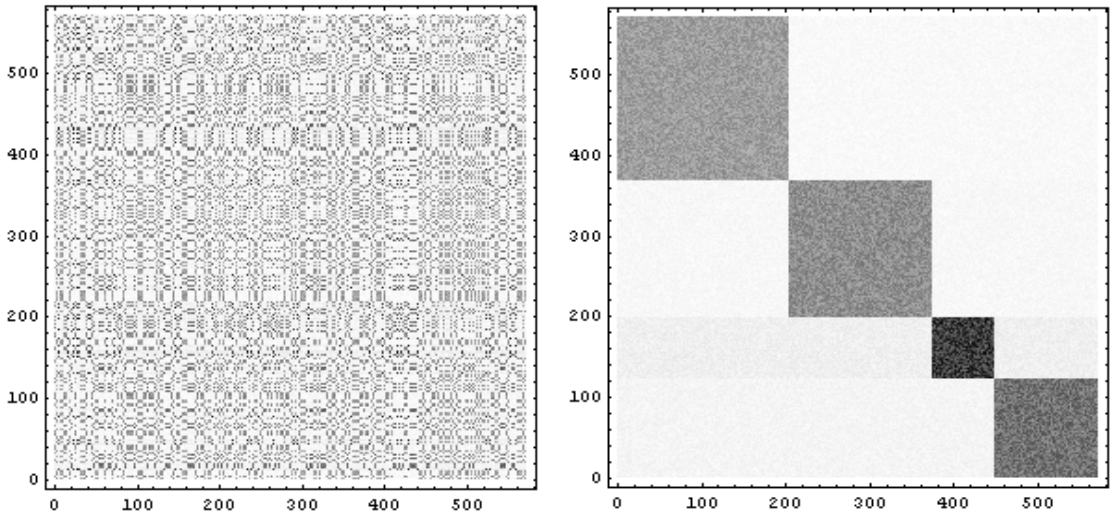
CHALMERS



Finding meta stable states in proteins

Identification of meta stable states in VPAL-peptide's dynamics

Master's Thesis in the Master Degree Programme,
Complex adaptive systems

SVITLANA RUZHYTSKA

Department of Energy and Environment
Division of Physical Resource Theory
Complex Systems Research Group
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden, 2012
Report No. 2012:4

REPORT NO. 2012:4

Finding meta stable states in proteins: Identification of meta stable states in VPAL-peptide's dynamics

SVITLANA RUZHYTSKA

Department of Energy and Environment, Division of Physical Resource Theory

CHALMERS UNIVERSITY OF TECHNOLOGY

Göteborg, Sweden 2012

Finding meta stable states in proteins: Identification of meta stable states in VPAL-peptide's dynamics

SVITLANA RUZHYTSKA

Technical Report No: 2012:4

Examiner: Martin Nilsson Jacobi
Department of Energy and Environment, *Division of Physical Resource Theory*
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31-772 10 00

Göteborg, Sweden, 2012.

ACKNOWLEDGEMENTS

I would like to thank all those who helped and supported me along this interesting very exciting and sometimes hard journey of master thesis research.

First of all I would like to thank Chalmers acceptance committee for accepting me to the wonderful master degree program, Complex adaptive systems, and giving me the chance to study in Sweden. These studies would be impossible without the financial support provided by the Swedish institute, who accepted me to the Visby program granting me scholarship, for which I am eternally grateful.

Great appreciation I would like to express to Prof. Kristian Lindgren, who accepted me to his research group, giving me such a great opportunity to carry out this work.

The most special thanks and sincere appreciation I would like to address to my supervisors Dr Martin Nilsson Jacobi and Dr Dmitry Nerukh, who played key roles in my research providing me with the method and subject of research. Their amazing collaboration, guidance, trust, support, and continual encouragement helped me to put together all pieces of the puzzle and carry out this important research not only for the academia but also for the worldwide scientific society in protein research through our publication in Journal of Chemical Physics. I would like to separately thank them for their patience, valuable comments and suggestions regarding this master's thesis report.

My next gratitude goes both ways to this Master thesis and a special person in my life, John Finér. Thanks to this master's program we have met, thanks to this master thesis we have became close, and now thanks to him I have completed them both. I would like to thank him for being a good listener and advisor, and for greatly helping me with my presentation. Thank you for being by my side.

Finally, I would like to thank my parents for all they have done for me, for being my inspiration and encouragement in life, for believing in me and helping me when I need it most. Thank you! I love you.

SUMMARY

Proteins are very large biological molecules, responsible for many functions in living cells and organisms. Ever since they were recognized as a distinct class in 1789 by Antoine Fourcroy and others, proteins were, and still are due to their vital importance and high complexity, subject of studies of scientists from various fields of science all over the world. It was discovered, that each protein consists of one or more chains of amino acids compactly folded in space. This fully folded structure protein attain from the initial unfolded structure through intermediate ones. Moreover, scientists believe that protein properties and functions strongly depend on these structures.

Fully folded structure, called native state, is the most preferable, due to the lowest overall free energy, so once it is reached, it holds for a long time, which increases chances for success in the experimental research of it (using, for example, *X-ray crystallography* or *Nuclear Magnetic Resonance* techniques). Unfortunately, this is not the case for intermediate structures, due to the either too fast (hard to record and notice) or too slow (hard to separate) folding process. That is when united physical, mathematical, and computer research dominates.

This master thesis work presents a mathematical approach for studying intermediate protein structures. These intermediate structures, usually called meta stable states, are usable in multiple fields of research. For example, biochemistry and pharmaceutics are using them to optimize the shape of the drug molecules for achieving the best possible binding properties with respect to the target molecules; In biophysics it helps to find the folding pathways in the potential energy landscape; and in computational biology to reduce the amount of data needed to be stored obtained from the molecular dynamics simulations and also so that the problem can be partitioned into smaller pieces and be run in parallel on multiple computers.

The goal of this work was to establish, if it is possible to apply spectral method for finding meta stable states of proteins. This was achieved through 3 major steps. First step was to obtain data for analysis, namely to perform molecular dynamics simulations resulting in the protein dynamics trajectory. Second step was to prepare this data for the analysis, namely to divide conformational trajectory into microstates and construct transition probability matrix. And, finally, the third step was to analyze data, by applying spectral method, which resulted in meta stable states.

Tests were performed on the small peptide (peptide is the name for protein with the length of the chain less than 100 amino acids) consisting of only 4 amino acids named Valine-Proline-Alanine-Liucine (VPAL) [32]. This peptide, despite of its small size, attains properties of a protein, i.e. (folds through / has) meta stable states, however, because of its size, it makes it possible to get results comparatively quickly.

Results obtained for VPAL-peptide were confirmed by those known in literature, which let us to conclude that it is possible to apply spectral method for finding meta stable states.

General structure outlining main topics of the performed Master thesis research is schematically shown on the figure 1.

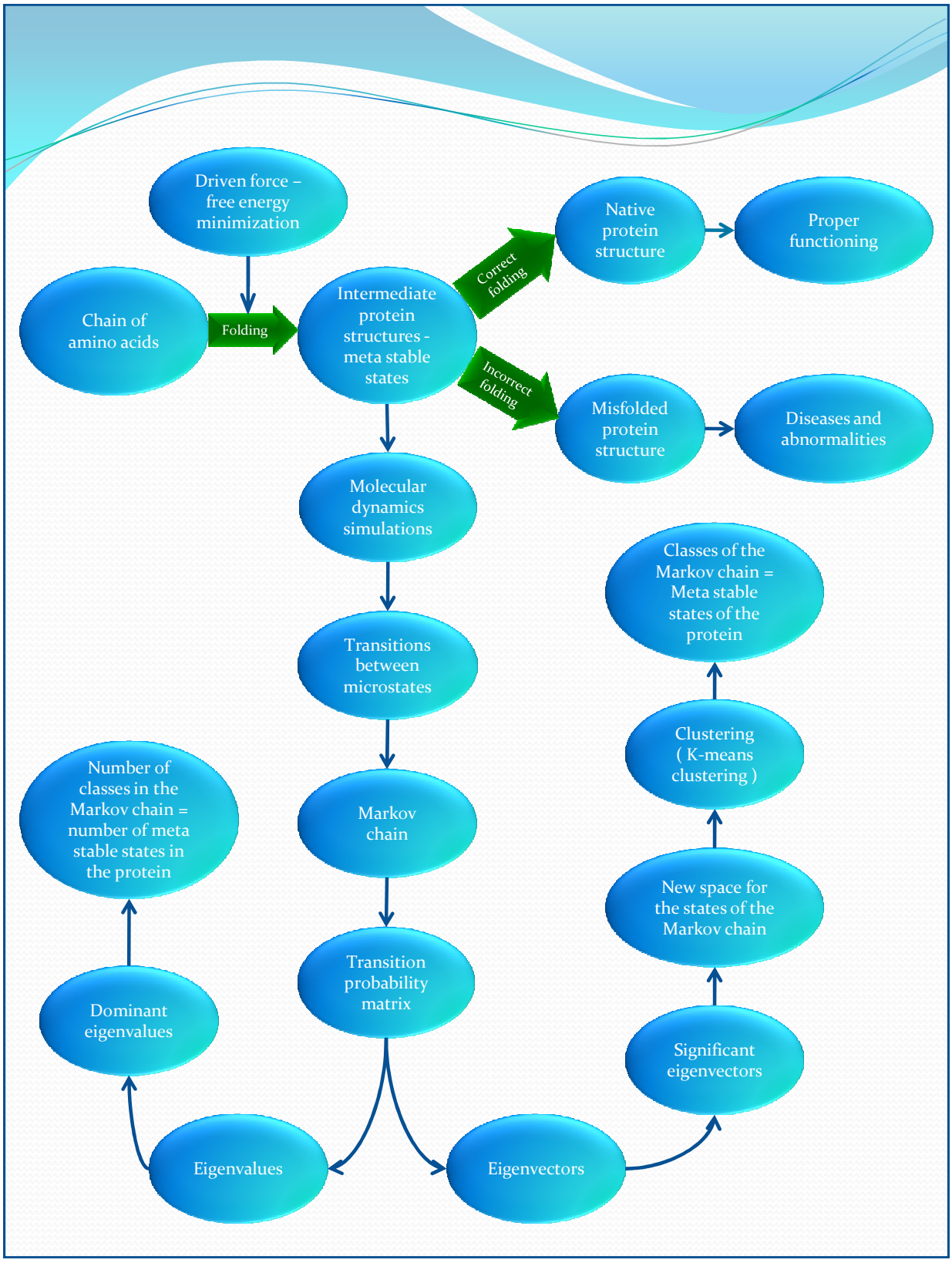


Figure 1. Key topics of the Master thesis research

CONTENTS

List of abbreviations

- DNA - DeoxyriboNucleic Acid,
- RNA - RiboNucleic Acid,
- CFTR protein - Cystic Fibrosis TRansmembrane conductance regulator,
- ADDLs (pronounced “addles”) - Amyloid ß-Derived Diffusible Ligands (the toxic proteins),
- SI - The International System of Units (abbreviated SI from French: Système International d'unités),
- AMBER - Assisted Model Building with Energy Refinement (molecular dynamics simulations package),
- GROMACS - GROningen MAchine for Chemical Simulations (molecular dynamics simulations package),
- CHARMM - Chemistry at HARvard Macromolecular Mechanics (molecular dynamics simulations package),
- SPC - Simple Point Charge (water models),
- SPC/E - Extended Simple Point Charge (water models),
- TIP3P - Transferable Intermolecular Potential 3 Point (water models),
- HOH - water molecule H_2O,
- VPAL - Valine-Proline-Alanine-Liucine (peptide).

List of symbols and notations

- G is the Gibbs free energy (SI unit is Joule [J]),
- H is the enthalpy (SI unit is Joule [J]),
- S is the entropy (SI unit is Joule per Kelvin [J/K]),
- T is the temperature (SI unit is Kelvin [K]),
- U is the internal energy (SI unit is Joule [J]),
- p is pressure (SI unit is Pascal [Pa]),
- $\mathcal{V}$ is volume (SI unit is cubic meter [m^3]),
- E_p is the potential energy (SI unit is Joule [J]),
- E_k is the kinetic energy (SI unit is Joule [J]),
- V is the potential energy (computer simulation units),
- r is the distance between the two connected atoms (computer simulation units),
- r_{ij} is the distance between the two unconnected atoms (computer simulation units),
- θ is the angle between the three connected atoms (computer simulation units),
- φ is the torsional angle between the four connected atoms (computer simulation units),
- q represents three Cartesian coordinates x, y, and z,
- N is the total number of atoms in the protein molecule,
- F is the forces acting on atoms (computer simulation units),
- t is the time (computer simulation units),
- m is the mass (computer simulation units),
- v is the velocity (computer simulation units),
- P is the transition probability matrix,
- A is any square matrix,
- I is the identity matrix,
- n is the total number of rows and columns in the square matrix,
- λ is the eigenvalue,
- ν is the eigenvector,
- k is the algebraic multiplicity,
- m is the number of linearly independent eigenvectors corresponding to one eigenvalue.

Introduction

The goal of the research was to find meta stable protein structures, appearing during the protein folding, applying spectral method. Chapter 1 contains short history of the protein research. Chapter 2 places meta stable structures among others; Chapter 3 shows the importance of the protein folding research, by explaining how the "correctly" and "incorrectly" folded proteins effect functions of living organisms; Chapter 4 explains why and how meta stable structures are formed (considering protein from the thermodynamic point of view); in Chapter 5 molecular dynamics simulations are discussed, to show how the data for the analysis, namely protein dynamics trajectory, is collected (considering protein from the mechanic point of view); Chapter 6 introduces some definitions and properties of Markov chains and describing them transition probability matrices; Chapter 7 reminds definitions of eigenvalues and eigenvectors and shows connections between their properties, properties of the transition probability matrices and properties of the Markov chains; and finally Chapter 8 introduces spectral method. Results obtained from the application of the spectral method for finding meta stable structures (or states) of the Valine-Proline-Alanine-Liucine peptide (VPAL-peptide) are published in the Journal of Chemical Physics (see [32]).

1 Short history of the protein research

Proteins were first recognized as a distinct class in 1789 by Antoine François de Fourcroy.

The word "protein" comes from Greek word *prota*, which means "of primary importance". This name was introduced by the Swedish chemist Jöns Jakob Berzelius in 1838 for large organic compounds with almost equivalent empirical formulas. This name was used because the studied organic compounds were primitive, but seemed to be very important for animal nutrition [1].

The next crucial step of the protein study was made by James B. Sumner in 1926 by showing that enzymes could be isolated and crystallized.

In 1955 Sir Frederick Sanger determined the complete amino acid sequence of the first protein - insulin. This was a first prove, that all proteins have specific structure.

In 1958 first three-dimensional structures were solved by X-ray diffraction analysis by Max Perutz (for hemoglobin) and Sir John Cowdery Kendrew (for myoglobin).

Since then scientists all over the world established amino acid sequences and three-dimensional structures for thousands of proteins and stored this information in *Protein Data Bank.*

2 Protein folding and structures

> "The amino acid sequences of polypeptide chains (…) only make functional sense when they are in the three dimensional arrangement that characterizes them in the native protein structure"
>
> C. B. Anfinsen, 1972, Nobel Lecture

Proteins are very complex biological molecules, so different scientists distinguish different types of protein structures, depending on the focus of their research. For example, one type of classification includes primary, secondary, tertiary, and quaternary structures, while another one separates initial, intermediate, and final structures. For our purposes we will consider the second type of classification. It is used for distinguishing structures, appearing during three major stages of protein folding, characteristics of which are summarized in table 1.

Table 1. Protein folding stages and their characteristics

Stage number	Structure Name	Chain description	Structure characteristics
Stage 1	Initial	Unfolded	Unstable
Stage 2	Intermediate	Partly folded	Meta stable
Stage 3	Final	Fully folded	(Quite) Stable

Our research is focused on the intermediate structures or meta stable structures appearing during the second stage. In the following subsections we will describe each stage in more detail introducing some common terms and pointing out some of the main difficulties arising in the studies of each stage.

2.1 Stage 1: Polypeptide chain and its characteristics

Proteins are made of amino acids linked into linear chains, called **polypeptide chains**. Proteins are formed by one or several such chains. These chains are very flexible, so one can think of them as of "open" beads necklaces. Just as the necklace they can have different length and sequence combinations.

First let's talk about the length. Polypeptide structures build from 2 to 100 amino acids are usually called **peptides**, longer structures are classified as **proteins**. [1] A typical protein contains 200-300 amino acids. But much larger proteins also exist. The largest to date is *titin*, a protein found in skeletal cardiac muscle; one version of it contains 34,350 amino acids in a single chain! [2]

As for the sequence combinations, the sequence of the polypeptide chain is defined by a gene with a genetic code. There are only 20 standard amino acids, that appear in living organisms. But in total the number of different proteins, which is possible to produce from 20 amino acids is enormous. For example, for a protein constructed of 10 amino acids it is possible to have 20^{10} different sequence combinations, which is approximately equal to 10^{13} or 10 trillions of different structures. For example, only in *E. coli* cell scientists established about 3000 different proteins. [1]

2.2 Stage 2: Protein folding and intermediate structures

After the chain is formed, it starts packing in space into a compact, fully folded structure, to minimize protein's overall free energy. This packing is called **protein folding**. It is not random, but through a specific time-ordered sequence of intermediate structures, known as “folding pathway”, which was confirmed theoretically and experimentally: theoretically in 1968 by Cyrus Levinthal, who pointed out that protein containing 100 amino acids would need 4×10^{9} years to fold, if it would test at least two conformational possibilities per amino acid (“Levinthal’s paradox” [3,4]), while instead it takes in general only $10^{-1}-10^{3}$ seconds; and experimentally in 1975 by Christian Anfinsen and Harold Scheraga, who established that a protein chain folds spontaneously and reproducibly to a unique three dimensional structure, when placed in aqueous solution [5].

2.3 Stage 3: Folded protein and native state

There exist two kinds of fully folded protein structures: native and misfolded. Native structure (or state) of the protein is the one responsible for the protein's proper functioning. Unlike the misfolded state, native state corresponds to the global minimum of the free energy.

Once proteins fold, they stay in such state for a long time, but not forever. After some time folding process is replaced by the process of unfolding, which makes proteins - dynamical systems, being in constant motion.

3 Importance of studying protein folding

It is a common knowledge, that proteins are very important for all living organisms. First of all, they are involved in almost all processes occurring in living organisms, and second of all, this involvement can be both positive, caused by the "correctly" folded proteins into native state, resulting in proper functioning and life, and negative, caused by the misfolded and damaged proteins, resulting in illness or even death.

Positive involvement is described in the next subsection, where a brief summary of proteins functions is presented (see table 2), for more detailed description see [1]. And negative involvement is briefly described in the subsection after that, where some known diseases are mentioned.

3.1 Performed functions

Table 2. List of the functions performed by different proteins

Functional group name	Description of the function	Examples of proteins
Enzymes	Proteins that catalyze chemical and biochemical reactions within living cell and outside.	*DNA*- and *RNA-polymerases*
Hormones	Proteins that are responsible for the regulation of many processes in organisms.	*insulin, endorfine*
Transport proteins	These proteins are transporting or store some other chemical compounds and ions.	*hemoglobin*, *myoglobin*, *albumin*

Antibodies	Proteins that involved into immune response of the organism to neutralize large foreign molecules, which can be a part of an infection.	
Structural proteins	These proteins are maintain structures of other biological components, like cells and tissues.	*collagen, elastin, α-keratin, sklerotin*
Motor proteins	These proteins can convert chemical energy into mechanical energy.	a*ctin, myosin*
Receptors	These proteins are responsible for signal detection and translation into other type of signal.	*id rhodopsin* - light detecting protein
Signaling proteins	This group of proteins is involved into signaling translation process.	*GTPases*
Storage proteins	These proteins contain energy, which can be released during metabolism processes in the organism.	e*gg ovalbumin*, *milk casein*

3.2 Caused diseases and abnormalities

Diseases caused by the misfolded and damaged proteins can be divided in two groups. Diseases of group number one are characterized by the absence or disappearing of a key protein, as, due to its misfolding, it has been recognized as disfunctional and eliminated by the cell's own machinery (see table 3). Diseases of group number two are characterized, on the opposite, by the presence or deposition of the misfolded proteins in the insoluble aggregates within the cell (see table 4). [6]

Table 3. Diseases of group number one, caused by the absence of a key protein

Name of the disease	**Name of the absent misfolded protein**
Cystic fibrosis	CFTR protein
Marfan syndrome	Fibrillin
Fabry disease	alpha galactosidase
Gaucher's disease	beta glucocerebrosidase
Retinitis pigmentosa 3	Rhodopsin
Some cancers	Different

Table 4. Diseases of group number two, caused by the presence of the misfolded protein

Name of the disease	**Name of the deposited misfolded protein**
Alzheimer's disease	amyloid beta and tau
Type II diabetes	Amylin
Parkinson's disease	alpha synuclein
Creutzfeldt-Jakob disease	prion protein
Congestive heart failure	transthyretin deposited in the heart tissue
Peripheral neuropathy	transthyretin deposited in the nerves tissue

The research team, led by William L. Klein, found, that toxic proteins, called “amyloid ß-derived diffusible ligands” (ADDLs, pronounced “addles”), from the brain tissue of individuals with Alzheimer’s disease specifically attack and disrupt synapses, the nerve cell sites responsible for information processing and memory formation, causing memory loss, loss of balance or tremors [7].

4 Free energy minimization as a driven force of protein folding

4.1 Free energy

Thermodynamically, proteins are described by the free energy, usually, **Gibbs free energy** [23]. It is defined as:

$$G = H - TS,$$

where

- H is the enthalpy (SI unit is Joule [J]),
- S is the entropy (SI unit is Joule per Kelvin [J/K]),
- T is the temperature (SI unit is Kelvin [K]).

4.1.1 Enthalpy and potential energy

The enthalpy is defined as [23]:

$$H = U + p\mathcal{V},$$

where

- U is the internal energy (SI unit is Joule [J]),
- p is pressure (SI unit is Pascal [Pa]),
- $\mathcal{V}$ is volume (SI unit is cubic meter [m^3]).

In turn, the internal energy is

$$U = E_p + E_k,$$

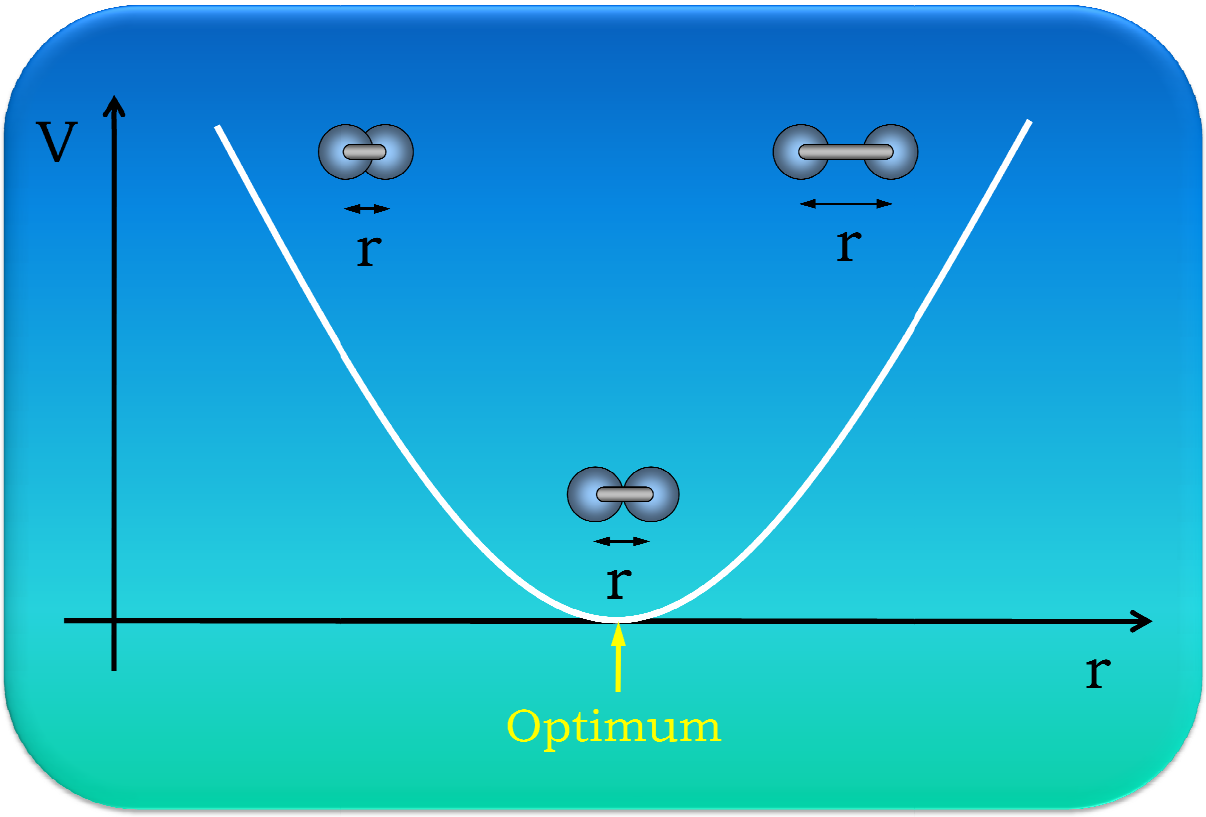
where

- E_p is the potential energy (SI unit is Joule [J]),
- E_k is the kinetic energy (SI unit is Joule [J]).

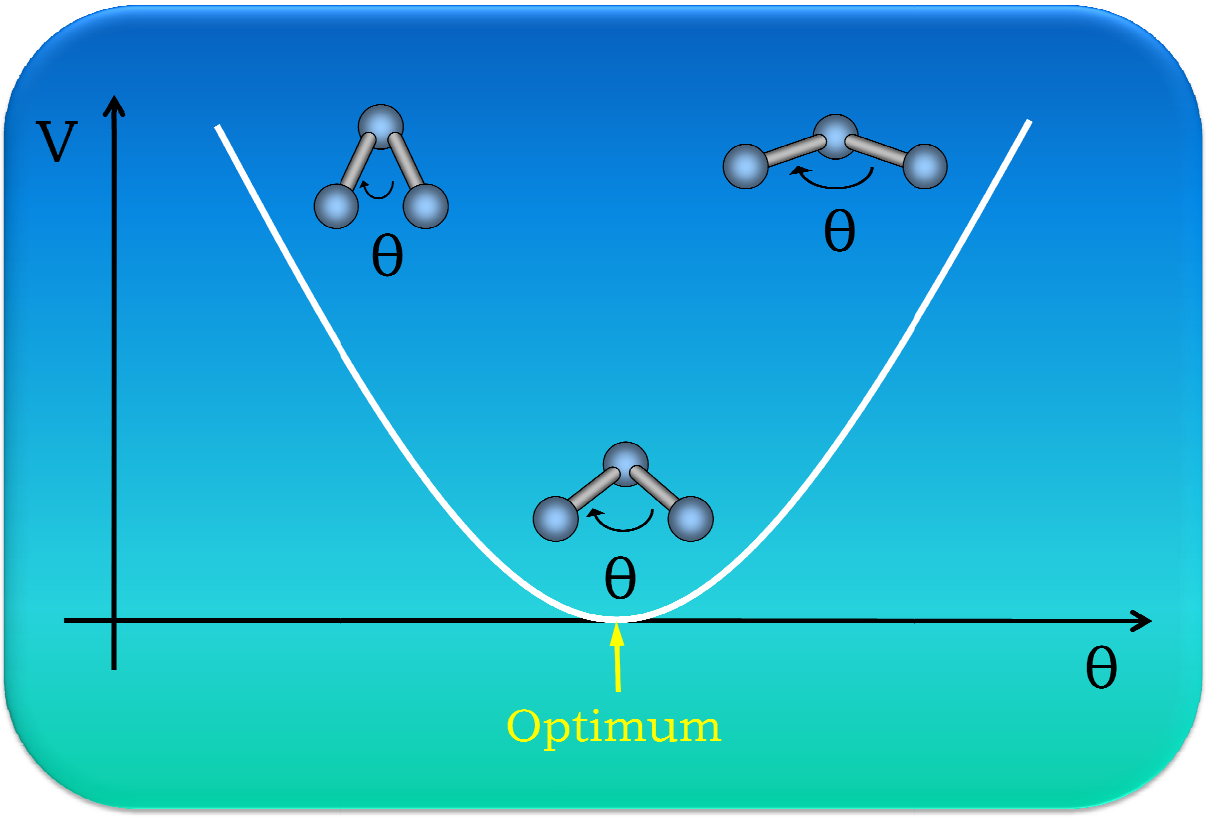
Note! *Further on, potential energy notation is changed from E_k to V.*

For proteins the major contribution to the enthalpy is made by the potential energy [26,28].

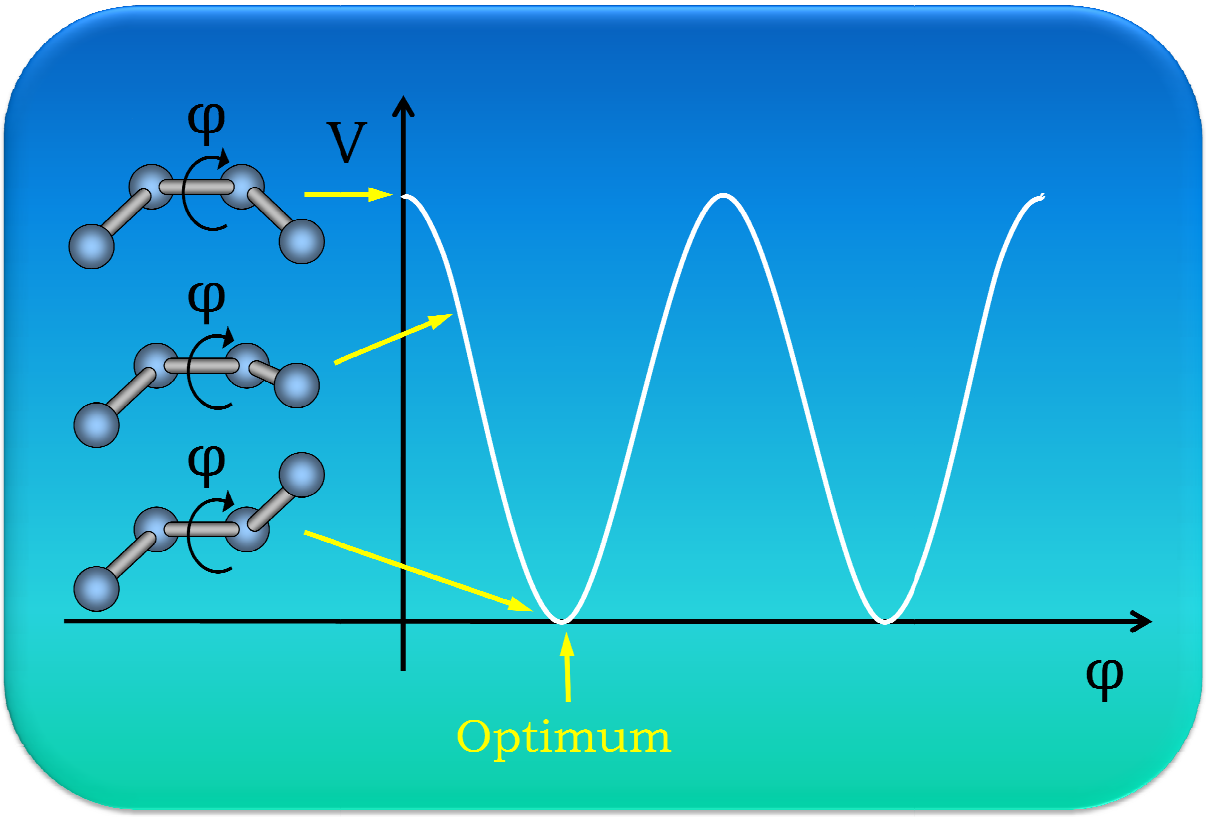
Potential energy function is usually written as a sum of potential energy "subfunctions", each of which describes specific type of the atom interaction. Such interactions are divided into two groups separating interactions between connected and unconnected atoms. Interactions between connected atoms may involve two (bond-length interactions), three (bond-angle interactions), and four atoms (torsional-angle interactions) at the same time, while interactions between unconnected atoms always involve pairs of atoms (Van der Waals and electrostatic interactions). On the figure 2 schematically shown relations between the positions of atoms and the curves of the potential energy "subfunctions", describing different types of the interactions. It is important to note, that there exist much more types of interactions then it is mentioned here and, depending on the number of them included in the potential energy function, researchers distinguish different kinds of the potential energy functions (see [8,9,10]).



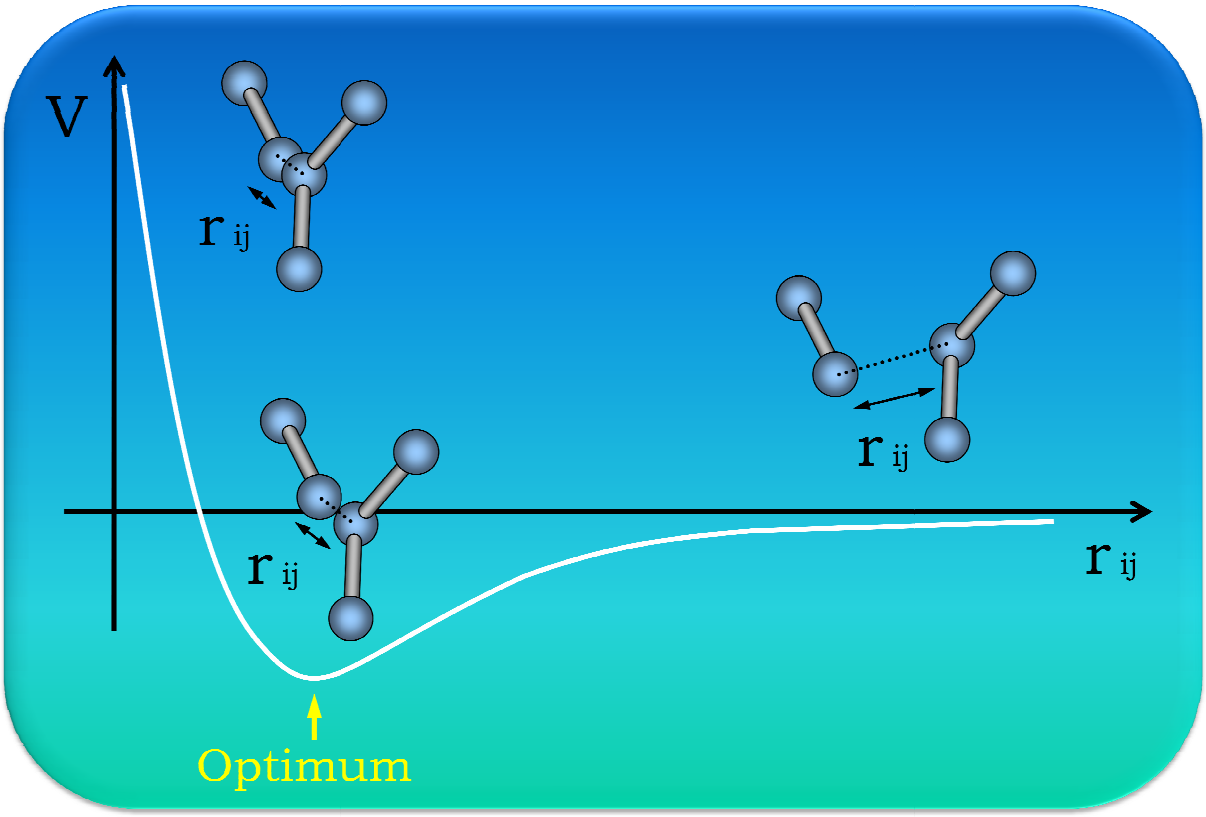
(a)



(b)



(c)



(d)

Figure 2. Schematic pictures of the relations between the positions of atoms and curves of the potential energy "subfunctions", describing *(a)* bond-length interactions, *(b)* bond-angle interactions, *(c)* torsional-angle interactions, and *(d)* non-bonded interactions

4.1.2 Entropy

For proteins, entropy describes two quantities. First quantity is order and disorder, or in other words variety of possible three dimensional structures (or configurations). The more protein is organized, the lower its entropy [24,25,26,27], as one might expect native structure has the lowest entropy. Second quantity is hydrophobic interactions. Protein's polypeptide chain contains hydrophilic and hydrophobic parts. In the folded protein hydrophobic parts are buried in the protein's interior, to minimize their contact with aqueous environment, which results in the lower entropy of the protein [28].

4.2 Protein folding and free energy landscape

As was shown above, during protein folding, both enthalpy and entropy decrease, and hence the overall free energy.

Free energy function constructs free energy surface, commonly known as **free energy landscape**. Each point of this landscape corresponds to the energy of the protein in some structure (or configuration) and belongs to the (*3N+1*)-dimensional space, where *N* is the number of atoms in the protein. High multidimensionality of the landscape, makes it impossible to perform it on paper, however, since experiments show, that there is only one most stable structure of any protein, scientists believe, that by its nature it is similar to a funnel, and not a smooth one (schematically in is shown on Fig.3 and Fig.4a). Global minimum of the energy landscape corresponds to the native structure, while local ones correspond to the meta stable structures, separated from each other by saddles and barriers.

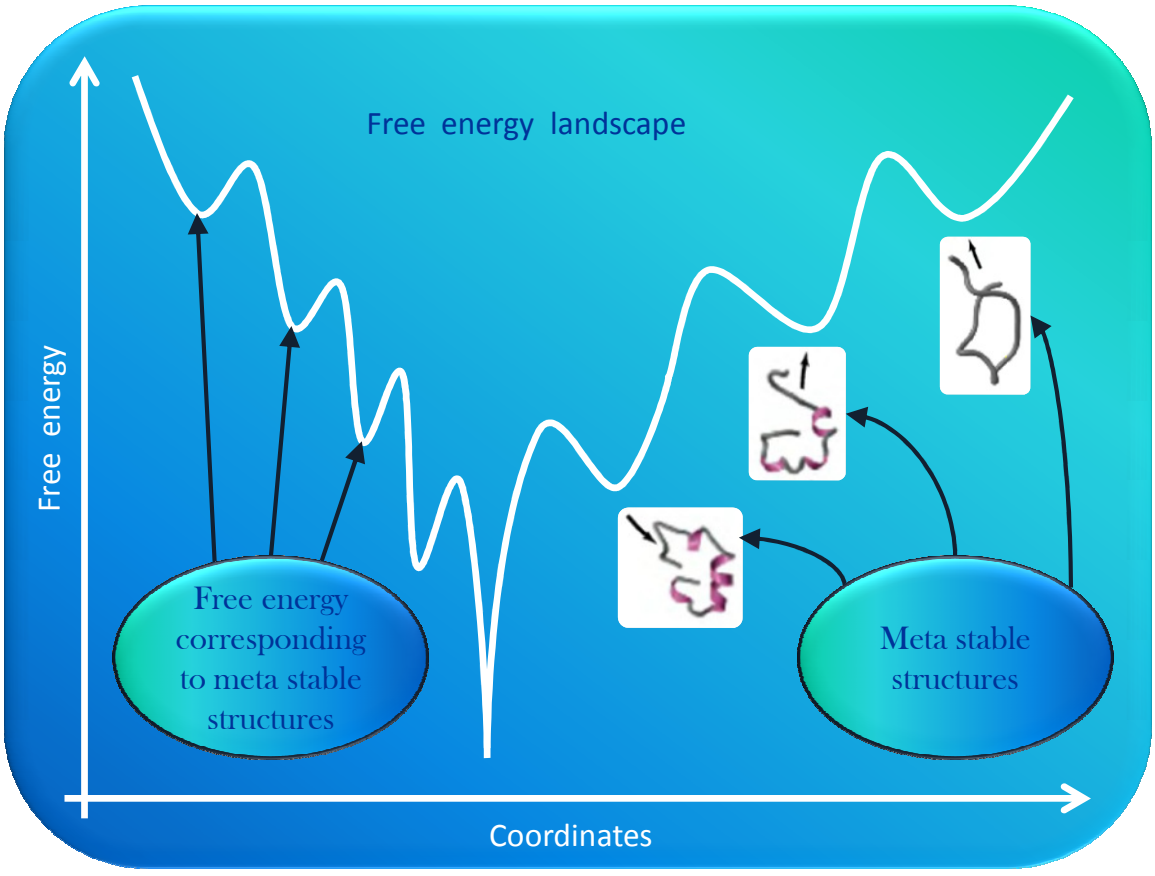
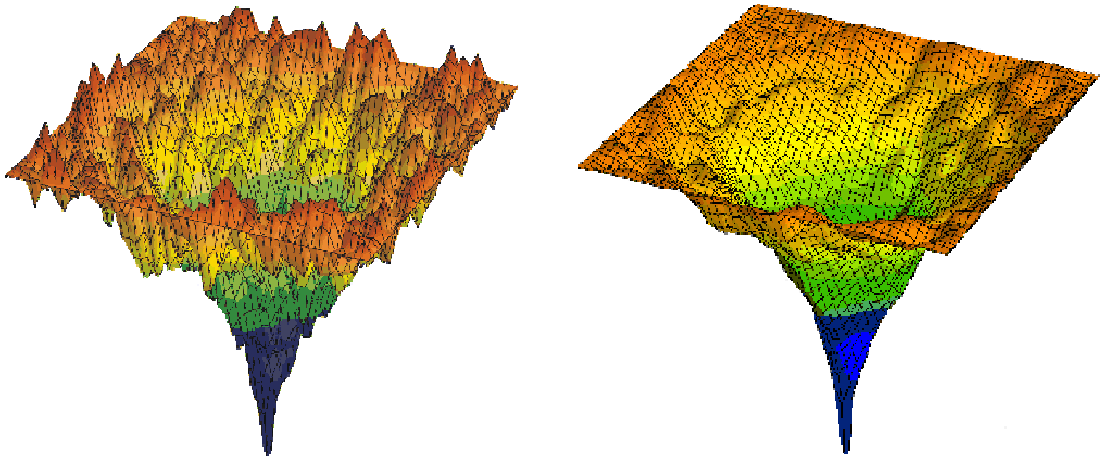


Figure 3. Schematic picture of the protein free energy landscape

Another important characteristic of the energy landscape, which can also be concluded from the experiments and (the above) observations, is that its shape is changed depending on the protein's environment. For example, when a protein is fully hydrated (or simply saying surrounded by water), its energy landscape is seen to be considerably smoothed (Fig.4b) (which was one of the reasons, why the test molecule, presented in the published article [32], was solvated in water). One more characteristic comes from the existence of the misfolded proteins, whose free energy landscape must have at least one more deep "funnel" resulting in a local minimum with such a high barrier, that it cannot be "overcome" for a long time.



(a) *(b)*

Figure 4. Schematic picture of the energy landscape for the non-hydrated *(a)* and hydrated *(b)* protein, i.e. protein without and with water. (taken from [29]) Dark blue color corresponds to the lowest energy and orange one to the highest energy.

5 Molecular dynamics simulations of protein folding

5.1 Ways to study protein folding

Protein folding is observed to occur during $10^{-1} - 10^{3}$ seconds both *in vivo* (Latin for "within the living", i.e. in live isolated cells) and *in vitro* ("within the glass", i.e. in a test tube) [5], which complicates a lot experimental study and collection of data about intermediate structures. However, accumulated knowledge about proteins and recent development of the computer technology resulted in the new techniques of protein studies, namely *in silico* (i.e. via computer simulations). These simulations are done by

the molecular dynamics simulation packages, among which Folding@Home, Rosetta@Home, AMBER, GROMACS, CHARMM, Abalone and others.

5.2 Molecular dynamics simulations

All molecular dynamics simulation packages follow roughly the same major simulation steps. But since data, used in this master thesis work, was obtained with the help of GROMACS, we will briefly describe major steps of this package. More detailed information can be found in [10].

In the most general way molecular dynamics simulations can be described in terms of the black box, shown on figure 5.

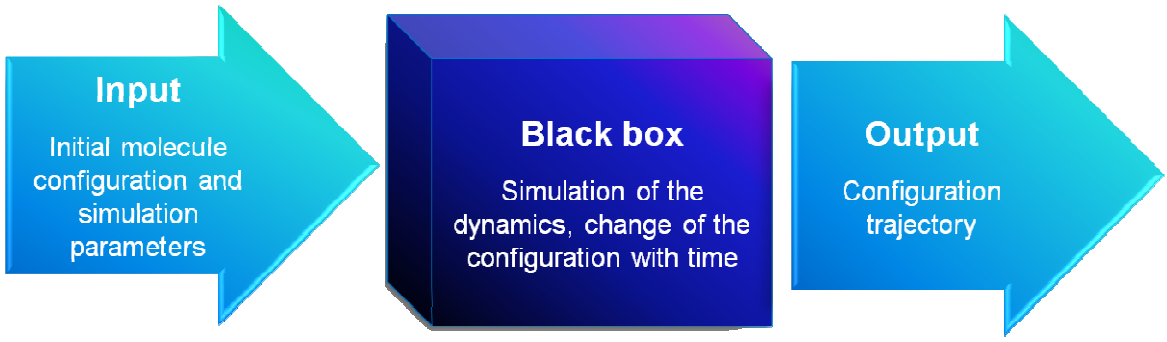


Figure 5. Black box scheme, describing molecular dynamics simulations in the most general way

5.2.1 Input

First step of simulations is a preparation of the input data and specification of all parameters of the system. By input data we understand initial positions and (optional) velocities of all atoms of the protein (or peptide) molecule, given by coordinates q_i and velocities v_i respectively. As for parameters, some of them are mentioned in table 5 (mostly the ones, which are mentioned in the published article [32]), while the full list with descriptions can be found in [10].

Table 5. List of the parameters with a short description.

Parameters	Short description	Parameters used in the simulations performed for the numerical experiment
Type of the potential energy function, referred to as **force field**	Potential energy functions differ from each other by the number of terms or "subfunctions" they contain, which in turn depends on the number of interactions taken into account. While choosing this function one should keep in mind two things: two little terms will result in unrealistic description of the system and too much terms will cause long simulation time.	53a6
Shape of the container	Protein (or peptide) molecule is usually surrounded by a box, filled with water molecules. To minimize simulation time, one can choose the box shape in such a way, that it "repeats" the shape of the molecule. Among the possible shapes are *cubic*, *octahedron*, and *rhombic dodecahedron*.	Cubic
Boundary conditions	It is common to apply periodic boundary conditions to avoid surface artifacts, so that a water molecule that exits to the right, reappears on the left.	Periodic boundary conditions
Size of the container (in nm)	If the boundary conditions are chosen to be periodic, it is important to choose a sufficiently large box, so that the water molecules would not interact significantly with their periodic copies.	3 *x* 3 *x* 3 (nm)

Type of the water molecules	Many water models are developed in a specific force field within a specific package, and then (sometimes correctly) adopted to other force fields. For example, SPC and SPC/E water models were developed for GROMACS, TIP3P for AMBER, and modified TIP3P for CHARMM [11]. To calculate interactions between atoms of the molecule in liquid water, one need a model of the individual water molecule that tells where the charges reside. SPC water molecule has three centers of concentrated charge, which leads to an incorrect value for the permanent dipole moment, which in turn is compensated by the increase of HOH angle to 109.47°, from the observed value of 104.5° [12].	Simple Point Charge (SPC)
Thermostat	System is normally coupled to a thermostat, that scales velocities during the integration to maintain constant temperature and container size to maintain constant pressure.	Berendsen thermostat
Output step	Output step indicates how often one wants to save molecule configuration to the output file, along with other chosen characteristics such as energy, temperature, pressure etc.	Every 0.5 ps or 0.0005 ns

5.2.2 The algorithm

Molecular dynamics simulations is an iterative process. Major steps of the first iteration of this process are presented on the figure 6. Following iterations are the same, but without the input step and possibly without the output step, depending on the specified value of the output step parameter.

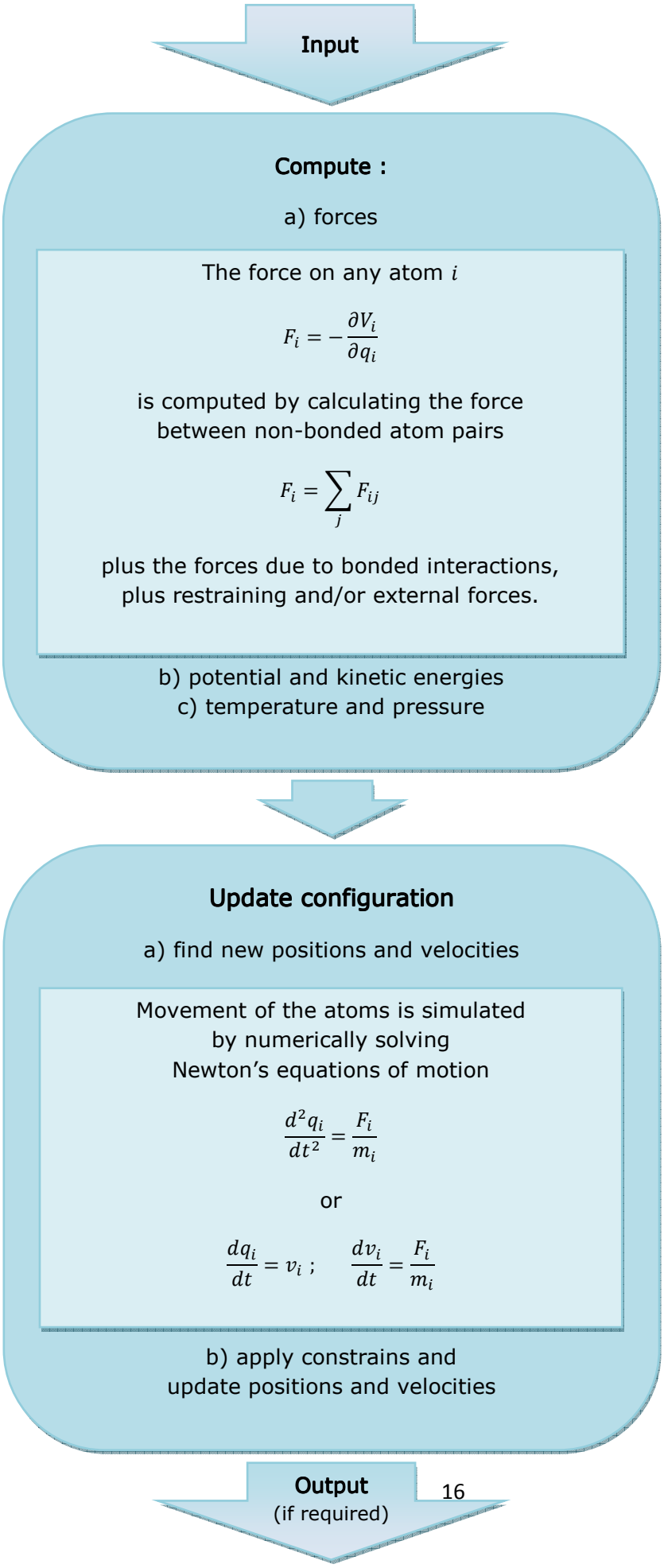


Figure 6. Global scheme of the first iteration of the molecular dynamics simulations [10].

Let's just briefly mention that forces are defined as the negative gradient of the potential energy function $V(q_1,q_2,...,q_N)$:

$$F_i = -\frac{\partial V}{\partial q_i}, \quad i = 1 \dots N,$$

where q represents three Cartesian coordinates x, y, and z, specifying positions of atoms.

The default molecular dynamics integrator in GROMACS is the so-called **leap-frog algorithm** [30] for the integration of the equations of motion. However, the equations of motion are modified for the temperature coupling and pressure coupling, and extended to include the conservation of constraints, so in this case **the velocity Verlet algorithm** [31] (also implemented in GROMACS) is more preferable [10].

5.2.3 Output

The most important output of the simulations is the **trajectory file**, which contains coordinates of all atoms of the molecule and (optionally) velocities recorded every time step specified by the "output step" value. In addition to that one can also have file with other characteristics of the system.

Intuitively, one might expect, that trajectory obtained from such simulations contains protein's "folding pathway", i.e. time-ordered sequence of intermediate structures ordered from the most unfolded to the most folded one. Unfortunately, this is not true, even though simulated dynamics globally does describe protein folding. Trajectory file, indeed, contains sequence of structures, but they are not organized. Schematically, obtained configuration trajectory projected on the energy landscape is shown on the figure 7. From the figure one can see, that it is almost impossible to capture meta stable structure of the protein, corresponding to one local minimum of the energy landscape, instead there is a group of structures "living" in the same valley. Moreover, structures belonging to the same valley appear (let's say for now) "randomly" (or not successively) in the trajectory sequence.

The goal of this Master thesis was to distinguish these groups of structures having the trajectory of the molecular dynamics simulations.

If one considers protein structures as states and is able to prove, that trajectory has Markov property, i.e. that transition to the next structure (or state) does not depend on the history of transitions between previous structures, then one can consider trajectory of structures as a Markov chain, and, applying spectral method, one can find groups of structures belonging to the same valley. In the table 6 term correspondence between the protein theory, Markov chain theory and the published article [32], containing numerical experiment, is presented.

Table 6. Term correspondence.

Protein theory	**Markov chain theory**	**Published article**
Protein structures	States	Microstates
Group of structures belonging to the same valley	Classes (almost closed recurrent classes)	Meta stable states

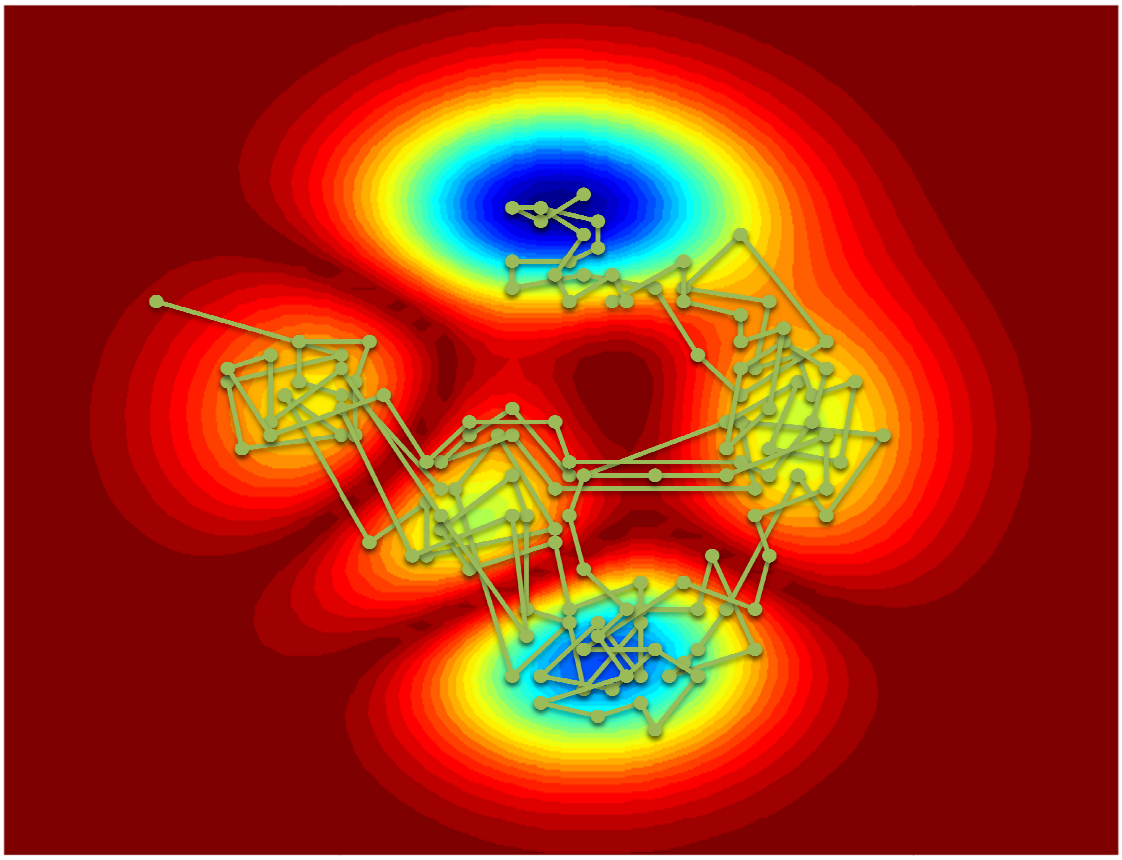


Figure 7. Schematic projection of the
molecular dynamics simulation trajectory
on the contour plot of the free energy landscape

6 Markov chain. Transition probability matrix

Let's begin by introducing some definitions. Let's consider system characterized by a set of "states". As time passes, this system evolves by transitioning from one state to another. If time is discrete and the number of states is countable, the system is called **discrete**. If transitions in such system occur randomly according to some probability distribution, then evolution of the system is called a **discrete stochastic process**. Such process is called **memoryless**, if the probability of the transition from one state to another does not depend on the history of the process (the sequence of previous states). Finally, process having all of these characteristics, namely memoryless discrete stochastic process with a finite number of states, is called a **finite Markov chain**. [13,14]

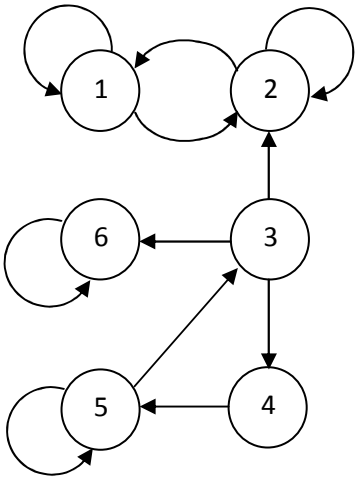
Finite Markov chain with *N* states is characterized by the *N x N* **transition probability matrix** *P*, whose entries p_{ij} describe probability of the transition from state i to state j, $i, j = 1,2, \dots, N$. Transition probability matrix *P* is called **row-stochastic**, if all of its entries are nonnegative real numbers ($p_{ij} > 0, \ \forall i, j$) and the sum of each row is one ($\sum_j p_{ij} = 1, \ \forall i, j$). [13]

Note! *From now on we will consider only row-stochastic transition probability matrices and we will refer to them as simply transition probability matrices.*

Note! *There are several different ways to construct transition probability matrix from the protein molecular dynamics simulations trajectory. One of the techniques is described in details in the published article [32].*

State *j* is **accessible** from state *i*, if $p_{ij} > 0$. States *i* and *j* **communicate**, if they are accessible to each other. Two states are said to be in the same **class**, if they communicate with each other (see Example 1). [19]

Example 1



Let we have a Markov chain represented by the following graph. This Markov chain has 3 classes:

- *Class 1 contains states 1 and 2*
- *Class 2 contains states 3, 4 and 5*
- *Class 3 contains state 6*

A state i is said to be **transient**, if, given that we start in state i, there is a non-zero probability that one will never return to i. [14] If the state is not transient, it is **recurrent**. Thus, a recurrent is the state one keep coming back to (after one or more transitions), while transient is the state one eventually leave forever. Furthermore, it is possible to prove, that all states belonging to the same class are either all transient or all recurrent [20]. Hence, if they all are transient, the class is called transient, and if they all are recurrent, the class is called recurrent (see Example 2). For finite Markov chains all recurrent classes are **closed** (and vice versa) [20], which means that once there was a transition to any state belonging to this class, all further transitions will be only between the states of this class.

The question we will consider now is how to determine whether the class is recurrent or transient. This property depends on the transitions between states, which in turn, are described by the transition probability matrix, so transition probability matrix should contain information on whether the class is recurrent or transient. And it does. However in some cases this information can be obtained solely from the structure of the matrix and in some cases much deeper analysis of the matrix should be performed to get it.

We know, that for the recurrent class any state of this class is "accessible" to any other state of this class, so transition probability between all states of the same class should be more than zero, we also know, that recurrent class is closed and there are no transitions to the states beyond this class, so transition probability to those states should be zero. This leads to a conclusion that, if the sequence of the states is such, that states belonging to the same class are standing next to each other, then recurrent classes are represented by square blocks along the main diagonal in the transition probability matrix (see Example 2).

But what happens if states belonging to the same class are not standing next to each other in the transition probability matrix? Then recurrent classes, if they exist, can be found with the help of the spectral method based on the analysis of eigenvalues and eigenvectors (Chapter 8).

Example 2

$$\begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 3/4 & 1/4 & 0 & 0 & 0 & 0 \\ 0 & 2/5 & 0 & 2/5 & 0 & 1/5 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 2/3 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Let we have transition probability matrix describing Markov chain from the example 1, constructed in such a way, that states belonging to the same class are standing next to each other. Then

- *Class 1 is recurrent and closed*
- *Class 2 is transient*
- *Class 3 is recurrent and closed*

7 Eigenvalues and eigenvectors

Eigenvalues and eigenvectors reflect certain properties of the matrices, which play key role in spectral method, so in this Chapter we will introduce some definitions and properties of eigenvalues and eigenvectors of different kinds of square matrices. In subsections 7.1 and 7.2 we will consider two types of square matrices (positive and block-diagonal respectively) and concentrate on answering four main questions :

1. What are the definitions of eigenvalue and eigenvector?
2. How can one find them?
3. How many eigenvalues does one square matrix has and what properties do they have?
4. How many eigenvectors does one square matrix has and what properties do they have?

In the following subsections of this Chapter we will concentrate on showing how properties of eigenvalues and eigenvectors of different types of transition probability matrices reflect properties of the Markov chains these matrices describe.

7.1 Eigenvalues and eigenvectors of a square matrix

Let's consider a square *n x n* matrix *A*, all elements of which are real and positive. Given a square matrix *A,* the number λ is said to be an **eigenvalue** of *A,* if there exists a nonzero vector *v* satisfying

$$Av=\lambda v.$$

In this case, *v* is called a **right eigenvector** of *A* corresponding to the eigenvalue λ [15].

Note! *Further on right eigenvector will be referred to by simply eigenvector.*

To find eigenvalues, one solves equation

$$\det(A-\lambda I)=0, \quad (1)$$

where *I* is the identity matrix of the same size as *A*. Once this is done, one can find corresponding eigenvectors by solving homogeneous system

$$(P-\lambda I)v=0 \quad (2)$$

for each found eigenvalue.

Now let's see how many eigenvalues one square matrix can have. According to the **Leibniz formula**, **determinant** of any square *n x n* matrix *A* can be found as

$$det(A) = \sum_{\sigma \in S_n} sgn(\sigma) \prod_{i=1}^{n} A_{i,\sigma(i)},$$

where the sum is computed over all permutations σ of the set $\{1,2,\dots,n\}$ and $sgn(\sigma)$ is equal to +1 for even permutations and -1 for odd permutations, depending on the number of switches of numbers performed for obtaining a new permutation sequence (for more details see [18]). This shows, that *det(A- λI)* must be a polynomial in λ of degree *n*, which means that equation (1) has *n* roots and consequently one square matrix *A* has *n* eigenvalues λ_i, i=1,2,...,n. These roots might be all different, all the same, or part of them can be different, while the other part (or parts) is (are) the same. In the case of *k* identical roots, it is said, that eigenvalue has **algebraic multiplicity** *k*. [17]

Some more information about eigenvalues. **Perron–Frobenius theorem**, proved by Oskar Perron (1907) and Georg Frobenius (1912), asserts that a real square matrix with positive entries has a unique largest real eigenvalue and that the corresponding eigenvector has strictly positive components [16]. Other eigenvalues, can be positive, negative or complex.

Now let's discuss eigenvectors. First of all let's point out that we are interested only in linearly independent eigenvectors. If there is no identical roots and all eigenvalues have algebraic multiplicity 1, then there are as many linearly independent eigenvectors as there are eigenvalues. However, if algebraic multiplicity of some eigenvalue is $k>1$, then there exist $m \leq k$ linearly independent eigenvectors corresponding to this eigenvalue, and hence there are $(n-k+m)$ linearly independent eigenvectors in total for the square $n \times n$ matrix. [21]

Proposition 7.1 If a square matrix A has a constant row sum r, then A has the eigenvalue r with the corresponding eigenvector x, whose entries are all 1. [21]

Proposition 7.2 If v is an eigenvector of some square matrix, then Cv, where C is some constant, is also an eigenvector of this matrix. [21]

7.2 Eigenvalues and eigenvectors of a block-diagonal square matrix

Let's consider a square $N \times N$ block-diagonal matrix A, such that

$$A = \begin{pmatrix} A^{(1)} & 0 & \cdots & 0 \\ 0 & A^{(2)} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & A^{(s)} \end{pmatrix},$$

where each block $A^{(i)}$ is a square matrix of size $n^{(i)} \times n^{(i)}$ $(i=1,2,...,s)$, such that

$$\sum_{i=1}^{s} n^{(i)} = N$$

Eigenvalues and eigenvectors of such matrix can be found in two ways both leading to the same results. Since matrix is square, one can use the same technique as the one described for square matrices in the subsection 7.2. Another way of finding eigenvalues and eigenvectors is the following :

- First, one finds eigenvectors and eigenvalues for each block $A^{(i)}$ $(i=1,2,...,s)$ separately;
- then, one adds $(N-n^{(i)})$ zero-elements to the eigenvectors so that

$$v_{j_1}^{(1)} = (\underbrace{v_{j_1,1}^{(1)},\ v_{j_1,2}^{(1)},\ \ldots,\ v_{j_1,n^{(1)}}^{(1)}}_{n^{(1)}},\ \underbrace{0,\ 0,\ \ldots,\ 0}_{n^{(2)}},\ \ldots,\ \underbrace{0,\ 0,\ \ldots,\ 0}_{n^{(s)}})^T,$$

$$v_{j_2}^{(2)} = (\underbrace{0,\ 0,\ \ldots,\ 0}_{n^{(1)}},\ \underbrace{v_{j_2,1}^{(2)},\ v_{j_2,2}^{(2)},\ \ldots,\ v_{j_2,n^{(2)}}^{(2)}}_{n^{(2)}},\ \ldots,\ \underbrace{0,\ 0,\ \ldots,\ 0}_{n^{(s)}})^T,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$v_{j_s}^{(s)} = (\underbrace{0,\ 0,\ \ldots,\ 0}_{n^{(1)}},\ \underbrace{0,\ 0,\ \ldots,\ 0}_{n^{(2)}},\ \ldots,\ \underbrace{v_{j_s,1}^{(s)},\ v_{j_s,2}^{(s)},\ \ldots,\ v_{j_s,n^{(s)}}^{(s)}}_{n^{(s)}})^T,$$

where $j_i = 1,2,\ldots,n^{(i)}$ and *i=1,2,...,s*, i.e. $\sum_{i=1}^{s} j_i = N$.

- after that, one unites all found eigenvalues and eigenvectors
- and finally, one re-arranges eigenvalues (along with the corresponding eigenvectors) in the descending order and re-names them

Example 3

First let's consider the following square block-diagonal matrices :

$$A^{(1)} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 1 \end{pmatrix}, \qquad A^{(2)} = \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}, \qquad A^{(3)} = \begin{pmatrix} 1 & 4 \\ 4 & 1 \end{pmatrix}.$$

Now let's find their eigenvalues and eigenvectors :

$$\lambda_1^{(1)} = 4,\quad v_1^{(1)} = (1 \quad 1 \quad 1)^T;$$

$$\lambda_2^{(1)} = 1,\quad v_2^{(1)} = (1 \quad -2 \quad 1)^T;$$

$$\lambda_3^{(1)} = -1,\quad v_3^{(1)} = (-1 \quad 0 \quad 1)^T;$$

$$\lambda_1^{(2)} = 4,\quad v_1^{(2)} = (1 \quad 1)^T;$$

$$\lambda_2^{(2)} = -2,\quad v_2^{(2)} = (-1 \quad 1)^T;$$

$$\lambda_1^{(3)} = 5,\quad v_1^{(3)} = (1 \quad 1)^T;$$

$$\lambda_2^{(3)} = -3,\quad v_2^{(3)} = (-1 \quad 1)^T;$$

Let's point out, that all considered matrices have constant row sum, so by proposition 7.1 each of them has one eigenvalue equal to this sum and the corresponding eigenvector, whose elements are all equal to 1.

Now, using these 3 matrices, let's construct one square block-diagonal matrix A :

$$A = \begin{pmatrix} 1 & 1 & 2 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 0 & 4 & 1 \end{pmatrix}$$

It has the following eigenvalues and eigenvectors :

a) before re-arranging and re-naming

$\lambda_1^{(1)} = 4, \quad v_1^{(1)} = (1 \ \ 1 \ \ 1 \ \ 0 \ \ 0 \ \ 0 \ \ 0)^T;$

$\lambda_2^{(1)} = 1, \quad v_2^{(1)} = (1 \ \ -2 \ \ 1 \ \ 0 \ \ 0 \ \ 0 \ \ 0)^T;$

$\lambda_3^{(1)} = -1, \quad v_3^{(1)} = (-1 \ \ 0 \ \ 1 \ \ 0 \ \ 0 \ \ 0 \ \ 0)^T;$

$\lambda_1^{(2)} = 4, \quad v_1^{(2)} = (0 \ \ 0 \ \ 0 \ \ 1 \ \ 1 \ \ 0 \ \ 0)^T;$

$\lambda_2^{(2)} = -2, \quad v_2^{(2)} = (0 \ \ 0 \ \ 0 \ \ -1 \ \ 1 \ \ 0 \ \ 0)^T;$

$\lambda_1^{(3)} = 5, \quad v_1^{(3)} = (0 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ 1 \ \ 1)^T;$

$\lambda_2^{(3)} = -3, \quad v_2^{(3)} = (0 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ -1 \ \ 1)^T;$

b) after re-arranging and re-naming

$\lambda_1 = 5, \quad v_1 = (0 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ 1 \ \ 1)^T;$

$\lambda_2 = 4, \quad v_2 = (0 \ \ 0 \ \ 0 \ \ 1 \ \ 1 \ \ 0 \ \ 0)^T;$

$\lambda_3 = 4, \quad v_3 = (1 \ \ 1 \ \ 1 \ \ 0 \ \ 0 \ \ 0 \ \ 0)^T;$

$\lambda_4 = 1, \quad v_4 = (1 \ \ -2 \ \ 1 \ \ 0 \ \ 0 \ \ 0 \ \ 0)^T;$

$\lambda_5 = -1, \quad v_5 = (-1 \ \ 0 \ \ 1 \ \ 0 \ \ 0 \ \ 0 \ \ 0)^T;$

$\lambda_6 = -2, \quad v_6 = (0 \ \ 0 \ \ 0 \ \ -1 \ \ 1 \ \ 0 \ \ 0)^T;$

$\lambda_7 = -3, \quad v_7 = (0 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ -1 \ \ 1)^T;$

Let's point out, that even though eigenvalue $\lambda = 4$ has algebraic multiplicity 2 for the matrix A, it "came" from different blocks, so it has two linearly independent eigenvectors.

7.3 Eigenvalues and eigenvectors of a transition probability matrix describing Markov chain with only one closed recurrent class

Let's consider a transition probability matrix describing Markov chain with only one closed recurrent class, i.e. square matrix with real positive elements, such that each row sums to 1. This matrix has a constant row sum 1, so by the proposition 7.1 the largest (also called dominant or Perron-Frobenius) eigenvalue is always 1 and its algebraic multiplicity is 1. Eigenvector corresponding to this eigenvalue has identical non-zero elements either all equal to 1 by proposition 7.1 or to some constant *C* by proposition 7.2, where *v=x*. (see example 4)

Example 4

Let's consider three transition probability matrices $P^{(1)}$, $P^{(2)}$, and $P^{(3)}$, each describing one recurrent class of some Markov chain, where

$$P^{(1)} = \begin{pmatrix} 1/5 & 2/5 & 2/5 \\ 2/6 & 3/6 & 1/6 \\ 5/8 & 1/8 & 2/8 \end{pmatrix} \qquad P^{(2)} = \begin{pmatrix} 1/3 & 2/3 \\ 1/4 & 3/4 \end{pmatrix} \qquad P^{(3)} = \begin{pmatrix} 3/8 & 5/8 \\ 3/4 & 1/4 \end{pmatrix}$$

Now let's find their eigenvalues and iegenvectors :

$$\lambda_1^{(1)} = 1,$$
$$v_1^{(1)} = (1 \quad 1 \quad 1)^T;$$

$$\lambda_2^{(1)} = -0.31,$$
$$v_2^{(1)} = (-0.93 \quad 0.18 \quad 1)^T;$$

$$\lambda_3^{(1)} = 0.26,$$
$$v_3^{(1)} = (0.21 \quad -0.97 \quad 1)^T;$$

$$\lambda_1^{(2)} = 1,$$
$$v_1^{(2)} = (1 \quad 1)^T;$$

$$\lambda_2^{(2)} = 0.08,$$
$$v_2^{(2)} = (-2.67 \quad 1)^T;$$

$$\lambda_1^{(3)} = 1,$$
$$v_1^{(3)} = (1 \quad 1)^T;$$

$$\lambda_2^{(3)} = -0.38,$$
$$v_2^{(3)} = (-0.83 \quad 1)^T;$$

7.4 Eigenvalues and eigenvectors of a transition probability matrix, describing Markov chain with two or more closed recurrent classes, with a block-diagonal structure

Let's consider a block-diagonal transition probability matrix describing Markov chain with two or more closed recurrent classes, i.e. square block-diagonal matrix with real positive elements inside each block and zero-elements outside blocks, such that each row of the matrix sums to 1 and each block describes one closed recurrent class. Recall, that eigenvalues of this matrix can be found as an ordered union of eigenvalues found for each block separately (subsection 7.2). Also recall, that dominant eigenvalue for each block should be 1 (subsection 7.3). So, algebraic multiplicity of the eigenvalue 1 of the matrix should be the same as the number of blocks or the number of recurrent classes in the Markov chain, which this matrix describes. Each eigenvector corresponding to this eigenvalue has block structure with two sets with identical values, one with zeros and another one either with ones by proposition 6.1, or with some constant *C* by proposition 7.2. Moreover, amount of non-zero identical elements is the same as the size of the block it "arrived" from, or, in other words, as the amount of states belonging to the recurrent class described by this block. (see example 5)

Example 5

Let's consider transition probability matrix describing Markov chain with three recurrent classes, such that states 1, 2, and 3 belong to the first class, states 4 and 5 to the second class, and states 6 and 7 to the third class, and let transitions in these classes are described by the matrices from the example 4:

$$P = \begin{pmatrix} 1/5 & 2/5 & 2/5 & 0 & 0 & 0 & 0 \\ 2/6 & 3/6 & 1/6 & 0 & 0 & 0 & 0 \\ 5/8 & 1/8 & 2/8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 2/3 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 3/4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3/8 & 5/8 \\ 0 & 0 & 0 & 0 & 0 & 3/4 & 1/4 \end{pmatrix}$$

It has the following eigenvalues and eigenvectors (already re-arranged and re-named)

$$\lambda_1 = 1, \qquad v_1 = (0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 1 \;\; 1)^T;$$

$$\lambda_2 = 1, \qquad v_2 = (0 \;\; 0 \;\; 0 \;\; 1 \;\; 1 \;\; 0 \;\; 0)^T;$$

$$\lambda_3 = 1, \qquad v_3 = (1 \;\; 1 \;\; 1 \;\; 0 \;\; 0 \;\; 0 \;\; 0)^T;$$

$$\lambda_4 = 0.26, \qquad v_4 = (0.21 \;\; -0.97 \;\; 1 \;\; 0 \;\; 0 \;\; 0 \;\; 0)^T;$$

$$\lambda_5 = 0.08, \qquad v_5 = (0 \;\; 0 \;\; 0 \;\; -2.67 \;\; 1 \;\; 0 \;\; 0)^T;$$

$$\lambda_6 = -0.31, \qquad v_6 = (-0.93 \;\; 0.18 \;\; 1 \;\; 0 \;\; 0 \;\; 0 \;\; 0)^T;$$

$$\lambda_7 = -0.38, \qquad v_7 = (0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; -0.83 \;\; 1)^T;$$

7.5 Eigenvalues and eigenvectors of a transition probability matrix, describing Markov chain with two or more almost closed recurrent classes, with a block-diagonal structure

First, let's define almost closed recurrent class of the Markov chain. We will call the class **almost closed recurrent**, if any state belonging to this class has much higher probability to transit to another state of the same class, then to the state belonging to another class. Block-diagonal transition probability matrix describing such class is such, that blocks on the main diagonal have positive real elements with values much higher than all others.

First of all, let's point out that, while both, eigenvalues and eigenvectors of such matrix, can be found using the technique described in the subsection 7.1, only eigenvalues can be found using the technique described in the subsection 7.2.

Eigenvalues of such matrix have three main properties:

1. As eigenvalues of the transition probability matrix describing Markov chain with only one closed recurrent class, the largest eigenvalue is 1 and its algebraic multiplicity is 1.
2. As eigenvalues of the block-diagonal transition probability matrix describing Markov chain with two or more closed recurrent classes, eigenvalues can be found separately for each block along the main diagonal and then united, re-arranged in the descending order and re-named.
3. And a unique property is that there are so many eigenvalues close to 1 (or just separated from all others by a gap), that with the eigenvalue 1 their amount becomes equal to the number of blocks along the diagonal, or to the amount of the almost closed recurrent classes in the Markov chain.

Eigenvector properties will be considered in more details in Chapter 8. Here we will just mention that, since there is only one eigenvalue equal to one, there is only one corresponding eigenvector with the identical non-zero elements either all equal to 1 by proposition 7.1 or to some constant *C* by proposition 7.2.

7.6 Eigenvalues and eigenvectors of a transition probability matrix, describing Markov chain with two or more closed recurrent classes, which does not have block-diagonal structure

Let's consider a transition probability matrix describing Markov chain with two or more closed recurrent classes, such that states belonging to the same class does not stand next to each other in the transition probability matrix, i.e. transition probability matrix does not have block-diagonal structure.

Eigenvalues and eigenvectors of such matrix can be found by using technique described in the subsection 7.1.

These eigenvalues and eigenvectors have the same properties as the ones described in the subsection 7.4, namely, the dominant eigenvalue is 1 and its algebraic multiplicity is the same as the amount of closed recurrent classes in the Markov chain; and eigenvectors corresponding to these eigenvalues have two sets of identical elements (zero and non-zero ones). Each eigenvector has the same amount of non-zero elements as the amount of states belonging to one of the classes, but since these states are spread out in the transition probability matrix, non-zero elements are also spread out, moreover, in the same way.

Example 6

Let's consider transition probability matrix describing Markov chain with three recurrent classes, whose states' transition probabilities are taken from the example 5, but the order of the sates comparing to that in the example 5 is changed to {1, 6, 4, 3, 5, 2, 7}:

$$P = \begin{pmatrix} 1/5 & 0 & 0 & 2/5 & 0 & 2/5 & 0 \\ 0 & 3/8 & 0 & 0 & 0 & 0 & 5/8 \\ 0 & 0 & 1/3 & 0 & 2/3 & 0 & 0 \\ 5/8 & 0 & 0 & 2/8 & 0 & 1/8 & 0 \\ 0 & 0 & 1/4 & 0 & 3/4 & 0 & 0 \\ 2/6 & 0 & 0 & 1/6 & 0 & 3/6 & 0 \\ 0 & 3/4 & 0 & 0 & 0 & 0 & 1/4 \end{pmatrix}$$

It has the following eigenvalues and eigenvectors

$$\lambda_1 = 1, \qquad v_1 = (0 \;\; 1 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 1)^T;$$

$$\lambda_2 = 1, \qquad v_2 = (0 \;\; 0 \;\; 1 \;\; 0 \;\; 1 \;\; 0 \;\; 0)^T;$$

$$\lambda_3 = 1, \qquad v_3 = (1 \;\; 0 \;\; 0 \;\; 1 \;\; 0 \;\; 1 \;\; 0)^T;$$

$$\lambda_4 = 0.26, \qquad v_4 = (-0.21 \;\; 0 \;\; 0 \;\; -1.03 \;\; 0 \;\; 1 \;\; 0)^T;$$

$$\lambda_5 = 0.08, \qquad v_5 = (0 \;\; 0 \;\; -2.67 \;\; 0 \;\; 1 \;\; 0 \;\; 0)^T;$$

$$\lambda_6 = -0.31, \qquad v_6 = (-5.26 \;\; 0 \;\; 0 \;\; 5.67 \;\; 0 \;\; 1 \;\; 0)^T;$$

$$\lambda_7 = -0.38, \qquad v_7 = (0 \;\; -0.83 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 1)^T;$$

Note, that eigenvalues of this matrix are the same as in example 5 and the corresponding them eigenvectors have also 2, 2, and 3 elements equal to 1. Moreover, states 6 and 7 belong to the same class and were placed on the positions 2 and 7 respectively in the transition probability matrix, at the same positions first eigenvector have non-zero elements equal to 1. The same observation can be made for the states {4, 5} and {1, 2, 3}.

7.7 Eigenvalues and eigenvectors of a transition probability matrix, describing Markov chain with two or more almost closed recurrent classes, which does not have block-diagonal structure

Let's consider a transition probability matrix describing Markov chain with two or more almost closed recurrent classes, such that it does not have block-diagonal structure.

Eigenvalues and eigenvectors of such matrix can be found by using technique described in the subsection 7.1.

These eigenvalues and eigenvectors have the same properties as the ones described in the subsection 7.5, except for the possibility to find eigenvalues for each block separately, as there is no blocks.

8 Spectral method

Spectral method is used for finding classes in Markov chains by considering describing them transition probability matrices and analyzing their eigenvalues and eigenvectors (Fig.8). Recall, that set of eigenvalues of the matrix is called **spectrum**, so, since the method is based on the analysis of eigenvalues, it got the name spectral.

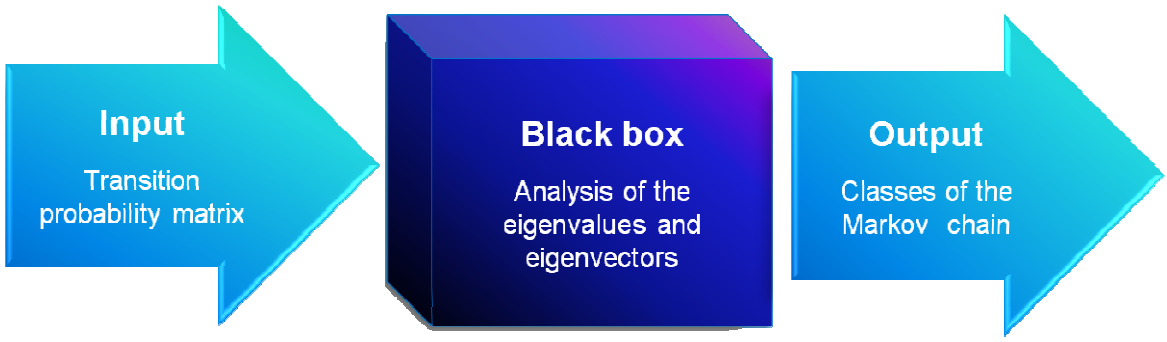


Figure 8. Black box scheme describing Spectral method

(Dominant) eigenvalues are used to answer the question of how many classes is there? While significant eigenvectors are used to answer the question of which state belongs to which class?

First question was discussed in details in Chapter 7, so now we will concentrate on the second one.

Let's start by introducing the definition of the **significant eigenvectors**. In the case, when Markov chain has two or more closed recurrent classes, dominant eigenvalue has algebraic multiplicity more than one, so we will call significant eigenvectors corresponding to this eigenvalue. In the case, when Markov chain has two or more almost closed recurrent classes, dominant eigenvalue has algebraic multiplicity one, so we will call significant eigenvectors corresponding to the eigenvalues closest to 1 (or separated from others by a gap).

The main idea of the method is to introduce a new space for the states of the Markov chain, where classes become more evident. In this space each state is placed according to the coordinates specified by the elements of the significant eigenvectors, so that each eigenvector contains coordinate in one direction. Recall, that number of elements in each eigenvector is the same as number of states in the Markov chain and observe, that dimension of the new space is defined by the amount of the significant eigenvectors.

Once all states are transformed to the new space, one performs clustering using, for example, K-means clustering method and finds which state belongs to which class.

Example 7

Let's consider Markov chain from the example 5 and 6. And let's assume, that we don't know how many classes does it contain and which state belong to which class. Suppose we want to find it out, given the transition probability matrix describing this Markov chain from the example 6.

Recall, that this matrix had the following eigenvalues :

$$\lambda_1 = 1, \quad \lambda_2 = 1, \quad \lambda_3 = 1, \quad \lambda_4 = 0.26,$$

$$\lambda_5 = 0.08, \quad \lambda_6 = -0.31, \quad \lambda_7 = -0.38,$$

from which we conclude first, that there are 3 classes in this Markov chain and second, that there are 3 significant eigenvectors :

$$v_1 = (0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1)^T;$$

$$v_2 = (0 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 0)^T;$$

$$v_3 = (1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0)^T,$$

which in turn means that we have to transform given states to the three dimensional space and assign them coordinates :

- *for the state 1 (0, 0, 1)*
- *for the state 6 (1, 0, 0)*
- *for the state 4 (0, 1, 0)*
- *for the state 3 (0, 0, 1)*
- *for the state 5 (0, 1, 0)*
- *for the state 2 (0, 0, 1)*
- *for the state 7 (1, 0, 0)*

We can see that states 1, 3, and 2 have the same coordinates, so they are transformed to the same point in the new space, and hence they belong to the same class. The same observation can be made for the states 6, 7 and then 4, 5.

Observe, that obtained classes are the same as introduced in the example 5.

Sometimes, to show (or to see) that the classes in the Markov chain are found correctly, it is common to re-arrange given transition probability matrix, so that states belonging to the same class would be next to each other in the matrix, which would reveal its block-diagonal structure.

Example 8

Let's consider Markov chain with 570 states, described by the transition probability matrix presented on the figure 9.

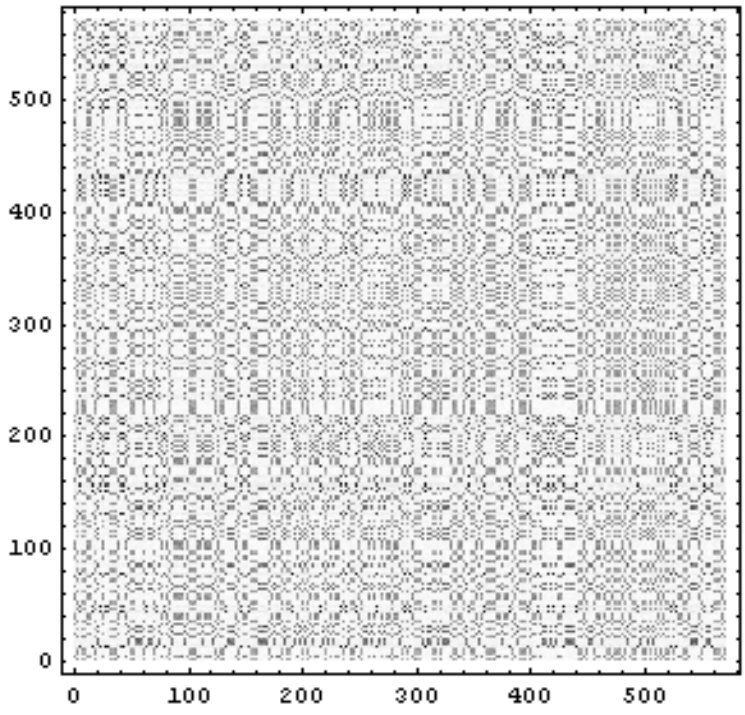


Figure 9. Transition probability matrix

It's eigenvalues presented on the figure 10.

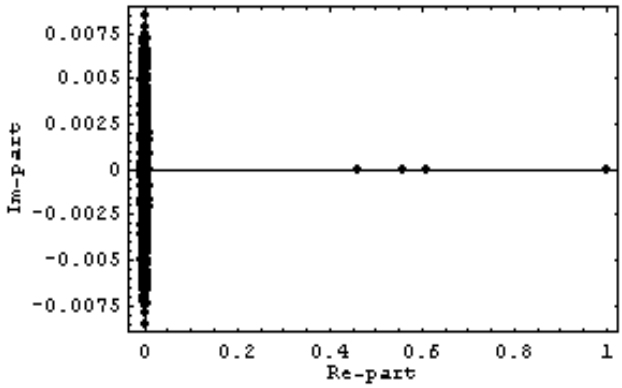


Figure 10. Eigenvalues of the transition probability matrix from the figure 6

There is one dominant eigenvalue equal to 1 and 3 more eigenvalues separated from the others by a gap, from which we conclude first of all, that there are 4 classes in the Markov chain, and second of all that there are 3 significant eigenvectors (second, third, and fourth). So, states are transformed to the three dimensional space and clustered using K-means clustering method. Results are shown on the figure 11, where different colors indicate different classes.

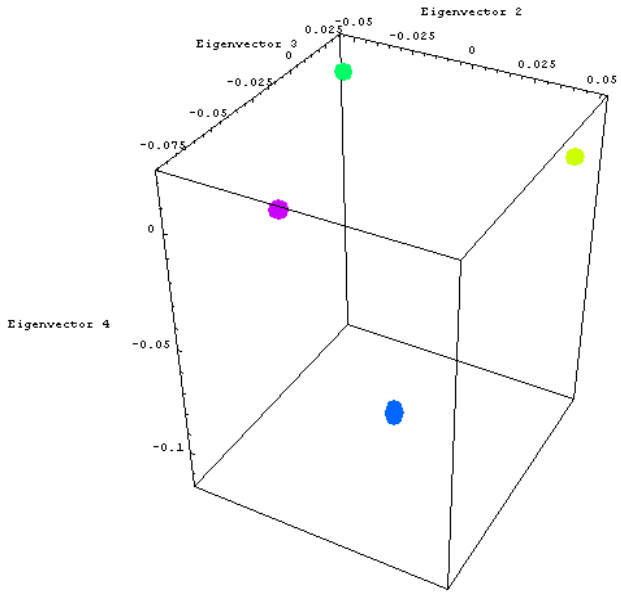


Figure 11. States in the new three dimensional space, where different colors indicate different classes

According to the obtained results, we re-arranged initial transition probability matrix, which revealed its block-diagonal structure (Fig.12), showing that classes were found correctly.

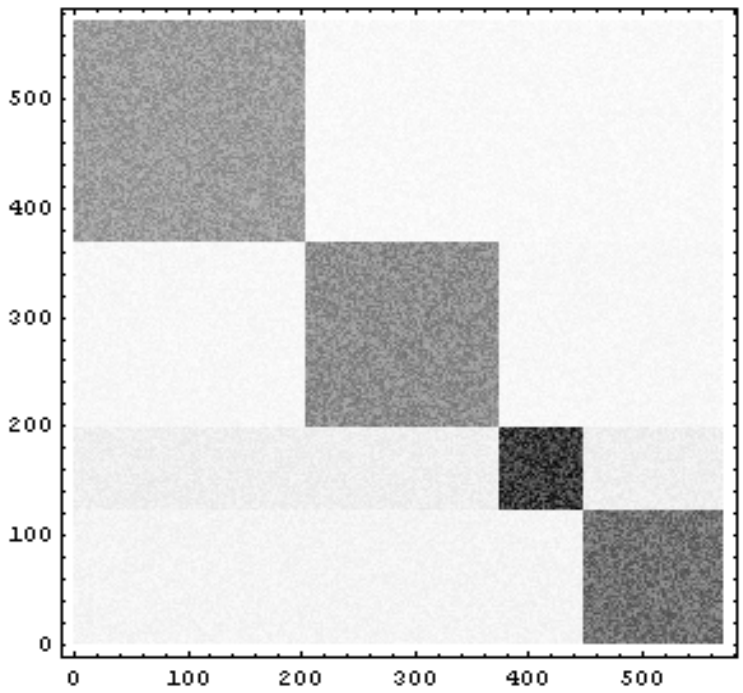


Figure 12. Re-arranged initial transition probability matrix, which now revealed its block-diagonal structure

Conclusions

Studying of intermediate protein structures, so called meta stable structures, appearing during protein folding process, may answer some of the most important questions in protein research, such as how come proteins fold this fast; what causes "correct" and "incorrect" protein folding; is it possible to predict three dimensional protein structure knowing only the amino acid sequence in the polypeptide chain.

This master thesis work presents spectral method, which, using Markov chain theory and interconnected properties of Markov chains, their transition probability matrices, and their eigenvalues and eigenvectors, allows to find meta stable structures of proteins, appearing during their molecular dynamics simulations.

Moreover, during this master thesis work spectral method was applied for finding meta stable states of the VPAL-peptide, giving quite good results, published in the Journal of Chemical Physics (see [32]), confirmed by those known in literature.

References

1. [retrieved: 2011-01-10] Accessible at: http://www.proteincrystallography.org/protein/

2. [retrieved: 2011-01-10] Accessible at: http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/P/Proteins.html

3. R. H. Garrett, C. M. Grisham. Biochemistry 2ed., 851p., 1999.

4. [retrieved: 2011-01-10] Accessible at: http://www-miller.ch.cam.ac.uk/levinthal/levinthal.html

5. C. B. Anfinsen, H. A. Scheraga. Experimental and theoretical aspects of protein folding. *Adv Protein Chem* 29, 205-300, 1975.

5. [retrieved: 2011-01-10] Accessible at: http://www.friedli.com/herbs/phytochem/proteins.html

6. [retrieved: 2011-01-10] Accessible at: http://www.beremans.com/pdf/The_misfolding_diseases_unfold.pdf

7. [retrieved: 2011-01-10] Accessible at: http://www.sciencedaily.com/releases/2004/12/041220004402.htm

8. C. Oostenbrink, A. Villa, A. E. Mark, W. F. Van Gunsteren A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos force-field parameter sets 53a5 and 53a6. *Journal of Computational Chemistry*. Vol. 25, No. 13, 2004.

9. T. Vietshans, D. Klimov, and D. Thirumalai, *Fold. Des.* 2, 1-22, 1997.

10. D. van der Spoel, E. Lindahl, B. Hess, A. R. van Buuren, E. Apol, P. J. Meulenhoff, D. P. Tieleman, A. L. T. M. Sijbers, K. A. Feenstra, R. van Drunen and H. J. C. Berendsen, Gromacs User Manual version 4.5.4, www.gromacs.org (2010)

11. [retrieved: 2011-01-10] Accessible at: http://xray.bmc.uu.se/~calle/md_phd/water_models.pdf

12. [retrieved: 2011-01-10] Accessible at: http://polymer.bu.edu/Wasser/robert/work/node10.html

13. [retrieved: 2011-01-10] Accessible at: http://www.classes.cs.uchicago.edu/archive/2005/fall/27100-1/Markov.pdf

14. [retrieved: 2011-01-10] Accessible at: http://en.wikipedia.org/wiki/Markov_chain

15. [retrieved: 2011-01-10] Accessible at: http://aix1.uottawa.ca/~jkhoury/markov.htm

16. [retrieved: 2011-01-10] Accessible at: http://en.wikipedia.org/wiki/Perron%E2%80%93Frobenius_theorem

17. [retrieved: 2011-01-10] Accessible at: http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-262-discrete-stochastic-processes-spring-2011/lecture-slides/MIT6_262S11_lec08.pdf

18. [retrieved: 2011-01-10] Accessible at: http://en.wikipedia.org/wiki/Determinant

19. [retrieved: 2011-01-10] Accessible at: http://ocw.usu.edu/Electrical_and_Computer_Engineering/Stochastic_Processes/lecture10_4.htm

20. [retrieved: 2011-01-10] Accessible at: http://www.statslab.cam.ac.uk/~james/Markov/s15.pdf

21. [retrieved: 2011-01-10] Accessible at: http://people.sc.fsu.edu/~jburkardt/html/linear_glossary.html

22. A. v David, A. Harville. Matrix algebra from a statistician's perspective (Lemma 21.2.3, p529)

23. [retrieved: 2011-01-10] Accessible at: http://en.wikipedia.org/wiki/Gibbs_free_energy

24. [retrieved: 2011-01-10] Accessible at: http://en.wikipedia.org/wiki/Entropy

25. L. Kavraki. Protein Folding, [*Connexions Web site*], May 9, 2006. [retrieved 2011-01-10] Accessible at: http://cnx.org/content/m11467/1.7/.

26. [retrieved: 2011-01-10] Accessible at: http://www.chaperone.sote.hu/Forces.html

26. [retrieved: 2011-01-10] Accessible at: http://www.chembio.uoguelph.ca/educmat/phy456/456lec02.htm

28. [retrieved: 2011-01-10] Accessible at: http://www.wiley.com/college/pratt/0471393878/student/review/thermodynamics/7_relationship.html

29. [retrieved: 2011-01-10] Accessible at:
http://www.btinternet.com/~martin.chaplin/protein2.html

30. R. W. Hockney, S. P. Goel, J. Eastwood, Quiet High Resolution Computer Models of a Plasma. *J. Comp. Phys.* 14, 148 - 158, 1974.

31. W. C. Swope, H. C. Andersen, P. H. Berens, K. R. Wilson, A computer-simulation method for the calculation of equilibrium-constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* 76, 637 - 649, 1982.

32. S. Ruzhytska, M. N. Jacobi, C. H. Jensen, and D. Nerukh, Identification of metastable states in peptide's dynamics. *J. Chem. Phys.* 133(16), 164102, 2010.