Clustering Approach to Protein Folding Problem and Dynamics of the Tropical Atmosphere

Mikhail Filippov

School of Physical and Mathematical Sciences

Nanyang Technological University

A thesis submitted to the Nanyang Technological University in fulfilment of the requirement for the degree of

Doctor of Philosophy

Singapore August 2015

I dedicate this thesis to my parents and to my fiancee. Thank you for all unconditional support along the way.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Assistant Professor CHEONG Siew Ann and to my co-supervisor Associate Professor KOH Tieh Yong for the guidance and the support of my research throughout my years as a PhD student.

Abstract

Many different multidimensional and multivariate data-sets are studied currently in almost all research areas. Usually, the aim of such studies is to reveal meaningful patterns and to gain an insights about special characteristics of the particular data-set. It is usually done by scanning over the elements of the data-set at large and bringing up groups or clusters of these elements that exhibit correlations across the variables or dimensions. In this thesis I use different techniques to address two problems: protein folding and the dynamics of the tropical atmosphere.

Biological research has generated vast quantities of protein sequences. One of the most important characteristics of the certain protein is its function. It is crucial to understand it in order to use protein for pharmaceutical, engineering or research purposes. Protein folding problems, which predict the spatial structure and the corresponding function of a protein from its amino acid sequence alone, are then critical in biological studies. In this work I study several proteins: start with a penta-alanine molecule, which consists of an acetyl cap group, a methylamide cap group and five residual alanine groups, I then study three polyalanine peptides. Using pairwise correlation, clustering techniques, minimal spanning tree and other techniques, I was able to identify the moments of folding and precursors of the folding event for some of the mentioned molecules.

In the Tropics seasonal variations are mainly governed by precipitation, since temperature and day length remain relatively constant throughout the year. The Madden-Julian oscillation (MJO) is the major tropical intraseasonal variability with extensive meteorological impacts, that keeps on puzzling the climate research community on both theoretical and modeling fronts. Together with monsoons, known as dominant water distributors, MJO is an important contributor to the rainfall cycle. In this work, I study dynamics of the Madden-Julian oscillation and its interaction with monsoons. I use satellite data collected by the Tropical Rainfall Measurement Mission (TRMM). Using clustering techniques I managed to identify MJO events, measure its parameters, show relation between magnitudes of MJO and monsoon.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Our world is complex: it can be envisioned as a network with connections between people, communities, countries, companies, cells, organisms. Such web would describe then the conditions our daily life on every level. Considering this complexity, it is difficult to imagine any interesting problem that could be addressed in separately, without taking into consideration the adequate model of both system integral components and their relationships. Information that is describing agents as well as context, environment of their interaction is buried in endless data-sets. Businesses, government organizations, scientific researchers, and individuals today are collecting them it in vast quantities: from news feeds and financial transactions to weather and traffic updates. Furthermore, things like email and instant messaging, address books, digital photos, and personal music and video libraries add to our growing individual data footprints. Managing all this data is becoming a real challenge. Disciplines like biology and medicine are rapidly transitioning into information sciences. Such techniques as, for example, large-scale DNA sequencing [1, 2, 3] or genotyping [4, 5, 6] allow researchers to easily obtain thousands of measurements of biological interest at once. Development of these technologies, as well as growing data pool make possible such projects as capturing human diversity in 1000 genomes [7]. However, often collected measurements lack obvious interpretations. As a result, advances are prodding biomedical discovery toward information management tasks. Other disciplines and fields are facing similar problems: growth in both quantity and data types generate the need to effectively organize, extract, retrieve, and interpret the information.

Machines, that help us collecting data, come to rescue again. People were interested in building machines that mimic human brain for a very long time. The development of machine learning is an attempt to recreate human's cognitive and problem solving skills and is an integral part of the development of artificial intelligence. One of the first important steps was the invention of the perceptron model in 1957 [8]. A decade later Marvin

Minsky analyzed the limitations of this model in expressing complex functions [9]. Hence researchers stopped using this model for another decade. Interest came back in 1980s with the introduction and development of the decision tree model [10, 11, 12, 13] and revival of neural networks [14, 15, 16]. Decisions trees and neural networks see wide applications from medicine [17, 18, 19] to business and finance [20, 21, 22]. With invention of the Internet, fast interaction on it and rapid growth of data, collected on it, both new and old methods and techniques have place. We truly see rise of machine learning as an important scientific field.

Usually machine learning methods divided into two groups:

- **Supervised learning**. The learning system starts with *labeled* instances and known desired output. During learning the algorithm gets feedback on its performance on the data and the human expert is usually involved. The goal of the algorithm then is to come up with general rules that map inputs to outputs. Supervised learning methods are widely used for multiple classes of problems from spam filtering [23] to ecological predictions [24].

- **Unsupervised learning**. The learning system starts with *unlabeled* instances and unknown desired output. The goal is to to discover structure of the data and to organize it in some way. Unsupervised learning could be a sole aim of the study (discovering hidden patterns in data) or one of the steps or approaches helping reach another goal. One of the most common unsupervised learning technique is cluster analysis. Its task is to find hidden patterns in data by grouping similar objects together. Unsupervised learning applications include sequence analysis [25] and clustering gene expression patterns [26] in bio-informatics; image segmentation [27] in medicine; and object recognition [28] in computer vision.

Apart from these two main groups there are also semi-supervised learning, reinforcement learning, transduction and others, depending on the taxonomy used.

In this thesis I employ machine learning methods to address two Big Data problems: protein folding and dynamics of the tropical atmosphere. These problems have been approached using many methods, including machine learning. For example, protein fold recognition [86] and prediction of the protein's secondary structure [85] by neural networks or protein classification by agglomerative hierarchical clustering [94, 95]. Meteorological applications include investigation of the dynamics of polar weather systems [87], study on temperature trends in the US [88], identification of the spatio-temporal patterns in climate data [89]. Machine learning techniques were also used to analysis and comparison of the

predictions from fully coupled circulation models [90] and for the development of a long range forecast model by the British Meteorological Service [91].

Main aims of this thesis are: 1) to establish a robust approach of the folding event identification, 2) to find out the reason(s) for the complete or incomplete folding of the particular protein molecule, 3) solve problem of identification and tracking of the Madden Julian Oscillation (MST) and 4) get an insight about interaction between MJO and monsoons.

Main papers produced during the course of this thesis are the following:

Mikhail FILIPPOV. Protein folding and transformations: insights form clustering techniques. International Conference on Computational Science (2011).

Mikhail FILIPPOV. Understanding Protein Folding Mechanisms: Machine Learning Study. The Institute of Physics Singapore Meeting (2015).

Mikhail FILIPPOV, Khipin CHUA, Jeremy HADIDJOJO, Jiali SHAO, Chong Eu LEE, Yuguang MU, Dawei ZHANG, Lock Yue CHEW, San Keong LAI, and Siew Ann CHEONG. Universal Correlational Fingerprints and Precursors for Protein Folding Into Alpha-Helices (in preparation).

Mikhail FILIPPOV, Siew Ann CHEONG, Tieh-Yong KOH. Slow variables in tropical atmospheric dynamics. (in preparation).

1.1 Clustering

The idea of grouping similar objects is not new. As people shape their environments, classifying objects within these environments makes this task easier. From city planning to organization of a personal library we are guided by groups, classes, their similarities and dissimilarities. The practice of classification of objects is the basis in many fields of science. Organization of data into meaningful groups and hierarchies is one of the most fundamental ways of understanding, learning and exploring.

One of the most popular machine learning approaches to group data is clustering. Basically, clustering is a method to identify a structure in a collection of unlabeled data. It is objectively organizing data into homogeneous groups with high similarity between the objects within one group and low similarity between the objects in different groups. The aim of cluster analysis is to explore a convenient and rational organization of the data-set, not to create a set of rules for separating particular data-set into categories. Clustering algorithms are aimed for finding a structure in the data.

Clustering techniques provide several advantages over simple manual grouping process. First, they can be used to apply a specified objective criterion consistently and automatically to construct clusters. People are especially good in detect patterns and groups in one, two

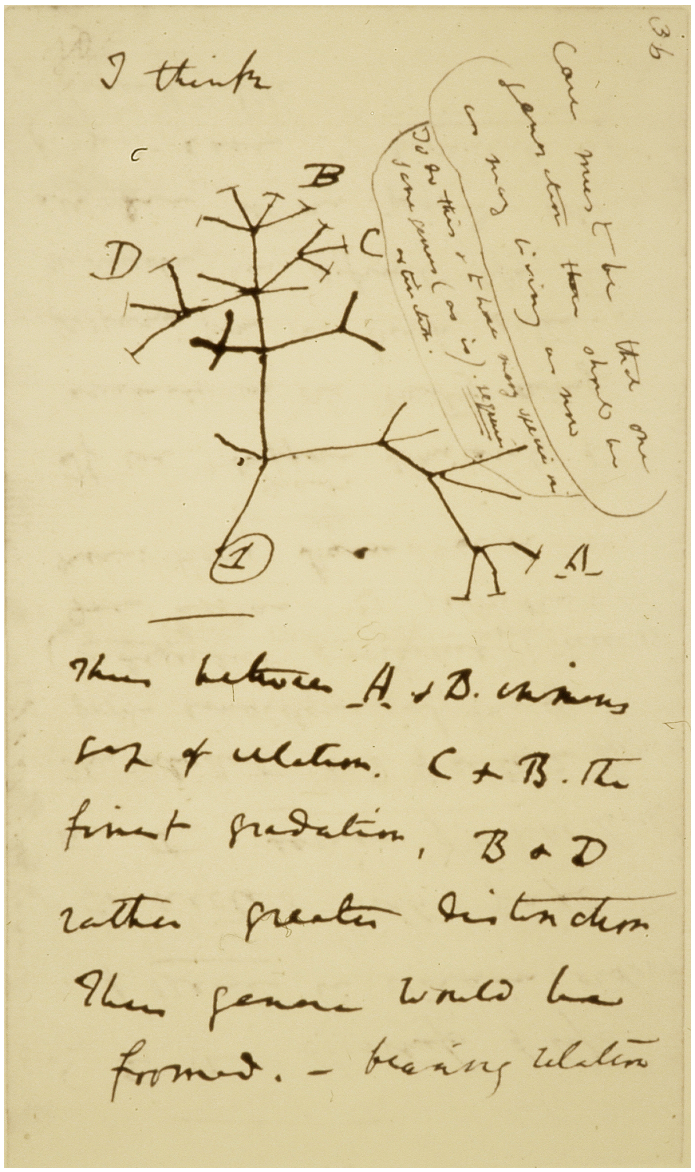

Figure 1.1: A sketch of the tree of life from Darwin's Notebook B. Dating from 1837. [114]

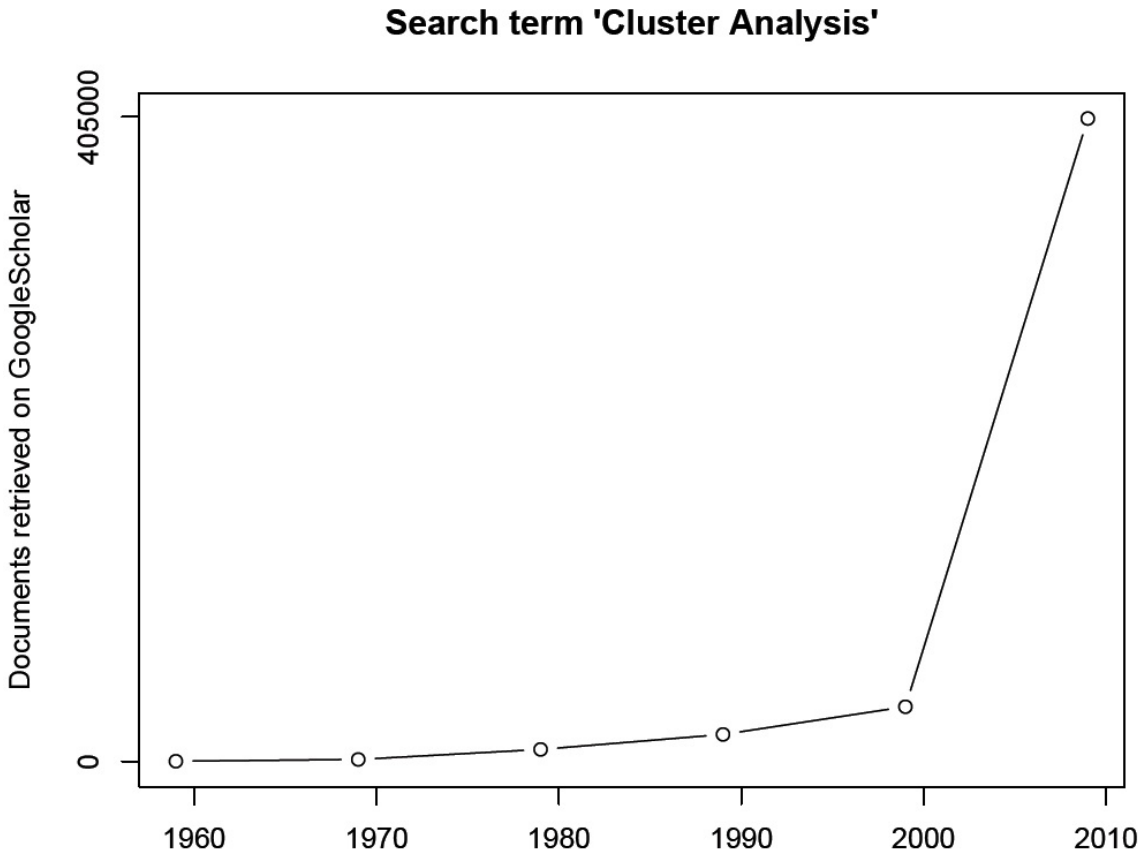

Figure 1.2: Google Scholar retrievals using search term “cluster analysis”, for the years 1950-1959, 1960-1960, etc., up to 2000-2009. (Data collected in September 2012.) [29]

and three dimensions, but different people may identify different clusters within the same data-set. The proximity measure defining similarity among objects depends on an individual's educational and cultural background. Thus it is quite common for different human subjects to form different groups in the same data, especially when the groups are not well separated. Second, a program with a clustering algorithm is capable of identifying clusters can in a fraction of time required by a manual action, especially if a list of attributes or is provided for each object. The consistency, speed, and constancy of clustering techniques in organizing data-sets together represent a good reason to use it. A clustering algorithm relieves a scientist or data analyst of the treacherous job of "looking" at a pattern matrix or a similarity matrix to detect clusters. A data analyst's time is better spent in analyzing or interpreting the results provided by a clustering algorithm.

Although papers on clustering techniques date as far back as 1960 [30, 31, 32], interest in this approach started to grow rapidly only after 2000 [Fig. 1.2]. General references with good overview on contemporary clustering techniques include [33-48]. Clustering approaches have been used in many disciplines: finance[238, 239, 240], medicine [49], chemistry [50] statistics [51], high-energy physics [52], computational biology [53, 241, 242] and others. Clustering is often used as a pattern recognition tool. A general introduction into this framework could be found in [54], while more statistical approaches are given in [55] and [56].

1.2 Protein folding

Proteins are the basic ingredients of all existing life forms. In Greek $\pi\rho\omega\tau\alpha$ (first), means "of prime importance" and it was first used to name proteins and describe them by Jöns Jacob Berzelius in 1838. Proteins serve not only as building blocks of cells and tissues, but are also very important for execution and controlling of many biological processes. Antibodies, transcription factors, enzymes, pumps and hormones are all examples from the long list of functions constantly performed by proteins in every living organism. However, a protein is only functional when it folds into a typical spatial structure, which is called the native state.

There are three main characteristics of a protein: structure, sequence and function. The structure of the protein describes its three dimensional state. The sequence of the protein represents the string of amino acids it includes. Finally, most important, but at the same time most ambiguous, is the function of the protein. it defines what role this particular protein will play in this particular organism. For most applications, such as genetic engineering, pharmaceutics or basic biological research it is crucial to understand the function of the protein. Protein folding problems, which predict the spatial structure and the cor-

responding function of a protein from the string of its amino acids alone, are then critical in biological studies. When proteins cannot fold correctly, there can be serious diseases caused. Alzheimer's, Parkinson's, mad cow diseases and many cancers are believed to be caused by the misfolding of proteins [57].

Primary or native state of the proteins is determined by the residue (sequence of amino acids). Amino acids may fold into local secondary structures including α-helix, β-sheet or non-regular coil [58, 59]. Regularly repeating elements of secondary structure are then packed into form tertiary structure and stabilized by hydrogen bonds, non-local interactions and side chain interactions between amino acids [60, 61 ,62] . The tertiary structures of several protein molecules can form complex that is known as quaternary structure by binding together.

Given a sequence of protein with 100 amino acids, and assuming that each residue can adopt two possible conformations, namely α-helix or β-sheet, the number of possible three-dimensional conformations of such a protein will be $2^{100} \approx 10^{30}$. The shortest time need for protein to make a change of conformation is of 1 picosecond order. Therefore the folding by random search in the conformation space will take 10^{18} seconds. However, most proteins fold in the order of milliseconds to seconds. This paradox was first described by Levinthal [70]. Therefore, there must be a conformational information stored in the primary structure of proteins which drives the protein toward the native state. Ever since Anifsen [71] first showed that proteins are sometimes folding in vitro without any other help such as folders or shapes, the self-assembly property of proteins with a small amount of atoms have been fascinating for scientists for almost half a century. The protein folding problem - how does a protein with certain amino acids sequence fold into the specific 3D structure? - have been the subject of extensive theoretical and experimental studies.

Those proteins that have tertiary native structure do interact within one cell to execute different kinds of biological functions, like storage, transport, mechanical support, coordinated motion, control of growth, immune protection, enzymatic catalysis, generation and transmission of nerve impulses [63]. Extensive biochemical experiments [60, 62, 64, 65] have shown that structure of a protein dictates its functions. Therefore, the identification of the structure of a protein is critical for understanding its function and for this reason it is one of the central problems in medical and biological sciences.

In order to predict the unknown function of a certain protein, biologists usually try to find proteins with known functions that resemble unknown protein and then deduce its function from their properties. This method, known in statistical learning theory as a "Nearest-Neighbor" is considered as one of the simplest. However, it is at the same time one of the most powerful and widely used prediction methods for supervised learning [92, 93]. If

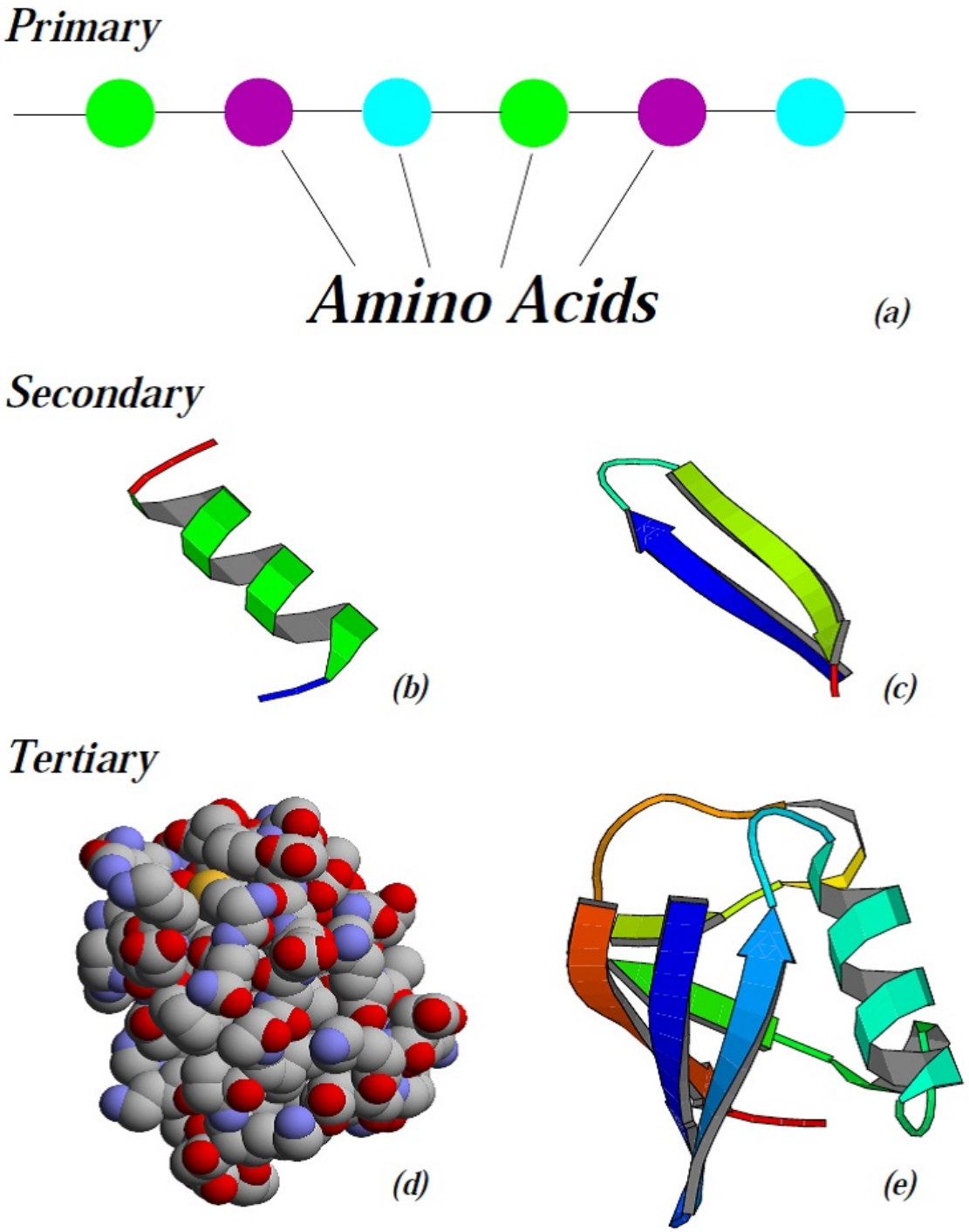

Figure 1.3: Illustration of the hierarchical composition of proteins. (a) Primary structure of the protein is the amino acid sequence of polypeptide chain. Different segments of protein form secondary structures: (b) α-helix and (c) β-sheet. These secondary structures are stabilized by hydrogen bonds between peptide backbones. The tertiary structure appears when several secondary structure elements are packing into the compact spherical units. Here we can see protein ubiquitin presented in the space-filled (d) and cartoon (e) representations.

data-bases of proteins are systematized into families, clusters or groups in a way to capture protein similarity, “Nearest-Neighbor” method can be also reinforced by protein classification or clustering methods.

Several experimental approaches are mainly used to determine protein structure: X-ray Crystallography [66, 67] and Nuclear Magnetic Resonance (NMR) spectroscopy [68, 69]. Although these experimental methods can provide high-resolution structural information about some proteins, computer simulations can be used to reveal important information that cannot be obtained experimentally.

Another way to anticipate protein’s native structure is by calculation from “first principles”. However its accuracy and reliability are inferior to comparative methods. Thermodynamic hypothesis states, that protein’s native structure of a protein is at the global minimum of free energy [66, 67], and can be further approximated by the global energy minimum. Predicting the native state of a protein can then be interpreted as a global minimization procedure. Most of the existing global optimization techniques can be used for the protein folding problem.

The Protein Data Bank (PDB) has currently approximately 30,000 proteins (with determined structure) deposited in the library [72]. Such a rich and diverse library of protein structures provide valuable data and helps to find out how exactly certain protein folds into its unique three dimensional structure and how we can foresee it from protein’s sequence [73].

During the last two decades one can observe an explosive growth in different protein databases. It is mostly due to appearance of effective protein sequencing techniques and as a result a large gap between sequence data and functional information has been created. Biological function of up to one half of sequenced proteins is still unrevealed.

In this thesis we are trying to find possible precursors of the protein folding event by studying evolution of the molecule during the simulation.

1.3 Madden Julian Oscillation

In the temperate regions, atmospheric dynamics is already well understood [74, 75, 76]. Quasi-geostrophic theory and its contemporary encapsulation in potential vorticity thinking give good results in extra-tropical regions. The main reason why these approaches work so well are relatively small frictional and diabatic effects on the short scale and quasi-balanced motion over a large range of scales. Due to these reasons, the dynamics of fundamental processes such as Rossby wave propagation and baroclinic instability are well understood. Furthermore, the evolution of the atmosphere in middle and high latitudes can be predicted

quite accurately by virtue of the steep energy spectrum of quasi-geostrophic turbulence. The dynamics of the tropical atmosphere is however poorly understood. In contrast, most phenomena that occur in the tropical atmosphere are neither quasi-balanced nor adiabatic. This is the reason why the tools work well in extra-tropical meteorological studies inappropriate for the Tropics. Finally, most of the belt between the Tropics of Cancer and Capricorn is covered by ocean, and its atmosphere was poorly observed until the advent of earth observation satellites. During the last decades, important advances have been made in data mining for meteorological applications [77, 78, 79, 80, 81]. However, it is still difficult to produce reliable analysis of the variables. Advanced data assimilation techniques work well in higher latitudes, but give questionable results in tropical regions. Additionally, the influence of convective and mesoscale phenomena complicates the analysis. In addition, data mining methods may unintentionally display outcomes which appear cogent and meaningful but which do not actually predict future behavior and could not be reproduced on another data set or new sample. Unlike the extra-tropical regions with strong variations in day length, temperature and precipitation, in the Tropics temperature and day length stay relatively constant during the whole year and seasonal variations are largely governed by precipitation. They are in turn, mainly influenced by the tropical rain belt (a portion of the Hadley cell) and with some geographical variations by El Nino-Southern Oscillation, the Indian Ocean Dipole, the Madden-Julian Oscillation and Monsoons. Tropical rainfall is the main spreader of heat through the atmospheric circulation. Understanding of the precipitation variability is crucial for our insight into tropical atmosphere dynamics. In addition to its effect on the climate, rainfall is the major source of fresh water and is crucial for human consumption, agriculture and industry. In this work we will investigate the dynamics of the Madden Julian Oscillation. This weather pattern is an important contributor to the rainfall cycle [82, 83, 84]. We will also examine its interaction with monsoons, that are known as a dominant water distributor.

The dynamics of the tropical atmosphere is regulated by many variations of different scale. The largest element in the intra-seasonal (30–90 days) variability is Madden-Julian Oscillation (MJO)[97, 98, 99, 100]. It is a large-scale coupling of patterns in the deep convection and atmospheric circulation. It is characterized by variations in such essential oceanic and atmospheric parameters as ocean surface evaporation, sea surface temperature (SST), rainfall, cloudiness and the speed and direction of the lower and upper level winds. The MJO first appears in the West Indian ocean as a positive convective anomaly. It propagates eastwards over the warm equatorial waters towards the maritime continent (where convection weakens) and further into the West and Central Pacific (convection strengthens again). The tropical rainfall pattern almost fades away over the East Pacific ocean, where

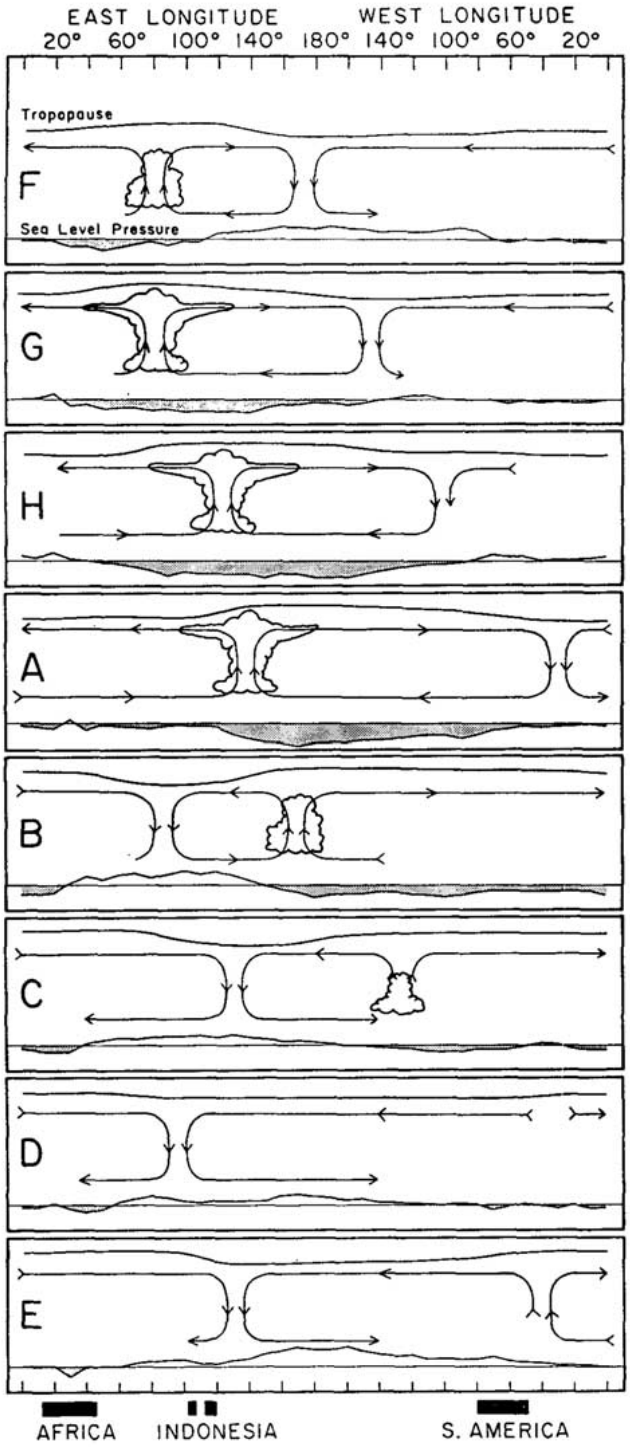

Figure 1.4: Illustration of the large-scale features of the Madden-Julian Oscillation (from top to bottom) life cycle along the equator. The mean zonal wind distribution is forming the circulation. The cloud symbols represent the convective center. The curves above and below the circulation represent perturbations in the upper tropospheric and sea level pressure. From Madden and Julian [97].

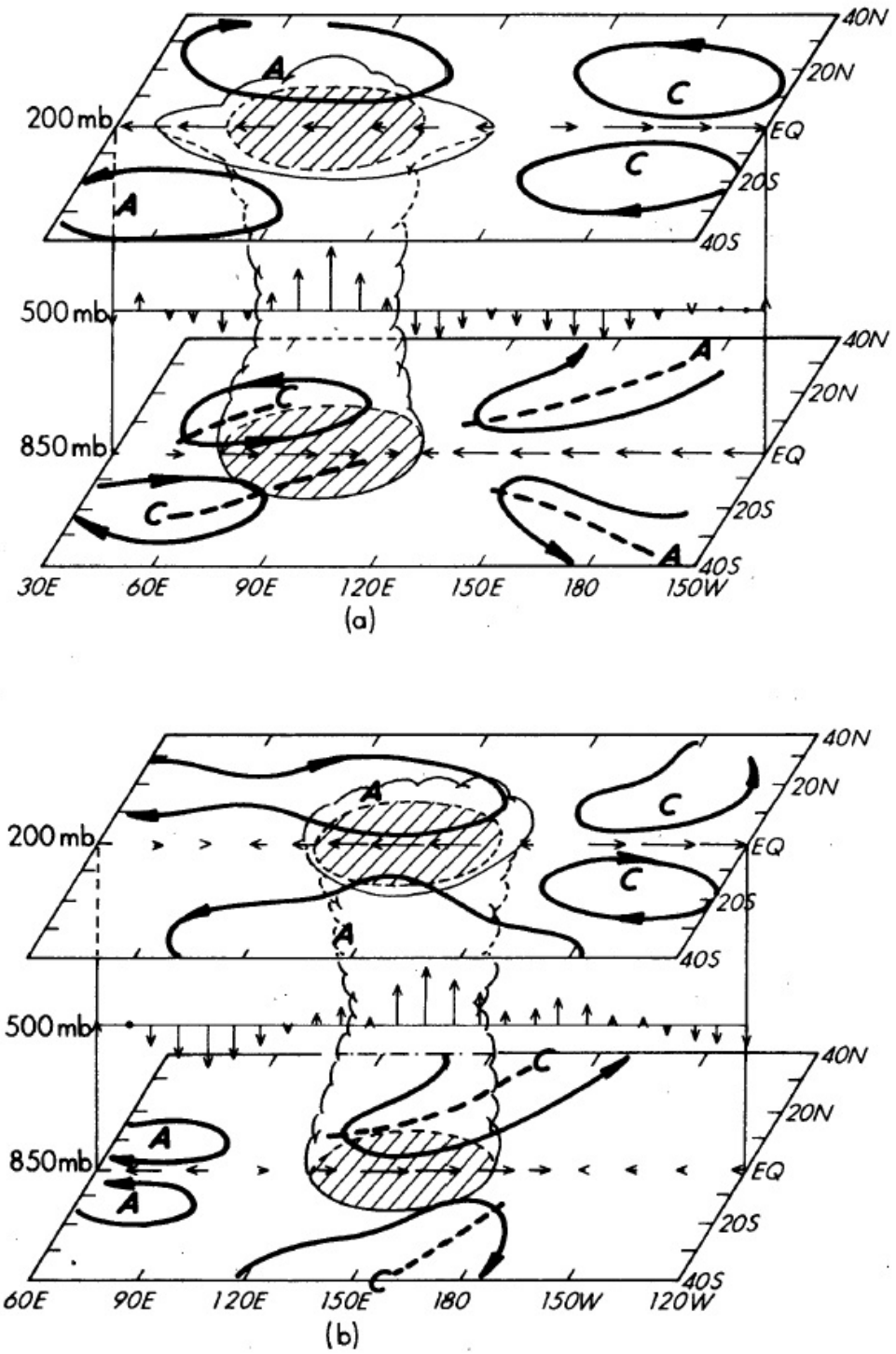

Figure 1.5: Schematic diagram of the vertical three-dimensional structure of an established MJO with anomalous convection center (shaded region) approximately passing through 90° E (a) and 150° E (b). The arrows represent zonal winds and vertical velocity anomalies. Areas C and A represent cyclonic and anticyclonic circulation centers. Black arrows represent wind direction and rising (sinking) motion. From Rui and Wang [96].

water is cooler, however it usually appears again later over the tropical regions of Atlantic ocean and Africa. Thus, the circulation travels and completes the whole circumference of the globe [99]. Figure 1.4 depicts vertical cross section of the MJO along the equator and its propagation eastwards around the global tropics.

The three dimensional structure of the MJO is illustrated on figure 1.5. First, there is an active convection in the Indian Ocean and over Indonesia: it forces anomalous easterlies (westerlies) to leave the enhanced convection area in the upper levels of the atmosphere Cyclonic gyres, in turn, leave the areas of suppressed convection behind in both hemispheres. Secondly, anomalous westerlies (easterlies) at low levels are visible behind (ahead) the region of the enhanced convection. Upper levels gyres are usually stronger than the low level gyres. Finally, with propagation of the such coupled pattern towards the central Pacific, the upper and lower level circulation anomalies are becoming less observable and consistent, however they still play a critical role in redistribution air and water masses over the Tropics.

In 1966, before the discovery of MJO, Matsuno [104] derived the theoretical model of equatorial waves based on a divergent barotropic model. These waves could be summarized as westward propagating Rossby waves, eastward propagating Kelvin waves, westward propagating mixed Rossby and inertio-gravity waves, western and eastern propagating inertio-gravity waves. MJO does not fit well into Matsuno's theory, since his waves are too small and propagate too fast. From the very beginning MJO has added complications to the prior theories and yet is crucial for our understanding of the tropical circulation. Liebmann et al. [109] were the first to establish this relationship for the Indian and western Pacific basins, noting that cyclonic vorticity and divergence anomalies westward and poleward of the MJO circulation cells seem to be the driving forces behind increased tropical cyclone activity. Courtney [110] showed that of the 18 tropical cyclones that formed in the southeast Indian Ocean and South Pacific between December 2002 and June 2003, 13 could be associated with convectively active phases of the MJO. Hall et al. [111] found that an active MJO, along with a well-established monsoon, contributed to the development of several typhoons in the Australian basin, and also noted that the relationship between the activity of tropical cyclones and MJO was even stronger during El Niño events. Hendon and Liebmann [107] studied an evolution of the monsoon's intraseasonal oscillation and its interaction with MJO in Australia. Yasunari [106] found a dominant periodicity of about 40 days in the cloudiness over the Asian monsoon area, indicating a relation between the break and active phases of the monsoon and the MJO. Krishnamurti and Subrahmanyam [112] presented northward migration of ridges and troughs at 850 hPa at periods of 30-50 days over India. However, not all the intraseasonal variabilities in monsoon regions are associated with the MJO. For example, Goswami [113] concluded for the Indian monsoon that although a fraction of its

intraseasonal variability is associated with the MJO, other independent northward moving disturbances represent a significant portion (up to 50%) of the variability.

Collectively, these studies support our belief that the Madden-Julian Oscillation is not only a major contributor to weather variability in the global tropics, but also extensively influencing other inter- and intra-seasonal oscillations in the region. However, there is no theory for the MJO and other intra-seasonal oscillations so far. Although we can observe its patterns, we do not know its origin on a fundamental level.

1.4 Structure of the thesis

The plan of this thesis is as follows:

- In **chapter 2** we give an overview of data sets for both case studies and main methods, that are used in our work.
- In c**hapter 3** we analyze the folding process of different sets of proteins.
- In **chapter 4** we study the atmospheric dynamics of Madden Julian oscillation and its interaction with monsoons.
- In **chapter 5** we summarize key contributions of this work, discuss open problems and future directions, and offer some concluding remarks.

Chapter 2

Data & Methods

In this chapter we give an overview of data sets for both case studies and main methods, that are used in our work. We introduce data-sets for single protein molecule (in low and high resolution) and three protein molecules. Overview of the Tropical Rainfall Measurement Mission and the details of the data, provided by it is given later in the chapter. Finally, main methods, used in this thesis are introduced: correlation, clustering analysis, minimal spanning tree.

2.1 Data

One of the hottest question currently in the experimental area is what is the role of the specific amino acid in the protein folding process and how the final stabilization in the native state occurs. It was shown by site-directed mutagenesis experiments [149, 150, 151] that the energy gap between unfolded or misfolded and folded is filled by the free energy of some amino acids and it is possible to name them. However, the complicacy of such experiments makes it very difficult to use them widely in different studies on the of individual amino acid's energetics.

Computer simulations could serve as a suitable replacement for such a sophisticated and complex experiments. With the main goals of identification of the kinetic role of individual amino acids in a protein and prediction of the thermodynamics of them, large number of computer experiments [152, 153, 154, 155, 156] have been conducted. However, in such experiments the folding time is too long to be obtainable at the atomistic level since the conformational space of the protein is multi-dimensional and known as very complex [148, 157, 158, 159]. The fastest known proteins fold in the millisecond time scale, while the shortest vibrational mode performs on a femtosecond time scale, since it corresponds to the atomic covalent bonds' perturbation.

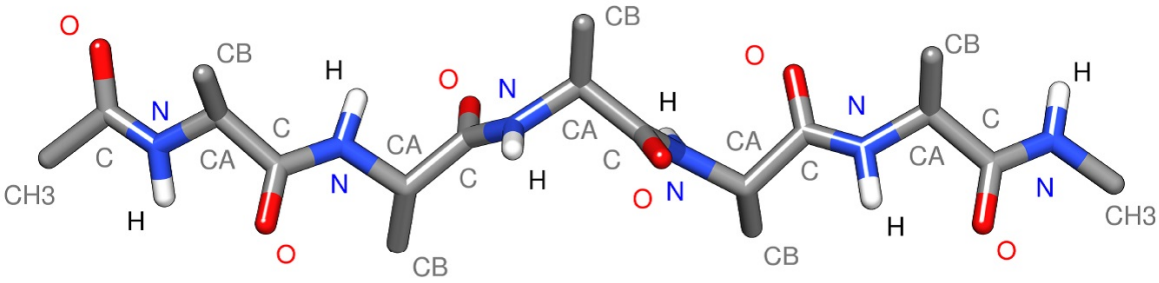

Figure 2.1: Penta-alanine (ALA-5) protein molecule structure at the beginning of the simulation. The CH3-, α-, and β-hydrogen atoms are not shown for clarity.

2.1.1 Protein data

In the first part of our work we study the dynamics of the protein molecule in order to identify the folding event. For this purpose we analyze data from molecular dynamics simulations.

2.1.1.1 Single protein

At first we study time-series that describe a penta-alanine molecule, which consists of an acetyl cap group, a methylamide cap group and five residual alanine groups. In total, there are 62 atoms. This data-set was produced by our collaborator Assistant Professor Mu Yaguang from the School of Biological Sciences (SBS), NTU [244].

The peptide consists of five residues of alanine, with a methyl group (-CH3) cap on the front end and a methylamide (NME) cap on the other end, shown in Figure 2.1. All these main parts of the molecule are shown in Table 1 below.

i	Group	Abbreviation
1-6	Acetyl cap group	Ace1
7-16	Alanine residue	Ala2
17-26	Alanine residue	Ala3
27-36	Alanine residue	Ala4
37-46	Alanine residue	Ala5
47-56	Alanine residue	Ala6
57-62	Methylamide cap group	Nme7

Table 2.1: Seven groups of penta-alanine molecule

The low resolution protein folding simulation data contains data of ALA-5 protein simulated in water for 5.0 ns duration. The names and coordinates (x_i, y_i, z_i) of all 62 atoms are

saved in 1.0 ps steps, resulting in a data of 5,001 data points concatenated in a single Protein Data Banks(.pdb) file.

The high resolution time series data of the same ALA-5 protein in water for 5.0 ns duration consists of frames, taken every 0.1 ps. Data-set contains a snapshot of the positions and velocities of every atom at that time. There are in total 50,001 frames concatenated one after another in a single Gromos87 .gro file format.

2.1.1.2 Three proteins

Another data-set was obtained by our collaborator Assistant Professor Zhang Dawei from Division of Chemistry and Biological Chemistry, School of Physical and Mathematical Sciences, NTU. This data-set of the peptides' time series was originally used to study the electrostatic polarization effect on the forming of α-helices from amino acid sequences (Fig. 2.2). By using AHBC in AMBER03 force field [147], it was shown that the α-helix folding is simulated by the electrostatic, at the same time, formed hydrogen bonds remained stable. This showed good agreement of theoretical and experimental results.

We use this data-set to study three polyalanine peptides. Their sole difference is in residues three and eight . The full sequences of the protein (Fig. 2.3) are the following:

Q: Ac-(AAQAA)2-GY-NH2

K: Ac-(AAKAA)2-GY-NH2

D: Ac-(AADAA)2-GY-NH2

with A standing for Alanine, G for Glycine, Y for Tyrosine, N for Nitrogen and H for Hydrogen; while Q, K and D being Glutamine (polar charged), Lysine (positively charged) and Aspartic acid (negatively charged). Using AHBC, average α-helical contents were found to be 0.783 for Q, 0.752 for K and 0.520 for D

This molecular dynamic simulation dataset contains the coordinates (x_i, y_i, z_i) of each atom in the peptides, spanning across 25.0 ns. The time series consists of 25000 steps, each step being 1.0 ps in a single .pdb file. An example showing the format of a portion of a PDB file is shown below in table 2.2. In order to illustrate the structure of the peptides we use the software VMD 1.9.1(Visual Molecular Dynamics) downloaded from "http://www.ks.uiuc.edu/Research/vmd/".

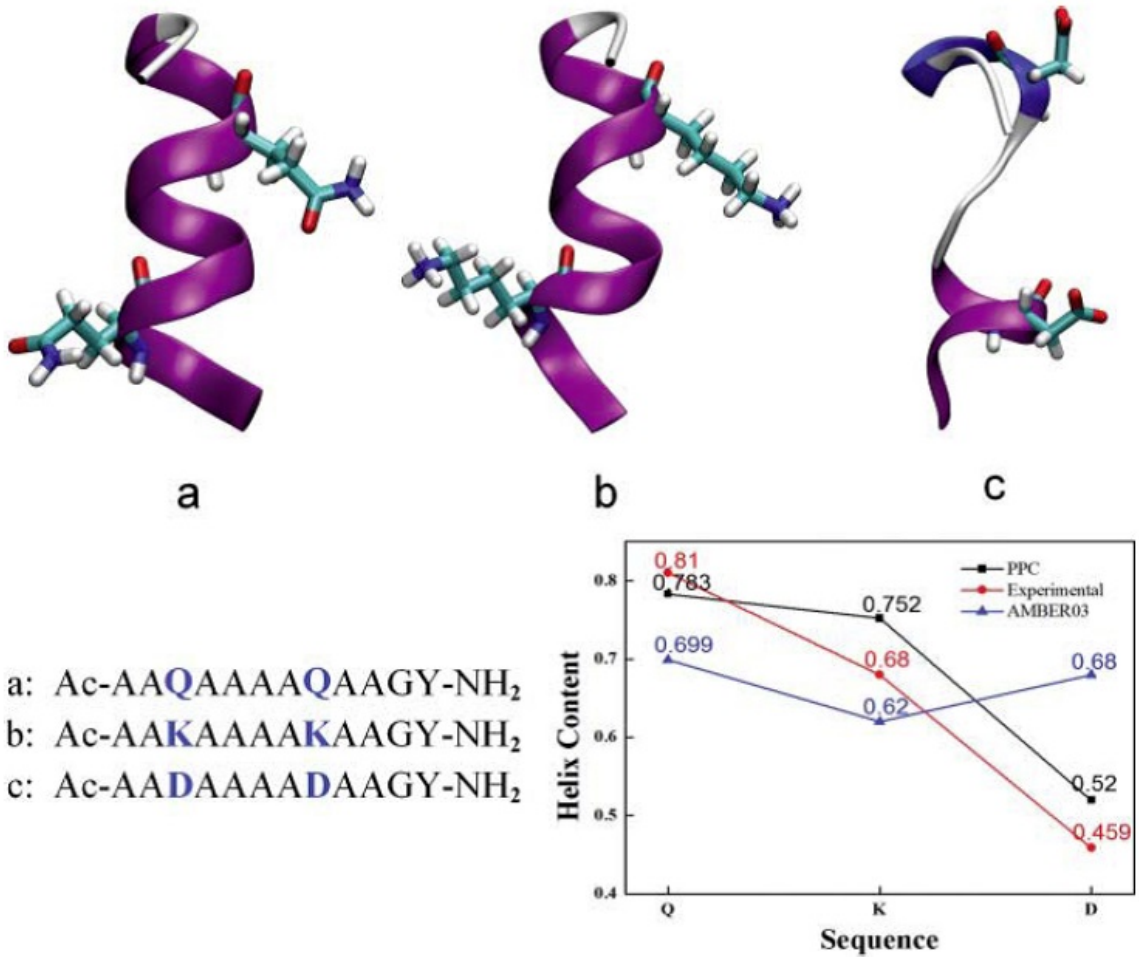

Figure 2.2: Lower left: Compositions of the predicted structures for three polyalanine peptides. The structures represent the geometry with the lowest free energy during the simulation under the AHBC. Label sites are the guest amino acids. Right: α-helical fractions for all peptides obtained from CD spectra and the predicted α-helical fractions based on the simulations under AHBC and AMBER03 charge. From Dawei [147].

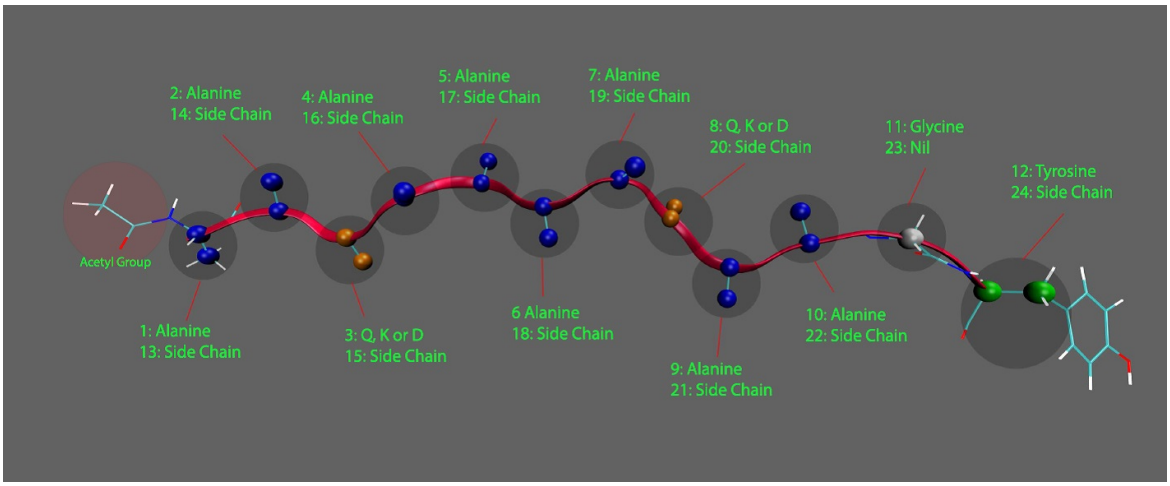

Figure 2.3: Illustration of the peptide chain in its initial state. The backbone of the chain is represented as a the red ribbon. Different kinds of amino acids are coloured differently. Plotted with VMD.

Atom	Index	Chemical Name	Amino Acid	Residue Index	x	y	z
Atom	27	N	ASP X	4	15.196	8.029	3.826
Atom	28	H	ASP X	4	15.389	7.556	4.698
Atom	29	CA	ASP X	4	15.077	9.493	3.893
Atom	30	HA	ASP X	4	14.204	9.802	3.318
Atom	31	CB	ASP X	4	14.672	9.942	5.34
Atom	32	HB2	ASP X	4	13.997	9.208	5.78
Atom	33	HB3	ASP X	4	15.608	10.074	5.882
Atom	34	CG	ASP X	4	13.905	11.325	5.315
Atom	35	OD1	ASP X	4	12.966	11.503	6.118
Atom	36	OD2	ASP X	4	14.175	12.141	4.417
Atom	37	C	ASP X	4	16.277	10.21	3.214
Atom	38	O	ASP X	4	17.447	9.738	3.295

Table 2.2: Example showing the format of the PDB file for residue Aspartic acid in protein

We assigned an index to each peptide residue (Fig. 2.3) in our study. Indices one to twelve being assigned to the residues AAXAA-AAXAA-GY (X being Q, K and D) respectively. Indices thirteen to twenty four being assigned to the respective residue's side chains. Glycine's side chain (index twenty three) has only a single hydrogen atom and hence this particular residue was abandoned for consistency.

Figure 2.4 below shows the final states of the Q, K and D proteins in the ribbon repre-

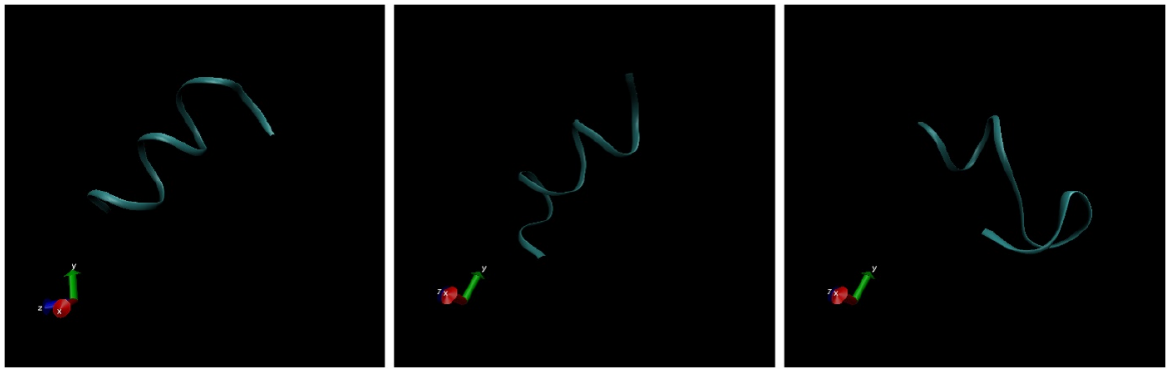

Figure 2.4: Illustration of the final states of Q, K and D proteins from left to right. Plotted with VMD.

sentation, using Visual Molecular Dynamics.

2.1.2 Atmospheric data

As we discussed in previous chapter, seasonal variations in Tropics are mainly governed by precipitation. One of the ever present challenges of meteorology is the collection of the accurate data-sets from around the world. However, in many regions resources are very limited and even existing network of stations is sometimes not maintained properly. In addition, most of the tropical region is covered by oceans. Due to this fact, it is highly problematic to have network of stations with good spatial resolution. The development of new satellite technologies and numerical methods that allow modification of existing and creation of new data-sets, help to fill the absence of in-situ data.

2.1.2.1 Overview of Tropical Rainfall Measurement Mission

As we stated earlier, one of the challenges for meteorology is the collection of accurate both temporal and spatial measurements. The problem is that the satellites with a high temporal resolution usually to have a low spatial resolution and vice versa. To collect data on precipitation, for example, “A-Train” - the satellite of NASA was build and equipped the a number of sensors. Yet during the 24 hours it crosses equator in certain location only twice due to the orbit and coverage. To address this problem, the groundwork validation initiative was brought by Xie and Arkin [131]. They started from comparing the data-sets collected by the satellites with multi gauge networks and found out that at least five gauges are required in order to “produce areal-averaged monthly rainfall for grids of 2.5° x 2.5° latitude/longitude with an accuracy of 10%”. Similar validation initiatives were realized recently with the results being of 20 required gauges [125] for accurate validations over

specific regions in order to preserve accuracy at the same level.

As a result of the collaboration between the National Aeronautics and Space Administration (NASA) and the Japan Aerospace Exploration Agency (JAXA) in 1997 satellite of the Tropical Rainfall Measurement Mission (TRMM) has been launched. The main sensors for precipitation detection are the TRMM Microwave Imager (TMI), the Precipitation Radar (PR), and the Visible and Infrared Radiometer System (VIRS) (Fig.2.5) [126]. The the Precipitation Radar is also the first space rain radar. Starting on the orbit at 350 km in November 1997, it is at 402.5 km since August 2001 in order to extend its observations beyond the original time frame of 2000 [127]. TRMM is on a 46 day precessing, circular, non-sun-synchronous orbit at an inclination of 35 degrees to the equator [GES DISC 2012]. It means that during 24 hours TRMM follows 16 orbits around the Earth and passes directly over a single point at a different local time during the 46 day precession.

To determine amount of rainfall around the world, TRMM uses both satellite precipitation products use both active and passive sensors. It first observes specific bands of the electromagnetic spectrum and afterwards converts the incoming photons into the data-set. The problem with this method is that the electromagnetic data should be converted into precipitation data and to come up with reliable algorithm extensive calibration is required. Such process may create a bias with the priority given to a certain cloud types or precipitation conditions.

In addition to the rain radar, TRMM is also introduced microwave radiometric data-set that covers 35 degrees North to 35 degrees South latitudes. It describes the vertical distribution of precipitation over the tropics and contributes greatly to our understanding of the air - sea - land masses interaction. Data provided by TRMM also helps to improve tropical rainfall models and hence to improve local and global rainfall prediction and its variations over different time scales [115].

In summary, we will state again several merits of using TRMM-based rainfall data-sets.

1. PR data inclusion. Compared to many satellite-based sensors designed to sense precipitation responsive wavelengths passively, precipitation radar was designed to to sense precipitation actively.

2. High spatial resolution of VIRS and TMI. It is possible to go to such resolution-detailed level as 2.2 km and 5 km respectively [126].

However, there are also disadvantages in considering TRMM-based data. Narrow satellite's orbit is defining the spatial limits of the data-set: observations are only available between 40 degrees North and 40 degrees South. Spatial limitation for the precipitation are based on the swath width which is only 215 km [126] and constrains the area covered during each

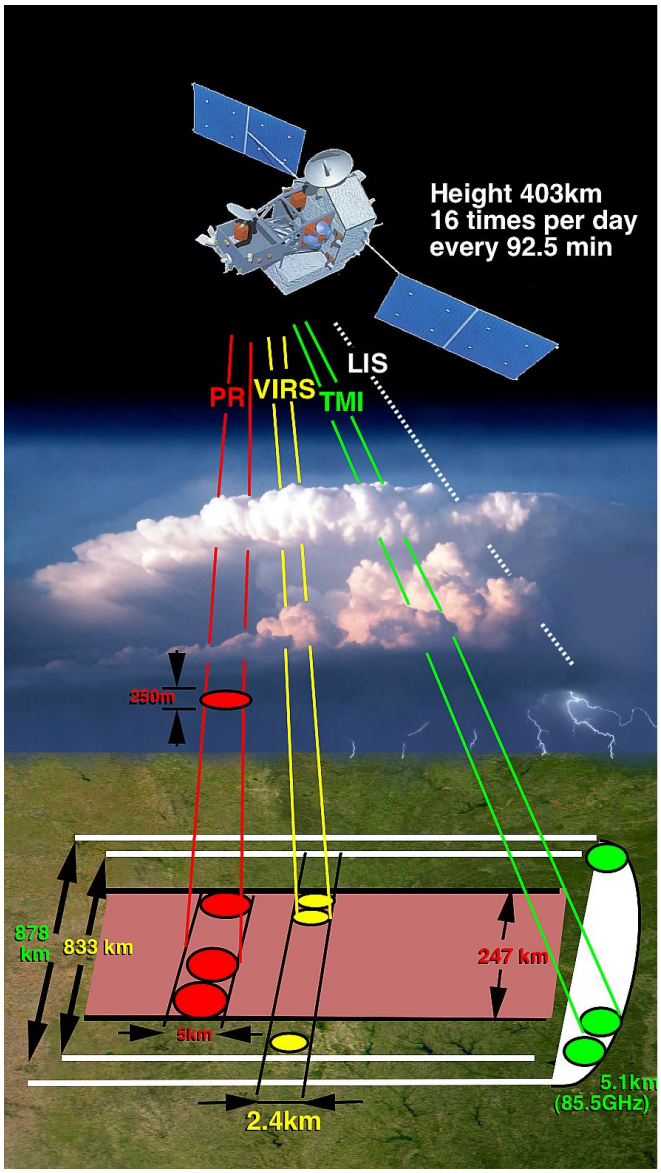

Figure 2.5: Illustration of the TRMM satellite instruments: Lightning Imaging Sensor (LIS), TRMM microwave imager (TMI), Precipitation Radar (PR), Visible and Infra Red Scanner (VIRS). Picture of PRECIPITATION MEASUREMENT MISSIONS of NASA (http://pmm.nasa.gov/image-gallery/diagram-trmm-instruments-measurement-path)

pass. Finally, temporal limitations are due to the rather recent launch of the satellite in 1997: earliest available data in for 1998.

"The *data file layout* and *data file access technique* details that were used in this work are provided for 3B42 at http://disc.gsfc.nasa.gov/precipitation/TRMM_README/TRMM_3B42_rea Reading of a binary TRMM HDF file of data has been done using a C toolkit from http://pps.gsfc.nasa.gov/ts "; "The 3B42 data fields provide a variety of information. It has been used for the quality control and data cleaning.

Index	Data field	Amino Acid	Residue Index
1	precipitation	precipitation	mm/hr
2	precipitation random error*	relativeError	mm/hr
3	satellite observation time	satObservationTime	min. from nominal
4	HQ precipitation	HQprecipitation	mm/hr
5	IR precipitation	IRprecipitation	mm/hr
6	satellite precipitation source	satPrecipitationSource	n/a

Table 2.3: List of data fields, their variable names (in the data structure), and the data units for 3B42 data files.

The coding in the source field matches that in the 3B42RT file, which is as follows: 0 = no observation; 1 = AMSU; 2 = TMI; 3 = AMSR; 4 = SSMI; 5 = F; 17 = SSMIS; 6 = MHS; 7 = TCI; 8 = MetOp-B; 9 = spare sounder; 10 = spare sounder; 11 = F; 16 = SSMIS; 12 = F; 18 = SSMIS; 13 = spare scanner; 30 = AMSU&MHS avg.; 31 = conical avg; 50 = IR; 1,2,. . . ,12 + 100 = sparse-sample HQ. Because the data are provided at nominal UTC hours, each 3B42 data set represents a nominal +/-90-minute span around the nominal hour. Thus, the 00 UTC images include data from the very end of the previous UTC day. For historical reasons, this coding is slightly different than that for the TMPA-RT.

In our work we use TRMM 3B42 data-set. The 3B42 algorithm was designed in order to produce root-mean-square precipitation-error estimates and merged-infrared precipitation from the original TRMM observations. The 3B42 retrieval algorithm used for this product is based on the techniques developed by Huffman [128, 129, 130] and consists of two steps.

1. Firstly, the TRMM VIRS and TMI orbit data (TRMM products 1B01 and 2A12) and the monthly TMI/TRMM Combined Instrument (TCI) calibration parameters (from TRMM product 3B31) are used to calculate monthly IR calibration parameters.

2. The IR parameters are now used to adjust the merged-IR precipitation data, which consists of GOES-W, GOES-E, GMS, Meteosat-5, Meteosat-7 and NOAA-12 data.

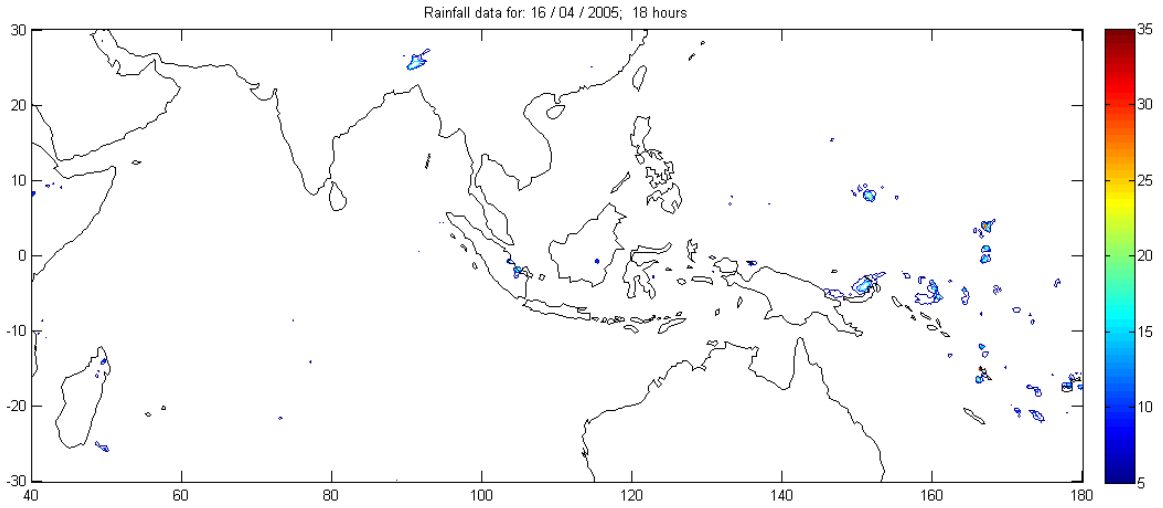

Figure 2.6: Example of the plot with TRMM 3B42 rainfall signal (mm)

In the end, precipitation-error estimates and adjusted and gridded merged-IR precipitation (mm/hr) have a one day temporal resolution and a 0.25-degree by 0.25-degree spatial resolution.

2.1.2.2 Details of the data-set

Main interest of our work is in studying the dynamics of the Madden Julian Oscillation (MJO) and its interaction with monsoon. This determines the region of our study: 30 south to 30 north in latitude and 40 east to 180 east in longitude (Fig. 2.6).

There is evidence that large-scale phenomena like El Niño-Southern Oscillation (ENSO) [117, 118, 119, 120] and Indian Ocean Dipole (IOD) [121, 122, 123] impact dynamics of MJO. To avoid bias from this phenomena, our data-set contains only those years, when they were in the neutral state. To identify these years we used work of Koh [124], where combinatorial probability test was applied to the probability tables of ENSO and IOD events. From the classification table, presented in this work we identified that during operational time of TRMM following years were neutral: 2000, 2001, 2003, 2004, 2005, 2008. Data-sets from these years we study further in this thesis. Since MJO is also affected by monsoons, to identify clear MJO signal we start with analysis of inter-monsoon periods: April-May and October-November. Finally, MJO is an irregular event, it appears sporadically and at discrete time [216]. Because of this we chose for our study those time intervals, when MJO was active and present in the region on Fig 2.6

2.2 Methods

2.2.1 Correlation

When two or more random variables (or quantities that can be considered as variables with some degree of accuracy) have relationship between each other, such relationship is called correlation. Such cases are usually of great interest in data science. Many scientists are interested in how strongly certain variables in their studies are related to each other. For example, two variables, X and Y, are considered to be related to each other if the assumed value of the first variable affects the distribution of the second variable. In turn, variables X and Y are considered to be independent if the changes of the value of X is not affecting the value of Y. Usually, the correlation coefficients express a monotonic connection between the variables. Correspondingly, if with the increase of the values of X the values of Y also do increase it is said that the positive correlation is occurring. Similarly, if with the increase of the values of X the values of Y decrease it is said that the negative correlation is taking place.

The most popular among different types of nonparametric correlation coefficients are: Kendall's tau correlation, Spearman's rank-order correlation and Pearson product moment correlation. In the case when the examined data-set has outliers Kendall's tau correlation and Spearman's rank correlation are usually chosen. When the examined data-set is extensive and with only few outliers, more practical statistical method is considered to be Pearson's correlation coefficient. The Pearson correlation coefficient indicates the strength of a linear relationship between two variables, but its value generally does not completely characterize their relationship [245, 246]. In particular, if the conditional mean of Y given X, denoted E(Y|X), is not linear in X, the correlation coefficient will not fully determine the form of E(Y|X). For the stated reason, we employed Pearson's correlation coefficient (PCC) to study the dynamics of the tropical atmosphere and the proteins folding problem.

2.2.1.1 Pearson's correlation coefficient

The Pearson's correlation coefficient is a common measure of linear dependence between two continuous variables X and Y. It is commonly designated by the Greek letter ρ and could be expressed through the formula:

$$\rho_{X,Y} = \frac{Cov(X;Y)}{\sigma_X \sigma_Y} \tag{2.1}$$

where $Cov(X;Y)$ is the covariance between the variables X and Y, σ_X and σ_Y are respectively standard deviations. In the first part of our study, for example, the variables X and Y

are three-dimensional atom coordinate vectors. To address this problem we need to derive the PCC for vectors. Since,

$$Cov(X;Y) = \sum_{i=1}^{n}((X_i - \overline{X})(Y_i - \overline{Y})), \quad \sigma_X \sigma_Y = \sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}. \tag{2.2}$$

the equation (2.1) can be written as:

$$\rho = \frac{\sum_{i=1}^{n}((X_i - \overline{X})(Y_i - \overline{Y}))}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{2.3}$$

where $\bar{X}$ and $\bar{Y}$ are:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \quad \overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}. \tag{2.4}$$

The Pearson's correlation coefficient for vectors is very similar to the scalar version:

$$\rho = \frac{Cov(\mathbf{q},\mathbf{p})}{\sigma_\mathbf{q}\sigma_\mathbf{p}} \tag{2.5}$$

with $\mathbf{q}$ and $\mathbf{p}$ being not scalars, but vectors. However, the covariance and standard deviation for vectors are defined differently:

$$Cov(\mathbf{q},\mathbf{p}) = \sum_{(q,p)=[(q_x,q_y,q_z),(p_x,p_y,p_z)]} \sum_{i=1}^{n}(q_i - \overline{q})(p_i - \overline{p}) \tag{2.6}$$

$$\sigma_\mathbf{q} = \sum_{q=(q_x,q_y,q_z)} \sum_{i=1}^{n}(q_i - \overline{q})^2 \tag{2.7}$$

$$\sigma_\mathbf{p} = \sum_{p=(p_x,p_y,p_z)} \sum_{i=1}^{n}(p_i - \overline{p})^2 \tag{2.8}$$

with x, y and z being the three components of the vectors $\mathbf{p}$ and $\mathbf{q}$. First, for each individual component the covariance between $\mathbf{q}$ and $\mathbf{p}$ is calculated and then the values are summed. The standard deviation is calculated similarly. The vector Pearson's correlation coefficient in contrast with the scalar one sums up the variations of individual components.

The Pearson's correlation coefficient takes a value in the range $[-1,+1]$ (Fig. 2.7). Values $\rho > 0$ represent positive monotonic association between two variables X and Y (X increases (decreases) as Y increases (decreases)), values $\rho > 0$ represent negative monotonic association (X increases (decreases) as Y decreases (increases)), finally value $\rho = 0$

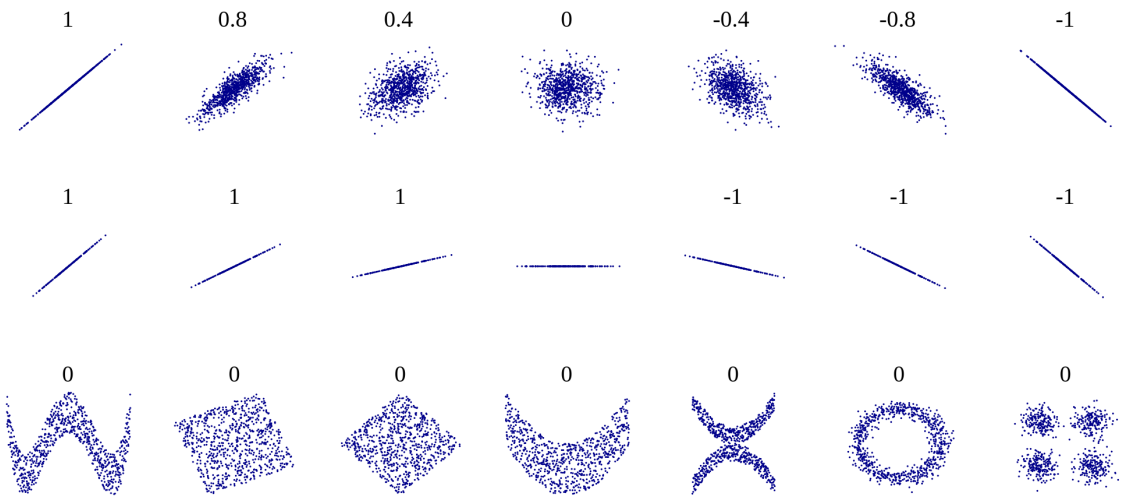

Figure 2.7: Several sets of points, representing variables X and Y with the correlation coefficient (ρ) for each set [160].

represents the absence of the monotonic association between the variables. In the case of bivariate normal data value $\rho = 0$ would imply absence of any association, however, for other bivariate distributions (e.g. $Y = X^2$, $(x \in (-1,1))$) ρ can be zero for dependent variables. As a result, the absolute value of ρ represents the strength of the linear correlation between the two variables: ρ of 1 implies total positive correlation (linear relationship: $Y = a + bX$).

2.2.2 Vector cross correlation

In our studies we work with the time series that is produced from the molecular dynamics simulation. Our main interest is concerning residues interaction. For this purpose we will calculate the correlations between all possible pairs of residues.

Given two scalar time series $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ with means μ_x and μ_y, respectively, the zero-lag scalar cross correlation between $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$C(\mathbf{x}, \mathbf{y}) = \frac{\langle (x - \mu_x)(y - \mu_y) \rangle}{\sigma_x \sigma_x} \tag{2.9}$$

where σ_x^2 and σ_y^2 are the respective variances of $\mathbf{x}$ and $\mathbf{y}$.

Generalizing into three-dimension and replacing σ_i^2 by the covariance matrix $\sum_i$, we immediately see that we have to calculate the inverse square-root of the covariance matrix $\sum_i^{-1/2}$. If $\blacksquare_i$ and $\mathbf{U}_i$ are the eigenvalue and eigenvector matrices of $\sum_i$, then its inverse

square-root is defined as:

$$\sum_{i}^{-1/2} = \mathbf{U}_i \blacksquare_i^{-1/2} \mathbf{U}_i^T \tag{2.10}$$

Hence, this tells us that if we define a scaled vector $\vec{\xi}_i = \sum_i^{-1/2}(\vec{v}_i - \vec{\mu}_i)$, the zero-lag vector cross correlation between vector time series $\mathbf{v}_i$ and $\mathbf{v}_j$ is then simply:

$$C_{ij} = \left\langle \vec{\xi}_i \cdot \vec{\xi}_j \right\rangle \tag{2.11}$$

This definition of vector cross correlation is basis independent, and thus truly represents the cross correlation between two vector quantities.

For n vector time series $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, the cross correlation matrix $\mathbf{C}$ is simply a symmetric, $n \times n$ matrix whose elements are pairwise vector cross correlation of the i-th and j-th time series:

$$\mathbf{C}_{i,j} = C_{ji}. \tag{2.12}$$

2.2.3 Data Clustering

Data Clustering, a major unsupervised machine learning technique, aims at forming the groups (clusters) of the data points in such a way, that elements within a group (cluster) have high similarity with each other, while being dissimilar to points in other groups (clusters) [139]. Figure 2.8 depicts the clustering of 2D points into 3 clusters. Each cluster can be represented by its center of mass, or average point (legend $*$), or an actual point referred to as medoid or exemplar (legend o).

2.2.3.1 Clustering for Exploratory Data Analysis

While clustering also applies to supervised data-sets (when each point is labeled after its class according to some oracle), it is more often used for exploring the structure of the data-set in an unsupervised way – provided that some similarity or distance between points is available [247, 248].

1. Group discovery. By grouping similar points or items into clusters, clustering provides some understanding of the data distribution, and defines a preliminary stage for a discriminant analysis, after the "divide to conquer" strategy.

2. Structure identification. A particular type of clustering approach, hierarchical clustering provides a clustering tree (as opposed to the partition in Fig 2.8). The clustering tree,

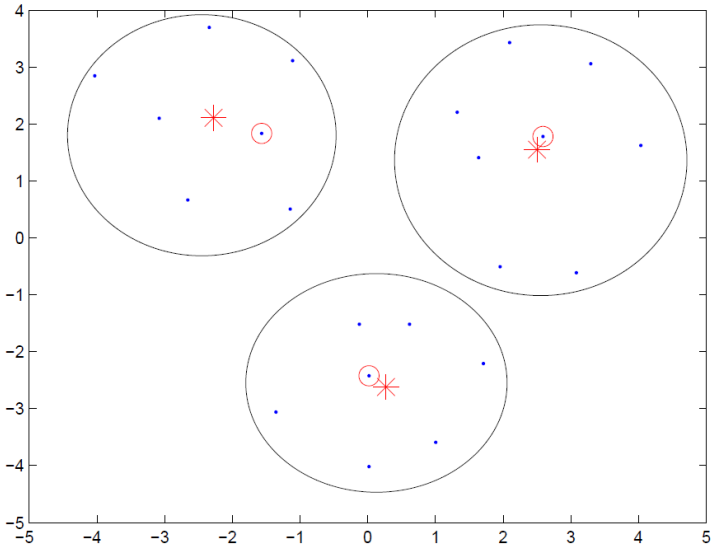

Figure 2.8: Example of data clustering: data set contain 3 clusters and for each cluster center of mass ($*$) is calculated and its exemplar ($\circ$).

also known as dendrogram, depicts the structure of the data distribution with different granularities; it is used in particular in the domain of biology to depict the structure of evolved organisms or genes [133].

3. Data compression. One functionality of clustering is to provide a summary of the data-set, representing each cluster from its most representative element, either an artifact (center of mass) or an actual point (exemplar). The cluster is also qualified by its size (number of elements), the radius (averaged distance between the elements and the center), and possibly its variance. Clustering thus allows the compression of N samples into K representatives, plus two or three parameters attached to each representative.

4. Dimensionality reduction or feature selection. In the case, when the amount of items in the data set is much smaller than the amount of features, dimensionality reduction or choice of feature is required as a preliminary step for most machine learning algorithm. One unsupervised approach to dimensionality reduction is based on clustering the features and retaining a single (average or exemplar) feature per cluster [134, 135].

5. Outlier detection. Many applications involve anomaly detection, e.g., intrusion detection [136], fraud detection [137], fault detection [138]. Anomaly detection can be achieved by means of outlier detection, where outliers are either points which are very far from their cluster center, or form a cluster with small size and large radius.

6. Data classification. Last but not least, clustering is sometimes used for discriminant learning, as an alternative to 1-nearest neighbor classification, by associating one point to the majority class in its cluster.

2.2.3.2 Formal Definition

Let $X = \{x_1, \ldots x_N\}$ be a set of points, and let $d(x_i, x_j)$ denote the distance or dissimilarity between items x_i and x_j. Let clustering on X be denoted by $C = \{c_1, \ldots c_K\}$. The quality of C is most often assessed from its distortion, defined as:

$$J(C) = \sum_{i=1}^{K} \sum_{x \in C_i} d^2(x, C_i) \tag{2.13}$$

where distance between x and cluster C is most often set to the distance between x and the center of mass $\mu_i = \frac{1}{n_C} \sum_{x \in C_i} x$ of cluster C. n_C denotes the size (number of items) in C.

The above criterion thus can be interpreted as the information loss incurred by representing X by the set of centers associated to C. It must be noted that the distortion of clusterings with different numbers of cluster cannot be compared: the distortion naturally decreases with the increasing number of clusters and the trivial solution associates one cluster to each point in X.

2.2.3.3 Distance

As it was shown in the part 2.2.3.2, clustering depends on the distance defined on the domain space. Distance learning is currently among the hottest topics in Machine Learning [142]. Since the elements of the data-set may be closer to each other or further from each other, depending on the definition of the distance, clusters would be shaped according to the choice of metric. For example, the distance between points (1,0) and (0,0) in 2D space is 1, according to the common norms, however the distance between (1,1) and (0,0) can be $\sqrt{2}$ if using Euclidean distance, 2 if using Manhattan distance or 1 if using maximum distance. Among the most commonly used definitions of the distance are the following [249]:

Euclidean distance

Let x_i and y_j be N-dimensional vectors. Then Euclidean distance will be calculated as:

$$d_{Ecl} = \sqrt{\sum_{k=1}^{N} (x_{ik} - y_{jk})^2}. \tag{2.14}$$

Root mean square distance

The root mean square distance (or average geometric distance) for the same vectors will be:

$$d_{RMS} = \frac{d_{Ecl}}{n}. \tag{2.15}$$

Minkowski distance

Minkowski metric is defined in a normed vector space, and represents a generalization of the Euclidean and Manhattan distances. Minkowski distance of order q ($q \in \mathbb{N}$) for the same pair of vectors is defined as

$$d_M = \sqrt[q]{\sum_{k=1}^{N} (x_{ik} - y_{jk})^q}. \tag{2.16}$$

Distance based on Pearson's correlation

For the same pair of N-dimensional vectors x_i and y_j Pearson's correlation will be defined as

$$C = \frac{\sum_{k=1}^{N}(x_{ik} - \mu_{x_{ik}})(y_{jk} - \mu_{y_{jk}})}{\sqrt{\sum_{k=1}^{N}(x_{ik} - \mu_{x_{ik}})}\sqrt{\sum_{k=1}^{N}(y_{jk} - \mu_{y_{jk}})}}. \tag{2.17}$$

Here μ_{x_i} and μ_{y_j} are mean values of x_i and y_j respectively, computed as $\mu_{x_i} = \frac{1}{N}\sum_{k=1}^{N} x_{ik}$ and $\mu_{y_j} = \frac{1}{N}\sum_{k=1}^{N} y_{jk}$.

There are several versions of cross-correlation-based distances. For example, in his work [132] Golay propose these two for the fuzzy c-means algorithm:

$$(a)\, d_C^1 = (\frac{1-C}{1+C})^{\beta}, \qquad (b)\, d_C^2 = 2(1-C). \tag{2.18}$$

In Eq. (2.18a), β takes positive values and has a analogous function as m in the fuzzy c-means algorithm (we will discuss it later). In turn, for the hierarchical clustering the following distances are proposed:

$$(a)\, d_C^3 = 1 - C, \qquad (b)\, d_C^4 = 1 - C^2. \tag{2.19}$$

Distance defined in Eq. (2.19b) is called the Pearson squared distance. It measures the correlation between two profiles, both positive and negative (Figure 2.9).

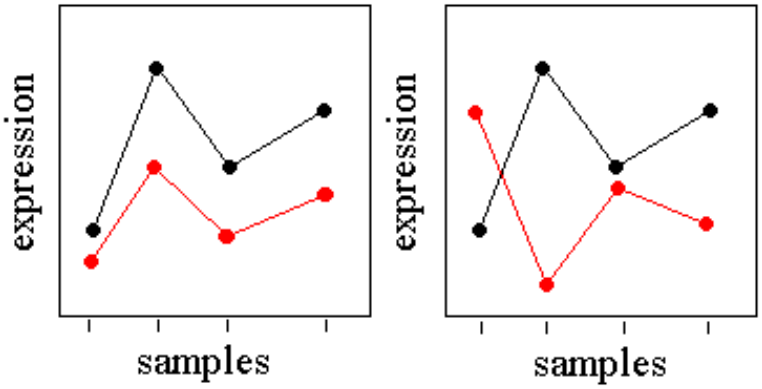

Figure 2.9: Example of correlated and anti-correlated time series.

In the left part of Fig. (2.9), both profiles show very high similarity and have almost perfect correlation despite the scale differences and shift in level. These time series would be clustered together if the distance will be defined as in Eq (2.19a) or Eq (19b). In right part of the Fig. (2.9), both profiles show almost perfect anti-correlation. Considering distanced_C^3 these time series would be grouped in remote clusters and would be grouped together if Pearson squared (d_C^4) is used.

In our work we use data-set, that consist of coordinates of the atoms. One of the questions we will be answering using this data-set is whether there are any consistent patterns of motion of the atoms. Such patterns may be revealed by both correlated and anti-correlated episodes. Hence, we will measure similarity between objects within certain data-set using Pearson correlation-based distances defined in Eq. (2.19). To study the dynamics of the protein molecule we will use Pearson squared distance d_C^3 (Eq. 2.19a) and to study the dynamics of intraseasonal phenomena of the tropical atmosphere we will use use d_C^4(Eq. 2.19b).

2.2.3.4 Clustering algorithms

The literature offers a large variety of different clustering techniques; the choice of a particular technique should reflect the nature of the data-set and incorporate all available knowledge of it. With no pretension to exhaustivity, we will briefly introduce most common [139, 140] categories of clustering algorithms.

Partitioning methods

A partitioning clustering method creates m partitions of the data-sets [250, 251]. It allocates the elements of the data-set into m clusters, that satisfy the following criteria:

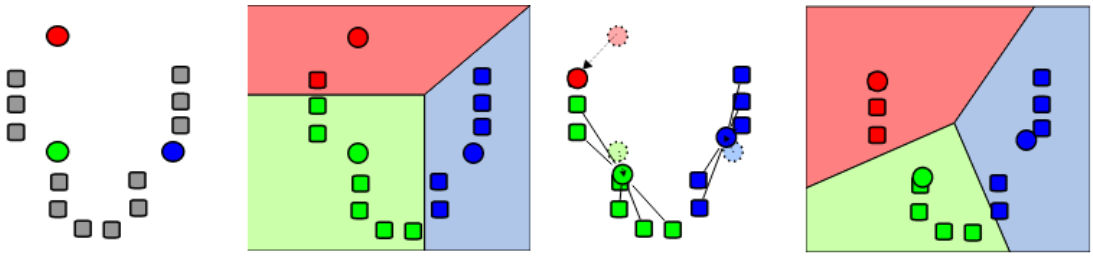

Figure 2.10: Demonstration of the standard algorithm for partitioning method.

• each cluster contains at minimum one element of the data-set,
• each element of the data-set must belong to exactly one cluster.

The most commonly used partitioning methods are K-means and K-medoids. Clustering procedure for this method consists of the following four steps (Fig. 2.10):

1. for desired number K of clusters, we randomly choose K points $x_1, ...x_K$ from X, and set $C_i = x_i$;

2. iteratively, associate each x in X to cluster C_i minimizing $d(x, C_i)$;

3. replace the initial collection of K points with the center of mass μ_i of clusters $C_1, ...C_K$;

4. go to step 2 and repeat until the partition of X is stable.

Clearly, the above procedure minimizes the clustering distortion, however there is no assurance that a global minimum will be reached. A better solution (albeit still not optimal) is obtained by executing the algorithm using different initializations and returning the best solution. K-median, another partitioning algorithm, is used in the situation when a center of mass cannot be calculated (e.g. when data points are structured entities, curves or molecules).

Hierarchical Clustering

Hierarchical clustering is a method that aims to build a hierarchy of clusters, cluster tree - dendrogram. There are two strategies of construction of the dendrogram: "bottom up" or agglomerative hierarchical clustering and "top down" divisive hierarchical clustering. Agglomerative hierarchical clustering procedure allocate each data point x in X to a separate cluster. If there are N data points in the set, then there will be N clusters initially. Then, sequentially, by connecting similar objects it results in one cluster, containing all objects. Agglomerative hierarchical clustering procedure consists of the following steps (Fig. 2.11):

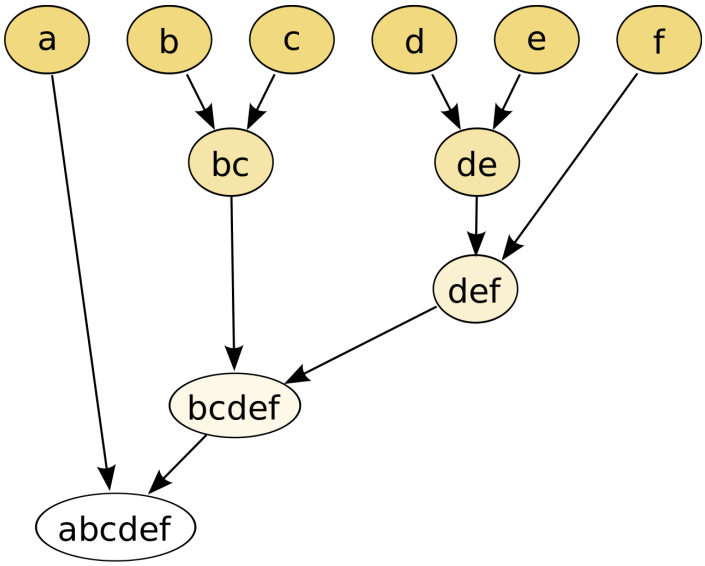

Figure 2.11: Demonstration of the standard algorithm for hierarchical clustering method.

1. For each pair (C_i, C_j) $(i \neq j)$ compute the inter-cluster distance $d(C_i, C_j)$;
2. find out the two clusters with minimal inter-cluster distance and merge them;
3. go to step 1, and repeat until the number of clusters is one, or the termination criterion is satisfied.

As exemplified on Fig. 2.10, the 6 initial clusters ({a}, {b}, {c}, {d}, {e}, {f}) become 4 by merging {b} and {c}, and {d} and {e}; next, clusters {d e} and {f} are merged; then {b c} and {d e f} are merged. And the last two clusters are finally merged.

In this way, agglomerative hierarchical clustering simply advance by establishing the most similar clusters and merging them. To determine distance between sets of observations, different linkage criterions are used [250, 252]. Some commonly used linkage criteria are (where $d(a,b)$ is the chosen distance from part 2.2.3):

- **Single linkage clustering**: minimum distance

$$d(C_i, C_j) = \min\{d(\mathbf{x}_i, \mathbf{x}_j) \mid \forall \mathbf{x}_i \in C_i, \forall \mathbf{x}_j \in C_j\} \tag{2.20}$$

- **Complete linkage clustering**: maximum distance

$$d(C_i, C_j) = \max\{d(\mathbf{x}_i, \mathbf{x}_j) \mid \forall \mathbf{x}_i \in C_i, \forall \mathbf{x}_j \in C_j\} \tag{2.21}$$

- **Mean linkage clustering**: mean distance

$$d(C_i, C_j) = d(\mu_i, \mu_j), \tag{2.22}$$

where $\mu_i = \frac{1}{|C_i|}\sum_{\mathbf{x}_i \in C_i} \mathbf{x}_i$ and $\mu_j = \frac{1}{|C_j|}\sum_{\mathbf{x}_j \in C_j} \mathbf{x}_j$.

- **Average linkage clustering**: average distance

$$d(C_i, C_j) = \frac{1}{\mid C_i \mid \times \mid C_j \mid} \sum_{\mathbf{x}_i \in C_i, \mathbf{x}_j \in C_j} d(\mathbf{x}_i, \mathbf{x}_j) \tag{2.23}$$

- **Average group linkage**: group average distance (assume that C_i and C_j are merged)

$$d(C_i, C_j) = \frac{1}{(\mid C_i \mid + \mid C_j \mid) \times (\mid C_i \mid + \mid C_j \mid -1)} \sum_{\mathbf{x}_i \in C_i \cup C_j, \mathbf{x}_j \in C_i \cup C_j} d(\mathbf{x}_i, \mathbf{x}_j) \tag{2.24}$$

- **Minimum energy clustering**

$$d(C_i, C_j) = \frac{2}{nm}\sum_{p,q=1}^{n,m} \| \mathbf{x}_{ip} - \mathbf{x}_{jq} \|_2 - \frac{1}{n^2}\sum_{p,q=1}^{n,m} \| \mathbf{x}_{ip} - \mathbf{x}_{iq} \|_2 - \frac{1}{m^2}\sum_{p,q=1}^{n,m} \| \mathbf{x}_{jp} - \mathbf{x}_{jq} \|_2 \tag{2.25}$$

Contrasting with agglomerative hierarchical clustering, divisive hierarchical clustering starts with a single cluster gathering the whole data set. In each iteration, one cluster splits into two clusters until reaching the state when each and every point is in a separate cluster (or the termination criterion is satisfied). The divisive hierarchical clustering criterion most often is the maximal diameter or the maximal distance between two closest neighbors in a cluster. Application-wise, agglomerative hierarchical clustering are much more popular than divisive hierarchical clustering, seemingly because the divisive hierarchical clustering criterion is less natural and more computationally expensive.

The dendrogram obtained by hierarchical clustering methods shows the structure of the data distribution, illustrating the relationship between items. Every level of dendrogram gives one possible partition of the data-set, enabling one to select the appropriate number of clusters a posterior. This is important distinction from partitioning methods, where number of cluster is set initially.

Density-based methods

The distance between the elements of the data-set is the main parameter of how the elements are clustered for the most of partitioning methods. It allows to find only spherical-shaped clusters using such methods and at the same time meet the difficulties in exploring clusters of arbitrary shapes. To address this problem, the notion of density was introduced as a main clustering parameter. Density-based clustering methods put the stress on exploring arbitrarily shaped clusters. This methods are grounded on the clustering assumption [141],

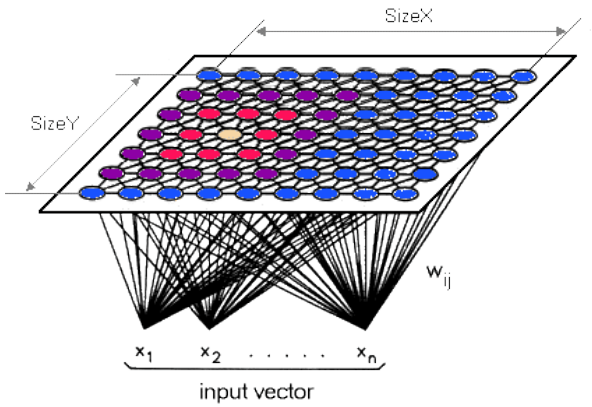

Figure 2.12: Illustration of the Self-Organizing Map.

that states that dense regions of the data-set are clusters, and clusters are separated by regions with low density. Such methods are used both for noise reduction and the discovery of the arbitrarily shaped clusters.

Neural network. Self-Organizing Map

Neural network is another popular clustering approach, specifically the Self-Organizing Map method. Kohonen, who developed self-organizing maps [143], came up with an idea of forming a class of neural networks, where neurons are organized in a low-dimensional (typically 2D) structure and then are iteratively trained using certain self-organizing procedure. The SOM model can be used in particular to visualize the data-set and explore its properties.

Self-organizing map is defined as a grid of interconnected nodes following a regular (quadratic, hexagonal, ...) topology (Figure 2.12). Each node is associated to a representative c usually uniformly initialized in the data space. The representatives are iteratively updated along a competitive process: for each data point x, the representative c_x most similar to x is updated by relaxation from x and the neighbor representatives are updated too. Clusters are defined by grouping the most similar representatives. When dealing with a large size grid, similar nodes are clustered (using e.g. k-means or agglomerative hierarchical clustering) in order to promote the quantitative analysis of the data and of the map [144].

2.2.3.5 Visualization of hierarchical clustering results

The results of any hierarchical clustering procedure can be represented in several ways: graphical or as a list of symbols depicting elements of the data-set and their relationships.

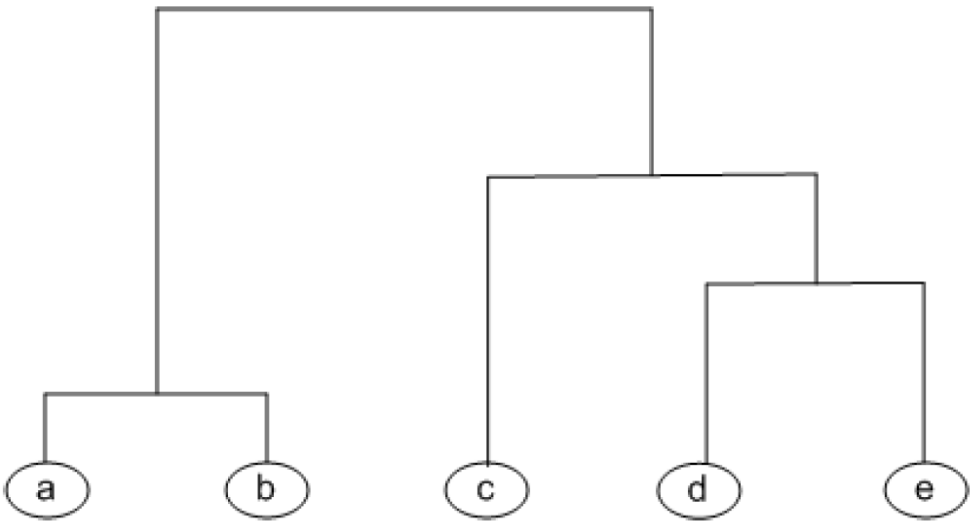

Figure 2.13: Illustration of hierarchical clustering dendrogram.

For a human being it is easier to operate with a graphical representation of hierarchical clustering results (dendrogram) and to analyze it. For a computer, in turn, a list of symbols is more preferable and could be used to boost the performance of the algorithm.

The dendrogram, graphical representation of the hierarchical clustering results comes from Greek words “tree” and “graphic”: *dendro* and *gramma*. This plot where depicts every step of the hierarchical clustering procedure as a merger of two branches of the tree into a single one. Each cluster produced at the certain step is represented as a branch. Figure 2.13 shows a dendrogram of a data-set of the objects: $\{a,b,c,d,e\}$.

Reingold and Tilford were the first [145] who introduced the classical hierarchical view for their algorithms. The algorithm computes independently the sub-tree’s relative positions, then connects them by putting sub-trees together. “Top down”, “left right” and grid like layouts could be produced. The algorithm is simple, fast, and predictable. Tree-maps were introduced by Johnson and Shneiderman [146]. In the tree-maps, the hierarchical structure is mapped to nested rectangles. Tree-map is constructed by recursive subdivision, i.e., a node is divided into some rectangles based on the children’s node size. The direction of subdivision changes, a rectangle is subdivided in one direction (for instance, horizontally), and for the next level this direction alternates. Tree-maps provide a compact visual representation of complex hierarchical data.

2.2.4 Minimal Spanning Tree (MST)

In Section 2.2, we introduced the cross correlation notion. For better understanding how the transition occurs we will be visualizing the cross correlation matrices by drawing the minimal spanning tree (MST) graphs [161, 162].

One of the most commonly used algorithms for complexity network visualization is the Minimal Spanning Tree (MST). It was introduced by Kruskal in 1956 [161]. Given a cross correlation matrix C, we can construct MST graph by first sorting the pairs $(i; j)$ based on their cross correlation value $C_{i;j}$ from the largest (most strongly-correlated pair) to the smallest. Then, starting from the top of the list, we draw a link between i and j. We continue linking the next most strongly-correlated pairs without allowing any loop in the graph. If connecting a pair from the list will result in loop formation, we will not draw the link and continue instead to the next pair. Once we have connected all the n nodes, we have completed the MST graph and stop drawing.

Chapter 3

Protein Folding Problem

In this chapter we analyze the folding process of different sets of proteins. We start with investigation of the single 62-atom protein molecule, its low (1.0 ps time step) and high (0.1 ps time step) resolutions. Later in the chapter we start inspecting three protein molecules in order to find possible folding events and precursors of such event.

3.1 Single molecule

3.1.1 Low-Resolution Study

We start with the analysis of the data-set described in subsection 2.1.1.1 For each atom, the positions (x_i, y_i, z_i) are recorded every picosecond for totally 5 nanoseconds. There are thus a total of 5001 time points. Taking the difference of two successive positions of the 5001 time points then gives the distance traveled for each atom in one picosecond, hence average velocity. The time series is then partitioned into ten time windows with 500 time steps each. Our strategy is to identify at first the period when the folding may take place by inspecting big parts of the time series (ten time windows) and then zoom at this moment and study it in more details. For each time window, we start with compute the Pearson vector correlation matrix C_k, $k = 1, 2, \ldots 10$. Each C_k is a 62-by-62 correlation matrix.

For each vector Pearson correlation matrix $C_k = [C_{kij}]$, we then calculate the distance matrices D_k, where the distance between two objects is calculated using the formula 2.19a from subsection 2.2.3.3: $d_{kij} = 1 - C_{kij}$. The resulting distance matrices D_k, is then the distance matrix for each time window, with the desired property of all zeros along the main diagonal. According to the hierarchical clustering algorithm described in 2.2.3.4 we then can calculate the clustering matrices and visualize results via dendrograms.

On the first iteration we produced ten dendrograms with non-overlapping windows of

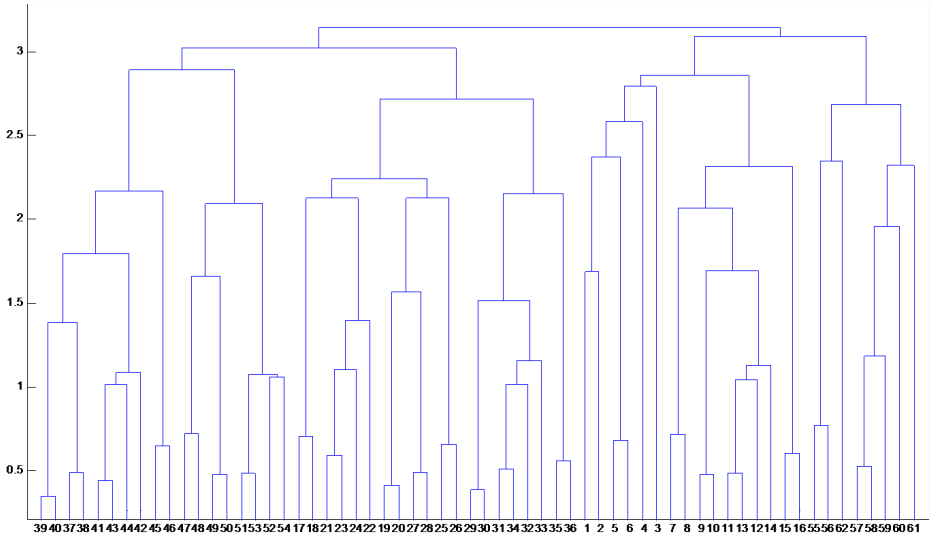

Figure 3.1: One of the ten dendrograms obtained from the ten vector Pearson correlation matrices of the average velocity time series for the 62-atom penta-alanine molecule. The correlation matrices were calculated over non-overlapping windows of 500 time steps each. The one presented on this figure correspond to the interval $t_1 = 4001$ ps to $t_2 = 4500$ ps.

500 time steps each (Fig. 3.1). From the ten dendrograms, we saw that the strongly correlated clusters of atoms, {7, ..., 16}, {17, ..., 26}, ... correspond to the five alanine residues. Based on the relationships between these five residues, we visually classified these ten dendrograms into three groups. In the first group Ala2 is in the same cluster as Ala3 and Ala4. In the second groups Ala2 is in a different cluster from Ala3 and Ala4. In the third group, Ala4 is in a different cluster from Ala2 and Ala3.

These dendrograms gave us a picture of how the protein molecule is evolving in time. For example, on the dendrogram from Figure 3.1, we see that Ala3, Ala4, Ala5, Ala6 are in one cluster, but Ala2 is in a different cluster with the terminal methyl and methylamide groups. This suggests that during these time window, the dynamics of Ala2 is decoupled from those of Ala3 and Ala4, and became more coupled to the dynamics of the terminal groups. Because the grouping of dendrograms into these three clusters is physically meaningful, this is the classification we wish to obtain automatically, using the clustering approaches explored below.

On this very first step we analyzed general behavior of the molecules visually due to low time resolution. For more systematic study of the underlying folding mechanism greater number of dendrograms had been classified. To automate this dendrogram based analysis,

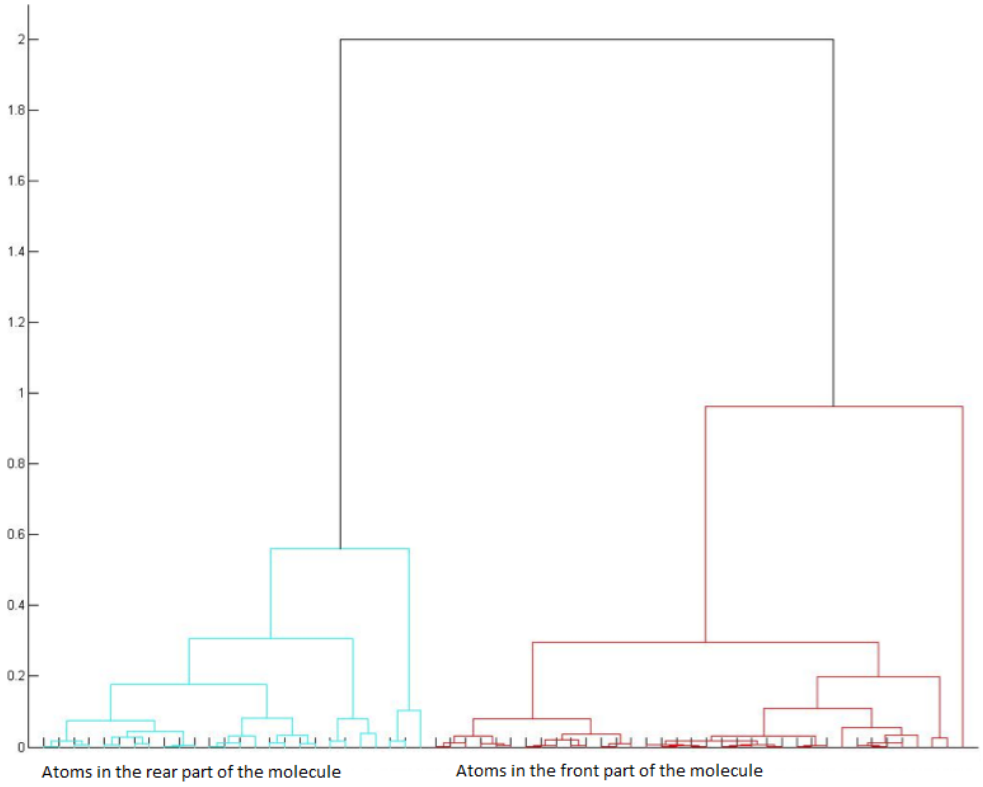

Figure 3.2: Complete-link dendrogram of cross correlation between atoms, showing two non-interacting main clusters of roughly equal in size. The red cluster contains atoms from roughly the front half of the protein (shown on the Fig.2.1), and the blue cluster contains atoms from the other half.

we looked at the problem of clustering the dendrograms themselves. Then both correlation-based distances and order-based distances were studied.

3.1.2 High Resolution Study

The high resolution study uses the same time series of ALA-5 protein in water for 5.0 ns duration consists of frames, taken every 0.1 ps. We followed the same procedure as in the low-resolution study with only difference being that the distance was defined as $d_{kij} = 1 - C_{kij}$ for convenience of visualization. New array of dendrograms was obtained, however their structure was already slightly different.

From the analysis of the dendrograms we found that, at different times, the protein can be represented by different number of effective clusters consisting of different atoms. At certain times, the atoms clusters into two robust clusters, for example shown on Figure 3.2, with the red cluster containing roughly the atoms from the front half of the molecule and

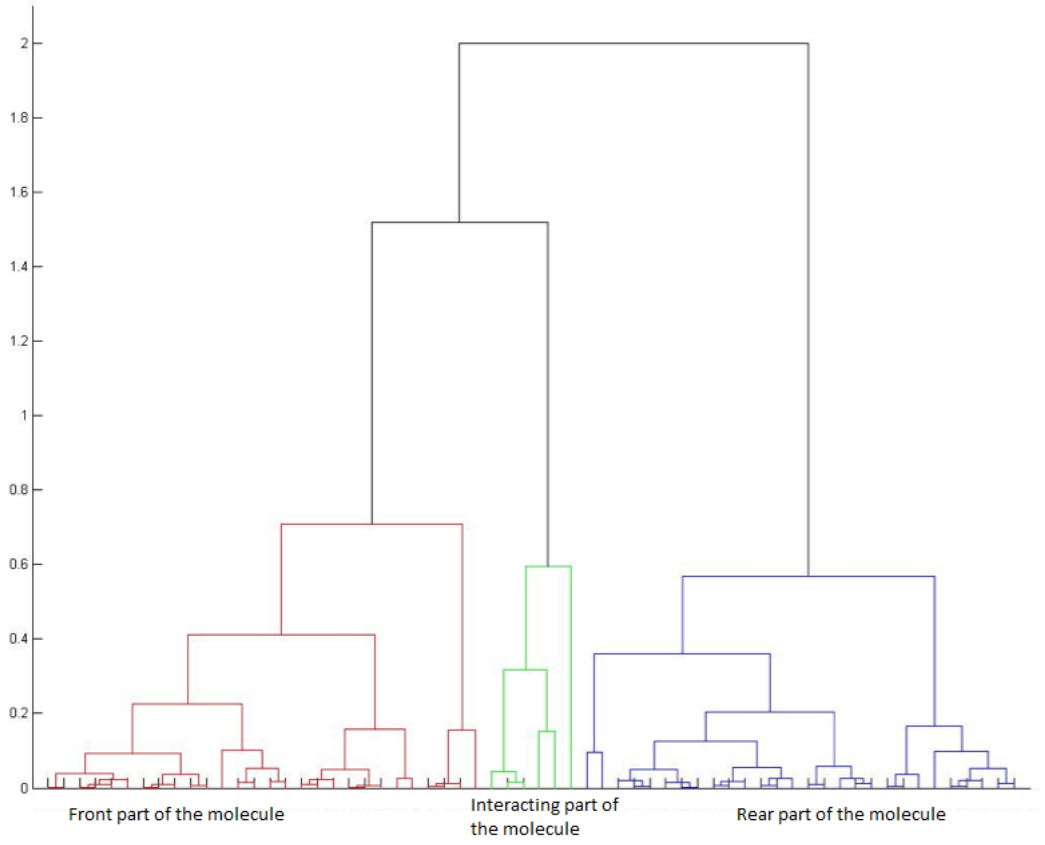

Figure 3.3: Complete-link dendrogram of cross correlation between atoms, showing many, less robust clusters of roughly equal in size. Note that the red and blue clusters shrink while the interaction green cluster grows in size, indicating strong interaction between atoms across the protein.

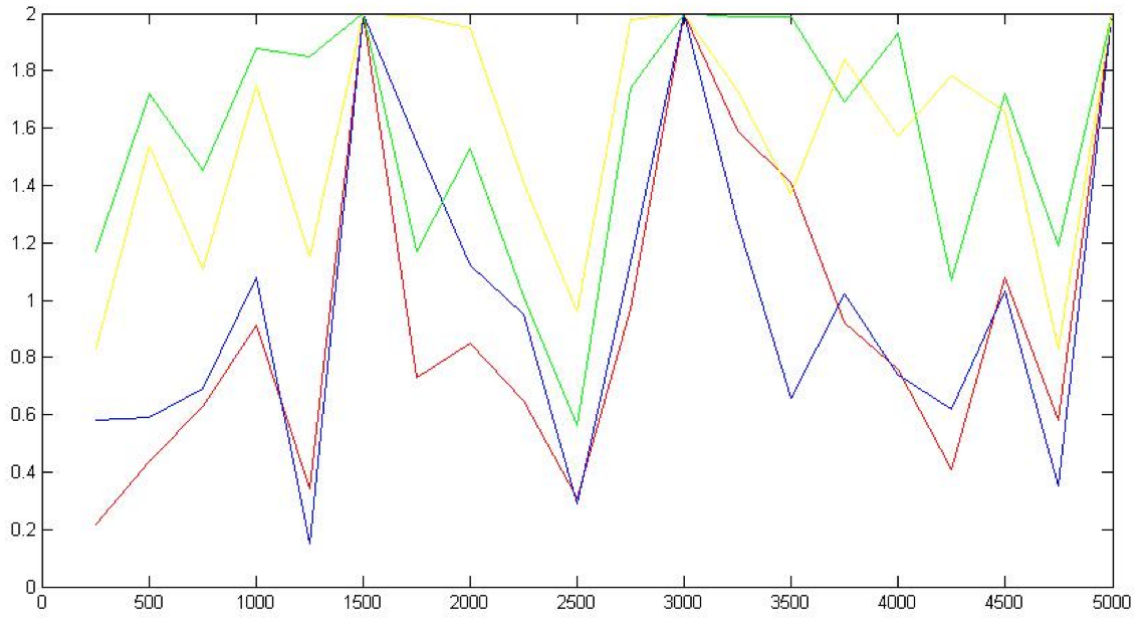

Figure 3.4: Clustering thresholds of two main clusters (red and blue) and two largest interaction clusters (green and yellow). Low thresholds imply robust clusters with strongly correlated members. Every time step on the horizontal axis represents 1 ns.

the blue cluster containing atoms from the other half. This indicates that at these times the protein can be thought of consisting of only two rigid, weakly-interacting main clusters roughly equal in size. In other times, we observe a sign of weak interaction between the two clusters, as shown in Figure 3.3. By interaction between clusters we imply exchange of atoms between them over several time-steps. Hence, weak interaction over certain period means, that clusters consist of the same atoms and strong interaction means, that the clusters exchange significant (> 10%) of their composing atoms. Here the red and blue clusters are the two main clusters seen previously, interacting through the smaller green interaction cluster which consists of atoms located in the middle of the protein between the red and blue clusters. In an strongly-interacting state, we found the main clusters break into many smaller, less robust clusters.

Besides these pictures, we also discovered an interesting fact from the cluster threshold values. Such a threshold is defined for a particular cluster as the maximum distance between elements within the cluster. Since the distances are defined from the cross correlation values, a low clustering threshold implies a robust cluster with strongly correlating members. The time evolution plot of the thresholds for the two main clusters and two largest interaction clusters is shown in Figure 3.4. Among other things, we are most interested in the significant drops of the four thresholds at around 1,250, 2,500, and 4,750 ns., indicating the

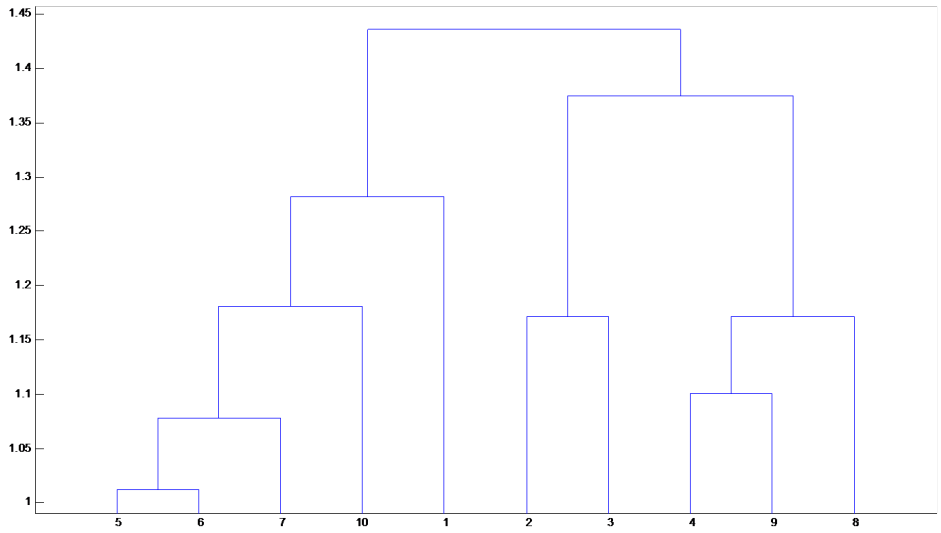

Figure 3.5: Dendrogram showing the relation between 10 clustering matrices, each covering 1000 time steps.

formations of robust, strongly-correlated clusters. We speculate that these drops may correspond to the folding of the protein. If this is true, then the brief increase of the threshold just before the drops may actually correspond to barrier crossings before folding, and that the significant increase afterwards are the unfolding process. Furthermore, the slight increase of only blue threshold at around time step 3,750 ns indicates a possible failed folding attempt where barrier crossing by the blue cluster was not followed by the red cluster. This picture, suggesting that the protein may actually folded and unfolded three times during the simulation, is in somewhat agreement with our previous finding where we observed plausible folding signatures at around time step 500 and 3,000.

So far, it seems to us that the general facts emerging from the two studies agree, while the details on when the folding actually take place do not. However, we also note that the time series clustering study alone only provide us with very crude results.

With the time series data containing only 50000 time steps and the current window size of 250 and 500 time steps, we have 100 to 200 data points and graphs to analyze. To approach this problem we decided to study evolution of the structure of the protein molecule by applying clustering analysis to the clustering matrices that were initially obtained. As a result we have new dendrograms where on the horizontal axis we have not the single time steps, but clustering matrices describing structure over certain period.

On the figure 3.5 we can see that there are two main clusters: one consists of the matrices

representing the structure of the molecule during the middle of the simulation plus very beginning and very end of it and another cluster consists of the matrices that represents the rest of the simulation. It agrees with the results we obtained from the threshold analysis, when we saw changes in the beginning, middle and the end of the simulation.

Assuming that we know where to look to find the folding event, we are also interested in the dynamics of the protein folding. And especially what events trigger the folding process. For this purpose we identified possible precursor segments visually from the simulation itself by looking for short segments that started slightly before the interesting events. The results are presented in the Table 3.1.

i	Atom Type	Atom No	Precursor Segment	Global Event
1	CB	11	2,159	2,500
2	O	16	2,135	2,500
3	CB	31	2,354	2,500
4	CB	51	2,171	2,500
5	O	16	11,320	11,700
6	N	17	11,330	11,700
7	CB	21	11,300	11,700
8	CB	41	11,480	11,700
9	CB	21	31,310	31,470
10	O	26	31,250	31,470
11	CB	31	31,110	31,470
12	CB	21	32,220	32,450
13	O	26	32,980	32,450
14	CB	31	32,250	32,450
15	CB	51	32,340	32,450
16	O	56	32,260	32,450
17	O	26	43,270	43,450
18	CB	31	43,260	43,450

Table 3.1: Visually detected segments that might serve as precursors before the global events. Fourth column consists of the time steps of the start of the precursor segment. Fifth column consists of the time steps when the global event occurred

Interestingly, the precursor atoms are predominantly β-C and O atoms. These are located near the middle of the protein, which is in agreement with what we have learned from the time series clustering study. Out of the five global events in total, we observed

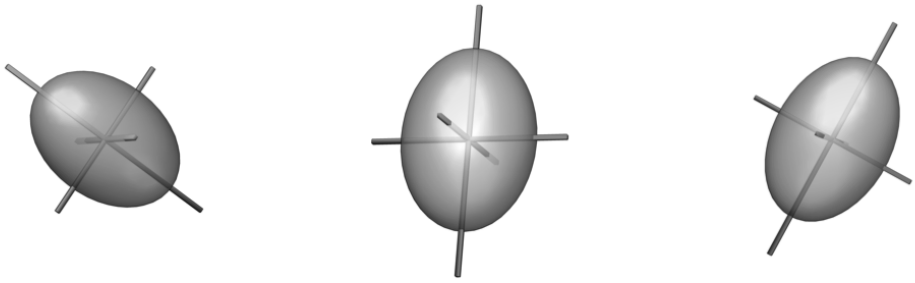

Figure 3.6: Fluctuation ellipsoids of CB(11) atom before precursor segment (left, from time step 1,650 to 2,159), within the precursor segment (middle, from time step 2,159 to 2,640), and after global event (right, from time step 2,640 to 2,996).

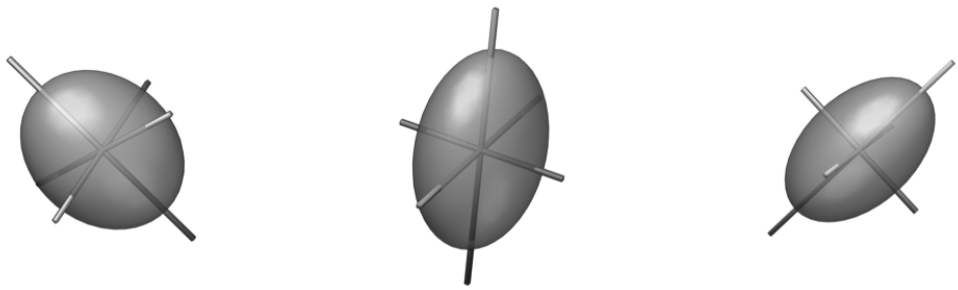

Figure 3.7: Fluctuation ellipsoids of CB(31) atom before precursor segment (left, from time step 524 to 2,354), within the precursor segment (middle, from time step 2,354 to 2,573), and after global event (right, from time step 2,573 to 3,042).

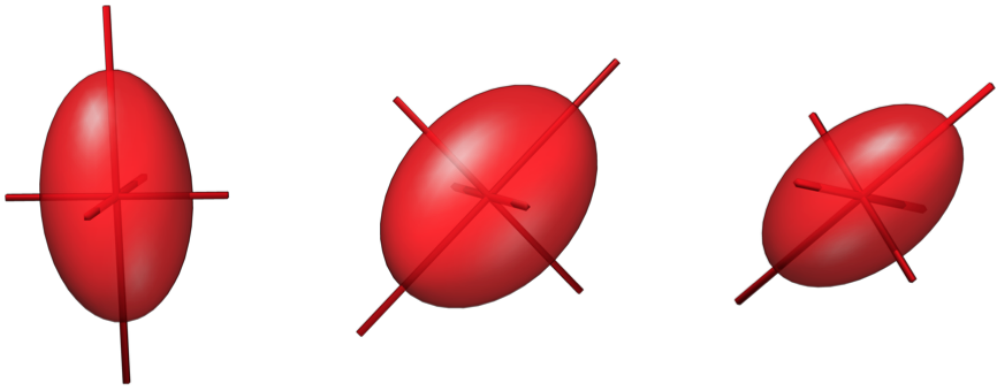

Figure 3.8: Fluctuation ellipsoids of O(16) atom before precursor segment (left, from time step 1,696 to 2,135), within the precursor segment (middle, from time step 2,135 to 2,649), and after global event (right, from time step 2,649 to 2,982).

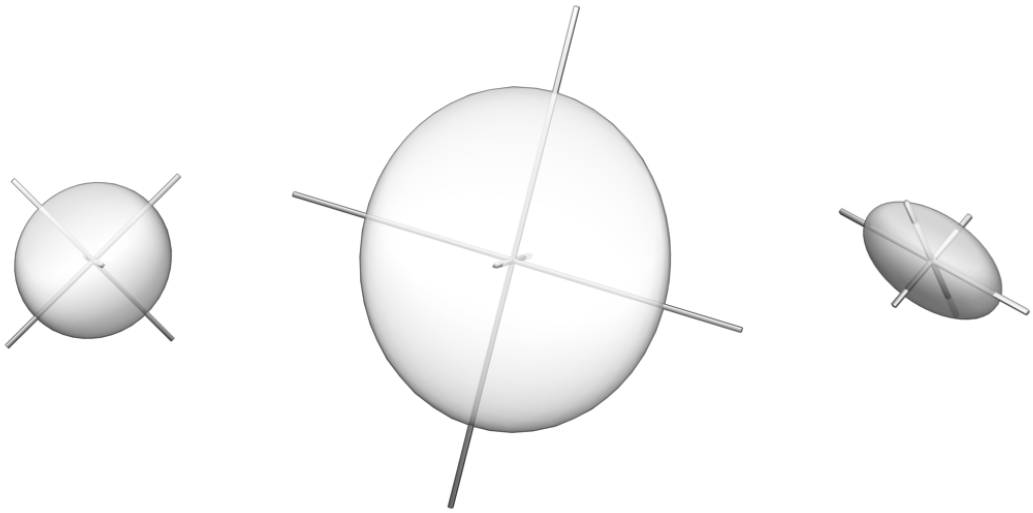

Figure 3.9: H(62) atom around bright burst near time step 16,100, showing fluctuation ellipsoids at: before (left, from time step 15,838 to 16,060), within (middle, from time step 16,060 to 16,090), and after the bright burst (right, from time step 16,090 to 16,351).

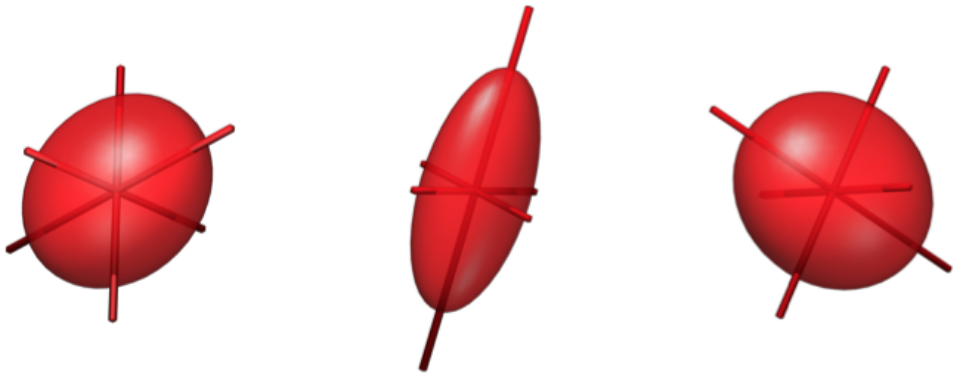

Figure 3.10: O(16) atom around bright red segment at time step 10,600, showing fluctuation ellipsoids at: before (left, from time step 10,377 to 10,590) within (middle, from time step 10,590 to 10,740), and after the red segment (right, from time step 10,740 to 11,770.)

four potential precursors involving CB(31) and three potential precursors involving O(26) and CB(21). These suggests that the folding processes must have started in the middle of the protein and thereafter propagates outwards. Fluctuation ellipsoids of precursor atoms leading to global event at time step 2,500 are visualized on Figures 3.6, 3,7 & 3.8.

We also visualized other segments, identified from the dendrograms, to better understand what happened during the evolution. Figure 3.9 shows the fluctuation ellipsoids of H(62) atom around a bright burst at time step 16,100, and Figure 3.10 shows the actuation ellipsoids of O(16) atoms around a bright red segment at time step 10,600. However, since we found no global events shortly following these segments, we do not consider them plausible precursors.

Finally, we calculated the correlation matrix C using equation (2.9) with the distance between two objects calculated using the formula 2.19a from subsection 2.2.3.3 and draw the Minimal Spanning Tree (MST) graph. As expected, the MST topology is identical to the structure of the protein, indicating that the strongest-correlating pairs are simply the pairs of atoms with structural bonds (Figures 3.11 & 3.12). These are the minimal spanning tree (MST) graphs of cross correlation matrix calculated between time steps 2,000 and 2,300, just before a global event at around time step 2,500. The topologies also remain the same for all MST graphs drawn at all different times in the simulation. However, we also found that MST graphs calculated from all other intervals before and after global events produce exactly the same topology, and thus do not provide further information. Apparently, the

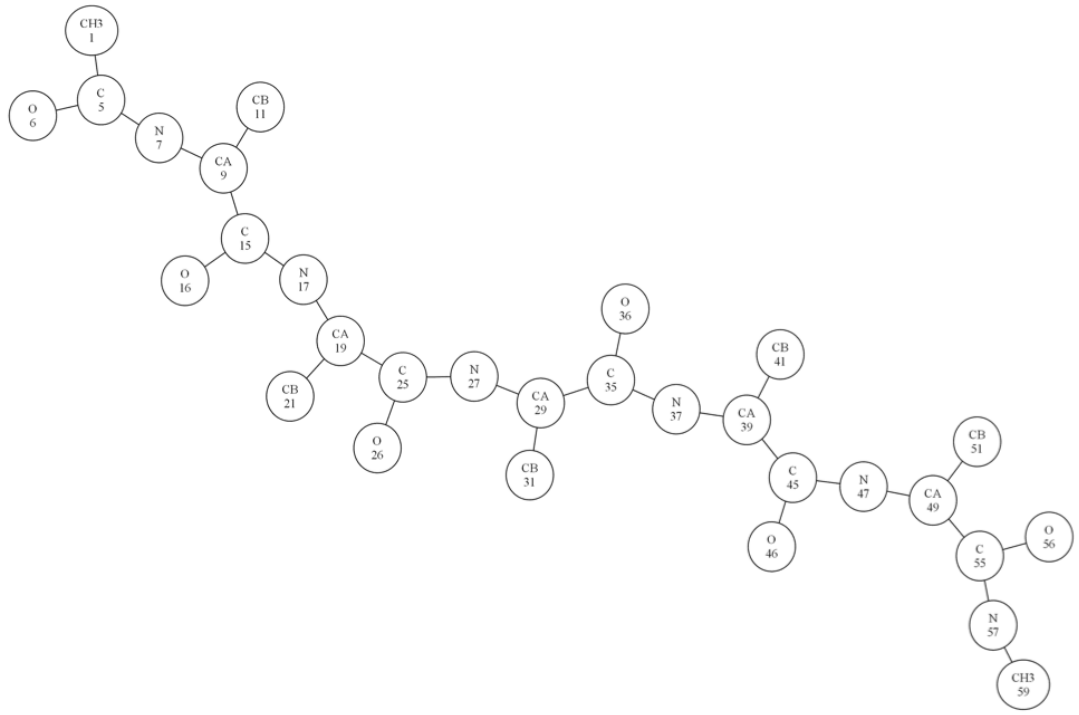

Figure 3.11: Minimal spanning tree of non-hydrogen atoms

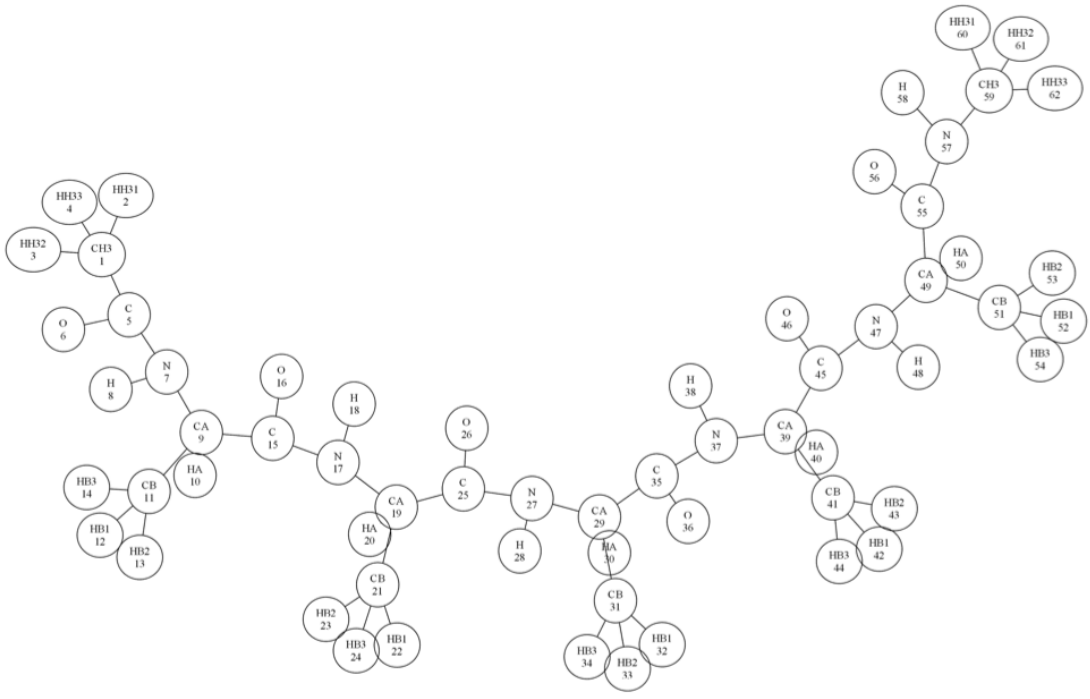

Figure 3.12: Minimal spanning tree of all atoms including hydrogen

cross correlations between the structurally-bonded atoms are always very high, such that any changes to the cross correlations between other pairs of atoms are constantly overwhelmed and do not show up in the MST graphs.

3.2 Three proteins

3.2.1 Hierarchical Clustering

In this part we discuss our study on another data-set, described in the subsection 2.1.1.2. It consists of three poly-alanine peptides and in the previous part this molecular dynamic simulation data-set contains the coordinates (x_i, y_i, z_i) of each atom in the peptides. It spans across 25.0 ns and consists of 25000 steps, each step being 1.0 ps.

The molecules in this study are bigger than in the previous, interactions between different parts of the molecule are much more complex (Fig. 3.13) and there is no such easy classification of the molecule structures as we obtained in previous section. Hence, we decided to proceed to automatic analysis of the clustering matrices and dendrograms. To do this we are again clustering the matrices of clusters obtained from the molecular simulation time series.

From the dendrogram on a figure 3.14 we can see that the structure of the molecule during 1 ns is very different from the rest of the time. Similarly, on a Figure 3.15 we can see, that the structure of the molecule during approximately first 3 ns is different from the later time. We can conclude then, that some principal changes happen around the mentioned time periods. Unfortunately, we cannot identify such transition in structure for the third molecule from the presented dendrogram.

3.2.2 Pairwise correlation

During the first part of our study we noticed that the results based on the average velocities are not giving good insights. It might have been because velocities are not independent (they are dependent on covalent bond which determine basic structure of the protein). For an alternative we draw eyes on the dihedral angles (Fig. 3.17) - they are not affected by the bond length. The dihedral angles are fined as a cross product between vectors of direction of the

- C-carbon and C-nitrogen bonding for the main chain angles and
- C-C and C-nitrogen bonding for the side chain angles.

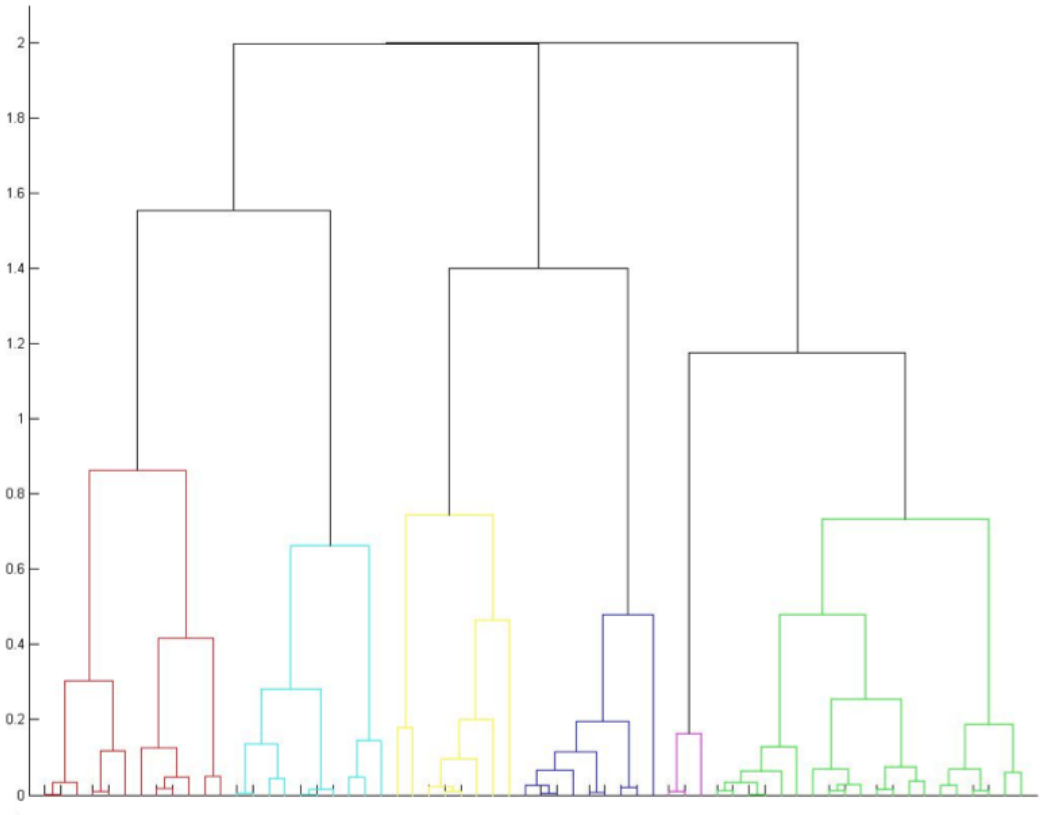

Figure 3.13: One of the complete-link dendrogram (representing the interval between t_1 = 12001 ps to t_2 = 12500 ps.) of cross correlation between atoms for of the Q molecule. It shows several clusters of approximately same size and not very robust structure. Note that the red and blue clusters shrink while the interaction green cluster grows in size, indicating strong interaction between atoms across the protein.

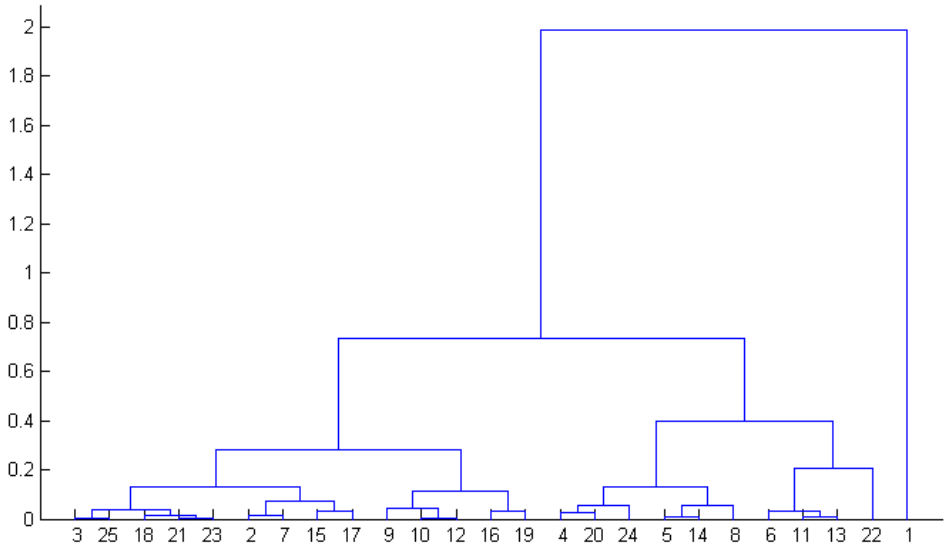

Figure 3.14: Dendrogram showing clustering analysis of the dendrograms sequentially produced for the time intervals with 1 ns step of the Q molecule simulation. On the x-axis each node represents a dendrogram computed for one of 25 time windows.

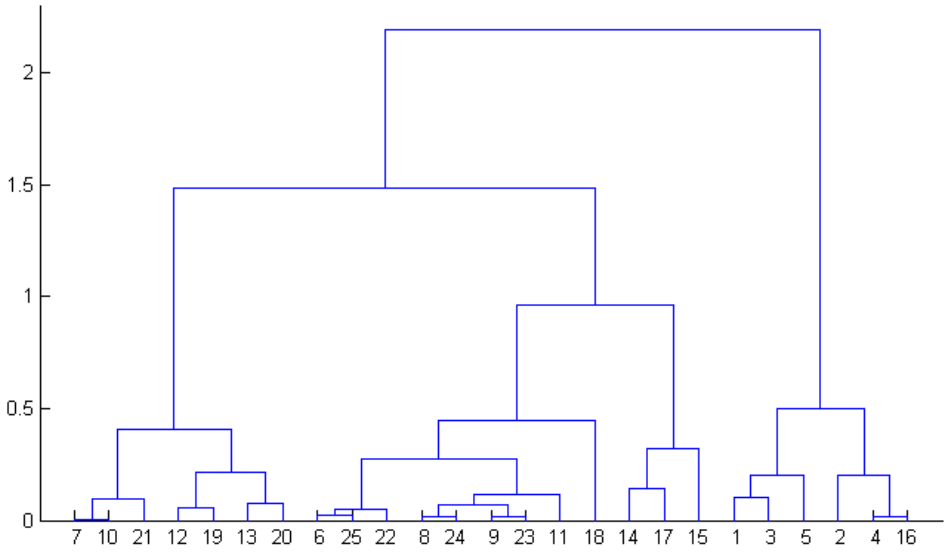

Figure 3.15: Dendrogram showing clustering analysis of the dendrograms sequentially produced for the time intervals with 1 ns step of the K molecule simulation.

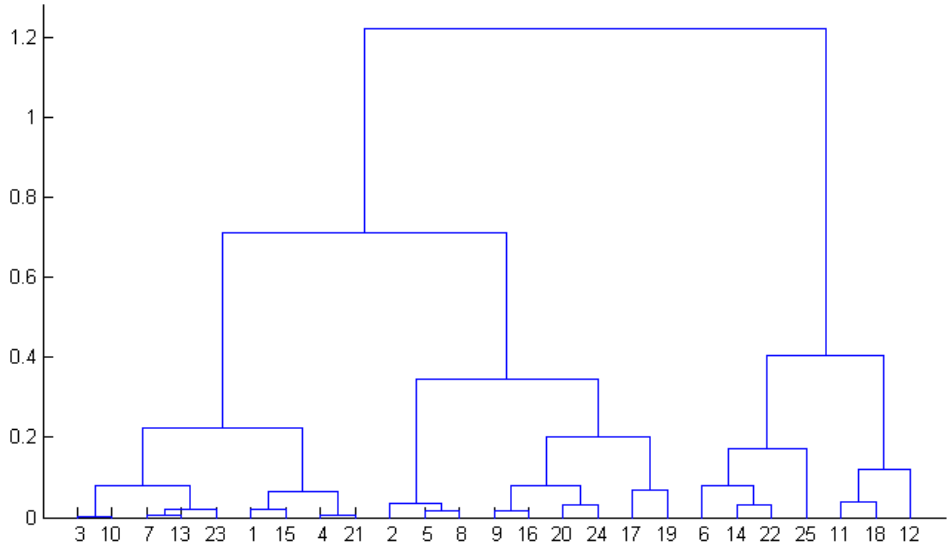

Figure 3.16: Dendrogram showing clustering analysis of the dendrograms sequentially produced for the time intervals with 1 ns step of the D molecule simulation.

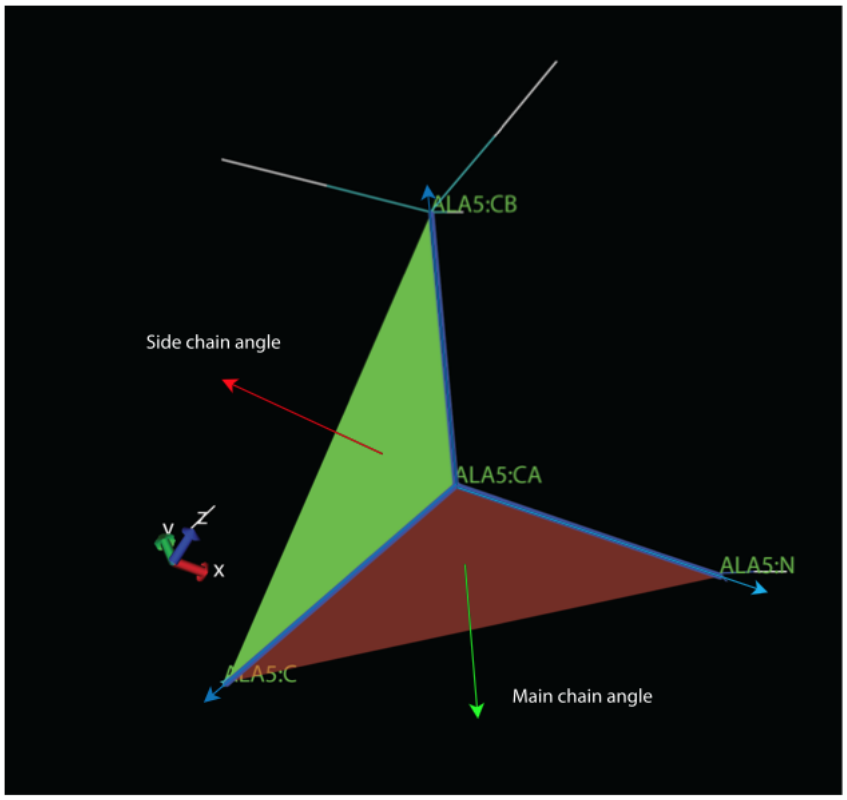

Figure 3.17: Illustration of the dihedral angles in an alanine residue. Plotted with VMD

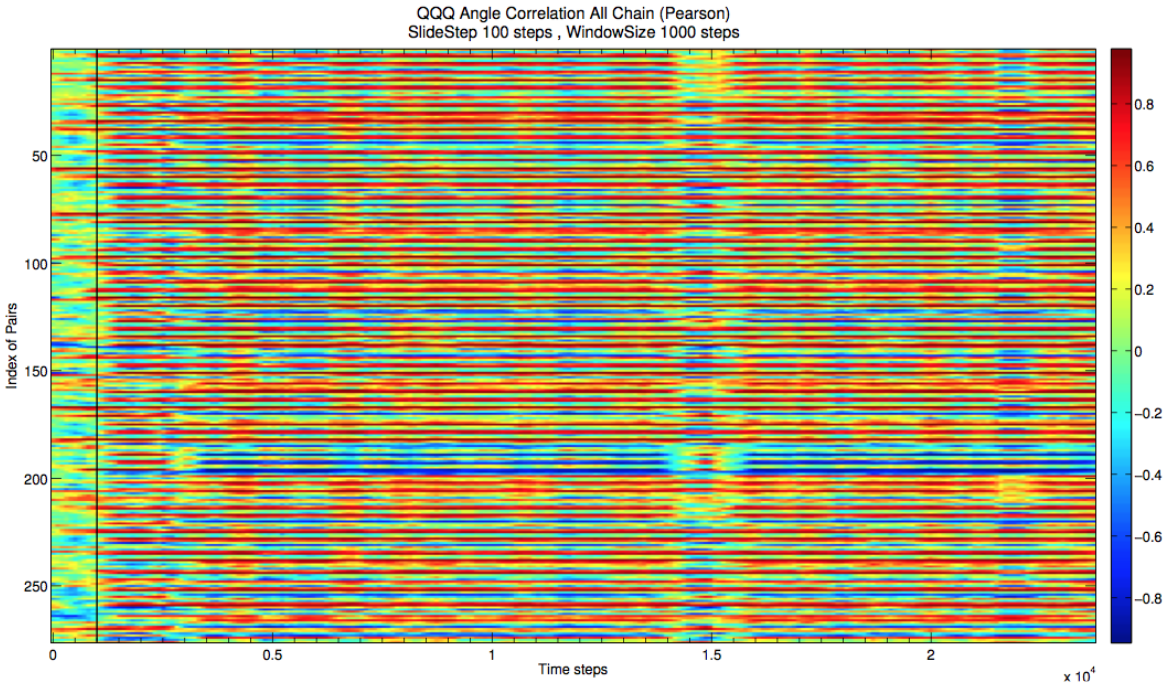

Figure 3.18: The pairwise correlation time-series for Q molecule

We calculated the pairwise correlation between all dihedral angles then plot the results: Figure 3.18 for Q molecule, Figure 3.19 for K molecule and 3.20 for D molecule. Here red colour represents positive correlation and blue colour represents negative correlation. The saturation of colour represents the strength of the correlation.

On the colour-maps 3.18, 3.19 and 3.20 we can observe sharp transition in colour after 1 ns for Q, K and D molecules many pairs are becoming more correlated, both positively and negatively. These correlations remained extremely stable towards the end of the time-series. These stable pairs represent the characteristic of the folded state of the proteins because a protein in the folded state is expected to be highly constrained both in its shapes and the pairs' interactions. These are the fingerprints of α-helix folding.

These colour-maps are showing a good agreement with the clustering analysis results. On the dendrograms we could see sharp separation of two groups of sub-clusters for Q and K molecules: those that represent dendrograms computed for the first few ns and all the rest. In the mean time, clustering of the dendrograms for the D molecule shows lower dissimilarity level between different sub-clusters and temporally less ordered structure. Additionally, it tells us about the evolution of the state of the molecules.

Once reached, strong correlation state does not change again and stays rather stable until the end of the simulation. Such stable correlation between pairs is signifying that the folded state is reached - once protein reach the folded state it is known to be very stable in shape,

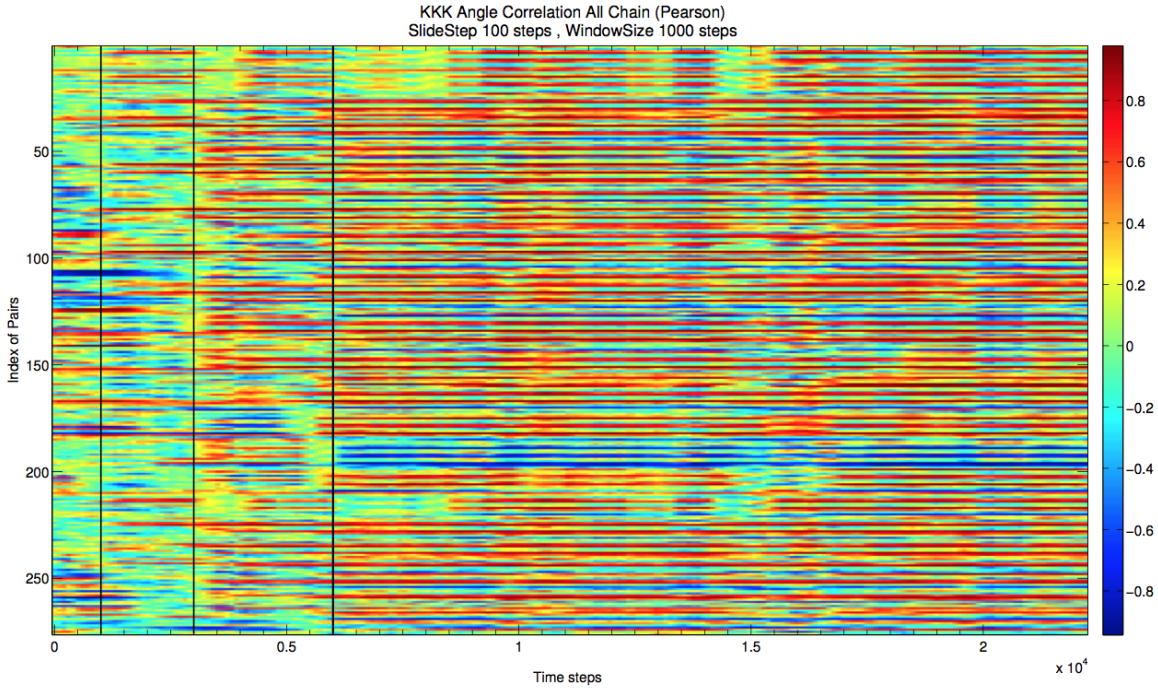

Figure 3.19: The pairwise correlation time-series for K molecule

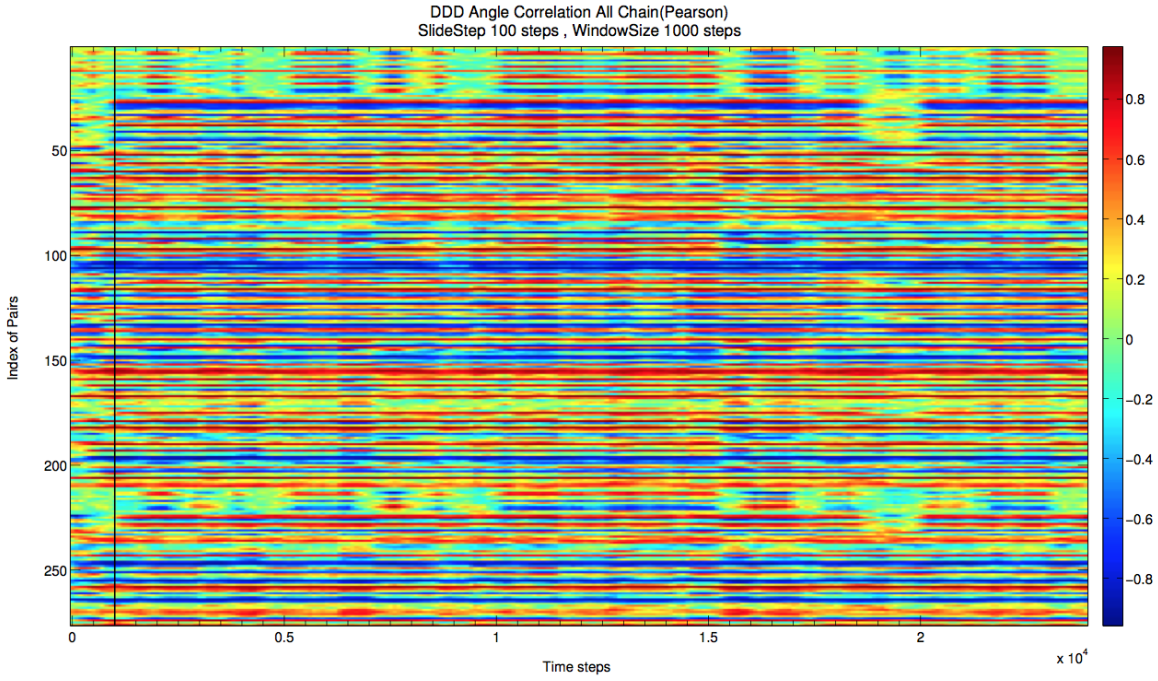

Figure 3.20: The pairwise correlation time-series for D molecule

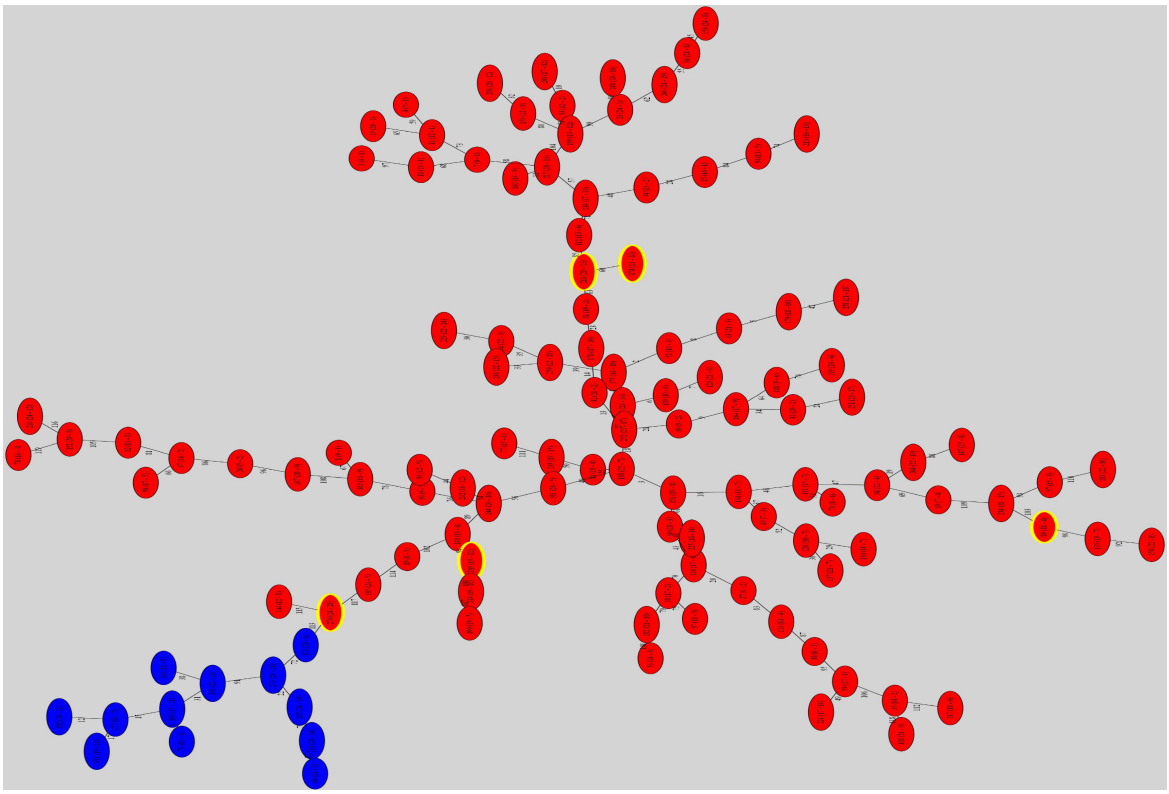

Figure 3.21: Minimal Spanning Tree diagram for protein Q.

and as a consequence, in pairs interactions.

3.2.3 Minimal Spanning Tree

To study interrelations between different pairs we visualize the discussed correlation results using the Minimal Spanning Tree graphs: Figures 3.21 for Q, 3.22 for K and 3.23 for D molecules. The red colour represents the fingerprints with positive correlation and the blue colour represents the fingerprints with negative correlation, yellow represents unique fingerprints.

From the MST graphs, we can observe that the positively correlated pair tend to be linked together, so are the negatively correlated pairs. Q and K molecules manifest many more pairs with positive correlation. D molecule, in contrast, have almost the same amount of the pair with positive and negative correlation. From here we can conclude that for the successful α-helix folding needs more positively correlated fingerprints.

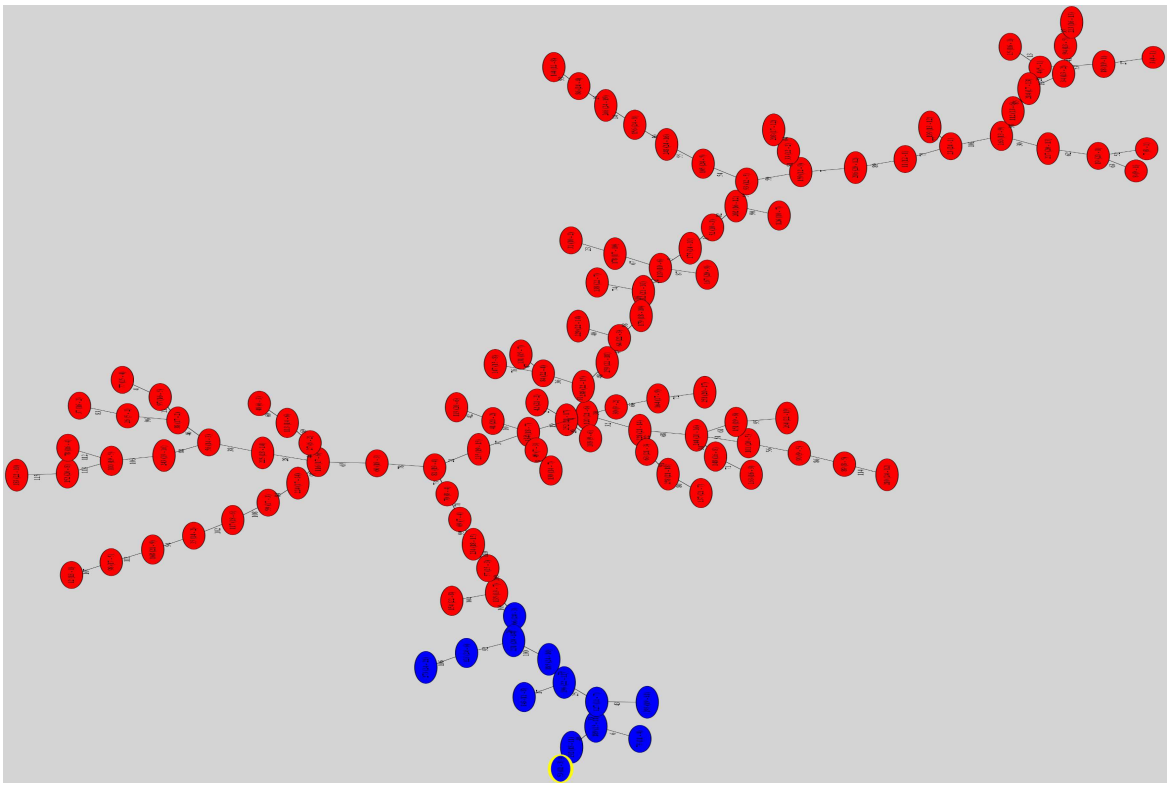

Figure 3.22: Minimal Spanning Tree diagram for protein K

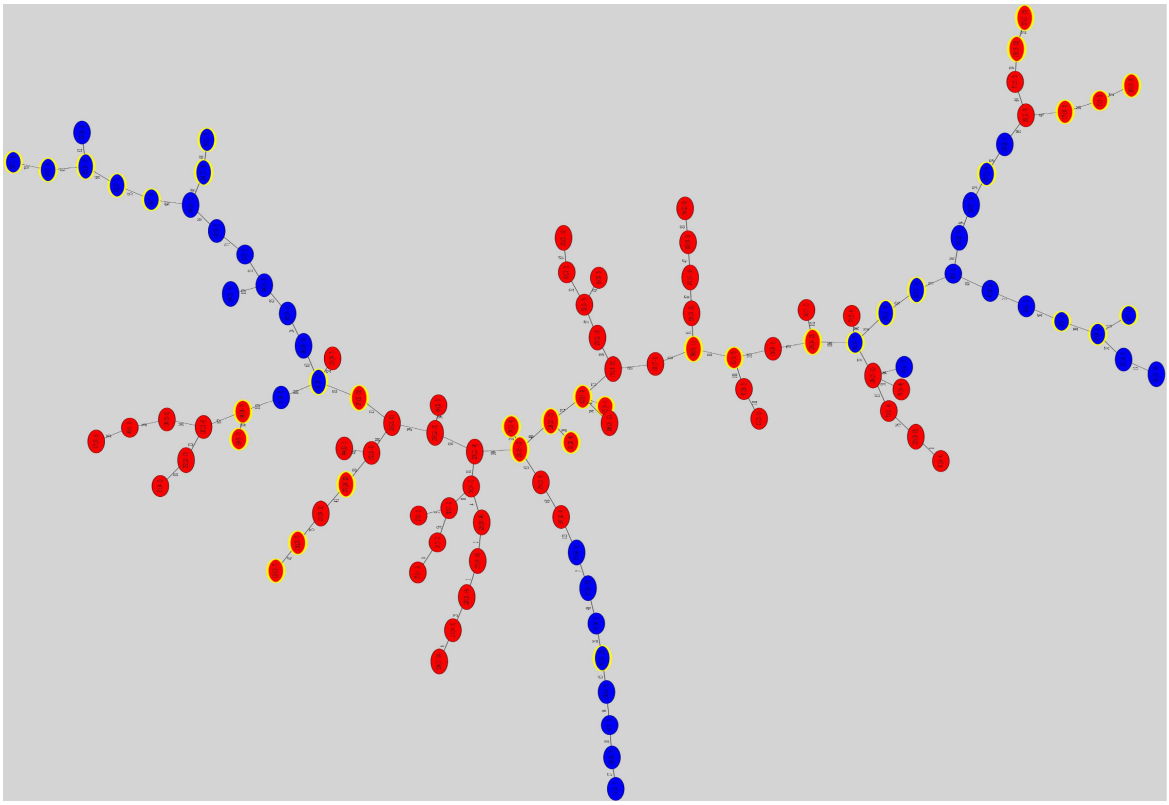

Figure 3.23: Minimal Spanning Tree diagram for protein D

Chapter 4

Dynamics of the Tropical Atmosphere. Madden Julian Oscillation.

In this chapter we study the atmospheric dynamics of Madden Julian oscillation and its interaction with monsoons. After introduction of this atmospheric phenomenon and its main characteristics, we show that it is possible to capture its pattern by using clustering technique on rainfall data. We start with inter-monsoon periods and compare our results with the classic RMM index. Later we investigation the MJO interaction with another principal phenomenon of the region - monsoon.

4.1 Introduction

MJO attracts interest for many reasons: it affects the fluctuations in rainfall over the Pacific Islands, over South and South East Asia [182, 183, 184], over Australia [185], along western coast of North America [186, 187, 188] and South America [189, 190] and over Africa [191]. Madden Julian Oscillation also affects the development of the tropical cyclones over the Pacific Ocean and the Caribbean [192, 193, 194, 195, 196], winds over the Atlantic Ocean [197] and the intensity of the convergence zone of the Southern Hemisphere [198]. Since this oscillation is a reason of periodic changes in emissions of heat into the atmosphere, it is a component of the mechanism of signal transduction from the tropics to temperate latitudes [199, 200, 201, 202]. It can also modulate the global angular momentum [203, 204, 205], and the electric and magnetic field of the Earth [206]. MJO interacts with other phenomena, such as monsoons and El Niño [207, 208, 209]. Interestingly, MJO might even left a mark in the history of mankind: strong irregular westerly surface winds, that last up to 30 days (currently known as part of the MJO fluctuations), could be those es-

sential weather conditions that helped brave Polynesian fishermen 4,500 years ago to travel to the east in the equatorial Pacific (in areas, usually dominated by the trade winds) [210]. It should be noted that MJO is one of the reasons for limited predictability in atmosphere both in nontropical and tropical latitudes [211, 212, 213, 214, 215].

Period of the MJO

Dominant period of the Madden Julian Oscillation is from 30 to 100 days, although the period, which accounts for maximum variability varies widely within time. Although this phenomenon is called "oscillation", in fact it is very irregular. MJO appears sporadically and discrete time [216].

Planetary scale

The typical length of zonal phenomenon MJO, which is defined as the extent of the areas occupied by positive or negative anomalies cloud cover ranges from 12,000 to 20,000 km [217]. Typically, in the tropics at a particular time, there is only one full phenomenon MJO. Infrequently there may be two weak phenomenon MJO, when one is in its infancy over the Indian Ocean, and the second fades in the central Pacific Ocean [164].

Easterly motion

Progression from west to east (at a rate of about 5 m/s) is one of the main characteristics of MJO, allowing to distinguish it from other propagating phenomena in tropics. For example, equatorial Kelvin waves, that also propagates to the east, but with a higher rate (on the order of 15-17 m/s) [210, 217]. The speed of propagation of the MJO could slightly change from cycle to cycle and during the various phases of the single cycle.

Seasonal variation

MJO is experiencing a pronounced seasonal changes in both intensity and latitudinal localization [215, 218, 219, 220]. The main seasonal intensity maximum occurs during southern hemisphere summer and autumn - the strongest signal is then observed to the south from the equator. The main peak of intensity during the southern hemisphere summer is associated with the Australian monsoon [184]. Seasonal latitudinal migration of MJO is expressed over the western Pacific better than over the Indian Ocean. In a narrow latitudinal band of 5 degrees north to 5 degrees south MJO is experiencing only one maximum in the seasonal course attributable to the southern hemisphere summer and autumn. In the eastern Pacific Ocean single maximum also exists - during northern hemisphere summer.

Inter-annual variability

Inter-annual variability in the zonal wind over the Pacific is better expressed in the lower troposphere than in the upper [221]. During the warm phase of ENSO (El Niño), the eastern end of the basin of warm water shifts further to the east [222], followed by the displacement of MJO area [206, 223, 224, 225, 226, 227]. MJO in the Pacific ocean is particularly intense in the period preceding the El Niño and is weakening after its culmination.

Interaction with monsoon

First work on the influence of the Madden Julian Oscillation on the summer monsoon was published by Yasunari in 1979 [178]. Since then, with the availability of new satellite data we learned more about this interaction.

The influence of the Madden Julian Oscillation on boreal summer and boreal winter rainfall variability is different: during the summer it manifest itself in extra-equatorial regions: South China Sea, Bay of Bengal and Western North Pacific. Maximum of the mean rainfall during the summer occurs in these regions. During this time, Asian summer monsoon is not only experience 30 to 60 days variations, but also 10 to 20 days quasi bi-weekly variations [167]. While most of the state-of-the-art models of global climate fail to simulate MJO, others, that are implementing coupling principles, were able to resemble some aspects of the MJO [228, 229] and even predict its behavior [230, 231].

Around one week (5 to 10 days) before a a strengthening of the precipitation over South Asia, that is believed to be related to the MJO, the sea surface temperatures (SST) are tend to be higher than usual [179]. Break and active phases of the Asian Monsoon are also connected to the MJO. During the break phase of monsoon MJO is usually found to be propagating towards the east over Indian ocean, Maritime continent and further into Pacific region [179].

RMM index

The most commonly used MJO index known as Real-time Multivariate MJO (RMM) index. The index was developed and introduced by Wheeler and Hendon [164]. RMM daily time resolution with all-season and global scale coverage is widely considered in many studies about regional weather impact of MJO , especially to rainfall [170, 171, 172, 173, 174, 175, 176, 177]. Furthermore, the approach of confirming MJO properties of RMM that was demonstrated by Wheeler and Hendon was utilized by U.S. Climate Variability and Predictability (CLIVAR) MJO Working Group in assessing the performances of Global Circulation Model (GCM) in simulating MJO event [177]. Based on such advantages RMM is

involved in this study as a benchmark of MJO activity.

4.2 MJO identification

The MJO does not oscillate regularly: the life cycle of a canonical MJO event is described in previous section, however, the location of initiation and propagation characteristics of individual MJO events vary significantly. In this section we will discuss methods that we use to identify and analyze an MJO event.

In chapter 2, we declared that our choice of the dataset is dictated by the interest in intraseasonal phenomena: Madden Julian Oscillation and Asian monsoon. We have chosen neutral years for the inter-seasonal phenomena like El Nino Southern Oscillation and Indian Ocean Dipole. Within these years we will first study Madden Julian Oscillation separately. To do this we will examine behavior of this phenomenon during inter-monsoon seasons: April-May and October-November. In the literature, the boundaries of the monsoon and inter-monsoon season are defined differently for different locations [205, 206, 209] and may change from year to year. In our work we analyze data over vast region, that includes patterns of both Indian and Australian monsoons and due to this reason we define monsoon season as a union of all subsets.

We start with the satellite rainfall data described in chapter 2, which is initially in a form of matrix, grid. It covers regions from 50 degrees south to 50 degrees north and from 30 degrees east to 180 degrees east with a spatial resolution 0.25-degree by 0.25-degree. Temporal resolution is 3-hourly. So, for every 3 hours we have a matrix with the values of the rainfall and at the first step, we transform a matrix into a vector. Then we form new matrices from such vectors. These new matrices cover 25 days in sequence and hence consist of 200 vectors. In our further studies, for better tracking of the signal, we overlap these matrices by 5 days, meaning 40 vectors. Then we shift the time window by 1 day or 8 vectors.

We begin with computing PCC for each 25-day rainfall matrix and visualizing the results as, for example, on the Figure 4.1. Dark blue means low correlation between rainfall data points, while red mean high correlation . We then use correlation matrix and proceed with the hierarchical clustering analysis (described in section 2.2.3.4), plotting the dendrogram for each time window of 25: example on the Figure 4.2. Based on the results of clustering analysis we then reassemble the correlation matrix in such a way that data points on the axes are lined according to their hierarchical clustering structure. As it could be seen from the comparison of figures 4.1 and 4.3, the size of rainfall pattern is becoming more visible.

After that we construct clustering histogram, as on Figure 4.4. We first extract part of

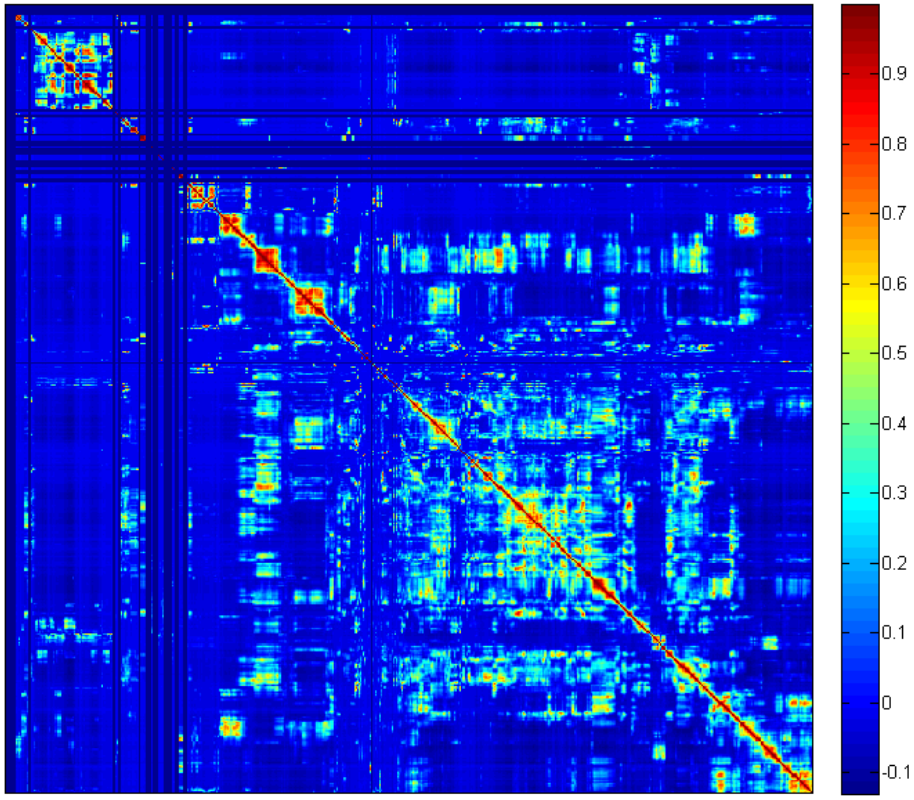

Figure 4.1: Unordered correlation matrix for rainfall data from 1 April to 25 April 2005

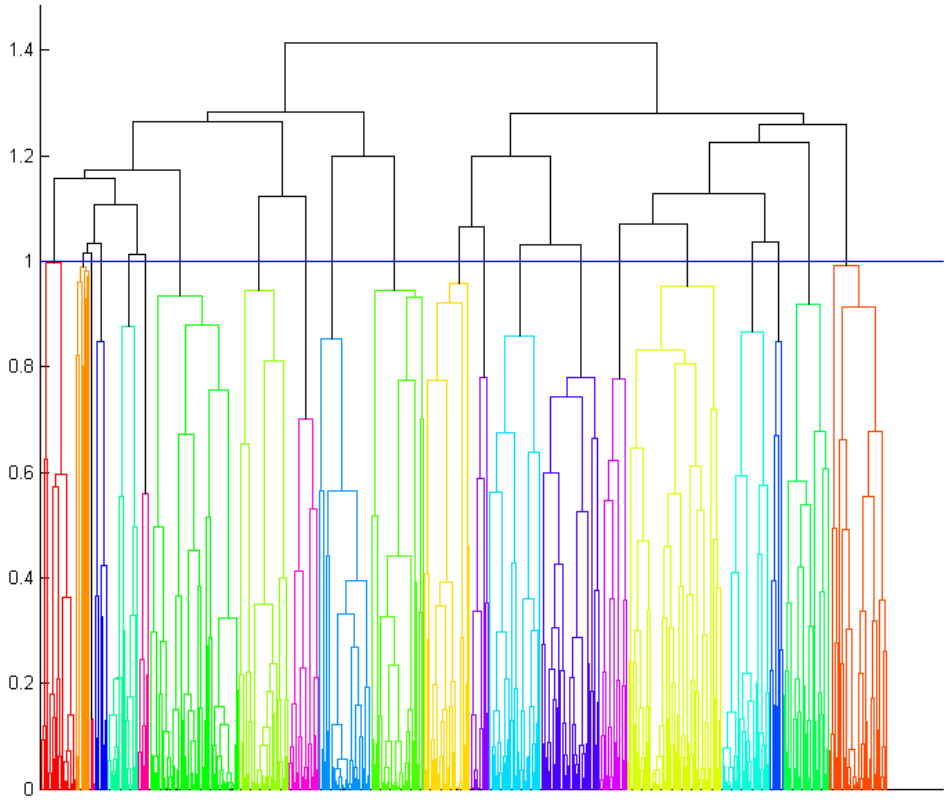

Figure 4.2: Dendrogram based on rainfall data for 1 April to 25 April 2005

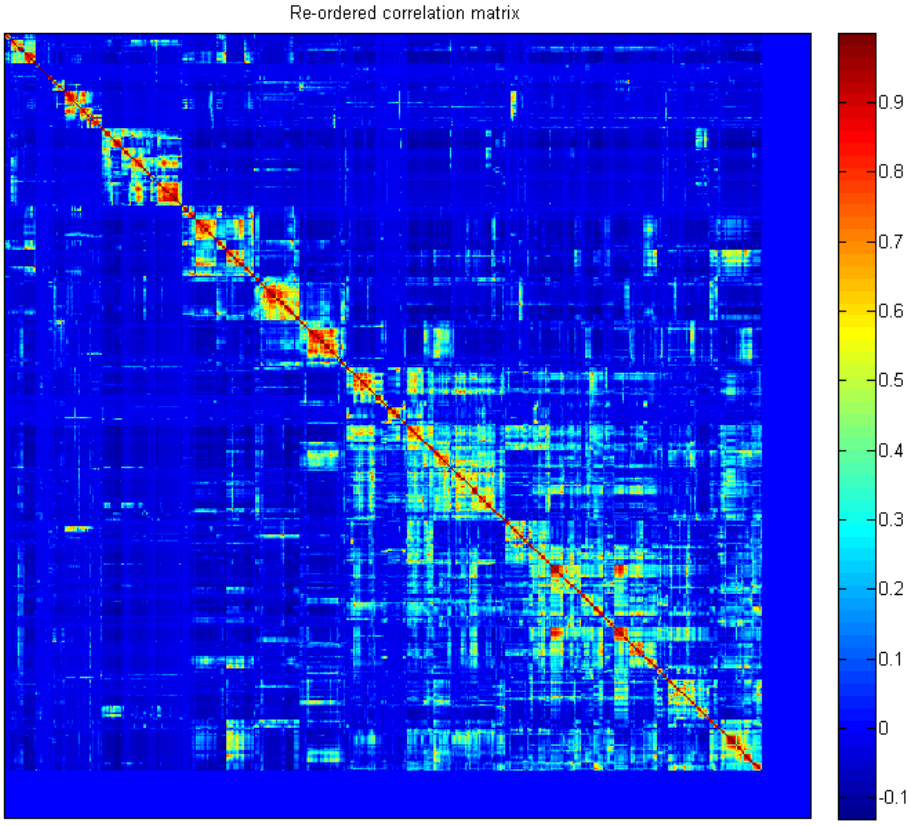

Figure 4.3: Reordered correlation matrix for rainfall data from 1 April to 25 April 2005

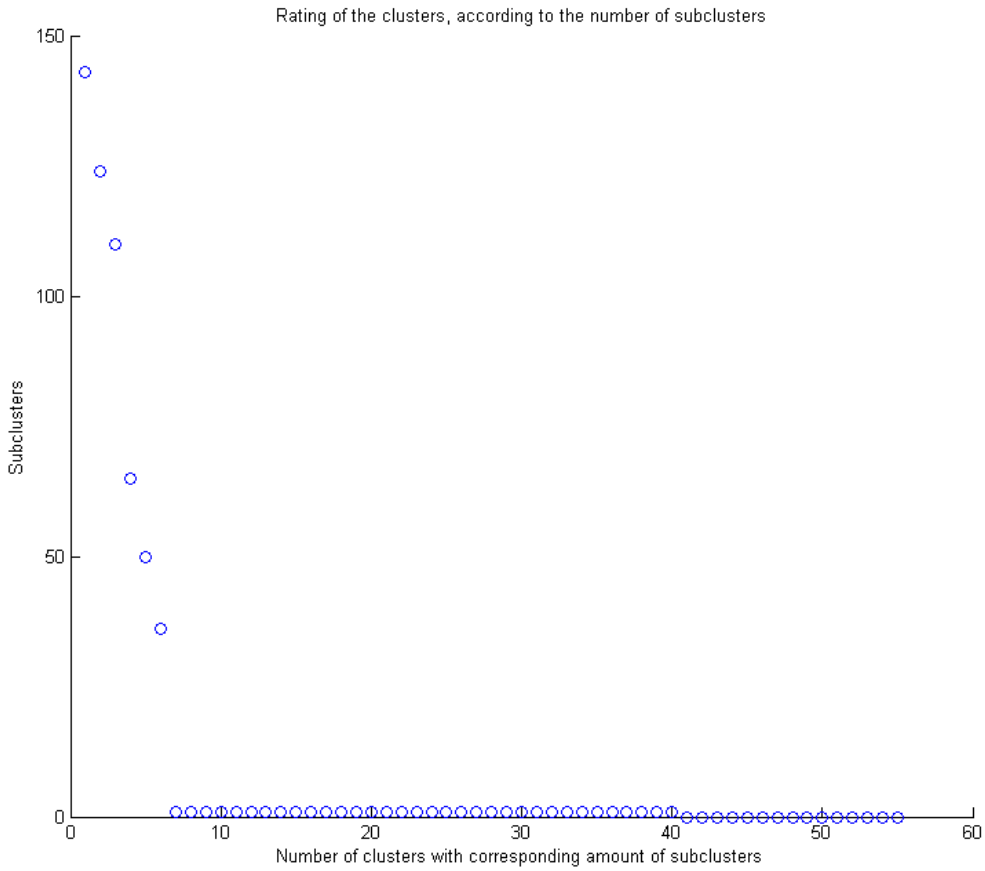

Figure 4.4: Cluster distribution, according to the number of robust sub-clusters for 1 April to 25 April 2005

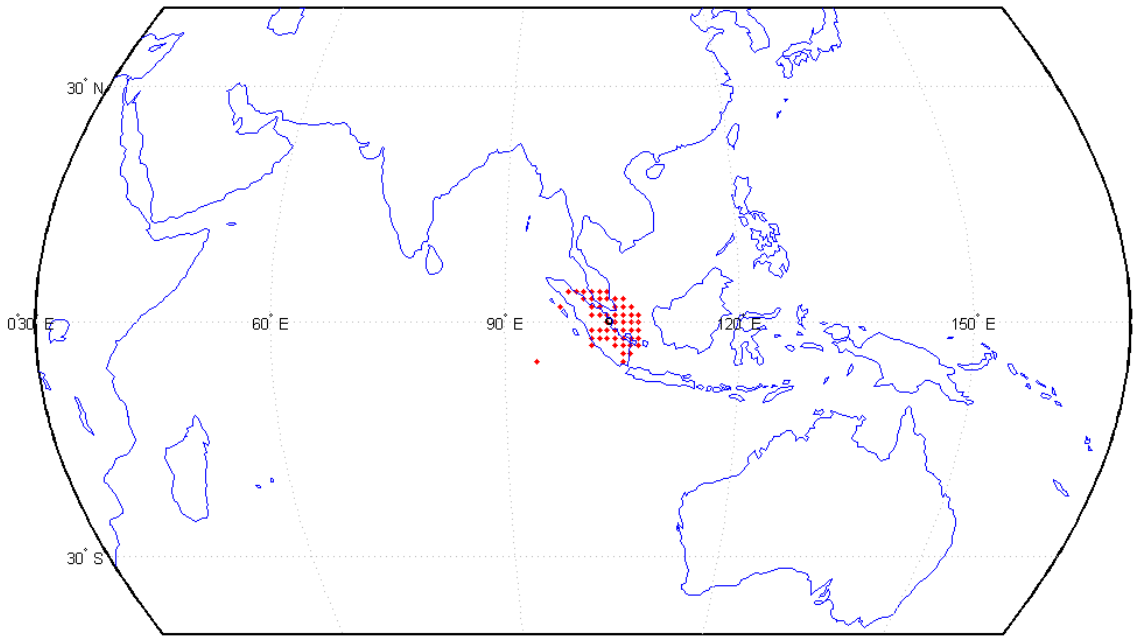

Figure 4.5: Elements of the biggest cluster (red dots) and center of mass of this cluster (black circle) for 1 April to 25 April 2005

the clustering matrix that describes clusters with having correlation more than 0.6. Then we simply plot the distribution of this clusters according to the number of sub-clusters they have. For example, from the Fig. 4.4 we can see, that there are in total 6 clusters, that have more than 100 sub-cluster each.

Finally, we visualize geographically structure of the clusters. For example, on the Figure 4.5 you can see the elements of the biggest rainfall cluster for the time interval 1 to 25 of April, 2005. Black circle represents the center of mass of the cluster. We calculate it for the every iteration in order to be able to track the evolution of clusters.

4.2.1 Inter-monsoon seasons

Our first goal was to capture the signal of the Madden Julian Oscillation. In order to do this we analyzed the rainfall time series over the neutral years during the inter-monsoon seasons. We followed the procedure described in the previous section, plus calculated the correlation matrices for the whole inter-monsoon season. On the Figure 4.6 one can see re-ordered correlation matrices for the April-May inter-monsoon seasons of 2000, 2001, 2003, 2004, 2005 and 2008.

Several patterns could be noticed here:

- for the years 2004 and 2005 big patterns are present: they are not uniformly highly correlated, but cover most of the time window;
- for the year 2000 we can see smaller, yet more strongly correlated rainfall pattern;
- finally, years 2001, 2003 and 2008 have only small patterns over this inter-monsoon season.

Let us compare these results with the distribution of the clusters (Figure 4.7). Interestingly, for the years 2004 and 2005 biggest clusters consist of 100 and more members, while for the other years this number is smaller.

Let us now consider October-November inter-monsoon period. The correlation matrices are plotted on the figure 4.8 and the clustering distributions are depicted on the figure 4.9. We cannot see such clear patterns as in previous case, only on the plots for years 2001, 2003 and 2004 small inter-correlated patterns are present, however they do not resemble any connection with the distributions of clusters. Let us note also that during the October-November inter-monsoon season of year 2005 correlation matrix looks very messy with many vectors being correlated, however these correlations are not very strong.

For better understanding of these results, let us compare them with established MJO index that is traditionally used for tracking the oscillation.

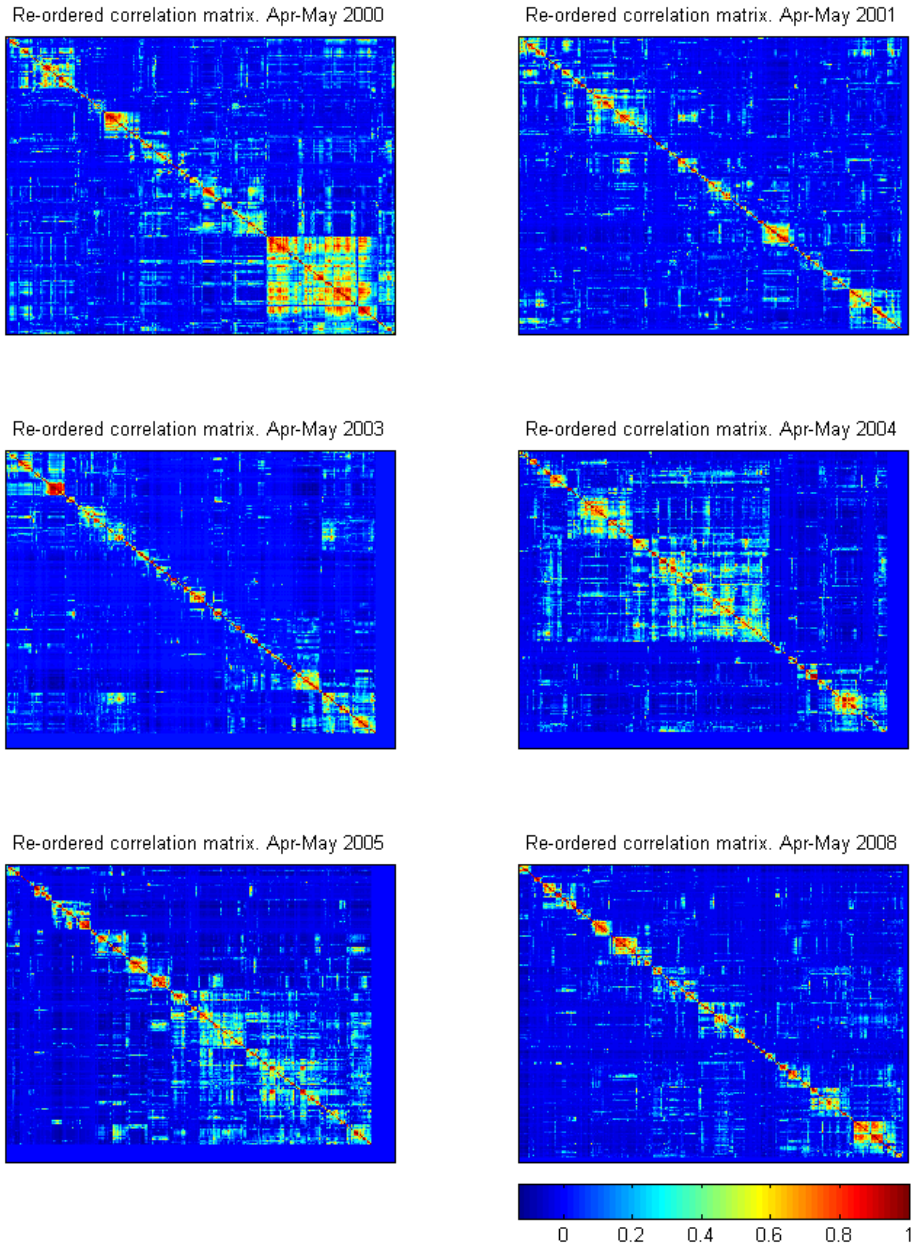

Figure 4.6: Re-ordered matrices for Apr-May

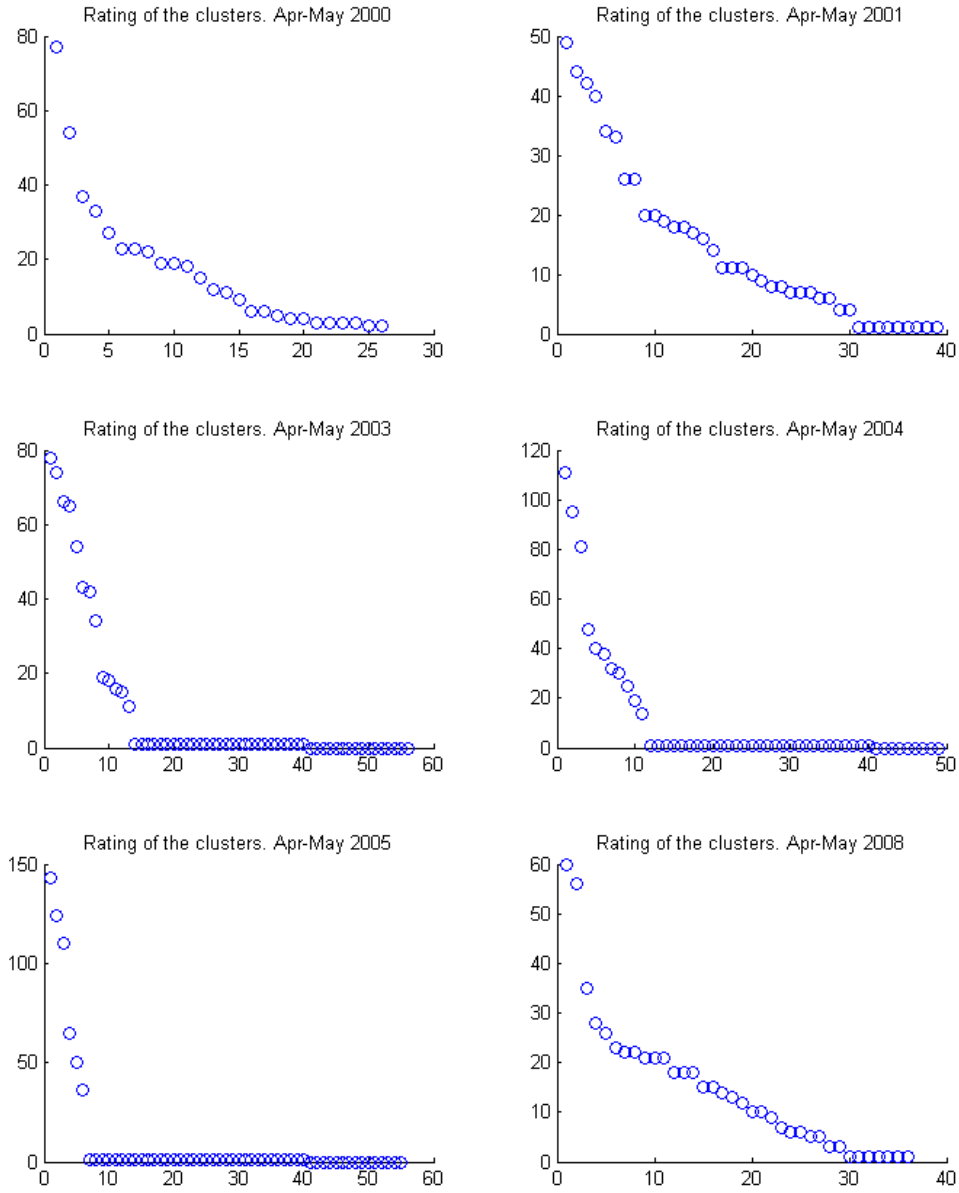

Figure 4.7: Cluster distribution, according to the number of robust sub-clusters for April-May

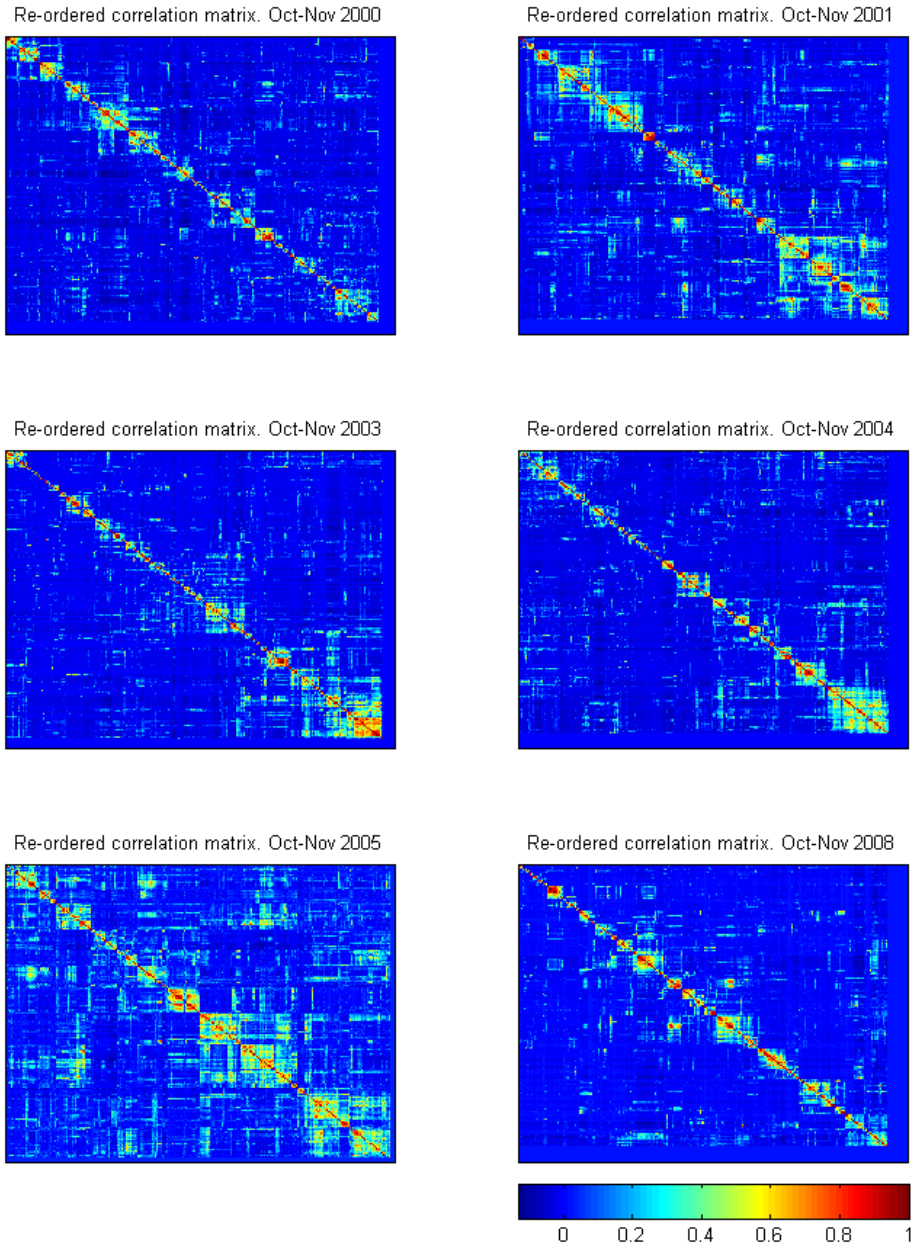

Figure 4.8: Re-ordered matrices for Oct-Nov

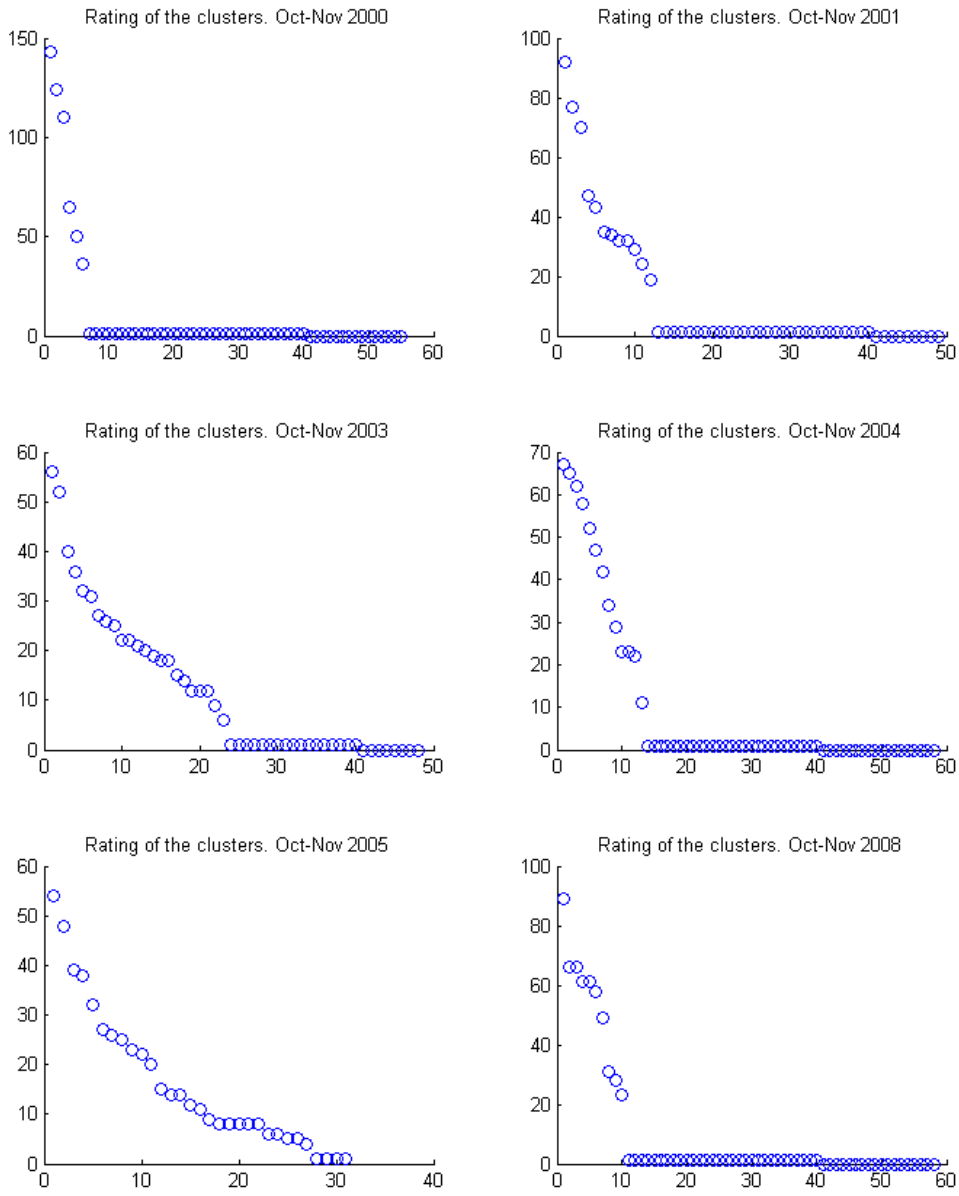

Figure 4.9: Cluster distribution, according to the number of robust sub-clusters for October-November

4.2.2 Comparing with RMM index

RMMs (RMM1 and RMM2) are derived from the first two Empirical Orthogonal Functions (EOFs) of the combined fields of: satellite-observed outgoing long-wave radiation (OLR), 200 hPa zonal wind and near-equatorially averaged 850 hPa zonal wind. The principal components are then extracted from the projections of the daily observed data onto EOFs. The annual cycle and the components of inter-annual variability are removed from these projections. As a result these principal components are mostly varying on the intraseasonal time scale and could be used as an effective index for real time use, while the the projection is used as a filter for the MJO signal. The time series of the form the index Real-time Multivariate MJO (RMM) and are separately called RMM1 and RMM2. To calculate RMMs we used the NCEP/NCAR Reanalyses and the NCEP Operational analyses for the zonal winds and the NOAA polar-orbitting satellite-based time-series for the OLR.

Figure 4.10 shows the spatial structures of the first two EOFs of the combined fields, which together account for 25 % of the total variance. It is important to construct a physical interpretation of the EOFs. The positive or the negative phase of EOF 1 represents enhanced and suppressed convection over the Maritime Continent respectively with low-level westerlies behind the convective center and low-level easterlies ahead of it; winds in the upper-troposphere are in the opposite direction to the low-level winds. The positive or the negative phase of EOF 2 represents enhanced or suppressed convection over the West Pacific and suppressed or enhanced convection over the Indian Ocean respectively.

Since the MJO is a propagating signal, the two eigenvectors have a quadrature relation in their spatial structure while the projections of the MJO on the two eigenvectors (RMM1 and RMM2) - are in temporal quadrature and possess roughly equal variance (Fig. 4.10 and 4.11).

The availability of this index in real time has facilitated the monitoring and prediction of the MJO and its various impacts but some limitations have been noted. One limitation stems from the use of just two EOFs to define the MJO, which necessarily just depict its canonical large-scale structure. The tacit assumption in the approach of RMM methodology is that the MJO has a consistent broad-scale expression in circulation and convection, which will be efficiently detected by the RMM indices. However, not every MJO event evolves with the same structure, even at the largest scales [234], so the use of a single pair of EOFs cannot capture all the nuances of every MJO event, especially at smaller scales where the MJO expresses itself on local weather. Development of the RMM indices without the use of a band-pass filter also means that there will be some contamination from higher-frequency "noise" [235]. The RMM indices may also not be optimal for detecting the initiation of some MJO events when a large-scale circulation signal is absent [236]. It has also been

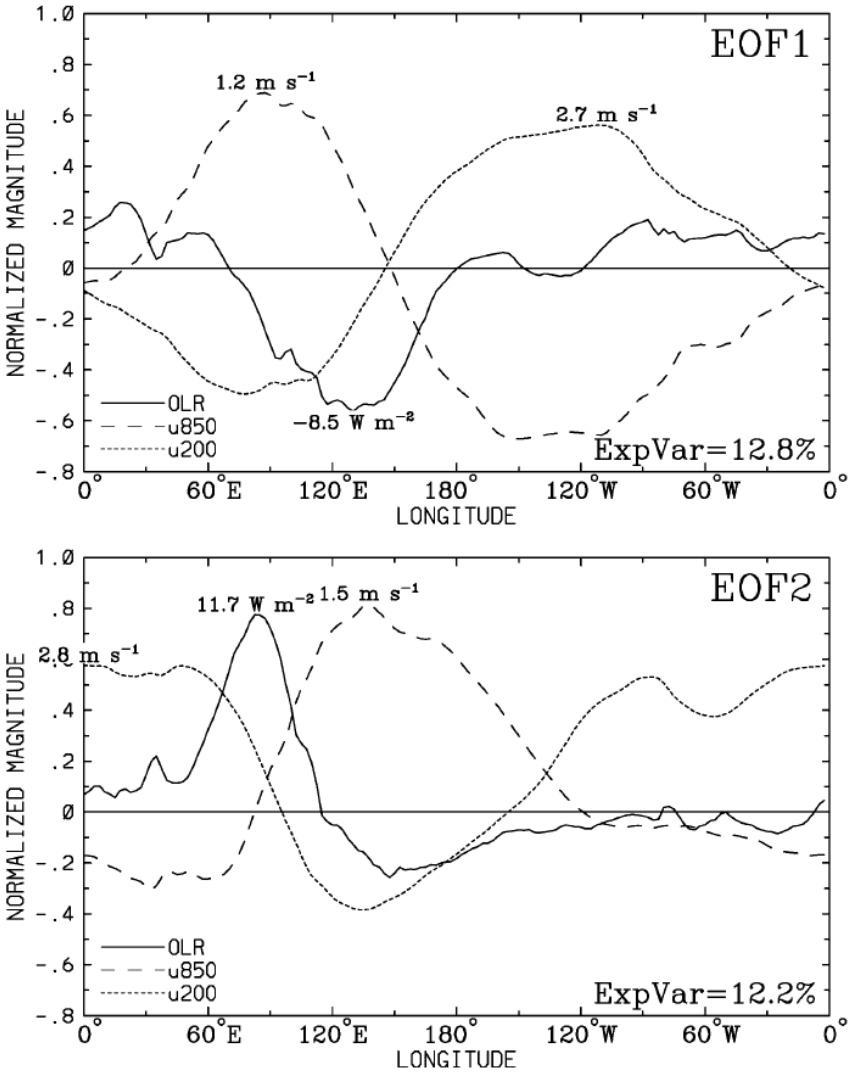

Figure 4.10: Spatial structures of EOFs 1 and 2 of the combined analysis of OLR and the zonal wind at geopotential heights 800 and 200 hectaPascals. A key for the field described by each curve is given. As each field is normalized by its global (all longitudes) variance before the EOF analysis, their magnitude may be plotted on the same relative axis. Multiplying each normalized magnitude by its global variance gives the field anomaly that occurs for a 1 std dev perturbation of the PC, as given for the absolute maxima of each field. The variance explained by the respective EOFs is 12.8% and 12.2%. From Wheeler and Hendon [164]

suggested that by using equatorially averaged OLR as input, the RMM indices may have limited capability of detecting the MJO when its convective signal shifts into one hemisphere [237]. Convective variations associated with the MJO are often asymmetric about the equator during the solistical seasons.

The RMM indices locate the region of enhanced convection associated with the MJO, and, plotted in RMM phase space, the propagation of individual events. Figure 4.12 shows the evolution of the April 2009 MJO event. The amplitude increases with distance from the origin; eastward propagation is represented as anti-clockwise rotation around the diagram. Weak MJO activity is defined as when the RMM amplitude is inside the unit circle.

Let us now use calculate RMM index for the neutral years and compare its dynamics with the results from the previous session. We picked one year as an example: 2005 looks particularly interesting, because the correlation matrices and clustering distribution over different inter-monsoon seasons are rather dissimilar.

From the data provided by NOAA, we first calculate RMM1 and RMM2 (Fig. 4.13) and then calculate RMM index [164] (Fig. 4.14). Finally, we plot the RMM phase portrait (figs. 4.15 and 4.16).

Let us first inspect the figure that shows the dynamics of the RMM's absolute value, that is resembling strength of the MJO. It is clear from the graph that during the April-May season MJO was active and much stronger than over the October-November season of the same year. To inspect the connection between the patterns on the correlation matrices and the cluster distribution we decided to plot the dynamics of the RMM index next to the graph that shows the number of the elements in the biggest cluster. For the April-May (Fig. 4.17) we see strong correlation between two graphs, while for the October-November (Fig. 4.18) no such connection exists. Reference line at 1 on figures 4.17 and 4.18 shows meaningful level of signal for sum of RMMs. Reference line at 60 was chosen by us as a threshold that is used to define main clusters.

To finish our introduction of the RMM index let us inspect the propagation of the MJO by plotting phase diagrams. Active MJO event during April-May 2005 could be clearly seen on figures 4.15 (for April) and 4.16 (for May). RMM signal over October-November 2005 is, in turn, saying that the MJO was weak during this period. It is shown of figures 4.19 (for October) and 4.20 (for November).

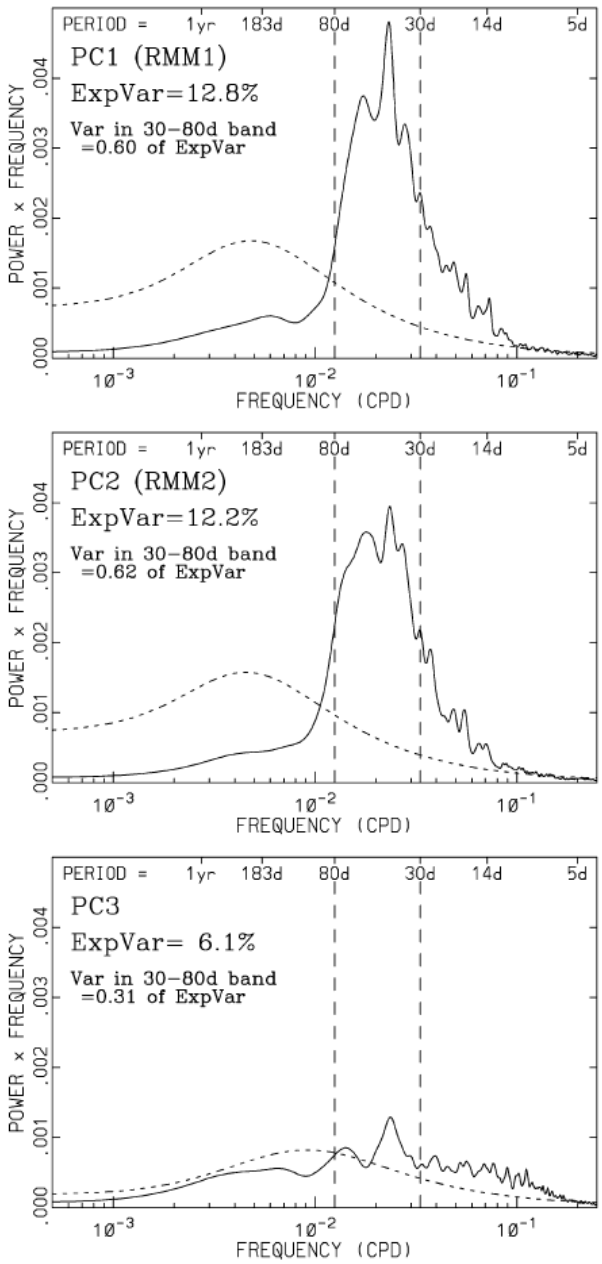

Figure 4.11: Power spectra of the PCs of the leading three EOFs of the combined analysis of Fig. 1, as calculated using the whole time series. The plotting format forces the area under the power curve in any frequency band to be equal to variance. The total area under each curve is scaled to equal the explained variance (Exp Var) by that EOF. The fraction of ExpVar in the 30- to 80-day band for each PC is given. The dashed curve is the red-noise spectrum computed from the lag 1 auto-correlation. Multiple passes of a 1–2–1 filter are applied to all spectra resulting in an effective bandwidth of 3.0×10^{-3} cpd (cycles per day). From Wheeler and Hendon [164]

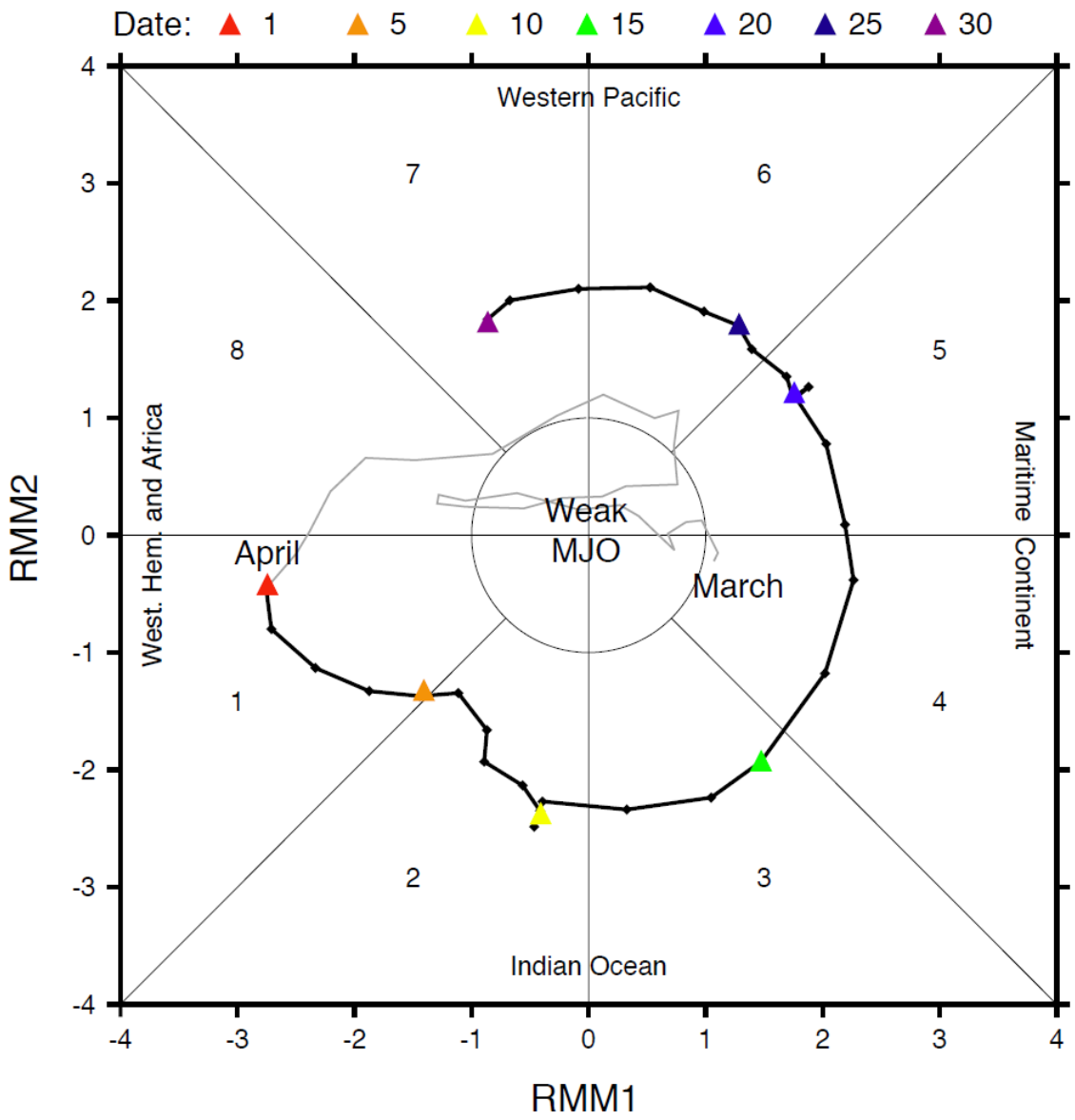

Figure 4.12: RMM1 and RMM2 of satellite-derived OLR and NCEP reanalysis zonal winds at 850 hPa and 200 hPa for a strong MJO event in April 2009, taken from Wheeler [69]. The black (grey) line indicates the evolution of MJO activity during April (March). Coloured triangles represent the date in April. Text labels indicate the approximate location of the enhanced convective signal of the MJO.

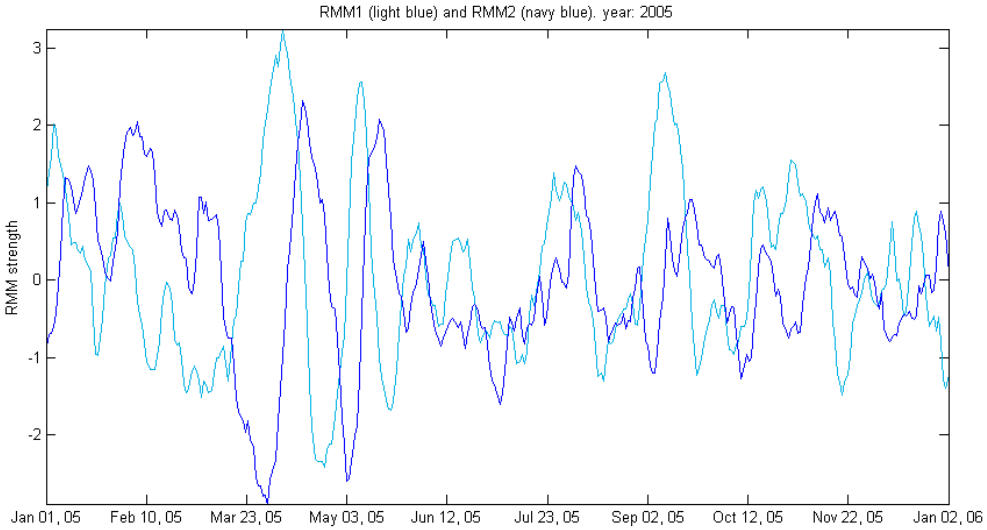

Figure 4.13: RMM1 and RMM2 signals for year 2005

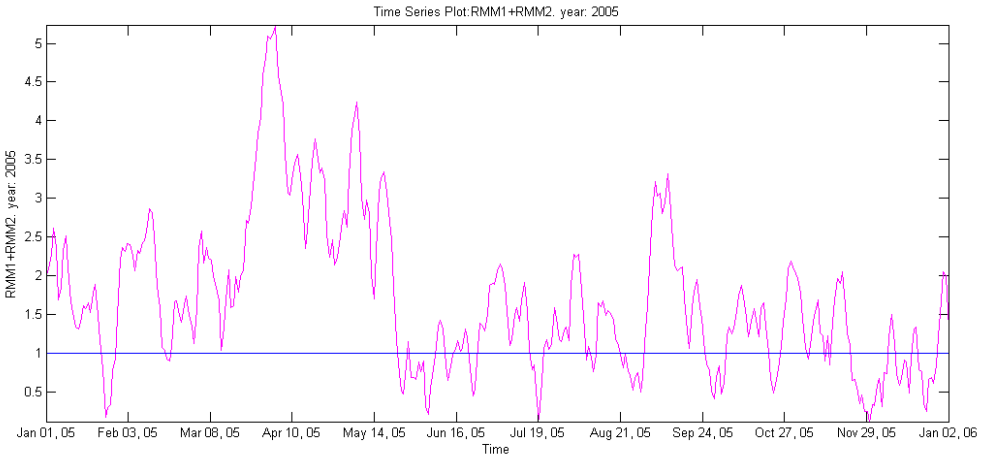

Figure 4.14: Sum of RMM1 and RMM2 signals for year 2005

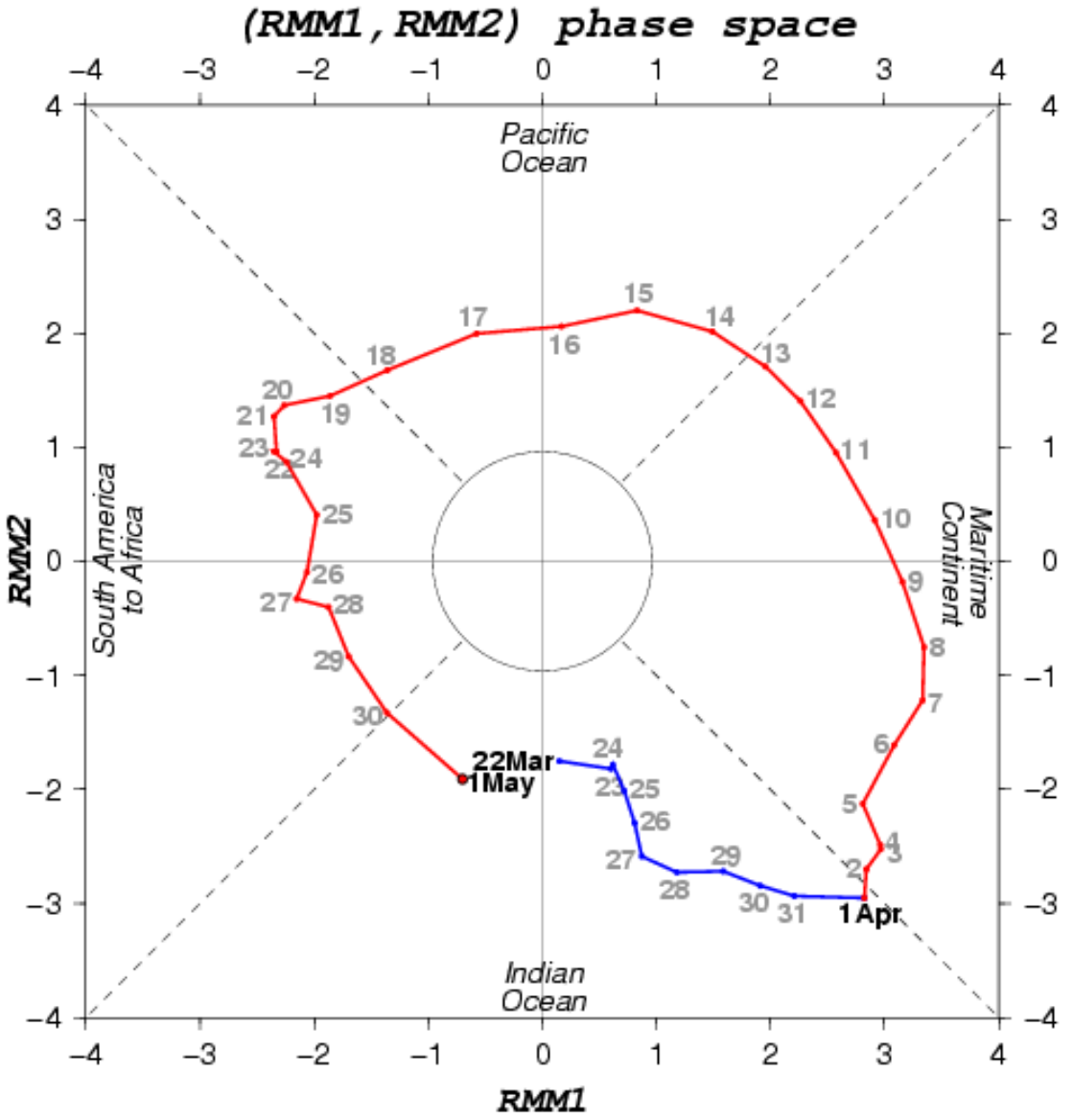

Figure 4.15: MJO phase space, based on multivariate EOF analysis, for 22.03.2005 to 01.05.2005.

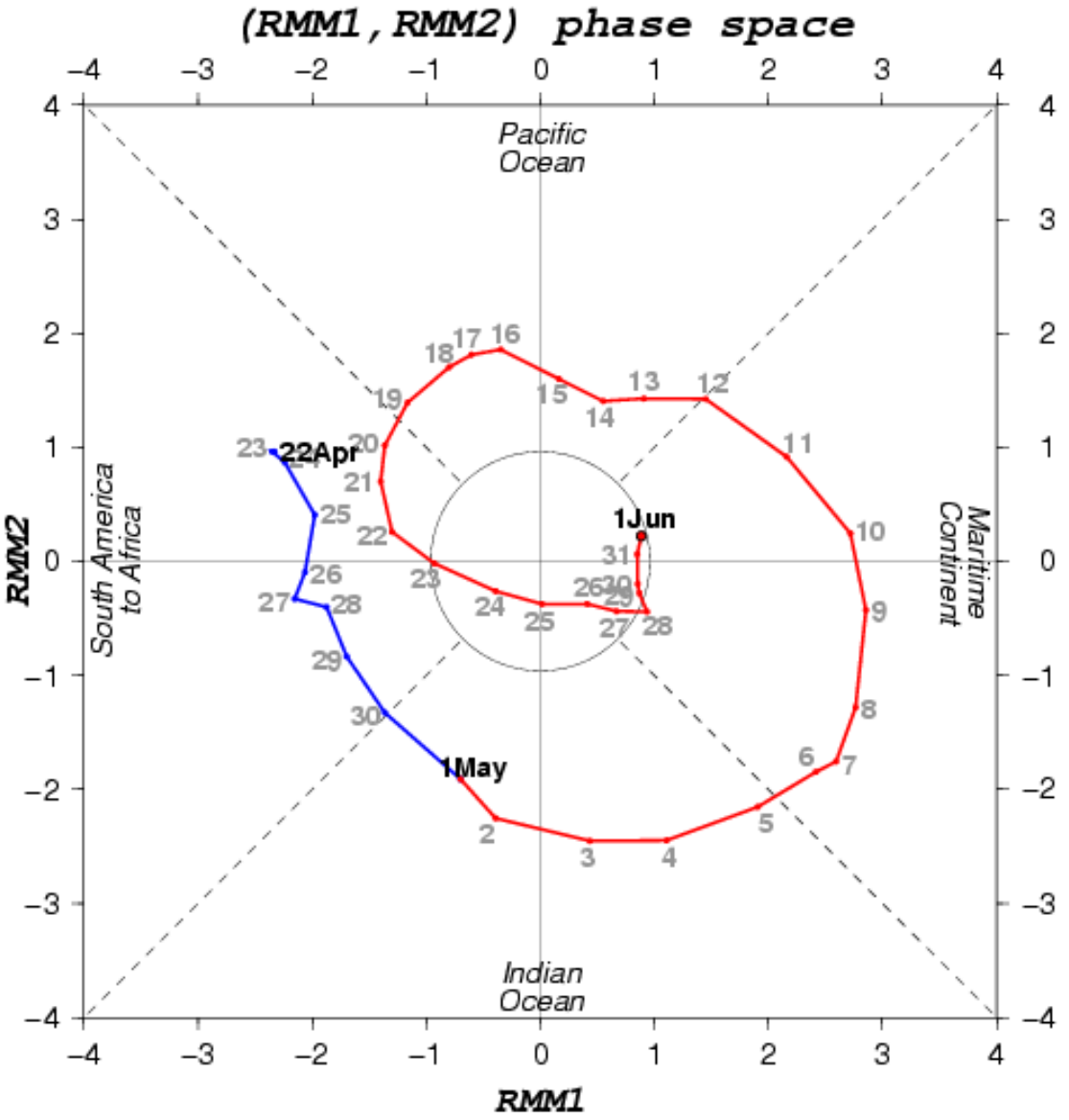

Figure 4.16: MJO phase space, based on multivariate EOF analysis, for 22.04.2005 to 01.06.2005.

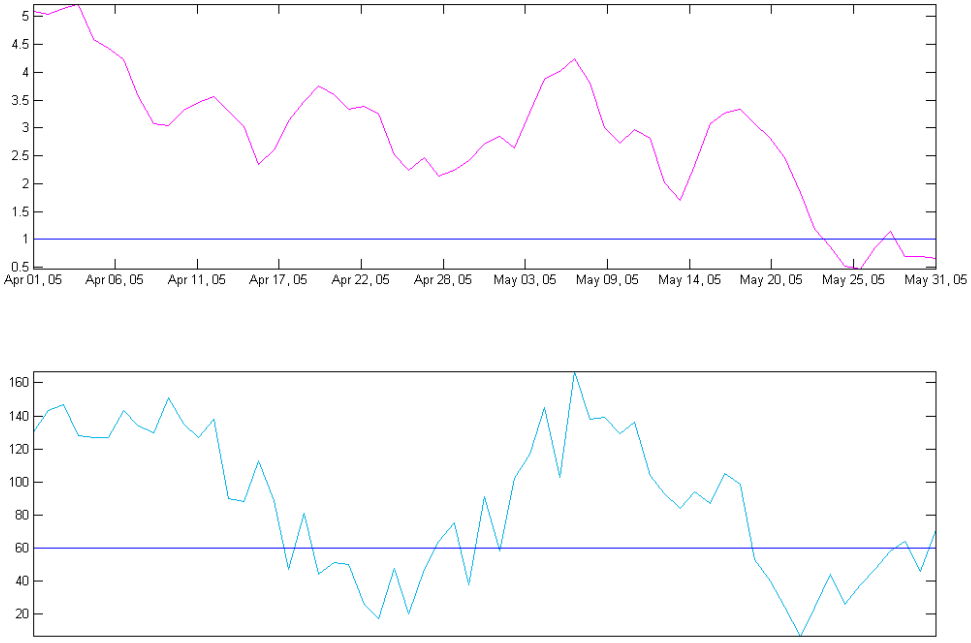

Figure 4.17: Sum of RMM1 and RMM2 signals (top) and number of elements in main cluster (bottom) for inter-monsoon season April-May 2005

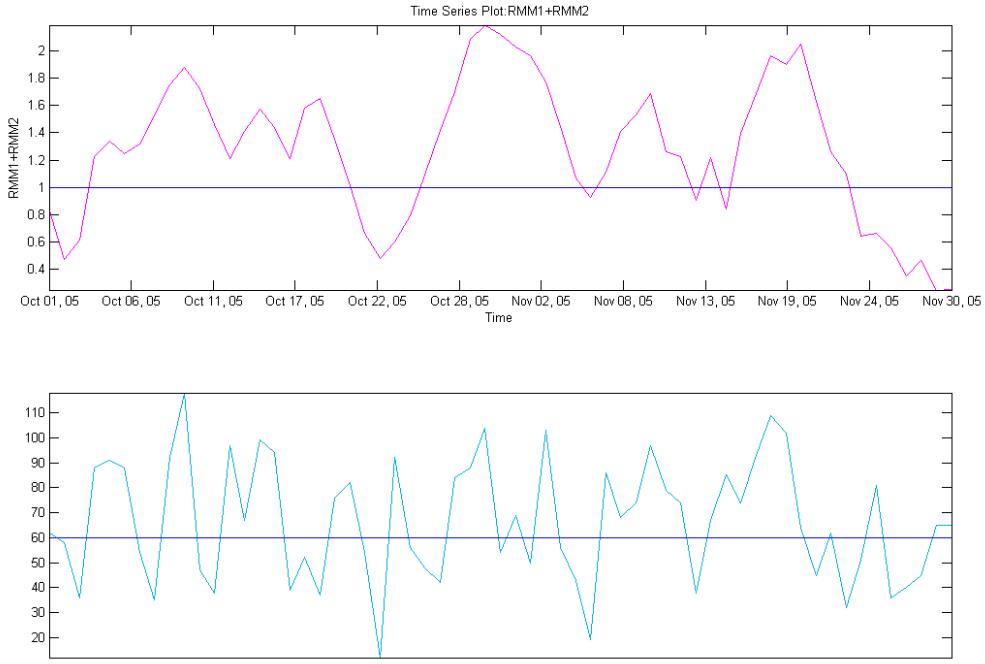

Figure 4.18: Sum of RMM1 and RMM2 signals (top) and number of elements in main cluster (bottom) for inter-monsoon season October-November 2005

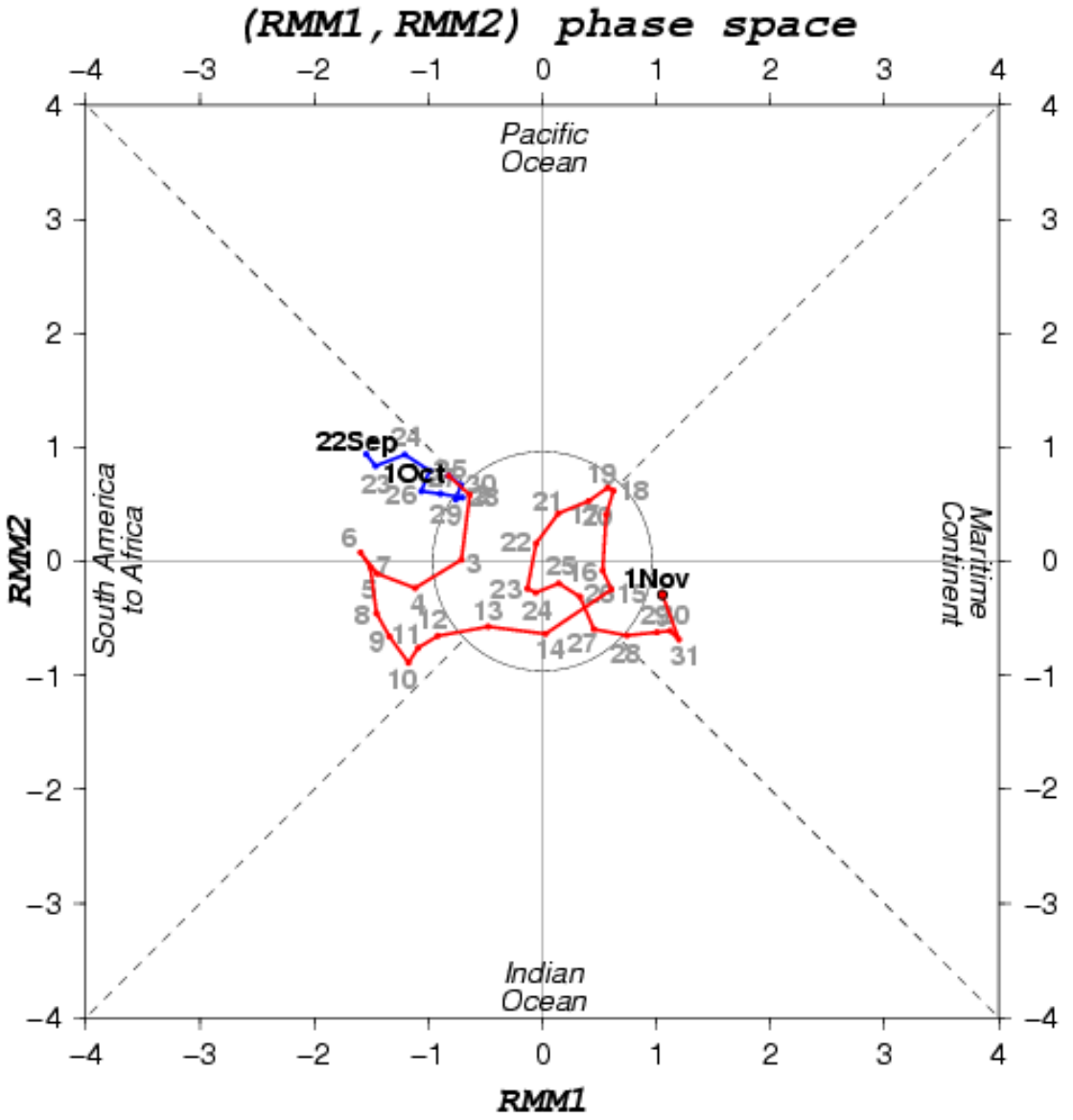

Figure 4.19: MJO phase space, based on multivariate EOF analysis, for 22.09.2005 to 01.11.2005.

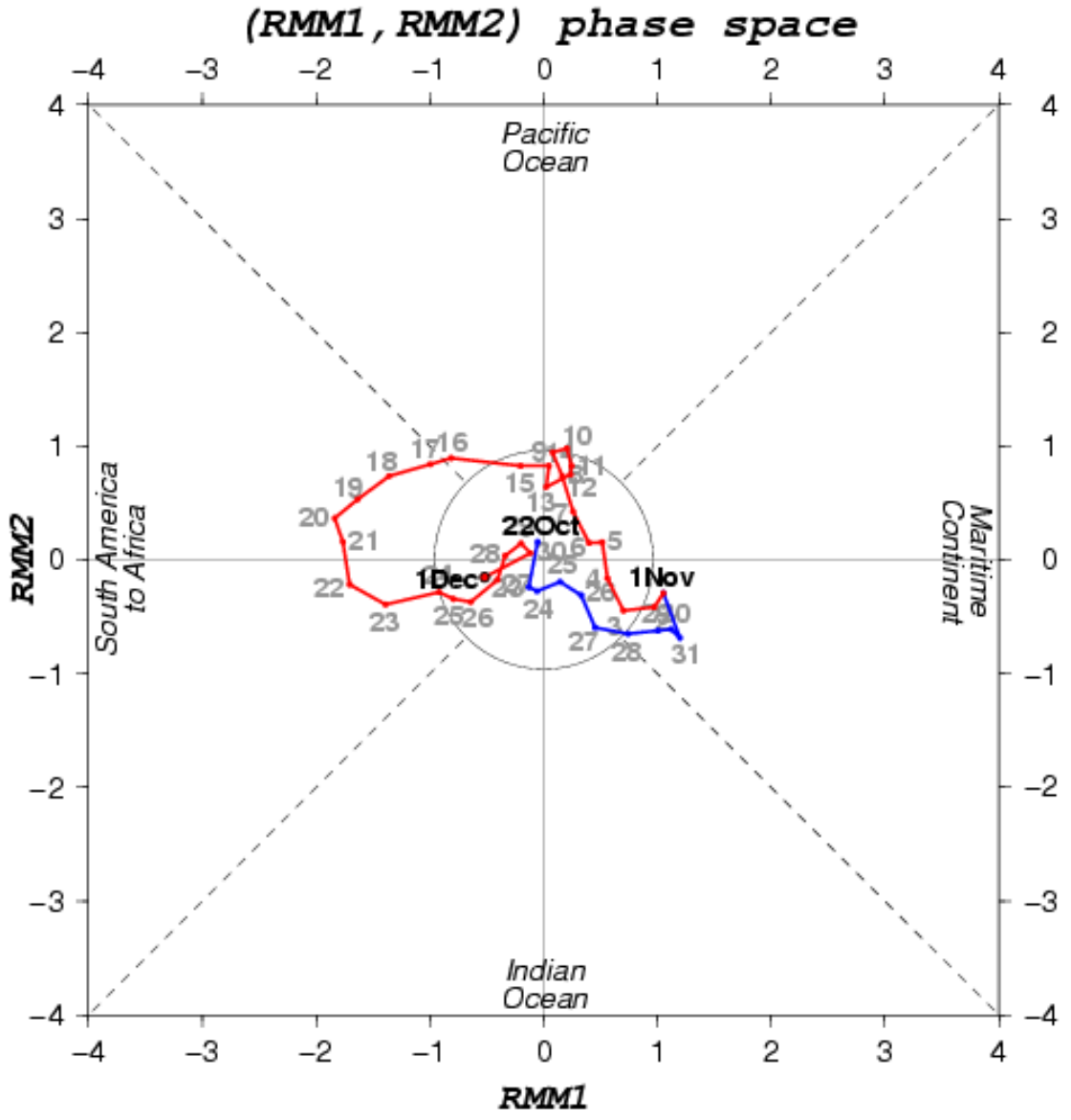

Figure 4.20: MJO phase space, based on multivariate EOF analysis, for 22.10.2005 to 01.12.2005.

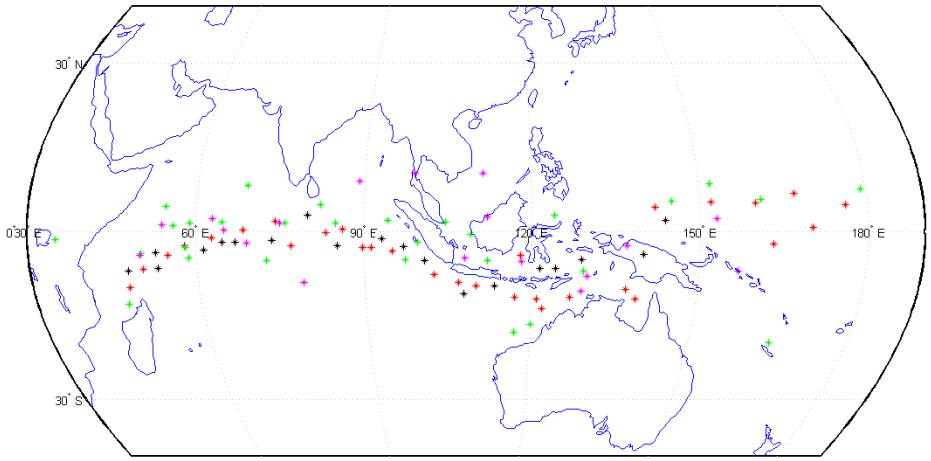

Figure 4.21: Centers of mass for the main clusters (red for first biggest, green for second biggest and magenta for third biggest, black for center of mass of raw rainfall data) for the inter-monsoon season April-May 2005.

4.2.3 Trajectories of the main clusters

In previous sections our main goal was to identify the MJO event. Now we will study its geographical evolution. To investigate the trajectories of the clusters we compute the centers of mass of the three biggest clusters for each time window. On the figure 4.21 trajectories of three biggest clusters for the April-May 2005 are present. It should be noticed that the trajectory of the biggest cluster, represented as red star for each time window, is in good agreement with observations and RMM phase diagram. In the figure 4.22 we can see centers of mass of the biggest cluster during October-November 2005 (for second and third biggest cluster data is not plotted for clarity). The plotted points do not resemble the trajectory of the MJO in this case, but we know already from the previous results that during this period no active events were present. Hence we can assume that the dots represent random rain pattern over this period of time.

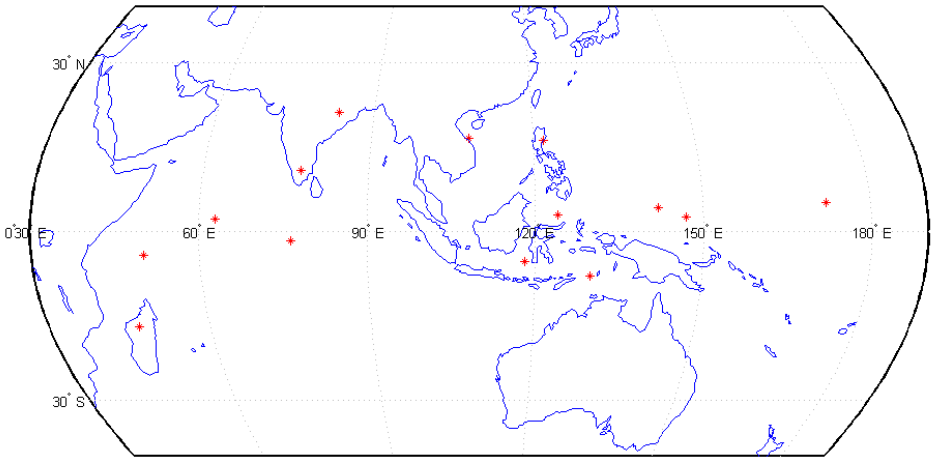

Figure 4.22: Centers of mass for the main cluster for the inter-monsoon season October-November 2005.

4.3 Monsoons cluster structure

Before proceeding to the study on the MJO-monsoon interaction it would be useful to understand how the clustered rainfall pattern of monsoon would look like. To answer this question, we analyzed one summer monsoon period (2008), when MJO signal is weak almost all over the summer. By performing the same analysis as we did to capture MJO we expect to find out the properties of the monsoon. On the figure 4.23 we can see both the correlation matrix and cluster distribution for the active monsoon season of June-July 2008.

From these plots we can see, that amount of the elements in biggest monsoon clusters and their general distribution is comparable to those of MJO.

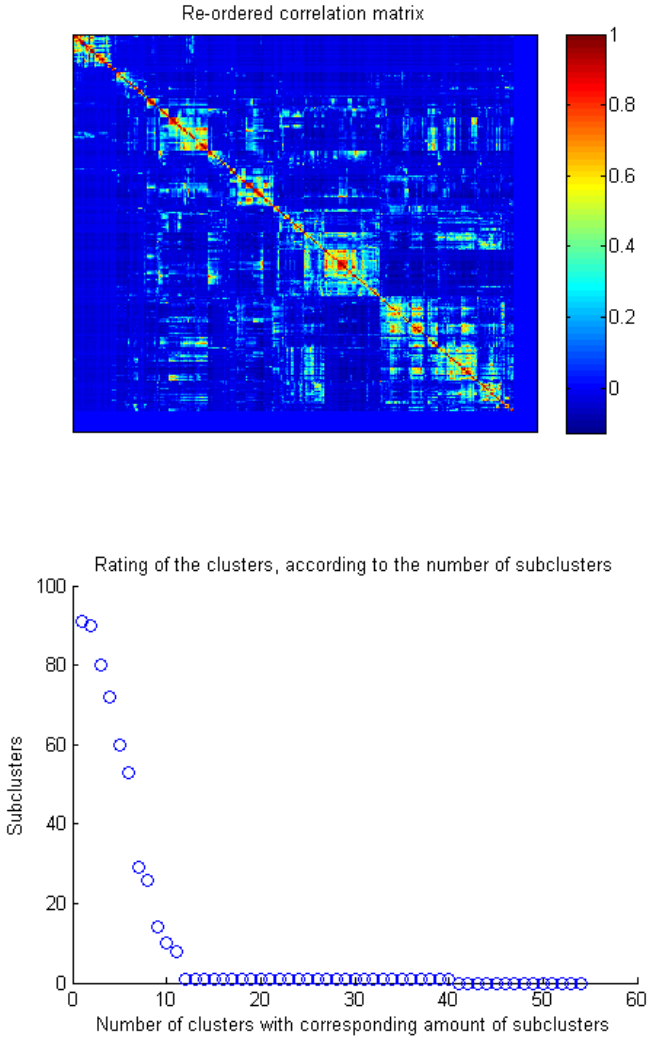

Figure 4.23: Re-ordered matrix and rating of clusters, according to number of sub-clusters for June-July 2005

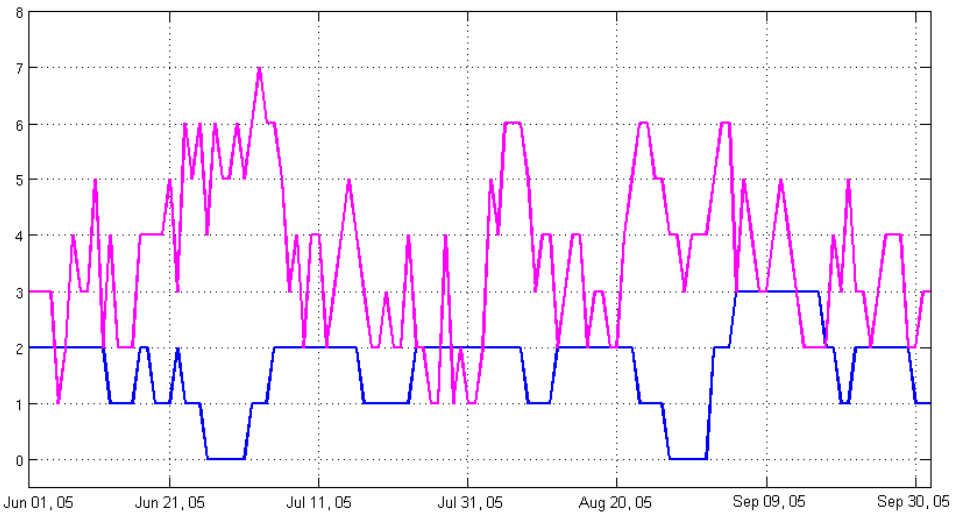

Figure 4.24: Dynamics of the number of elements in the first biggest and second biggest clusters for the summer monsoon 2005

4.4 MJO-Monsoon interaction

Let us finally investigate MJO-monsoon interaction. Knowing that MJO and monsoon have the same magnitude of the rainfall signal, we decided to look at dynamics of the clusters of different size. Three groups were created: clusters with 100 and more elements were considered as big clusters and formed first group, clusters with 80 to 100 elements were considered to be medium and formed second group, finally, clusters with less than 80 elements were considered small and formed third group. Started the study with three groups, we later understood that only first two groups provide robust signal and can serve as a proxy. The evolution of these groups is represented on the figure 4.24.

It is interesting to notice that the behavior of two plots is very different - while medium sized group is losing and gaining members all the time, large sized group shows rather stable dynamics with several peaks (we will examine their nature later). This is going in agreement with the trajectories of the centers of mass: while second and third biggest clusters paved winding paths, biggest trajectory of the biggest cluster followed the trajectory of the MJO.

There are two moments on the graph which did draw our attention: around 30 June and 9 September. Amount of clusters with more than 100 elements is suddenly dropping and naturally increasing the number of smaller clusters. One more event, when amount of smaller clusters is raising occur around 5 August, however during this time at least one big cluster is present. To explain this we will compare it with the RMM signal.

On the figure 4.25 sum of RMM1 and RMM2 is plotted. According to the graph, during

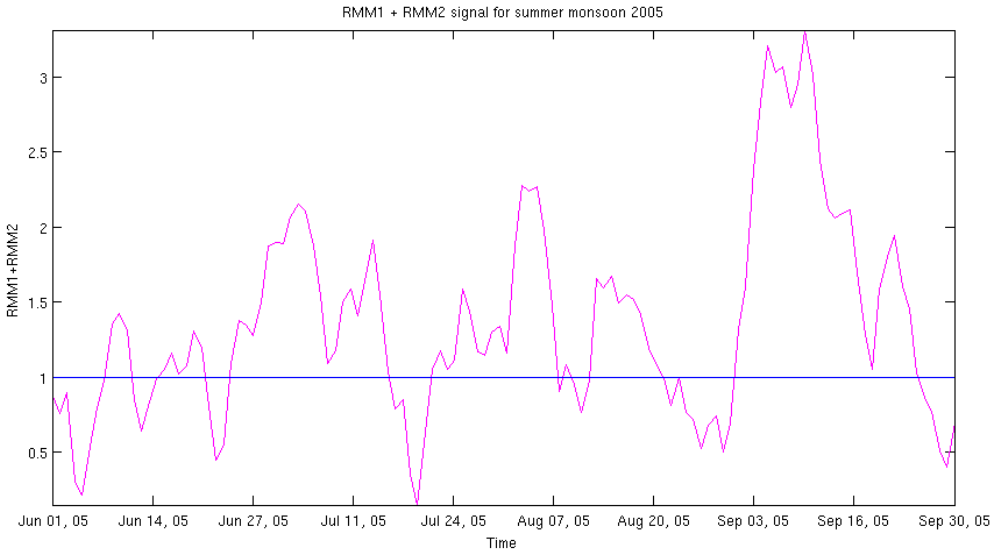

Figure 4.25: Sum of two RMMs for summer monsoon 2005

the summer monsoon season of the 2005 there were three moments when MJO was especially strong: around 30 June, 5 August and 9 September. Peaks span over several days, however there is a strong similarity between RMM signal and rainfall clusters dynamics described right before. For more detailed picture, we plotted phase diagram of the RMM signal (Fig. 4.26). Now we can also see not only when, but also where MJO was particularly active: eastern coast of Africa during the end of June, Western Pacific during August and over Maritime continent during September.

After comparison with the RMM signals we can say that we captured several features of the MJO dynamics during this monsoon season. Presence of the big rainfall patterns is disturbed during active phases of the MJO, when these big clusters disintegrate into several smaller ones and assemble back once MJO is gone. So far, we discussed the impact of MJO to the results on the figure 4.24 by comparing them with the RMM signal, however as we found before clustering pattern of the monsoon is very similar and to study its contribution we will take a look at the monsoon indices.

4.4.1 Comparison with monsoon indices

As a benchmark for monsoon signal we have chosen Asian monsoon monitoring indices based on area averaged vertical zonal wind shear (Fig. 4.27) [181] and OLR (Fig. 4.28) [165]. These products are useful in monitoring the strength and expansion of the summer monsoon. Equatorial zonal wind index is defined as area-averaged zonal wind anomalies

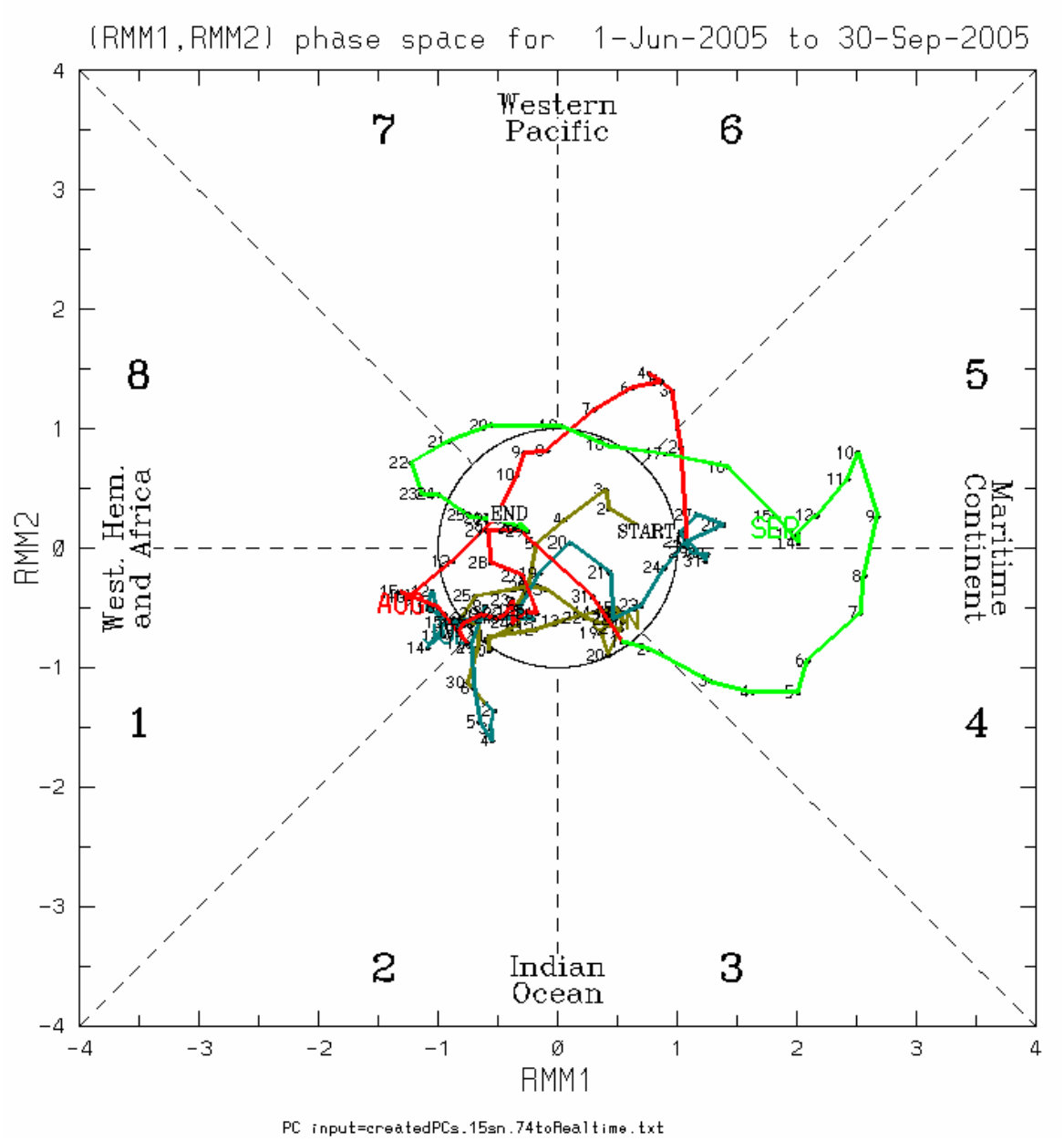

Figure 4.26: Phase portrait for RMM signal for summer monsoon 2005. Trajectory of the MJO during June is coloured dark green, during July - navy blue, during August - red and during September - light green.

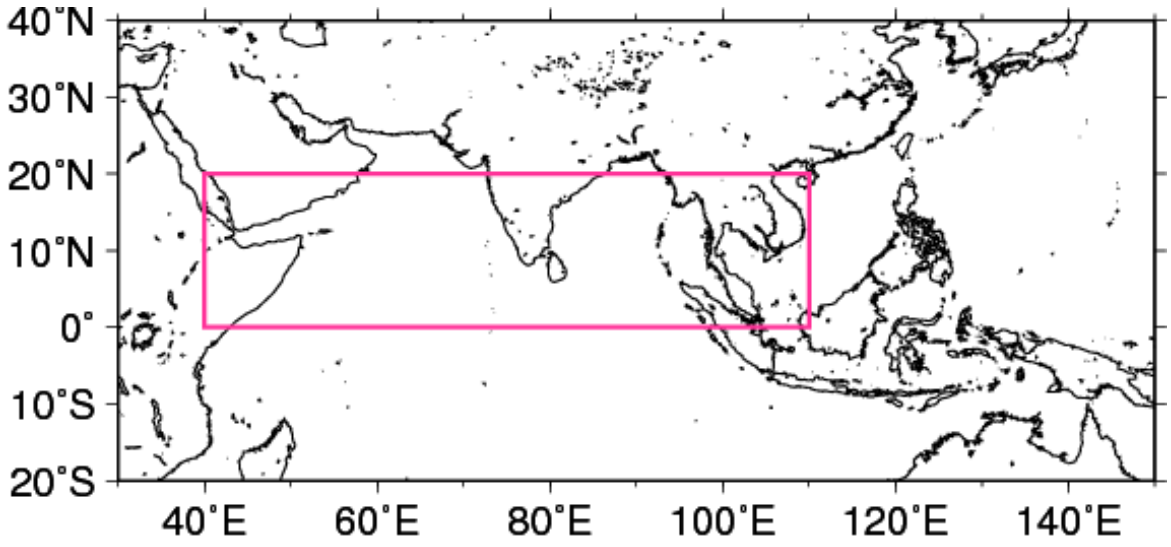

Figure 4.27: The area enclosed by the pink line indicates the area for the wind-based index

normalized by its standard deviations. OLR index is defined as area-averaged OLR anomaly with reverse sign, normalized by its standard deviations. Positive (negative) values of the OLR index indicate enhanced (suppressed) convective activity. The same initial data-sets for the OLR and zonal wind are used as previously for the RMM calculation.

On the figures 4.29 and 4.30 the signals of the wind-based and OLR-based monsoon indices are plotted for the summer monsoon and two inter-monsoon seasons of 2005. It is interesting to notice that both monsoon indices are showing very similar variations with the RMM. Peaks in the end of June, beginning of August and beginning of September are present. Such a good agreement could be explained through the high similarity between the rainfall patterns and their mutual influence. Considering all the above we believe that the results obtained from the clustering technique provide considerable insight into the dynamics of the MJO and its interaction with the summer monsoon.

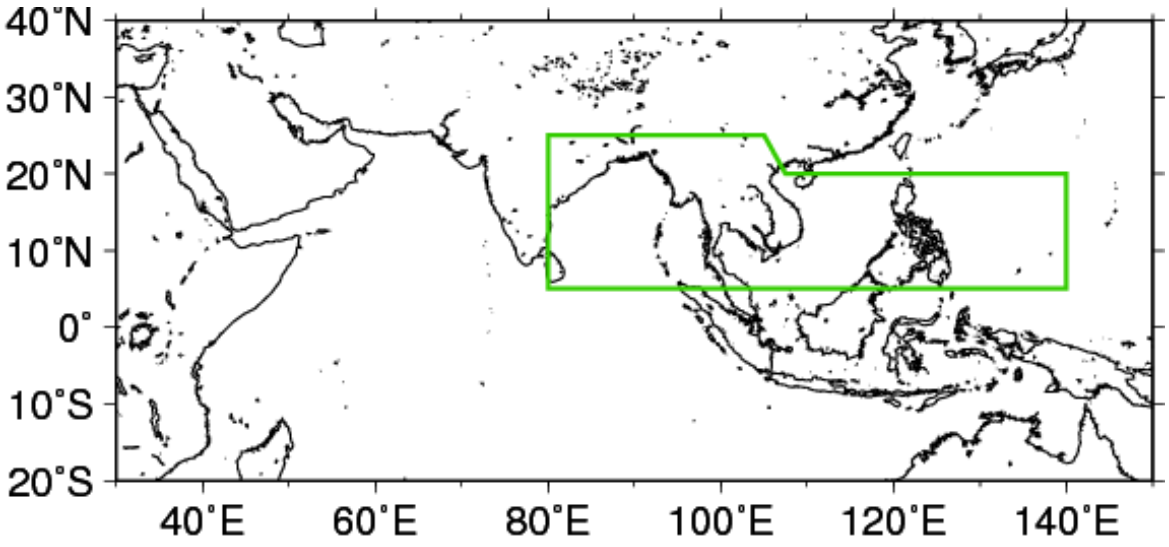

Figure 4.28: The area enclosed by the green line indicates the area for the OLR-based index

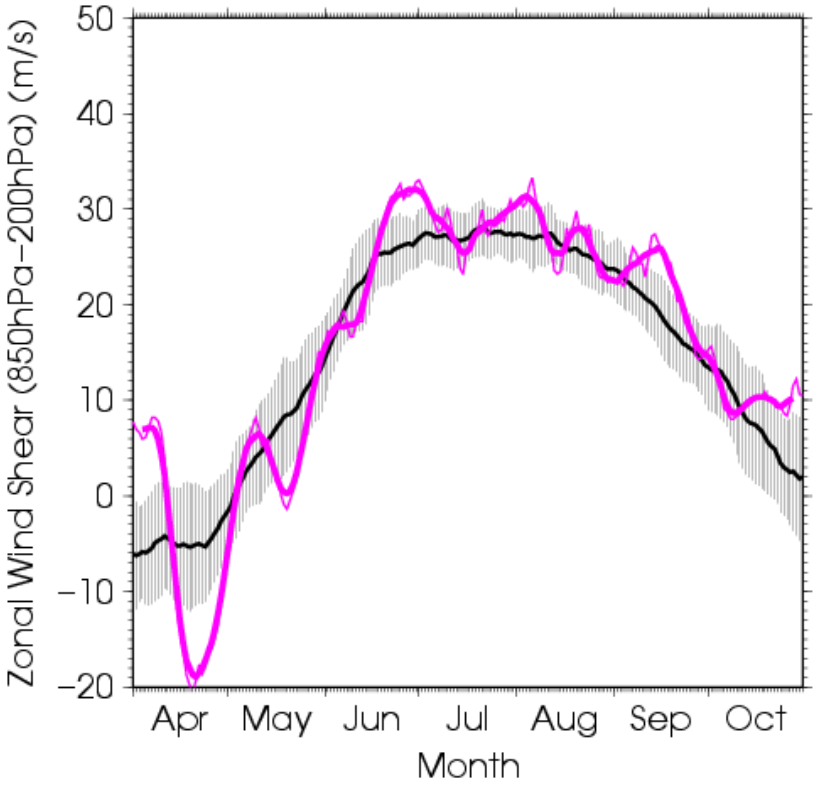

Figure 4.29: Asian monsoon index (wind-based) for April-October 2005. The thick and thin pink lines indicate seven-day running mean and daily mean values, respectively. The black line denotes the normal (the 1981 - 2010 average), and the gray shading shows the range of the standard deviation calculated for the time period of the normal.

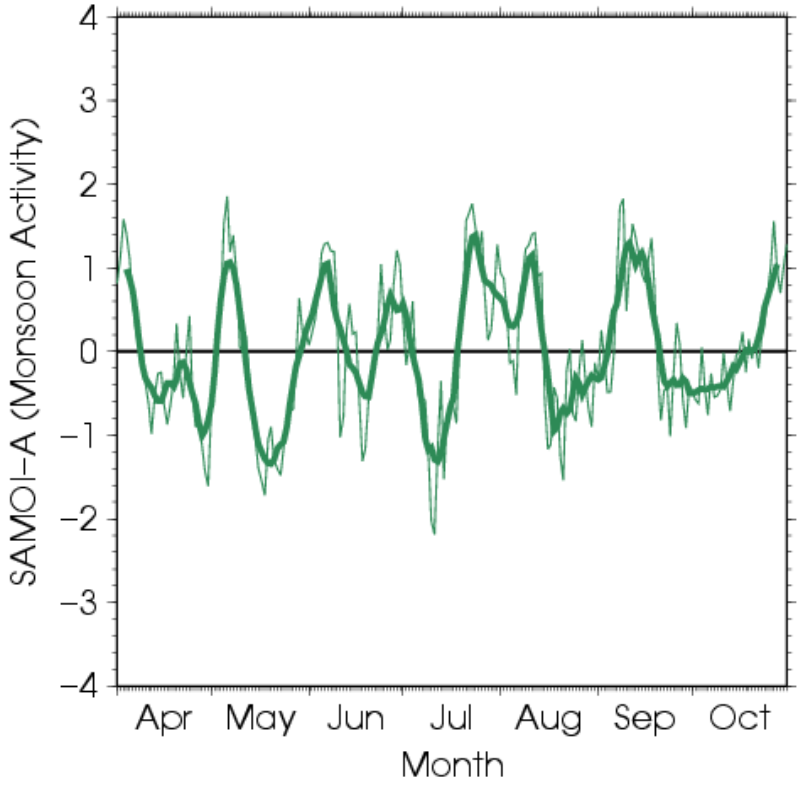

Figure 4.30: Asian monsoon index (OLR-based) for April-October 2005. The thick and thin green lines indicate seven-day running mean and daily mean values, respectively.

Chapter 5

Conclusion & Discussion

Learning is one of the most wonderful abilities, manifestation of intelligence. We, humans, not only have capability to learn more than other animals but also to create algorithms that will help us with wide range of problems. Machine learning techniques are aimed not only mimic human-like behavior but also to extend our learning abilities using computers. In this work we used multivariate statistical methods and machine learning techniques to reveal hidden patterns and connection within different data-set. Although all these methods have certain limitations, such as certain dependence on linkage criteria or problem with hierarchical structure: nature of the analysis means that early 'bad judgements' cannot be rectified, we combined and improved them in a way that allowed us to obtain valuable results.

Protein folding problem

We started with the protein folding problem. Our goal was to understand protein folding mechanism from the statistical point of view. We worked with two simulation data-sets: penta-alanine protein in water with low and high resolution (produced by our collaborator Assistant Professor Mu Yaguang) and three polyalanine peptides, Q, K and D, that eventually folded into α-helices (given by Assistant. Professor Zhang Dawei).

In the beginning we started to work with the dendrograms that we produced based on a low resolution data. We classified the complete-linkage dendrograms constructed from the correlation matrices visually and found from that the dendrograms they into three categories: (a) only two major clusters, (b) between three to four clusters, and (c) more than four clusters. While this could be done when the number of dendrograms to be considered is small, a more automatic procedure was desirable for a more systematic study of the molecule folding mechanism. Since low resolution data could not provide enough data-

points for detailed study, we then proceed to the high-resolution study of the same molecule. There we discovered that some interesting insights could be obtained by studying From the time evolution plot of the thresholds identified critical events at 1,250, 2,500, and 4,750 ns., indicating the formations of robust, strongly-correlated clusters. This might correspond to the folding of the protein. Later we supported this finding by clustering dendrograms obtained from the molecule's analysis. Indeed, three time the structure of the molecule was cardinally changed. To identify precursors of the possible folding events we studied the evolution of the molecule. It was then found out that the precursor atoms (β-C and O atoms) are positioned in the middle of the molecule which supported our findings again. Finally, we build a minimal spanning tree graphs only to see that the MST topology is identical to the structure of the protein, indicating that the strongest-correlating pairs are simply the pairs of atoms with structural bonds.

The study in the second part of the chapter we again started with the clustering analysis. Automatical analysis of the dendrograms helped us identify several moments with sharp transition in structure of the molecules Q and K. Since we did not reveal enough details about the nature of the processes in that molecule we decided to work with another attribute: dihedral angles. From the pairwise correlation of them we understood that the transition in structure indeed takes place after approximately 1 ns for the Q molecule and after approximately 6 ns for the K molecule. After that we used the minimal spanning tree graphs again to learn more about connections between the pairs of the correlation time series. From the tree structure it could be seen that the pairs that are both strong positively and negatively correlated are tend to be linked together and for proteins Q and K positively correlated pairs for the majority. In contrast, in D molecule negatively and positively correlated pairs are equally distribute.

In this part we successfully identified folding events and the precursors by using different techniques. The findings are in agreement with the results of Dawei [147], who was also closely exploring this date-set. Additionally, I would like to express my gratitude to Jeremy Hadidjojo and Chua Khi Pin - our discussions helped this study.

Dynamics of the tropical atmosphere

In this part we were exploring the dynamics of the tropical atmosphere through the prism of the one of the most interesting intraseasonal phenomena: Madden Julian Oscillation. It has coupled features of tropical deep convection and atmospheric circulation and is usually tracked by several parameters: near-equatorially averaged 850 hPa zonal wind, 200 hPa zonal wind, outgoing long-wave radiation (OLR) etc. However, in our study of MJO we used rainfall as a proxy for two reasons: enhanced or suppressed tropical rainfall is one of

the main characteristics of MJO and very consistent and reliable satellite-observed rainfall database TRMM.

We started with the analysis of the data-sets that are spanning the time when other atmospheric phenomena are neutral. By using correlation and clustering analysis we were able to identify MJO signal. We also noticed interesting distribution in rainfall cluster sizes with and without MJO. To explore this regularity we calculated and plotted the RMM index for the MJO, which is based on wand and OLR data. This gave us the understanding that changes in cluster distribution at least during the neutral seasons are indeed resembling evolution of the MJO and are in a good agreement with the classic RMM index.

Our next interest was lying in interaction of Madden Julian Oscillation and Monsoons. From cluster analysis we saw that the magnitude of the monsoon rainfall clusters is comparable to those of MJO. Our next step was to explore the behavior of clusters during the monsoon seasons and by this explore the interaction of two atmospheric phenomena. We showed that cluster structure represents MJO and monsoon indices, even though they are based on different variables. Finally, from the clustering analysis of the cluster matrices we obtain a picture that shows us that there are two regimes: time periods without strong MJO when monsoon is smooth and active MJO periods when level of dissimilarity between two circulations. These results are in agreement with other works that are stating that strong MJO activity may influence “active” and “break” monsoon rainfall regimes [163, 178].

Many questions remain unanswered and one of the promising future direction in this study is detailed geographical distribution of the clusters. Among others, we plan to examine in the future interaction between clusters by tracking the list of the elements in a certain cluster, fluctuations of the center of mass, variations in the area, associated with one cluster - we believe that this approaches may help us to understand the dynamics of the tropical atmosphere better.

Contribution and publications

In this thesis we have shown the benefit of multivariate statistical methods and machine learning techniques for studies of the protein dynamics and atmospheric dynamics. We have also provided some insights in our case studies that might have immediate applications: we investigated several proteins and their folding precursors are now identified, which might be used in the protein data bank. However, to get more fundamental understanding of the folding process more molecules have to studied. Problem of the protein folding has not been solved yet theoretically, but probably numerical solutions will change the game. Combination of the proposed multivariate statistical methods with more advanced machine learning techniques as artificial neural network might give very good results in this direction. The

same is foreseen for the MJO studies. This relatively new phenomena is not yet described theoretically. Building such theory on the outcome of our empirical analyses a thorough theoretical investigation is desirable. In the mean time, proposed rainfall identification and tracking system may be used to construct better numerical model of the phenomenon.

Main papers produced during the course of this thesis are the following:

- **Mikhail FILIPPOV**. Protein folding and transformations: insights form clustering techniques. International Conference on Computational Science (2011). In this conference paper first results of the protein study were presented. Low and high resolution studies were discussed.

- **Mikhail FILIPPOV**. Understanding Protein Folding Mechanisms: Machine Learning Study. The Institute of Physics Singapore Meeting (2015). In this conference paper results on three proteins study were discussed. Identification of the folding events and the precursors was presented with evidence from clustering techniques, MST and segmentation.

- **Mikhail FILIPPOV, Khipin CHUA, Jeremy HADIDJOJO, Jiali SHAO, Chong Eu LEE, Yuguang MU, Dawei ZHANG, Lock Yue CHEW, San Keong LAI, and Siew Ann CHEONG**. Universal Correlational Fingerprints and Precursors for Protein Folding Into Alpha-Helices (in preparation).

- **Mikhail FILIPPOV, Siew Ann CHEONG, Tieh-Yong KOH**. Slow variables in tropical atmospheric dynamics. (in preparation).

Apart from the mentioned papers several presentations were made on the seminars in Nanyang Technological University and Ecole Polytechnique.

References

[1] Large-scale DNA sequencing. Tim Hunkapiller, Robert J. Kaiser, Ben F. Koop, Leroy Hood. Current Opinion in Biotechnology, Vol. 2, issue 1, February 1991, pp. 92–101

[2] The Complete Genome Sequence of Escherichia coli K-12. Frederick R. Blattner, Guy Plunkett III, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau, Ying Shao. Science, 5 September 1997: Vol. 277 no. 5331 pp. 1453-1462

[3] Next-generation DNA sequencing. Jay Shendure & Hanlee Ji. Nature Biotechnology 26, 1135 - 1145 (2008)

[4] A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. Paul Scheet, Matthew Stephens. The American Journal of Human Genetics, Volume 78, Issue 4, April 2006, Pages 629–644.

[5] Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. David G. Wang, Jian-Bing Fan, Chia-Jen Siao, Anthony Berno, Peter Young, Ron Sapolsky, Ghassan Ghandour, Nancy Perkins, Ellen Winchester, Jessica Spencer, Leonid Kruglyak, Lincoln Stein, Linda Hsie, Thodoros Topaloglou, Earl Hubbell, Elizabeth Robinson, Michael Mittmann, Macdonald S. Morris, Naiping Shen, Dan Kilburn, John Rioux, Chad Nusbaum, Steve Rozen, Thomas J. Hudson, Robert Lipshutz, Mark Chee, Eric S. Lander. Science 15 May 1998: Vol. 280 no. 5366 pp. 1077-1082

[6] Large-scale genotyping of complex DNA. Giulia C Kennedy, Hajime Matsuzaki, Shoulian Dong, Wei-min Liu, Jing Huang, Guoying Liu, Xing Su, Manqiu Cao, Wenwei Chen, Jane Zhang, Weiwei Liu, Geoffrey Yang, Xiaojun Di, Thomas Ryder, Zhijun He, Urvashi Surti, Michael S Phillips, Michael T Boyce-Jacino, Stephen PA Fodor & Keith W Jones. Nature Biotechnology , 2003, Vol. 21, pp. 1233 - 1237

[7] A Plan to Capture Human Diversity in 1000 Genomes. Jocelyn Kaiser. Science 25 January 2008: Vol. 319, p. 395.

[8] The perceptron: A probabilistic model for information storage and organization in the brain. Rosenblatt, F. Psychological Review, Vol. 65(6), Nov 1958, pp. 386-408

[9] Computation: finite and infinite machines. Marvin L. Minsky. Computation: finite and infinite machines Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1967

[10] Lower bounds for algebraic computation trees. Michael Ben-Or. Proceeding STOC '83 Proceedings of the fifteenth annual ACM symposium on Theory of computing Pages 80-86 ACM New York, NY, USA,1983

[11] Decision trees and random access machines. W Paul, J Simon, 1980

[12] Lower Bounds for Algebraic Decision Trees. J Michael Steele, Andrew C Yao. Journal of Algorithms, Vol .3, Issue 1, March 1982, pp. 1–8

[13] On the complexity of varying sets of primitives. D. Dobkin and R. J. Lipton. .I. Comput. System Sci., Vol. 18 (1980), pp. 86-91.

[14] Multilayer feedforward networks are universal approximators. Kurt Hornik, Maxwell Stinchcombe, Halbert White. Neural Networks Volume 2, Issue 5, 1989, pp. 359–366

[15] Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. Kurt Hornik, Maxwell Stinchcombe, Halbert White. Neural Networks Volume 3, Issue 5, 1990, pp. 551–560

[16] Optimal unsupervised learning in a single-layer linear feedforward neural network. Terence D. Sanger. Neural Networks Volume 2, Issue 6, 1989, Pages 459–473

[17] Boosted Decision Tree Analysis of Surface-enhanced Laser Desorption/Ionization Mass Spectral Serum Profiles Discriminates Prostate Cancer from Noncancer Patients. Yinsheng Qu, Bao-Ling Adam, Yutaka Yasui, Michael D. Ward, Lisa H. Cazares, Paul F. Schellhammer, Ziding Feng, O. John Semmes and George L. Wright Jr. Clinical Chemistry October 2002 vol. 48 no. 10 1835-1843

[18] Drug-related morbidity and mortality: updating the cost-of-illness model. Ernst FR, Grizzle AJ. Journal of the American Pharmaceutical Association, 2001, Vol. 41(2), pp.192-199

[19] Application of artificial neural networks to clinical medicine. W.G Baxt. The Lancet Volume 346, Issue 8983, 28 October 1995, Pages 1135–1138

[20] Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real World Performance. Robert R. Trippi, Efraim Turban. Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real World Performance McGraw-Hill, Inc. New York, NY, USA 1992

[21] Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). Edward I. Altman, Giancarlo Marco, Franco Varetto. Journal of Banking & Finance Volume 18, Issue 3, May 1994, Pages 505–529

[22] What is the right organization structure? Decision tree analysis provides the answer. Robert Duncan. Organizational Dynamics Volume 7, Issue 3, Winter 1979, Pages 59–80

[23] Machine learning in automated text categorization. Fabrizio Sebastiani. ACM Computing Surveys (CSUR), Volume 34, Issue 1, March 2002 Pages 1-47

[24] A review of supervised machine learning algorithms and their applications to ecological data. C. Criscia, B. Ghattasb, G. Perera. Ecological Modelling Volume 240, 10 August 2012, Pages 113–122

[25] Hidden Markov models for sequence analysis: extension and analysis of the basic method. Richard Hughey, Anders Krogh. Comput Appl Biosci, 1996 Vol.12 (2), pp. 95-107.

[26] A hierarchical unsupervised growing neural network for clustering gene expression patterns. Javier Herrero, Alfonso Valencia, Joaquín Dopazo. Bioinformatics, 2001 Vol. 17 (2), pp, 126-136.

[27] Current methods in medical image segmentation. Dzung L. Pham, Chenyang Xu, Jerry L. Prince. Annual Review of Biomedical Engineering, 2000, Vol. 2, pp. 315-337

[28] Object class recognition by unsupervised scale-invariant learning. Fergus, R., Perona, P., Zisserman, A. Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference, Volume 2, 18-20 June 2003, II-264 - II-271

[29] A History of Cluster Analysis Using the Classification Society's Bibliography Over Four Decades. Fionn Murtagh, Michael J. Kurtz. arXiv:1209.0125

[30] Hierarchical clustering schemes. Stephen C. Johnson. Psychometrika September 1967, Volume 32, Issue 3, pp 241-254

[31] A new approach to clustering. Enrique H. Ruspini. Information and Control Volume 15, Issue 1, July 1969, Pages 22–32

[32] A clustering technique for summarizing multivariate data. Geoffrey H. Ball, David J. Hall. Behavioral Science Volume 12, Issue 2, pages 153–155, March 1967

[33] Clustering Algorithms. Hartigan, J. John Wiley & Sons, New York, NY. 1975.

[34] Cluster Analysis Algorithms.Spath H. Ellis Horwood, Chichester, England. 1980.

[35] Algorithms for Clustering Data. Jain, A. and Dubes, R. 1988, Prentice-Hall, Englewood Cliffs, NJ.

[36] Finding Groups in Data: An Introduction to Cluster Analysis. Kaufman, L. and Rousseeuw, P. John Wiley and Sons, New York, NY. 1990.

[37] Cluster Analysis and Related Issues. Dubes, R.C. In Chen, C.H., Pau, L.F., and Wang, P.S. (Eds.) Handbook of Pattern Recognition and Computer Vision, 3-32, World Scientific Publishing Co., River Edge, NJ. 1993.

[38] Cluster Analysis (3 rd ed.). Everitt, B. Edward Arnold, London, UK. 1993.

[39] Mathematic Classification and Clustering. Mirkin, B. Kluwer Academic Publishers. 1996.

[40] Data clustering: a review. Jain, A.K, Murty, M.N., and Flynn P.J. ACM Computing Surveys, 31, 3, 264-323. 1999.

[41] An analysis of recent work on clustering algorithms. Fasulo, D. Technical Report UW-CSE01 -03-02, University of Washington. 1999.

[42] Clustering Algorithms for Spatial Databases: A Survey. Kolatch, E. 2001. *PDF is available on the Web.*

[43] Data Mining. Han, J. and Kamber, M. Morgan Kaufmann Publishers. 2001.

[44] Clustering of time series data—a survey. T. Warren Liao. Pattern Recognition Volume 38, Issue 11, November 2005, Pages 1857–1874

[45] Data clustering: algorithms and applications. Charu C. Aggarwal, Chandan K. Reddy. 2014 .Taylor & Francis Group

[46] Mining data streams: a review. Mohamed Medhat Gaber, Arkady Zaslavsky, Shonali Krishnaswamy. ACM SIGMOD Record, Volume 34, Issue 2, June 2005, Pages 18 - 26

[47] A review of robust clustering methods. Luis Angel García-Escudero, Alfonso Gordaliza, Carlos Matrán, Agustín Mayo-Iscar. Advances in Data Analysis and Classification September 2010, Volume 4, Issue 2-3, pp 89-109

[48] Subspace clustering for high dimensional data: a review. Lance Parsons, Ehtesham Haque, Huan Liu. ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets, Volume 6, Issue 1, June 2004, Pages 90-105

[49] Clustering Algorithms in Biomedical Research: A Review. Rui Xu; Wunsch, D.C. Biomedical Engineering, IEEE Reviews, Volume 3, Page(s): 120 - 154, 04 October 2010.

[50] The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis. MASSART, D., KAUFMAN, L. John Wiley & Sons, New York, NY. 1983.

[51] An overview of combinatorial data analysis, in: Arabie, P., Hubert, L.J., and Soete, G.D. (Eds.) Clustering and Classification, 5-63, World Scientific Publishing Co., NJ. 1996

[52] Longitudinally-invariant k-clustering algorithms for hadron-hadron collisions. S. Catani, Yu.L. Dokshitzera, M.H. Seymour B.R. Webber. Nuclear Physics B Volume 406, Issues 1–2, 27 September 1993, Pages 187–224

[53] Clustering gene expression patterns. Ben-Dor, A. and Yakhini, Z. In Proceedings of the 3 rd Annual International Conference on Computational Molecular Biology (RECOMB 1999), 11-14, Lyon, France.

[54] Pattern Classification and Scene Analysis. Duda, R. and Hart, P. 1973. John Wiley & Sons, New York, NY.

[55] Maximum likelihood from incomplete data via the EM algorithm. Dempster, A., Laird, N., and Rubin, D. 1977. Journal of the Royal Statistical Society, Series B, 39, 1, 1-38.

[56] Introduction to Statistical Pattern Recognition. Fukunaga, K. 1990. Academic Press, San Diego, CA.

[57] Minimal model for studying prion-like folding pathways J. Z. Y. Chen, A. S. Lemak, J. R. Lepock, and J. P. Kemp. Proteins: Structure, Function, and Bioinformatics Volume 51, Issue 2, pages 283–288, 1 May 2003.

[58] The structure of proteins: Two hydrogenbonded helical congurations of the polypeptide chain. L. Pauling, R.B. Corey, and H.R. Branson. Proc. Nat. Acad. Sci. USA, Vol. 37, pp. 205-211, 1951.

[59] The pleated sheet, a new layer conguration of the polypeptide chain. L. Pauling and R.B. Corey. Proc. Nat. Acad. Sci. USA, 37:251{256, 1951.

[60] Structure of myoglobin: a three-dimensional fourier synthesis at 2a resolution. J.C. Kendrew, R.E. Dickerson, B.E. Strandberg, R.J. Hart, D.R. Davies, D.C. Phillips, and V.C. Shore. Nature, Vol. 185 pp. 422-427, 1960.

[61] Dominant forces in protein folding. K.A. Dill. Biochemistry, Vol. 31, pp. 7134-7155, 1990.

[62] Structure of haemoglobin: a three-dimensional fourier synthesis at 5.5a resolution, obtained by x-ray analysis. M.F. Perutz, M.G. Rossmann, A.F. Cullis, G. Muirhead, G. Will, and A.T. North. Nature, Vol. 185 pp. 416-422, 1960.

[63] From protein structure to biochemical function? R.A. Laskowski, J.D. Watson, and J.M. Thornton. J. Struct. Funct. Genomics, Vol. 4 pp.167-177, 2003.

[64] Structure, function and diversity of class I major histocompatibility complex molecules. P.J. Bjorkman and P. Parham. Ann. Rev. Biochem, Vol. 59 pp.:253-288, 1990.

[65] DNA conformation and protein binding. A. Travers. Ann. Rev. Biochem, Vol. 58 pp.427- 452, 1989.

[66] Protein Crystallography. T.L. Blundell and L.H. Johnson. Academic Press, New York, 1976.

[67] The development of x-ray analysis. Sir Lawrence Bragg. G. Bell and Sons, London, 1975.

[68] Crystal structure of interleukin 8: Symbiosis of NMR and crystallography. E.N. Baldwin, I.T. Weber, R.S. Charles, J. Xuan, E. Appella, M. Yamada, K. Matsushima, B.F.P. Edwards, G.M. Clore, A.M. Gronenborn, and A. Wlodawar. Proc. Nat. Acad. Sci. USA, Vol. 88 pp. :502-506, 1991.

[69] NMR of Proteins and Nucleic Acids. K. Wuthrich. John Wiley & Sons, New York, 1986.

[70] Are there pathways for protein folding? C. Levinthal. J. Chem. Phys. 65, 44. 1968.

[71] Principles that goven the folding of protein chains. Anifsen CB. Science 1973; Vol 181, pp. 223-230

[72] The Protein Data Bank. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. Nucl. Acids Res., Vol. 28, pp. 235-242, 2000.

[73] The impact of structural genomics: expectations and outcomes. J.M. Chandonia and S.E. Brenner. Science, Vol. 311, pp.347-351, 2006.

[74] Mid-Latitude Atmospheric Dynamics: A First Course. J.E. Martin. John Wiley and Sons, 2006

[75] Mesoscale Meteorology in Midlatitudes. P. Markowski, Y. Richardson. John Wiley and Sons, 2010

[76] Mid-latitude weather systems. T. Carlson. Routledge, 1992

[77] Implementation of Data Mining Techniques for meteorological applications. A. Cofino, J. Gutierrez, B. Jakubiak, M. Melonek. World Scientific: pp. 215–240., 2003

[78] Precipitation Estimation from Remotely Sensed Imagery Using an Artificial Neural. Y. Hong , K. Hsu, S. Sorooshian, X. Gao. Journal of Applied Meteorology Vol 43: pp. 1834–1852., 2004

[79] Techniques and Experience in Mining Remotely Sensed Satellite Data. Hinke, T. H., J. Rushing, H. Ranganath and S. J. Graves. Artificial Intelligence Review (AIRE, S4): Issues on the Application of Data Mining, pp 503-531, 2001.

[80] The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. Xue, M., D.-H. Wang, J.-D. Gao, K. Brewster, and K. K. Droegemeier. Meteor. and Atmos. Physics, 82, 139-170. 2003

[81] Trends in extreme daily rainfall and temperature in Southeast Asia and the South Pacific: 1961–1998. Manton, M.J., Della-Marta, P.M., Haylock, M.R., Hennessy, K.J., Nicholls, N., Chambers, L.E., Collins, D.A., Daw, G., Finet, A., Gunawan, D., Inape, K., Isobe, H., Kestin, T.S., Lefale, P., Leyu, C.H., Lwin, T., Maitrepierre, L., Ouprasitwong, N., Page, C.M., Pahalad, J., Plummer, N., Salinger, M.J., Suppiah, R., Tran, V.L., Trewin, B., Tibig, I. and Yee, D. Int. J. Climatol., 21: 269–284.

[82] Near-global impact of the Madden-Julian Oscillation on rainfall. Alexis Donald, Holger Meinke, Brendan Power, Aline de H. N. Maia, Matthew C. Wheeler, Neil White, Roger C. Stone and Joachim Ribbe. Geophysical Research Letters Volume 33, Issue 9, May 2006

[83] Global Occurrences of Extreme Precipitation and the Madden–Julian Oscillation: Observations and Predictability. Charles Jones, Duane E. Waliser, K. M. Lau, and W. Stern.

J. Climate, 17, 4575–4589. 2004

[84] Modulation of Daily Precipitation over Southwest Asia by the Madden–Julian Oscillation. Mathew Barlow, Matthew Wheeler, Bradfield Lyon, and Heidi Cullen. Mon. Wea. Rev., 133, 3579–3594. 2005

[85] Combining evolutionary information and neural networks to predict protein secondary structure. Rost B., Sander C. Proteins. 1994 May;Vol. 19(1), pp. :55-72.

[86] Multi-class protein fold recognition using support vector machines and neural networks. Chris H.Q. Ding and Inna Dubchak. Bioinformatics (2001) 17 (4): 349-358.

[87] The dynamic climatology of the beaufort to laptev sea sector of the polar basin for the winters of 1975 and 1976. Ledrew, E. F. (1985) J. Climatol., 5: 253–272.

[88] Spatial analysis of secular temperature fluctuations. Lawson, M. P., Balling, R. C., Peters, A. J. and Rundquist, D. C. (1981), J. Climatol., 1: 325–332.

[89] Finding spatio-temporal patterns in climate data using clustering. Sap, M.N Md and Awan, A.M (2005) Proceedings - 2005 International Conference on Cyberworlds, CW 2005, 23th -25th Nov. 2005.

[90] Using Clustered Climate Regimes to Analyze and Compare Predictions from Fully Coupled General Circulation Models. Hoffman, Forrest M., William W. Hargrove, David J. Erickson, Robert J. Oglesby, 2005: Earth Interact., 9, 1–27.

[91] A multivariate statistical model for forecasting anomalies of half-monthly mean surface pressure. Maryon, R. H. and Storey, A. M. (1985), J. Climatol., 5: 561–578.

[92] Prediction of Protein Secondary Structure by Combining Nearest-neighbor Algorithms and Multiple Sequence Alignments. Asaf A. Salamov, Victor V. Solovyev. Journal of Molecular Biology Volume 247, Issue 1, 17 March 1995, Pages 11–15

[93] Better Prediction of Protein Cellular Localization Sites with the k Nearest Neighbors Classifier. P. Horton, K. Nakai. Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology, Pages 147-152, 1997

[94] ProtoMap: automatic classification of protein sequences and hierarchy of protein families. Yona, G., Linial, N., and Linial, M. (2000). Nucleic Acids Res. 28, 49-55.

[95] The COG database: new developments in phylogenetic classification of proteins from complete genomes. Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. (2001). Nucleic Acids Res 29, 22-28.

[96] Development characteristics and dynamic structure of tropical intraseasonal convection anomalies. H. Rui, B. Wang. Journal of the Atmospheric Sciences, 47, 357–379. 1990

[97] Observations of the 40-50 day tropical oscillation: a review. Madden, R. A., and P. R. Julian. 1994: Mon. Wea. Rev., 122, 814-837.

[98] Detection of a 40-50 day oscillation in the zonal wind in the tropical Pacific. Madden, R.A., and P.R. Julian. 1971: J. Atmos. Sci., 28, 702-708.

[99] Description of global-scale circulation cells in the tropics with a 40-50 day period. Madden, R.A., and P.R. Julian. 1972: J. Atmos. Sci., (29), 1109-1123.

[100] Madden-Julian Oscillation. Zhang, C. Rev. Geophys., 43, RG2003, 2005.

[101] Wavenumber-frequency spectra of satellite-measured brightness in tropics. Gruber, A. Journal of the Atmospheric Sciences 31(6): 1675-1680. 1974

[102] Global-scale intraseasonal oscillations of outgoing longwave radiation and 250-mb zonal wind during northern hemisphere winter. Thomas R. Knutson, Klaus M. Weickmann, and John E. Kutzbach. Monthly Weather Review 114(3): 605-623. 1986

[103] Tropical super clusters within intraseasonal variations over the Western Pacific. Nakazawa, T. Journal of the Meteorological Society of Japan 66(6): 823-839. 1988

[104] Quasi-geostrophic motions in the equatorial area. T Matsuno. J. Meteor. Soc. Japan, vol. 44, no. 1, pp 25-43. 1966

[105] The Boreal Summer Intraseasonal Oscillation: Relationship between Northward and Eastward Movement of Convection. David M. Lawrence and Peter J. Webster. 2002: J. Atmos. Sci., 59, 1593–1606.

[106] Cloudiness Fluctuations Associated with the Northern Hemisphere Summer Monsoon. Tetsuzo Yasunari. Journal of the Meteorological Society of Japan Vol. 57, No. 3. 1979

[107] The Intraseasonal (30–50 day) Oscillation of the Australian Summer Monsoon. Harry H. Hendon and Brant Liebmann, 1990: J. Atmos. Sci., 47, 2909–2924.

[108] Intraseasonal Variability over Tropical Africa during Northern Summer. Adrian J. Matthews, 2004: J. Climate, 17, 2427–2440.

[109] The relationship between tropical cyclones of the Western Pacific and Indian Oceans and the Madden-Julian Oscillation. Liebmann, B., H. H. Hendon, et al. (1994). Journal of the Meteorological Society of Japan 72(3): 401-412.

[110] The South Pacific and southeast Indian Ocean tropical cyclone season 2002-03. Courtney, J. B. (2005). Australian Meteorological Magazine 54(2): 137-150.

[111] The modulation of tropical cyclone activity in the Australian region by the Madden-Julian oscillation. Hall, J. D., A. J. Matthews, et al. (2001). Monthly Weather Review 129(12): 2970- 2982.

[112] The 30–50 Day Mode at 850 mb During MONEX. T. N. Krishnamurti and D. Subrahmanyam, 1982: J. Atmos. Sci., 39, 2088–2095.

[113] South Asian monsoon. B. N. Goswami. Intraseasonal Variability in the Atmosphere-Ocean Climate System Springer Praxis Books 2005, pp 19-61

[114] Notebook B on the transmutation of species. Charles Darwin. 1837–1838

[115] TRMM and Other Data Precipitation Data Set Documentation. George J. Huffman, David T. Bolvin. 25 May 2014

[116] An All-Season Real-Time Multivariate MJO Index: Development of an Index for Monitoring and Prediction. Matthew C. Wheeler and Harry H. Hendon. Mon. Wea. Rev., 2004, 132, pages 1917–1932.

[117] On the predictability of the interannual behaviour of the Madden-Julian oscillation and its relationship with el Niño. Slingo, J. M., Rowell, D. P., Sperber, K. R. and Nortley, F. ,1999, Q.J.R. Meteorol. Soc., 125: 583–609.

[118] Madden Julian Oscillation Impacts. J. Gottschalck and W. 2008, Climate Prediction Center Review.

[119] Analysis of a Reconstructed Oceanic Kelvin Wave Dynamic Height Dataset for the Period 1974–2005. Paul E. Roundy and George N. Kiladis, 2007, J. Climate, Vol. 20,pp. 4341–4355.

[120] The Association of the Evolution of Intraseasonal Oscillations to ENSO. Paul E. Roundy and Joseph R. Kravitz, 2009, Phase. J. Climate, Vol. 22, pp. 381–395.

[121] Low and high frequency Madden-Julian oscillations in austral summer: interannual variations. Izumo Takeshi, Masson Sebastien, Vialard Jerome, De Boyer Montegut Clement, Behera Swadhin K., Madec Gurvan, Takahashi Keiko, Yamagata Toshio, 2010, Climate Dynamics, Vol. 35(4), pp. 669-683.

[122] Observations of the Madden Julian Oscillation during Indian Ocean Dipole events. Wilson, E. A., A. L. Gordon, and D. Kim, 2013, J. Geophys. Res. Atmos., Vol. 118, pp. 2588–2599.

[123] Impact of Indian Ocean Dipole on high-frequency atmospheric variability over the Indian Ocean. Kug, S.-J., K.P. Sooraj, Fei-Fei Jina, Jing-Jia Luo, Minho Kwon, 2009, Atmos. Res., Vol. 94, pp. 134-139.

[124] Statistical Evidence for ENSO – IOD Interaction. Koh T-Y. 2011. Journal of Applied Meteorology and Climatology.

[125] Rainfall Estimation in the Sahel. Part II: Evaluation of Rain Gauge Networks in the CILSS Countries and Objective Intercomparison of Rainfall Products. Ali, Abdou, Abou Amani, Arona Diedhiou, Thierry Lebel, 2005, J. Appl. Meteor., Vol. 44, pp. 1707–1722.

[126] The Tropical Rainfall Measuring Mission (TRMM) Sensor Package. Kummerow, C., W. Barnes, T. Kozu, J. Shiue, J. Simpson, 1998, Journal of Atmospheric and Oceanic Technology: 15:3, 809–817.

[127] Ground Validation for the Tropical Rainfall Measuring Mission (TRMM). Wolff, David B., D. A. Marks, E. Amitai, D. S. Silberstein, B. L. Fisher, A. Tokay, J. Wang, J. L. Pippitt, 2005, J. Atmos. Oceanic Technol., Vol. 22, pp. 365–380.

[128] Global precipitation estimates based on a technique for combining satellite-based estimates, rain gauge analysis, and NWP model precipitation information. Huffman, G.J., R.F. Adler, B. Rudolf, U. Schneider, and P.R. Keehn, 1995, J. Climate, 8, 1284-1295.

[129] Estimates of root-mean-square random error for finite samples of estimated precipitation. Huffman, G.J, 1997, J. Appl. Meteor., pp. 1191-1201.

[130] The global precipitation climatology project (GPCP) combined precipitation dataset. Huffman, G.J., R.F. Adler, P. Arkin, A. Chang, R. Ferraro, A. Gruber, J. Janowiak, A. McNab, B. Rudolph, and U. Schneider, 1997, Bull. Amer. Meteor. Soc. , Vol 78, pp. 5-20.

[131] An Intercomparison of Gauge Observations and Satellite Estimates of Monthly Precipitation. Pingping Xie and Phillip A. Arkin, 1995, J. Appl. Meteor., Vol. 34, pp. 1143–1160.

[132] A new correlation-based fuzzy logic clustering algorithm for fMRI. X. Golay, S. Kollias, G. Stoll, D. Meier, A. Valavanis, P. Boesiger, Mag. Resonance Med., 1998, Vol 40, pp. 249–260.

[133] Cluster analysis and display of genome-wide expression patterns. Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Proceedings of the National Academy of Science of the United States of America, 95(25):14863–14868, 1998.

[134] On feature selection through clustering. Richard Butterworth, Gregory Piatetsky-Shapiro, and Dan A. Simovici. In ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining, pages 581–584, 2005.

[135] Stability-based model selection. Tilman Lange, Mikio L. Braun, Volker Roth, and Joachim M. Buhmann. In Advances in Neural Information Processing Systems (NIPS), pages 617–624, 2003.

[136] Computer system intrusion detection: a survey. Anita K. Jones and Robert S. Sielken. Technical report, Computer Science, University of Virginia, 2000.

[137] Statistical fraud detection: A review. Richard J. Bolton and David J. Hand. Statistical Science, 17:235–255, 2002.

[138] Fault detection in an ethernet network using anomaly signature matching. Frank Feather, Dan Siewiorek, and Roy Maxion. ACM SIGCOMM Computer Communication Review, 23(4):279–288, 1993.

[139] Data Mining - Concepts & Techniques. Jiawei Han and Micheline Kamber. Morghan Kaufmann publishers, 2001.

[140] Clustering of time series data—a survey. T.Warren Liao. Pattern Recognition 38 (2005) 1857 – 1874

[141] A densitybased algorithm for discovering clusters in large spatial databases with noise. Martin Ester, Hans peter Kriegel, Jorg Sander, and Xiaowei Xu. In Proceedings of International Conference on Knowledge Discovery and Data Mining(KDD), pages 226–231, 1996.

[142] Distance metric learning for large margin nearest neighbor classification. Kilian Q. Weinberger, John Blitzer, and Lawrence K. Saul. In Advances in Neural Information Processing Systems (NIPS), 2005.

[143] Automatic formation of topological maps of patterns in a self-organizing system. Teuvo Kohonen. In Proceedings of the 2nd Scandinavian Conference on Image Analysis, pages 214–220, 1981.

[144] Clustering of the self-organizing map. Juha Vesanto and Esa Alhoniemi. IEEE Transactions on Neural Networks, 11(3):586´ lC600, 2000.

[145] Tidier drawings of trees. Edward M. Reingold and John S. Tilford. IEEE Transactioons on Software Engineering, 7(2):223–238, March 1981.

[146] Tree-maps: a space-filling approach to the visualization of hierarchical information structures. Brian Johnson and Ben Shneiderman. In Proceedings of the 2nd conference of Visualization '91, pages 284–291, 1991.

[147] The electrostatic polarization is essential to dierentiate the helical propensity in polyalanine mutants. C. Wei, D. Tung, Y. M. Yip, Y. Mei, and D. Zhang, Communication. The Journal of Chemical Physics, vol. 134, no. 171101, 2011.

[148] Are there pathways for protein folding? C. Levinthal. J. Chem. Phys., Vol 65, p.44-47, 1968.

[149] Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. V. Villegas, J. C. Martinez, F. X. Aviles, and L. Serrano. J. Mol. Biol., Vol.283, pp. 1027-1036, 1998.

[150] Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. F. Chiti, N. T addei, P . M White, M. Buccian tini, F. Magherini, M. Stefani, and C. M. Dobson. Nature Struct. Biol. , 6:1005-1009, 1999.

[151] Effects of core mutations on the folding of a beta-sheet protein: implications for backbone organization in the i-state. M. Lorch, J. Mason, A. Clark e, and M. Parker. Biochemistry, 38:1377-1385, 1999

[152] Theory for the folding and stability of globular proteins K. A. Dill. Biochemistry, 24:1501-1509, 1985.

[153] Molecular dynamics studies of folding of a protein-like model. N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley , and E. I. Shakhnovich. Folding & Design , 3:577-587, 1998.

[154] Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. E. Alm and D. Baker. Proc. Natl. A c ad. Sci. USA , 96:11305-11310, 1999.

[155] A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. O. V. Galzitskaya and A. V. Finkelstein. Proc. Natl. A c ad. Sci. USA , 96:11299-11304, 1999.

[156] A simple model for calculating the kinetics of protein folding from three- dimensional structures. V. Munoz and W. A. Eaton. Proc. Natl. Acad. Sci. USA , 96:11311-11316, 1999.

[157] Non-interacting local-structure model of folding and unfolding transition in globular proteins. I. formulation. N. Go and H. Abe. Biopolymers , 20:991-1011, 1981.

[158] Kinetics of protein folding. A lattice model study of the requirements for folding to the native state.A. Sali, M. Karplus and E. I. Shakhnovich. Journal of molecular biology,1994, Vol. 235.5, pp: 1614-1636

[159] Theoretical studies of protein-folding thermo dynamics and kinetics.

E. I. Shakhnovich. Curr. Opin. Struc. Biol. , Vol 7, pp. 29-40, 1997.

[160] Pearson product-moment correlation coefficient. Wikipedia. 2014.

[161] On the shortest spanning subtree of a graph and the traveling salesman problem. J. Kruskal, Proceedings of the American Mathematical society, vol. 7, no. 1, pp. 48-50, 1956.

[162] Shortest connection networks and some generalizations. R. Prim Bell system technical journal, vol. 36, no. 6, pp. 1389-1401, 1957.

[163] Active/break cycles: Diagnosis of the intraseasonal variability of the Asian summer monsoon. Annamalai, H. and J.M. Slingo. Climate Dynamics, Vol 18, 1-2 (2001), pp. 85-102

[164] An all-season real time multivariate MJO index: development of an index for monitoring and prediction, Wheeler, M.C. and H.H. Hendon, 2004: Mon. Wea. Rev., 32, 1917-1932.

[165] Choice of South Asian summer monsoon indices. Wang, B. and Z. Fan, 1999: Bull. Amer. Meteor. Soc., 80, 629-638.

[166] Interannual variability of Asian summer monsoon: Contrast between the Indian and western North Pacific-East Asian monsoons. Wang, B., R. Wu, K.-M. Lau, 2001: J. Climate, 14, 4073-4090.

[167] Boreal summer quasi-monthly oscillation in the global tropics. Wang, B., P. Webster, K. Kikuchi, T. Yasunari, and Y. Qi, 2006: Clim. Dyn., 27, 661-675.

[168] Choice of South Asian summer monsoon indices. Wang, B., Z. Fan, 1999: Bull. Amer. Meteor. Soc., 80, 629-638.

[169] Interannual variability of Asian summer monsoon: Contrast between the Indian and western North Pacific-East Asian monsoons. Wang, B., R. Wu, K.-M. Lau, 2001: J. Climate, 14, 4073-4090.

[170] Near-global impact of the Madden-Julian Oscillation on rainfall, Donald, A., H. Meinke, B. Power, A. de H. N. Maia, M. C. Wheeler, N. White, R. C. Stone, and J. Ribbe (2006), Geophys. Res. Lett., 33.

[171] The Influence of the Madden–Julian Oscillation on Canadian Wintertime Surface Air Temperature. Lin Hai, Gilbert Brunet, 2009: Mon. Wea. Rev., 137, 2250–2262.

[172] Impact of the Madden–Julian Oscillation on Summer Rainfall in Southeast China. Zhang Lina, Bizheng Wang, Qingcun Zeng, 2009: J. Climate, 22, 201–216.

[173] Impacts of the Madden–Julian Oscillation on Australian Rainfall and Circulation. Wheeler, Matthew C., Harry H. Hendon, Sam Cleland, Holger Meinke, Alexis Donald, 2009: J. Climate, 22, 1482–1498.

[174] Influence of the Madden-Julian Oscillation on Indonesian rainfall variability in austral summer. Hidayat, R., Kizu, S, 2010: International Journal of Climatology, 30:12, 1816-1825.

[175] Impacts of the MJO on winter rainfall and circulation in China. Jia, X., L. J. Chen, F. M. Ren, and C. Y. Li, 2011: Adv. Atmos. Sci., 28(3), 521–533.

[176] Climate extremes in Malaysia and the equatorial South China Sea, Salahuddin A, Curtis S, 2011, Global and Planetary Change 78, 83-91.

[177] MJO simulation diagnostics. Waliser, D., and Coauthors, 2009: J. Climate, 22, 3006–3030.

[178] Cloudiness fluctuations associated with the Nothern Hemisphere summer monsoon. Yasunari, T., 1979: J. Meteor. Soc. Japan, 57, 227-242.

[179] Ocean Temperatures Affect Intensity of the South Asian Monsoon and Rainfall". NASA GSFC. Goddard Space Flight Center (2002). National Aeronautics and Space Administration. Retrieved 2009-11-06.

[180] 2011: The Japanese 55-year Reanalysis "JRA-55": an interim report. Ebita, A., S. Kobayashi, Y. Ota, M. Moriya, R. Kumabe, K. Onogi, Y. Harada, S. Yasui, K. Miyaoka, K. Takahashi, H. Kamahori, C. Kobayashi, H. Endo, M. Soma, Y. Oikawa and T. Ishimizu, , SOLA, 7, 149-152.

[181] Monsoon and ENSO: Selectively Interactive Systems. Webster, P. J. and S. Yang, 1992: Quart. J. Roy. Meteor. Soc., 118, 877-926.

[182] Aspects of the 40–50 day oscillation during the northern summer as inferred from

outgoing longwave radiation, Lau, K.-M., and P. H. Chan (1986), Mon. Weather Rev., 114, 1354–1367.

[183] Multiple phenomena in the tropical atmosphere over the western Pacific, Sui, C.-H., and K. M. Lau (1992), Mon. Weather Rev., 120, 407–430.

[184] The boreal summer intraseasonal oscillation: Relationship between northward and eastward movement of convection, Lawrence, D. M., and P. J. Webster (2002), J. Atmos. Sci., 59, 1593–1606.

[185] The intraseasonal (30– 50 day) oscillation of the Australian summer monsoon, Hendon, H. H., and B. Liebmann (1990), J. Atmos. Sci., 47, 2909–2923.

[186] Tropical convection and precipitation regimes in the western United States, Mo, K., and R. W. Higgins (1998), J. Clim., 10, 3028–3046.

[187] Occurrence of extreme precipitation events in California and relationships with the Madden-Julian Oscillation, Jones, C. (2000), J. Clim., 13, 3576–3587.

[188] The influence of the Madden- Julian Oscillation on precipitation in Oregon and Washington, Bond, N. A., and G. A. Vecchi (2003), Weather Forecasting, 18, 600–613.

[189] Intraseasonal modulation of South American summer precipitation, Paegle, J. N., L. A. Byerle, and K. C. Mo (2000), Mon. Weather Rev., 128, 837–850.

[190] Subseasonal variations of rainfall in the vicinity of the South American low-level jet stream and comparison to those in the South Atlantic Convergence Zone, Liebmann, B., G. N. Kiladis, C. S. Vera, A. C. Saulo, and L. M. V. Carvalho (2004), J. Clim., 17, 3829–3842.

[191] Intraseasonal variability over tropical Africa during northern summer, Matthews, A. J. (2004), J. Clim., 17, 2427–2440.

[192] The relationship between tropical cyclones of the western Pacific and Indian oceans and the Madden-Julian Oscillation, Liebmann, B., H. Hendon, and J. Glick (1994), J. Meteorol. Soc. Jpn., 72, 401–411.

[193] Dynamical aspects of twin tropical cyclones associated with the Madden- Julian Oscillation, Nieto Ferreira, R., W. H. Schubert, and J. J. Hack (1996), J. Atmos. Sci., 53, 929–945.

[194] Modulation of eastern North Pacific hurricanes by the Madden-Julian Oscillation, Maloney, E. D., and D. L. Hartmann (2000), J. Clim., 13, 1451–1460.

[195] The modulation of tropical cyclone activity in the Australian region by the Madden-Julian Oscillation, Hall, J. D., A. J. Matthews, and D. J. Karoly (2001), Mon. Weather Rev., 129, 2970– 2982.

[196] Intercomparison of the principal modes of interannual and intraseasonal variability of the North American monsoon system, Higgins, R. W., and W. Shi (2001), J. Clim., 14,

403–417.

[197] The 30–70 day oscillations in the tropical Atlantic. Foltz, G. R., and M. J. McPhaden (2004), Geophys. Res. Lett., 31

[198] More extreme swings of the South Pacific convergence zone due to greenhouse warming. CaI W., Lengaigne M, Borlace S., Collins M., Cowan T., McPhaden M.J., Timmermann A., Power S, Brown J., Menkes C., Ngari A., Vincent E.M., Widlansky M.J (2012). Nature, vol. 488, pp. 365-370

[199] Intraseasonal interactions between the Tropics and Extratropics in the Southern Hemisphere. Berbery E. H., Nogue's-Paegle J. (1993), J. Atmos. Sci., vol. 50, pp. 1950-1965.

[200] Sensitivity of Australian rainfall to inter-El Niño variations. Wang G., Hendon H. H.(2007), J. Climate, vol. 20, pp. 4211-4226.

[201] Intraseasonal (30-60 day) fluctuations of outgoing longwave radiation and 250 mb streamfunction during northern winter. Weickmann, K. M., Lussky G. R. and Kutzbach J. E.(1985) Mon. Wea. Rev., vol. 113, pp. 941-961.

[202] Atmospheric angular momentum and the length of the day: A common fluctuation with a period of 50 days. Langley R. B., King R. W., Shapiro I. I., Rosen R. D. and Salstein D.A. // Nature, 1981, vol. 294, pp.730–732.

[203] Exchange of momentum among atmosphere, ocean, and solid earth associated with the Madden-Julian Oscillation. Gutzler D. S. and Ponte R. M. (1990) J. Geophys. Res., vol. 95, pp. 18,679 –18,686.

[204] The dynamics of intraseasonal atmospheric angular momentum oscillations.

Weickmann K. M., Kiladis G. N. and Sardeshmukh P. D. (1997) J. Atmos. Sci., vol. 54, pp. 1445–1461.

[205] The Manifestation of the Madden Julian Oscillation in Global Deep Convection and in the Schumann Resonance Intensity. Anyamba E., Williams E., Susskind J., Fraser Smith A. and Fullerkrug M. (2000) American Meteorology Society, vol.5 , No.8, pp. 1029-44.

[206] Intraseasonal Air-Sea Interactions at the Onset of El Nino.Bergman J. W., Hendon H. H. and Weickmann K. M. (2001) J. Climate, vol. 14, pp. 1702-1719.

[207] Forcing of intraseasonal Kelvin waves in the equatorial Pacific.Kessler W. S., McPhaden M. J. and Weickmann K. M. (1995) J. Geophys. Res., vol. 100, pp. 10613-10631.

[208] Abrupt termination of the 1997- 98 El Nino in response to a Madden-Julian oscillation. Takayabu Y.N., Iguchi T., Kachi M., Shibata A., Kanzawa H., (1999) Nature, vol. 402, pp. 279-282.

[209] Voyage of Rediscovery, Finney B. 1994, Univ. of Calif. Press, Berkeley, 401 p.

[210] Convectively coupled equatorial waves: Analysis of clouds and temperature in the wavenumber–frequency domain. Wheeler M. C. and Kiladis G. N. 1999J. Atmos. Sci., vol. 56, pp. 374-399.

[211] Prospects for improved forecasts of weather and short-term climate variability on subseasonal (2 week to 2 month) time scales. Schubert S., Dole R., Dool H.v.d., Suarez M., Waliser D., 2002 Proceedings of the workshop, Mitchellville, MD, NASA/TM 2002-104606, vol. 23, p. 71.

[212] Predictability and Forecasting. Intraseasonal Variability of the Atmosphere-Ocean Climate System, W. K. M. Lau, and D. E. Waliser. (Eds.), Springer, Heidelberg, Germany, 474 pp. 351.

[213] Potential Predictability of the Madden-Julian Oscillation. Waliser D. E., Lau K. M., Stern W. and Jones C. (2003) Bull. Amer. Meteor. Soc., vol. 84, pp. 33-50.

[214] Real-time monitoring and prediction of modes of coherent synoptic to intraseasonal tropical variability. Wheeler M. and Weickmann K.M. (2001) Mon. Wea. Rev., vol. 129, pp. 2677-2694.

[215] Planetary scale circulations in the presence of climatological and wave induced heating. Salby M. L., Garcia R. R. and Hendon H. H. (1994) J. Atmos. Sci., vol. 51, pp. 2344–2367.

[216] Development characteristics and dynamic structure of tropical intraseasonal convection anomalies. Rui, H. and Wang B. (1990) J. Atmos. Sci., vol. 47, pp. 357–379.

[217] Observations of a convectively coupled Kelvin wave in the eastern Pacific ITCZ. Straub K. H. and Kiladis G. N. (2002) J. Atmos. Sci., vol. 59, pp. 30-53.

[218] Seasonal variations of the 40–50 day oscillation in the tropic. Madden R. A. (1986) J. Atmos. Sci., vol. 43, pp. 3138–3158.

[219] Seasonal variations in the spatial structure of intraseasonal tropical wind fluctuations. Gutzler D. S. and Madden R. A. (1989) J. Atmos. Sci., vol. 46, pp. 641–660.

[220] Seasonality of the Madden-Julian Oscillation, Zhang C. and Dong M. (2004) J. Clim., vol.17, pp. 3169–3180.

[221] Interannual fluctuations of intraseasonal variance of near-equatorial zonal winds. Gutzler D. S. (1991) J. Geophys. Res., vol. 96, pp. 3173–3185.

[222] Mechanism of the zonal displacements of the Pacific warm pool: Implications for ENSO. Picaut, J., Ioualalen M., Menkes C., Delcroix T. and McPhaden M. J.. (1996) Science, vol. 274, pp. 1486–1489.

[223] Temporal variability of the 40–50-day oscillation in tropical convection. Anyamba E. K., and Weare B. C. (1995) Int. J. Climatol., vol., 15, pp. 379-402.

[224] Some potential forcing mechanisms of the year-to-year variability of the tropical convection and its intraseasonal (25 ± 70-day) variability. Fink, A., and Speth P. (1997) Int. J. Climatol., vol. 17, pp. 1513–1534.

[225] The interannual variability of the Madden-Julian Oscillation in an ensemble of GCM simulations. Gualdi, S., Navarra A. and Tinarelli G. (1999) Clim. Dyn., vol. 15, pp. 643–658.

[226] Interannual variation of the Madden–Julian oscillation during austral summer. Hendon H. H., Zhang C. and Glick J. D. (1999) J. Climate, vol.12, pp. 2538–2550.

[227] The relationship between convection and sea surface temperature on intraseasonal timescales. Woolnough S. J., Slingo J. M. and Hoskins B. J. (2000) J. Clim., vol. 13, pp. 2086–2104.

[228] Coupled model simulations of boreal summer intraseasonal (30-50 day) variability, Part I: Systematic errors and caution on use of metrics. Sperber, K. R., and H. Annamalai, (2008) Clim. Dyn., 31: 345-372

[229] Interannual variations of the boreal summer intraseasonal variability predicted by ten atmosphere-ocean coupled models. Kim, H.-M., I.-S. Kang, B. Wang, and J. Y. Lee, (2007) Clim. Dyn., 2008, Vol 30, Issue 5, pp 485-496

[230] Impact of atmosphere-ocean coupling on the predictability of monsoon intraseasonal oscillations. Fu, X., B. Wang, D. Waliser, and L. Tao, (2007) J. Atmos. Sci., 64, 157-173.

[231] The role of the ocean in the Madden-Julian Oscillation: Sensitivity of an MJO forecast to ocean coupling. Woolnough, S. J., F. Vitart, and M. A. Balmaseda, (2007) Quart. J. Roy. Meteor. Soc., 133, 117-128.

[232] An All-Seasonal Real-Time Multivariate MJO Index. Wheeler, M. (2009).

http://www.cawcr.gov.au/sta/mwheeler/maproom/RMM/.

[233] Active/break cycles: Diagnosis of the intraseasonal variability of the Asian summer monsoon. Annamalai, H. and J.M. Slingo. Climate Dynamics, Vol 18, 1-2 (2001), pp. 85-102

[234] Goulet, L., and J.-P. Duvel, 2000: A new approach to detect and characterize intermittent atmospheric oscillations: Application to the intraseasonal oscillation. J. Atmos. Sci., 57, 2397– 2416.

[235] Roundy, P. E., C. J. Schreck III, and M. A. Janiga, 2009: Contributions of convectively coupled equatorial Rossby waves and Kelvin waves to the Real-Time Multivariate MJO Indices. Mon. Wea. Rev., 137, 469–478.

[236] Straub, K., 2013: MJO initiation in the Realtime Multivariate MJO Index. J. Climate, 26, 1130–1151

[237] Ventrice, M. J., C. D. Thorncroft, and P. E. Roundy, 2011: The Madden–Julian Oscillation's influence on African easterly waves and downstream tropical cyclogenesis. Mon. Wea. Rev., 139, 2704–2722.

[238] Musmeci, Nicoló, Tomaso Aste, and Tiziana Di Matteo. "Relation between financial market structure and the real economy: comparison between clustering methods." PloS one 10.3 (2015): e0116201

[239] Musmeci, Nicoló, Tomaso Aste, and Tiziana Di Matteo. "Risk diversification: a study of persistence with a filtered correlation-network approach." arXiv preprint arXiv:1410.5621 (2014).

[240] Song, Won-Min, T. Di Matteo, and Tomaso Aste. "Hierarchical information clustering by means of topologically embedded graphs." PLoS One 7.3 (2012): e31929.

[241] Ashburner, Michael, et al. "Gene Ontology: tool for the unification of biology." Nature genetics 25.1 (2000): 25-29.

[242] Langfelder, Peter, Bin Zhang, and Steve Horvath. "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R." Bioinformatics 24.5 (2008): 719-720.

[243] Schmid, Friedrich, and Rafael Schmidt. "Multivariate conditional versions of Spearman's rho and related measures of tail dependence." Journal of Multivariate Analysis 98.6 (2007): 1123-1140.

[244] Y. Mu, P. H. Nguyen, and G. Stock, "Energy landscape of a small peptide revealed by dihedral angle principal component analysis.," Proteins, vol. 58, no. 1, pp. 45-52, 2005.

[245] Mahdavi Damghani, Babak (2012). "The Misleading Value of Measured Correlation". Wilmott 2012 (1): 64–73.

[246] Dietrich, Cornelius Frank (1991) Uncertainty, Calibration and Probability: The Statistics of Scientific and Industrial Measurement 2nd Edition, A. Higler. [4] Aitken, Alexander Craig (1957) Statistical Mathematics 8th Edition. Oliver & Boyd.

[247] Friedman, Jerome H. (1998). "Data Mining and Statistics: What's the connection?". Computing Science and Statistics 29 (1): 3–9.

[248] Russell, Stuart; Norvig, Peter (2003) [1995]. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall. ISBN 978-0137903955.

[249] Deza, Elena; Deza, Michel Marie (2009). Encyclopedia of Distances. Springer. p. 94.

[250] Kaufman, Leonard, and Peter J. Rousseeuw. Finding groups in data: an introduction to cluster analysis. Vol. 344. John Wiley & Sons, 2009.

[251] Fraley, Chris, and Adrian E. Raftery. "How many clusters? Which clustering method? Answers via model-based cluster analysis." The computer journal 41.8 (1998):

578-588.

[252] Anderberg, Michael R. Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks. Vol. 19. Academic press, 2014.