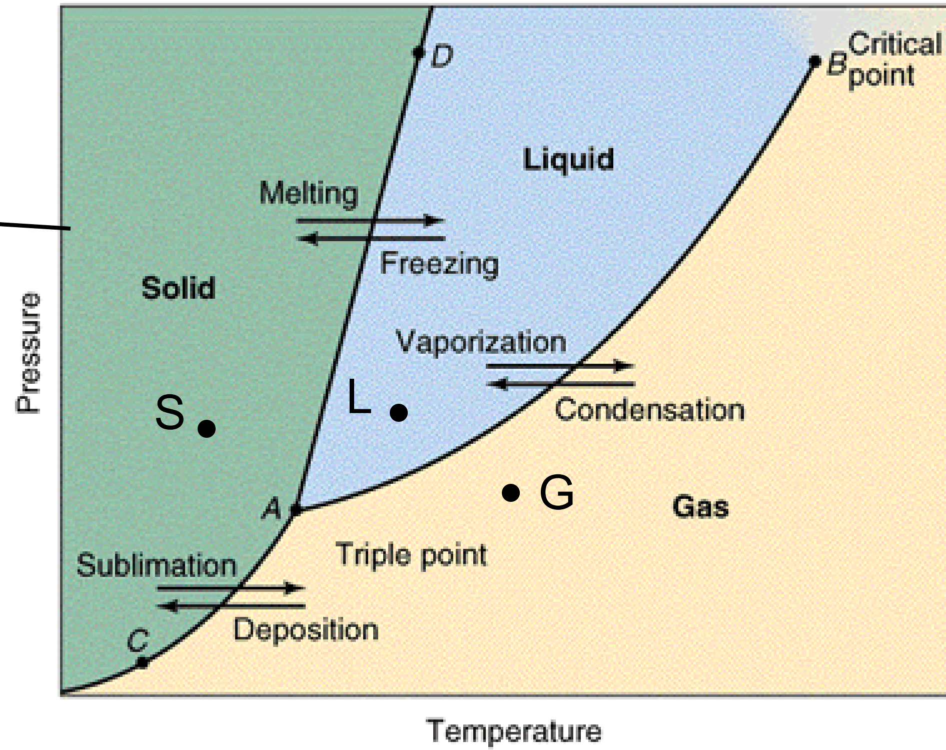
Time Series Approaches to Understanding Protein Dynamics

CHEONG Siew Ann

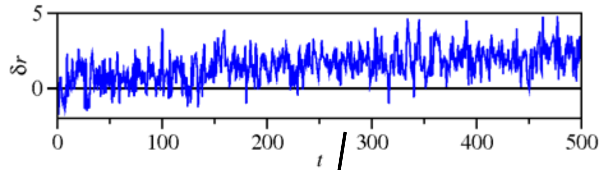
Division of Physics and Applied Physics
School of Physical and Mathematical Sciences
Email: cheongsa@ntu.edu.sg
Homepage: http://www1.spms.ntu.edu.sg/~cheongsa/

Macroscopic Thermal Physics

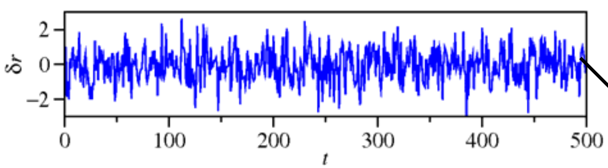


- Macroscopic order parameters differentiate
  - Solid (S)
  - Liquid (L)
  - Gas (G)

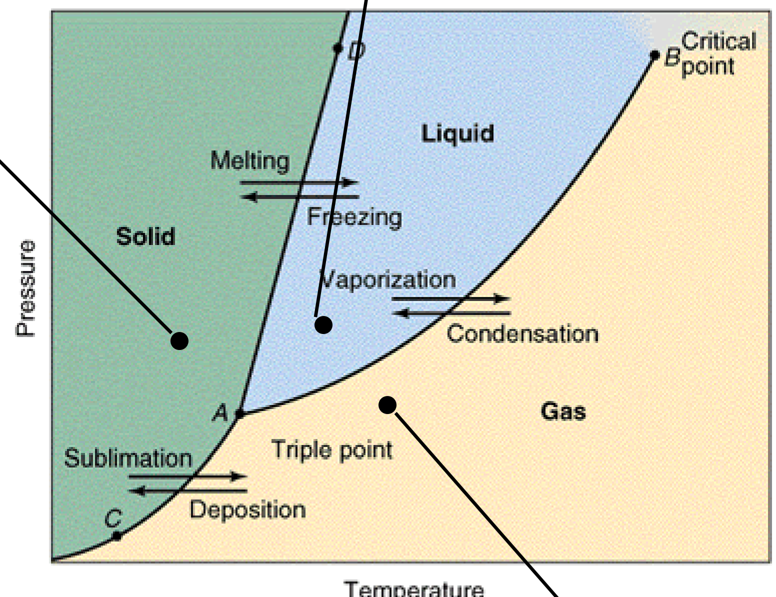
Microscopic Statistical Physics



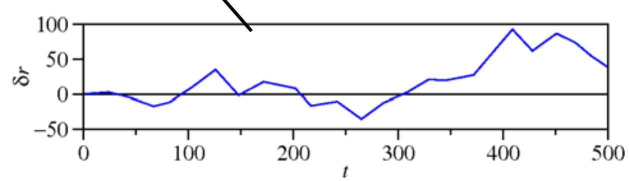
diffusive trajectories,
δr² increases with time



δr fluctuates about 0,
$\delta r^2 = \alpha T$ time-independent

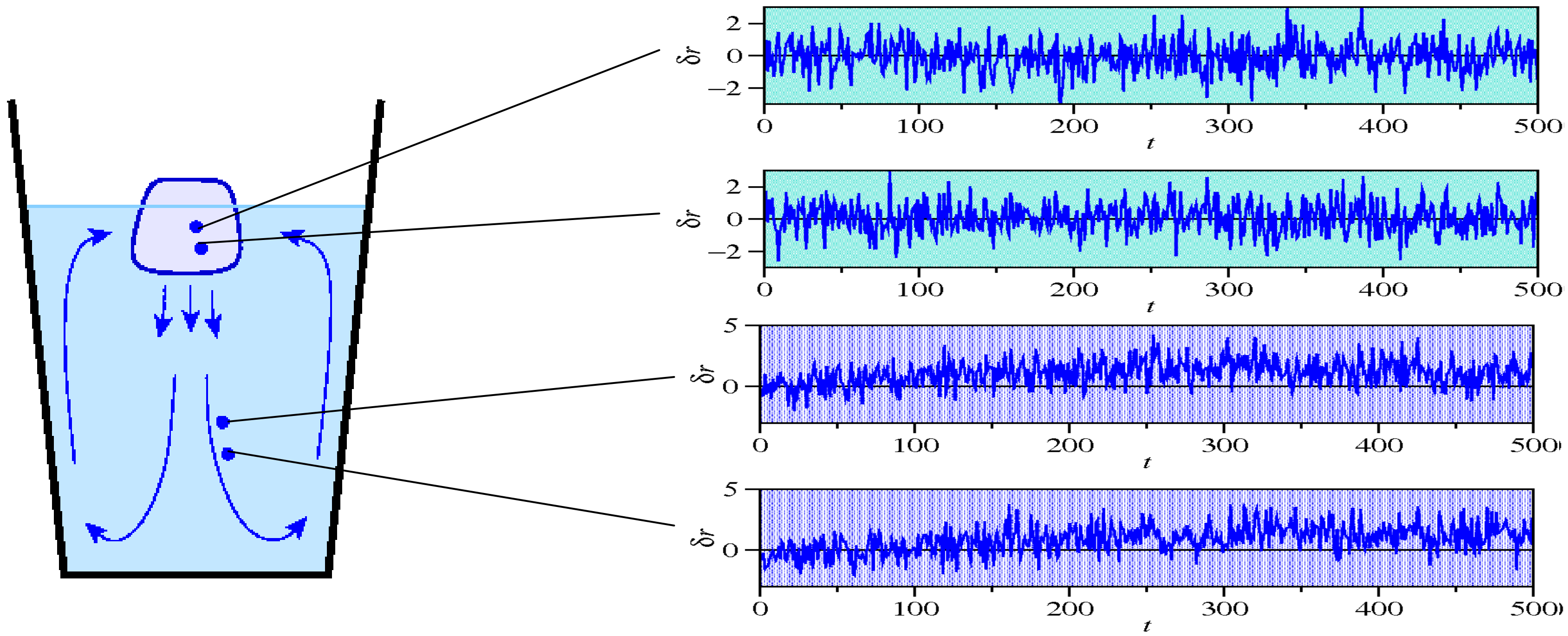


ballistic trajectories,
infrequent collisions



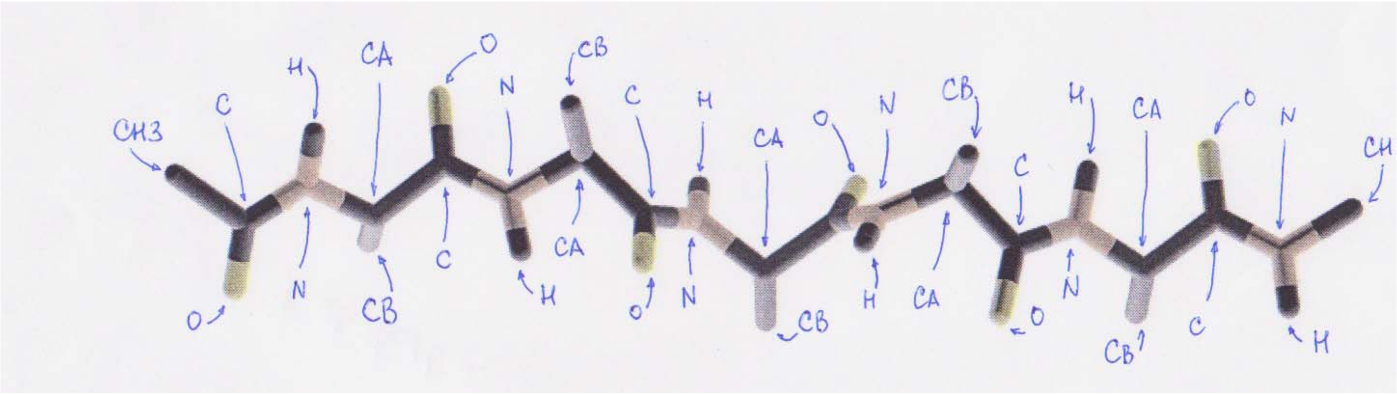
- S, L, G time series distinguishable
- S, L, G phase within single time series distinguishable

From Micro to Macro



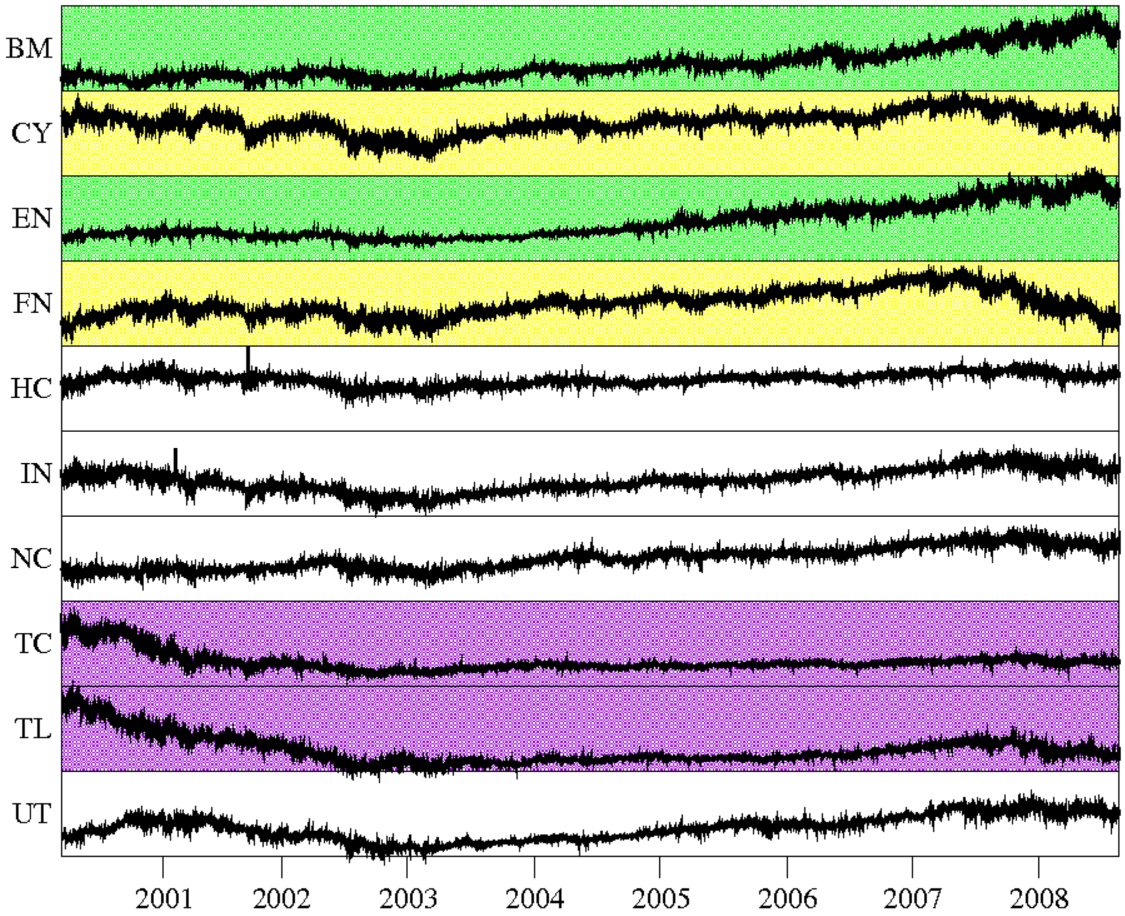
- Group statistically similar time series
- Discover presence of different phases

Dynamics of a Small Protein



- ACE-(ALA)$_5$-NME
  - 5 alanine repeated units
  - Capped by acetyl and methylamide groups
  - 62 atoms in all
- Simulation in water
  - MU Yuguang, School of Biological Sciences, NTU
- Fold into α-helix twice during simulation

Cross Correlations Between Time Series



Dow Jones US economic sector indices

- Pearson
- Spearman
- Digital

$$C_{ij} = \overline{\left(\frac{x_i(t) - \bar{x}_i}{\sigma_i}\right)\left(\frac{x_j(t) - \bar{x}_j}{\sigma_j}\right)}$$

Vector Pearson Correlations

- Velocity time series $\mathbf{v}_i(t), \mathbf{v}_j(t)$
  - Means $\bar{\mathbf{v}}_i, \bar{\mathbf{v}}_j$
  - Covariances Σ_i, Σ_j
- Basis-independent square deviations

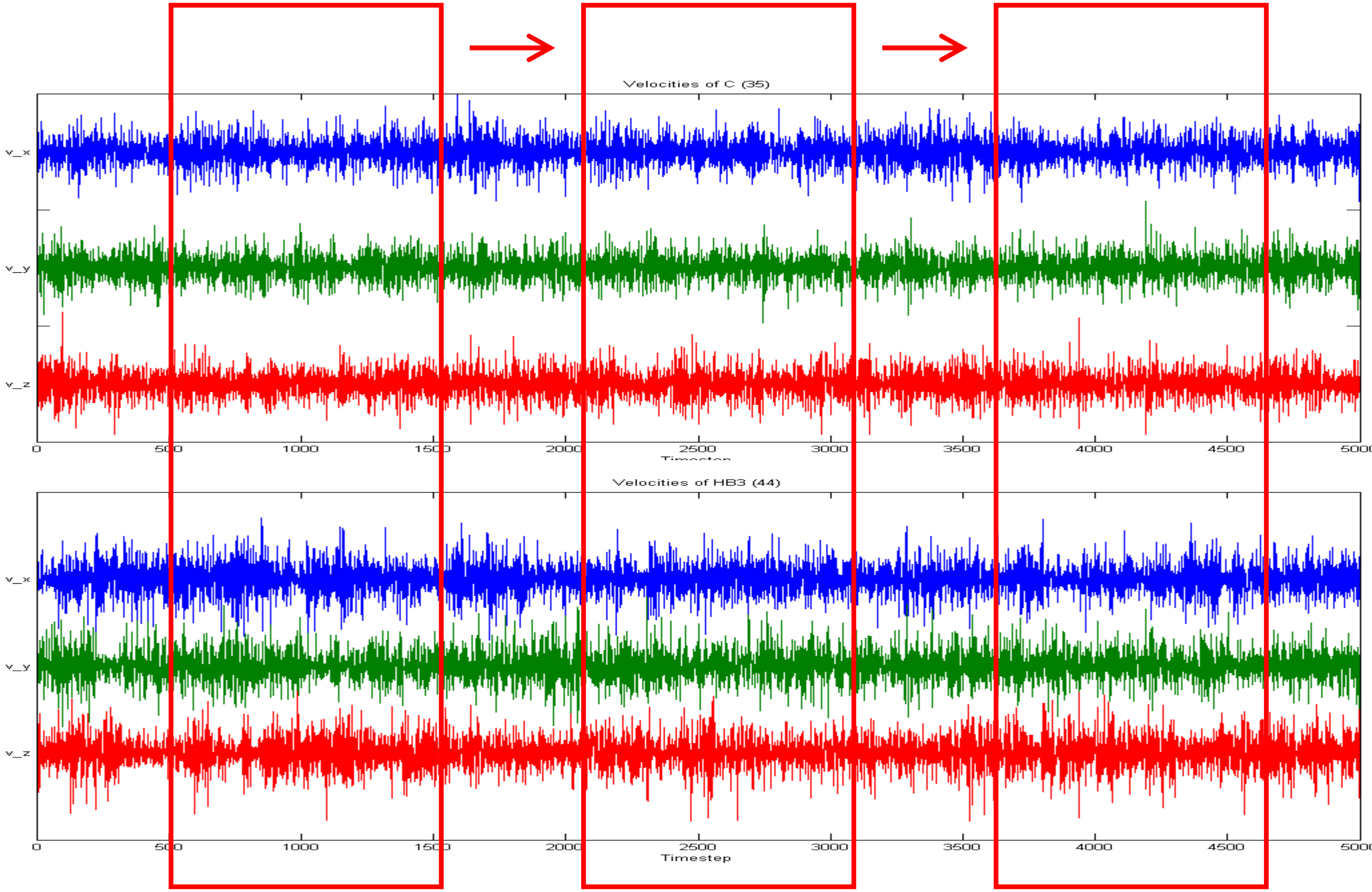
$$\left[\mathbf{v}_i(t)-\bar{\mathbf{v}}_i\right]^T \Sigma_i^{-1}\left[\mathbf{v}_i(t)-\bar{\mathbf{v}}_i\right], \left[\mathbf{v}_j(t)-\bar{\mathbf{v}}_j\right]^T \Sigma_j^{-1}\left[\mathbf{v}_j(t)-\bar{\mathbf{v}}_j\right]$$

- Scaled deviations $\vec{\xi}_i = \Sigma_i^{-1/2}\left[\mathbf{v}_i(t)-\bar{\mathbf{v}}_i\right],$

$$\vec{\xi}_j = \Sigma_j^{-1/2}\left[\mathbf{v}_j(t)-\bar{\mathbf{v}}_j\right]$$

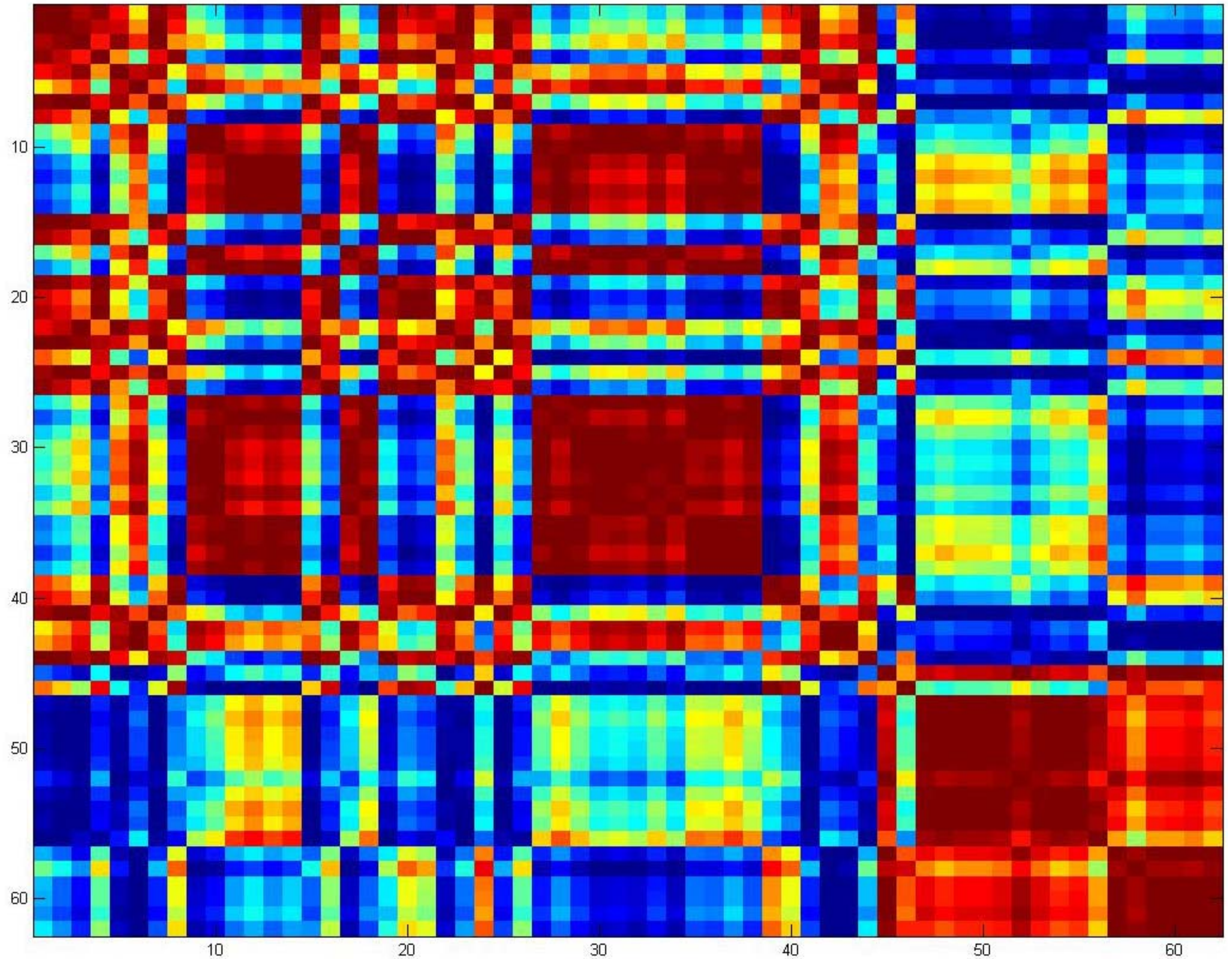
- Vector correlations $C_{ij} = \overline{\vec{\xi}_i \cdot \vec{\xi}_j}$

Sliding Windows



Correlation Matrix

(1250, 1500)



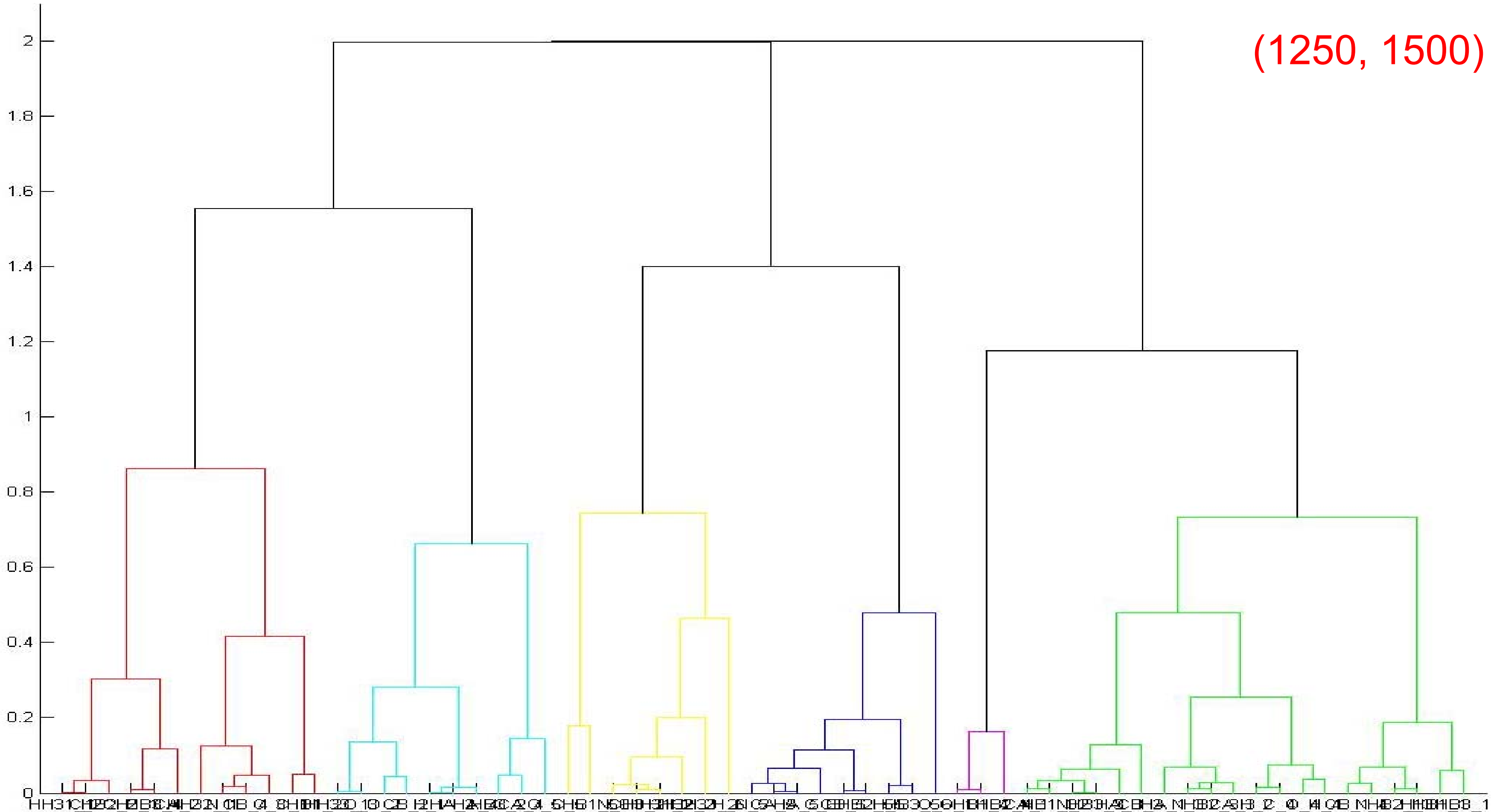
Hierarchical Clustering

- Complete linkage algorithm
- Pairwise distance

$$d_{ij} = \sqrt{2\left(1 - C_{ij}\right)}$$

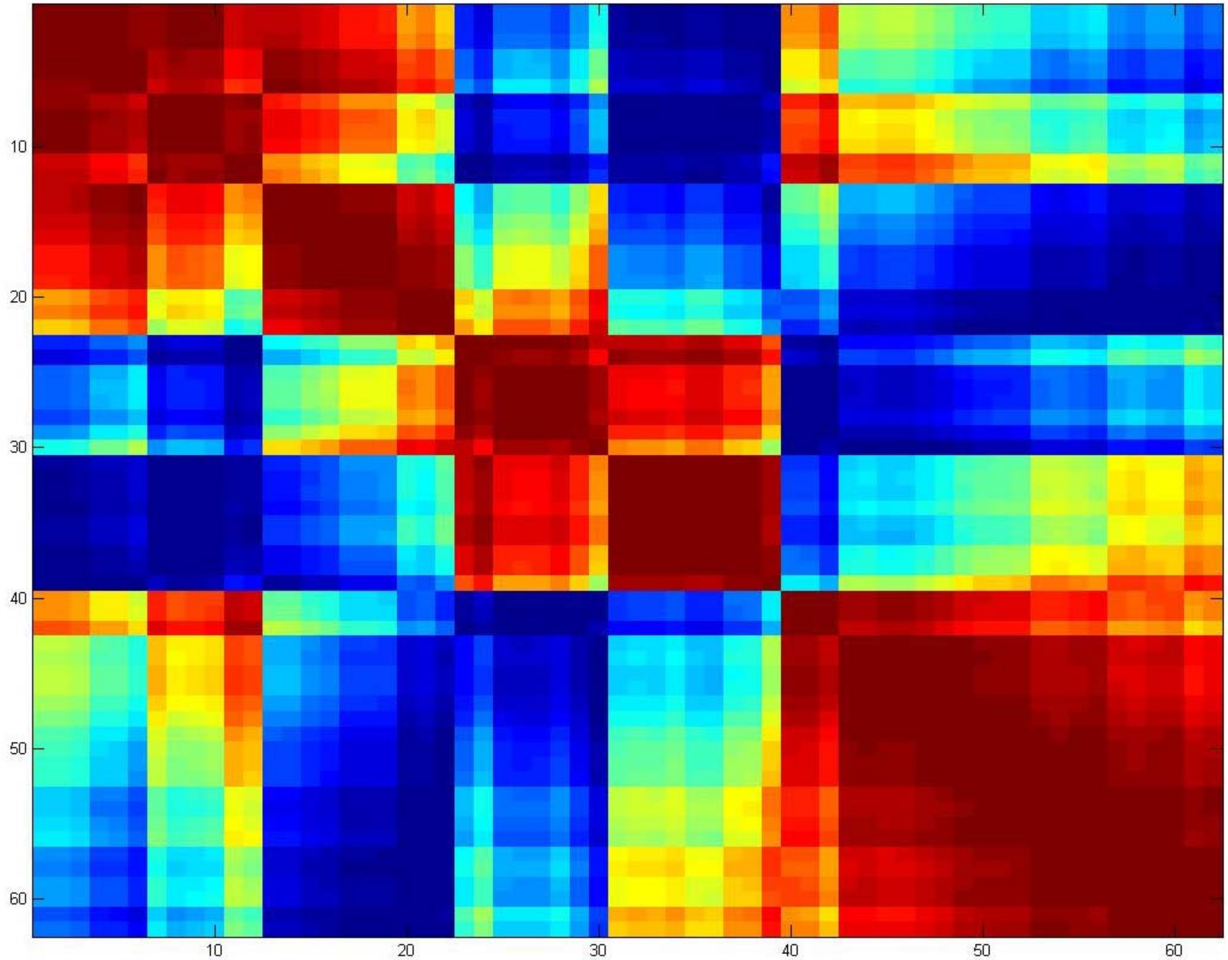
- Determine as function of time
  - Number of clusters
  - Composition of clusters
  - Thresholds of clusters

Complete Linkage Dendrogram

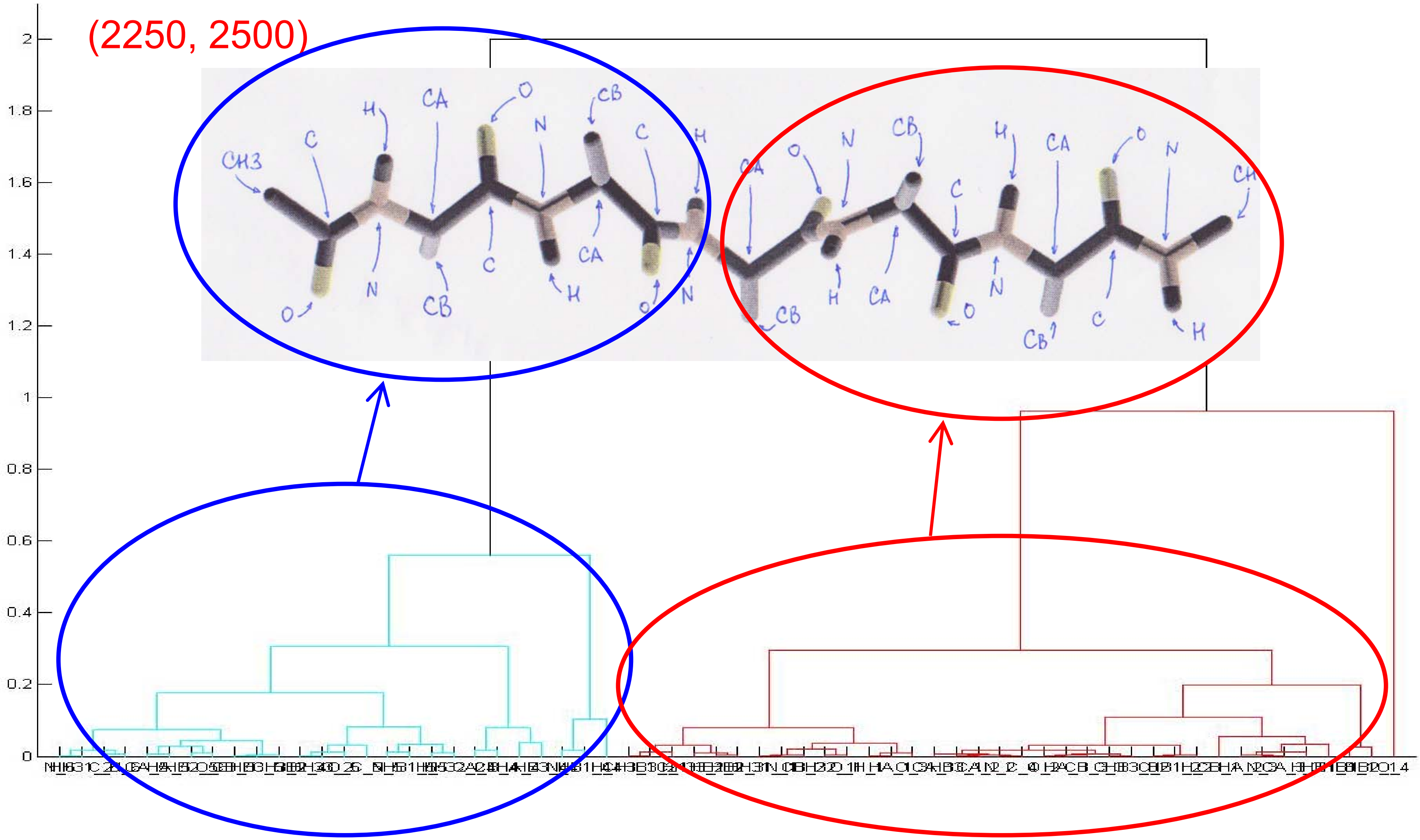


Reordered Correlation Matrix

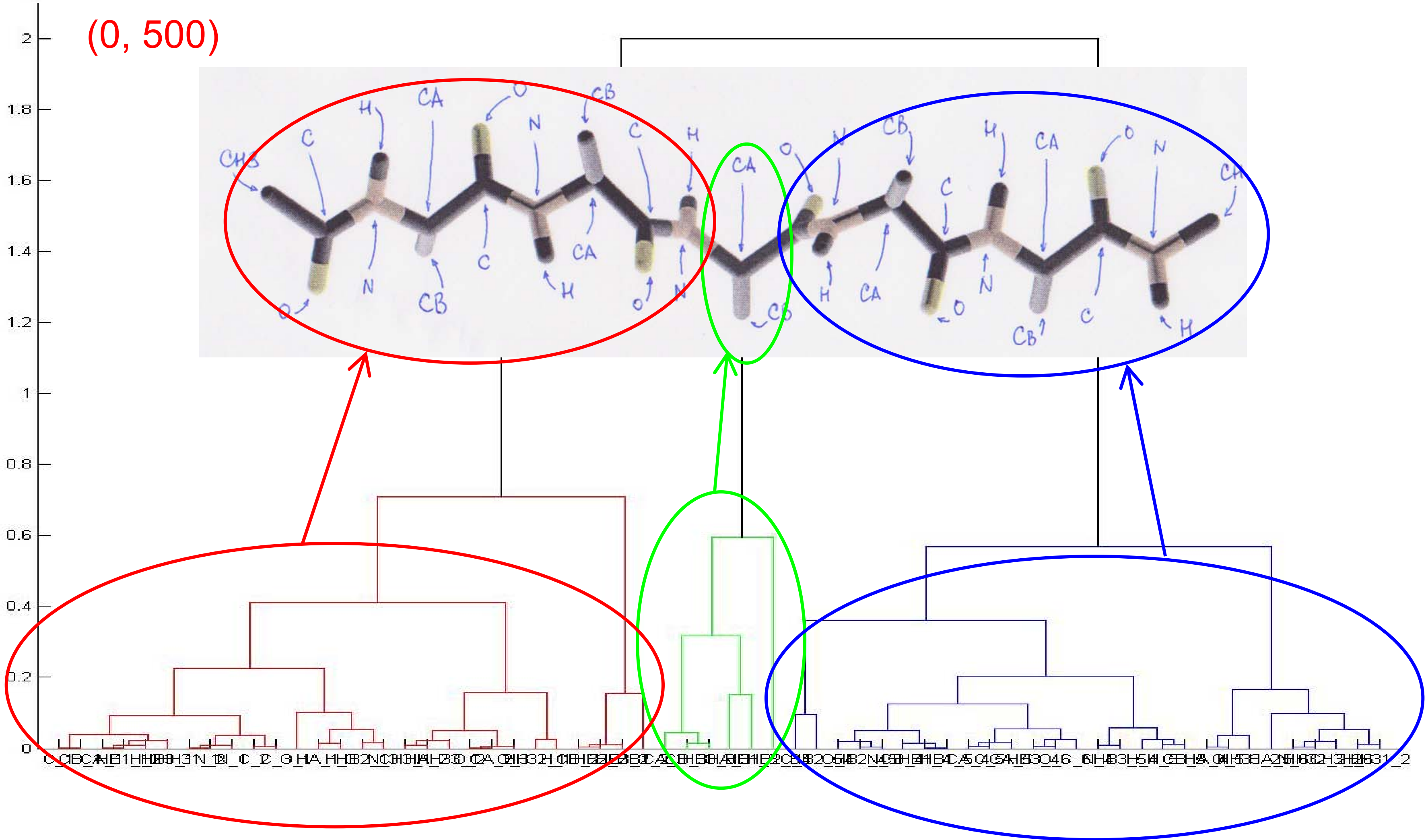
(1250, 1500)



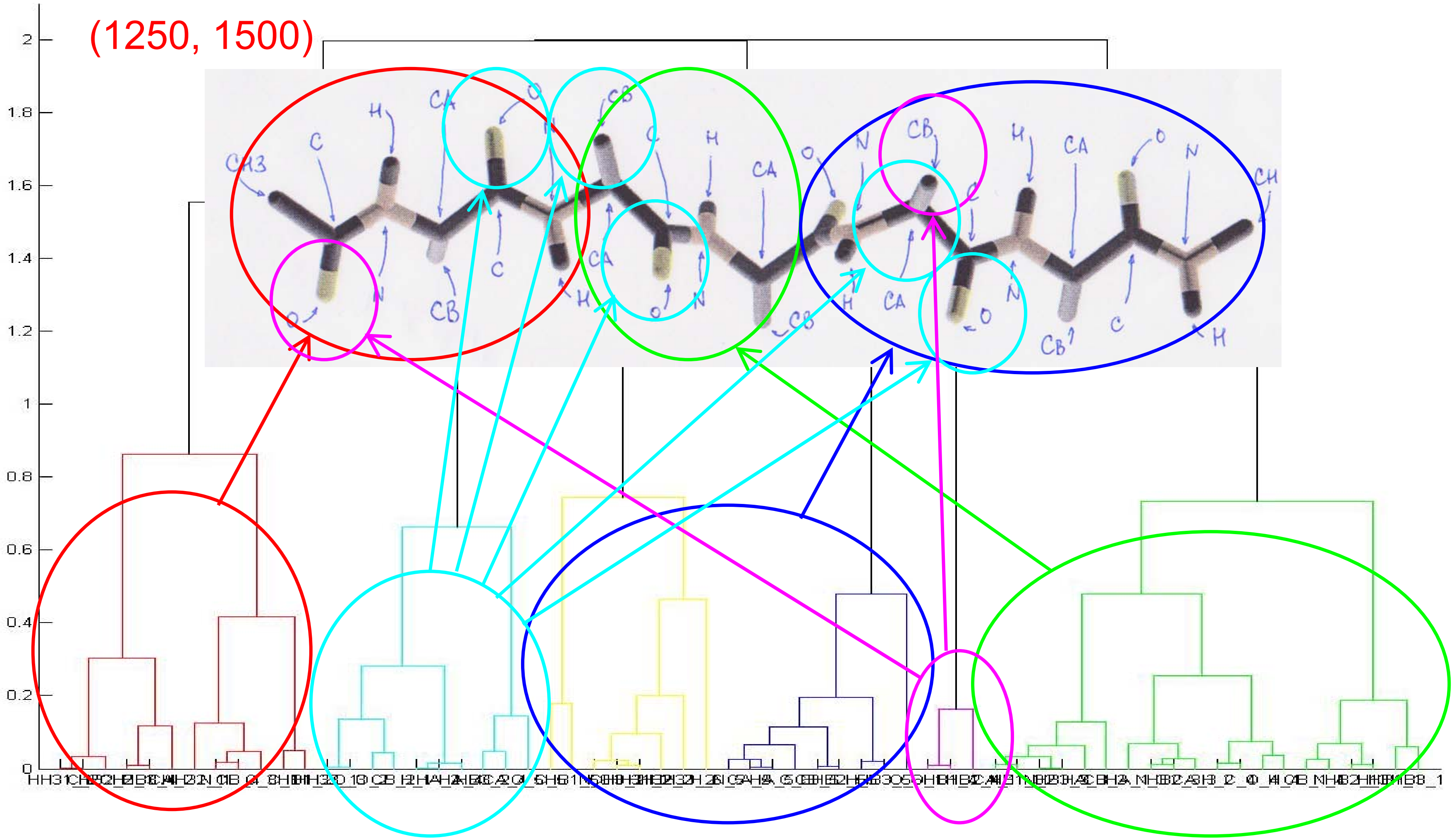
Effective Variables



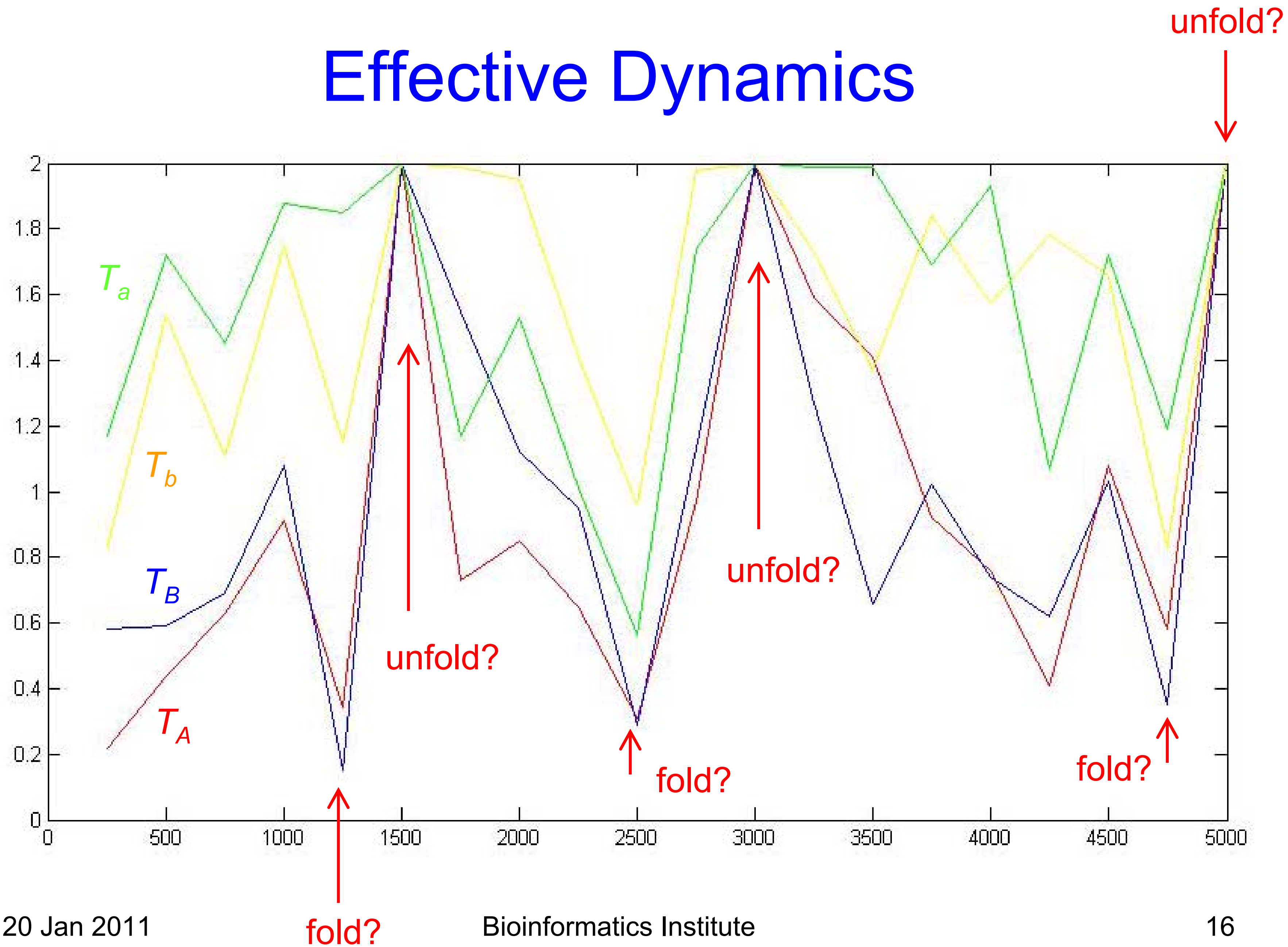
Effective Interactions



Effective Interactions



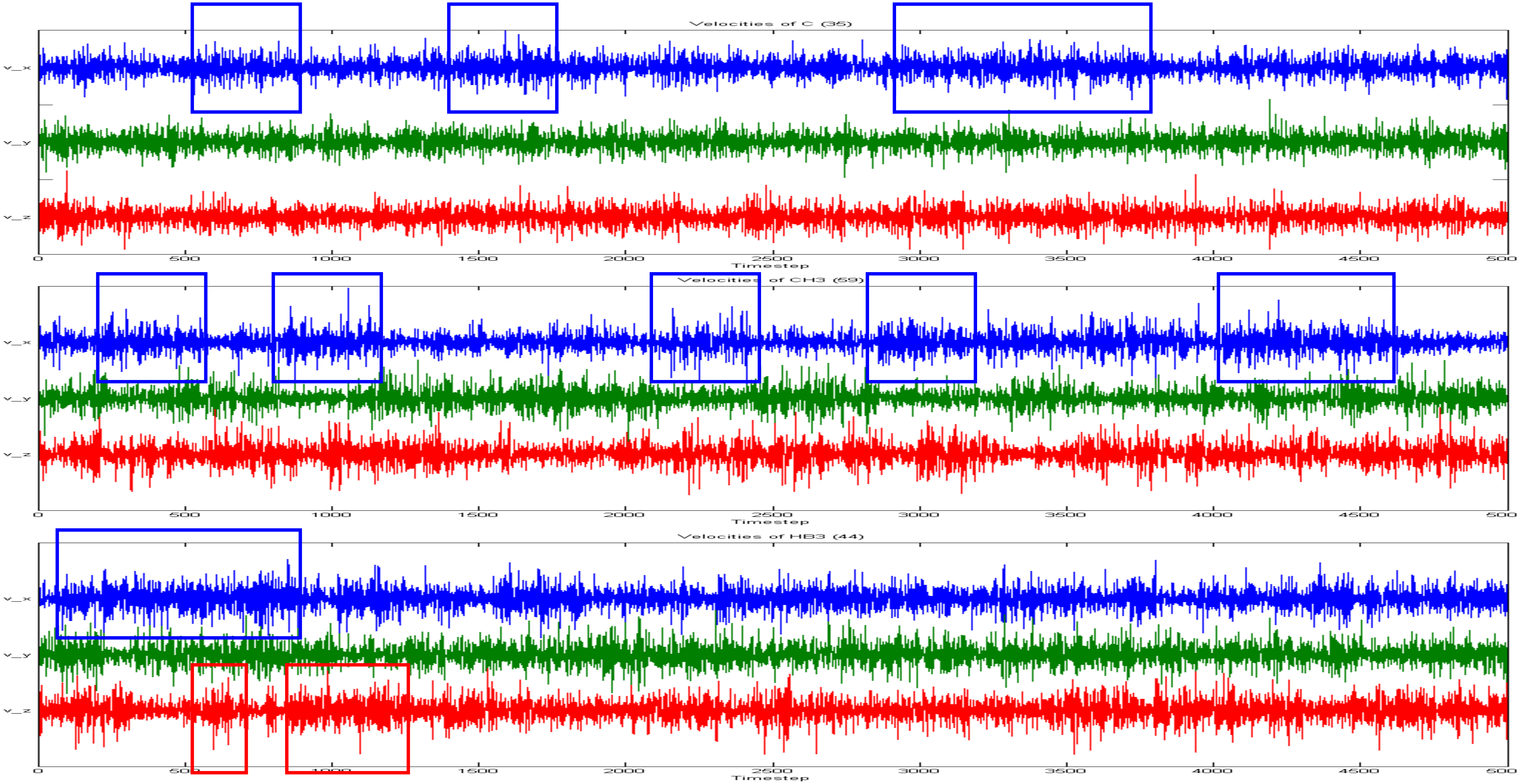
Effective Dynamics



Segmentation vs Clustering

- Time Series Clustering
  - Discover effective mesoscopic variables in given time window
  - Discover slow time evolution of effective variables by sliding time window
- Time Series Segmentation
  - Discover number/type of macroscopic phases
  - Discover lifetimes of macroscopic phases
  - Discover time scales of transitions between macroscopic phases

Nonstationarity in Time Series



Modeling Nonstationary Time Series

- Assume non-stationary time series
  - ($\mathbf{v}(1)$, $\mathbf{v}(2)$, …, $\mathbf{v}(t)$, …, $\mathbf{v}(N)$)
  - M stationary segments
  - In segment m, data points drawn from ($\boldsymbol{\mu}_m$, Σ_m) Gaussian distribution
- Recursive segmentation
  - One time series → two segments
  - Each segment → two subsegments
  - Iterate + optimize
  - Terminate

Jensen-Shannon Divergence

- Single-segment likelihood for (**v**(1), **v**(2), …, **v**(*N*))

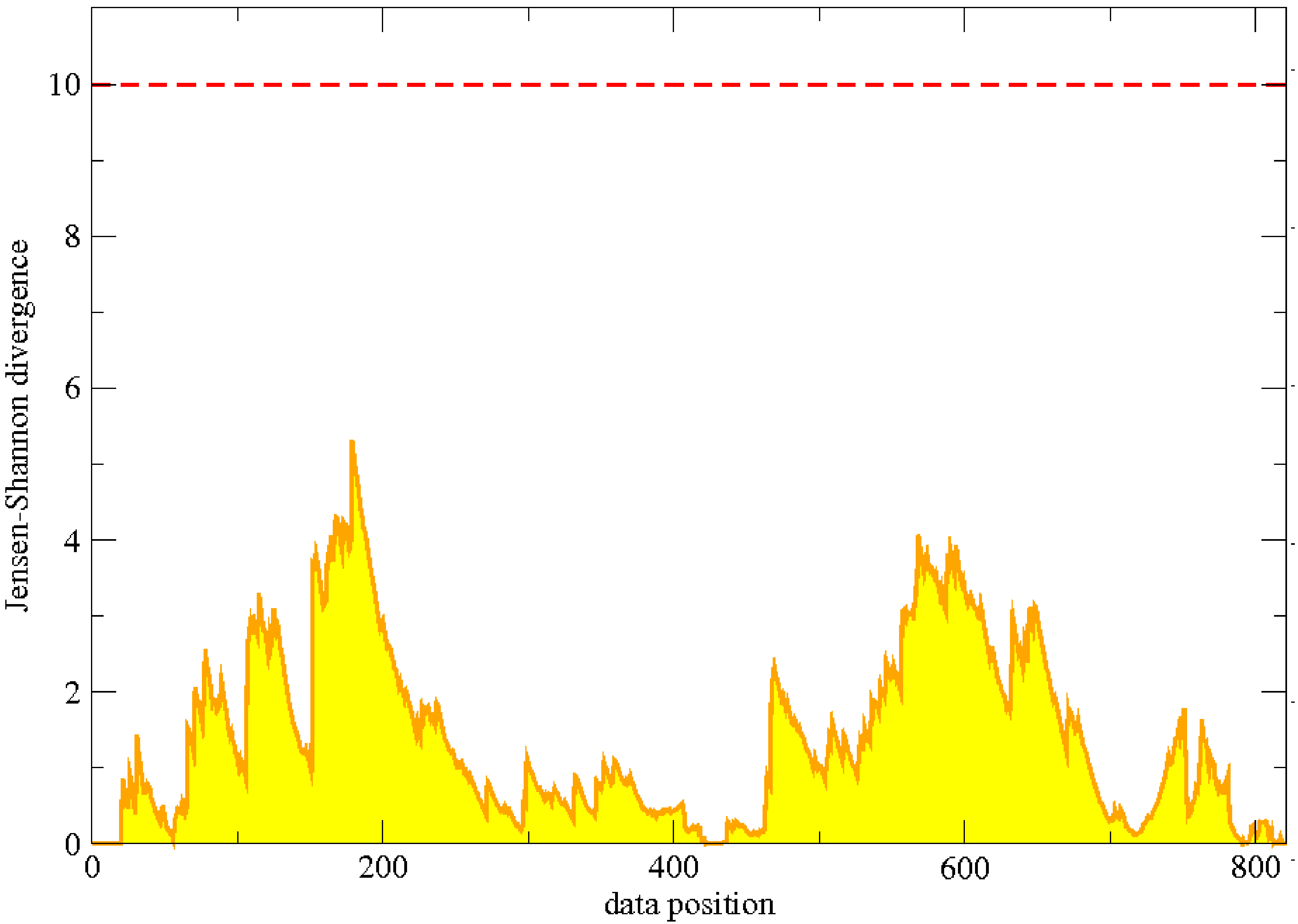
$$L_1 = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left\{-[\mathbf{v}(i)-\boldsymbol{\mu}]^T \Sigma^{-1} [\mathbf{v}(i)-\boldsymbol{\mu}]\right\}$$

- Two-segment likelihood for (**v**(1), …, **v**(*t*), **v**(*t*+1), …, **v**(N))

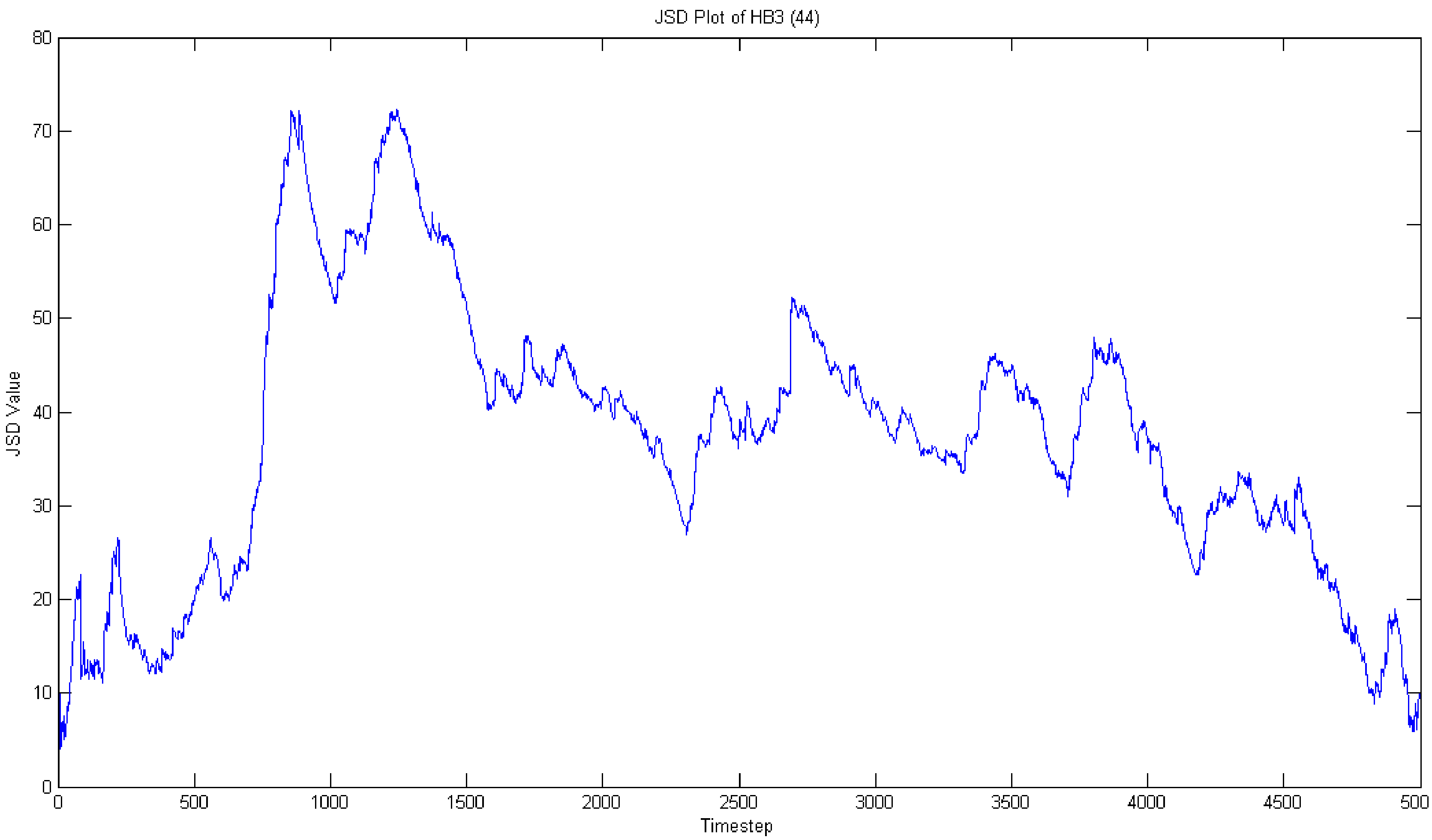
$$L_2(t) = \prod_{i=1}^{t} \frac{1}{\sqrt{2\pi|\Sigma_L|}} \exp\left\{-[\mathbf{v}(i)-\boldsymbol{\mu}_L]^T \Sigma_L^{-1} [\mathbf{v}(i)-\boldsymbol{\mu}_L]\right\} \prod_{i=t+1}^{N} \frac{1}{\sqrt{2\pi|\Sigma_R|}} \exp\left\{-[\mathbf{v}(i)-\boldsymbol{\mu}_R]^T \Sigma_R^{-1} [\mathbf{v}(i)-\boldsymbol{\mu}_R]\right\}$$

- ML estimates $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_L, \hat{\boldsymbol{\mu}}_R, \hat{\Sigma}, \hat{\Sigma}_L, \hat{\Sigma}_R$
- Jensen-Shannon divergence $\Delta(t) = \ln \frac{L_2(t)}{L_1} \geq 0$

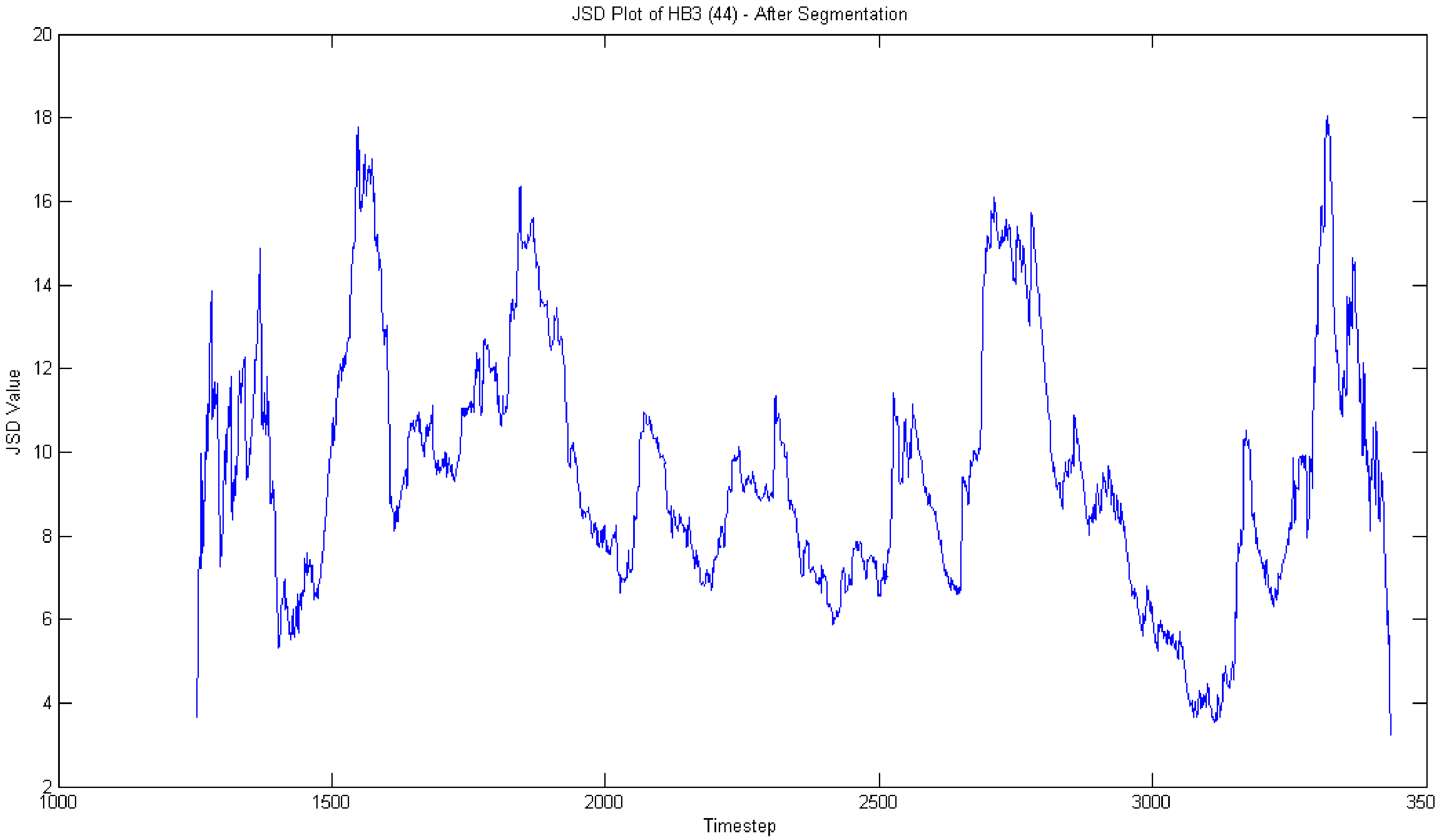
Recursive Segmentation



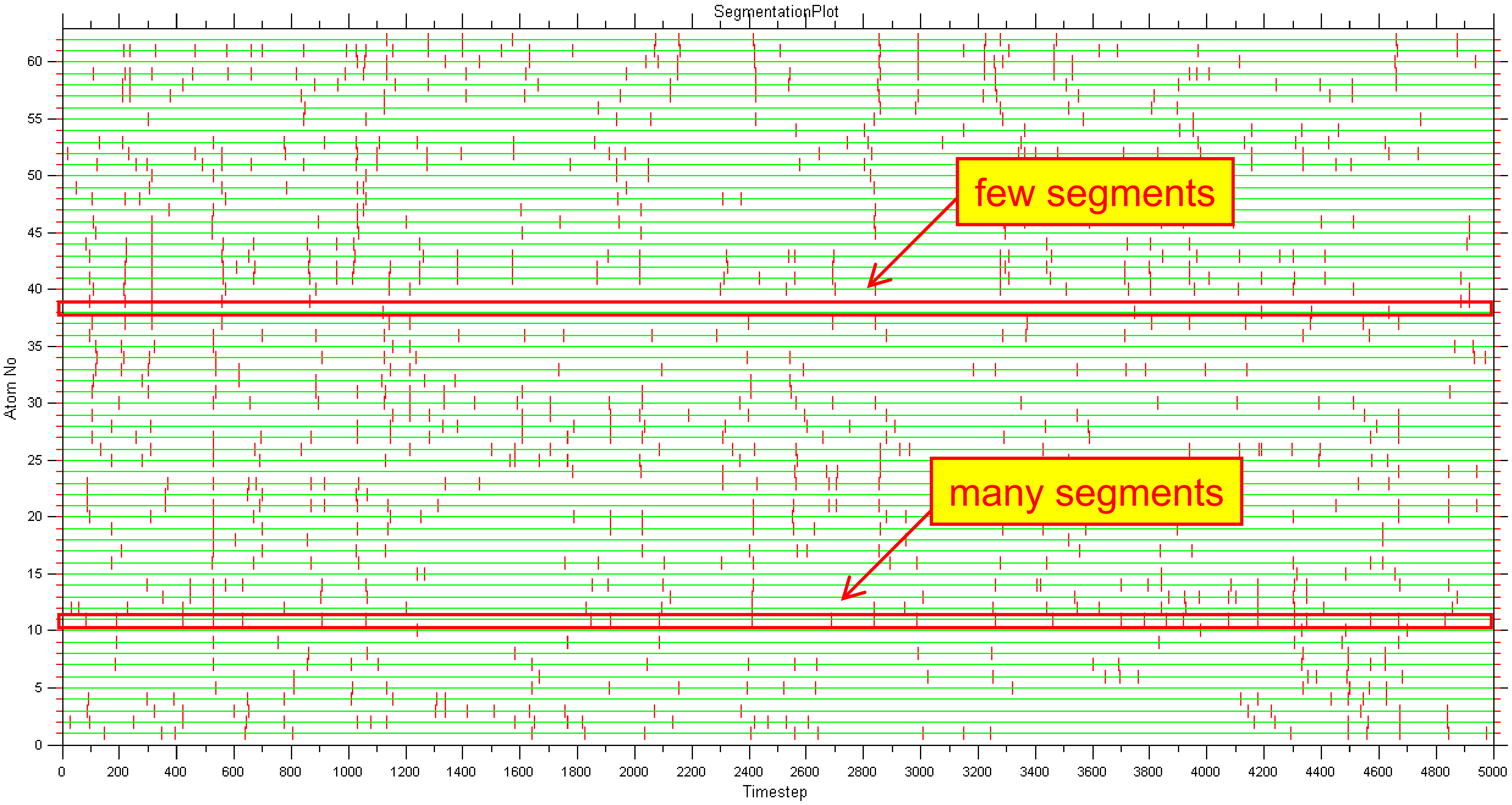
Start of Segmentation



End of Segmentation



Final Segmentations



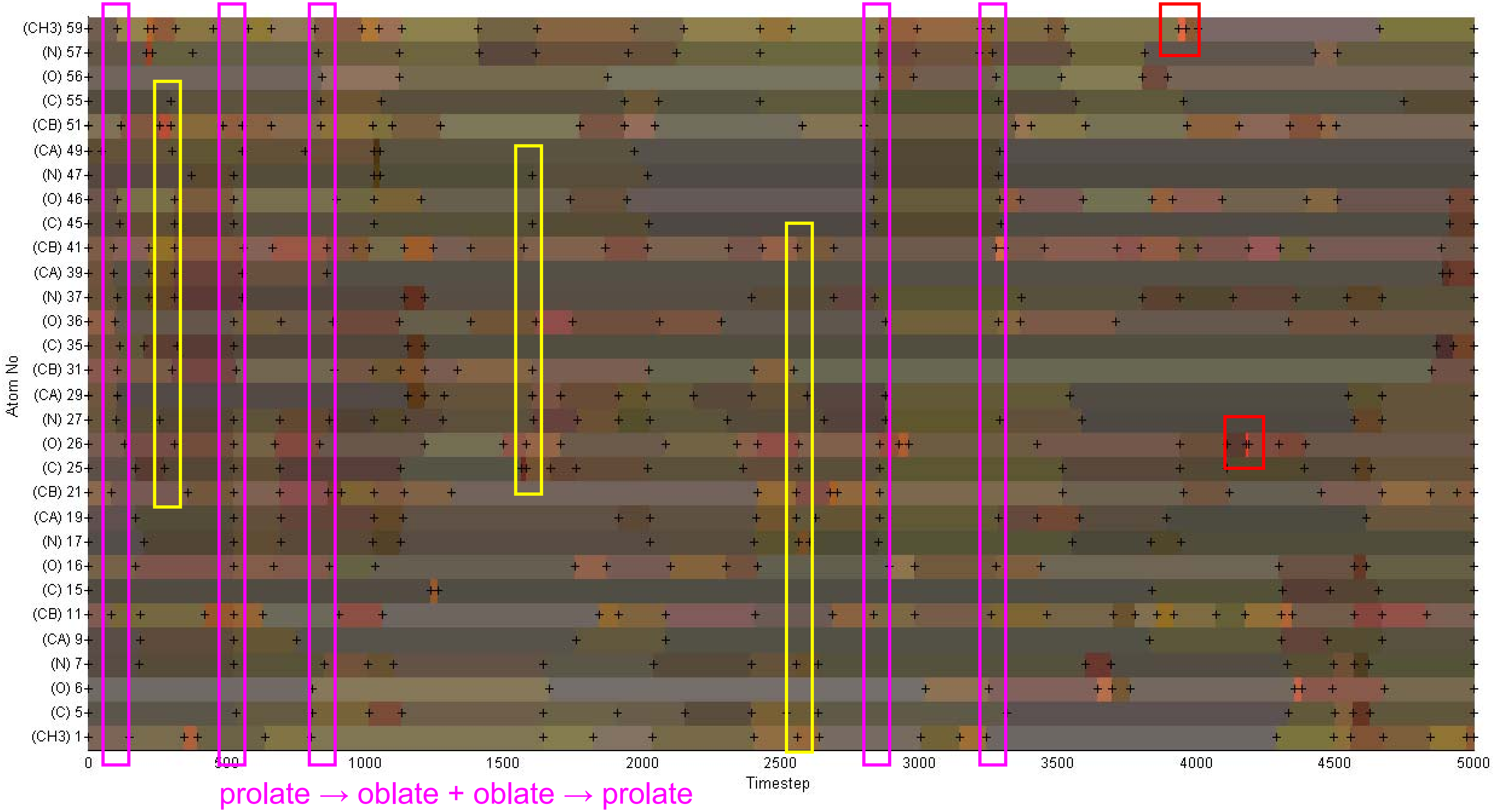
Fluctuations Ellipsoid and Color Map

- Velocity fluctuations in **v**(t) characterized by covariance matrix Σ
- Eigenvalues λ_1, λ_2, λ_3, semi-axes of fluctuations ellipsoid
- Color map for segments

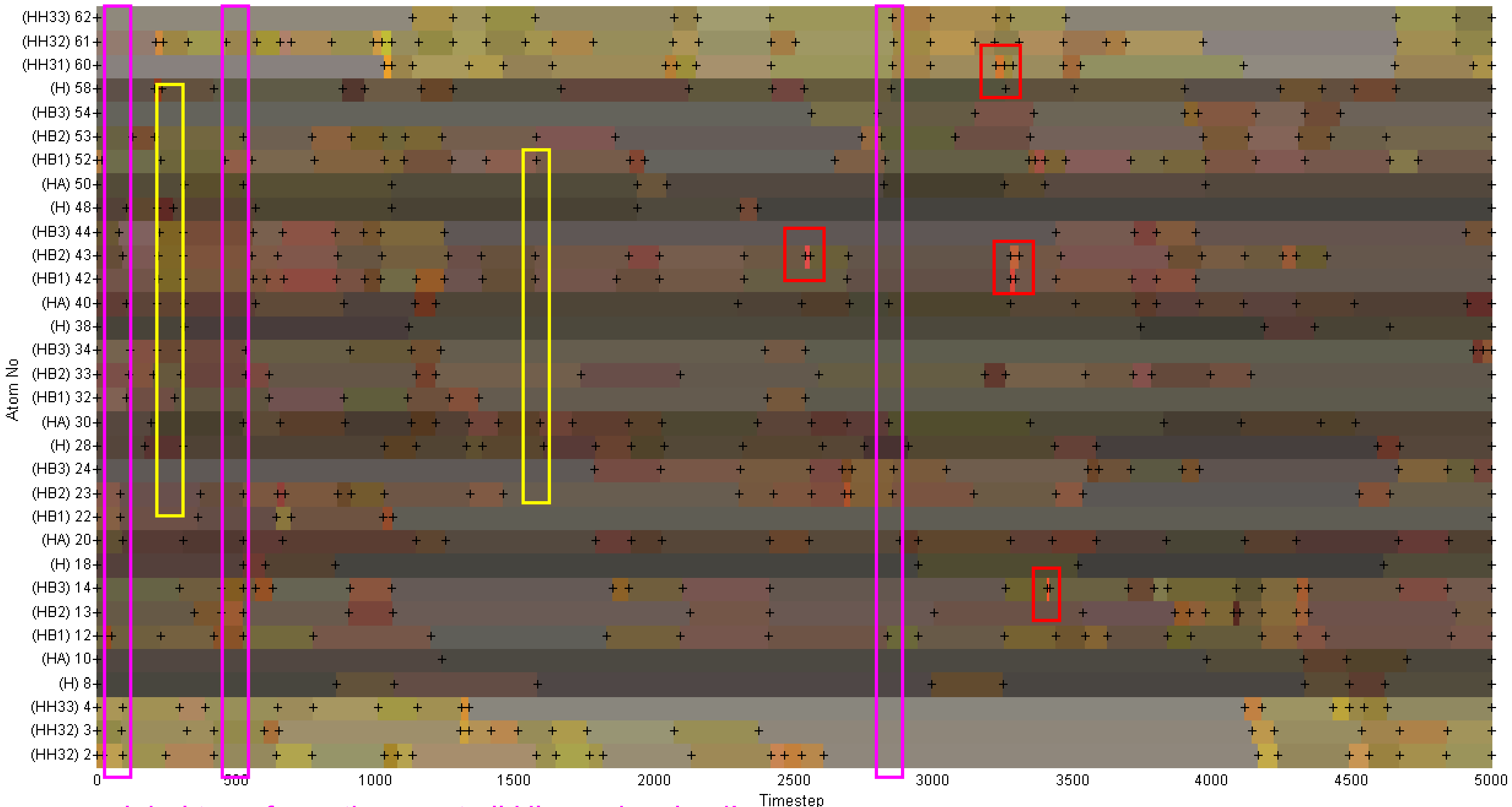
$$(R, G, B) = \left(\sqrt{\frac{\lambda_1}{\lambda_{1,\max}}}, \sqrt{\frac{\lambda_2}{\lambda_{1,\max}}}, \sqrt{\frac{\lambda_3}{\lambda_{1,\max}}} \right)$$

  - Gray = spherical
  - Reddish/purplish = prolate
  - Greenish = oblate

Temporal Distributions of Eccentricity



Temporal Distributions of Eccentricity



global transformations: not all H's are involved!

Conclusions

- Dynamics of small protein from microscopic time series
- Time series clustering
  - Two synchronized clusters
    - Interaction clusters
  - Effective dynamics from thresholds
    - Identify folding & unfolding events
  - Nucleation from midpoint of protein
- Time series segmentation
  - Precisely identified global vs local events
  - Changes in fluctuations ellipsoid
    - Potential to understand mechanisms
  - Nucleation from middle of protein

Acknowledgments

- Time Series Clustering
  - Mikhail FILIPPOV (PhD)
- Time Series Segmentation
  - Jeremy HADIDJOJO (PAP/4)

Thank You!