Wiesław Nowak

IV. Modeling of Mechanical and Electrical Properties of Bionanosystems

WN-L1 (2h)

Introduction to nanotechnology of biological systems. Central Dogma of Molecular Biology.

Let us consider the phenomenon of life. Life lasts continuously some 2 billion years. There are millions of various living species. Biologist classify them into kingdoms, orders, families, etc. Constant evolution modifies these organisms. Changes may be positive (better fitting to environment) or negative. Quite often some species (dinosaurs) extinct, but others (cyanobacteria) last for millions years. The very basic fact of life is reproduction – some individuals don't survive but others have a "progeny". Sometimes a period of hibernation is involved before an organism shows all its capabilities (seeds). Sometimes survival of the progeny critically depends on the performance of parents (mammals). How does information ("life plan") goes from one organism ("parents") to another ("kids"). What chemical mechanisms allow for better fitness and adaptation to the local environment? Why some 60 mln years ago large dinosaurs have became extinct (an impact of large meteorite??), but small mammals have survived? Due to progress in science (biology, chemistry, physics, computer science) we may try to answer these fundamental questions.

Knowing well biological mechanisms and structures we can mimic certain solutions "invented' by the Nature through the millions of years of the evolution. We may try to transfer these solutions into technology – even on nanoscale. This was the motivation for the newly developed of science: bionanotechnology [Ehud Gazit, Plenty of room for biology at the bottom: An introduction to bionanotechnology. Imperial College Press, 2007, a free chapter: http://www.worldscientific.com/doi/suppl/10.1142/p465/suppl_file/p465_chap01.pdf].

People put a lot of effort in "improving the Nature". And here it comes: computer modeling of bionanosystems.

Living organisms are (usually) composed of tissues, and tissues are made of cells. Main metabolic processes, such as intake of energy-rich compounds or structural compounds, building new cell organelles, generation of energy, etc., are localized in the cell. Typical cells are small, the diameter is of the order of one micrometer. They are made of biopolymers, and the interior contains a dens solution of membranes, proteins, signaling or energetic compounds. Structural or enzymatic proteins (and perhaps phospolipides present in membranes) play a major role in cell metabolism. Proteins are synthesized in specially designed regions of cells – in ribosomes. Ribosomes are complex "nanofactories". In each human being there are estimated 1,000,000,000,000,000,000 ribosomes. The 3D structure of the ribosome – extremely complicated macromolecular complex – has been determined at the end of XXth century and for this discovery a Nobel Prize has been awarded (see photo below).

The Nobel Prize in Chemistry 2009
Venkatraman Ramakrishnan, Thomas
A. Steitz, Ada E. Yonath

Torun, 10/09/2012

Please enjoy viewing the structure of a ribosome here:
http://www.rcsb.org/pdb/101/motm.do?momID=121
or here:
http://www.dnatube.com/video/2425/Translation-Atomic-View).

How this biochemical machine works? How it comes that from such a simple components – amino acids – so complex proteins are made (hemoglobin – governs oxygen transport, actin - muscles,…)? One may answer these questions using results of computer simulations of proteins dynamics. The largest supercomputers are used for the modeling , since it is very CPU intensive job. Movies presenting the action of the ribosome are shown here:
http://www.dnatube.com/video/1228/tRNARibosome-Molecular-Dynamics-Simulation,

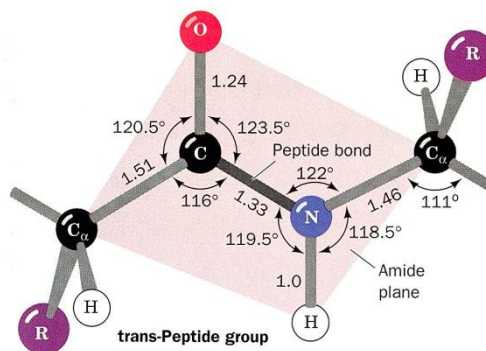Proteins are bioplymers built from amino acids. Amino acids are linked via peptide bond:

Fig. 4.1. The peptide Bond (source: D. Voet, J. Voet, Biochemistry, 2n Ed, 1995)

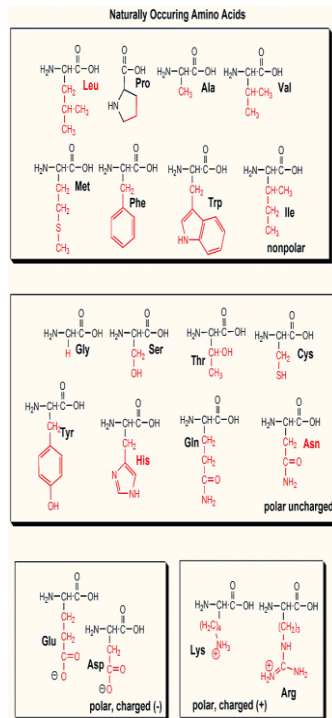There are 20 basic amino acids (AA) present in natural proteins:

Fig. 4.2. Structures of basic amino acids (Internet,
http://employees.csbsju.edu/hjakubowski/classes/ch112/proteins/aminoacidprot1.gif).

Practically all proteins are made of these 20 (or 22) building block. Of course, chemists know much more various amino acids (thousands), but only such a small fractions is used in construction of the living matter.

The order of connection of AA into a protein chain is critically important. The particular arrangement affects physical, chemical and eventually biological properties of a protein. This is possible, since the amino acids are different: they have different polarities, charges, hydrophobicity. Nanobiotechnologists may use such diverse building blocks to create new systems. The sequence (a primary structure) is important, but the fold of a protein finally determine its properties. We know nearly 1000 different folds – specific arrangements of a protein's polymer into a 3D functional structure. Should a protein fold spontaneously into a wrong shape during the synthesis – it will be tagged by special enzymes and recycled in the proteosome. Please note, that not all proteins necessary for life have defined 3D fold. There are also intrinsically disordered proteins. In recent years this became a field of vigorous research.

Proteins may be, *inter alia*, enzymes and/or structural proteins. Wool, silk or spider fibers are made (mainly) from proteins.

What is a flow of information into a ribosome showing what protein has to be made? In what order AA should be assembled? The factory receives a precise recipe from DNA via RNA. Cells contain condensed DNA in their nuclei. This fantastic molecule is also classified as a biopolymer. It is made from nucleotides: there are only four types of nucleotides (letters) used.

The nucleotide is composed of DNA (or RNA) base) a five ring sugar (deoxyribose or ribose) and a phosphate core. The chemical bases involved are called purines (two rings) and pirimidynes (one ring). The 3D structure of DNA has been discovered in 50. by J.Watson and F.Crick (a Nobel Prize). That time they built rather metal then computer models of DNA. A typical D-DNA has alpha-helical structure (twisted stairs):
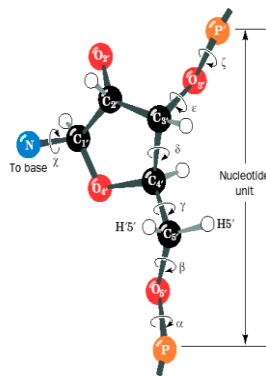
FIGURE 29-7   The conformation of a nucleotide unit is determined by the seven indicated torsion angles.

Fig. 4.3. A scheme of nucleotide (źródło: D.Voet, J. Voet, Biochemistry, 2n Ed, 1995).

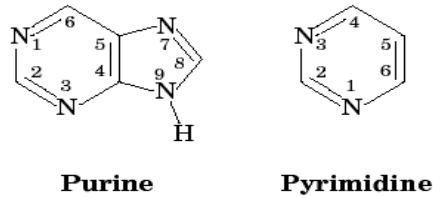**Purine**   **Pyrimidine**

Fig. 4.4. A scheme of puryne and pirymidyne rings. These structures code information in DNA and RNA.
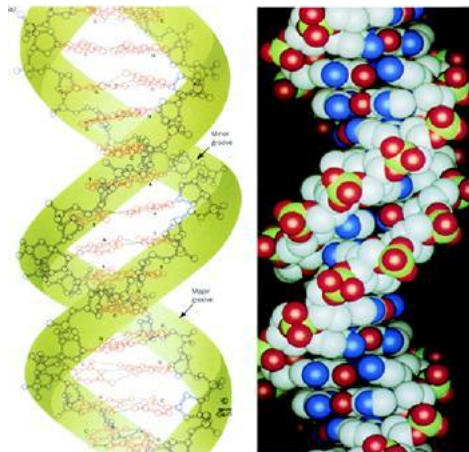

Fig. 4.5. A structure of B_DNA (D.Voet, J. Voet, Biochemistry, 2n Ed, 1995).

There is only one copy of nuclear DNA in the cell. It is a very long molecule – 1.5 meter, 3, 000, 000, 000 pairs of bases are there. It is spitted into 23 chromosomes. The diameter of DNA jest very small,

mere 2-3 nm. There are two complementary chains of DNA polymers. The standard pairing of DNA bases (via 2 or 3 hydrogen bonds) is A-T, C-G.

Certain fragments of DNA contain information how each protein should be synthesize. These fragments are not contiguous, data for one protein (a gene) are usually scattered over a large span of DNA chain. The coding regions are called exons, non-coding ones: introns.
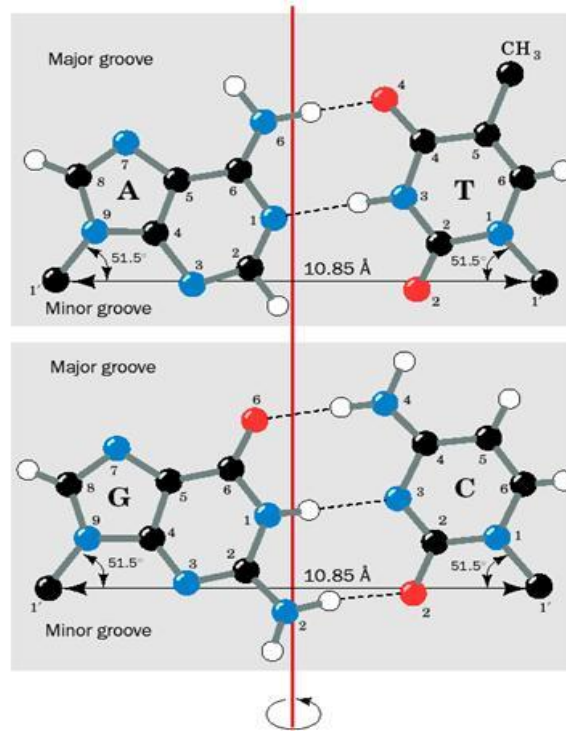

Fig. 4.6. Standard Watson_Crick (WC) base pairs (bp) (source: D.Voet, J. Voet, Biochemistry, 2n Ed, 1995).

The median length of an exon is 12 bp, on average a gene contains 8 exons (up to 350), the average length of a coding sequence is 447 AA. Thus in 3 bln letters of human DNA there are more non-coding regions than coding ones. In 2000 the human genome has been red out, it came out theta the number of human genes is smaller than initially expected: 20 687 (as of year 2012).

When a new protein has to be synthesized, for example for formation of the progenitor cell, a fragment of the DNA polymer winds up. This process of getting access to a tightly packed polymer is catalyzed by appropriate enzymes. Then on a single, locally „isolated" DNA strand (for example, 3000 bp) a synthesis of a new, complementary RNA chain occurs. His short RNA chain is called messenger RNA (mRNA). RNA is chemically quite similar to DNA, instead of T we have uracil (U) here, and deoxyribose is replaced by a similar sugar (ribose as well) – ribose. The mRNA leaves nucleus and goes to ribosome. There information encoded in the mRNA sequence (the same or modified by a process of splicing) is red (3 letters at once) like a magnetic tape. It has been found that the universal genetic code is used in all living organisms (humans, animals, plants, ..). A given sequence of 3 RNA letters denotes an amino acid. The code is shown in the table. Note, that the code is redundant (there are 64 combinations of 3 element string composed of 4letters, but only 20 AA), this helps to protect critical protein regions against spontaneous (or chemically induced – cigarettes smoke!) mutations which destroy precious genetic information.

Fig. 4.7. The genetic code (Internet)

Redundancy helps to protect DNA. The ribosome check 3 letters (i.e. AUU) and allows for creation of the new peptide bond if and only if t-RNA brings into the synthesis place the corresponding amino acids coded by this sequence (Ile). All other possibilities (wrong amino acids statistically delivered to the site) are rejected. This process is very fast, it is controlled by complex enzymes.
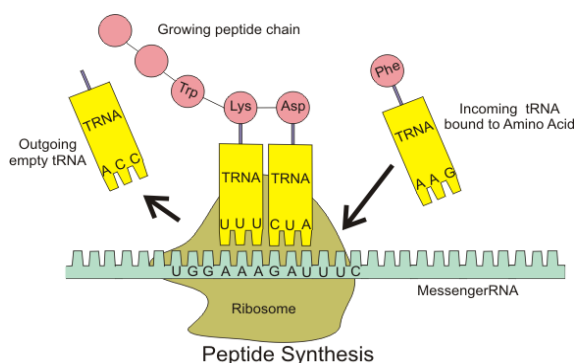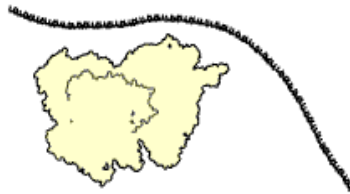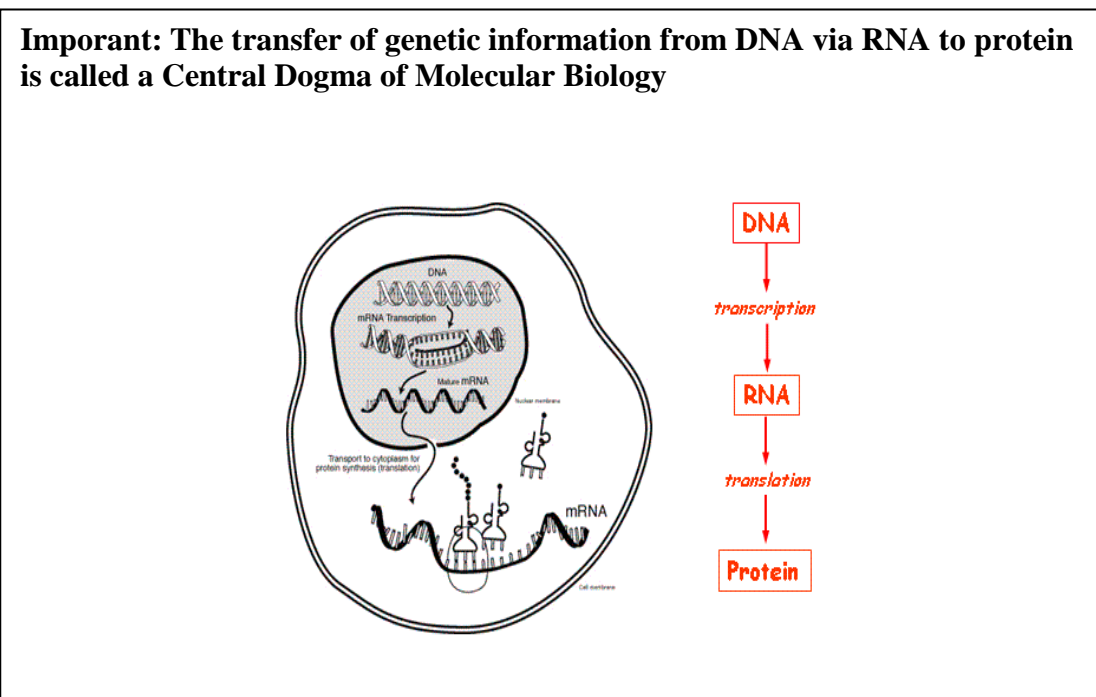

Fig. 4.8. Synteza białka w nanomaszynie jaką jest rybosom (Internet) .

It is recommended to watch the following computer animation of biotechnological process and a nano-machine in action (it is not a MD trajectory!)

Fig. 4.9. An animation showing synthesis of a protein. It is not MD trajectory but just computer graphics animation!

Imporant: The transfer of genetic information from DNA via RNA to protein is called a Central Dogma of Molecular Biology

It is worth to know that after synthesis a protein chain may be further post-translationally modified, for example, serines may get phosphorylated.

The Central Dogma of Molecular Biology was the topic of our first lecture. Now we understand that changes in genes (DNA sequence) may lead to production of altered protein products, which in turn can result in disturbance of cell metabolism, often manifesting as a disease. Although lethal changes result in death of the cell concerned, there are also mutations that render the cell immortal, ie. able to multiply indefinitely. This mechanism is involved in neoplasm formation (oncogenesis). Thus, there are questions that remain unanswered - how can we prevent mutagenesis, what means should we

use to stop uncontrollable growth of cells? How can we protect organisms from „falsification" of genetic information by environmental factors, such as ultraviolet radiation (eg. sunbathing, which a risk factor for melanoma) or tobacco smoke (lung cancer)?

Health is a sought-after good, which explains why rich societies spend considerable sums on research into molecular basis of diseases and new treatments. Because biological molecules (such as proteins and DNA) are so small that nanometers are used to describe their size, the new discipline of practical research concerning their properties is called bionanotechnology. It employs not only physical methods, such as X-ray diffraction, NMR or AFM (atomic force microscope), but also computer simulations.

It is worth noticing that modern methods of molecular biology enable us to make a range of organisms (bacteria, yeast) produce virtually any protein. We are also able to use enzymes to edit the DNA sequences as we like. Automated machines let us describe the composition of biopolymers and bioinformatics analysis helps us in optimizing their physicochemical properties. Within bionanotechnological framework scientists try to merge biopolymers with other materials or objects, such as quantum dots, in hope that this will lead to new advances, eg. materials efficiently transforming light into electrical current. In the following lectures methods of computing analysis of structural, electric and mechanical properties of such nanomaterials will be presented.

Methods of modeling – a formal basis.

Introduction

The whole idea of computer simulations is rather simple: We construct a mathematical model of a biopolymer (protein, DNA, phospholipids, a hormone, a drug, etc.). The model has to be based on some empirical data: dimensions of the molecule, masses of atoms, charges, presence of the environment having some counter-ions, temperature, pressure and a definite pH. One of the main goals of modeling is a search for optimal structures of these biopolymers. Quite often we are interested in dynamics: how fast and how often these biomolecule conformation change one into another. It is possible to monitor how a protein interacts with the other one (proteomics). Models are useful in testing effects of a substitution of one amino acid for another in an enzyme: what happens to the electric field in the catalytic clef – is it stronger, weaker or the same after mutation? One can dock to such a model small chemicals (drugs) and can check what is the interaction energy. It is believed that strong binders are good candidates for inhibitors (new drugs). One may even tailor molecules having desired biochemical properties, thus having predefined physiological effects (rational drug design). A number o research group tests putative mechanisms of enzymatic reactions, for example, those involved in biotechnological production of beer or widely used chemical acrylamide.

Computer modeling of biopolymers is recognized research method. It allows us to obtain relatively quickly new, unique information in a rather inexpensive way. Computer modeling has already been used in drug design processes, several compounds (such as anti-virial drugs) are marketed. It has also been useful in design of innovative diagnostics tests and in green chemistry (biotechnology). Using computer modeling we gather new information on molecular biology of cell, breeding processes, evolution, metabolism, photosynthesis, some pathological process. There is a continuous quest for new materials, for example hybrid materials. To this end biomimetic approach exploits "experience" gathered through millions of years of evolution.

The modeling requires solid physical background, a lot of computer science skills (informatics) and a good deal of math. Commercial software suites may cost hundred thousand dollars. Software houses may employ up to 200 persons with PhD degrees to write such specialized software. Computer specialists are also hired in this high-tech field.

Currently some companies dominate computer modeling commercial market, but there is a lot of academic software, or freeware, which may do a very good job for us. We will present fundamentals of some computational methods and will learn (during tutorials) how to use the **ArgusLab** code (semi-empirical quantum chemistry) and the NAMD+VMD (classical molecular dynamics) suite.

A map of molecular modeling methods - quantum vs classical approach

In computer modeling of chemical systems (molecules or their complexes) and/or biopolymers we are interested mainly in two questions:
  (a) What is their molecular structure?
  (b) What interactions with the environment are involved?

We answer these questions using calculations, not measurements (these are difficult to perform and expensive). Quantum mechanics is the physical theory which should be applied here. However, biomolecules are too large to be studied using a full quantum mechanical approach. They simply contain too many electrons. Thus, we adopt more approximate and simplified description and models. An atom is the smallest piece of matter analyzed separately in the classical approach. Sometimes (coarse grained models) it is even a group of atoms ("a bead"). In the technical and scientific literature there are numerous acronyms referring to the level of approximation used in the studies. In molecular quantum mechanics (quantum chemistry) a crude map of the methods is as follows:

Methods of molecular quantum mechanics:
- *„ab initio"* – based on the variational principle [the Hatree-Fock method (HF), the configuration interaction method (CI),] or the perturbation theory [Many Body Perturbation theory, (MBPT), the Coupled Cluster method (CC, CC SD), MP2]
- DFT – Density Functional Theory (DFT B3LYP, LSDA, TD DFT , etc.)
- Semi-empirical methods (the Huckel method, the PPP method, the Tight Binding method, the Austin Model 1 (AM1) method, Parametric Method 3 (PM3, PM6), ZINDO/S method).

These methods have different levels of approximations adopted, however, each has its own advantages and disadvantages. The positive aspects of semi-empirical methods that we plan to use during tutorials (ArgusLab code and AM1, PM3, MNDO) is their efficiency, speed, good quality of optimized geometries, applicability to very large systems, up to hundreds of atoms.

The following packages for quantum chemical calculations are perhaps the most popular nowadays: Gaussian09, Gamess, NWCHEM, VASP, Jaguar, Hyperchem. The license fees (one year) for some of them are higher than a thousand of US dollars.

A 3D structure of a biomolecule, important for assessment of its biological or technological role, corresponds to the minimum of potential energy surface (PES). We will introduce this concept formally, since this a basis of the whole bionanotechnological modeling.

Potential Energy Surface (PES)

Quantum mechanics predicts that the whole information about a molecule is present in its wavefunction in Dirac's notation $|\Phi\rangle$. The wavefunction one can obtain solving the Schroedinger equation:

$$\hat{H}|\Phi\rangle = \varepsilon|\Phi\rangle$$

The operator of energy, Hamiltonian $\widehat{H}$, depends on both electronic (*i,j*) and nuclear degrees of freedom (A,B), in atomic units we have:

$$\hat{H} = -\sum_{i=1}^{N}\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2$$

$$-\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{i>j}^{M}\frac{1}{r_{ij}} +$$

$$+\sum_{A}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B}{R_{AB}}$$

Atomic units (a.u.):

$$\hbar = 1$$

$$"c = 1" \quad (137.)$$

$$m = 1$$

Usually so called Born-Oppenheimer approximation is adopted. W assume that the atomic nuclei are infinitively heavy, thus they don't move. In that way a separation of electronic and nuclear degrees of freedom is possible:

$$\hat{H}_{el}\Phi_{el} = \varepsilon_{el}\Phi_{el}$$

Here the electronic Hamiltonian reads:

$$\hat{H}_{el} = -\sum_{i=1}\frac{1}{2}\nabla_i^2 - \sum_{i=1}^{N}\sum_{j=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{i>j}^{M}\frac{1}{r_{ij}}$$

and the electron wavefuction Φ_{el} depends on positions of atoms { $\vec{R}_A$ } in a parametric way (semicolon denotes such a dependance):

$$\Phi_{el} = \Phi_{el}\left(\{\vec{r}_i\};\{\vec{R}_A\}\right)$$

$$\varepsilon_{el} = \varepsilon_{el}\left(\{\vec{R}_A\}\right)$$

We see that the electronic energy of the biomolecule studied ε_{el} depends on positions of atoms. For each fixed individual position of the nuclei we have distinct wavefunction Φ_{el} (it depends on electronic degrees of freedom $\vec{r}_i$). Once the interatomic distances R_{AB} are set, the total enenrgy of the biomolecule may be obtained:

$$\varepsilon_{tot}^{fix N} = \varepsilon_{el} + \sum_{A}^{N}\sum_{B>A}^{M}\frac{Z_A Z_B}{R_{AB}}$$

Nuclear Hamiltonian $\hat{H}_{nucl}$, on the other hand, reads:

$$\hat{H}_{nucl} = -\sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2 + \left\langle\hat{H}_{nucl}\right\rangle_{\Phi_{el}} + \sum_{A}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B}{R_{AB}} =$$

$$-\sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2 + \underbrace{\varepsilon_{el}\{R_A\} + \sum_{A}^{N}\sum_{B}^{M}\frac{Z_A Z_B}{R_{AB}}}_{} =$$

$$-\sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2 + \varepsilon_{tot}^{fix N}\{R_A\}.$$

In this expression we see a term $\varepsilon_{tot}^{fix\,N}$ that is a potential in which atomic nuclei move. This term is called potential energy surface (PES). In classical molecular dynamics the quantum character of PES is negleced and this surface is approximated by analytical potentials (force fields). In numerous laboratories force fields have been developed. They are used to model different nanosystems, for example, proteins or crystals. Of course each type of nanosystem requires a dedicated, special force field.

Let us notice that solving (using QM) the nuclear Schroedinger equation:

$$\underbrace{\hat{H}_{nucl}\Phi_{nucl} = \varepsilon_{nucl}\Phi_{nucl}}$$

one can obtain information on molecular oscillations, rotations, and translations. We can obtain the same data, not refering to the QM, but just to mere calssical mechanics, provided that a realistic formula for PES is known. Such surfaces (force fields) are subject of both commercial and academic developement based on experiments and quantum chemical calculations.

Force Fields (FF)

Applications of force fields in nanosystems modeling are based on an underlying assumption that these parametrized analytical formulas are good approximations to real PES. There is yet another important assumption, that the parameters of FF describing a group of atoms (amino acid) in one biomolecule may be transferred without modifications to another one**.** The general formula for FF is the following (note that there are even more complex formulas):

$$V = \sum_{bonds\ i} V_i^B + \sum_{bond\ angles\ j} V_j^A + \sum_{torsional\ angles\ k} V_k^D + \sum_{improper\ angles\ l} V_l^E$$

$$+ \sum_{atom\ pairs\ (r,s)} (V_{r,s}^C + V_{r,s}^P + V_{r,s}^{VdW})$$

The physical meaning of individual energy terms are described in the figure caption.

The main terms correspond to energies of chemical bonds (V_i^B), energies of bond angles deformations, (V_j^A, V_k^D, V_l^E) and two-body interactions between all pairs of atoms present in the system. The nonbonding interactions contain Coulomb energy terms (electrostatics, partial atomic charges) (V^C) and Van der Waals (V^P+V^{VdW}) interactions.

Let us repeat, that in computer modeling not only "all atom" force fields are used, but crude coarse-grained models (a group of atoms is a centre of potential) show growing popularity.
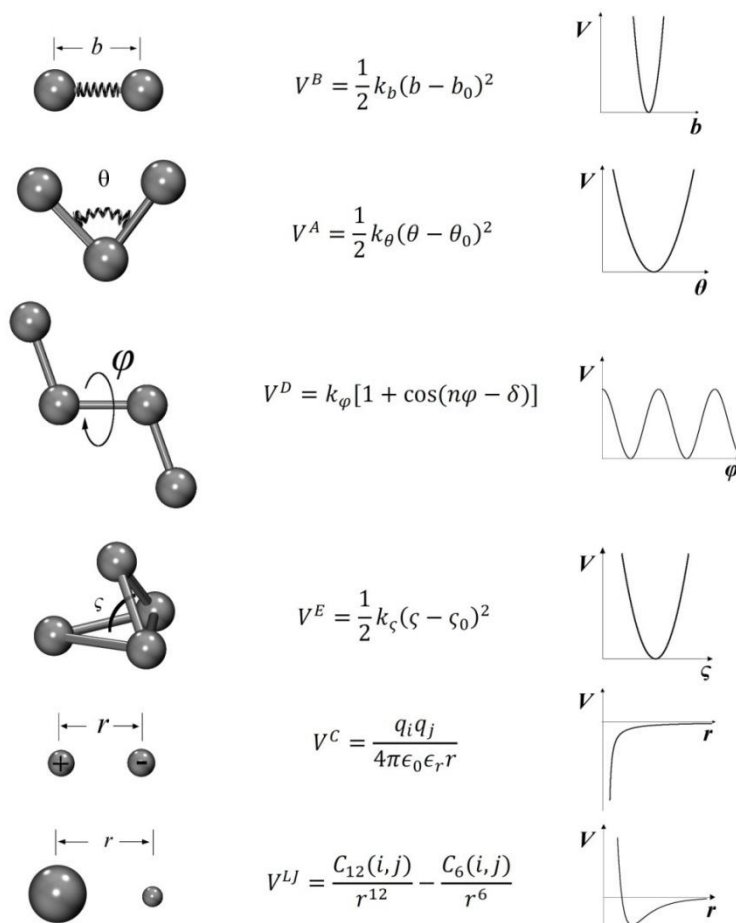
Fig. 4.10. (based on H.Grubmueller, "Proteins as Molecular Machines: Force Probe Simulations", published in Computational Soft Matter: From Synthetic Polymers to Proteins, Lecture Notes, Norbert Attig, Kurt Binder, Helmut Grubmueller, Kurt Kremer (Eds.), John von Neumann Institute for Computing, Julich, Germany, NIC Series, Vol. 23, ISBN 3-00-012641-4, pp. 401-422, 2004. © 2004 by John von Neumann Institute for Computing), Interaction contributions to a typical force field. Bond stretch vibrations are described by a harmonic potential V^B, the minimum of which is at the equilibrium distance b_0 between the two atoms connected by chemical bond i (the indices i, j etc. are not shown in the Figure). Bond angles and out-of-plane (improper) angles are also described by harmonic potential terms, V^A and V^E, where Θ_0 and ζ_0 denote the respective equilibrium angles. Dihedral twists (torsional angles) are subjected to a periodic potential V^D; the respective force constants are denoted by k's with appropriate indices. Non-bonded forces are described by Coulomb interactions, V^C, and Lennard-Jones potentials, $V^{LJ} = V^P + V^{VdW}$, where the latter includes the Pauli repulsion, $V^P \sim r^{-12}$ and the van der Waals interaction, $V^{VdW} \sim -r^{-6}$, respectively.

In protein and DNA modeling the most popular force fields are: CHARMM, AMBER, GROMOS, OPLS. References for their full formal description may be found in Wikipedia. We will use CHARMM v.27 in our tutorial exercises.

Geometry optimization (=energy minimalization)

According to Anfinsen hypothesis native proteins adopt conformations corresponding to minimum of their free energy F. The free energy function, defined as F=E-TS, takes into account a classical potential energy E (taken from the FF) and entropic factors (S) and a temperature (T) of a system. Quite often we neglect entropic factors. Then finding the minimum of F is equivalent to finding the corresponding minimum of E. Thus, we use computers to answer the fundamental question: for what atomic positions the energy E of a biomolecule is the lowest. Since E depends on positions of so many atoms, and we need take into account water as biological environment, the task is not trivial. We need to minimize a function of many, many variables: a rough estimate for hemoglobin (20 000 atoms) gives 60 000 degrees of freedom. All this has to be optimized! Fortunately it is not a hopeless problem: despite lack of a general algorithm for finding the global minimum of such multidimensional function, there are methods for finding some local minima.

„Brute force" minimization

When the number of degrees of freedom is not large, the systematic scanning of the whole conformational space may lead to good results. For biomolecules it is a dead end, since the number of conformers exponentially grows with the number of atoms.

Steepest descent methods

We will present the steepest descent method (SD) for one dimensional function f(x).

1. Select the starting point x_0
2. For point $x_k = x_0$ calculate the negative gradient:

$$-\nabla f(x_k)$$

3. The next point is determined from the formula:

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k)$$

4. The step size (in a given direction) is determined by linear minimization in this direction, we select this factor γ_k, that the following condition is fulfilled:

$$f(x_k - \gamma_k \nabla f(x_k)) = min_{\gamma > 0} f(x_k - \gamma_k \nabla f(x_k))$$

5. Iterate points 2-4 until stop criterion is met:

$$\nabla f(x_k) \leq \varepsilon \text{ and } f(x_k) - f(x_{k-1}) \leq \varepsilon'$$

The scheme may be generalized into multidimensional functions.

The Conjugated Gradients (CG) method offers better efficiency than SD. Here the direction of the minimization is determined not form the pure gradient, but form a "mixture" of the current gradient and the previous direction of the minimization.

1. Find the value of the steepest descent (sd)

2. If sd < ε, end the search
3. If it is a step no. 1 – remember the direction of the descent (for step no 1 it will be a direction of the gradient)
4. For a step no. K, K>1, determine the direction of the descent. It will be a combination of the current gradient and the previous step direction of the descent (conjugate directions). Find a minimum in that direction. Go to this minimum.
5. Go to point no. 1.

There are numerous popular variants of the conjugate gradient methods: Fletcher-Reeves, Polak-Ribiere, Hestenes-Steifel (see:

http://www.mymathlib.com/optimization/nonlinear/unconstrained/fletcher_reeves.html

The CG algorithm requires moderate memory size, is relatively fast converging, but does not work well for structures far from the minimum. In such a case the SD algorithm is better.

The main drawback of these algorithms (SD CG and similar) applied to molecular mechanic problems is the need for calculations of energy function gradients. In typical force Fields such a function has to be numerically differentiated. This is a computationally intensive task. Back in 1970s another minimization algorithm has been introduced, similar to CG, but in this method there is no need for grad E calculations. It is called Powell algorithm, since it is widely used in MD simulations, it is recommended to study details of this approach:

 (See Press, WH; Teukolsky, SA; Vetterling, WT; Flannery, BP (2007). "Section 10.7. Direction Set (Powell's) Methods in Multidimensions". *Numerical Recipes: The Art of Scientific Computing* (3rd ed.). New York: Cambridge University Press. or http://people.equars.com/marco/poli/phd/node55.html)

2nd order methods (Newton-Raphson, N-R)

Much faster convergence may be achieved (in 1 step for quadratic surfaces!) when the information of second derivatives of the optimized function (PES) may be included.

Les us put:

x_k – a point in configuration space in step k

d_k – direction of minimization in step k,

-gradient:

$$-\nabla f(x_k) = -g_k$$

ω, λ – scalar coefficients

H –matrix of second derivatives (Hessian matrix)

V – an approximation (matrix) to the Hessian reciprocal matrix

According to the N-R method we have:

$$x_{min} = x_0 - H_0^{-1}\nabla f(x_0)$$

***

An example (1 –dim)

$f(x) = a + bx + cx^2$

$f'(x) = b + 2cx$

$f''(x) = 2c$

In minimum we have $f'(x_{min}) = 0$ thus $b + 2cx_{min} = 0$ and $x_{min} = -b/2c$

Therefore:

$x_{min} = (-f'(x) + 2cx)/2c = x - f'(x)/f''(x)$.

***

Calculations of second derivatives for biopolymers are highly impractical, since very large RAM spaces are required. There are, however, numerous approximate N-R methods: quasi-Newton (Davidon-Fletcher-Powell or Broyden-Fletcher-Goldfarb-Fano). These methods are used in the ArgusLab code (semi-empirical quantum chemistry) to optimize geometry of medium-sized molecules.

Integration of equations of motion (Newton, Langevin)

In the classical molecular dynamics we assume that time evolution of the molecular biosystem is given by Newton equations of motion:

$$\boldsymbol{a}_i = \boldsymbol{F}_i / m_i$$

In order to determine the position x_i of an atom *i* in time point *t* we have to know its mass m_i, the force acting on the atom $\boldsymbol{F}_i$ (t) and integrate the equation twice with respect to time. We need to repeat this procedure for each atom one by one, sometimes 20 000 or 200 000 times (N). If we know the force field (and we know it , indeed) we may calculate the gradient of E, in a point of the configurational space that we study in the current moment. Thus we can obtain required forces $\boldsymbol{F}_i$ acting on atoms:

$$\boldsymbol{F}_i = -\nabla E(\boldsymbol{x})$$

We need to integrate a differential equation for the acceleration $\boldsymbol{a}_i$ twice, in order to find a trajectory for a given atom x(t). In the numerical analysis methods we search for the most effective way of performing this integration (using a computer, of course). Among many methods the most popular one is perhaps an algorithm developed by Verlet. We present a few variants of this algorithm.

In this approach time is not continuous anymore: it is divided (discretized) into equidistant points. We solve the problem on a mesh. The time step depends on mechanical problem studied. Too small time step is expensive and redundant, to large leads to a crash of simulations. Usually we require that the time step is not smaller than the fraction 1/10th of the period of the fastest periodic motion in the system under study. In practice, for proteins at 300K, the fast test are C-H oscillations, and we use 1 fs time step (sometimes 2 fs will do as well). During practical exercises it is worth checking how

big a time step may be before your system crashes. If one sees that the energy of the system suddenly rises up, and MD simulation stops before reaching assigned number of steps, then it is recommended to repeat this job with the time step reduced by a factor of 2.

We want to simulate motions of biomolecular atoms. Atomic motions are related to equilibrium temperature of the system and surroundings. It is safe to use the statistical concept of the temperature, since our system contains usually thousands of atoms. We initiate MD calculation by selecting of initial velocities of atoms. We require conservation of the momentum:

$$\mathbf{P} = \sum_{i=1}^{N} m_i \mathbf{v}_i = 0$$

Directions of the velocity vectors are random. Velocity components are generated from the Maxwell-Boltzmann distribution:

$$p(v_{xi}) = \sqrt{\frac{m_i}{2\pi k_B T}} \exp\left[-\frac{m_i v_{xi}^2}{2 k_B T}\right]$$

$$T = \frac{1}{3N} \sum_{i=1}^{N} \frac{|p_i|}{2m_i}$$

The „positional" Verlet algorithm uses information on atomic positions in time points t and $t - \delta t$ in order to determine new positions in time point $t + \delta t$. One can easily generate this formula using the following Taylor expansions of r(t):

$$\left.\begin{aligned} r(t + \delta t) &= r(t) + v(t)\delta t + \frac{1}{2} a(t)\delta t^2 \\ r(t - \delta t) &= r(t) - v(t)\delta t + \frac{1}{2} a(t)\delta t^2 \end{aligned}\right| -$$

Limiting the expansions to the second order terms, and subtracting both series one has the recipe for trajectory propagation:

$$r(t + \delta t) = r(t) - r(t - \delta t) + \frac{1}{2} a(t)\delta t^2$$

The „Leap-frog" algorithm

This is another popular variant of Verlet algorithm. Like in a childrens game „leap-frog" in this algorithm we calculate velocities for time points $t + \delta t/2$, then positions are determined for $t + \delta t$.

$$\begin{cases} v\left(t + \frac{1}{2}\delta t\right) = v\left(t - \frac{1}{2}\delta t\right) + a(t)\delta t \\ r(t + \delta t) = r(t) - v\left(t + \frac{1}{2}\delta t\right)\delta t \end{cases}$$

The advantage of LP variant of Verlet is the possibility of storing velocities (we don't work in the configurational space but in the phase space). Note, however, that the positions and velocities are shifted in time, often we use average for velocities:

$$v(t) = \frac{1}{2}\left[v\left(t - \frac{1}{2}\delta t\right) + v\left(t + \frac{1}{2}\delta t\right)\right]$$

Sometimes „velocity" variant of Verlet algorithm is used:

$$\begin{cases} r(t + \delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 \\ v(t + \delta t) = v(t) + \frac{1}{2}[a(t) + a(t + \delta t)]\delta t \end{cases}$$

Practical aspects of molecular dynamics simulations. A study on a mechanical strength of nanocomposite biomaterial.

Temperature in SMD simulations

While integrating equations of motions we need to conserve the assumed statistical properties of the modeled system. We may want to keep constant the total energy (canonical ensemble), or temperature (microcanonical ensemble, a contact with thermal bath is assumed). During our simulations we will keep temperature constant (Langevin method). The simple method of velocity scaling also helps to maintain temperature during MD simulations. We use energy equipartion theorem – in each degree of freedom, in equilibrium, an energy of ½ kT is stored :

$$E_{kin} = \frac{3}{2} \, NkT = \sum_{i=1}^{N} \frac{m_i^1 v_i^2}{2}$$

One can immediately see that the proper temperature may be maintained through atomic velocity scaling.

Simulation of an environment

In bionanotechnology the main solvent is water. Physicists have developed more than 20 computers models of water, of which all have some advantages but none in perfect (http://en.wikipedia.org/wiki/Water_model , Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K, Pekka Mark and Lennart Nilsson , *J. Phys. Chem. A*, 2001, 105 (43), pp 9954–9960). The water molecule is particularly difficult to describe. We will use TIP3P model, similar to TIP4P presented below in a figure (fig. from the Internet: u-tokyo/~murayama)):
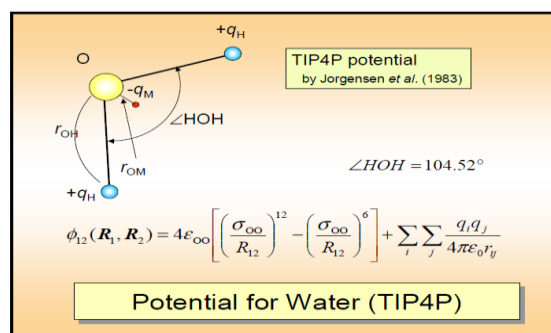

Fig. 4.11. TIP4P computer model of water molecule.

Sometimes we prefer to embed our studied protein in a droplet of water. This is good for SMD. More popular, especially for the system equilibration phase, is an approach in which Periodic Boundary

Conditions (PBC) are imposed. The space is divided into 27 cubes, the central cube contains modeled system, the 26 cubes have a copy (an image) of the primary box. If an atom leaves from the central cube, its image enters back to the primary box from the appropriate image box. In that way a number of atoms is kept constant and the central model "feels" an environment similar to that in an infinite crystal.

The main technical obstacles in practical MD simulations are very long computations times. Existing codes use up to 70% of CPU time for calculations of nonbonded interactions. The cost growths as N^2 with the size of the system. Since VdW or electrostatic interactions decay rather fast with a distance, these forces are simply neglected above certain cut-off distance. In practice a cut-off of 12-16 Ang. is sufficient for a protein modeling. Of course, one should carefully modify border regions of potentials, in order to maintain the proper asymptotic behavior.

A general scheme of MD simulations

1. Select a model (what do you want to study, what for, what is a source of your atomic coordinates)
   (a) A good source of structures is pdb
   (b) Protein should be "tuned" using special codes (VMD may help), missing atoms or missing residues should be added, immerse protein into a box of computer water (VMD again) TIP3P is a popular model
   (c) Remove waters located too close to protein atoms (? 2.3 Ang.)
   (d) Check waters inside the protein model. Keep that molecules which have at least 3 h-bonds
   (e) Determine the charge of charged amino acids: Glu, Asp, Arg, Lys, His – it is related to pH
   (f) Check protonation of His residues, check protonation of all charged residues with respect to assumed pH
2. Select the force field. Check that you have all necessary parameters.
3. Select cut-offs, size of the PBC box, frequency of non-bonded list regeneration.
4. Impose constrains on a protein. Equilibrate water shell/box for some 0.5-1 ns (up to 300K)
5. Heat gradually the whole system protein+water
6. Check the structure on a graphics (VMD)
7. Equilibrate the whole system without any constrains, in desired temperature, for at least 1 ns (discard these trajectories – useful for diagnostics only)
8. Generate the proper trajectory for the required time (1 ns – 10 ns , 1 fs time step is typical) Save snapshots (frames) from the trajectory each 100, or 1000 steps. Control energy, temperature etc. "on-the-fly".
9. Analyze collected trajectories (sometime up to 20 Gb of data) using your favorite graphical desktop and code, for example, VMD, Maestro, PyMol, Yasara, etc.
10. Repeat calculations using different code/ force field to reduce possible artefacts.

A study on mechanical strength of nanocompsite biomaterial – a caddisfly fiber

Some insects, such as the silkworm, spiders and caddisfly, can produce very strong thread. It serves to catch prey (spider web) and for protection. The composition of these materials remains partially unknown. They contain a big proportion of protein, together with mineral ingredients. In general, such silk is a composite material characterized by interesting mechanical properties. The toughness

of spider silk may surpass that of Kevlar (bullet resistant vests). Bionanotechnology often employs biomimetic approach, which is based on observing the nature. Our aim here is to understand why natural materials have mechanical properties that are so perfect.

We may use classical modeling MD variant called steered molecular dynamics. The idea is very simple: to some of the atoms of a protein a harmonic potential is added, holding them still. To other atom, or atoms (eg. on the other side of the molecule), a force is applied through a spring having a spring constant k (it links the towed atom to an additional atom to which the force is applied). When the towing atom moves away from the towed atom, the spring's length increases and, in accordance with Hooke's law and the third Newton's law of motion, spring acts on the towed atom with increasing force. The force leads to unwinding of the protein, which we observe on a computer screen. This way we can study mechanical properties of different proteins or protein domains. Such a study will be a topic of our exercises.
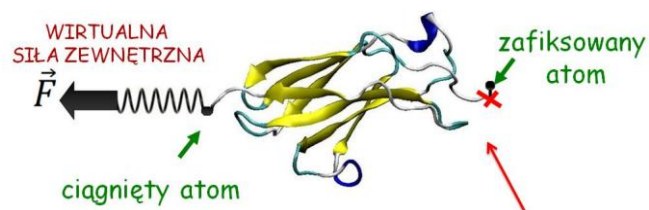
Fig. 4.12. A principle of Steered MD.

We will study a fragment of caddisfly protein (Trichoptera). These insect make a glue working in water. We hope that better understanding of this material may lead to new types of surgical tissue adhesives. The caddisfly fibers were studied experimentally and computationally in the Department of Biophysics and Medical Physics, NCU in Torun, Poland:

(J. W. Strzelecki, J. Strzelecka, K. Mikulska, M. Trzydel, A. Balter, W. Nowak, *Nanomechanics of new materials – AFM and computer modelling studies of Trichoptera silk*, Cent. Eur. J. Phys. 9 (2011) 482 - 491.)

Fig. 4.13. *Trichoptera* make a fiber which glue stones.

Let us have a look at a spider silk fiber structure:

SEM (M. Tszydel)

Oprzęd pająka (Nature Materials, 2010 r) [2]
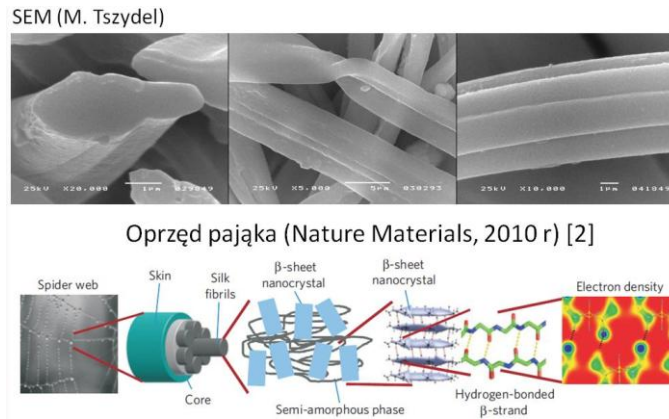
1710   *Nanomechanics of spider silk proteins*   S. Keten and M. J. Buehler
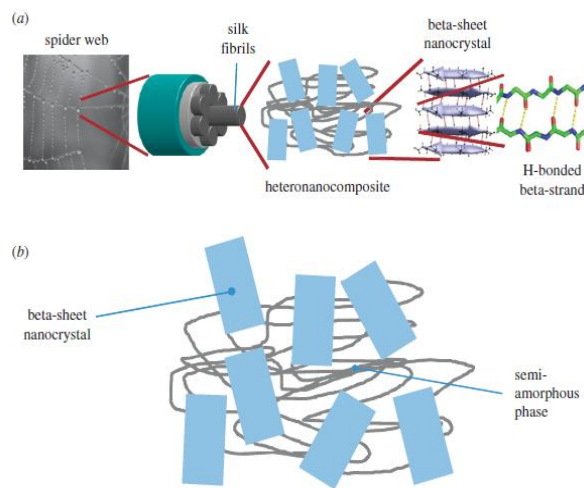
Fig. 4.14. The complex structure of caddisfly fiber. It is nanocompopsite biomaterial.

For simulations we select only a small, but characteristic, fragment of fibroin. It is a main protein in the fiber:

KISRSASYSVERIVTPTVITRISGSHSVSAE

(H fibroina, Fib-H)
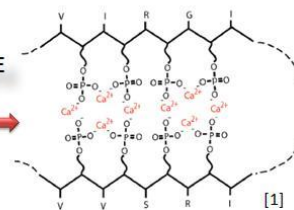
Fig. 4.15. A fragment of H- fibroin (H-f) protein sequence will be used in computer simulations (KM).

Note that the H-f protein fragment contains 8 serine residues (S, Ser). This amino acid contains the OH group in its side chain. It is prone to enzymatic phosphorylation. The PO^{4-} groups (negatively charged) are added. If Ca^{+2} ions for some reasons are present in model protein, they may "glue" the negatively charged phosphorylated beta strands. W expect that this phenomenon should increase the mechanical strength of the fiber. <u>We will test this hypothesis computationally</u>. We will use the SMD method and will attempt to rupture the protein in at least two cases:

(1) when serines are phosphorylated and contain calcium ions
(2) when serines are phosphorylated but there is no calcium

Of course one may expect that the two remaining cases of SMD simulations are also informative, ie. when serines are not phosphorylated and there is calcium (3) or calcium is absent (4). All these cases were studied in the past, during practical exercises we will limit studies to cases no. (1) and (2).
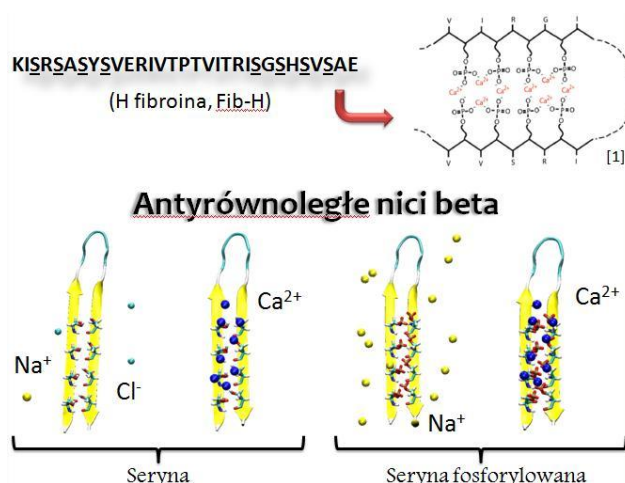

Fig. 4.16. Four models of beta H-fibroin fragments, used to study on the role of phosphorylation and Ca^{+2} in fiber nanomechnics (fig. KM and WN).

We need to select SMD parameters, some hints are presented in instructions to the exercises:
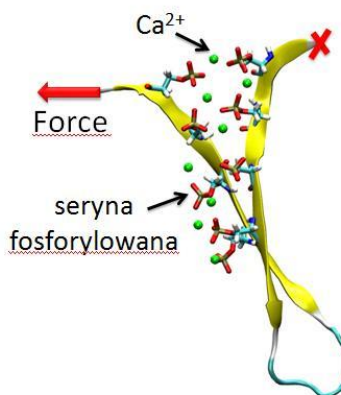

Fig. 4.17. Suggested set-up for simulations (fig. KM).

It is always recommended that one should know in advance possible outcomes of computer experiments. Here are our earlier results of mechanical unfolding of fibroin fragments:
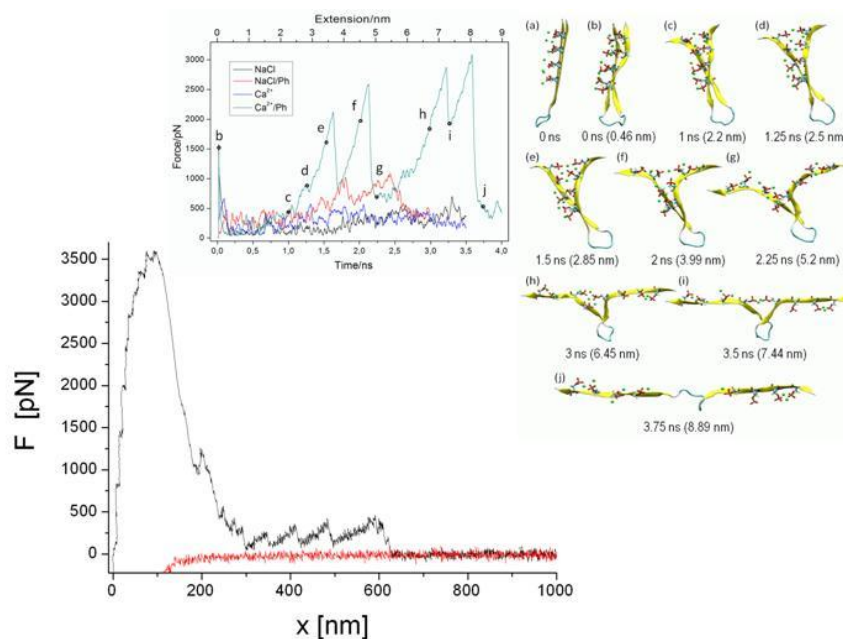

Fig. 4.18. Selected results of earlier SMD simulations of H-F fragments and AFM measurements of a single fiber (J. Strzelecka). Note how phosphorylation and a presence/absence of calcium ions affect the Maxim force. How the calculated force compares with the AFM curve?

Our conclusions were as follows: phosphorylation alone increases mechanical strength of a fiber by 100%. The phosphorylation in the presence of Ca^{+2} in the model increases a mechanical resistance by 800%-1100%. It would be interesting to know to what extent these results are sensitive to details of computational protocol. Should we accept these finding or reject? Good luck in computational invesitigations ☺.

Literature :

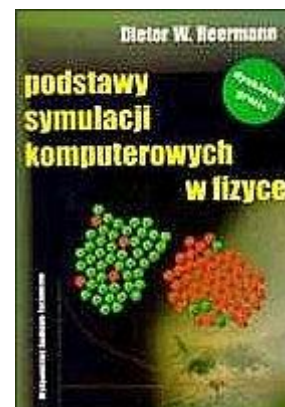1. Podstawy symulacji komputerowych w fizyce   14.00zł

   Autor: Dieter W. Heermann
   ISBN: 83-204-2047-4
   Ilość stron: 173
   Data wydania: 1997

2. Manuale VMD i NAMDA:
   http://www.ks.uiuc.edu/Research/vmd/current/ug/
   http://www.ks.uiuc.edu/Research/namd/2.7/ug/

3. Handouts
4. L. Piela , Idee chemii kwantowej. (fragm.)
   Wydawnictwo Naukowe PWN 2005, ok. 55 zł, ISBN: 83-01-14000-3

5. Wiesław Nowak,
    "Applications of computational methods to simulations of proteins
    dynamics." ,
    in "Handbook of Computational Chemistry", Springer, 2012 – a book
    chapter, pp.1129-1114, ISBN 978-94-007-0712-2