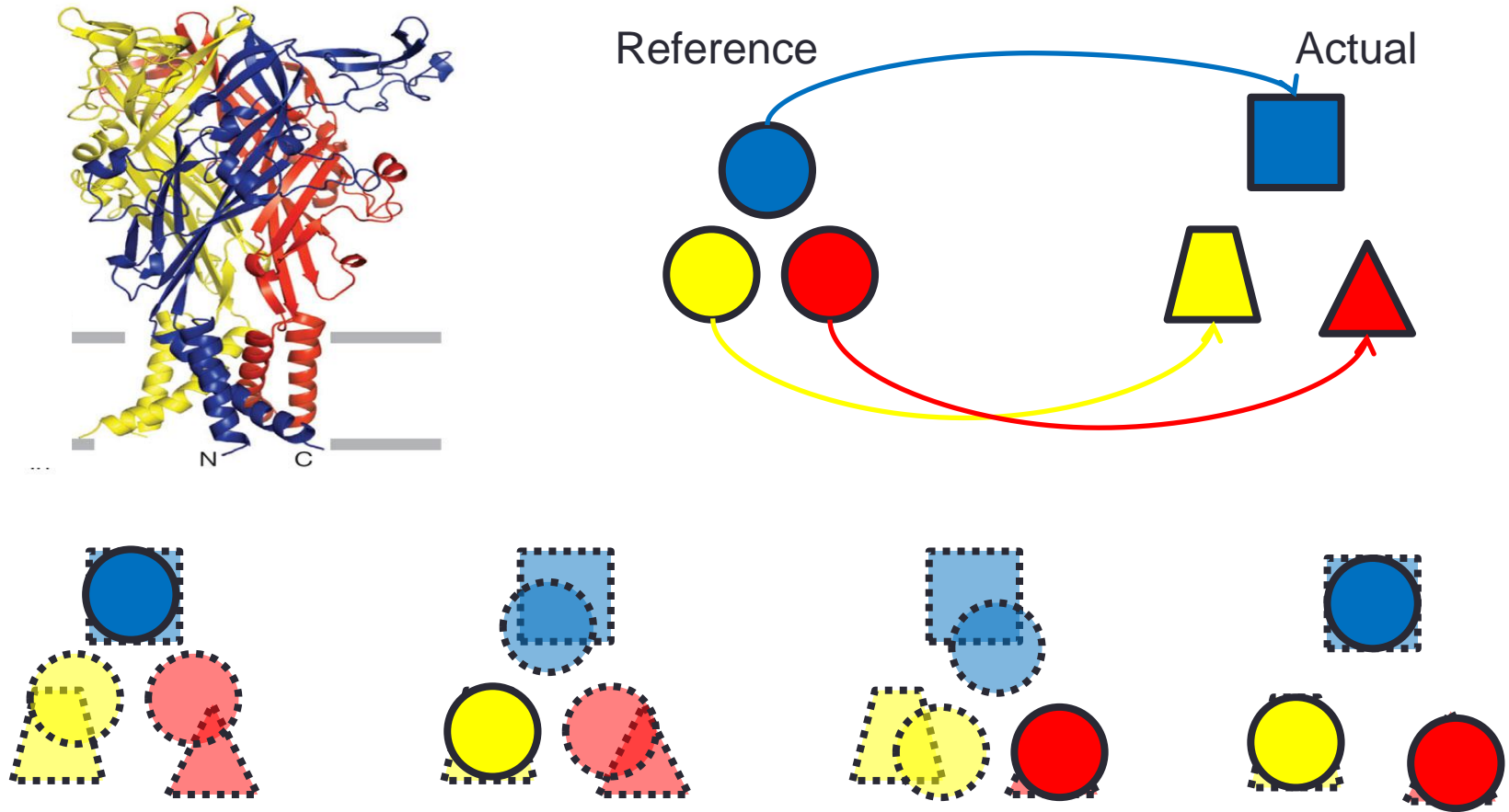
Application to P2X4

P2x4 is a membrane channel

- Activated by the union of three molecules of ATP
- It is a homotrimer



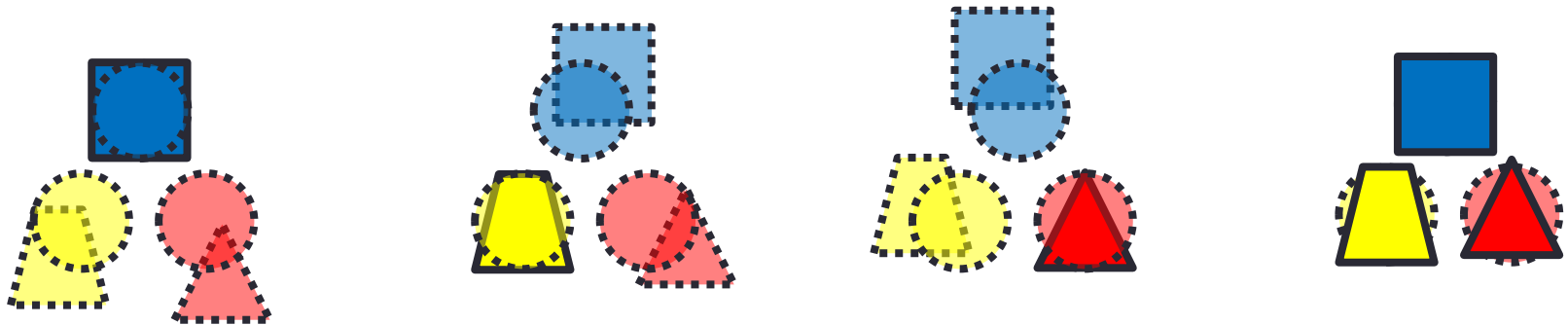
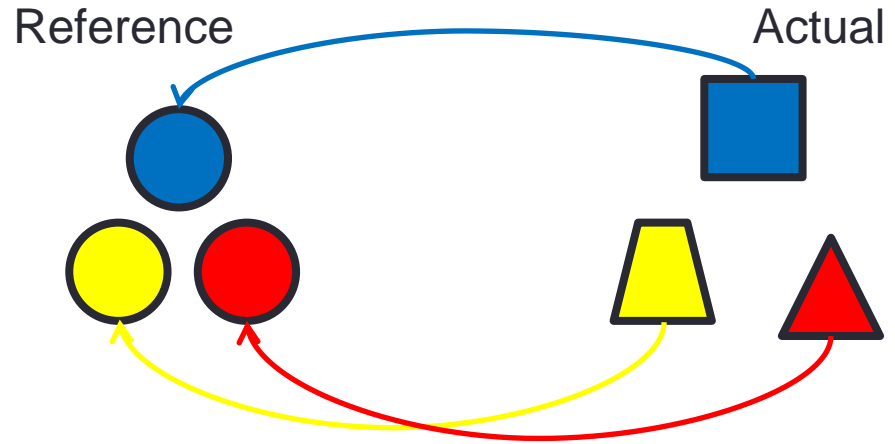
Uncover inter-chain motions



M. D. Vesper and B. L de Groot,
Plos Comput. Biol. 9(9): e1003232.

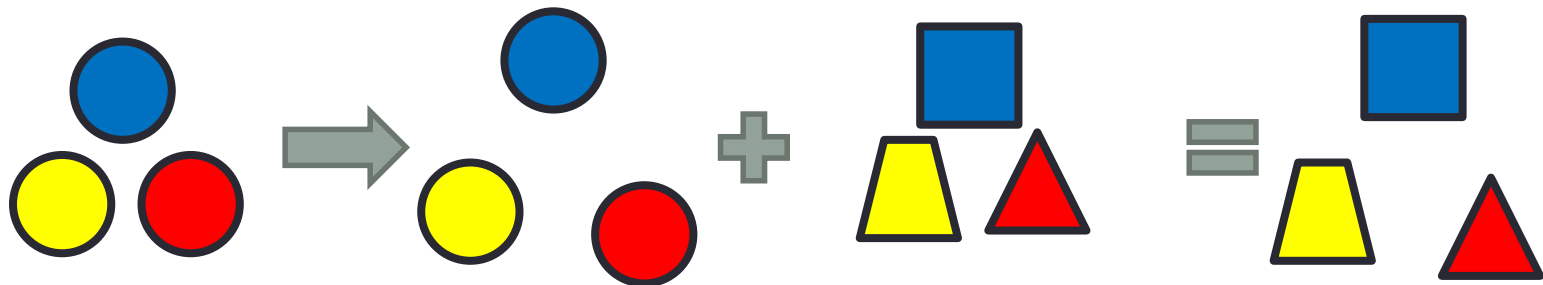
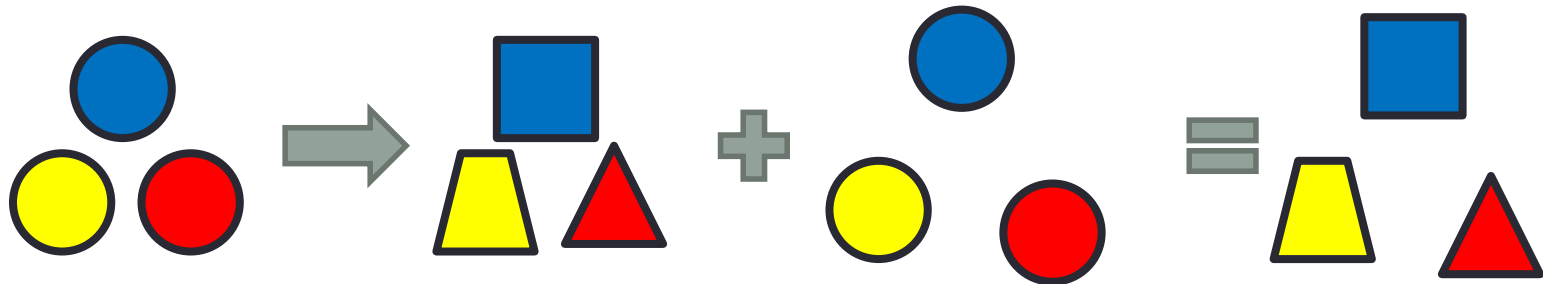
Inter chain
motions

Uncover intra-chain deformations



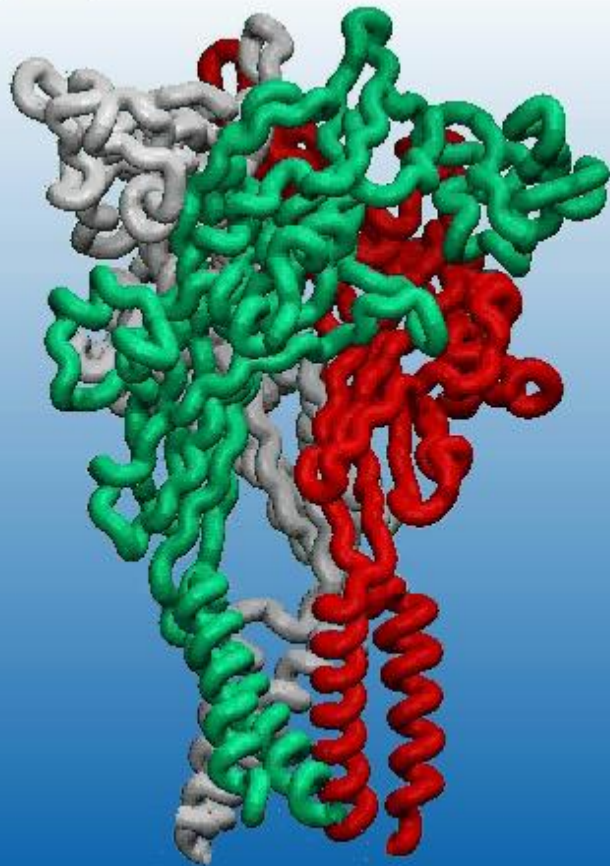
Intra chain
deformations

Altogether account for all motions

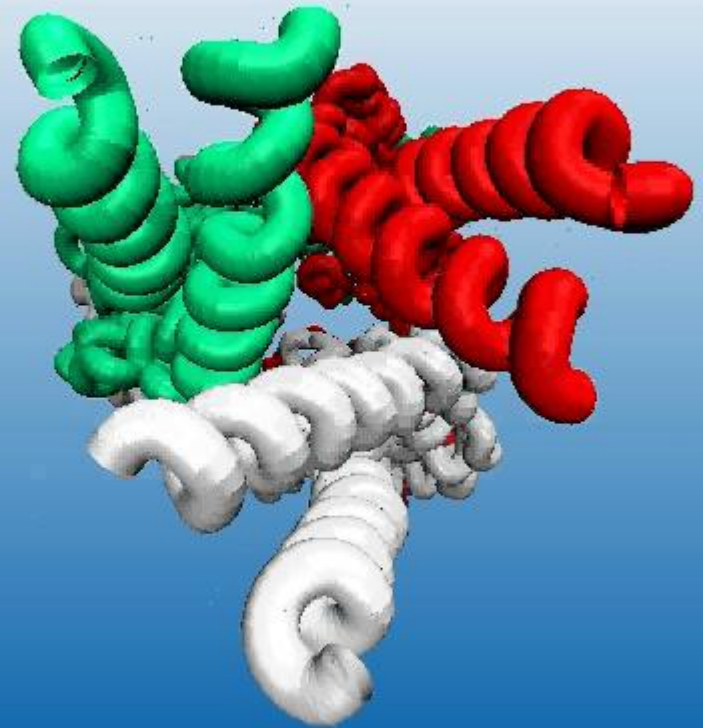


Inter-chain motions

VideoMach unregistered

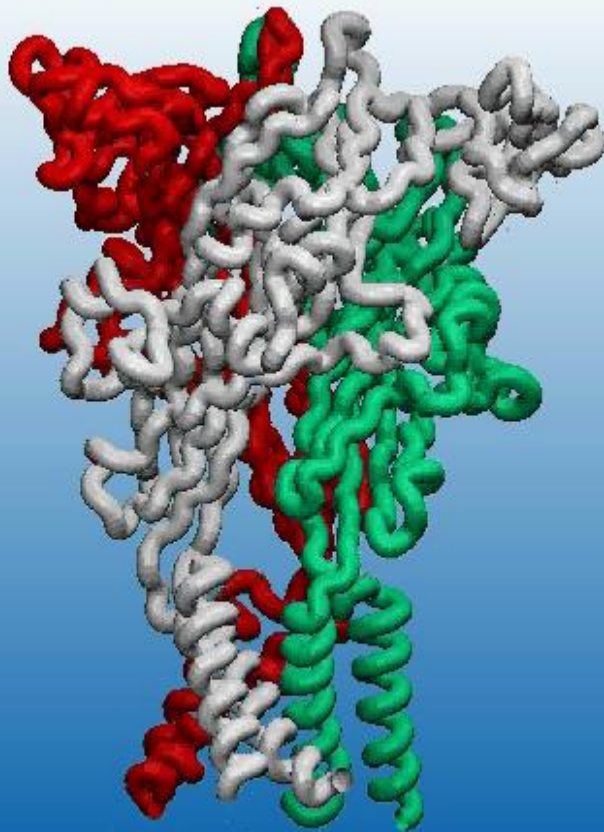


VideoMach unregistered

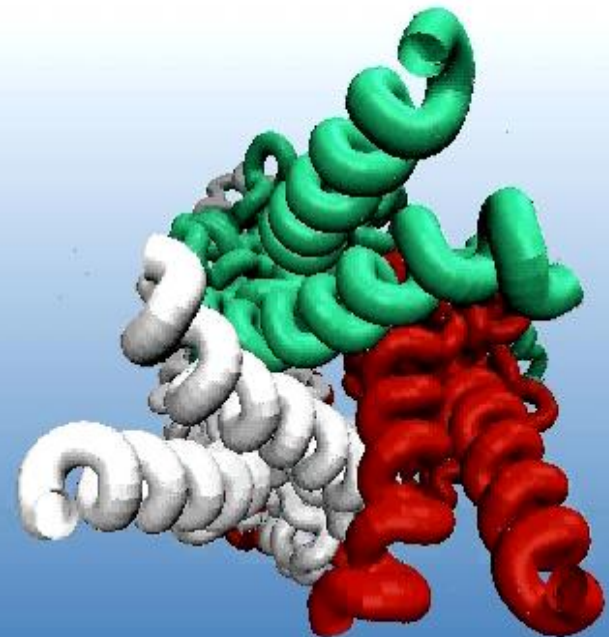


Intra-chain deformations

VideoMach unregistered



VideoMach unregistered



The opening of the pore

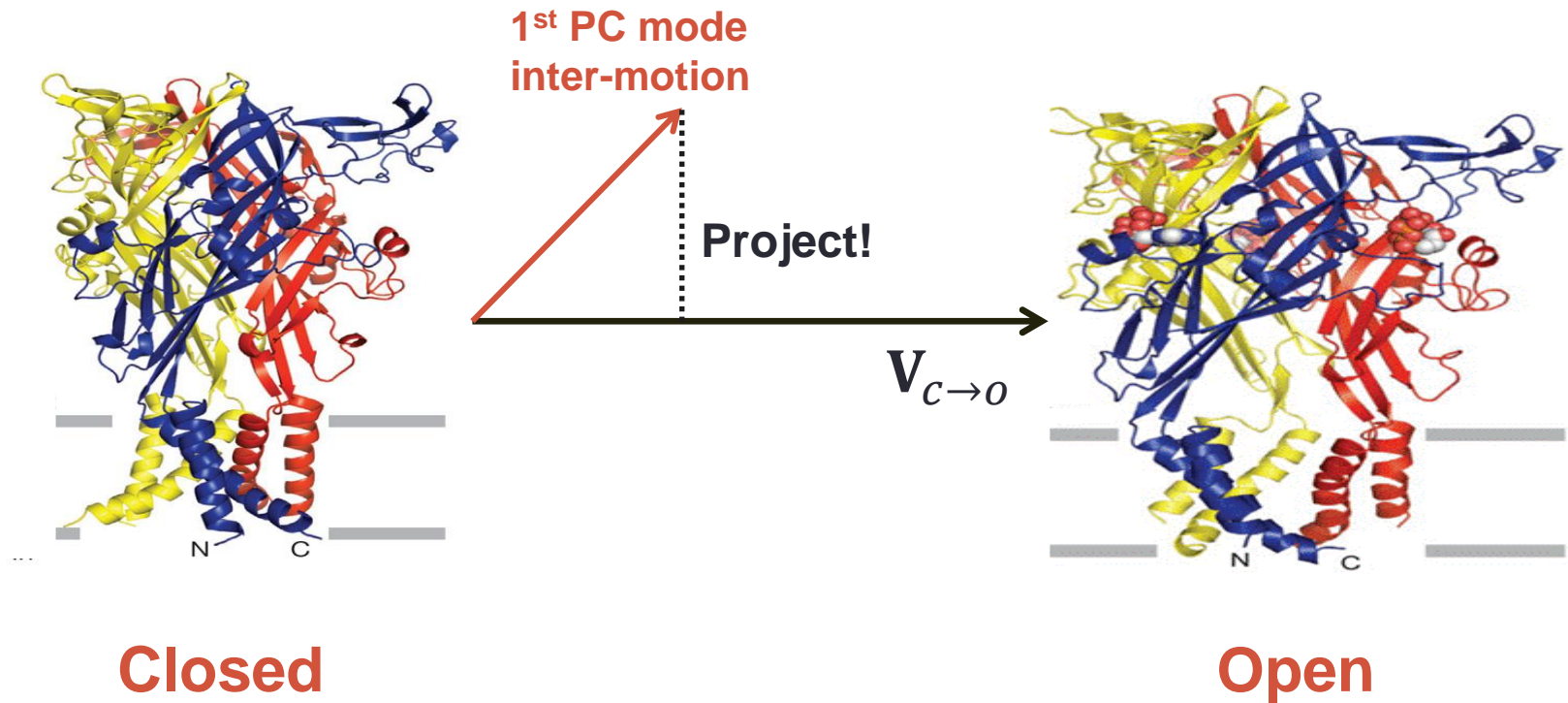
- Is mainly caused by inter-chain motions

Measures narrowest part of the pore



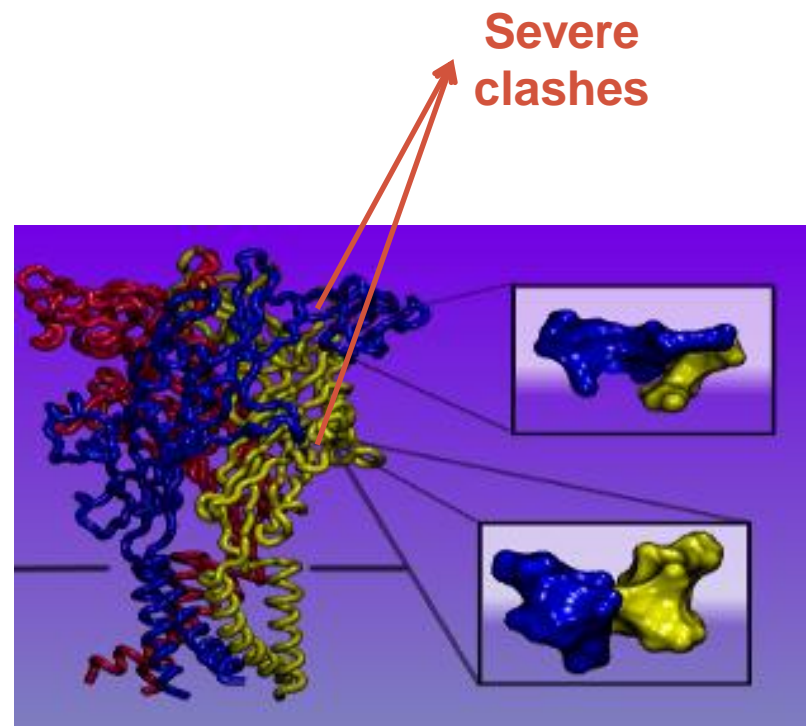
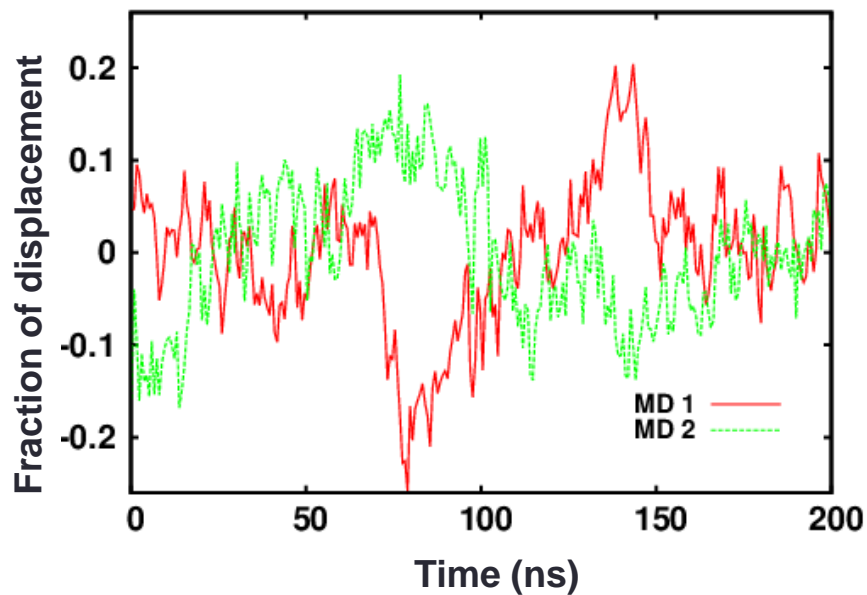
	$d[\text{axis-C}_\beta(\text{Ala 347})] / \text{\AA}$
$V_{c \rightarrow o}$	0.8 \rightarrow 3.60
Intra-chain deformations	0.8 \rightarrow 1.10
Inter-chain movements	0.8 \rightarrow 3.26

- Does this movement occurs in the absence of ATP?

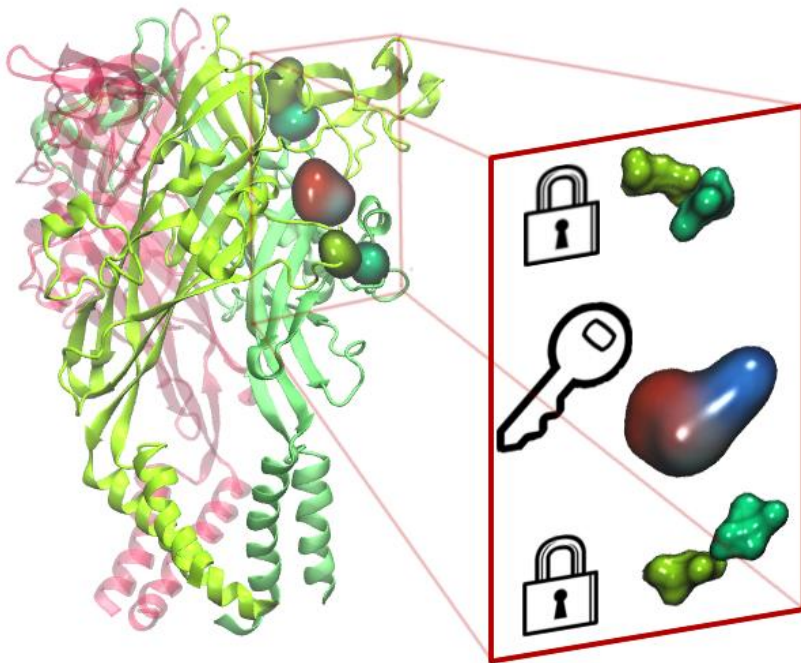


- The projection is large.
- Inter-chain motions are aligned with the $V_{C \rightarrow O}$ vector.
- But...

The amplitude is not large enough



Possible role of ATP



The screenshot shows the Biophysical Journal website interface. The main article featured is "The Dynamic Behavior of the P2X₄ Ion Channel in the Closed Conformation" by Gustavo Pierdominici-Sotillo, Luciano Moffatt, and Juliana Palma. The page includes a search bar, navigation tabs (Explore, Online Now, Current Issue, Archive, Journal Information, For Authors, Biophysical Society), and a "Current Issue" section for Vol. 111 Iss. 12, December 20, 2016. A "Submit Manuscript" button is also visible.

Biophysical Journal
Article

Biophysical Society

The Dynamic Behavior of the P2X₄ Ion Channel in the Closed Conformation

Gustavo Pierdominici-Sotillo,^{1,*} Luciano Moffatt,² and Juliana Palma¹

¹Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, CONICET, Buenos Aires, Argentina; and ²Instituto de Química Física de los Materiales, Medio Ambiente y Energía, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina

ABSTRACT We present the results of a detailed molecular dynamics study of the closed form of the P2X₄ receptor. The fluctuations observed in the simulations were compared with the changes that occur in the transition from the closed to the open structure. To get further insight on the opening mechanism, the actual displacements were decomposed into interchain motions and intrachain deformations. This analysis revealed that the iris-like expansion of the transmembrane helices mainly results from interchain motions that already take place in the closed conformation. However, these movements cannot reach the amplitude required for the opening of the channel because they are impeded by interactions occurring around the ATP binding pocket. This suggests that the union of ATP produces distortions in the chains that eliminate the restrictions on the interchain displacements, leading to the opening of the pore.

Conclusions

- We have established a clear meaning for concatenated PC-modes.
- We have shown that consistent / reproducible PC-modes can be obtained by concatenating equivalent trajectories.
- We have shown the usefulness of separating inter-chain from intra-chain displacements in analysing proteins with a quaternary structure.
- We have presented a hypothesis for the role of ATP in the activation of the P2X4 channel.

Acknowledgements

- **Group**
 - Gustavo Pierdominici-Sottile.
 - Rodrigo Cossio-Pérez.
 - Vanesa Racigh.
 - Agustín Ormazabal.



Acknowledgements



Thank you for your attention!