Thèse de Doctorat de l'Université d'Evry-Val-d'Essonne

Discipline : Biologie Moléculaire et Cellulaire

Spécialité : Biochimie

Présentée par **Vladislav KLYASHTORNYY**

Pour obtenir le grade de

Docteur de l'Université d'Evry-Val-d'Essonne

PRINCIPLES OF PROTEIN:NUCLEIC ACID RECOGNITION ON THE EXAMPLES OF THE RIBOSOMAL PROTEIN L1 AND THE COLD SHOCK DOMAIN OF YB-1 PROTEIN

Directeurs de Thèse : **Stanislav NIKONOV** et **Philippe MANIVET**

Soutenue le vendredi 13 janvier 2012

Devant le jury composé de

Olivier MAUFFRET	Rapporteur
Alexei RAK	Rapporteur
Stanilav NIKONOV	Examinateur
Lev Pavlovitch OVCHINNIKOV	Examinateur
Patrick A. CURMI	Examinateur
Philippe MANIVET	Examinateur

Content

Introduction

Molecular recognition between biological macromolecules has been one of the key problems of molecular biology for a long time. How macromolecules find and recognize each other in the cells? Why some protein complexes break just after formation whereas other ones are stable for several minutes and even hours? This is far from a full list of questions to be answered by modern explorers.

The interactions between proteins and nucleic acids play a central role in molecular biology. These interactions determine many key processes in living cells: division, gene expression and regulation, mRNA translation and regulation, DNA repair etc… To that end proteins should posses either sequence specific or nonspecific affinity for nucleic acids which depends on the process to which they participate. Indeed, some proteins bind to only highly specific nucleic acids (DNA or RNA), while other proteins show a high variability of the sequences they can recognize. The first group includes proteins which participate to dedicated special processes like the regulation of a gene expression or mRNA translation. This group of proteins also comprises proteins which function in multimolecular complexes with RNA/DNA. For these proteins this requires an ability to bind strongly and very specifically to exactly defined place on the targeted nucleic acid. The best example of such complexes is ribosome which represents a large complex of RNA and different proteins (*ribosomal proteins*) and works as a machinery to synthesize proteins in the cells. Some of the ribosomal proteins have dual function: i) they are involved in ribosomal function; ii) as well as to the regulation of mRNA translation. The second group of the proteins usually includes proteins involved in many different processes in the cells. YB-1 protein is one of the brightest examples of such proteins. It participates in most of the processes which are important for the survival of eukaryotic cells showing a high specific and nonspecific affinity for both DNA and RNA.

In the present work, we study some aspects of the mechanism of protein-nucleic acids interactions using the examples of two proteins that either interact specifically with RNA (*ribosomal protein L1*) or show a high affinity for different DNA/RNA sequences (*YB-1 protein*). To study the interaction between nucleic acid and a protein one usually uses point mutation to explore the region of the interface. However it is impossible to say a priori how a substitution will change the molecular structure. Therefore the analysis of binding assay experiments brings sometimes hypothetic character if it is not associated with structural explorations. In the first part of the present work, the structures of L1 mutants were determined by X-ray crystallography method to reveal the influence of some point mutations at the binding site on the RNA-binding properties of L1 protein.

In the second part we studied the nucleic acids binding properties of YB-1 Cold Shock Domain (CSD) based on Molecular Dynamics simulation due to difficulties of the structural exploration of these complexes with experimental methods. The results obtained contribute to enlighten some of the principles of protein:nucleic acids recognition and the factors responsible for the stability of these macromolecular complexes.

The thesis consists of Introduction, Literature review, Material and Methods as well as Results and Discussion sections and contains a reference list. The Literature review section is devoted to the structure-function description of the two proteins studied here: L1 (section 1.1) and YB-1 (section 1.2). Both of them possess a high specific affinity to nucleic acids. YB-1 also shows high unspecific affinity to nucleic acids. The RNA binding properties of L1 was analyzed with X-ray method, so the section 2.2 is devoted to detailed description of this method. The nucleic acid binding properties of YB-1 were studied with molecular dynamics simulation method. In the section 2.3 we give the detailed description of this method, too.

In the experimental part, the main steps of the structure determination and molecular dynamics simulation are summarized. In the Results and Discussion section, the principles of macromolecular recognition are demonstrated using the examples of the L1 and YB-1 proteins. We underlined the key role of intermolecular H-bonds that are inaccessible to solvent molecules. The work was performed at the Institute for protein research (Pushchino, Russia) and at the University of Evry (Evry, France). The data shown in the work were published in several papers and reported at different conferences and seminars.

Chapter 1. Literature review

1.1. Ribosomal protein L1

L1 is one of the biggest ribosomal proteins (molecular weight 25kDa). In the cell it has two functions: *ribosomal* and *regulatory of translation*. Homologies of L1 protein were found in all the life domains: bacteria, archea and eukaryotes.

1.1.1. Activity of L1 protein

Ribosomal protein L1 from bacterium *Escherichia coli* (EcoL1) specifically binds to 23S rRNA independently of other proteins and protects a 100 nt length fragment from nuclease activity [Zimmerman 1980, Branlant *et al.* 1981]. EcoL1 also has a significant affinity to different bacterial and archeal 23S rRNAs [Stanley *et al.* 1978, Baier *et al.* 1989], as well as to eukaryotic 26S/28S rRNAs [Gourse *et al.* 1981]. Moreover L1 from archea *Methanococcus vannielii* was shown to replace functionally EcoL1 in *E. coli* ribosome *in vitro* [Baier *et al.* 1990]. L1 is a primary binding ribosomal protein and in the complex with 23S rRNA it forms the so-called L1-protuberance. This is a very labile structure and its labiality is important for the releasing of deacylated tRNA from ribosomal E-site during translation [Wower *et al.* 2000, Kirillov *et al.* 2002].

In addition to its functions in the ribosome, L1 has one more activity. L1 from *Escherichia coli* (EcoL1) mediates autogenous regulation of translation by binding to a region within the leader sequence, close to the Shine–Dalgarno sequence, of the mRNA of the L11 operon coding for ribosomal proteins L1 and L11. L1 from *M. vannielii* (MvaL1) was shown to be an autoregulator of the MvaL1 operon encoding ribosomal proteins L1, L10 and L12 [Mayer *et al.* 1998]. It was also shown that EcoL1 can inhibit the in vitro translation of MvaL1 polycistronic mRNA and, conversely, that MvaL1 can inhibit the synthesis of both L11 and L1 proteins of *E. coli* [Hanner *et al.* 1994]. There are experimental data that L1 from *E. coli* can inhibit translation of MvaL1 mRNA in vitro, contrariwise, L1 from *M. vannielii* can inhibit the synthesis of L11 and L1proteins from *E. coli* It means that bacterial and archaeal L1 proteins are functionally interchangeable both in the ribosome and in the repression of translation. Moreover the L1-binding sites exhibit high similarity in both sequence and secondary structure to the L1 binding site on the 23S rRNA [Draper 1989]. Based on these biochemical data it was proposed that L1 binds to rRNA and mRNA in a similar manner and RNA-binding sites on the surface of L1 and

L1-binding sites on the surface of RNAs are structurally homological among different sources [Sor *et al.* 1987, Hanner *et al.* 1994].

L1 proteins from mesophilic and thermophilic bacteria and archaea bind to specific site on 23S rRNA with 5- to 10-fold higher affinity than to their regulatory binding site on mRNA [Kohrer *et al.* 1998]. This difference fits the requirements of classical regulation of ribosomal synthesis (*feedback inhibition*) based on direct competition between the two binding sites. L1 likewise many other ribosomal protein is not necessary for the cells to survive as the mutants lacking L1 protein, despite a slow growth and only a half rate of the protein synthesis, does not show any other defects as compared with wild-type ribosome [Subramanian *et al.* 1980].

1.1.2. Structural characteristics of L1 protein

Up to date some structures of isolated L1 proteins from different sources have been determined at high resolution [Nikonov *et al.* 1996, Nevskaya *et al.* 2000, Nikulin *et al.* 2003, Nevskaya *et al.* 2005, Nevskaya *et al.* 2002]. Description of two such proteins, from bacterium *Thermus thermophilus* and from archea *Methanococcus jannaschii*, is given in two next sections.

1.1.2.1. L1 from *Thermus thermophilus*

The spatial structure of isolated L1 protein from *Thermus thermophilus* (TthL1) was determined at 1.85 Å resolution [Nikonov *et al.* 1996**Erreur ! Signet non défini.**]. TthL1 contains 228 amino acids (49% sequence homology with EcoL1) and has a molecular weight of 24.7 kDa. This protein consists of two structural domains. N- and C-termini of the protein are located in domain I. This domain includes residues 1-67 and 160-228, the connectivity scheme of the secondary structure elements for this domain is α_1-α_2-β_1-β_2-β_7-β_8-α_7-β_9-β_{10} (fig. 1). This domain contains a well known structural motif, the split β-α-β motif [Orengo *et al.* 1993] or abc/ad unit [Efimov 1994]. This motif consists of a three-stranded (β_1, β_8, β_9) antiparallel β-sheet and one α-helix (α_7). Two other α-helixes (α_1 and α_2) are found N-terminal to the motif. The polypeptide chain runs through the first strand of the motif (β_1) then strand β_2, the whole domain II and stand β_7 and subsequently returns back to the second and third strands (β_8 and β_9) of the motif. An additional C-terminal strand (β_{10}) extends the β-sheet of the motif.

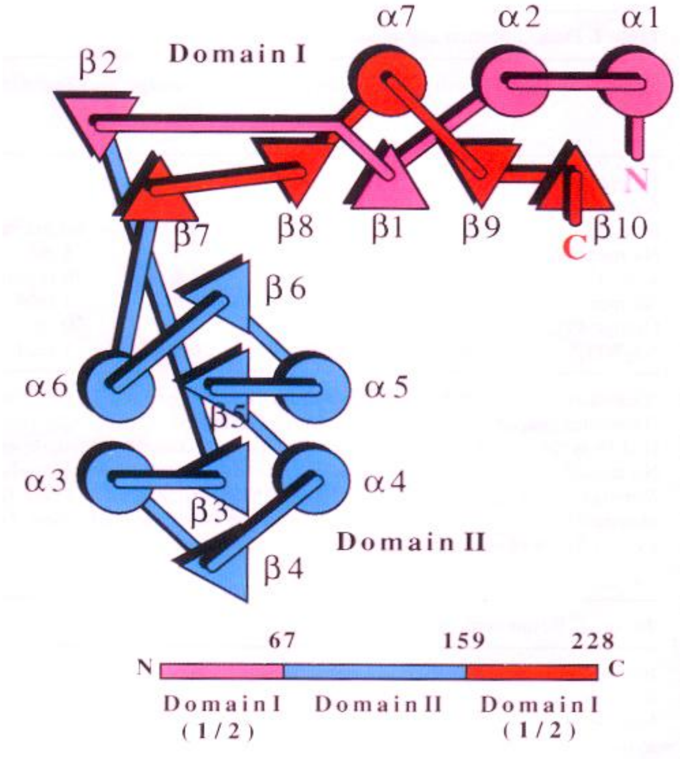

Figure 1. Schematic representation of the structure of ribosomal protein L1 from *Thermus thermophilus*. The figure is taken form [Nikonov *et al.* 1996].

Helix α_1 is quite separated from the globular part of the molecule and is associated with helix α_2 and strand β_{10} by hydrophobic interactions as well as by a salt bridge between Lys13 and Glu31. Eight N-terminal amino acids are very flexible which is confirmed by a low-quality electron density map in this region. The loop at positions 216-219 contains highly conserved residues and protrudes into the interdomain region. Residues 58-63 and 160-165 form a double-stranded (β_2, β_7) antiparallel β-sheet which is separated from the main sheet of domain I and makes the covalent connections to domain II. Domain II (residues 68-159) contains two helices on each side of a four-stranded parallel β-sheet with an overall Rossmann fold topology. The connectivity scheme is β_3-α_3-β_4-α4-β_5-α_5-β_6-α_6 (fig. 1).

Most of the conserved residues of TthL1 are located at the interface between the two domains and form two protrusions which are closed to each other. One of these protrusions is located at the C-terminal end of helix α_5 in domain II. The other protrusion is formed by the loop connecting strands β_9 and β_{10} in domain I. In close proximity to these protrusions there is the strictly conserved Phe37 belonging to the loop between helix α_2 and strand β_1 of domain I. Some residues in the interdomain region are involved in interdomain interactions, but this interaction is rather loose which results in an unstable mutual location of two domains relatively each other. Between domains there is a small but clear cavity. According to that the relative orientation of domains can be called “*closed*” in contrast to “*opened*” conformation found for L1 protein from *Methanococcus jannaschii*.

1.1.2.2. L1 from *Methanococcus jannaschii*

The spatial structure of ribosomal protein L1 from archea *Methanococcus jannaschii* (MjaL1) was determined at 1.85 Å resolution [Nevskaya *et al.* 2000]. The polypeptide chain contains 219 amino acid residues and shows 29% sequence homology with TthL1. An alignment of the primary sequences of these two proteins with secondary structure information is presented on figure 2. MjaL1 is an elongated molecule with two domains and overall dimensions of 57 Å x 45 Å x 32 Å. The entire structure is well ordered, and the only region with significant flexibility is the C-terminus (residues 213-219) where the electron density is weak. Domain I includes residues 1-56 and 149-219 and contains both C- and N-termini. Domain II spans residues 57-148. The connectivity schemes of the secondary structure elements are α_1-β_1-β_2-β_7-β_8-α_7-β_9-β_{10} and β_3-α_2-β_4-α3-α_4-β_5-α_5-β_6-α_6 for domain I and II respectively (fig. 1).

Despite the rather low sequence homology between MjaL1 and TthL1 their overall structures are very similar. Each domain of archeal L1 has approximately the same dimensions and is closely related topologically to its counterpart from *T. thermophilus*, although there are essential differences. The N-terminal helix of TthL1 is absent in the MjaL1 structure due to a shorter amino acid sequence at the N-terminus, the archeal protein also has seven extra residues in domain II forming helix α_4. Bacterial and archeal structures in the isolated state significantly differ in the relative orientation of two domains as well as in spatial orientation of α-helixes and β-sheets of both domain I and domain II. As a result they superpose with a rms of about 3.0 Å for each domain. However the structure of the β-sheets for these proteins is highly conserved (rms for C_α-atoms is equal to 0.35 Å). The specific feature of TthL1 structure, a bend of helix α_4, is also kept for archeal L1 (helix α_5).

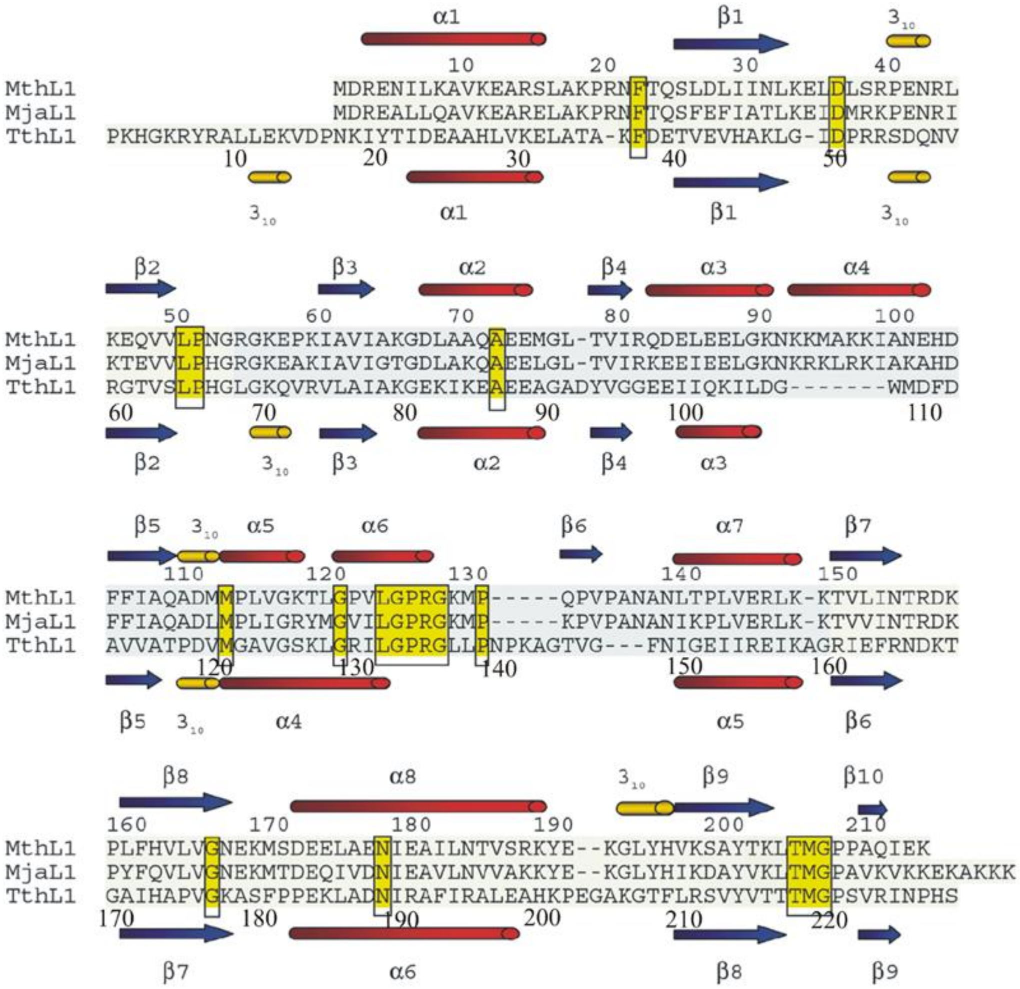
α1
β1
3₁₀
10 20 30 40
MthL1 MDRENILKAVKEARSLAKPRNFTQSLDLIINLKELDLSRPENRL
MjaL1 MDREALLQAVKEARELAKPRNFTQSFEFIATLKEIDMRKPENRI
TthL1 PKHGKRYRALLEKVDPNKIYTIDEAAHLVKELATA-KFDETVEVHAKLG-IDPRRSDQNV
10 20 30 40 50
3₁₀
α1
β1
3₁₀
β2
β3
α2
β4
α3
α4
50 60 70 80 90 100
MthL1 KEQVVLPNGRGKEPKIAVIAKGDLAAQAEEMGL-TVIRQDELEELGKNKKMAKKIANEHD
MjaL1 KTEVVLPHGRGKEAKIAVIGTGDLAKQAEELGL-TVIRKEEIEELGKNKRKLRKIAKAHD
TthL1 RGTVSLPHGLGKQVRVLAIAKGEKIKEAEEAGADYVGGEEIIQKILDG-------WMDFD
60 70 80 90 100 110
β2
3₁₀
β3
α2
β4
α3
β5
3₁₀
α5
α6
β6
α7
β7
110 120 130 140 150
MthL1 FFIAQADMMPLVGKTLGPVLGPRGKMP-----QPVPANANLTPLVERLK-KTVLINTRDK
MjaL1 FFIAQADLMPLIGRYMGVILGPRGKMP-----KPVPANANIKPLVERLK-KTVVINTRDK
TthL1 AVVATPDVMGAVGSKLGRILGPRGLLPNPKAGTVG---FNIGEIIREIKAGRIEFRNDKT
120 130 140 150 160
β5
3₁₀
α4
α5
β6
β8
α8
3₁₀
β9
β10
160 170 180 190 200 210
MthL1 PLFHVLVGNEKMSDEELAENIEAILNTVSRKYE--KGLYHVKSAYTKLTMGPPAQIEK
MjaL1 PYFQVLVGNEKMTDEQIVDNIEAVLNVVAKKYE--KGLYHIKDAYVKLTMGPAVKVKKEKAKKK
TthL1 GAIHAPVGKASFPPEKLADNIRAFIRALEAHKPEGAKGTFLRSVYVTTTMGPSVRINPHS
170 180 190 200 210 220
β7
α6
β8
β9

Figure 2. Sequence alignment of the L1 proteins. Residues invariant in all known bacteria and archaea are shown in yellow. Numbering of the secondary structure elements correspond to MjaL1 (over the sequence) and to TthL1 (below the sequence). The figure is taken from [Nevskaya *et al.* 2002].

1.1.2.3. Two conformations for L1 proteins

The main difference between the structures of TthL1 and MjaL1, as mentioned above, is that in TthL1 its two domains are closed to each other and make a contact. In MjaL1 the contact between the two domains is observed only at the region of the intermolecular hinge. As a result of that the interdomain cavity observed for TthL1 structure is almost absent for the archeal protein and the domains are significantly distant from each other and the molecule adopts an *opened* conformation in contrast to the *closed* conformation found for TthL1 (fig. 3). The conformation of interdomain hinge is also changed especially for its long chain (residues 50-58). The relative orientation of the two domains is stabilized by several hydrogen bonds and hydrophobic core that goes through entire molecule and contains the residues from the interface on the surface between two domains (Phe104, Leu146 and Val151).

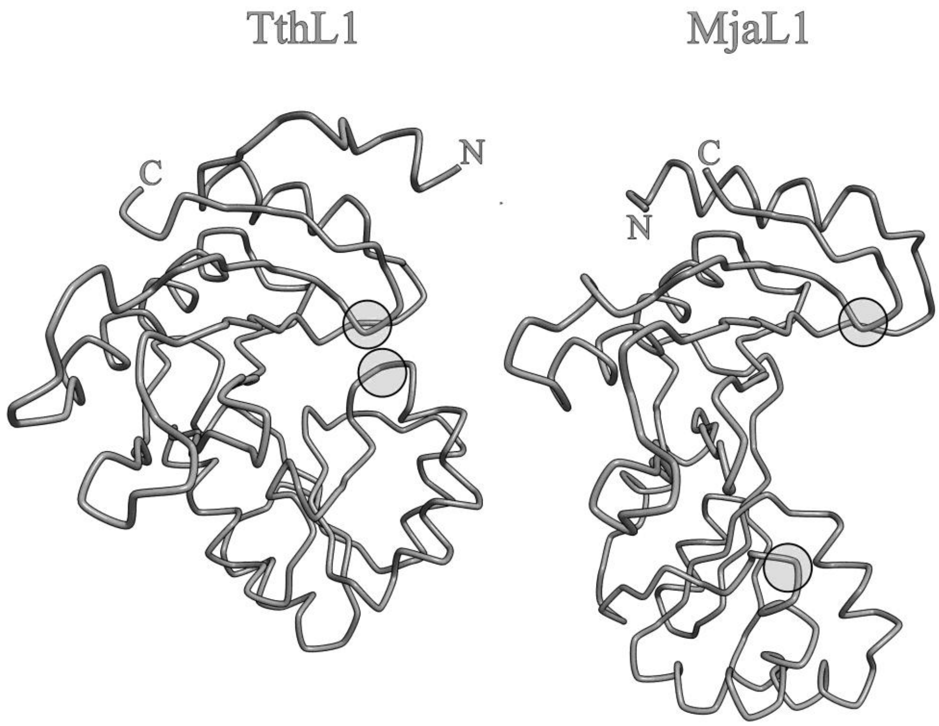

Figure 3. The spatial structures of isolated ribosomal protein L1. TthL1 (left) shows a *closed* conformation: the two domains are closed to each other and make many contacts between each other. MjaL1 has a fully *opened* conformation: the domains do not contact each other. Clusters of highly conserved residues in domains I and II are shown with circles.

A structural analysis of TthL& shows that there is no visible limitations interrupting transition of the closed conformation to opened one [Nevskaya *et al.* 2000]. It was suggested that this transition should take place for TthL1 protein upon RNA binding. Determination of the 3D structure of L1 in complex with different RNA fragments [Nikulin *et al.* 2003, Nevskaya *et al.* 2005, Nevskaya *et al.* 2006] confirmed this suggestion.

1.1.3. Interaction between L1 and RNA

Up to date some structures of complexes of ribosomal protein L1 with different mRNA and rRNA fragments are solved. First of them was the structure of L1 protein from *Sulfolobus acidocaldarius* in complex with rRNA fragment *Thermus thermophilus*.

1.1.3.1. The ribosomal complex

The crystal structure of ribosomal protein L1 from archea *Sulfolobus acidocaldarius* (SacL1) in complex with specific 55nt-length 23S rRNA fragment from *Thermus thermophilus* was determined at 2.65 Å resolution [Nikulin *et al.* 2003]. 3D structure of SacL1 is very similar to MjaL1 structure, relative domain orientation corresponds to opened conformation with the distance between two conserved clusters about 25 Å. 23S rRNA fragment contains helix 77, shortened helices 76 and 78 as well as connecting loops A and B (fig. 4A). Helix 78 is capped with a tetraloop GCAA, which does not affect L1 binding or interacts structurally with other parts of the complex.

The structure of the RNA fragment is stabilized by stacking interactions and a network of hydrogen bonds. Helices 76 and 77 are joined in one helical structure perpendicular to helix 78 (fig. 4B, C). Interconnecting loops A and B interact with each other. The loops are linked by three base triplets, A2114▪A2119▪G2168, G2115▪A2171-U2167 and A2117-U2172▪G2166, which interact via an extensive network of hydrogen bonds. The bases of interconnecting loops A and B form three stacking lines G2112▪A2117, A2169-U2172 and G2168▪G2162, the last of which is bifurcated into lines G2127-C2129 and C2161-G2159 of helix 78. The bases of G2127 and U2172 are pulled out of the stacking lines of helix 77 to stack on bases in helix 78 and loop B and to form canonical base pairs with C2161 and A2117, respectively. The ribose moieties of A2126 and G2127 are approximately perpendicular to each other, with a minimal distance of 2.95 Å between them. As a result, the RNA backbone bends sharply at the 5’ P atom of G2127. This turn is stabilized by a network of hydrogen bonds, including those to the base moiety of the highly conserved A2173. The RNA backbone between A2126 and C2129 resembles a loop terminating the shallow groove of helix 77. Another sharp bend of the RNA backbone is possible due to the formation of the canonical U2172-A2117 base pair and results in the bulging of G2116 and U2118, the positions of which are stabilized by hydrogen bonds. In addition, U2118 is stacked onto C2111 (fig. 4B). The bases of nucleotides C2111, G2116, U2118, C2163, C2164 and G2165 are largely exposed to the solvent.

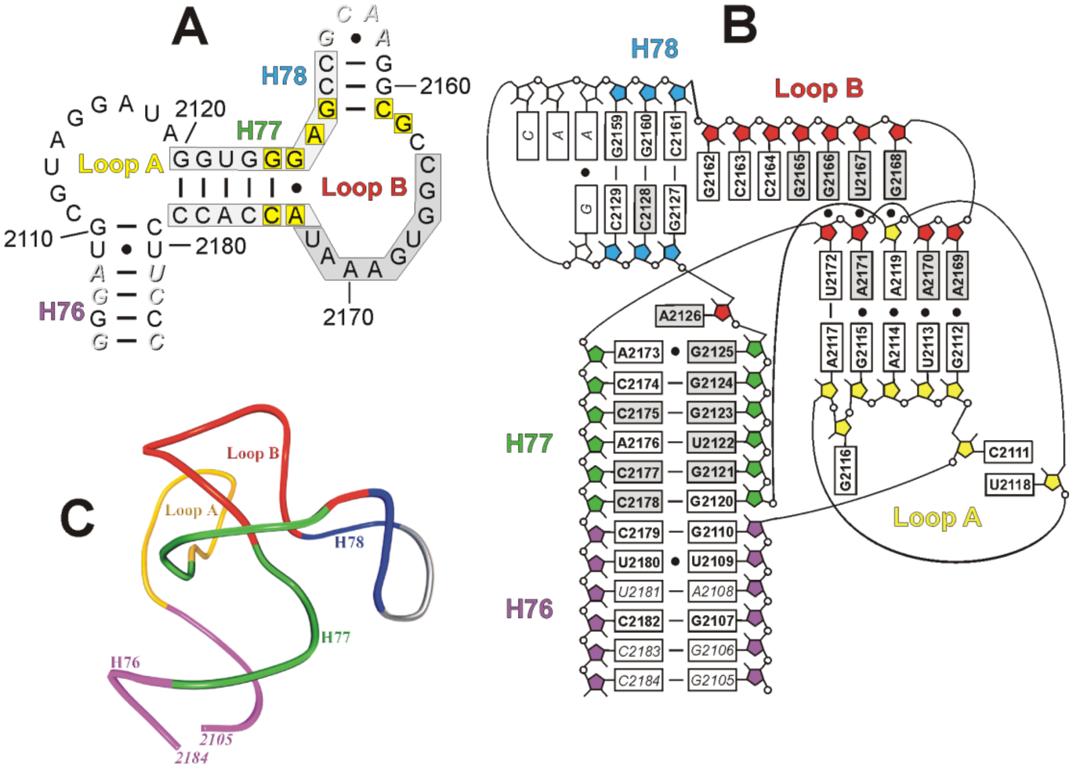

Figure 4. 23S rRNA fragment from *T. thermophilus*: **A** – secondary structure. Nucleotides artificially implemented to the structure are shown in italics. Nucleotides contacting domain I and domain II of the protein are shown with light-gray and dark-gray, respectively, yellow background indicates the nucleotides conserved in the binding sites of all 23S rRNAs and mRNAs; **B** – diagram of stacking interactions; **C** – spatial structure. The figure is taken from [Nikulin *et al.* 2003].

Protein L1 interacts with the 23S rRNA through both of its domains. From domain I, over 20 residues are involved in protein-RNA contacts, whereas the number of such residues in domain II is almost three times less. Domain I contacts 23S rRNA through the inner face of the β-sheet, consisting of strands β_1, β_7, β_8 and β_9 as well as loops α_1-β_1, β_7-β_8 and β_9-β_{10}, which together form a slightly concave surface. Loops α_1-β_1 and β_9-β_{10} contain residues that are identical in all known L1 sequences. Contacts between domain II and the RNA are formed by helix α_4 and loop α_5-β_6. The surfaces of the contacting regions in the two domains are approximately perpendicular to each other. Only one residue in the interdomain connector (Lys60) makes contact with 23S rRNA.

Two regions on the surface of the RNA, separated by a deep negatively charged cavity containing one Mg^{2+} ion and water molecules, participate in the interaction with L1 protein (fig. 5). The first region which includes helix 77 and one strand of helix 78 interacts with domain I of the protein. In particular the shallow groove of helix 77 which is terminated by one strand of helix 78 interacts with the concave region of domain I mimicking the turn of the helix. Strands β_1, β_8 and β_9 of the protein are aligned with this shallow groove. The RNA backbone of the highly conserved nucleotides G2124, G2125, G2127 and C2175 form an approximately planar platform that contacts the strictly conserved residues Phe26, Thr208, Met209 and Gly210 of SacL1.

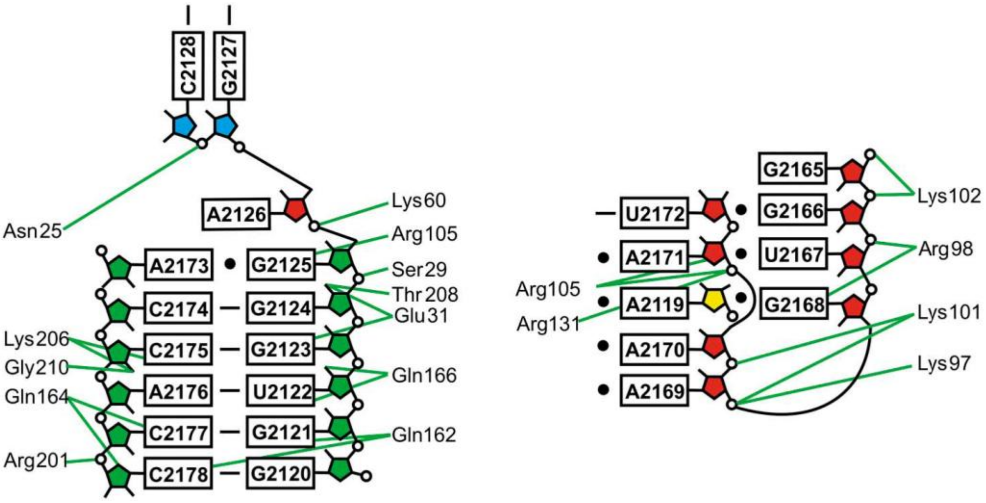

Figure 5. Diagram of RNA-protein hydrogen bonds (shown in green) for the complex between SacL1 and 23S rRNA fragment from *Thermus thermophilus*. The colors of the riboses correspond to those as at the fig. 4. The figure is taken from [Nikulin *et al.* 2003].

The second region of interaction formed by nucleotides G2165-A2171 of loop B imitates the shallow groove of a RNA A-helix and interacts with a cluster of positively charged amino acids including six residues from domain II. With the exception of G2168 this region of the RNA contacts protein L1 only through its backbone. Comparison of the structure of the complex L1/rRNA with the structure of the complex L1/mRNA showed that RNA-protein interface is well conserved.

1.1.3.2. The regulator complex

The structure of the complex between L1 protein from *Thermus thermophilus* and 38nt-length mRNA fragment from *Methanococcus vannielii* was determined at 2.6 Å resolution [Nevskaya *et al.* 2006]. The ribbon model of the complex is shown at fig. 6.

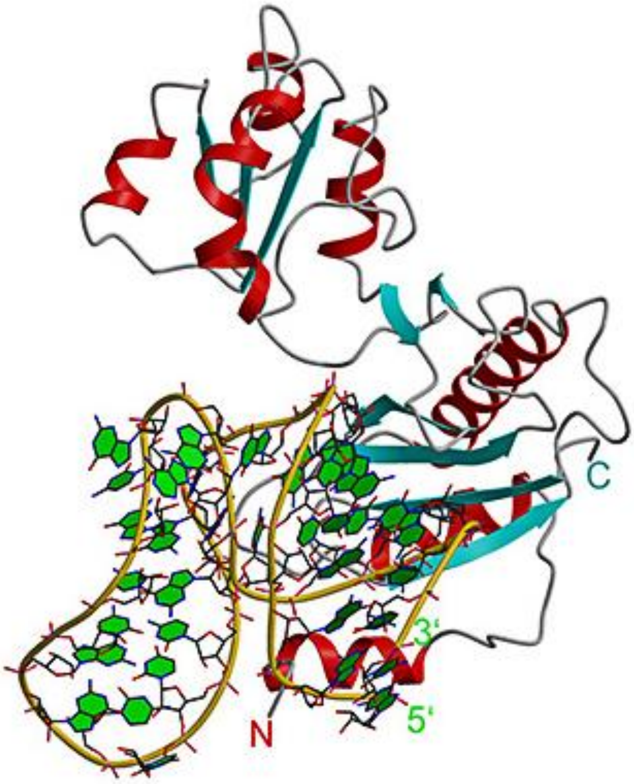

Figure 6. Ribbon model of the complex of TthL1 with mRNA fragment.

The structure of each domain is essentially preserved between TthL1 in the complex and in the isolated molecule. However, as it was suggested earlier, it was observed a transition from the closed conformation of isolated protein to the opened conformation when the protein is bound to RNA (fig. 7). If we superpose the domains I of both structures, then domain II of TthL1 in the complex rotates approximately 75° around an axis that passes near the C_{α}-atom of Ile161 roughly parallel to the β-sheet of domain II and perpendicular to its β-stands. The crucial residue is His66, which C-O bond noticeably changes direction and induces conformational alterations in the succeeding residues of the hinge. The second chain of the hinge (158-160) changes its position in the complex but retains the same torsion angles as the isolated protein. Although both domains do not significantly change their structure and the domain-by-domain superposition yields a rms deviations of 0.88 Å and 1.08 Å for C_{α}-atoms of domain I and domain II respectively. The most significant deviations were found in the loop 33-37 (with maximal C_{α}-atom displacement of 6.3 Å for Ala35; without this loop the rms is equal to 0.59 Å). The structure of domain II also preserves its conformation and significant alterations observed only for the hinge region. Without this region rms deviation is 0.45 Å.

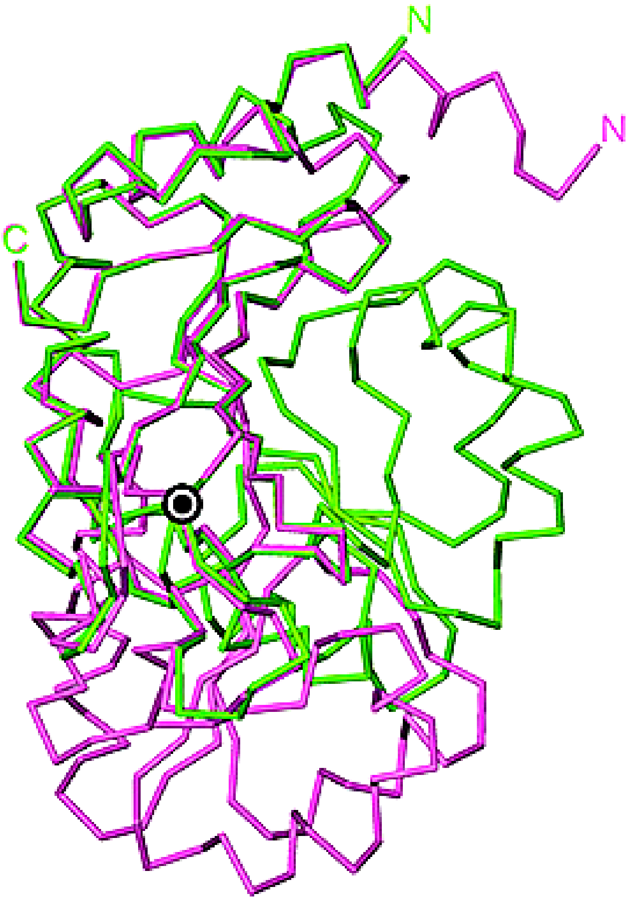

Figure 7. Superposition of the structures of TthL1 in the isolated (green) and RNA-bound form (magenta). To switch between the *closed* and the *open* conformations, domain II rotates around an axis which is shown by the black bulls-eye. The figure is taken from [Nevskaya *et al.* 2006].

The mRNA fragment form two regular A-helices that are practically perpendicular to each other. The long (10 base pairs) helix is closed with the non-canonical pair G10▪A27 (fig. 8A). The second short (3 base pairs) helix is capped with a UUCG tetraloop. The RNA backbone bends sharply at position G12, whereas nucleotides G19-C36 form a loop C23-A26 in the middle of the fragment (fig. 8B). As a result, the ribose group of G13 and C28 are closed to each other. The surface of the first mRNA helix is complementary to the surface of the β-sheet of TthL1 first domain.

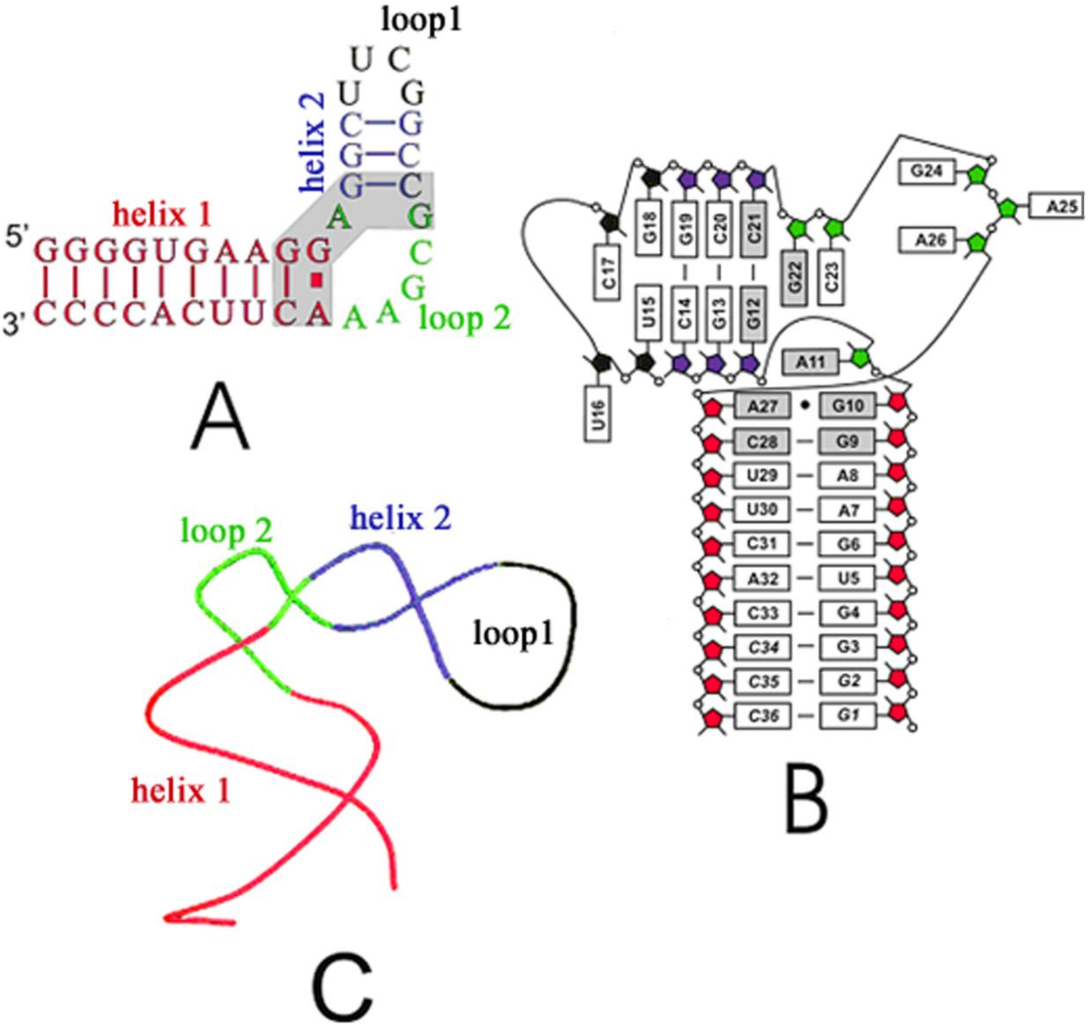

Figure 8. mRNA fragment from *M. vannielii*: **A** – secondary structure. Nucleotides in a gray background are conserved in the binding sites of 23S rRNAs and mRNAs; **B** – diagram of stacking interactions; **C** – spatial structure.

Comparative analyses of the 3D structures of two the complexes confirmed the preceding suggestion about similarity of the binding sites on L1 protein and RNA surfaces.

1.1.3.3. Structural comparison of the L1/mRNA and L1/rRNA complexes

The proteins in both RNA-protein complexes have similar 3D-structures corresponding to the opened conformation. Ribosomal RNA has more complicated spatial organization as both of its loops interact with each other. In mRNA, one of the loop (A) is absent, and another one (B) is shortened (fig. 8C). In spite of these differences both RNAs have similar unique region where two helices joined. In the rRNA fragment, the helices 76 and 77 make one helix perpendicular to helix 78, similar to the structure formed by two perpendicular helices of mRNA (fig. 9). In both RNAs the junction of the two helices contains nucleotides strictly conserved in all L1-binding sites on rRNAs as well as on mRNAs from bacteria and archea which specifically bind L1 and for which feedback regulation was shown. These nucleotides are connected by a network of conserved hydrogen bonds, most of which are inaccessible to solvent molecules, that stabilizes the unique 3D-structure of this region. These sites are also structurally conserved and yield a rms deviation of about 0.16 Å for P-atoms (fig. 9C).

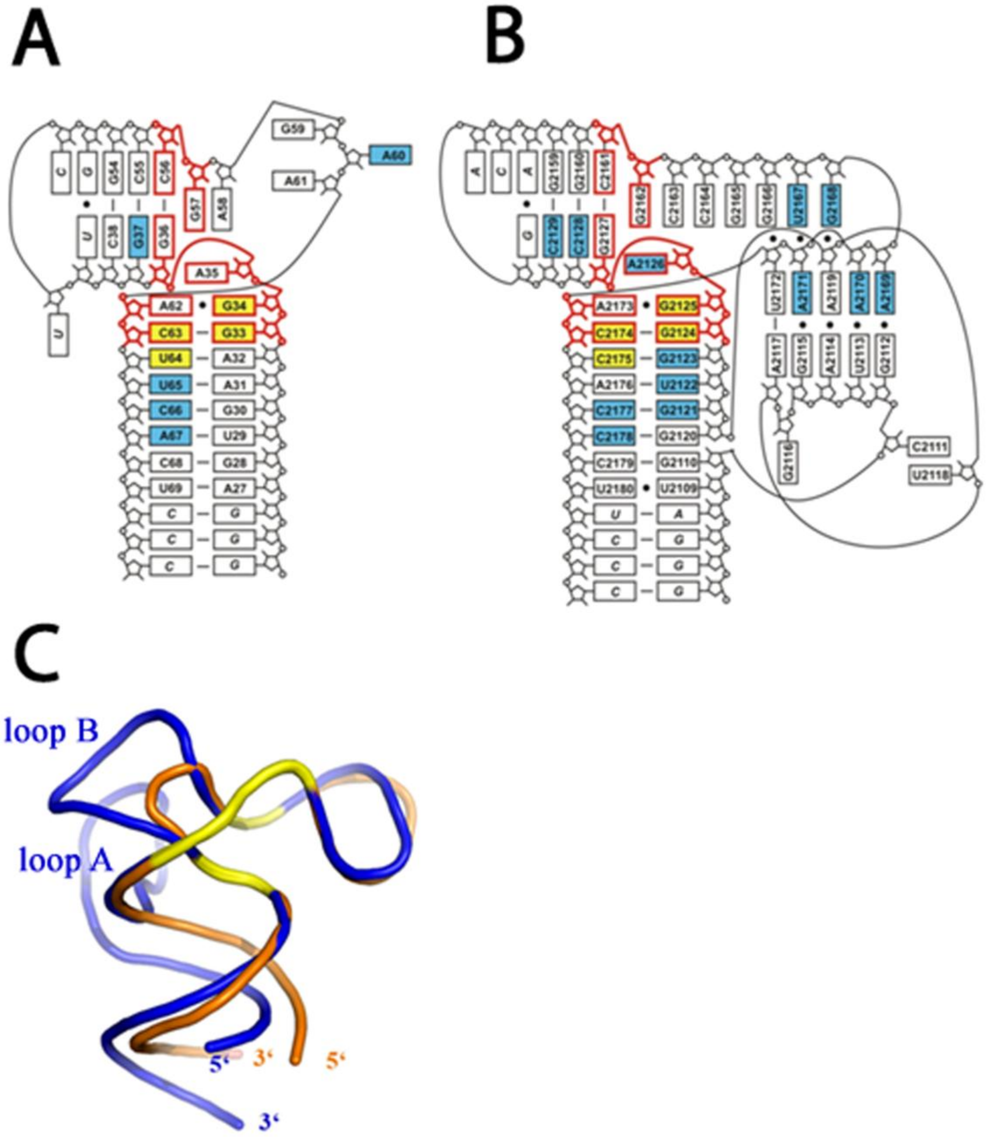

Figure 9. Spatial organization of the mRNA fragment (**A**) from the TthL1/mRNA complex and the rRNA fragment (**B**) from the SacL1/rRNA complex. Nucleotides in blue form RNA-protein hydrogen bonds; nucleotides in yellow interact with RNA-recognizing module of the protein. The structurally invariant site of RNA is marked in red. **C** – superposition of the structures of mRNA (orange) and rRNA (dark-blue).

This invariant site presents a main structural element recognized by L1 protein in ribosome and in regulation complexes. It was suggested earlier that conserved residues, making H-bonds inaccessible to solvent and forming invariant sites on the surfaces of both isolated and RNA-bound L1 proteins, play a main role in specific recognition of the RNA targets. Comparison of ribosomal and regulator complexes of L1 protein revealed five such amino acid residues (Thr40, Glu42, Thr217, Met218, Gly219). They participate in six RNA-protein H-bonds identical in both the complexes (table 1).

Table 1. L1-RNA H-bonds present in both regulationand ribosomal complexes. Numeration of the mRNA nucleotides corresponds to mRNA$_{MvaL1}$.

TthL1/mRNA				SacL1/rRNA			
H-bond L1 - mRNA			H-bond length, Å	H-bond L1 - 23S rRNA			H-bond length, Å
Thr40	OG1 - O2P	G34	2.81	Ser29	OG-O2P	G2125	2.77
Glu42	OE2 – O2’	G33	2.62	Glu31	OE2 – O2’	G2124	2.86
Thr217	OG1 – O2’	G33	2.93	Thr208	OG1 – O2’	G2124	2.84
Thr217	O – N2	G33	3.37	Thr208	O – N2	G2124	3.22
Met218	S – O2’	C63	3.45	Met209	S – O2’	C2174	3.29
Gly219	O – O2’	U64	2.83	Gly210	O – O2’	C2175	2.68

From the side of RNA, four nucleotides contribute to these H-bonds, three of them are involved in the structurally invariant site at the same time (fig. 9). Amino acid residues and nucleotides making these H-bonds are shown on fig. 10A. Apart from these residues, the strictly conserved Phe37 is also very important despite it does not make any H-bonds with RNA atoms. In all known structures of the complexes, this residue shields RNA-protein interface from solvent molecules providing stability to the complex. Conserved amino acid residues forming H-bonds inaccessible to solvent and Phe37 together build structurally stable site on the protein surface which seems to be responsible for the specific recognition of the complementary RNA surface (fig. 10B).

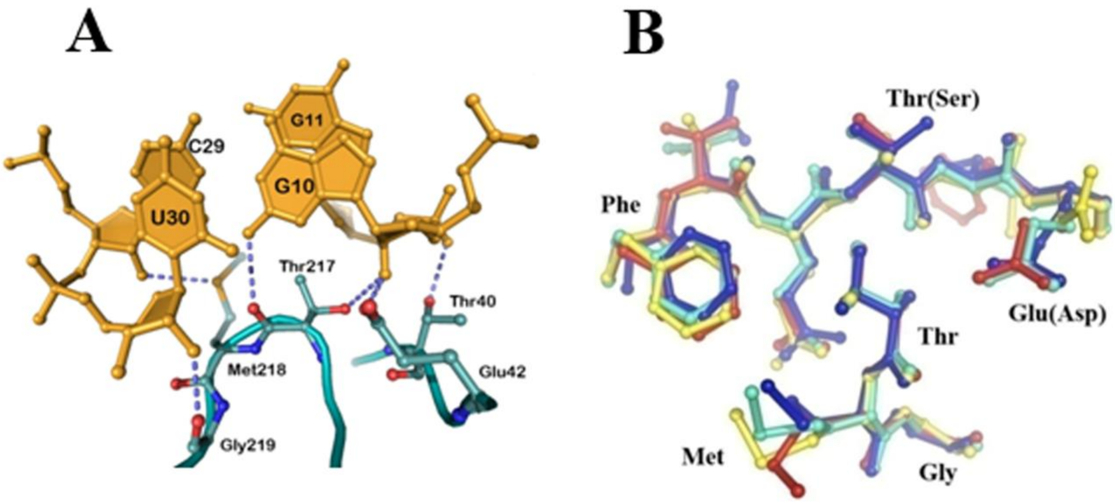

Figure 10. **A** – RNA-protein interface. The conserved amino acid residues and nucleotides are shown in blue and in orange respectively. **B** – Superposition of RNA-recognizing modules of L1 proteins from *T. thermophilus* (dark-blue), *M. thermolithotrophicus* (yellow), *M. jannaschii* (red), *S. acidocaldarius* (blue).

15 of 38 RNA nucleotides and 25 amino acid residues are involved in the interactions formed in the regulation complex. Protein TthL1 interact with mRNA mainly through domain I. The total surface of the RNA-protein contact is about 2500 Å^2 (whereas in SacL1/rRNA complex this value is about 3100 Å^2). mRNA fragment interacts with TthL1 mainly via RNA backbone, only two H-bonds are formed by the nucleotide base atoms. Those are N2 atoms of G6 and G9 (table 1). In the ribosome complex both protein domains make contacts with rRNA (fig. 11). However the number of contacts formed by domain II is much lower as compared with domain I. Domain I interacts with RNA via a slightly concave surface formed by the inner side of the β-sheet and two spatially closed loops containing residues identical in all known L1 proteins. In the regulator complex N-terminal helix α_1 of TthL1 also participates in the interactions with mRNA.

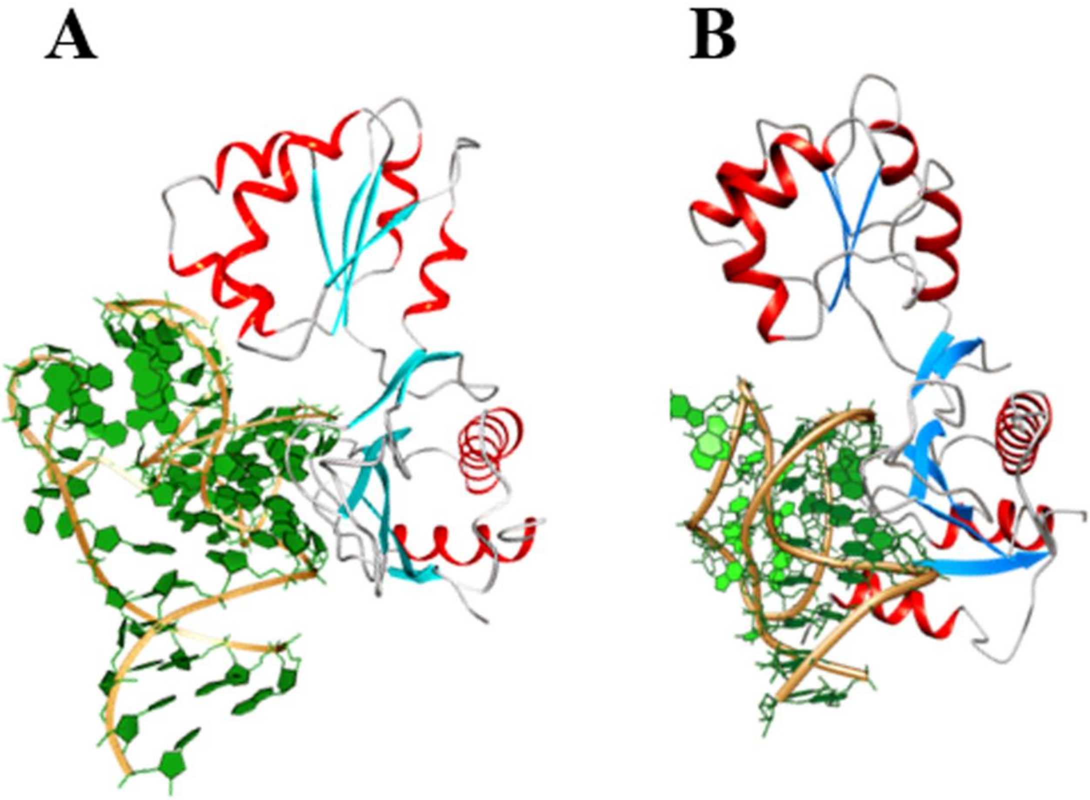

Figure 11. The complexes between protein TthL1 and 49-nt long rRNA (**A**) and 38-nt long mRNA (**B**). The proteins are shown in the same orientation.

Contact area between L1 protein and RNA is much more extended in the ribosomal complex than in the regulation one (fig. 11). This is related to the differences in the spatial organization of two RNAs. In both complexes, there is a common site of L1 binding. In mRNA it includes mainly nucleotides of the first helix and the loop. On the surface of rRNA, an analogue site is formed by helix 77 and one of the chains of helix 78. Both RNAs interact through this site with domain I of the protein. Highly conserved nucleotides of this site (G9, G10, G12, C28 in mRNA and G2124, G2125, G2127, C2174 in rRNA) make H-bonds with strictly conserved amino acid residues located in strain β_1 and the spatially closed loop β_9-β_{10} of L1 protein.

Furthermore in the L1/rRNA complex there is a second contact area formed by loop B. Nucleotides of this site interact with residues of domain II. In mRNA the connecting loop B is almost half the length compared to rRNA. This makes the RNA-protein contacts almost impossible in regulation complex. 3’- and 5’-termini nucleotides forming four base pairs in the first helix of mRNA are not involved to the interactions with L1. However when these nucleotides are absent, the RNA-protein complex cannot be formed. Hence these four nucleotide pairs are necessary for maintaining the unique spatial structure of the L1-binding site on RNA surface.

Based on the structural analysis of the complexes of L1 protein with mRNA and rRNA we proposed that domain I is sufficient for RNA recognition and making the stable complex. According to that; we proposed to determine the structure of domain I both in isolated state and in the complex with RNA. Indeed, domain I represents a much easier subject as compared with intact protein and more suitable for investigation of RNA-protein interactions. Earlier when investigated the structures of MjaL1/mRNA complex [Nevskaya *et al.* 2005], it was implemented three point mutations: Thr204Gly (analogue Thr217 in TthL1), Met205Gly (analogue Met218 in TthL1), Met205Asp and described their affinity to RNA. Analysis showed that, in the case of rRNA the affinity is lower as compared with the wild-type protein whereas for mRNA the binding was not detected for all the mutants above. Structural analysis with molecular graphics software Coot shows that these mutant proteins lose one of the conserved hydrogen bonds with RNA. These observations indicate the conserved RNA-protein H-bonds as one of the critical factors determining the recognition and the complex stability. However this suggestion is based on the maintenance of the structure near the substitution point. To check this suggestion we determined the spatial structures of the mutant forms of L1 protein with replacement of highly conserved amino acid residues and evaluated their RNA-binding properties.

1.2. Y-box Binding Protein 1

The multifunctional vertebrate Y-box-binding protein 1 (YB-1) is a member of a large protein family which contains an evolutionally ancient cold-shock domain. YB-1 is involved in a number of cellular processes including proliferation, differentiation and stress response. The YB-1 protein is performing its functions both in the cytoplasm and in the cell nucleus. It can also be secreted from cells, and, by binding to receptors on cell surface; it can activate intracellular signaling.

YB-1 is a DNA- and RNA-binding protein that has properties of a nucleic acid chaperone. It also interacts with a great variety of other proteins. By binding to nucleic acids, YB-1 is involved almost in all DNA- and mRNA-dependent processes including DNA replication and repair, transcription, splicing and mRNA translation. It packs and stabilizes mRNA as well as completes global and specific regulation of gene expression at different levels. Inasmuch as the content of YB-1 drastically increases in tumor cells, this protein is considered to be one of the most intense markers of malignant tumors.

YB-1 can translocate from the cytoplasm to the cell nucleus, and then activate transcription of genes, coding for several protective proteins, including proteins which provide multiple drug resistance to cells. An increase in the concentration of YB-1 in the cytoplasm prevents oncogenic cell transformation by the PI3K/Akt kinase signaling pathway and simultaneously it can promote transformation of differentiated epithelial cells into mesenchymal ones with higher migration activity. Thus, the YB-1 protein can also be a marker of metastasis of cancerous tumors in remote organs [Eliseeva *et al.* 2011].

1.2.1. Properties and structure-function organization of Y-box binding proteins

1.2.1.1. General properties of Y-box binding proteins

The basic peculiarities of all members of the three subfamilies of vertebrate Y-box binding proteins are as follows:

(1) a high content of alanine and proline in the N-terminal domain (hence its other name is the A/P domain);

(2) the presence of a cold shock domain (CSD);

(3) an elongated C-terminal domain containing alternating clusters of positively and negatively charged amino acid residues.

A comparison of the sequences of YB-1 from various species has demonstrated that their cold shock domains are identical by more than 90% (fig. 12A) while in the other part of the protein no essential homology is observed. Within a subfamily the homology is rather high. For example, human protein YB-1 is 96% identical to mouse protein MSY-1, 80% identical to protein FRGY1 from *X. laevis* and 67% identical to fish protein YB-1 from *Danio rerio*.

In line with the prediction of secondary structure, the N- and C-terminal domains are disordered. Probably this is the reason why there has been no success in determining the three-dimensional structure of full-size Y-box binding proteins. One hypothesis is that the conformation of these domains is fixed only upon binding to ligands and may vary in complexes with different ligands. It would be thus interesting to study three-dimensional structure using X-ray analysis of complexes of Y-box binding proteins with various associates.

The three-dimensional structure of the CspA protein which is 44% identical to YB-1 CSD was determined by X-ray crystallography and nuclear magnetic resonance (NMR) quite long ago [Newkirk *et al.* 1994, Schindelin *et al.* 1994], however the structure of the human YB-1 CSD was determined using NMR just about ten years ago [Kloks *et al.* 2002]. The 3D-structures of YB-1 and CspA CSD turned out to be very close, which could be suggested from the high homology of these proteins. The YB-1 and CspA CSDs consist of five β-strands packed antiparallel in a β-barrel, at the top and bottom of which are loops connecting the strands. CSD has the so-called consensus sequences of RNP-1 (K/N-G-F/Y-G-F-I/V) and RNP-2 (V-F-V-H-F) [Landsman 1992] (fig. 12A) thanks to which it can specifically and non-specifically bind DNA [Tafuri *et al.* 1992] and RNA [Ladomery *et al.* 1994, Bouvet *et al.* 1995].

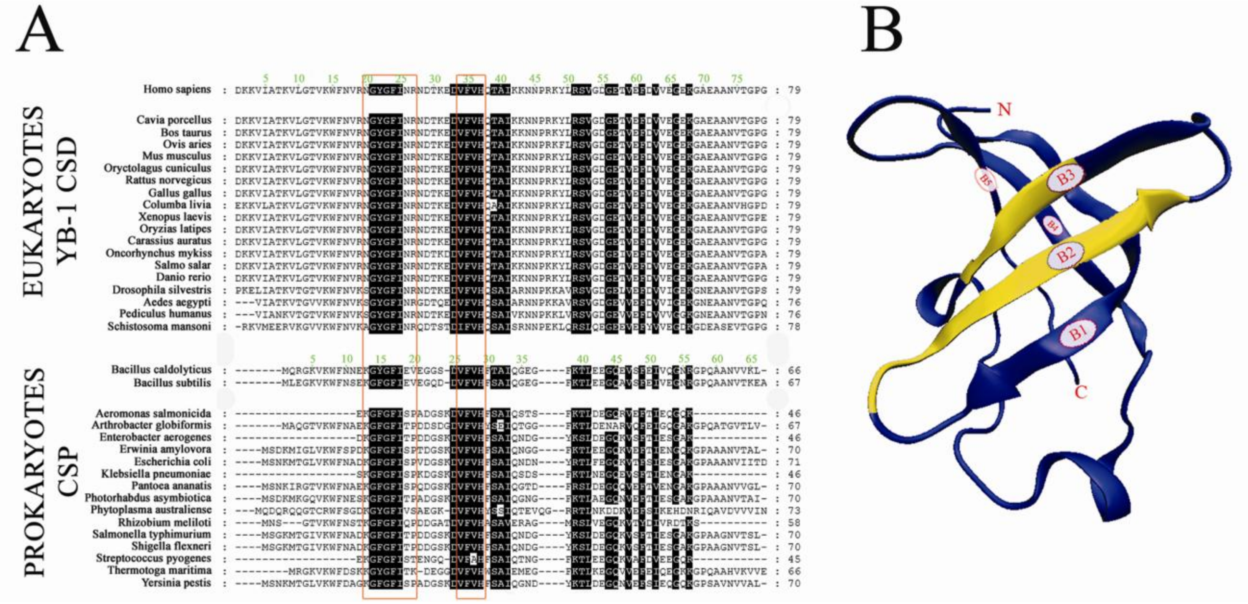

Figure 12. **A** - Sequence alignment of CSDs from eukaryotic Y-box proteins and prokaryotic CSPs. The RNA-binding motifs are boxed. Highly conserved amino acid residues (homology between CSDs and CPSs is 95%) are shown in black. **B** – 3D structure of the CSD from YB-1. RNP-1 and RNP-2 motifs are shown in yellow.

1.2.1.2. Properties of Y-Box Binding Protein 1 (YB-1)

Human YB-1 consists of 324 amino acid residues, the predominating ones being Arg (11.7%), Gly (12%), Pro (11%), and Glu (8.3%). Its molecular mass calculated from the amino acid sequence is about 35.9 kDa, but during SDS-gel electrophoresis YB-1 migrates as a protein with an apparent mass of about 50 kDa, i.e. behaves anomalously. A specific feature of YB-1 is an extremely high isoelectric point of about 9.5 [Minich *et al.* 1993].

1.2.1.3. Peculiarities of YB-1 interaction with DNA and RNA

YB-1/DNA:

YB-1 was discovered as a DNA-binding protein specifically interacting with the Y-box (5’-CTG**ATTGG**$^{C}/_{T}{}^{C}/_{T}$AA-3’) motif, however later it was clarified that it can bind to various sequences in DNA [Hasegawa *et al.* 1991, Grant *et al.* 1993, Zasedateleva *et al.* 2002]. When analyzing the interaction with oligodesoxyribonucleotides immobilized microchip, it was found that YB-1 has a greatest preference to the single-chain motif GGGG, then to one- and two-chain motifs CACC and CATC, and a lesser affinity to the sequences occurring in Y-boxes [Zasedateleva *et al.* 2002]. By binding to DNA, YB-1 essentially decreases the melting temperature of double helices, by three orders accelerates the formation of DNA double helices from mutually complementary chains in physiological conditions, and also catalyzes the exchange of complementary chains in incomplete duplexes to generate the most elongated and complete double helices, i.e. YB-1 reveals features of a DNA-chaperone [Skabkin *et al.* 2004, Zasedateleva *et al.* 2002, Skabkin *et al.* 2001]. It was shown that YB-1 has a far higher affinity for single-stranded DNA than for double-stranded one. Besides, YB-1 has an increased affinity for DNA containing apurine sites as well as to DNA damaged by cysplatin or containing unpaired bases [Hasegawa *et al.* 1991, Izumi *et al.* 2001, Lenz *et al.* 1990, Ise *et al.* 1999, Gaudreault et al. 2004]. Therefore it can be assumed that the discovered peculiarities of interaction of YB-1 with DNA control its functional activity in such processes as DNA transcription and repair.

YB-1/RNA:

The function of YB-1 in mRNA splicing, translation, stabilization, and packing is dependent on its ability to bind RNA. It was demonstrated that YB-1 has a high non-specific affinity to a wide variety of sequences, though showing inclination to some of them. When binding to homopolyribonucleotides, YB-1 had the highest affinity to poly(G) and gradually decreasing affinity to poly(U), poly(A) and poly(C). The affinity of YB-1 for globine mRNA and 16S rRNA is 4×10^{-9} M [Minich *et al.* 1993, Minich *et al.* 1992]. The specific sequence with which homologues of YB-1 *X. laevis* (FGRY1 and FRGY2) preferably interact was determined by the SELEX method. It is the hexanucleotide sequence 5’-AA**CA**U**C**-3’ called YRS (FRGY recognized sequence) [Bouvet *et al.* 1995]. Then similar sequences, to which YB-1 from different organisms specifically binds, were found using footprinting in the *YB-1* mRNA (5’-UC**CA**A/G**G**A-3’) [Skabkina *et al.* 2005], *protamine* mRNA (5’-UC**CA**U**C**A-3’) [Giorgini *et al.* 2001], *VEGF* mRNA (Vascular endothelial growth factor) (5'-AACC/UUCU-3') [Coles *et al.* 2004], Rous sarcoma virus RNA (5'-GUAC**CA**C**C**-3') [Swamynathan et al. 2000], and Dengue virus (+)PHK (5'-UC**CA**GGCA-3') [Paranjape et al. 2007]. It is seen that all of them are rich in A and C, and in addition, as shown by point mutagenesis, C in the third position, A in the fourth and C in the sixth (bold typed) are nucleotides determining a higher affinity of YB-1 to these sequences [Bouvet *et al.* 1995, Giorgini *et al.* 2001].

When YB-1 binds to RNA, it melts its secondary structure, yet the melting is incomplete (during interaction with YB-1 up to 60% of the initial secondary structure in globin mRNA is melted) [Evdokimova *et al.* 1995]. Under physiological conditions, YB-1 accelerates annealing and catalyzes the exchange of complementary RNA strands to generate the most elongated and completed duplexes, i.e. performs as an RNA chaperone [Skabkin *et al.* 2001]. It is important to note that the ratio of RNA-melting and RNA-annealing activities of YB-1 is dependent on the YB-1/RNA ratio within the complex: RNA-annealing activity is prevailing in complexes not saturated with the protein, whereas RNA-melting activity is prevailing in protein-saturated complexes [Skabkin *et al.* 2001]. It is probable that at low YB-1/mRNA ratio, YB-1 helps mRNA to adopt a conformation which facilitates recognition of mRNA by RNA-binding factors. It is thought that the C-terminal domain of YB-1 (CTD) is responsible for its non-specific binding with RNA, though according to some data, the CTD prefers pyrimidine-rich sequences [Ladomery *et al.* 1994]. The presence of the CTD provides YB-1 with a high affinity for nucleic acids. The cold shock domain accounts for the specific binding with RNA while CTD and perhaps A/P enhance and stabilize this interaction [Manival *et al.* 2001, Bouvet *et al.* 1995, Matsumoto *et al.* 1996].

1.2.1.4. Interaction of YB-1 with Proteins

All the three YB-1 domains are involved in the interaction of YB-1 with proteins (fig. 13). The A/P domain contains actin-binding regions [Ruzanov *et al.* 1999], regions of splicing factor SRp30c [Raffetseder *et al.* 2003], regions of transcription factor p53 [Okamoto *et al.* 2000], and cycline D1 regions [Khandelwal et al. 2009]. CSD can interact with kinase Akt [Sutherland *et al.* 2005] and E3 ubiquitin ligase FBX33 [Lutz *et al.* 2006]. CTD supports protein homomultimerization [Tafuri *et al.* 1992, Bouvet *et al.* 1995, Murray 1994]. This domain has binding regions of some important regulatory proteins such as hnRNP K [Shnyreva *et al.* 2000], hnRNP D [Moraes *et al.* 2003], TATA-binding protein TBP [Shnyreva *et al.* 2000], transcription factor p53 [Okamoto *et al.* 2000], YBAP1 (Y-box protein-associated acidic protein 1) [Matsumoto *et al.* 2005] and some others.

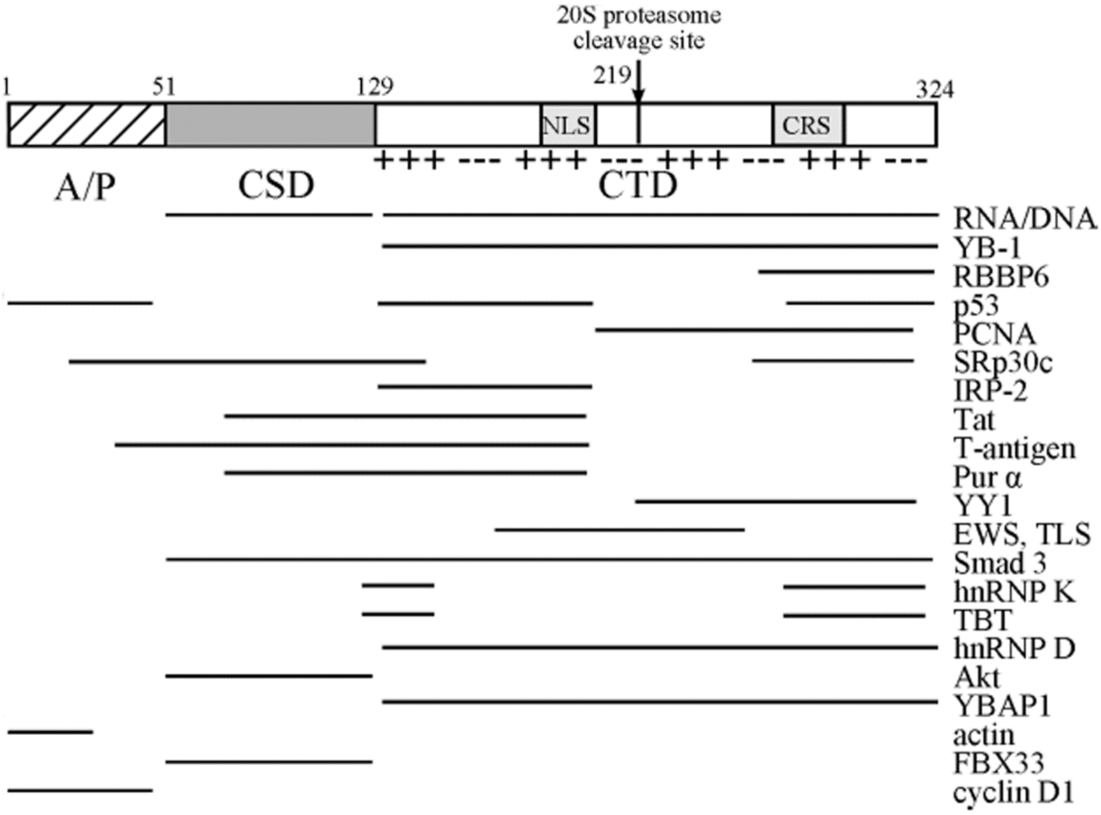

Figure 13. Scheme of binding sites of YB-1 partner proteins. Plus and minus indicate positions of clusters of positively and negatively charged amino acids. The figure is taken from [Eliseeva *et al*. 2011].

It is known that YB-1 also interacts with transcription factor Sox1 [Ohba *et al.* 2004], CARP (Cardiac ankyrin repeat protein) [Zou *et al.* 1997], tubulin [Chernov *et al.* 2008], Ankrd2 (Ankyrin repeat domain-containing protein 2) [Kojic *et al.* 2004] and some others. However regions of YB-1 molecule involved in these interactions have not been clearly determined.

1.2.1.5. Post-translational Modifications of YB-1

YB-1 is subjected to phosphorylation, ubiquitination, and probably acetylation. The total mass-spectrometry studies of phosphoproteome show that YB-1 is phosphorylated at the following amino acid residues: Ser165 and/or Ser167, Ser174 and/or Ser176, Ser313 and/or Ser314, and Tyr162 [Olsen *et al.* 2006, Molina *et al.* 2007, Dephoure *et al.* 2008, Coles *et al.* 2005]. YB-1 can be phosphorylated by kinases Erk2 and GSK3β, such phosphorylation enhancing the YB-1 binding to the promoter of the *VEGF* gene [Evdokimova *et al.* 2006]. YB-1 is phosphorylated at Ser102 *in vitro* and *in vivo* by kinase Akt [Sutherland et al. 2005, Evdokimova *et al.* 2006, Stratford *et al.* 2008] as well as pseudokinase RSK [Sorokin *et al.* 2005].

YB-1 can be completely cleaved by the 26S proteasome after ubiquitination [Lutz *et al.*

2006] and undergo limited proteolysis (processing) [Stenina *et al.* 2001] by the 20S proteasome. In the latter case, the cleavage of YB-1 into two fragments after Glu219 is ATP- and ubiquitin-independent [Sorokin *et al.* 2005].

1.2.2. Functions of YB-1 in the nucleus

When translocated into the nucleus, YB-1 is involved in the transcription of various genes, in DNA repair and replication and in pre-mRNA splicing.

1.2.2.1. The Role of YB-1 in Transcription

YB-1 can affect transcription of many genes including virus ones [Ohga *et al.* 1996]. In particular, YB-1 regulates the activity of genes whose products take part in cell division [Coles *et al.* 2002, Coles *et al.* 2005, Stenina *et al.* 2001], apoptosis [Lasham *et al.* 2000], immune response [Ansari *et al.* 1999, Sawaya *et al.* 1998], development of multiple drug resistance [Ohga *et al.* 1996, Stein *et al.* 2001, Sengupta *et al.* 2011, Stein *et al.* 2005], stress response [Li *et al.* 1997] and tumor growth [Raffetseder *et al.* 2009, Stratford *et al.* 2007]. The effect of YB-1 on transcription can be both stimulating (positive) and inhibiting (negative). It is proposed that the effect of YB-1 on transcription can result from its direct interaction with specific Y-box-containing regions in gene promoters as well as with single-stranded DNA regions that can have no Y-box sequence at all. Having formed a complex with DNA, YB-1 may attract other proteins in this complex. Moreover, it can interact with DNA only when associated with other proteins, or be involved in complexes with DNA via other proteins that have already bound to DNA. A detailed mechanism of the effect of YB-1 on transcription has not been yet established in any single case, although regulation of transcription of certain genes has been studied quite thoroughly.

Originally it was thought that YB-1 mediated transcription is dependent on the binding of YB-1 to Y-box sequence in double-stranded regions of gene promoters. But there is an ever increasing body of data on the binding of YB-1 to single-stranded sequences, including those varying from Y-boxes. For a given type of transcription regulation, the site for YB-1 binding is represented by sequences greatly asymmetrical in their distribution of purine and pyrimidine bases. This promotes the DNA transition to a single-stranded state under the action of YB-1 because the latter binds predominantly to the pyrimidine-rich sequences [Hasegawa *et al.* 1991, MacDonald et al. 1995, Izumi *et al.* 2001, Coles *et al.* 2002]. Stabilization of the single-stranded state prevents binding of the transcription factors that interact with the double-stranded DNA.

1.2.2.2. YB-1 in DNA Repair

The assumption of the involvement of YB-1 in DNA repair was made in 1991 when Lenz and Hasegawa with their colleagues identified YB-1 as a protein possessing a higher affinity to DNA-containing apurinic sites [Lenz *et al.* 1990]. This assumption is also corroborated by data showing that YB-1 has a higher affinity to DNA damaged with cisplastin or containing unpaired bases, as well as data on the ability of YB-1 to effectively melt duplexes of such a DNA [Izumi *et al.* 2001, Ise *et al.* 1999, Gaudreault *et al.* 2004]. The assumption on the involvement of YB-1 in DNA repair is also compatible with data on the ability of YB-1 to exhibit weak 3'-5'-exonuclease activity on the single-stranded DNA and weak endonuclease activity on double-stranded DNA. It is believed that nuclease activity of YB-1 is comparable to that of protein p53 and should, presumably, be strongly dependent on DNA sequence and structure [Izumi *et al.* 2001, Gaudreault *et al.* 2004, Guay et al. 2008]. It was also demonstrated that YB-1 enhances cell survival under stress conditions, where it is able to move to the nucleus and, possibly, activate transcription of some genes implicated in repair [Fukada *et al.* 2003, Ohga *et al.* 1998]. In addition, YB-1 interacts *in vivo* and *in vitro* with different proteins involved in DNA repair and can affect the activity of some of them and thus be involved practically in all types of repair. For example, YB-1 interacts with most of the base excision repair proteins (fig. 14), which suggests that it plays an most important role in this type of repair.

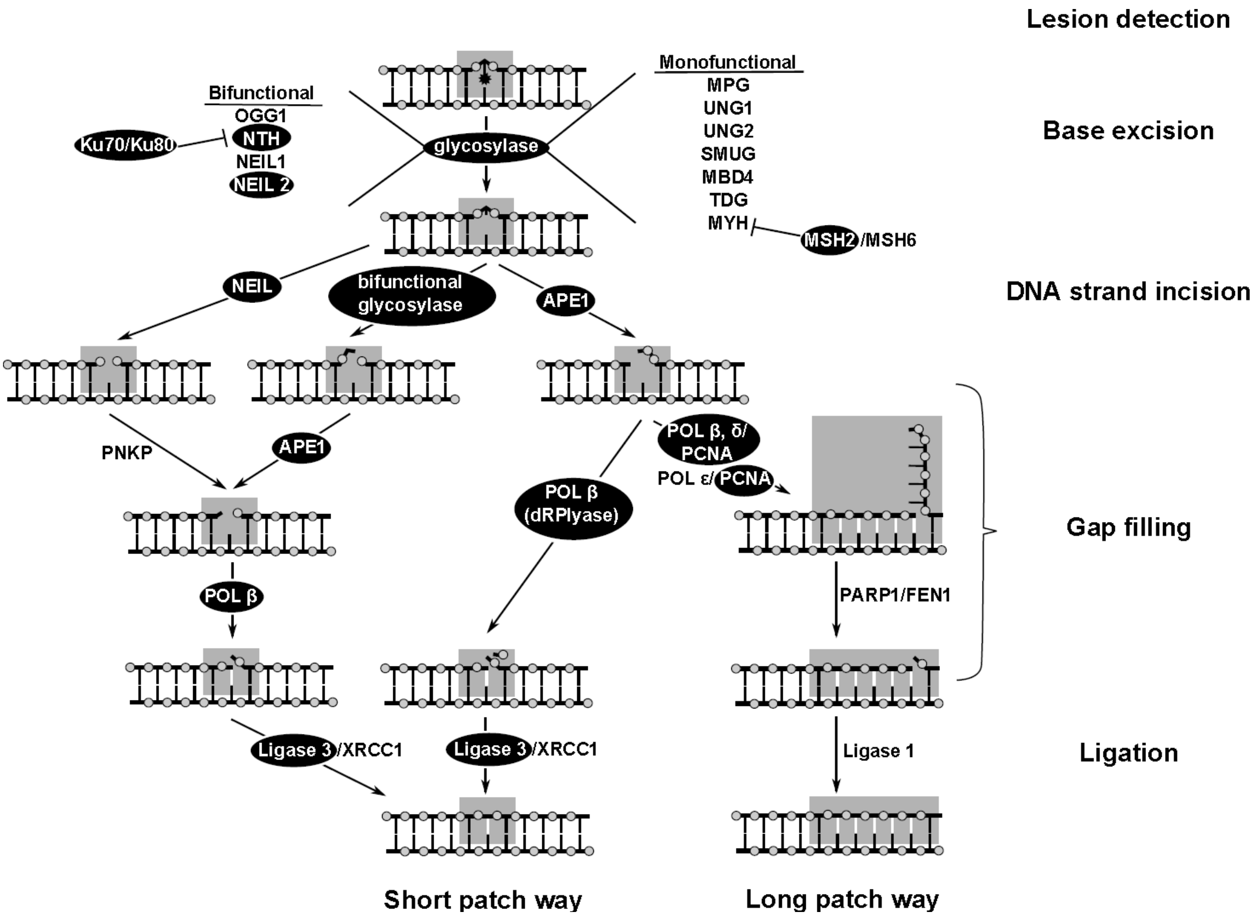

Figure 14. Scheme of base excision repair. Figure is taken from [Eliseeva *et al.* 2011].

1.2.2.3. YB-1 in DNA replication

It is believed that YB-1 participates to DNA replication. Some indirect data support this assumption. Thus, YB-1 passes into the nucleus during the cell-division cycle at the boundary of the G1/S phases [Finkbeiner *et al.* 2009]. The increase of YB-1 amount in the cell correlates with the growing level of PCNA, DNA topoisomerase IIα and DNA-polymerase α [Fukada *et al.* 2003, Finkbeiner *et al.* 2009, van Roeyen *et al.* 2005, Soop et al. 2003]. It was also shown that the lowering of YB-1 amount in cells is accompanied by termination of their proliferation [Hartmuth *et al.* 2002]. YB-1 can have a positive effect not only of the replication of cell DNA, but also on replication of the adenovirus genome [Dooley *et al.* 2006].

1.2.3. Functions of YB-1 in the cytoplasm

In the cytoplasm, YB-1 is the cardinal packing protein of mRNPs. It regulates mRNA translation, is necessary for its stability and is involved in its localization. It should be noted that all these functions of YB-1 are linked to each other.

1.2.3.1. YB-1 as a packing protein of mRNPs

As known, the entire mRNA in the cytoplasm of eukaryotic cells exists as mRNPs. These particles with unique physicochemical characteristics fall into two classes: free cytoplasmic (non-translated) and translated mRNPs of polysomes [Spirin 1964, Ovchinnikov *et al.* 1969, Perry *et al.* 1968, Ovchinnikov *et al.* 2001, Minich *et al.* 1989]. Both classes of mRNPs have a narrow density distribution and a low buoyant density value in CsCl. In free mRNPs it is about 1.39 g/cm^3 which is in line with a very high protein/RNA ratio of 3:1 (nearly 75% protein). In polysomal mRNPs the buoyant density value is somehow higher – 1.45 g/cm^3 which corresponds to a protein/RNA ratio of 2:1 (nearly 65% protein). It was observed that the buoyant density value of mRNPs (the protein/RNA ratio) does not essentially depend on the mRNA size. This indicates that the protein should be more or less uniformly distributed along the whole length of the mRNA. In spite of the very high content of protein in mRNPs of both classes, their mRNA is extremely sensitive to the action of endoribonucleases. This argues for an exposed, surface position of mRNA in the particles [Davydova *et al.* 1997].

The composition of mRNPs includes a great variety of proteins recognizing specific sequences and/or specific elements of three-dimensional structure of individual mRNAs (mostly in 5'- and 3'-untranslated regions – UTR). These proteins are responsible for selective translational control, regulation of the life-time of individual mRNAs and their specific intracellular distribution. Such proteins look like minors in protein preparations of total mRNPs. The better known are two major mRNP proteins in mammalian cells: YB-1 (or its homologues) and poly(A)-binding proteins PABP (Poly(A) binding protein) [Morel *et al.* 1971, Morel *et al.* 1973, Blobel 1972]. PABP is associated mainly with poly(A)-tails of poly(A)$^+$ mRNA polysomes. YB-1 is found in the composition of both polysomal and free mRNPs, in the latter its amount calculated per weight unit of mRNA being twice as large as that in polysomal mRNPs [Minich *et al.* 1992, Dong *et al.* 2009]. So, the mRNA transition to polysomes is accompanied by dissociation of about half of the initial amount of YB-1 and association with PABP. It was demonstrated that YB-1 is one of the mRNP proteins most strongly associated with mRNA: a large portion of it remains on mRNA at high concentrations of monovalent cations where other mRNP proteins are dissociated [Minich *et al.* 1992]. Thus, YB-1 and its homologues are universal proteins associated with all or many mRNAs existing both in untranslated and translated states.

The analysis of YB-1 complexes with mRNA *in vitro* [Skabkin *et al.* 2004] demonstrated that at a relatively low YB-1/mRNA ratio as observed for polysomal translated mRNPs, YB-1 is bound to mRNA as a monomer by two RNA-binding domains – CSD and CTD. This leads to unfolding of mRNPs which may render their mRNA accessible for interaction with translation initiation factors and ribosomes. At a high YB-1/mRNA ratio specific for free untranslated mRNPs, YB-1 molecules interact with mRNA only through CSD, whereas CTDs from different YB-1 molecules probably interact with each other resulting in the formation of large multimer YB-1 complexes consisting of approximately 15-18 YB-1 molecules. These multimers are about 20 nm in diameter and 7 nm high and pack an mRNA fragment of about 600-700 nucleotide residues on their surface. In such a complex, the mRNA ends possibly become inaccessible for interaction with proteins of the translation initiation apparatus and exonucleases [Skabkin *et al.* 2004]. As a result, by packing mRNA, YB-1 can affect its translation status and the life-time in the cell (fig. 15). Translation can thus be regulated by YB-1 positively and negatively depending on the YB-1/mRNA ratio.

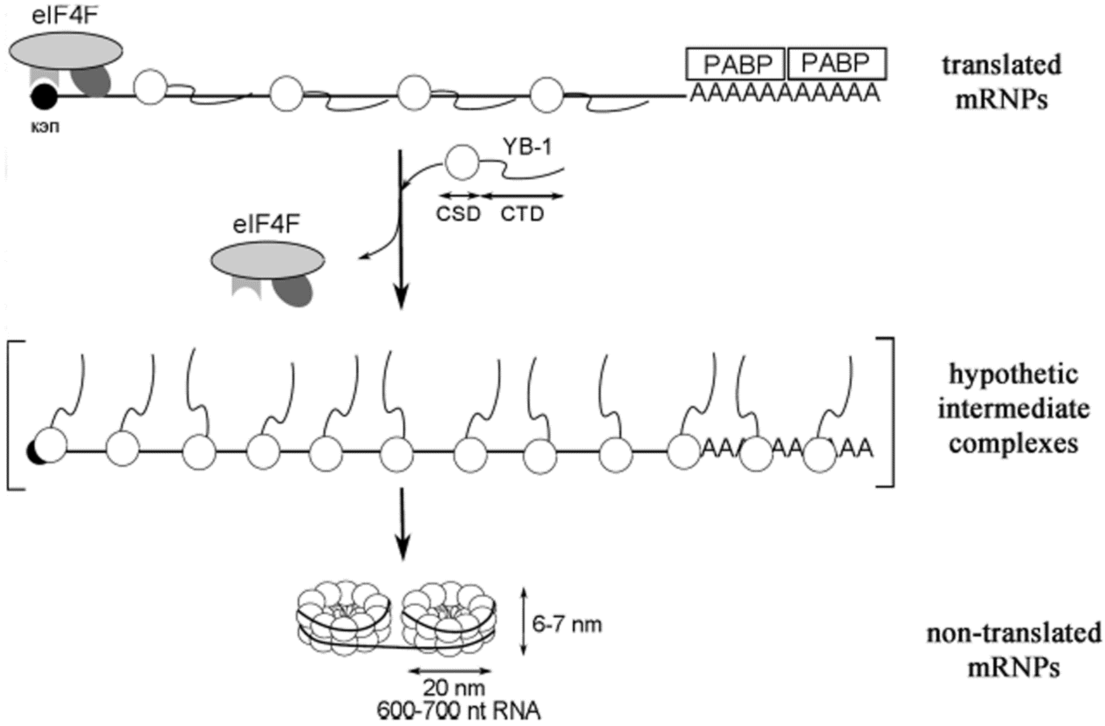

Figure 15. Scheme demonstrating peculiarities of the structure of unsaturated (translated) and saturated with protein YB-1 (untranslated) mRNA complexes with major proteins of mRNP and the initiation factor eIF4F. The transition of mRNP from the translated to the untranslated state is accompanied by a two-fold increase in the number of YB-1 molecules, and a displacement of eIF4F and PABP from mRNA as well as by compaction of the complex at the expense of multimerization of YB-1. Numerals show linear dimensions of the YB-1 multimer on mRNA in untranslated saturated complexes obtained by atomic force and electron microscopy and also the dimension of the mRNA segment packed on the surface of a multimeric protein globule. The figure is reproduced with modifications from [Evdokimova *et al.* 2001].

1.2.3.2. Effect of YB-1 on translation

The effect of YB-1 on translation depends on the YB-1/mRNA ratio. At high ratios, YB-1 inhibits translation, and on the contrary at low ratios it activates translation. YB-1 is able to inhibit the translation process both in cell-free systems and in mammalian cell culture [Minich *et al.* 1992, Minich *et al.* 1990, Bader *et al.* 2003]. Inhibition of translation is observed only at the initiation stage prior to the association of the small ribosomal subparticle with mRNA so that mRNA is found as free mRNPs. The CTD part of YB-1 is mostly responsible for this inhibition. This domain, like the full-size YB-1 displaces translation initiation factor eIF4G from mRNA [Nekrasov *et al*. 2003, Svitkin *et al.* 2009]. It was shown that the CSD of YB-1 can interact with the cap-structure or with the region adjoining it, which results in displacement of eIF4E and eIF4B [Evdokimova *et al.* 2001]. Thus, YB-1 can displace subunits of the eIF4F (eIF4G, eIF4E and eIF4B) factor from mRNA and inhibit translation at the initiation stage.

It is remarkable that when the YB-1 concentration in cell-free translation system increases, stimulation of translation remains possible by the action of PABP [Evdokimova *et al.* 1998]. This can be explained, on the one hand, by the competition between YB-1 and eIF4F for the binding to mRNA, and on the other hand, by the interaction of PABP with eIF4G enhancing the affinity of eIF4G to mRNA [Pisarev *et al.* 2002, Svitkin *et al.* 1996]. At low YB-1 concentrations, eIF4F binds effectively to mRNA in the cap-structure region and assures active translation of mRNA even at a low concentration of PABP. At high YB-1 concentrations, this protein displaces eIF4F from the complex with mRNA and inhibits translation. In these cases, an increase in the PABP amount and its interaction with eIF4G enhances the eIF4F affinity to mRNA and, as a consequence, its competitive ability to bind to mRNA. As a result, eIF4F displaces YB-1 from the complex with mRNA which leads to activation of translation under the action of PABP [Evdokimova *et al.* 1998].

As mentioned above, the involvement of YB-1 in the regulation of translation is slightly more complex. When YB-1 is removed from a lysate or when mRNA concentrations are increased, the translation process is terminated. An addition of YB-1 to such lysates leads to activation of translation [Minich *et al.* 1992, Matsumoto *et al.* 1996, Evdokimova *et al.* 1998, Jenkins *et al.* 2010]. YB-1 stimulates the protein synthesis only at the initiation stage without any effect on elongation and termination [Evdokimova *et al.* 1998, Jenkins *et al.* 2010]. So, YB-1 exerts a double action on translation: at a relatively low ratio of YB-1 to mRNA (up to the ratio observed in polysomal mRNPs), YB-1 promotes translation, whereas at mRNA saturation with the protein (as in free mRNPs) it acts as a repressor of translation. In addition to global protein synthesis regulation, YB-1 can be involved in selective regulation of translation of some mRNAs.

YB-1 is able to affect translation occurring by the cap-independent mechanism, including the mechanism of internal docking of the ribosome to a special site of mRNA enriched with secondary structure called IRES (Internal Ribosome Entry Site). Thus, it is established that YB-1 positively regulates translation of IRES-containing mRNAs of protooncogenes of the *myc* family [Evdokimova *et al.* 2009, Parker *et al.* 2004]. Moreover it was found that YB-1 is involved in translation regulation of a number of mRNA genes responsible for the epithelial mesenchymal transition (EMT) of *Snail1* mRNA [Skalweit *et al.* 2003, Chen *et al.* 2000]. This mRNA has a highly structured 5' UTR and initiation of its translation proceeds by the cap-independent mechanism at higher YB-1 concentrations than the optimal ones for translation of most of cellular mRNAs by the cap-dependent mechanism [Chen *et al.* 2000]. Hence, regulation of translation of 'weak' templates by protein YB-1 can be most likely observed for a great variety of mRNAs translated both by the cap-dependent and cap-independent mechanisms.

1.2.3.4. Stabilization of mRNA by YB-1

YB-1 and its homologues can efficiently stabilize mRNAs [Evdokimova *et al.* 2001] preventing their dissociation in cells and cell lysates due to its CSD domain. The maximum stabilization of mRNAs is achieved at a high YB-1/mRNA ratio which is associated with the mRNA release from polysomes and termination of their translation. It is notable that an efficient stabilization of mRNAs was observed both for long-lived and short-lived mRNAs such as *TNF*α mRNA [Gross *et al.* 2003]. In other words, stabilization of mRNAs caused by the action of YB-1 proceeds by an universal mechanism independent of destabilizing AU-rich elements ARE (AU-rich element) in the 3' UTR of mRNA but dependent on the YB-1/RNA ratio.

In experiments on UV cross-linking of YB-1 and mRNA with a radio labeled cap structure and in experiments on cap-sepharose affinity binding, it was demonstrated that CSD and probably the first half of the C-terminal domain of YB-1 interact with the cap structure and/or its adjacent region [Gross *et al.* 2003, Kahvejian *et al.* 2005] which leads to mRNA stabilization. At first glance, stabilization of mRNA caused by YB-1 looks paradoxical, since, as shown earlier, mRNA within mRNPs is exposed and highly sensitive to endoribonucleases [Evdokimova *et al.* 1995]. The paradox can be explained if we remember that mRNA in the cell is usually disrupted by exoribonucleases at two termini [Capowski *et al.* 2001], and assume that the structure of mRNPs in which mRNA is enriched with YB-1 is such that upon general exposure of mRNA its both termini are buried and inaccessible for the action of exoribonucleases the same as for interaction with other proteins including translation initiation factors.

In addition to the general stabilization of mRNA, YB-1 is able to selectively protect some mRNAs from degradation. So, it can enhance the stability of renin mRNA by binding specifically to its CU-rich elements in the 3'-UTR. Thus, YB-1 can stabilize mRNA in two ways. First, YB-1 forms saturated complexes with mRNAs in which 5'- and 3'-termini of molecules are buried inside mRNP globules and are inaccessible to the action of exoribonucleases. Second, YB-1 recognizes specific sequences in some mRNAs, and when in complex with other proteins it stabilizes them by an unspecified mechanism.

Thus, using the examples of the two proteins described above we studied some of the principles of macromolecular recognition in biological systems. Structural influence on the interactions of the L1 protein with RNA were explored, determined and analyzed with X-ray crystallography (see section 2.2) combined with the method of point mutagenesis at the binding site. The complexes between YB-1 CSD and nucleic acids were studied with molecular dynamics simulation (see section 2.3).

The results obtained on these protein:nucleic acids complexes allowed us to point out some critical factors responsible for the specific recognition and to evaluate their contribution to the overall stability of these complexes. These factors identified in the context of protein:NA interactions should also play an important role for other types of intermolecular interactions in water environment.

Chapter 2. Materials and methods: basic techniques for macromolecular structure investigation

During the last twenty years, we observed an exponential growth of the number of discovered structures of proteins, nucleic acids and of their complexes [http://www.pdb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100]. X-ray and NMR methods played a major role in filling the protein data bank PDB [Berman *et al.* 2000].

New theoretic methods as quantum chemistry and molecular dynamics have been developed in parallel to the progress of experimental investigations. As far as computational resources become more and more powerful, computational approaches methods become more and more accessible for exploration of the complex systems as proteins in water environment. Modern CPU clusters allow performing simulations of dynamic behavior of protein molecules with a quite good time resolution (fs). This allows receiving extra information about main conformational changes in a macromolecule, and these conformational changes could be very important for functions of the macromolecule. Thus, the combined application of experimental and theoretical methods to study macromolecular structure significantly broadens the horizons of our understanding of the biological processes at the molecular level.

2.1. Experimental methods

Biology as an experimental science always considered observations as a main source of information about living systems. Initially, the only tool for an observer was his own eyes, and the subject of observation was a whole organism. With the invention of the microscopes, biologists were able to observe the living systems at the cell level and the level of isolated cell organelles. Electron microscopy then revealed the counters of various molecular machineries and even of isolated proteins. With the appearance and development of NMR spectroscopy and X-ray diffraction methods and their application to biological macromolecules, observers received new tools to increase the resolution till isolated atom groups in a protein molecule.

NMR spectroscopy has become a standard method to determine the high-resolution structure of biomolecules including protein, nucleic acids and their complexes. Conventional NMR structural determination is based on calculation of the distance restraints obtained from proton–proton nuclear Overhauser effects (NOEs), which give approximate distances between interacting protons close in space by less than 5–6 Å [Wutrich 1986], and the torsion angles through vicinal spin couplings. These NMR parameters all give short-range structural information. Although the maximal distance observable by NOEs is limited basically to less than 5 Å, these short-range structural constraints can connect parts of a molecule that are far away in the primary sequence, but closed in 3D space. In this way, the short-range structural constraints successfully allowed accurate determination of secondary and tertiary molecular geometry of globular proteins [Wutrich 1986]. It is evident that several structures of a defined protein can fit all the NMR constrains at once. This explains why the final NMR model contains a set of structures. Some of these structures may correspond to functionally important states of the macromolecule [Mchaourab *et al.* 1997, de Groot *et al.* 1998].

In spite of the great success of conventional NMR for structural determination of globular proteins, there are significant limitations in determining multidomain protein structures due to their high molecular weight. In addition such proteins possess hinge regions between their domains. In these cases, the number of proton distances may be insufficient to define the spatial arrangement of the respective domains. Furthermore, the NOE-derived distance restraints have limited accuracy; the parts defined by the sparse NOE interactions are not fully reliable. Thus, the relative positions of distant parts of extended or modular proteins are often poorly defined. On the other hand, X-ray analyses are more suitable for large multidomain proteins and macromolecular complexes. This method, in opposite to NMR, has no theoretical limitation concerning the molecular weight of the subject. However, X-ray crystallography mainly deals with a static structure obtained from the crystal. The next section is devoted to the detailed description of this method as the main method used in this work to determine the spatial structure.

2.2. X-ray analyses as one of the most powerful methods for biological macromolecule structure determination

Obtaining the first X-ray patterns for the inorganic salt crystals [Bragg 1913] and confirmation of the wave nature of X-rays allowed drawing an analogy between X-ray diffraction and the appearance of an image by visible light. In the optic microscopy a subject is illuminated by a beam light and scatters this light. The scattered rays are then collected by the objective lenses. With recombining they give the image of the subject. In the case of X-rays, the first part of the image creation is the same. The crystal is illuminated by X-rays and the rays scattered by the electrons of the sample are registered by detector. However, due to very low refractivity of X-rays, analogue lenses for it are absent. The role of such a lens in this case is played by the mathematics apparatus (in particular Fourier transformation).

The establishment of X-ray analyses as a method for macromolecule structure determination happened in 1950-1960. During that time, the first structure of myoglobin was discovered [Kendrew *et al.* 1958]. Further the lysozim structure was determined [Blake *et al.* 1965] followed by some other enzymes: ribonuclease, chymotrypsin and carboxypeptidase. However, from the beginning of the X-ray method, it faced the question about the influence of the crystalline contacts on the protein structure. Fortunately, crystals contacts seemed not to influence noticeably the protein structure, which represents a stable equilibrium molecule conformation. This was explained by the fact that forces binding the molecules in a crystal are significantly weaker than that determining the spatial structure of proteins. It means that dramatic change of the protein conformation is almost unlikely to occur during crystallization. In some cases the same protein crystallized in different conditions and different space groups showed a very similar crystal structure. However some small conformation differences exist between the protein in the solution and in the crystal although these differences are not critical as the proteins in the crystal preserve their biological activity [Quiocho *et al.* 1972].

The popularity of X-ray analyses was associated with the intensive development of the techniques for obtaining X-ray radiation. Today three main sources of X-rays are known: i) X-ray pipe with stationary anode, ii) X-ray pipe with rotating anode and iii) synchrotrons. The principle of X-ray production in both pipe types is the same; the difference lies only in the ways of heat withdrawal, which results in different specific power produced by the anodes. This power is proportional to the intensity of the radiation source. Synchrotrons belong to the most powerful X-ray sources. Nowadays in their huge rings, electrons roll with linear velocities near the light velocity taking the energy from the transmitters working in a range of radiofrequencies. Rolling motion of the particles is provided by strong electromagnets. The charged particles emit the energy (synchrotron radiation) while changing the vectors of their speed. Accelerators of particles produce X-ray radiation range. Additional facilities (called *wigglers*) provide extra bending the electron beam enhancing the radiation intensity.

Almost all methods of X-ray diffraction data collection require monochromatic X-ray radiation. As monochromators, filters are often made of thin metal foil with an atomic number less by one of the atomic number of the target. Such filters effectively suppress K_β radiation. In monochromatic synchrotron radiation, automatically regulated systems of focusing mirrors and crystal monochromators, allow to chose any wave-length in the 0.3 to 3.0 Å range with an accuracy of the forth sign after point.

2.2.1. Basics of the method

When an X-ray beam interacts with material, scattering is realized in two ways: a) Thomson (or coherent) scattering; b) Compton (or incoherent) scattering. Thomson scattering is provoked by the next circumstance: When an X-ray radiation encounters a free electron, the influence of the varying electro-magnetic field of the X-ray wave invokes the electron to oscillate with the same frequency as the incident X-ray. An oscillating charge is a source of secondary scattered radiation which has the same wave-length as the exciting radiation but differs from it in the phase by 180°. All the rays scattered by electron have the same shift in the phase relatively to the primary beam (the scattering is coherent). In the second kind of scattering (Compton) the corpuscular nature of X-ray radiation plays a main role. An incident photon collides with a relatively weak bound electron and deviates from the initial direction of the motion with losing a part of its energy.

Compton scattering by an atom may be noticeable as compared to the Thompson scattering especially under large scattering angles. Although when a crystal diffracts X-rays, the cooperative coherent scattering by many atoms is much more significant than the sum of incoherent contributions. Therefore in X-ray crystallography, the incoherent scattering is usually neglected. Thus, a wave scattered by a crystal can be described as superposition of a huge amount of the waves, each of which is scattered by only electron. This means that scattering of X-rays depends on a number of electrons and their space distribution. While scattering on the crystals, the size and type of unit cells determine the direction of the diffracted rays that is the reflection positions on X-ray pattern. Whereas the intensities of these reflections are determined by the structure and surrounding of the molecule in the unit cell.

The subject spatial structure (more precisely the subject electron density) presents Fourier transformation of the X-ray pattern. The distribution of the electron density in a crystal can be expressed as:

$$\rho(x,y,z) = \frac{1}{V}\sum_{h=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty}\mathbf{F}(hkl)\exp(-2\pi i(hx+ky+lz)), \qquad (1)$$

where ρ is the electron density in a point (x, y, z) , *V* is the volume of the unit-cell, *h,k,l* are the integer indexes determining the positions of the diffraction maxima in the space (Miller indexes). The complex coefficient **F**(*hkl*) is called structure factor; it can be expressed as $\mathbf{F}(hkl) = F(hkl)\exp(i\varphi_{hkl})$, where *F*(*hkl*) is a module of the structure factor and φ_{hkl} is its phase. It is important to note, that for the structure factors **F**(*hkl*) and **F**(-*h*-*k*-*l*), the next condition is true: these vectors have the same length (amplitude) but opposite direction (phase angles).

This condition is called a Friedel law and the corresponding reflection pairs, (*hkl*) and (-*h*-*k*-*l*), named Friedel pairs. To be accurate, this law is true only while anomal scattering is absent, which is discussed next. This law leads to an important consequence: an X-ray pattern is always centrosymmetric even if a centre of symmetry is absent in the crystal structure.

Thus, using the expression (1), we could calculate the value of the electron density at any point (x, y, z). However at this step, we face a problem: to recreate the subject image from X-ray pattern one needs both the phase and intensity of each diffracted ray. In X-ray experiment, one can only measure the intensities of the diffracted rays, from which their amplitudes can be calculated, whereas all the information about the relative phases is being lost. Therefore it is impossible to determine the structure with using only experimental X-ray pattern because some important information is missed. Fortunately it is possible to solve this problem, using some dedicated techniques. These approaches will be reviewed in the following chapter 2.2.2.3.

2.2.2. From a crystal to a final model: main steps of 3D-structure determination

3D structure discovering is a time and labor consuming process which is conjugated with some difficulties and features linked to the nature of biological macromolecules. First of all protein crystals contain in average 50% of water. In addition, the protein molecules are much less ordered in the crystal as compared with small molecules. Protein molecules usually contain a large amount of atoms that requires a lot of diffraction measurements. Moreover the diffracted ray intensities are relatively weak as compared with the intensities of the crystals built from small molecules. These difficulties also lead to technical problems with data collection. The crystal has to be irradiated for long period, which leads in turn to its gradual breakage and distortion of the X-ray pattern [Blake *et al.* 1962, Hendrickson 1976].

A successful experiment to determine the 3D-structure of a protein includes the following steps: isolation and purification of the protein and with a high degree of homogenity; crystallization; collection of the diffraction data from the native protein crystal, as well as from its heavy atom derivatives if necessary; diffraction data reduction and determination of the crystal parameters; solution of the phase problem with one of the approaches available and calculation of the electron density maps; model building and refinement.

2.2.2.1. Isolation, purification and crystallization of macromolecules

The first stage of the macromolecule structure study by X-ray method is crystallization of the subject of interest. This step is limiting and often determines success of a whole structure exploration. In the case of proteins, successful crystallization sometimes requires large amount of material (often milligrams or even tens of milligram). Frequently the proteins of interest are present in the cells at very small concentration and to isolate and purify them directly from the cells in enough amounts is rather problematic. The optimal solution to this problem stems on cloning of the protein genes into bacterial vectors under the control of strong promoters and their further expression in an appropriate bacterial strain.

The protein used for crystallization should possess enough degree of purity that often requires several steps of purification. The procedures of protein isolation and purification begin from disruption of the cells of the overproducing strain. Ultrasound or French press techniques are commonly used to that end. Then, cell debris and large organelles are removed with low-speed and high-speed centrifugation. Then the preparation is purified using different kinds of chromatographic approaches. In most cases it is needed to reach a homogenous protein preparation with purity of at least 90%.

The protein preparation obtained is then screened for crystallization assays under different conditions (pH, temperature, different additions of both organic and inorganic compounds). When the first protein crystals are obtained, the conditions are then optimized. The goal is to improve the crystal quality, enlarge size and resolution. To successfully solve the structure in some cases (using MAD or MIR, methods see below) it requires additionally to obtain crystals with a heavy atom incorporated into the crystal. To that end, selenomethionine is frequently used where the sulfur atom is replaced by a heavier selenium atom. To produce such protein the cells of an overproducing strain is grown in minimal environment containing selenomethionine instead of methionine. All the cell proteins in this case will include the selenomethionine in place of methionine. The next step of X-ray study is diffraction data collection.

2.2.2.2. X-ray data collection and reduction

Just after receiving the first several X-ray images it is possible to evaluate most of the main crystal parameters. Among them are resolution limits, size and unit-cell parameters and a crystal space group. It is also possible to find the best strategy for data collection. One tries to collect the diffraction data set with maximal completeness and at as high as possible resolution. The nature of the space group of the crystal is one of the main factors to determine the choice for data collection strategy. Because the space group set-ups a minimal angle to be passed to collect a full data set. The higher symmetry means the fewer images to be collected to get a full set and the less time necessary for the collection process. However if a crystallographer is not limited by time, the best solution would be a data set where each reflection is measured several times, i.e. with some redundancy. This usually leads to improved data set characteristics, as it allows more precise measurement of the intensity values for all the reflections.

Data collection is best performed as a highly interactive process. Immediate data processing provides fast feedback during data collection. Visualization of the data plays an important role for the quality evaluation, too. Sometimes it may reveal some problems with the reflections, which leads to discard the collection of such data. For example if the crystal is twinned, the resulting image will be the superposition of two diffraction patterns and the reflections will be also doubled. In most of the case such data sets will be irresolvable. This is why, already at this stage one should stop collecting, which allows decreasing time spending (it is especially actual when synchrotron sources used). The most popular reduction data programs are Mosflm [Leslie 1992], HKL-3000 [Otwinowsk *et al.* 1997] and XDS [Kabsch 2001].

The image reduction is started from selection of the strongest reflections and their further indexing that is denoting the Miller indexes (*hkl*). Generally the crystal space group and unit-cell size is based on these reflections. While indexing reflections one tries to perform it in as high as possible symmetry, which minimally distorts the appropriate triclinic unit cell. The location of the weak reflections is suggested based on the chosen point symmetry and corresponding unit-cell parameters. After indexing all the reflections one starts their integration i.e. the measurement of the diffraction maxima intensity minus the average background in the neighborhood of the reflection. The integration process simultaneously refines some geometric parameters like the crystal-detector distance, beam line direction and centre position also it refines the crystal parameters and its orientation. The last step of data reduction is scaling and merging the diffraction data as well as calculation of the data set statistics.

There are several statistical functions and criteria to validate the data set reduction. The most commonly used are χ^2-criterium [Pearson 1900] and different R-factors. R-factor describes the divergence for the data. In particular XDS software outputs statistics for structural amplitudes of symmetry related reflections, R_{sym}:

$$R_{sym} = \frac{\sum_{hkl} \sum_i \left| |F_i(hkl)| - |\overline{F(hkl)}| \right|}{\sum_{hkl} \sum_i |F_i(hkl)|}, \qquad (2)$$

where a summing is performing for each reflection with indexes (*hkl*) and averaging, denoted as $|\overline{F(hkl)}|$, is realized for all i reflections symmetry related with a defined reflection.

One more important characteristic is divergence under averaging for several measurements for the reflection, R_{mrgd} [Diederichs *et al.* 1997, Weiss *et al.* 1997]:

$$R_{mrgd} = \frac{\sum_{hkl} \sum_{j=1}^{N} \left| |F_{hkl}| - |F_{hkl}(j)| \right|}{\sum_{hkl} \sum_{j=1}^{N} |F_{hkl}(j)|}, \qquad (3)$$

where $|F_{hkl}|$ is a finite amplitude value under averaging for j observations of a reflection (*hkl*).

Thus after diffraction data reduction user receive a half necessary information to determine the spatial structure of the subject studied. The missing information about the phases for each reflection (*hkl*) can be defined with one of the known approaches described below.

2.2.2.3. Phase problem and ways to solve it

The phase problem is a central problem of crystallography. It arises from the fact that in an X-ray experiment only diffraction spot intensities are registered whereas the phase information is lost. The most popular and widely used methods to solve the phase problem are molecular replacement method (MR), isomorphous replacement method (MIR) and anomal scattering method (MAD).

In the molecular replacement method to search for the phases, we use a known structure homologous (or partly homologous) to the one of interest. In general it is assumed that the higher homology between two molecules corresponds to the best model for a desired structure. The task of molecular replacement is to find the position (or positions) of the homologous molecule which could as best as possible approximate the position of the desired molecule in the unit cell. In fact the search for such positions is a six-dimensional task which is usually separated in two stages. At the first stage one searches an orthogonal transformation providing the matching for the model orientation with the desired molecule. At the second stage, one performs a search for a translation vector giving the position of the orientated model in the crystal unit cell.

Presence of several macromolecule copies in the unit cell makes it more difficult to find the right solution of molecular replacement task. Furthermore we are frequently forced to use only a part of the molecule as a start model while searching. This situation is possible for instance when the protein consists of several domains and due to their flexibility it is rather impossible to predict the location of the domains relatively to each other. In this case, firstly they use only one of the protein domains. If the desired structure is a protein-protein or a RNA-protein complex and the known structure contains only one of the complex components, the start model is also incomplete. In both the cases we face the low ratio for the start model volume to the total unit cell volume. Due to that the correct solutions could be noised and not among the list of the potential solutions.

The most frequently used methods in molecular replacement tasks are based on calculation of a Patterson function [Patterson 1935]. This function is determined by the next expression

$$P(u,v,w) = \int_V \rho(xyz)\rho(x+u, y+v, z+w)d\upsilon \quad , \qquad (4)$$

where $\rho(xyz)$ is the electron cloud density in a given point, V is the unit cell volume. The function value in a point **u**(u, v, w) can be calculated as a production of two electron density values in the points **x** and **x**+**u** (those related by the vector **u**) summed in all the unit cell volume. This function is also called an interatomic vector function, as it has maxima if (u, v, w) are the values of a vectors **u** connecting two atom centers. The Patterson function is also defined as

$$P(u,v,w) = \frac{2}{V}\sum_{h=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} F(hkl)^2 \cos 2\pi(hu + kv + lw), \qquad (5)$$

where V is the unit cell volume, h,k,l are Miler indexes, $F(hkl)$ are the structure factor amplitudes. Basically this expression provides an important property of a Patterson function: as in (5) only structure factor amplitudes are included, $P(u,v,w)$ could always be calculated from experimental diffraction data. Moreover the Patterson map can be calculated from a model with using expression (4). Thus the conclusion about matching between the model and the real structure could be made from comparison their Paterson maps calculated from the model and from experimental diffraction data.

The methods based on the use of Patterson functions are implemented in the well known programs Amore [Navaza 2001], CNS [Brunger *et al.* 1997], MolRep [Vagin *et al.* 1997]. In spite of the fact that a great amount of structures were successfully solved with these softwares, the methods applying Patterson functions face great problems while using incomplete models or presence of several macromolecule copies in the unit cell. Some times ago the maximum likelihood methods [Pannu *et al.* 1996, McCoy *et al.* 2004, Read 2001, Storoni *et al.* 2004, McCoy *et al.* 2005, McCoy 2004] were successfully applied in many areas of macromolecular crystallography. Regarding the molecular replacement, this approach was realized first in the program Beast [Read 2001] and then it was further developed in the program Phaser [McCoy *et al.* 2005]. The principles which lie in a base of the maximum likelihood method allow to manage the problems above in many cases. However, the limitation of the molecular replacement method mainly depends on the start model quality and is less dependent on the specific method realization. In most cases the molecular replacement method could be successfully applied if the standard deviation of the atomic positions between the start model and the desired structure does not significantly override 1 Å [McCoy 2004, Brunger 1997].

The isomorphous replacement method is one of the main methods to solve the phase problem *ab initio* [Green *et al.* 1954]. This method is based on the use of the effect of changing diffraction maxima (reflections) intensities when heavy atoms are incorporated in the crystal. The changing of the reflection amplitudes is used to determine the heavy atom coordinates and then to calculate the experimental phase set. A necessary condition to extract the phases by this method is absolute isomorphism of the heavy atom derivative that is the macromolecule conformation, its position and orientation in the crystal. If the isomorphous derivative is perfect, its electron density differs from the native crystal electron density only by the presence of a peak in the position taken earlier by solvent and replaced by a heavy atom. To determine the positions of a heavy atom, the Patterson maps mentioned above are applied. These maps can be calculated as a difference of the structure amplitudes for the native protein, $\mathbf{F}_{\mathrm{P}}$ and for the heavy atom derivative, $\mathbf{F}_{\mathrm{PH}}$

$$\Delta|F|_{iso} = |F_{PH}| - |F_P| \qquad (6)$$

The technique to calculate the structure factors of the heavy atom, $\mathbf{F}_{\mathrm{H}}$ is illustrated on figure 16. In the triangle ***ECB*** the side ***CE*** is determined by the expression CE=$|F_H|$cos(α_{PH}-α_{H}). In general α_{PH}-α_{H} is small, because for most reflections $|F_H| \ll |F_P|$ and $|F_{PH}|$. Therefore, CE $\cong \Delta|F|_{iso}$ and the result is that

$$\Delta|F|_{iso} = |F_H|\cos(\alpha_{PH} - \alpha_H). \qquad (7)$$

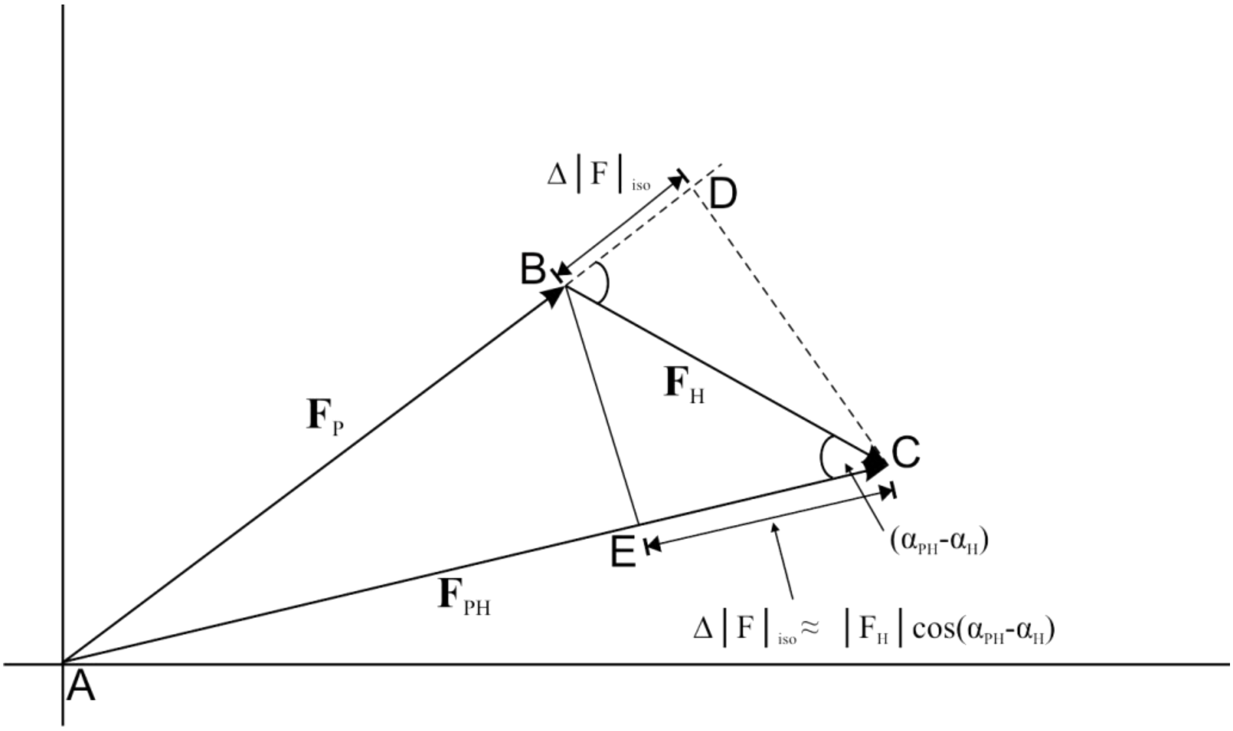
$\Delta|F|_{iso}$
D
B
$\mathbf{F}_{\mathrm{P}}$
$\mathbf{F}_{\mathrm{H}}$
C
$(\alpha_{PH}$-$\alpha_H)$
E
$\mathbf{F}_{\mathrm{PH}}$
$\Delta|F|_{iso} \approx |F_H|\cos(\alpha_{PH}$-$\alpha_H)$
A

Figure 16. The structure factor triangle for isomorphous replacement.

The result is that a Patterson summation $(\Delta|F|_{iso})^2$ as the coefficients will be in fact a Patterson summation with coefficients $\Delta|F_H|^2 cos^2(\alpha_{PH} - \alpha_H)$. Since

$$cos^2(\alpha_{PH} - \alpha_H) = \frac{1}{2} + \frac{1}{2}\cos 2(\alpha_{PH} - \alpha_H),$$

we obtain

$$\Delta|F_H|^2 cos^2(\alpha_{PH} - \alpha_H) = \frac{1}{2}|F_H|^2 + \frac{1}{2}|F_H|^2 \cos 2(\alpha_{PH} - \alpha_H). \qquad (8)$$

Because the angles α_{PH} and α_H are not correlated, the second term on the right will contribute only a noise to the Patterson map. However, the first term will give the Patterson function for the heavy atom structure on half the scale. Because a Patterson map is centrosymmetric, the choice is between two sets of heavy atom positions, which are centrosymmetrically related. It is not yet known what the correct one is, but for the moment this does not matter either of the two sets can be chosen. This problem will be discussed next.

It is worth to mention that the knowledge for the heavy atom coordinates allows calculation for their structure factors, $\mathbf{F}_H$, that is their structure amplitude and phase angle. Although after that it remains a phase ambiguity, which can be demonstrated with the Harker construction [Harker 1956]. Draw a circle with radius$|F_P|$. From the center of this circle a vector -$\mathbf{F}_H$ is drawn and then the second circle with radius $|F_{PH}|$ is added (fig. 17). The intersections of the two circles correspond to two equally probable protein phase angles, because for both points, the triangle $\mathbf{F}_{PH}$=$\mathbf{F}_P$+$\mathbf{F}_H$ closes exactly. With a second heavy atom derivative one can, in principle, distinguish between these two alternatives. However, because of errors, an exact intersection of the three circles with radii $|F_P|$, $|F_{PH1}|$ and $|F_{PH2}|$ will usually not be obtained, and some uncertainty as to the correct phase angle α_P remains. These errors are introduced in X-ray intensity data collection or by poor isomorphism. In practice more than two derivatives are used, if they are available, and, therefore, this method is called multiple isomorphous replacement (MIR).

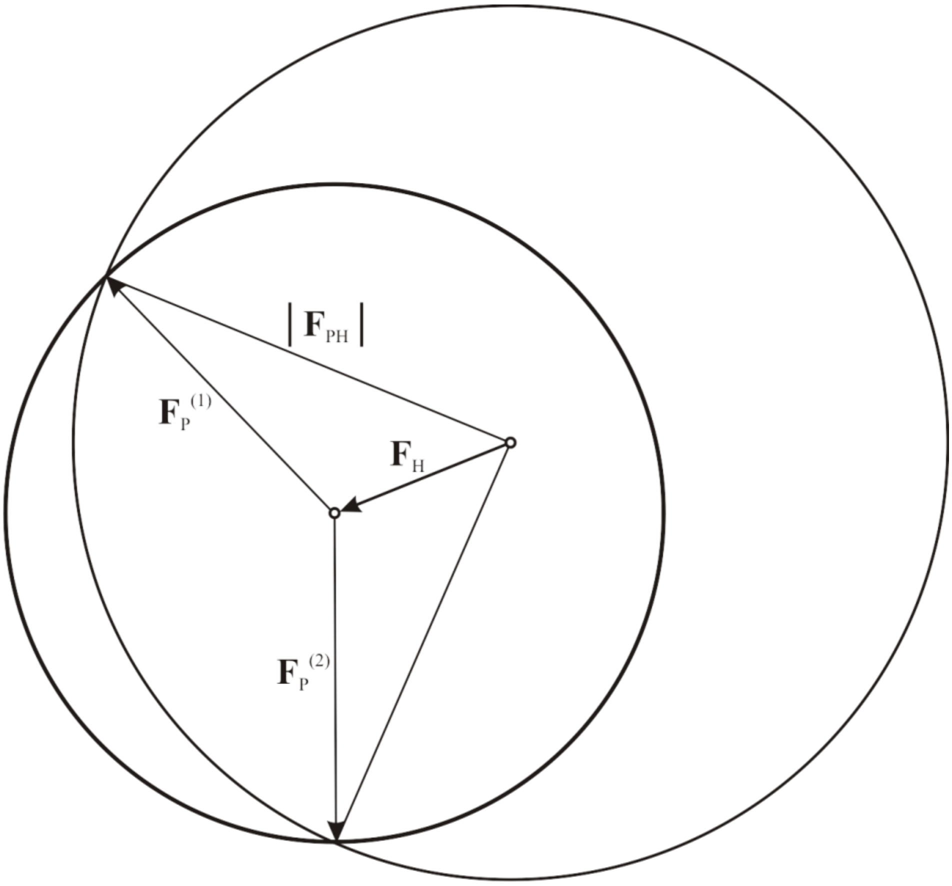

Figure 17. Harker construction for protein phase determination, In the isomorphous replacement method each heavy atom derivative gives two possibilities for the protein phase angle α_P, corresponding to the two vectors $\mathbf{F}_P^{(1)}$ и $\mathbf{F}_P^{(2)}$.

Due to the development of the techniques of synchrotron radiation and freezing, the multiwave anomal scattering method [Hendrickson *et al.* 1988, Murthy *et al.* 1988] is finding growing applications. This method also requires the incorporation of heavy atoms in the crystals, but in contrast to the preceding method it has no limitation related to isomorphism as the diffraction data are being collected from the same crystal but at different wave-lengths. One of them is taken in order the heavy atoms of the crystal may show the highest anomal scattering effect.

This effect is due to the fact that under X-ray wavelengths closed to the atom absorption edge the photon scattering cannot be considered as a scattering on free electrons. Under such wavelengths outer-shell (K-shell) electrons are excited with the next emitting lower energy photons when coming back to the K-shell from L-shell. The emitted photon has a phase angle different from the absorbed photon phase angle. This leads to the breakage of a Friedel law that is the structure factors $\mathbf{F}_{PH}(hkl)$ and $\mathbf{F}_{PH}(-h\text{-}k\text{-}l)$ for every heavy atom derivatives are not equal anymore and have different phase angles. This deference $|F|_{ano}$ can be used in a search for the heavy atom positions:

$$\Delta|F|_{ano} = \{|F_{PH}(+)| - |F_{PH}(-)|\}\frac{f'}{2f''} \qquad (9)$$

$|F_{PH}(+)|$ represents the amplitude of the structure factor for a reflection (*hkl*), and $|F_{PH}(-)|$ is the amplitude for the reflection (*-h-k-l*). $f'$and f'' are the factors of an atomic scattering; these parameters are determined in the MAD experiment. From the anomalous Patterson map, calculated with $(\Delta|F|_{ano})^2$, the location of anomals scatterers can be derived. It is worth to note that combined use of the anomal and isomorphous differences leads to less noisy Patterson maps, than if they are applied separately. Without anomal scattering, the isomorphous replacement method leads to either a correct protein structure or to mirror one. However if the resolution of the electron density map is quite high, the configuration of amino acid C_{α}-atoms can be easily checked. It should correspond to the L-configuration for the correct protein structure. In a protein molecule the right and not left α-helixes should be observed.

Despite on the increased requirements to diffraction data set quality and more complex organization of the experiment to collect them, the MAD method tolerates inisomorphism of the derivative crystals and allows solving the phase problem fast and with significant fewer expenses. This method is especially suitable while studying macromolecular complexes structures where the molecules in the crystals possess a high flexibility that often makes it almost impossible to obtain isomorphous heavy atom derivative crystals. To the shortcomings of MAD one can put a necessity to collect the data at three wave lengths that increases the time of collection and as a circumstance a risk of radiation damage of the crystal. These shortcomings are far less significant if the crystal structure is determined using the data from one crystal collected at one wave length (SAD). Although to successfully solve the structure with SAD the crystal should contain a heavy atom which provides strong enough anomal signal.

Solving the phase problem with one of the above mentioned methods allows to obtain this information and to calculate the electron density maps. After that, the electron density maps should be interpreted, and either the model of the subject should be built or the existing model should be corrected (if the molecular replacement method was applied).

2.2.2.4. Model refinement

The quality of the electron density maps obtained depends on many factors: the quality of the diffraction data, resolution, percentage of found heavy atoms in the case of MIR and MAD or percentage of the unit cell filled by the homological molecules in the case of MR.

Model refinement presents iterative auto and manual correction process and recalculation of the electron density maps with refined phases. The first step of electron density maps interpretation is the molecule main chain building (or electron density tracing).

On the next step, one adds the side chains and fits the known macromolecule amino acid (nucleotide) sequences containing in the unit cell. At the final stage the extra density is being described by adding ligands, solvent molecules, ions and by refinement of isotropic (or anisotropic) B-factors, and for a ultrahigh resolution (better than 1 Å) the model includes the hydrogen atoms, too.

Crystallographic refinement and validation of the model are the final stages of the macromolecular structure determination. With automatic correction, one can apply some different requirements and restrictions on the atomic model. They are fitting to the X-ray diffraction data, matching to the energy and stereochemical criteria and so on. To improve the atomic model some algorithms and refinement programs were developed. They include many different techniques and approaches:

- "Soft" *restraints* on the stereochemical parameters of the model [Konnert 1976, Hendrickson *et al.* 1980],
- "Hard" *constraints* on the defined stereochemical conditions [Sussman *et al.* 1977, Sussman 1985],
- Molecular mechanics force fields [Jack *et al.* 1978],
- Fast Fourier transformation methods to promote the calculations [Agarwal 1978],
- Molecular dynamics methods [Brunger *et al.* 1987, Brunger *et al.* 1989, Kuriyn *et al.* 1989, Brunger *et al.* 1990],
- Maximum likelihood method [Pannu *et al.* 1996],
- Building with molecular graphics software based on the electron density maps [Jones *et al.* 1991].

As far as the formal criteria and varying model parameters are chosen, the refinement problem consists of local minimization of a many variable function. A local character of the minimization is determined by the fact that the function has a great amount of almost equal minima and due to the experimental mistakes and scattering theory inaccuracies the deepest minimum may not correspond to the correct solution. Finally, during determination of the structure there could appear some mistakes related with the experiment as well as some mistakes in the interpretation of mediate results. With macromolecule structure refinement one can face some problems generally specific for the kind of the structures under consideration.

Macromolecule crystals have a large unit cell and it requires to collect a large amount of experimental data. At the same time these data show a low value for the signal/noise ratio. Therefore it becomes more difficult to collect the data sets at atomic resolution as compared to the structures with low molecular weight. Moreover the data set available often contains both systematic and occasional mistakes due to the crystal sizes and technical problems of the data collection.

The high solvent content in macromolecular crystals causes high thermal flexibility and dynamic disorder of the possible local conformations, decreasing maximum achievable resolution. These physical features of the macromolecular crystals cause the fast decreasing the diffracted beam intensities with increasing the scatter angles, that leads to resolution limitation (basically till 2-3 Å) at which the experimental data could be measured.

Due to the limitation on the resolution of the data set available, the ratio for the number of the structure amplitudes to the number of parameters to be refined is getting too small in order to provide the convergence and stability of the usual minimization method used during refinement, the least-root square method. At the same time the accuracy and reliability of the atomic model obtained in the crystallographic refinement process strongly depends on the level of overdetermination of the least root-square minimization task. The higher ratio for the number of experimental data to the number of refined model parameters leads to lower mistakes in the variables under minimization. In the case of low molecular structures this ratio reaches 10:1, even if the variable set includes six anisotropic thermal factors for each atom. The refinement performed with such redundancy of experimental data gives a solution with a high accuracy. However for the macromolecule crystals such overdetermination can be reached very rarely.

Some approaches increase the level of refinement overdetermination at the expense of decreasing the number of parameters refined. The most frequently used approach is to incorporate extra information, first of all data about chemical and physical regularities specific for macromolecular systems. These regularities include connectivity of the units in a whole polypeptide or nucleotide chain, stereochemical parameters specific for macromolecules, crystal packing and noncrystallographic symmetry. Thus it is common to combine the diffraction and stereochemical information while determination of the macromolecular structure.

Two methods are then used to incorporate stereochemical information into the refinement process. In the first approach “soft” *restraints* are included to the minimized function with an appropriate weight. In the second approach «strict» *constraints* are applied, and the geometry of some parts of the model is always being kept ideal, the appropriate variables are then excluded from the refinement. Although using *constrains* more effectively improves the observed/refined parameters ratio, *restrains* have their own advantages. The models with *restraints* behave more realistic and different kinds of restraints can be weighted in a different way providing more flexible refinement procedure.

Regarding the special types of the stereochemical information we now use some data about the geometry and conformation of different units forming macromolecules and about some specific stereochemical features of biopolymers. This information is taken from different sources including chemical analyses, theoretic studies, determination of crystal structures of basic chemical units and oligomers. This information contains accurate values for the bond lengths and angles, chirality of asymmetric centers, some group planarity, conformational preferences of torsion angles, Van-der-Vaal's interactions, possible hydrogen bonds, geometry of different secondary structure elements and noncrystallographic symmetry. Geometric consideration also takes place while defining the limits of thermal factor and occupancy variations.

Another way to decrease the number of the formal parameters is a switching from the individual atom characteristics to the parameters describing rigid atom groups (translation vectors, orientations and thermal factors for atom groups, or conformational rotate angles around ordinary bonds). Large secondary structure elements could be used as rigid groups, although such a switching becomes inefficient while refinement at a high resolution. One of the approaches used in the program CNS is the switching from refinement in the Cartesian coordinate system (with three coordinates x, y, z for each atom) to the coordinate system using torsion angles φ and ψ [Diamond 1971]. During such protein structure refinement the peptide planes are being kept planar and bond lengths and angles have fixed values. As far as the molecular model is built and refined it can still contain some mistakes caused by incorrect electron density interpretation, especially in the areas where this density is weak.

2.2.3. Method characteristics and model validation

Nowadays, X-ray analyses is one of the major and popular experimental methods for determination of the atomic structure of low-molecular compounds as well as large macromolecular complexes. However, X-ray experiments often can be very difficult to repeat by other groups. Thus, there is a real need to adequately validate the correctness and reliability of the structures solved and to choose the effective criteria for that [Dodson 1995]. A great amount of available structural information allows to define requirements for a newly determined structure based on the regularities found for known structures. These requirements can be divided into several categories:

-***chemical****:* bond lengths and angles, chirality, improper angles, hydrogen bonds;

-***physical****:* Van-der-Vaal's contacts, electrostatic interactions, distribution of the hydrophobic and hydrophilic amino acids;

-***protein molecule structure regularities***: distribution of the secondary structure elements and interaction between each other, distribution of the φ and ψ torsion angles in the Ramachandran plot [Ramachandran *et al.* 1963], distribution of the χ torsion angles of the side chains, configuration of the peptide groups, disposition of the water molecules and ions and their interaction with the macromolecule surface;

-***statistical****:* matching to the experimental diffraction data.

To evaluate the total correspondence of the model to the experimental data, the crystallographic R-factor given by the expression below is widely used:

$$R = \frac{\sum_h |F_0(h) - F_c(h)|}{\sum_h F_0(h)}, \qquad (10)$$

where *R* is R-factor, F_o represents the experimental values of the structure factor amplitudes, F_c are the structure factor amplitudes calculated from the model.

At each step of the structure refinement it is necessary to check a current model by the criteria above. A good model should meet all of them, though it cannot give *per se* a guarantee that the structure is absolutely correct [Briinden *et al.* 1990, Kleywegt *et al.* 1995]. A higher data set resolution is supposed to correspond to a higher accuracy of the structure determination. However, for R-factor there is no unambiguous dependence between its value and the structure reliability: a low R-factor value is necessary but not enough condition for the model accuracy. It is clear that with free moving atoms it is possible to achieve very low R-factor values, but such model will be wrong due to significant geometrical errors. On the other hand using strict geometric restrains makes it difficult to search for the conformation providing the minimal R-factors. Therefore refinement is used to find a compromise between these two criteria. Electron density maps allow a visual inspection of how good the atom model corresponds to experimental data.

Independent measurement for the correspondence between a model and the experimental structure factor amplitudes is so-called R_{free} factor value proposed by Brunger [Brunger 1992, Brunger 1993]. To calculate R_{free} value one uses an occasional chosen part of the experimental data set (usually 5-10%) which is excluded from the refinement and used only as a control. A good model should reproduce not only the experimental data that were used to build it, but also it should be in a good agreement with the independent data subset. Too high differences between R and R_{free} factor values could indicate some significant errors in the model. R_{free} is also a useful measurement to evaluate the progress and refinement results, this can be demonstrated by the next example. If adding new model parameters (adding water molecules or switching from isotropic thermal factors to anisotropic ones) leads to the equal decreasing of both R and R_{free} factor values, this procedure improves the model. But if R value decreases whereas R_{free} stays at the same level or even increases, the addition of new parameters is illegal and does not improve the model.

Nowadays, an X-ray analysis becomes more and more a standard procedure. Today no serious biological study can manage without structural information. However, the X-ray method still has a potential for development. In the next section we discuss the perspectives and main challenges for this method.

2.2.4. Next challenges and perspectives

As any method, X-ray analyses has its own limitations and areas for best application. The main limitation of the method restraining its further popularization is structure determination is only possible when the molecules can be crystallized and give more or less ordered crystals. Large complexes and molecular ensembles are the most difficult subjects to obtain crystals acceptable for data set collection. Therefore future progress in this direction is directly associated with the development of techniques for crystallization. Another important aspect concerns the further progress in the development of more powerful X-ray radiation sources.

On a technical point of view, attention should also be paid to the development of new approaches for phasing based on the usage of radiation damage of the crystals [Ravelli *et al.* 2003], implantation of atoms of noble gases (Kr, Xe) [Cohen *et al.* 2001] etc... The creation of X-ray laser would directly overcome the phase problem as the phases of the diffracted rays would be measured directly from the experiment. An important shortage which should be taken into account in X-ray analyses is the fact that this method does not allow reproducing different short-term states adopted by a protein for its function. This method shows only the time-averaged and ensemble-averaged structures. A possible way to obtain more information in this direction could be the use of combined approaches for example with the use of NMR data. Despite these difficulties, the X-ray analyses method is still one of the most powerful tools for biological macromolecule structure determination.

2.3. Theoretic methods

In parallel with experimental methods, theoretic methods are developed to study macromolecule structure. The impetuous evolution of computing engines widens the boundaries of application of such methods as quantum chemistry and molecular dynamics. Today the mechanism of biochemical reactions is being studied with quantum-mechanical approaches [Senn *et al.* 2009, Siegbahn *et al.* 2009]. Modern molecular dynamics now explore bigger and bigger protein and RNA-protein complexes [Adcock *et al.* 2006]. These approaches recommended themselves as a reliable and the simplest way for obtaining the information about the energy contributions of inter and intramolecular interactions [Tiana *et al.* 2004], energy barriers accompanying various conformational changes [Scheraga *et al.* 2007], entropic changes while various transformation [Scheraga *et al.* 2007]. In many cases the theoretic approaches are the only way to get such information.

In quantum chemistry the system could be described by a wave function, ψ depending on position and time. This wave function meets the Schrodinger differential equation. For an ensemble of interacting particles with a potential function U and masses m_k it is of the form:

$$-\frac{\hbar^2}{2}\sum_{k=1}^{N}\frac{1}{m^k}\left[\frac{\partial^2\Psi}{\partial x_k^2}+\frac{\partial^2\Psi}{\partial y_k^2}+\frac{\partial^2\Psi}{\partial z_k^2}\right]+U\Psi=i\hbar\frac{\partial\Psi}{\partial t}, \qquad (11)$$

where summation is performed for all the particles. Symbol i denotes imaginary unit and ħ is the Plank constant. The left expression acting to the wave function is called *Hamiltonian*.

If the system potential function does not depend on time (that is constant) the probability to find the particle in a space area does not also depend on time but only on the location of this area. Wave function ψ is characterized by a set of eigenvectors. This set corresponds to the set of eigenvalues of E. The physical meaning of the parameter E is that it indicates the total energy values accessible for the given system. For any Schrodinger equation corresponding to specific system there is an indefinite amount of the E values. These values can be *continuous* (for free moving particles) or *discrete*, if the particles are localized in a small space area. Discrete E values are called the energy levels. There are a lot of different theories and approximations allowing solving the Schrodinger equation for a system [Thiel 2009]. Among them, the most used are Hartree-Fock self-consistent method (MP2) [Møller *et al.* 1934] and density functional theory (DFT) [Burke *et al.* 2005].

Quantum-mechanical methods are necessary for correct description of the chemical reactions and other electron processes such as a charge transfer or an electron excitation. However these methods are basically limited to the systems up to several hundred atoms. Whereas the size and conformational complexness of biopolymer requires applying methods which can effectively calculate the behavior of several hundred thousand atoms on the time scale of several nanoseconds. This can be achieved using the highly effective techniques of molecular dynamics based on force fields. Thus, to simulate large biomolecules two methods are combined: quantum mechanics for the chemically active areas (for instance, substrates or cofactors in a enzymatic reaction) and molecular mechanics for surrounding (protein and solvent). Such approaches for studying the mechanism of reactions occurring in the biological systems could provide the necessary accuracy with reasonable computational expanses. When biological processes result from pure interactions without bond breakage or formation, there is no need to use quantum chemical methods. One of the most successfully applied approaches in this case is molecular dynamics. This method is discussed in the next section in details.

2.3.1. Molecular dynamics simulation method as one of the most powerful approaches for investigation of the conformational changes in biological systems

Molecular dynamics (MD) method is one of the most powerful approach to study the dynamic behavior of the macromolecular systems. This method allows collecting the information about the conformational changes in the observed system, many of which can be important for the function of this system [Berendsen *et al.* 2000].

Power of modern CPU clusters makes it possible to perform exploration in water surrounding describing the solvent not like an inert atmosphere (*implicit* models), as it was 10-20 years ago, but like a valuable component of the system. So-called *explicit* water surrounding models directly take into account the interactions not only between soluble molecule and waters but the interactions between water molecules, too. That noticeable increases the accuracy of MD simulation and removes some mistakes in electrostatic interaction evaluation caused by averaged dielectric conductivity in *implicit* models [Simonson 2001, Feig *et al.* 2004].

2.3.2. Basics of the method

Molecular dynamics simulation method assumes solution of the Newtonian equations of motion for the system of N interacting atoms:

$$m_i \frac{\partial^2 r_i}{\partial t^2} = F_\mathrm{i}, i = 1 \ldots N \quad , \qquad (12)$$

where strengths, F_i are the negative gradients of a potential-energy function $U(r_1,r_2,\ldots,r_N)$:

$$F_\mathrm{i} = -\frac{\partial U}{\partial r_i} \; . \qquad (13)$$

Choice of an appropriate energy function for describing the intermolecular and intramolecular interactions is critical for a successful (i.e., valid yet tractable) molecular dynamics simulation. [Ponder *et al.* 2003, Mackerell 2004]. In conventional MD simulations the energy function for nonbonded interactions tends to be a simple pairwise additive function (for computational reasons) of nuclear coordinates only. This use of a single nuclear coordinate to represent atoms is justified in terms of the Born-Oppenheimer approximation [Born *et al.* 1927].

The energy functions usually consist of a large number of parameterized terms. These parameters are mainly obtained from experimental and/or quantum mechanical studies of small molecules or fragments, and it is assumed that such parameters may be transferred to the large molecule of interest. The set of functions along with the associated set of parameters is termed a force field. A variety of force fields have been developed specifically for simulation of proteins and nucleic acids. There are notable exceptions, but it is usual for a force field to be purely additive. For instance, bond lengths are not considered to be dependent on the bond angles, and atomic partial charges are fixed in magnitude. Most of the widely applied force fields consist of several discrete terms. Each of these terms possesses a simple functional form and describes an intermolecular or intramolecular force exhibited within the system given the set of relative atomic coordinates

$$U = \sum_{bonds} Kd\,(d - d_0)^2 + \sum_{Urey-Bradly} K_{UB}(S - S_0)^2 + \sum_{angle} K_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} K_\chi\,(1 + \cos(n\chi - \chi_0)) + \sum_{impropers} K_\varphi\,(\varphi - \varphi_0)^2 + \sum_{nonbond} \left\{ \epsilon_{ij} \left[\left(\frac{R_{ij}^{min}}{r_{ij}} \right)^{12} - \left(\frac{R_{ij}^{min}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\epsilon_l r_{ij}} \right\} \qquad (14)$$

where K_d, K_{UB}, K_θ, K_χ, and K_φ are the bond length, Urey-Bradley (1-3 bond length), bond angle, dihedral angle, and improper dihedral angle force constants, respectively. Likewise, d, S, θ, χ, and φ are the bond length, Urey-Bradley (1-3 bond length), bond angle, dihedral angle, and improper dihedral angle values exhibited in the current configuration, and the zero subscript represents the reference, or equilibrium, values for each of those. These terms represent the bonded interactions. The final term in the function represents the nonbonded interactions, incorporating Coulombic and Lennard-Jones interactions. ϵ_{ij} relates to the Lennard-Jones well depth, R_{ij}^{min} is the distance at which the Lennard-Jones potential is zero, q_i is the partial atomic charge of atom i, ϵ_l is the effective dielectric constant, and r_{ij} is the distance between atoms i and j. Next rules are used to obtain the necessary Lennard-Jones parameters for each pair of different atoms: ϵ_{ij}values are the geometric mean of the ϵ_{ii} and ϵ_{jj}, values, while R_{ij}^{min} values are the arithmetic mean of the R_{ii}^{min} and R_{jj}^{min} values. Values for the atomic partial charges, q_i, are determined from a template-based scheme, with charges often modified to reproduce dielectric shielding effects (i.e., to mimic some of the effects of shielding from a high dielectric constant solvent). This ϵ_l is usually set to unity for simulations incorporating explicit solvent representations. The nonbonded terms are applied to all atoms except those attached through one or two covalent bonds. In certain specific cases, the Lennard-Jones term is adjusted for atoms connected through three covalent bonds in order to accurately reproduce experimentally observed structures. An example of such cases is the nitrogen and oxygen atoms of amides. For the purposes of MD, it is advantageous for the force field to have efficiently accessible first and second derivatives with respect to atomic position (which corresponds to the physical characteristics of atomic force and force gradients, respectively), and this is one of the more notable reasons for the very simple mathematical forms generally chosen.

The Newtonian equations of motion are being solved with numerous algorithms with a short and finite interval, Δ*t*; the temperature and pressure are being maintained at predefined levels. Given the position with respect to a single component of vector r_i, (that is the position along a single dimension, x) at a specific time, t, then the position after Δ*t*, is given by a standard Taylor series:

$$x(t+\Delta t) = x(t) + \frac{dx(t)}{dt}\Delta t + \frac{d^2x(t)}{dt^2}\frac{\Delta t^2}{2} + \cdots \qquad (15)$$

The position x(t), the velocity dx(t)/dt, and the acceleration $d^2x(t)/dt^2$ are sufficient for numerical solution to the equations of motion if some approximation to account for higher order terms in the Taylor series can be made. For this single dimension, Newton's second law describes the acceleration:

$$\frac{d^2x(t)}{dt^2} = \frac{F_x}{m}, \qquad (16)$$

where m is atomic mass, F_x is the component of the net force acting on the atom parallel to the direction of x. The simplest approach is to assume that the higher Taylor terms sum to zero, effectively truncating the Taylor expansion at the second derivative, the acceleration. In the general case, this is a very poor approximation as highlighted by consideration of Newton's third law. The net force acting in the entire system should be zero, resulting in conservation of the total energy (i.e., kinetic plus potential energies) and conservation of the total momentum. With the simple approximation suggested, significant fluctuations and drifting over time occur in the total energy of the system as a simulation progresses. A wide range of improvements to this simple approximation are used in modern molecular dynamics software [Adcock *et al.* 2006]. Numerous algorithms for integrating the equations of motion [Verlet 1967, Beeman 1976, Swope *et al.* 1982, Gear 1971] differ in accuracy and stability which mainly determined by the last term of the Taylor expansion they include. One of the commonly used algorithms, the Verlet integrator, is a fourth-order method with terms beyond Δt^4 truncated.

2.3.3. From static to dynamic: main steps of the MD simulation

MD trajectory calculation is a long-time process consisting of several steps. Each of them is discussed below.

2.3.3.1. System preparation

The earliest protein simulations considered the molecules as isolated entities, effectively in a vacuum. Representing the protein environment is however very important to describe correctly its properties. Thus, modern simulations include explicit water and neighboring protein molecules as in a crystal environment. It is now conventional to replicate the system periodically in all directions to represent an essentially infinite system. Typically, a cubic lattice is used for replication of the central cubic box (although it is possible to use any lattice available). The atoms outside the central box are simply images of the atoms simulated in that box. So-called *periodic boundary conditions* ensure that all simulated atoms are surrounded by neighboring atoms, whether those neighbors are images or not.

From a fixed amount of computation, the length of a simulation is determined by a number of factors including the cost of evaluating interactions, number of interactions that need to be evaluated at each time step, period of that time step, and number of degrees of freedom that need to be propagated. To increase the efficiency of a computer simulation, any of those four aspects might be improved. Improvements in efficiency are often obtained through freezing the fastest modes of vibration by constraining the bonds to hydrogen atoms to fixed lengths using algorithms such as SHAKE [van Gunsteren *et al.* 1977, Ryckaert *et al.* 1977], RATTLE [Andersen 1983] and LINCS [Hess *et al.* 1997]. The use of such algorithms and fixing of bond lengths involving hydrogen atoms allow the use of larger time-step (Δt) sizes without any significant amount of degradation in the quality of the trajectory (or in the accuracy of the simulation).

The simulation can be performed in different experimental conditions. But most commonly isobaric-isothermal conditions are used i.e. with constant pressure and the temperature. During a simulation at constant energy, the temperature will be observed to fluctuate due to the spontaneous interconversion of the kinetic and potential components of the total energy. The instantaneous temperature may be evaluated from the atomic velocities using

$$3k_bT = \frac{\sum_{i=1}^{N} m_i v_i^2}{N} \quad , \tag{17}$$

where k_b is Boltzmann's constant, m_i and v_i are the mass and velocity of atom *i*, respectively, and N is the total number of atoms. The atomic velocities can be rescaled to keep the temperature constant during the course of a simulation. To maintain a constant pressure during a simulation, the volume needs to be allowed to fluctuate by adjusting the dimensions of the periodic box and rescaling the atomic positions accordingly.

Thus, a start point for MD trajectory calculation is the molecule of interest placed in the solvent environment (usually water environment for biological molecules). Physical-chemical properties of the molecules of interest under physiological conditions should also be taken into account. Particularly a protein molecule is taken in a zwitterion form, that is with N- and C-terminus charged; polar residues Asp, Glu, Arg, Lys also carry proper charges and either deprotonated (Asp, Glu) or protonated (Arg, Lys). The total charge of a protein is usually neutralized by monovalent ions (Na^+, Cl^-), that is the system is totally electro neutral, which increases stability and accuracy of the simulation. However it is necessary to relax the system (by energy minimization) before assignment of the start velocities and launching the integration.

2.3.3.2. Energy minimization

Given a set of N independent variables, $r=(r_1, r_2, r_3, \ldots , r_N)$, the task is to find the values for each of these variables, termed r_{min}, for which a particular function, U(r), has its global minimum. In the case of a molecular mechanics protein model, N is typically three times the number of atoms (resulting from three degrees of freedom per atomic coordinate), r encodes the atomic coordinates (e.g., the Cartesian coordinates), and U is the potential energy as given by an equation (14). It is an extremely difficult task to locate a global minimum for such functions consisting of even ten variables. Typical biomolecular systems with as few as a hundred atoms will be described with on the order of 300 variables; thus, it is usually impossible to provably locate the global minimum. Also, while energy minimization methods may be used to efficiently refine molecular structures, they are totally inadequate for sampling conformational space. Given an unrefined molecular structure with bond angles and lengths distorted from their respective minima or with steric clashes between atoms, energy minimization methods can be very useful for correcting these flaws and are therefore routinely applied to protein systems. The most popular methods include those that use derivatives of various orders, including the first-order (i.e., utilizes first order derivatives) steepest descent [Fedoryuk 2001] and conjugate gradient methods [Hestenes *et al.* 1952] and the second-order (i.e., utilizes second-order derivatives) Newton-Raphson method [Ypma 1995].

The steepest descent method is one of several first-order iterative descent methods. All of these utilize the gradient of the potential-energy surface to guide a search path toward the nearest energy minimum. Because this corresponds to reducing the potential energy by moving atoms in response the force applied on them by the remainder of the system, this method is attractive as it may be considered to have a behavior that is physically meaningful. The algorithm of the minima search used in this method uses the next expression:

$$x(k) = x(k-1) + \lambda(k)F(k), \tag{18}$$

where the vector x represents the 3N dimensional configuration, λ(k) is a step size, and F(k) is the force vector. The step size for the first iteration is usually selected arbitrarily or by some simple heuristic. After every iteration, this step size is adjusted according to whether the overall potential energy of the system was reduced or increased by that step. If the energy increased, it is assumed that the step size was sufficiently large to jump over the local minimum along the search direction, and accordingly the step size is reduced by some multiplicative factor, typically 0.5. In the event where the energy was indeed reduced, the step size is increased by some factor, typically around 1.2. This continuous adjustment of the step size keeps it roughly appropriate for the particular curvature of the potential-energy function in the region of interest. While the steepest descent method is highly inefficient for multidimensional problems with irregular potential surfaces with multiple local minima, it is robust to locate the closest local minimum. Consequently, the global motions required to locate the global energy minimum will not be observed. Nonetheless, it is very effective in removing steric conflicts and relaxing bond lengths and bond angles to their canonical values.

The Newton-Raphson method is a popular second derivative method. The basic method relies on the assumption that, at least in the region of the minima, the potential energy is quadratically related to the individual variables $U(x_i) \approx a + bx_i + cx_i^2$, where a, b, and c are constants. This leads to first and second derivatives of

$$\frac{dU(x)}{dx} = b + 2cx \text{ and } \frac{d^2U(x)}{dx^2} = 2c \tag{19}$$

At the minimum $dU(x_{min})/dx = 0$, so x_{min} may be calculated using

$$x_{min} = \frac{-b}{2c} = \frac{\left(2cx - \frac{dU(x)}{dx}\right)}{\left(\frac{d^2U(x)}{dx^2}\right)} = x - \frac{\left(\frac{dU(x)}{dx}\right)}{\left(\frac{d^2U(x)}{dx^2}\right)} \tag{20}$$

For quadratic surfaces, no iterative searching is necessary since the exact minimum may be determined from the current configuration and the derivatives at that configuration. Unfortunately, biomolecular energy surfaces tend to be extremely nonquadratic and also contain many local minima. These characteristics render the basic Newton-Raphson method less useful. However, it has found widespread use as a method for efficiently completing the optimization performed via an alternative method. One modified form of the method, "Adopted Basis set Newton-Raphson" (ABNR), is very effective for large biomolecular systems [Brooks *et al.* 1983]. A minimized system contains the relaxed protein molecule surrounded by the solvent molecules and neutralized by the counter ions located as a rule at least 5 Å far away from each other and from the protein surface. The next step of MD simulation is the system equilibration.

2.3.3.3. System equilibration

At the system equilibration stage, the start velocities are assigned under required temperature and then a short MD trajectory is being calculated. This step is realized in different ways, in different programs. In the widely used MD package CHARMM [Brooks *et al.* 1983] the velocities are being assigned under temperature close to 0 K. Then the kinetic energy is gradually added to the system by a slow increase of temperature. The purpose of this procedure is homogenous distribution of the extra energy into all the degrees of freedom. At this step the water surrounding and the protein are being relaxed. The atoms can find and take positions closed to equilibrium. The length of this stage depends on the system size and may vary although for most of the biological systems a time of the order of 100 ps is usually enough for complete distribution of the extra energy.

Achievement of the equilibrium is checked by the energy and geometric parameters such as the total system energy, the geometric deviation from the start minimized structure, gyration radius etc... If these characteristics are stable for a long time and oscillate in a short range, the system can be considered to be equilibrated. After that, one starts the collection of data on the atomic motions in the system, that is the measure of the microscopic system properties which are associated with numerous macroscopic system properties such as heat capacity, free energy, oscillation spectrum etc...

2.3.3.4. MD-trajectory calculation

MD trajectory (i.e., the progress of simulated structure with respect to time) generally provides data only at the level of atomic positions, velocities, and single-point energies. To obtain the macroscopic properties requires the application of *statistical mechanics*, which connects microscopic simulations and macroscopic observables. Statistical mechanics provides a rigorous framework of mathematical expressions that relate the distributions and motions of atoms and molecules to macroscopic observables such as pressure, heat capacity, and free energies. Extraction of these macroscopic observables is therefore possible from the microscopic data, and one can predict, for instance, changes in the binding free energy of a particular drug candidate or the mechanisms and energetic consequences of conformational change in a particular protein. Specific aspects of biomolecular structure, kinetics, and thermodynamics that may be investigated via MD include, for example, macromolecular stability [Tiana *et al.* 2004], conformational and allosteric properties [Kim *et al.* 1994], the role of dynamics in enzyme activity [Wang *et al.* 2001, Warshel 2003], molecular recognition and the properties of complexes [Wang *et al.* 2001, Brooijmans *et al.* 2003], ion and small molecule transport [Roux

2002, Bond *et al.* 2004], protein association [Elcock 2004], protein folding [Daggett 2006, Day *et al.* 2003], and protein hydration [Pettitt *et al.* 1998]. MD, therefore, provides the opportunity to perform a variety of studies including molecular design (drug design [Wong *et al.* 2003] and protein design [Koehl *et al.* 1999]) and structure determination and refinement (X-ray [Brunger *et al.* 2002], NMR [Linge *et al.* 2003], and modeling [Fan *et al.* 2004]).

Tasks of a MD exploration produce some requirements to the quantity and quality of the information to be collected during the MD trajectory. The frequency of conformation change events in the system completely depends on the time we follow the system, that is trajectory length. For instance, studying protein folding often requires trajectory length of tens or even hundreds microsecond. Whereas a trajectory of several nanoseconds is enough to study the stability and binding energies of low molecular ligands on the protein surface.

Computational resources are one more significant aspect while choosing the length of the trajectory to be calculated, as MD simulation is a quite power-used process and requires large CPU resources. Thus, the calculation of 1ns length MD trajectory by CHARMM with one-node processor on 3gHz frequency takes about one month. Memory and disk space are also important especially for large systems. MD simulation outputs a file containing the trajectory of the system evaluation during the time predetermined. All the assigned information (atomic velocities, energies, forces and force gradients) are also being saved while a simulation. Thus, in any moment user can break the simulation and restart it then from any saved point. This is quite convenient especially when the task has been abortively interrupted caused by user independent reasons (for example, power cut). The large volume of the information saved during MD exploration also requires efficient analyses methods.

2.3.3.5. Analyses of the data obtained

An analysis of the MD trajectories begins usually with their visualization. There are some commonly used molecular graphics programs like VMD [Humphrey *et al.* 1996] and PYMOL [De Lano 2002]. It is also a common practice to analyze such geometric characteristics as gyration radius, root-square deviation from start structure, atomic fluctuations, solvent accessible surface, inter and intramolecular hydrogen bonds, secondary structure and so on. Among the most difficult types of analyses are principal component analyses [Jolliffe 1986], quasi-harmonic entropy calculation and clustering algorithms.

It is not a secret that mainly cooperative motions of different protein areas are very important and may determine their function like subtract binding and product release, regulation and allosteric behavior, as well as contractile and motor functions. In this sense separation of the functionally important motions from MD trajectory is a very urgent task. Principal component analyses (PCA) presents a very useful technique to separate such large-amplitude motions from random fluctuations along the MD trajectory [Amadei *et al.* 1993].

For any random data set where there is a significant correlation among individual variables. The first principal component represents linear combination of these variables describing maximal number of variations in these data. The second and next ones describe the variations remaining after exclusion of more principal component. Principal component analyses in application to MD trajectories [Garcia 1992] consist in diagonalising the covariant matrixes of atomic displacements relatively to the average structure. For the displacement vectors Δr_i and Δr_j of atoms i and j this matrix C_{ij} is defined by the next expression:

$$C_{ij} = \frac{\langle \Delta r_i \times \Delta r_j \rangle}{\left(\langle \Delta r_i^2 \rangle \langle \Delta r_j^2 \rangle\right)^{1/2}} \tag{21}$$

In this expression curly brackets denote the time averaging. The elements of the normalized covariant (correlation) matrix may lie in a range from -1 to +1. A positive value indicates correlation between the motions of atoms i and j, whereas a negative value means anticorrilation in their motions. This matrix is symmetric and can be diagonalized by orthogonal transformation, T that converts the matrix to diagonal one, Λ of the eigenvalues, λ_i. These eigenvalues represent total root-square fluctuation of the system along the corresponding eigenvectors. Each eigenvector is one correlated displacement of atom group in multidimensional space and eigenvalues of these vectors correspond to an amplitude of this displacement. Eigenvectors are being sorted in decrement order of their eigenvalues. Thus the first several eigenvectors describe principal components of the motion, i.e. correspond to the axis of maximal displacement in the system. All other motions including thermal fluctuations will be filtered out. The motions describing by each of the eigenvectors may be visualized with projection of the coordinates at all the time moment along the trajectory to these eigenvectors. This allows analyzing only main motions in the system which can have a biological meaning [Mongan 2004].

Quasi-harmonic entropy calculation is based on the diagonalizing the mass-weighted covariant matrixes of atomic fluctuations [Andricioaei *et al.* 2001]. 3n-6 nonzero eigenvalues λ_i (where n is a number of atoms) are related with quasi-harmonic frequencies:

$$\omega_i = \sqrt{\frac{kT}{\lambda_i},} \tag{22}$$

where T is simulation temperature, k is Boltzmann constant. These frequencies in turn give vibration entropy

$$S = k \sum_i^{3n-6} \frac{\hbar\omega_i/kT}{e^{-\hbar\omega_i/kT} - 1} - \ln\left(1 - e^{-\hbar\omega_i/kT}\right) \tag{23}$$

and quasi-harmonic vibration energy

$$E = \sum_i^{3n-6} \left(\frac{\hbar\omega_i}{2} + \frac{\hbar\omega_i}{e^{\hbar\omega_i/kT} - 1} \right). \tag{24}$$

The entropy measured like that is directly related with conformation space which was explored by the system while MD. It found a wide application from the evaluation of the entropic effect of mutations in a native protein to the determination of the relative entropy of ligands in the binding pockets [Thorpe *et al.* 2004].

2.3.4. Method characteristics and result validation

Trajectory validation is a very important aspect of the MD simulation. There are some criteria which allow revealing significant mistakes in the simulation process. Conservation of energy characteristics (potential, kinetic and their sum) is one of such tests indicating the stability of the integrators during the trajectory. To check it one usually makes a plot of these parameters at every trajectory step. Under correct working the total system energy and the temperature should be ranged without noticeable outliers.

If the force-field parameters was chosen correctly, the macromolecular structure should fluctuate around its equilibrium and not significantly differ from the start one determined by NMR or X-ray methods. For this comparison it is used the root-square superposition of each snapshot along the trajectory on the experimental (or minimized) structure. For the NMR structures the appropriate NMR spectrum are usually available [http://bmrb.wisc.edu/]; these data can also be calculated from the trajectories. Any structural characteristic (secondary structure elements, intra and intermolecular contacts, solvent accessible surface) can be used for the comparison with experimental values. For the structure validation one can also use the standard checking with such software as Procheck [Laskowski *et al.* 1993] and Whatcheck [Hooft *et al.* 1996]. These programs validate the stereochemical parameters like the bond length and angles, Ramachandran plot [Ramachandran *et al.* 1963], improper angles, Van-der-Vaal's contact and so on.

2.3.5. Next challenges and perspectives

Standard MD methods often fail to explore adequately the configuration space for the accurate evaluation of thermodynamic and kinetic properties of proteins. This is partly because such systems typically have enthalpic and entropic barriers that are significantly higher than the thermal energy at physiologically relevant temperatures. When systems are trapped in local regions of the configuration space over the time scale of a simulation, due to high free-energy barriers, they appear nonergodic [Palmer 1982]. That is, for these systems the time averages of observable characteristics do not equal the corresponding ensemble averages. The simple fact that the low-frequency motions of proteins typically correspond to the larger conformational changes, and these are often the more interesting motions, aggravates the issue. Such motions sometimes do not involve crossing of a very high energy barrier but may have a slow, diffusion character. Thus, the problem is just a matter of sampling for an inadequate length of time. Many different enhanced sampling methods have been introduced in the literature [Adcock *et al.* 2006] to overcome this problem. However, no perfect solution has been devised to date. Indeed, certain approaches are better suited to specific systems or observable characteristics than others. Future progress to resolve this issue will be of great interest.

Besides the development of improved sampling protocols, simply enhancing the efficiency of MD routines will increase its practical scope. For example, improvements to integrators might allow larger time steps to be used. Likewise, improved methods for long-range force evaluation, particularly in terms of computational parallelization, would lead to more efficient simulations. Improvements to the speed and accuracy of calculations regarding solvent will be particularly beneficial. Speed increases will be useful because a major portion of a typical simulation will concern the solvent. Accuracy improvements are important because the solvent often mediates essential aspects of protein structure, dynamics, and function. Currently, specific interactions, water shells, and long-range order are ignored in most implicit models, although some hybrid methods do seek to resolve this. In particular, the poor correlation between apolar solvation forces exhibited in explicit solvent simulations and implicit solvent simulations needs to be addressed.

Chapter 3. Experimental part

3.1. Determination of the 3D structures of L1 protein mutants by X-ray crystallography

In this work the 3D structures of the isolated TthL1 domain I (TthL1_d1) and its complex with mRNA fragment have been determined and analyzed. Additionally, the structures of some MjaL1 and TthL1 mutants with substitutions of amino acid residues involved in the RNA-recognizing module have been solved in the isolated state. For the MjaL1 protein, single substitutions were E27A, T204A, T204F and M205D and double substitution was E27A/T204A. For the TthL1 protein the substitutions included F37I, S211A, T217A, M218L as well as T217V for domain I (d1_T217V).

All the biochemical part of the work was performed by Nikonova E., Kostareve O. and Tishchenko S. in the laboratory of the structural analyses of the translation machinery (the Institute for Protein Research, Pushchino, Russia). They cloned and expressed the genes of mutant proteins in *E. coli* cells (strain BL21). To isolate and purify the proteins they used low and high-speed centrifugation followed by several chromatography essays. The purified proteins were then crystallized and the crystals suitable for data set collection were obtained. These crystals were given to me for the X-ray analyses experiments. My participation in this work started from collection of the data sets and further reduction of these data.

3.1.1. Collection and reduction of the diffraction data

To collect the diffraction data sets for each crystal of the mutant forms above we used synchrotron sources: DESY in Hamburg (lines EMBL X12 and EMBL BW7B) and SLS in Switzerland (lines PX and X06SA). All the data sets obtained have been processed with XDS software [Kabsch 2001]. Statistics of the data sets is given at the tables 2-4.

Table 2. Statistics for the diffraction data sets of TthL1 domain I in the isolated state and in complex with mRNA (values for a high resolution shell are given in brackets).

Parameters	TthL1_d1 (I)	TthL1_d1 (II)	TthL1_d1-mRNA	d1_T217V
Space group	P21	P31	P21212	P31
Unit cell parameters, Å, °	a=31.7	a=79.2	a=76.1	a=77.7
	b=45.3	b=79.2	b=144.4	b=77.7
	c=37.9	c=47.6	c=56.2	c=48.1
	α=90.0	α=90.0	α=90.0	α=90.0

	β=100.7	β=90.0	β=90.0	β=90.0
	γ=90.0	γ=120.0	γ=90.0	γ=120.0
Resolution limits, Å	15.00-2.55	15.00-2.30	20.00-2.30	15.00-2.45
	(2.60-2.55)	(2.40-2.30)	(2.40-2.30)	(2.55-2.45)
Wavelength, Å	0.950	0.950	0.842	0.900
Total amount of reflections	11175 (531)	65385 (7800)	114748 (13666)	51588 (7424)
Number of unique reflections	3392 (180)	14608 (1754)	26880 (3108)	10209 (1504)
Redundancy	3.3 (3.0)	4.6 (4.4)	4.3 (4.4)	5.1 (4.9)
Completeness, %	94.4 (89.1)	96.8 (98.4)	95.1 (94.3)	96.2 (88.5)
I/σ	8.8 (3.5)	7.9 (3.2)	12.9 (2.6)	12.4 (2.6)
R_{sym}	11.9 (35.1)	13.9 (46.1)	11.7 (42.1)	9.6 (49.7)

Table 3. Statistics for the diffraction data sets of MjaL1 mutant forms (values for a high resolution shell are given in brackets).

Parameters	E27A	T204A	E27A/T204A	T204F	M205D
Space group	P1	P1	P1	P1	C2
Unit cell parameters, Å, °	a=33.8	a=34.0	a=34.0	a=33.9	a=133.3
	b=39.0	b=38.5	b=39.1	b=38.4	b=33.7
	c=54.8	c=55.2	c=55.3	c=55.1	c=123.3
	α=83.0	α=82.3	α=83.1	α=83.0	α=90.0
	β=79.5	β=78.8	β=79.7	β=79.3	β=114.2
	γ=75.5	γ=76.7	γ=75.3	γ=76.2	γ=90.0
Resolution limits, Å	25.00-2.10	16.00-2.03	25.00-2.14	16.00-1.90	25.00-2.87
	(2.20-2.10)	(2.10-2.03)	(2.20-2.14)	(1.95-1.90)	(3.00-2.87)
Wavelength, Å	0.843	0.843	0.843	0.843	1.072
Total amount of reflections	40754 (5166)	45184 (3997)	56067 (3520)	56301 (4116)	43089 (4950)
Number of unique reflections	14635 (1874)	15709 (1407)	14609 (913)	19756 (1445)	11696 (1361)
Redundancy	2.8 (2.8)	2.9 (2.8)	3.8 (3.9)	2.8 (2.8)	3.7 (3.6)
Completness, %	94.5 (94.5)	91.5 (85.6)	92.6 (94.8)	94.8 (93.5)	97.4 (94.3)
I/σ	13.7 (3.2)	19.3 (3.8)	25.4 (5.4)	20.2 (5.6)	12.2 (3.5)
R_{sym}	6.6 (32.7)	3.8 (27.5)	3.0 (26.8)	3.9 (21.9)	8.5 (41.1)

Table 4. Statistics for the diffraction data sets of TthL1 mutant forms (values for a high resolution shell are given in brackets).

Parameters	F37I	S211A	T217A	M218L
Space group	P21212	P21212	P21212	P21212
Unit cell parameters, Å, °	a=80.8	a=80.9	a=80.9	a=74.9
	b=62.2	b=62.0	b=62.1	b=60.0

	c=43.8	c=43.7	c=43.6	c=43.9
	α=β=γ=90.0	α=β=γ=90.0	α=β=γ=90.0	α=β=γ=90.0
	20.00-1.31	20.00-1.46	20.00-1.46	15.00-2.00
	(1.40-1.31)	(1.60-1.46)	(1.60-1.46)	(2.10-2.00)
Resolution limits, Å	1.072	1.072	1.072	0.950
	363312 (58631)	270941 (60747)	268278 (59044)	48907 (6530)
Wavelength, Å	52215 (8922)	38396 (8844)	38090 (8632)	13468 (1740)
Total amount of reflections	7.0 (6.6)	7.1 (6.9)	7.0 (6.8)	3.6 (3.8)
Number of unique reflections	97.0 (93.2)	98.7 (96.5)	97.8 (94.1)	97.0 (95.1)
Redundancy	22.6 (9.1)	21.0 (7.4)	25.2 (9.4)	9.0 (4.0)
Completeness, %	5.2 (19.3)	5.0 (18.0)	4.5 (19.5)	9.7 (42.4)
I/σ				
R_{sym}				

3.1.2. Phase problem solving and structure refinement

The isolated domain I of TthL1 has been crystallized in two space groups $P2_1$ and $P3_1$ (table 2). The crystals belonging to the space group $P2_1$ difracted the X-rays to 2.55 Å resolution. The phase problem was solved by MR method with using the domain I from the full-length L1 as a start model. After that, the electron density maps were obtained that allowed building almost a complete model of the protein except for the first eight N-terminal amino acid residues. This model has been refined till R_{cryst} of 17.7% and R_{free} of 27.0% values (table 5) and deposited to the PDB [Berman *et al.* 2000] under entry 2OUM.

A detailed analysis of the diffraction data collected from the P31 domain I crystal has revealed the presence of merohedral twinning [Hurlbut *et al.* 1985] with a twin law of h+k, -k, -l and fraction of 0.41. The trials to remove this twinning with DETWIN from software CCP4 [Collaborative Computational Project, Number 4 1994] failed due to significant worsening of the data set statistics. Therefore we have decided to use twinned data for the phasing and next structure refinement. Three clear solutions were found for the molecular replacement task with Phaser [McCoy *et al.* 2005]. The specific CNS algorithms for twinned data [Brunger *et al.* 1997] were used for the structure refinement. The final model was refined till R_{cryst} of 19.3% and R_{free} of 23.3% (table 5) and contains 152 water molecules in addition to the protein molecule. The structure was deposited to PDB (entry 2OV7).

To achieve the molecular replacement task for TthL1_d1 in complex with mRNA, we used the structure of the isolated TthL1_d1 as well as the structure of the mRNA fragment which was determined before in the complex with a whole MjaL1 protein. The solution for two copies of the complex in asymmetric unit was found with Phaser software. The electron density was of a good quality and allowed building a complete model for both the complex molecules with the exception of two nucleotides on one of the mRNA chains. After several manual and auto correction cycles, the final model was refined till R_{cryst} of 20.9% and R_{free} of 27.3% (table 5). This model includes 274 amino acid residues, 94 nucleotides, 328 water molecules and 2 K^+ ions and is available in PDB under entry 2VPL.

Table 5. Refinement statistics for the TthL1 domain I in the isolated state and in complex with mRNA (values for a high resolution shell are given in brackets).

Parameters	TthL1_d1 (I)	TthL1_d1 (II)	TthL1_d1-mRNA	TthL1_d1_T217V
Resolution range, Å	15.00-2.55 (2.62-2.55)	15.00-2.30 (2.40-2.30)	15.00-2.30 (2.36-2.30)	20.00-2.45 2.60-2.45
Number of the reflections used in refinement	3356 (241)	14468 (1476)	26514 (1905)	11819 (1850)
R_{cryst}, %	17.7 (22.7)	19.3 (21.0)	20.9 (26.7)	19.6 (28.2)
R_{free}, %	27.0 (38.9)	23.3 (23.4)	27.4 (34.6)	25.3 (32.0)
Test subset size, %	5.0	5.0	5.0	5.0
Number of atoms in asymmetric unit:	1025	3228	4513	3097
protein monomers	1	3	2*	3
water molecules	23	152	328	40
An average B-factor, $Å^2$	34.3	60.7	59.3	56.4
Root-square deviation from the standard values:				
bond lengths, Å	0.010	0.006	0.006	0.001
bond angles, °	1.3	0.8	1.4	0.4
twinning:				
law	-	h+k, -k, -l	-	h+k, -k, -l
fraction	-	0.41	-	0.43

* for the structure of TthL1_d1-mRNA it is noted a number of complex monomers

All the structures of the protein L1 mutant forms in the isolated state have also been determined with the molecular replacement method using program Phaser. As a start model we used the appropriate wild type protein structure. After phasing and electron density map calculation the models were corrected manually and with automatic refinement in REFMAC [Murshudov *et al.* 1997]. The refinement statistics is given in the tables 6 and 7. The electron density maps for MjaL1 mutants were of excellent quality except for the C-terminal area (residues 213-219). The final models contain from 152 to 232 water molecules.

The model of wild type TthL1 protein did not include the eight N-terminal amino acid residues. In the case of the mutants of that protein a higher resolution allowed us to obtain models which contain residues number 4 to 228. All the models built possess good steriochemical parameters and do not contain residues located in the forbidden areas of Ramachandran plot [Ramachandran *et al.* 1963].

Table 6. Refinement statistics for the MjaL1mutants (values for a high resolution shell are given in brackets).

Parameters	E27A	T204A	E27A/T204A	T204F	M205D
Resolution range, Å	18.00-2.10 (2.15-2.10)	16.00-2.03 (2.08-2.03)	15.00-2.14 (2.19-2.14)	15.00-1.90 (1.95-1.90)	15.00-2.87 (2.94-2.87)
Number of the reflections used in refinement	14265 (1000)	15722 (1047)	13727 (965)	19158 (1389)	11125 (752)
R_{cryst}, %	15.7 (18.1)	19.4 (29.7)	19.9 (23.0)	18.3 (25.9)	21.2 (30.1)
R_{free}, %	25.5 (32.3)	25.6 (33.6)	25.2 (25.8)	25.3 (37.4)	26.4 (39.7)
Test subset size, %	5.0	5.1	5.0	5.1	5.3
Number of atoms in asymmetric unit:	1869	1855	1874	1934	3400
protein monomers	1	1	1	1	2
water molecules	164	165	153	232	-
An average B-factor, Å^2	34.4	36.5	47.4	29.1	60.9
Root-square deviation from the standard values:					
bond lengths, Å	0.009	0.010	0.009	0.007	0.008
bond angles, °	1.2	1.2	1.4	1.4	1.1

The TthL1 domain I with substitution T217V was crystallized in the space group P3_1 with the unit cell parameters similar to those of crystal form (II) for TthL1_d1. Twinning was found in these crystals, too (table 5). Presence of three monomers in the asymmetric unit suggested that they could be located similarly to the crystals of TthL1_d1 in the space group P3_1. Therefore all three monomers were used as a rigid body during the molecular replacement task solving. The refinement was performed with the Phenix software [Adams *et al.* 2002] applying the target functions for the refinement under twinning. After several cycles of manual and auto correction, the final model was refined till R_{cryst} of 19.6% and R_{free} of 25.3% (table 5).

Table 7. Refinement statistics for the TthL1mutants (values for a high resolution shell are

given in brackets).

Parameters	F37I	S211A	T217A	M218L
Resolution range, Å	15.00-1.31 (1.35-1.31)	20.00-1.46 (1.50-1.46)	15.00-1.46 (1.50-1.46)	15.00-2.00 (2.10-2.00)
Number of the reflections used in refinement	49928 (3598)	36718 (2558)	36430 (2488)	12987 (884)
R_{cryst}, %	18.1 (22.6)	13.8 (17.9)	16.4 (25.2)	17.9 (18.2)
R_{free}, %	20.6 (24.1)	19.1 (22.7)	20.2 (24.8)	24.6 (17.2)
Test subset size, %	5.1	5.0	5.0	5.1
Number of atoms in asymmetric unit:	2411	2464	2434	2123
protein monomers	1	1	1	1
water molecules	295	365	310	136
An average B-factor, Å^2	18.8	20.4	22.5	28.4
Root-square deviation from the standard values:				
bond lengths, Å	0.007	0.009	0.011	0.009
bond angles, °	1.6	1.3	1.4	1.4

3.2. Molecular dynamics simulation of YB-1 CSD and its complexes with oligonucleotides

The nucleic acid binding properties of YB-1 CSD were studied with molecular-dynamics simulation method using the Gromacs software [Hess *et al.* 2008]. The NMR structure of isolated CSD was used to create start models of the complexes.

3.2.1. System preparation

To create the start models we used the crystal structures of the complexes between CSP (a structural homolog of YB-1 CSD) and dT6 (PDB-entries 2HAX, 2ES2) as well as the NMR structure of YB-1 CSD (PDB-entry 1H95). The high sequence and structural homology of nucleic acid recognizing motifs between these two proteins (rmsd = 0.54 Å for 12 C_{α}-atoms, fig. 18) suggests a very similar mode for the nucleic acid binding. However it should be noted that there are some significant differences between these two binding sites. Gln38 of human YB-1 is replaced by Phe30 in the bacterial proteins. This Phe30 also participates in stacking interactions with nucleotide bases. Moreover, bacterial CSPs contain additional Phe38 (fig. 18) which is also stacked to the nucleotide bases leading to a much higher affinity of the bacterial proteins compared to isolated CSD of YB-1. However, the bacterial complexes can provide a quite reasonable start point to create the required models. Assuming that contacts are conserved between CSD and nucleotides at positions 2, 3 and 4 (fig. 18) we superposed the structure of human CSD into the CSP/dT6 complexes using the strongly conserved residues. The bacterial numeration of nucleotides bound to CSP has been used in this work. The side chains of His29 (His37 in CSD), Phe27 (Phe35 in CSD) and Phe17 (Phe24) are referred, respectively, as binding sites 2, 3 and 4 of the protein and appropriate bases of oligonucleotides are referred as N2, N3, N4 (fig. 18). We kept the local conformation of the oligonucleotide backbone during the substitutions of the nucleotide bases. One extra nucleotide has been added to both 5' and 3'-terminii and the total length of oligonucleotides in models was nine bases.

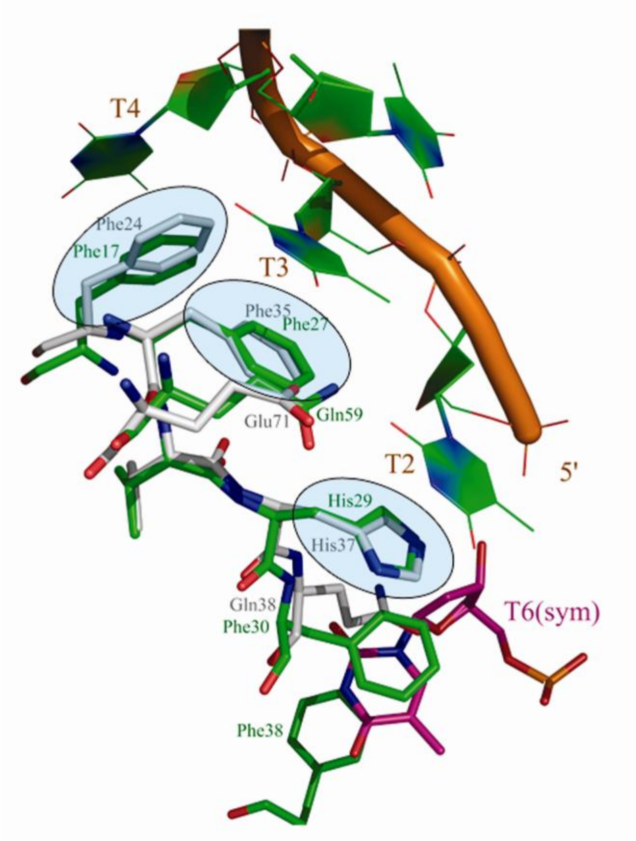

Figure 18. **S**uperposition of YB-1 CSD (gray) and bacterial CSP (green) nucleic acid binding site. Nucleotide binding sites are circled. The complex between bacterial CSP (green) and oligonucleotide dT6 (orange) corresponds to experimental structure. The nucleotide shown in magenta represents the symmetrically related molecule of the complex and shows the additional nucleotide binding site formed by Phe30 and Phe38.

MD simulations have been performed with the Gromacs 4.5 software [Hess *et al.* 2008] using Charmm27 force-fields [MacKerell *et al.* 2004, MacKerell *et al.* 1998]. The start models of the complexes were put to the orthogonal water box of the TIP3 type [Jorgensen *et al.* 1983]. The sizes of the box were calculated as follows: to the minimal and maximal coordinates of the molecules in all three directions we added 15 Å. As a result we obtained a system solvated by at least five water layers. The acid and basic residues were considered to be charged so that Glu and Asp residues had a COO$^-$-group, Arg and Lys residues were fully protonated and carried a charge +1. The phosphate groups of oligonucleotides were deprotonated and carried a -1 charge. The total charge of the systems was neutralized by the addition of seven sodium ions. For every model, the periodic boundary conditions were used. Before MD simulation the systems were minimized.

3.2.2. Energy minimization

Energy minimization was performed with 2000 steps of the steepest descend minimization [Fedoryuk 2001]. Van-der-Vaal's interactions were calculated for atom pairs located not beyond 14 Ǻ from each other. For electrostatic interaction the particle mesh Ewald method (PME) [Ewald 1921] was used with a cut-off of 10 Ǻ. The goal of the minimization was to remove atom clashes and unfavorable Van-der-Vaal's contacts. The achievement of a local minimum was controlled visually with the potential energy plots (fig. 19).

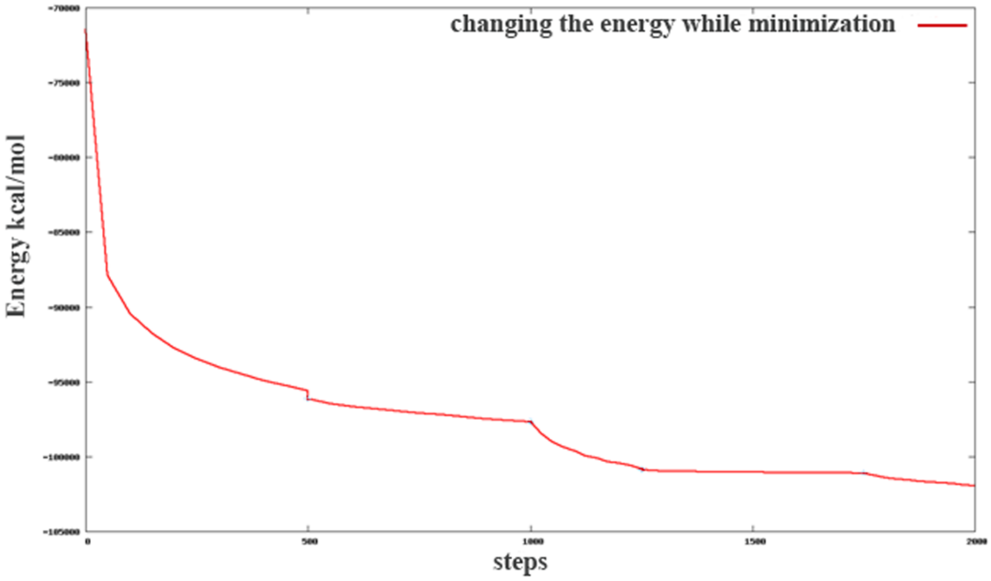

Figure 19. The potential energy minimization profile was controlled before launching MD simulation.

The next step of the simulation was system equilibration.

3.2.3. System equilibration

To equilibrate the system, we used a 50 ps length simulation protocol with maintaining the pressure and temperature with Berendsen algorithm [Berendsen *et al.* 1984]. It allowed the water molecules and bad contacts to be relaxed. The achievement of the equilibrium was controlled by the rmsd plots. Also we checked the energy stability of the system during this process. These initial steps were followed by constant pressure-temperature (CPT) dynamics simulations.

3.2.4. Calculation the MD trajectory

For the system of interest we calculated MD trajectories of 10 ns length using 1 fs integration time steps. Bond-length constrains were used only for water molecules. The long-term nonbonded interactions were evaluated using a cut-off of 14 Å. The temperature was controlled with the velocity rescaling method [Bussi *et al.* 2007] and pressure with the Berendsen algorithm [Berendsen *et al.* 1984]. Each 0.5 ps, the coordinates were saved and then analyzed. To validate all the trajectories, the NMR data available in BioMagResBank [Doreleijers *et al.* 2009] (entry number 4147) were used. The comparison between experimental and calculated with Sparta software [Shen *et al.* 2007] chemical shifts for C_α and C_β atoms is presented on figure 20.

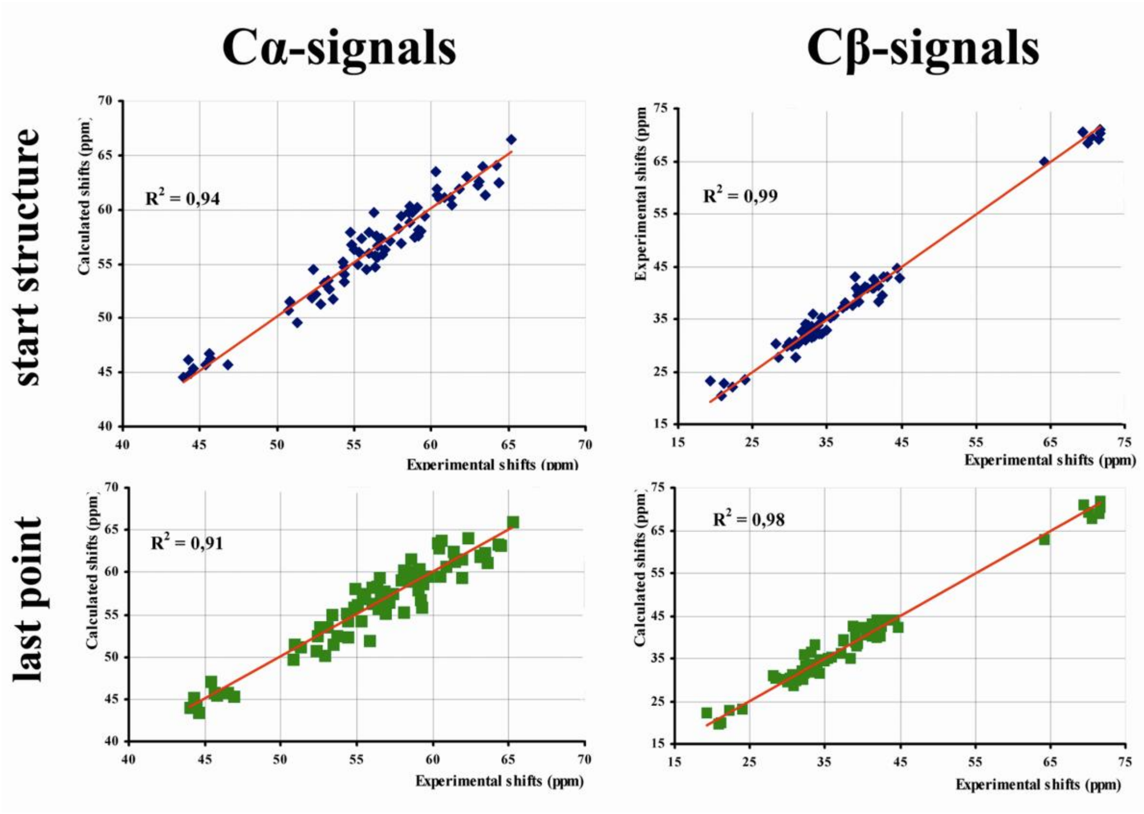

Figure 20. Correlation plots between experimental and calculated NMR chemical shifts obtained from different structures for $^{13}C_\alpha$ (left) and $^{13}C_\beta$ (right) atoms of CSD of human YB-1: initial structure available in PDB (top), the last point structure after simulation (bottom). Prediction of the chemical shifts was realized using the Sparta software [Shen *et al.* 2007].

The stereochemical parameters of the models at the different stages of the trajectories were also checked with the Procheck packages [Laskowski *et al.* 1993]. The Ramachandran plot example is shown on figure 21.

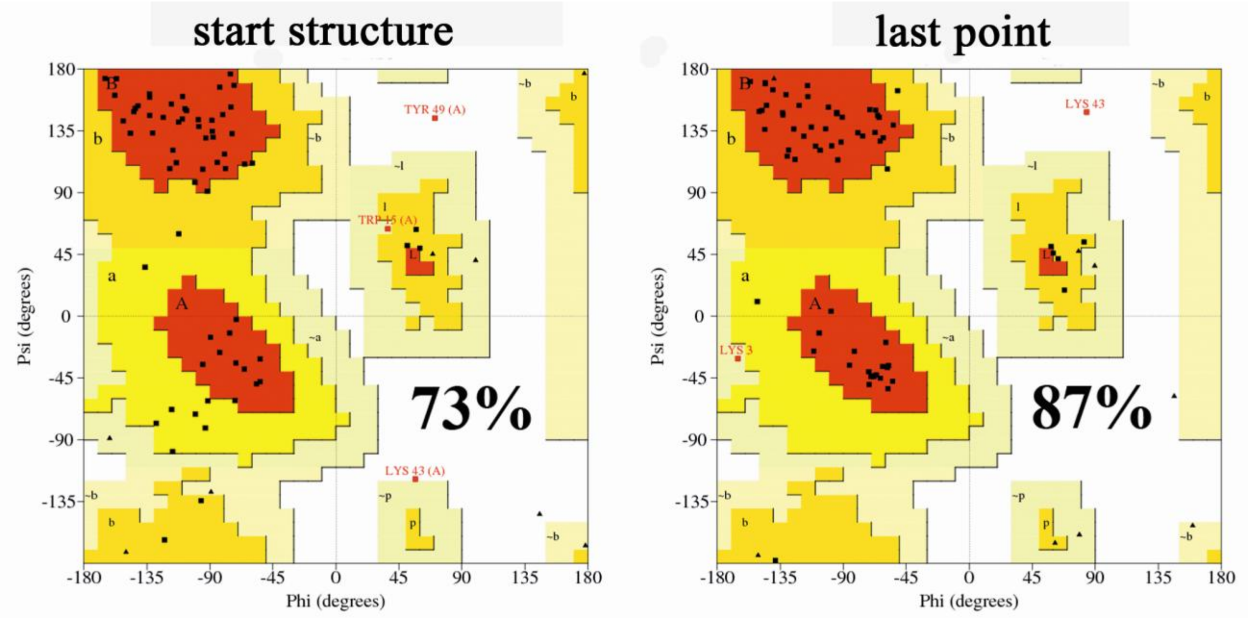

Figure 21. The Ramachandran plots of CSD of human YB-1 allow us to check the stereochemical quality of the models at different stages during trajectories: initial structure (left) and the structure after simulation (right). The percentage of the aminoacid residues localized in the most preferable regions of the diagram is shown for each model.

3.2.5. Evaluation of interactions between CSD and oligonucleotides

The binding properties could be described in terms of polar and nonpolar interactions. The first type includes H-bonds and electrostatic interactions between the nucleic acid (NA) and the protein. The second one represents Van-der-Vaal's contacts and stacking of aromatic groups. In the present work, we evaluated by simulation the protein-NA and NA-NA (H-bonds as well as stacking interactions) formed by amino acid residues and different nucleotides at positions 2 to 4 (fig. 18).

H-bonds were evaluated as follows. During the trajectory, at each snapshot we measured the distance between the appropriate H-bond donor and acceptor. We considered an H-bond as detectable if this distance was below 3.5 Å. For every H-bond we calculated the so-called occupancy (Occ). It represents the percentage of snapshots along the trajectory where a considered H-bond is found:

$$Occ\ (\%) = \frac{N}{T} \times 100, \tag{25}$$

where N – the number of snapshots with the detectable H-bond, T – the total number of snapshots. We took into account only H-bonds with occupancy over 10%. Another important parameter is accessibility of the H-bond to the solvent molecules (SA, Å^2). The H-bond less accessible to the solvent contributes more to the stability of the complex [Lim *et al.* 2006]. Thus, the relative strength of H-bond depends on its occupancy and its accessibility to the water molecules. In this work we calculated the relative strength of the H-bonds as the product of its occupancy and its relative accessibility

$$\frac{SA_0 - SA}{SA_0} \times \frac{Occ}{100}. \qquad (26)$$

Here SA_0 corresponds to the maximal possible solvent accessibility (empirically found as about 50 Å^2 from our simulations). Thus, the relative strength could vary from 0 (fully accessible to solvent H-bond) to 1 (fully inaccessible to solvent H-bond with occupancy 100%).

Concerning the stacking interactions, the average distance between geometric centers of both the base ring and the aromatic residue during the trajectories was measured (fig. 22). Two more determinants were used to properly describe stacking interactions. The first one is an angle between the planes of the aromatic rings stacked. The second parameter is a correlation in the movement of two vectors perpendicular to these planes (fig. 22). High correlation value indicates that the two rings move together along the simulation. Well stacked rings show correlation coefficient over 0.4 with angle less than 30 degrees and an average distance of about 4.2-4.5 Å along the trajectory.

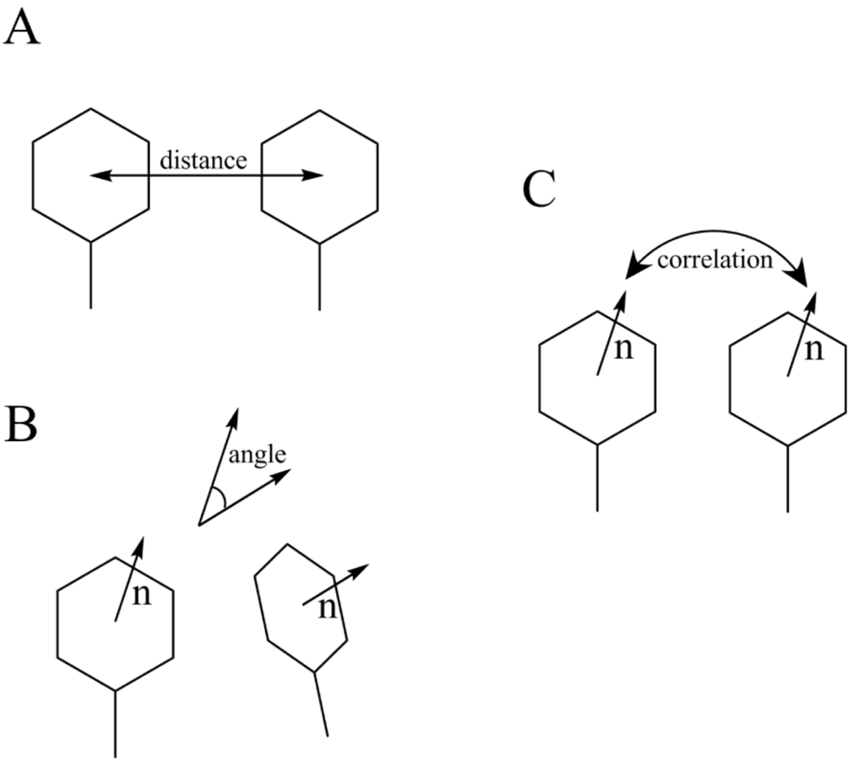

Figure 22. Structural determinants used to evaluate stacking interactions: **A** – average distance between geometrical centers of stacked rings during trajectory; **B** – average angle between planes of stacked rings; **C** – average correlation in the movement of normals to the planes.

For all the complexes the density of the protein-NA contacts as a function of time was evaluated as well. The plots represent the number of nucleotide atoms that are closer than a certain distance to the center of CSD mass. As the different nucleotides have a different number of atoms this plots are normalized by dividing by the number of the nucleotide atoms. Thus, the 100% value indicates that all the nucleotide atoms are within a determined radius from CSD mass centre.

Chapter 4. Results and discussion

4.1. The crystal structure analyses of L1 protein

The crystal structures determined in this work at a high resolution allow a detailed analysis of the structural changes that occurred as a result of amino acid substitutions. It also reveals the possible reasons supporting a decreased affinity of the protein for the coincident RNAs. The results described in this section were obtained at the IPR and published in the 5 papers that appear in annexe I.

4.1.1 Analyses of the structure of the TthL1 isolated domain I and its complex with mRNA

In TthL1, the flexible site connecting the two domains includes the amino acid residues 66-71 and 159-160. In the opened conformation, two glycines (Gly67 and Gly159) are located 4 Å apart from each other that allows constructing a peptide bond between them without significant changing the main chain course. Therefore the positions of these two glycines are the best choice when removing the second domain. Thus the TthL1 domain I contains the residues 1-67 and 159-228 (fig. 23). As in the case of TthL1 protein, N-terminal α-helix is highly flexible and cannot be located at the electron density maps of both crystal form $P2_1$ and $P3_1$.

Comparison of the domain I in the whole TthL1 and in the isolated state shows that the overall 3D structure is perfectly preserved (fig. 24). Only the flexible loops involved to the crystal contacts show some differences. The superposition of these domain structures yields a rmsd value, calculated on C_α-atoms, of around 0.9 Å (without the flexible loops this value falls to 0.5 Å). The competition binding experiments [Tishchenko *et al.* 2007] show that the TthL1 domain I and the intact protein interact with mRNA with a similar affinity. At the same time the affinity of TthL1 domain I for the ribosomal RNA is lower than that of the whole protein. This confirms preceding suggestions about the role of the protein domain II to increase stability of the TthL1/rRNA (*ribosomal*) complex relatively to the TthL1/mRNA (*regulator*) complex, which is necessary for the feedback translation regulation.

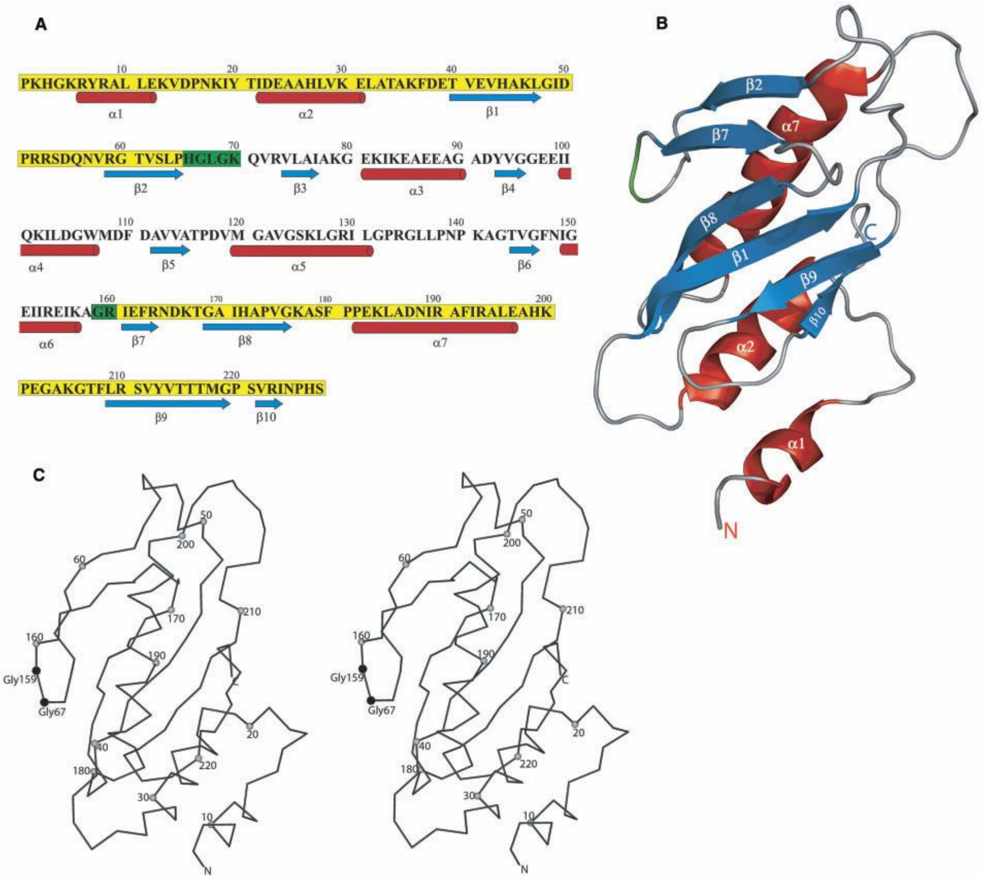
A
10 20 30 40 50
PKHGKRYRAL LEKVDPNKIY TIDEAAHLVK ELATAKFDET VEVHAKLGID
α1 α2 β1
60 70 80 90 100
PRRSDQNVRG TVSLPHGLGK QVRVLAIAKG EKIKEAEEAG ADYVGGEEII
β2 β3 α3 β4
110 120 130 140 150
QKILDGWMDF DAVVATPDVM GAVGSKLGRI LGPRGLLPNP KAGTVGFNIG
α4 β5 α5 β6
160 170 180 190 200
EIIREIKAGR IEFRNDKTGA IHAPVGKASF PPEKLADNIR AFIRALEAHK
α6 β7 β8 α7
210 220
PEGAKGTFLR SVYVTTTMGP SVRINPHS
β9 β10
B
β2
β7
α7
β8
β1
C
β9
β10
α2
α1
N
C
50
200
60
170
210
160
190
Gly159
Gly67
C
40
20
180
220
30
10
N

Figure 23. **A** – amino acid sequence of the TthL1 protein. The residues of the domain I and the flexible hinge with the domain II are shown in yellow and green respectively. **B** –Ribbon model of the isolated domain I. The secondary structure element numeration corresponds to that used for the whole TthL1. **C** – A stereoview of C_{α}-skeleton of the TthL1_d1 structure.

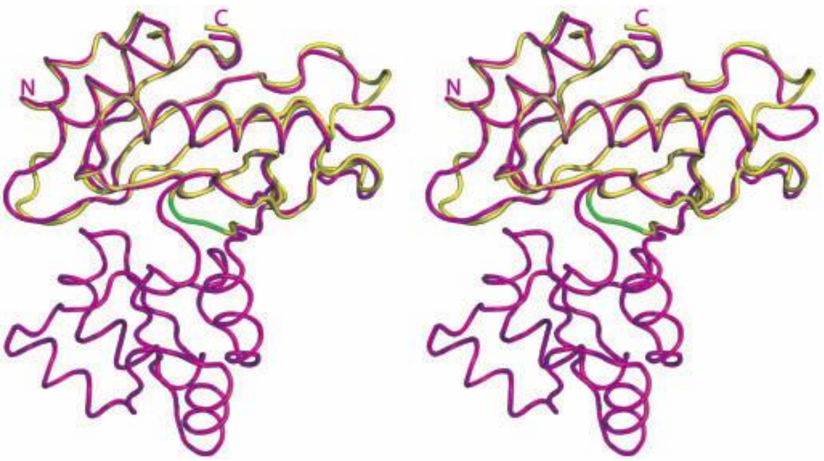

Figure 24. Superposition of the structures of the isolated domain I (yellow) and the whole TthL1 protein (magenta) with the least root-square method. Two Gly67 and 159 are in green.

Determination of the 3D structure of the complex of the domain I with specific mRNA fragment confirmed that this domain has possibility to form RNA-protein complex completely analogous to that formed with the intact protein (fig. 25). The analyses showed that the structure of domain I is not significantly changed while binding to mRNA.

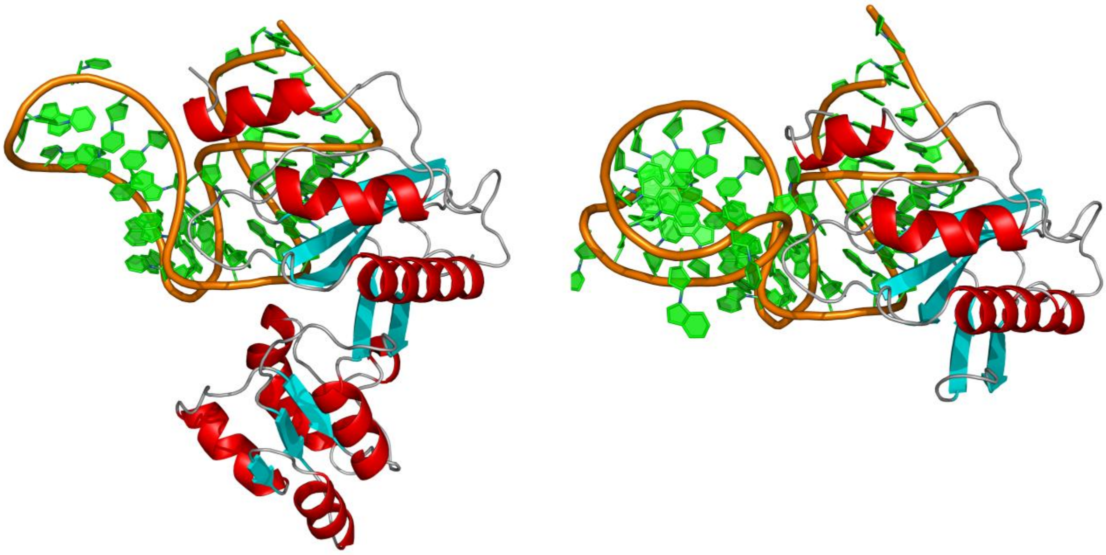

Figure 25. The complex of the full-size TthL1 protein with 38-nt length fragment mRNA (left) and 49-nt length fragment mRNA (right). The first domains of the proteins are shown in the same orientation.

Comparison of two mRNA-protein complexes demonstrates that the domain I in both the complexes can be perfectly superposed with a rmsd of about 0.45 Å for all C_{α}-atoms including the N-terminal α-helix. The total area of the RNA-protein contacts in both the complexes is similar and equal to approximately 1050 Å^2. The shape of the interacting surfaces appears also similar. All this together indicates that the interaction of TthL1 and TthL1_d1 with mRNA processes in a similar way. The RNA-protein hydrogen bond network is also almost identical in the complexes TthL1/mRNA and TthL1_d1/mRNA. In table 8 it is shown the list of all the RNA-protein hydrogen bonds with indication of solvent accessibility of the atoms forming these H-bonds.

Table 8. RNA-protein H-bonds in the complexes of full-sized TthL1 protein and its domain I with mRNA. In the brackets we report solvent accessibility for the appropriate atom (Å^2).

The complex TthL1_d1/mRNA			The complex TthL1/mRNA		
The protein atom	The RNA atom	Length, Å	The protein atom	The RNA atom	Length, Å
His3 ND1 (0.0)	U41 O1P (3.2)	2.4	His3 ND1 (0.0)	U30 O1P (3.3)	2.7
Lys5 N (0.0)	U16 O1P (10.1)	3.1	Lys5 N (0.0)	U16 O1P (13.9)	3.1
Arg6 NE (0.0)	C15 O1P (1.4)	2.9	Arg6 NE (0.0)	C15 O1P (1.4)	2.9
Tyr7 OH (10.9)	U42 O1P (8.8)	2.3	Tyr7 OH (10.9)	U31 O1P (8.3)	2.5
Lys36 N (0.0)	G14 O1P (0.9)	2.8	Lys36 N (0.0)	G14 O1P (0.6)	2.6
Thr40 OG1 (0.1)	G11 O1P (20.3)	2.6	Thr40 OG1 (0.6)	G11 O1P (3.9)	2.6

Glu42 OE2 (0.0)	G10 O2' (0.0)	2.7	Glu42 OE2 (0.0)	G10 O2' (0.0)	2.6
Lys46 NZ (10.2)	A44 O2' (0.0)	3.0	Lys46 NZ (8.7)	A33 O2' (0.0)	3.0
			Lys70 NZ (21.1)	G11 O1P (3.9)	3.1
Asp166 OD2 (0.0)	G7 N2 (0.0)	3.0	Asp166 OD2 (0.0)	G7 N2 (0.0)	3.0
Thr168 OG1 (0.0)	A44 O2' (0.0)	2.8	Thr168 OG1 (0.0)	A33 O2' (0.0)	2.6
Ser211 OG (0.0)	A44 O1P (15.1)	2.7	Ser211 OG (1.2)	A33 O1P (14.5)	2.8
Thr217 O (0.0)	G10 N2 (0.0)	3.3	Thr217 O (0.0)	G10 N2 (0.0)	3.0
Thr217 OG1 (0.0)	G10 O2' (0.0)	2.7	Thr217 OG1 (0.0)	G10 O2' (0.0)	2.8
Gly219 O (0.0)	U41 O2' (0.0)	2.4	Gly219 O (0.0)	U30 O2' (0.0)	2.7
Ser221 OG (8.0)	C43 O1P (17.8)	2.8	Ser221 OG (8.7)	C32 O1P (19.9)	2.6
a number of H-bonds		15	a number of H-bonds		16
a number of H-bonds inaccessible to the solvent		9	a number of H-bonds inaccessible to the solvent		10

From table 8 one can see that, in the complex TthL1_d1/mRNA one H-bond formed by Lys70 is absent due to the missing of this residue in the structure if the TthL1 domain I. Analyses of the solvent accessibility for the different amino acids involved in the interaction with mRNA did not reveal any significant differences between two complexes. The only important exclusion is a hydrogen bond formed by Thr40. In the full-sized complex this H-bond is screened from the solvent by the Lys70 side chain. As a result of the removing domain II, this H-bond becomes accessible to the solvent. The last circumstance could play a critical role and lead to the observed changing in the density of the RNA-protein contact for the complex (fig. 26).

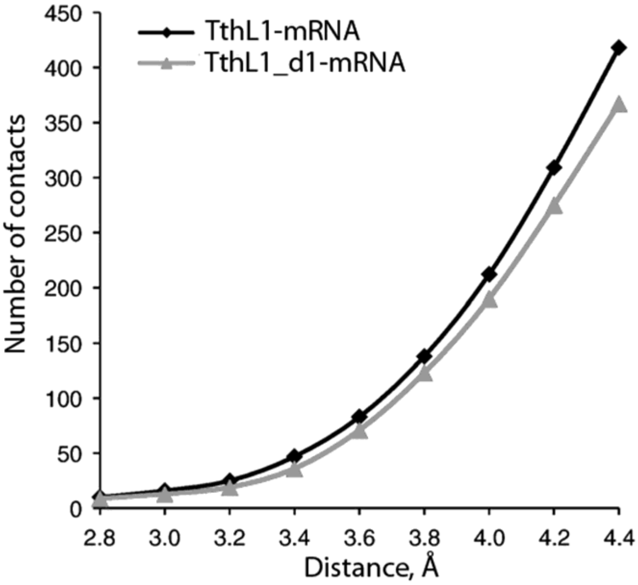

Figure 26. The number of RNA-protein interactions in the complex of full-sized TthL1 with mRNA (black) increases with increasing the interatom distances faster than in the case of the complex of the domain I with mRNA (grey).

4.1.2. Analyses of the ribosomal protein L1 mutant forms

The contact between the protein and the RNA is stabilized by a network of hydrogen bonds which exist only if the distance between the polar atoms ranges from 2.5 to 3.5 Å. The increasing intermolecular distances by 0.3-0.5 Å should lead to the breakage of H-bonds and of the complex. It is clear that any point mutation changes the landscape of the surface around the mutation point. However the influence of a point mutation on the closest areas is hard to evaluate without structure determination. In this section we consider the changes of the surface relief and the polar atom distribution at the L1-RNA contact area caused by amino acid point mutations in this area.

4.1.2.1. Point mutations in MjaL1

In the MjaL1 protein we substituted the conserved amino acid residues involved to the RNA-binding site (E27A, T204A, T204F, M205D and E27A/T204A) and determined their crystal structures. These structures were compared with the crystal structure of the wild type MjaL1 determined before [Nevskaya *et al.* 2000].

Analyses showed that the mutations E27A and T204A change neither the conformation of the protein backbone or the positions of the side chains of the non-mutated amino acid residues. The position of the C_β-atom of the alanine in the mutant proteins coincides with its position in the wild-type protein. The influence of these substitutions on the RNA-binding properties of the protein can be explained mainly by the loss of the appropriate H-bond and appearance the dehydratated polar atoms inside the interface. It is known that the presence such atoms in solvent inaccessible area destabilizes the structure. This effect is especially strong while double mutation; the mutant form E27A/T204A makes a complex neither with mRNA or rRNA. The dissociation constant in this case can increase by seven orders of magnitude compared to the wild-type complex [Lim *et al.* 2006].

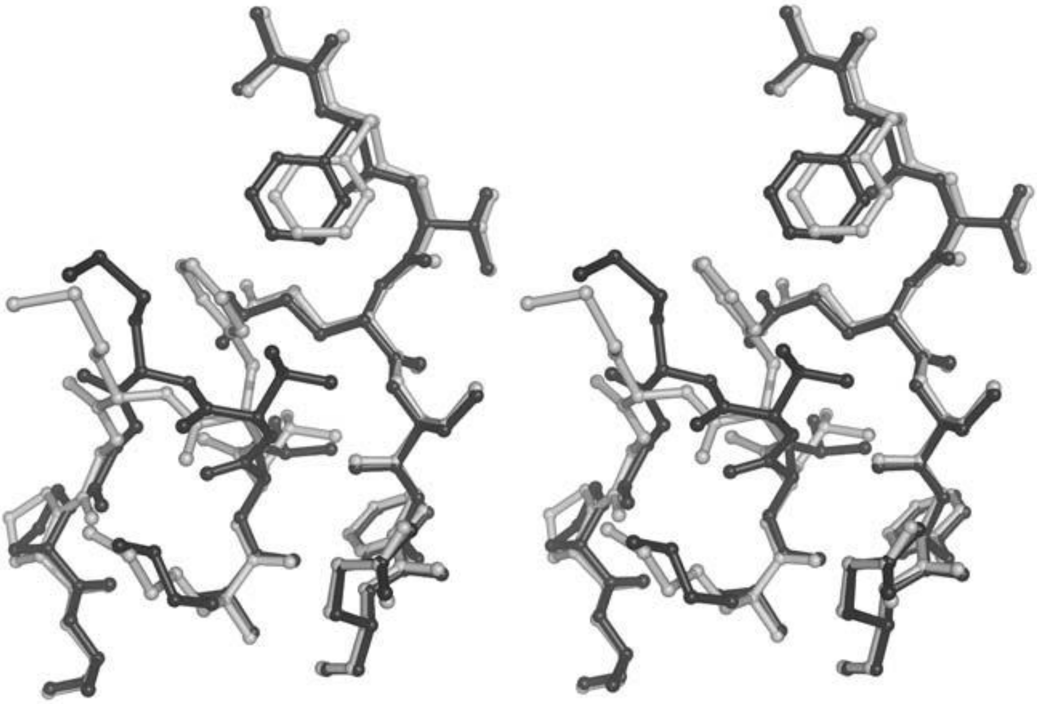

Figure 27. Changes in the MjaL1 surface region containing the T204F point mutation. Stereo view of the wild-type (black) and mutated (gray) MjaL1 fragment.

T204F has a more pronounced effect on the spatial structure of MjaL1. This results from the fact that among several possible phenylalanine rotamers, the protein accepts the one that places the side chain at a position similar to the threonine side chain in the wild type protein. As the side chain of F204 is larger than that of threonine, it pushes out the side chain of M205, thereby shifting the backbone in the region of F22. As a result, the loops containing F22 and M205 change their relative arrangement (fig. 27). The modified surface site is stabilized by stacking interactions and intramolecular hydrogen bonds. In the area occupied by M205 a bulge is formed of around 1 Å which can prevent specific H-bond observed in the wild type protein. Such changing in the binding site should lead to the disturbance of the RNA and protein surface complementarities. For this mutant form, we did not detect a stable complex with mRNA although the complex with rRNA could be formed.

The M205D MjaL1 mutant was crystallized in a space group C2, in contrast to wild-type MjaL1 and the other mutants of this protein which were crystallized in P1. The residue M205 in the wild type crystals participates in the crystal contacts with symmetry related molecules. Its substitution changes these contacts and results in the changing the space group. An analysis showed that the domains changed their relative orientation in the mutant structure. In addition, essential changes were observed in the structure of domain I for loops distant from the mutation point in the sequence and certain side chains involved in the crystal contacts (fig. 28). Whereas the structure nearby the mutation point is preserved. The positions of C_γ-atoms in the side chains of methionine and aspartat virtually coincide. Superposition of the mutant and wild-type proteins yielded rmsd of 1.8 Å for all $C_α$-atoms and 1.3 Å for domain I. As for the structure of domain II, it is preserved considerably better (rmsd = 0.5 Å).

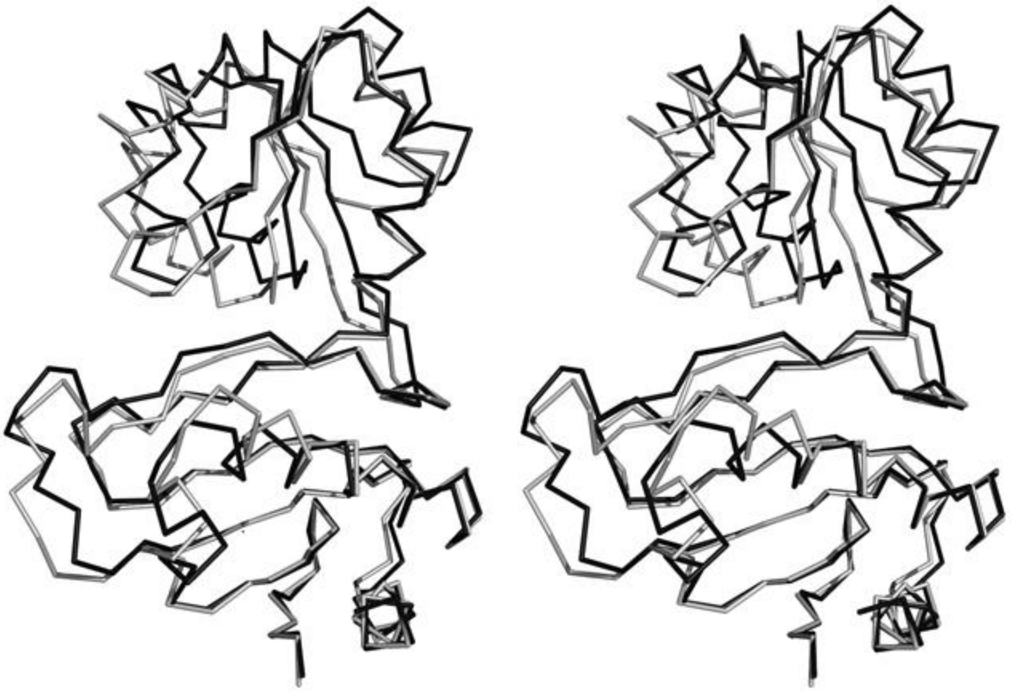

Figure 28. Change in the conformation of the MjaL1 backbone caused by the M205D mutation. The models of mutant (grey) and wild-type (black) proteins are superposed with a least square minimization of differences in the CA atom coordinates of the β-strands of domain I.

In table 9 below, we have summarized the rmsd values for $C_α$-atoms after the superposition of the MjaL1 mutant forms on either the isolated L1 protein or the L1 protein in the complex with RNA. The superposition is made with using residues 25-31 and 198-209 (MjaL1 numeration) that make the RNA-recognizing site. These areas correspond to residues 40-46 and 211-222 in TthL1 and to 29-35 and 202-213 in SacL1.

Table 9. Rmsd values (Å) of $C_α$-atoms obtained after superposition of the MjaL1 mutants on the L1 protein and L1 in the complex with RNA.

A mutant form	MjaL1*	MjaL1_mRNA*	SacL1_rRNA	TthL1_wt*	TthL1_mRNA
E27A	0.13 (0.21)	0.46 (1.27)	0.33	0.41	0.49
T204A	0.25 (0.37)	0.60 (1.29)	0.36	0.50	0.55
T204F	0.32 (0.28)	0.64 (1.31)	0.41	0.54	0.59
E27A/T204A	0.29 (0.32)	0.65 (1.30)	0.43	0.57	0.61
M205D	0.39 (1.81)	0.56 (1.94)	0.51	0.60	0.65
wt	-	0.47 (1.23)	0.31	0.43	0.50

* for MjaL1, the rmsd values in brackets correspond to superposition of $C_α$-atoms of residues 3-210.

4.1.2.1. Point mutations in TthL1

Four TthL1 mutants were prepared: F37I, S211A, T217A and M218L. The crystal structures of the mutant proteins were determined, comprehensively analyzed, and compared with the structure of wild-type TthL1. It was found that in the F37I and S211A mutants, virtually no change occurred in both the backbone conformation and the positions of the side chains of non-mutated amino acid residues. Moreover, the $C_{\beta-}$ and C_{γ}-atoms in the side chains of the substituted residues retained their positions. The I37 side chain adopted the position occupied by the plane of the phenylalnine ring of the wild-type protein. The rmsd values of the C_{α}-atom coordinates in the mutant proteins relative to wild-type TthL1 do not exceed the coordinate error for both the total protein and the region of the mutation. However, we observed in the F37I mutant the presence of a novel small cavity on the protein surface where water molecules could reach the RNA:protein interface. As the affinity of this mutant to rRNA decreases by one order of magnitude whereas the mRNA binding falls by more than two orders of magnitude, it is tempting to attribute these damages to the structure modifications that we observed.

The residue S211 is located at the RNA:protein interface. The substitution S211A allows us to evaluate the contribution of one solvent accessible H-bond to the stability of the complex. It was proposed that this substitution should not lead to dramatic consequences as the H-bond between RNA and S211 could be replaced by an H-bond between RNA and water molecules. The structural data confirm this suggestion. Moreover biochemical experiments indicate that the affinity of such mutant protein to both rRNA and mRNA remains almost at the same level. Thus, a solvent accessible H-bond does not play a significant role in the RNA-protein interaction.

The situation is different in the case of the T217A mutant. This mutation shifts the β_9-β_{10} loop. It was rather difficult to predict this structural change a priori, as the position of this loop in the wild-type protein is stabilized by three hydrogen bonds. Substitution T217A precludes only the bond involving the T217 OG atom. Nonetheless, the loop region 215–219 shifts along the unrealized bond. However, this shifted loop position is unstable and the loop can readily adopt a position corresponding to that observed in the wild-type protein. Presumably, both positions can exist in solution but one of them is more probable in the mutant. The absence of the threonine side chain gives the opportunity for two water molecules (W1 and W2 on fig. 29) to penetrate to the RNA-protein interface, to stabilize the structure of the RNA-binding site and to make on its surface a bulge preventing the formation of solid contacts with RNA.

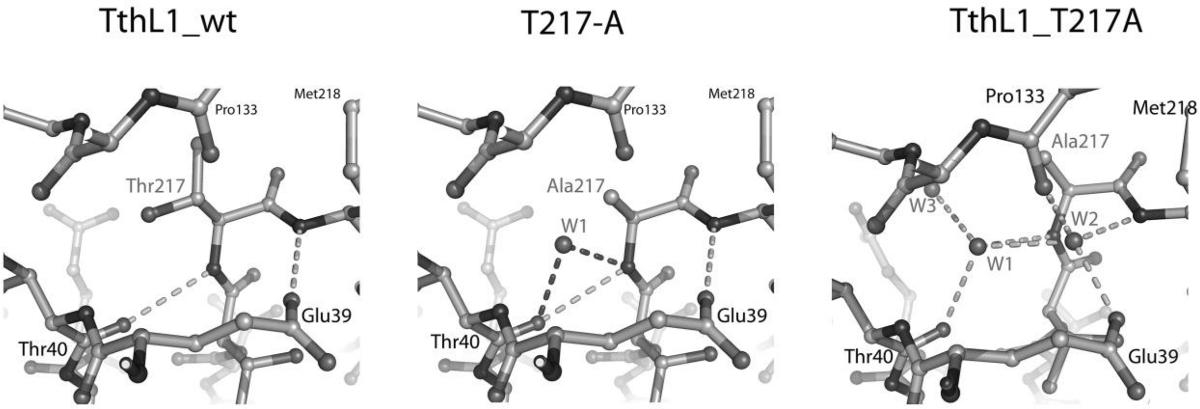

Figure 29. The T217A mutation allows three water molecules to occupy cavities leading to the conformational changes in region of the RNA-protein interface. Two of these waters (W1 and W2) make H-bonds with the protein polar atoms, W3 only interacts with W1.

It is worthy to note that such crucial effects were not observed in the MjaL1 protein with homological substitution T204A. As it was shown earlier, this substitution had almost no influence on the RNA-protein contact region. Probably such different influence of homological mutation of proteins from different sources could be explained by a nonconserved environment of the mutation point. These data again confirm the necessity of the determination of a mutant protein structure to correctly interpret the results of biochemical experiments.

The M218L substitution, like the M205D substitution in MjaL1, essentially changes the structure even beyond the region of the mutation. The relative arrangement of the domains changes in this mutant, considerably altering the unit cell parameters (table 4). In the region of the mutation, loop β_9-β_{10} is changed. Residues 215–219 are essentially shifted relatively to the corresponding region in the wild-type protein leading to changing the relief of the RNA-binding site and disturbance of its complementarity to the appropriate RNA site. The methionine is a strictly conserved residue in all the L1 family proteins and it becomes inaccessible to solvent after the RNA:protein complex is formed. Though a single rather weak H-bond with ribose phosphate backbone of the RNA [Nikulin *et al.* 2003] exists, it seems not critical for the RNA:protein complex stabilization. Nevertheless, our data show that the substitution of this residue can change the interactions between protein and RNA leading to strong destabilization of the complex.

In table 10 we summarize the rmsd values for C_α-atoms observed after superposition the TthL1 mutant forms on the isolated L1 protein and the L1 protein in the complex with RNA. The superposition is made with using the residues 40-46 and 211-222 (TthL1 numeration) making the RNA-recognizing site. These regions correspond to residues 25-31 and 198-209 in MjaL1 as well as 29-35 and 202-213 in SacL1. The conformation of these sites is the most conserved in all the known L1 structures.

Table 10. Rmsd values (Å) observed after superposition of the TthL1 mutants on the isolated L1 protein and L1 in complex with RNA.

A mutant form	TthL1_wt*	TthL1_mRNA	SacL1_rRNA	MjaL1	MjaL1_mRNA
F37I	0.185 (0.493)	0.288	0.444	0.426	0.633
S211A	0.195 (0.437)	0.328	0.436	0.422	0.620
T217A	0.429 (0.475)	0.569	0.743	0.730	0.819
M218L	0.334 (0.792)	0.377	0.495	0.520	0.747
wt	-	0.264	0.440	0.427	0.600

*rmsd values after the superposition with using C_α-atoms of the residues 11-228 are given in brackets.

The T217V substitution in TthL1_d1 leads to a significant decrease of affinity for RNA. The association constant drops by one order of magnitude whereas the dissociation constant increases by three orders as compared with TthL1_d1. In the wild-type protein the side chain of T217 makes an H-bond with nitrogen atom of the T40 main chain. Additionally, the mutual position of the loop β_9-β_{10}, N-terminal helix and the loop α_1-β_1 is strongly stabilized by the H-bond network. An analysis shows that the substitution of T217 hydroxyl group by the methyl group shifts the β_9-β_{10} loop (fig. 30) in the direction of M218 by 1Å forming a bulge on the protein surface. As a result, the α_1-β_1 loop containing the strictly conserved F37 residue obtains an extra flexibility. Increased flexibility of this region allows water molecules to penetrate to the inner cavity of the interface leading to complex destabilization.

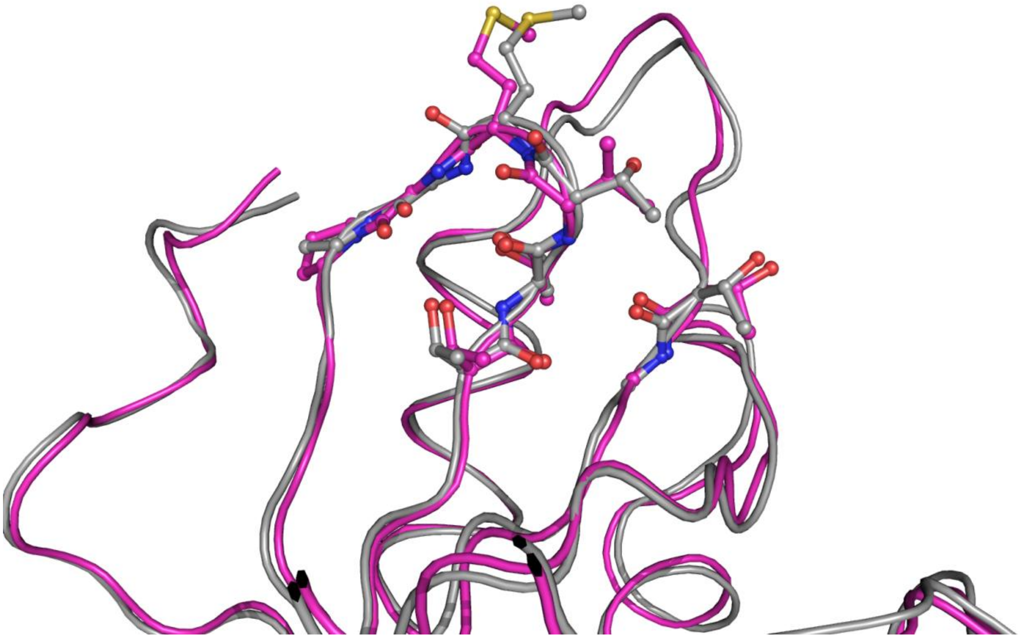

Figure 30. Superposition of the structures of TthL1_d1 (gray) and TthL1_d1 with substitution T217V (magenta). The valine side chain takes the same position as the threonine side chain.

All the structural studies performed here thus indicate possible distortions of the structure of the RNA:protein interface in the complexes formed by the mutants of L1 protein. The network of solvent inaccessible H-bonds should be significantly distorted, too. In turn it leads to the observed decrease of the stability of the complexes or in some cases even inability to form them. Thus, the data obtained indicate a key role of the specific H-bonds inaccessible to solvent molecules and complementarity of the interacting surfaces. These factors appeared to be also very important for the complexes between CSD and oligonucleotides, as it is discussed in the next section.

Annexe I

4.2. The complexes between YB-1 CSD and nucleic acids by MD simulation

MD simulations of YB-1 CSD in complex with different oligonucleotides were performed as a tool to overcome experimental investigations on these structures. The nucleic acids binding properties of CSD are discussed next. This section will show the results obtained in Evry and which were used to the preparation of a paper to be submitted to NAR (see Annexe II).

4.2.1. Affinity of CSD to DNA and RNA sequences

To address whether YB-1 CSD (named CSD hereafter) has any specificity to nucleic acid sequences and whether it has a similar affinity to RNA and DNA we explored in details its DNA and RNA binding properties.

4.2.1.1. DNA-binding

The DNA-binding properties of the CSD have been evaluated using different oligonucleotides bound to it (see 3.2.1. section for the definition of the parameters). The results indicate a sequence-dependent binding for positions labeled 2 to 4 (fig. 31). The oligoC sequence appeared as the worst CSD ligand (table 11).

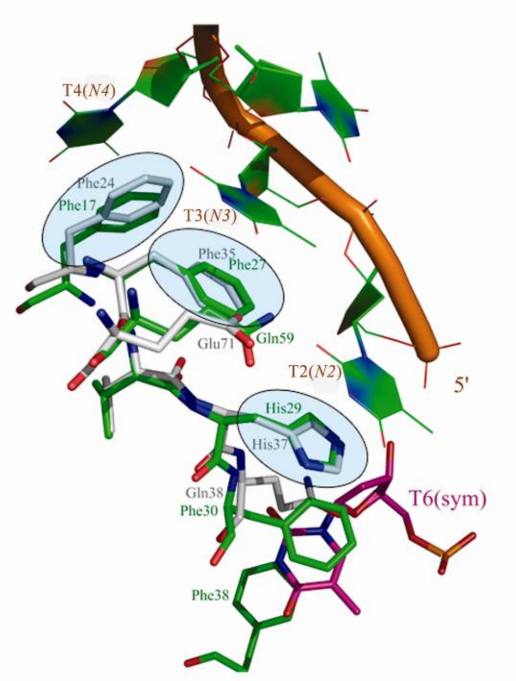

Figure 31. **S**uperposition of YB-1 CSD (gray) and bacterial CSP (green) nucleic acid binding site. Nucleotide binding sites are circled. The complex between bacterial CSP (green) and oligonucleotide dT6 (orange) corresponds to experimental structure. The nucleotide shown in magenta represents the symmetrically related molecule of the complex and shows the additional nucleotide binding site formed by Phe30 and Phe38.

Table 11. Binding of different desoxyribonucleotides at positions 2-4 on the surface of CSD.

nucleotide	stacking			Strength of H-bonds		Comment
	distance, Å	angle, deg	correlation	protein-NA	NA-NA	
A2	3.6±0.2	11±7	0.42±0.13	+		Strong interactions
A3	4.0±0.3	11±9	0.37±0.13	++		
A4	4.4±1.3	33±9	0.04±0.18	+		
C2	9.1±2.7	67±42	-0.03±0.11			unstable
C3	8.5±3.7	49±35	0.09±0.15	+		
C4	11.5±2.8	69±28	0.09±0.09			
G2	5.2±1.4	42±24	0.21±0.17	+++		Strong interactions
G3	3.9±0.4	15±10	0.56±0.13		++	
G4	4.9±0.5	56±20	-0.05±0.19			
T2	5.5±1.0	47±22	-0.11±0.13	++		unstable
T3	7.1±2.2	40±23	-0.02±0.16		+	
T4	6.0±0.9	55±13	0.03±0.28	++	+	
G2	4.3±0.5	23±15	0.26±0.19	++++	+	Strong interactions
G3	4.6±0.7	19±11	0.26±0.23		++	
T4	7.4±3.9	65±42	-0.10±0.10			
G2*	6.0±0.9	78±28	-0.08±0.11	++	+	Much less stable as compared with direct orientation
G3*	4.0±0.4	16±9	0.02±0.28			
T4*	7.3±0.9	88±18	0.02±0.25			

*reversed bound oligonucleotide (see explanation below)

**relative strength of H-bonds:	<0.2	0.2-0.7	0.7-1.2	1.2-2.0	>2.0
		+	++	+++	++++

The complex between CSD and oligoC was broken very soon during the MD, just after 100 ps of simulation. The residues His37, Phe35, Phe24 initially stacked with bases of the C2, C3 and C4 nucleotides, respectively lose completely their interactions with these bases. Very poor H-bonds were detected during the trajectory between the protein and the oligonucleotide with almost no H-bonds between the nucleotides (table 11). In the case of oligoT, we also observed very unstable interactions at positions 2 and 3. These thymidines broke their stacking to aromatic residues His37 and Phe35 within the first 100 ps and created a stable stacking to other nucleotide bases indicating the unfavourable interactions to the protein residues. T4 showed a better stacking with Phe24 thanks to additional H-bonds between the nucleotide base and the Lys14-Asp33 pair (fig. 32A).

Adenines in positions 2 and 3 were persistently stacked with His37 and Phe35 during the trajectory. The stacking of A3 was accompanied by a persistent hydrogen bond between its N6 and the main chain oxygen atom of Ala70. However, the strongest binding was observed in the case of guanine at the same positions (table 11). Additionally to the stacking we detected strong H-bonds specific to guanine between Glu71 and G2 (fig. 32B). We also found specific H-bonds between G2 and G3. The latter was not observed while other nucleotide pairs. Specific interactions between nucleotides at positions 2 and 3 can indicate that these nucleotides are sensitive to each other and could have a cooperative effect upon binding. The contact density plots also show that oligoC and oligoT are the less preferable ligands as compared to oligoG that provides the most stable complex with CSD (fig. 33). According to that the atomic fluctuation plots for the CSD:oligoG complex lie below the others (fig. 34).

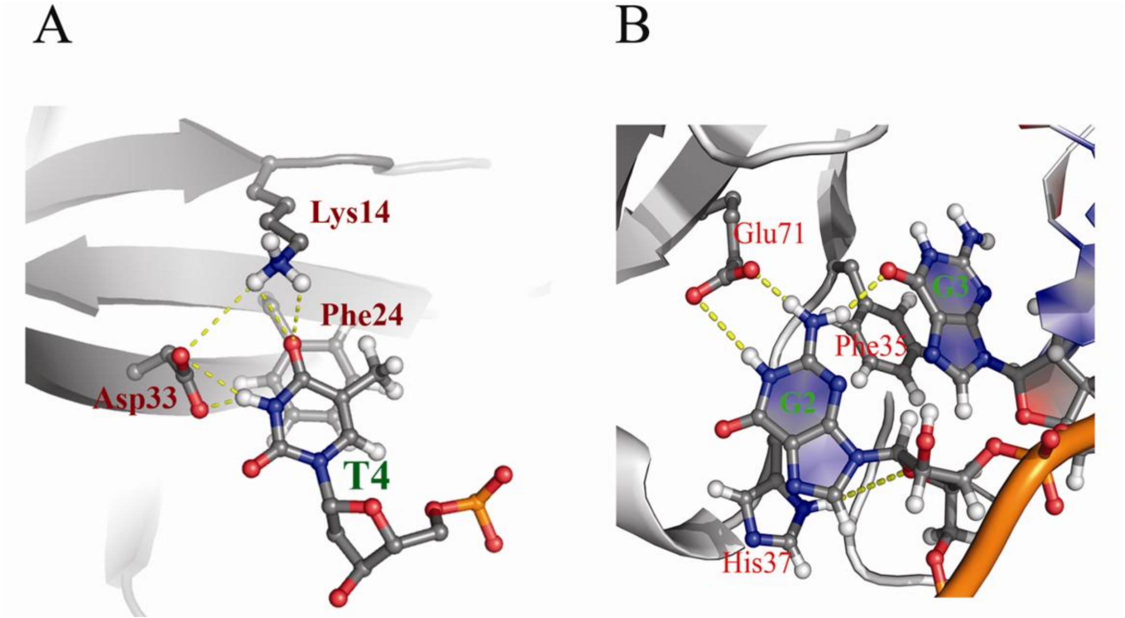

Figure 32. **A** – YB-1 CSD:T4 interactions: T4 stacks Phe24 and engaged by additional H-bonding with Asp33-Lys14; **B** – G2 and G3 engaged strong H-bond: the specific H-bond between G2 and G3 nucleotides bound to CSD in addition to H-bond between Glu71 and G2.

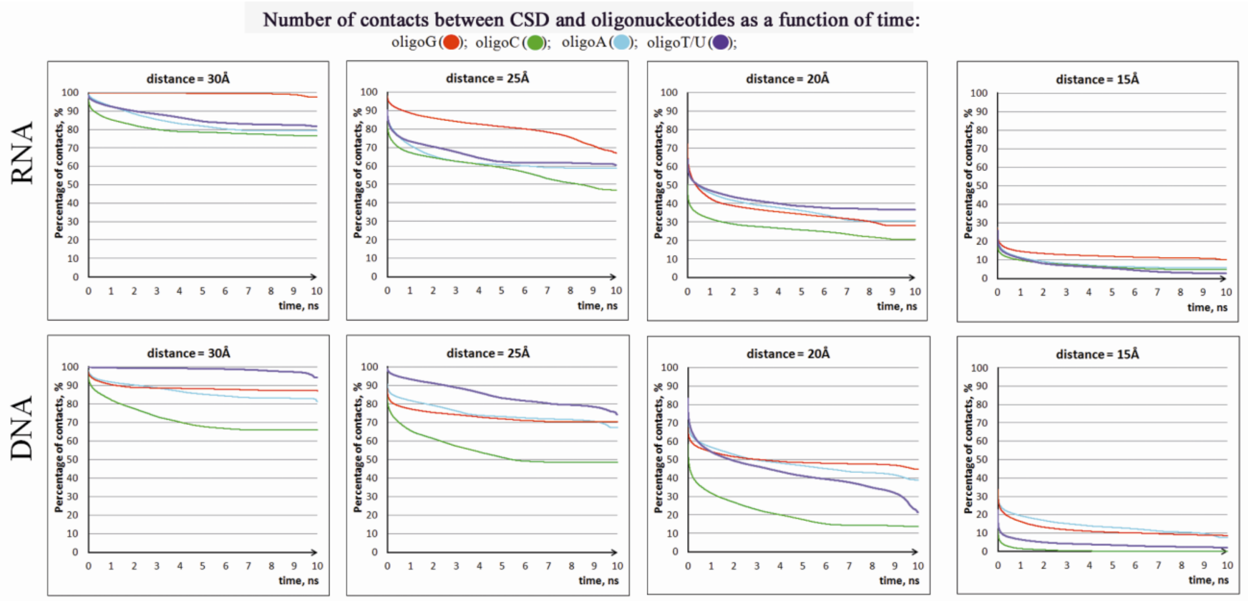

Figure 33. The contact density plots for the complexes between CSD and different oligonucleotides along the trajectories. More stable complexes show more smooth and plain graphs.

All these data allowed us to conclude that GGT seems to be the most preferable DNA-triplet to interact with CSD. Thus, the preference of the full-length YB-1 to G-rich single strand DNA-sequences [Zasedateleva *et al.* 2002] can be significantly provided by CSD. On the other hand the Y-box consensus sequence 5′-CTGATTGG-3′ seems to be recognized by other parts of YB-1 rather than CSD. From our simulation we can roughly evaluate dissociation constants from 10^{-3}-10^{-4} M for the less stable complexes to 10^{-5}-10^{-6} M for the most stable ones that is still significantly less as compared with the complexes formed by the bacterial CSP (10^{-8}-10^{-9} M) [Max *et al.* 2006, Max *et al.* 2007]. Possibly, the main reason of such difference is the missing of one stacking site for CSD.

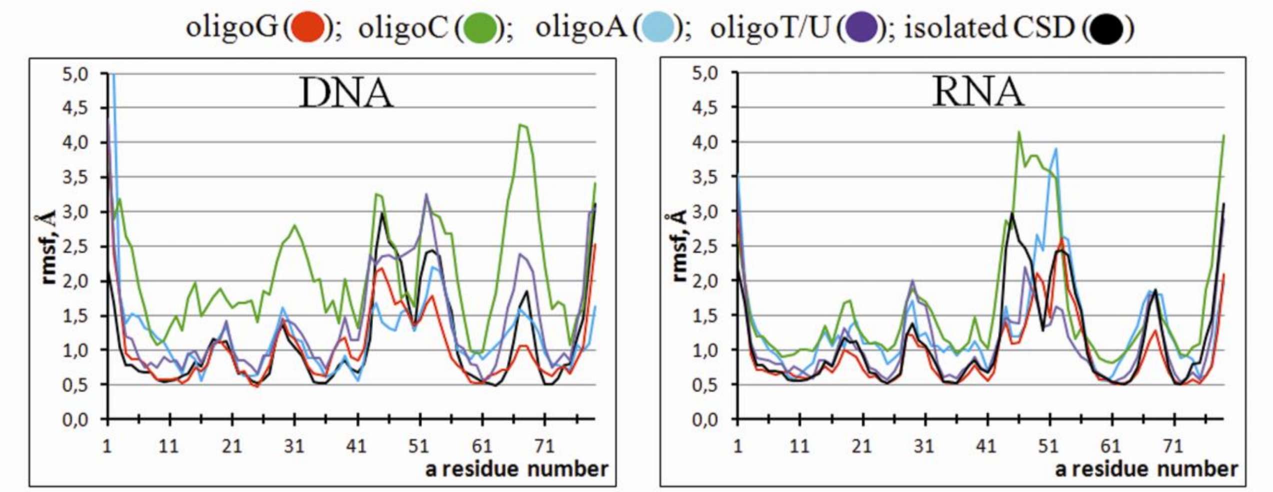

Figure 34. Atomic fluctuation (rmsf) of the protein C_α-atoms along the trajectories. The highest fluctuations are observed for less stable complexes due to disruption of the intermolecular contacts.

4.2.1.2. RNA-binding

Similar to the CSD:DNA complexes we also detected a strong binding of the G2-G3 and A2-A3 pairs in the case of CSD:RNA complexes (table 12). The significant difference between binding of RNA and DNA by the CSD was found for the U/T at positions 2-4. For desoxyribonucleitodes the interactions were very unstable with the only exception of dT4. At the same time oligoU (and even hypothetical oligo(r)T) can be bound by the CSD quiet properly. What is the structural explanation of such behavior? It can be clearly seen that during the trajectory, the OH groups of the ribose help to keep the appropriate local conformation of RNA participating in RNA:RNA H-bonding (table 12, fig. 35A). At position 4, only U can be found quite strong owing to the strong H-bond with Asp33 and Lys14 similar to those we observed for dT4 (fig. 32A). Analysis of the contact density plots along the trajectories show that the CSD:oligoC complex is breaking much faster as compared with the other complexes (fig. 33). The complex of the CSD with oligoG shows the highest stability. This is also confirmed by the atomic fluctuation plots indicating highest fluctuation for CSD:oligoC followed by CSD:oligoA (fig. 34). These results are in a good agreement with preceding biochemical data for RNA-binding properties of whole YB-1 [Minich *et al.* 1993] which found the highest affinity of YB-1 to polyG followed by polyU and polyA with polyC as the worst YB-1 ligand. It could imply that CSD mainly determines RNA-sequence specificity whereas the other YB-1 domains increase the affinity of the whole YB-1 protein to RNA and participates in its packing and masking [Skabkin *et al.* 2006].

Table 12. Binding of different ribonucleotides at the binding sites 2-4 on the surface of CSD.

nucleotide	stacking			Strength of H-bonds		Comment
	distance, Å	angle, deg	correlation	protein-NA	NA-NA	
A	3.8±0.4	15±11	0.12±0.31	+		Strong interactions
A	4.1±0.3	11±6	0.20±0.35	++	++	
A	12.6±3.8	89±34	0.08±0.15		+	
C	6.4±1.6	32±15	0.00±0.28		+++	unstable
C	4.0±0.3	14±8	0.07±0.37	++	++	
C	11.0±3.4	59±25	-0.13±0.09		+	
G	4.4±0.3	17±8	0.03±0.32	++++	+++	Strong interactions
G	3.7±0.2	10±6	0.13±0.37	+	++	
G	13.7±3.4	55±39	0.08±0.10			
U	7.2±2.8	38±18	0.29±0.15	++	++	Strong interactions
U	3.9±0.6	14±11	0.27±0.28		+	
U	4.1±0.5	25±19	0.14±0.18	+	+	
U*	3.7±0.3	18±11	-0.01±0.32	+	+	Less stable as compared with direct orientation
U*	7.4±1.5	69±35	-0.01±0.18	+++	+	
U*	5.8±1.8	50±25	-0.08±0.19	++	+	
G	4.3±0.4	18±8	0.19±0.22	++++	++	Strong interactions
G	3.9±0.3	11±7	0.18±0.26	+	++	
U	5.7±2.4	90±51	-0.21±0.14		+	
G*	7.6±1.8	35±22	0.21±0.13		+	Much less stable as compared with direct orientation
G*	6.3±1.5	57±28	-0.12±0.07	+++	++++	
U*	14.7±3.0	70±39	0.28±0.17		+	

* reversed bound oligonucleotide (see explanation below)

**relative strength of H-bonds:	< 0.2	0.2-0.7	0.7-1.2	1.2-2.0	>2.0
		+	++	+++	++++

In addition to sequence-dependent interactions at positions 2-4, we found some H-bonds between CSD and the sugar-phosphate backbone of the nucleotides at other positions. Several stable H-bonds have been detected between CSD amino acid residues (Asn20, Tyr22, Gln38, Glu65, Gly66 and Lys68) and ribose atoms at position 1, 5, 7 and 8. Additional H-bonds between the ribose OH group and CSD along with the stabilization of the local conformation of RNA by inner interactions should contribute significantly to the preference of YB-1 to RNA as compared to DNA. The difference in the CSD affinity to homologous sequences of RNA and DNA can be evaluated from our simulation as at least one order of magnitude. Another issue addressed by our simulation is whether CSD could bind oligonucleotides similarly in both orientations.

4.2.2. Does CSD orientate RNA/DNA under binding?

The structural analysis shows that there are no physical limit for the CSD to bind nucleic acids in both direct (as that in the bacterial complexes CSP:oligoT) and opposite orientations. Nevertheless, only one orientation appears possible for all known Oligonucleotide/ Oligosaccharide-Binding (OB) Fold proteins [Douglas *et al.* 2003]. The question about the polarity RNA/DNA upon binding to CSD is of interest since different OB-fold proteins bind nucleotides in different orientations. To resolve this question we have simulated the complexes of CSD with both directly orientated (as that found in bacterial complex CSP:dT6, fig. 35B) and reversely orientated (fig. 35C) oligonucleotides. We have used the most stable complexes which contained GGU and UUU triplets for RNA and GGT triplet for DNA at the positions 2-4. The MD simulation revealed that the complexes with the reverse orientated oligonucleotides are significantly less stable (at least one order of magnitude) as compared to direct orientated ones (table 11, 12). This results from unfavorable local interactions formed by nucleotides at position 4 (position 2 in the reversed form on fig. 35C); some polar atoms of RNA being dehydrated under this conformation and this leads to a separation of the protein and RNA molecules in this area and to the breaking of most of the intermolecular contacts. As a result, a nucleotide at position 4 is very poorly bound and the RNA:protein interface becomes accessible to the solvent. The difference in stability of the CSD:DNA complexes containing the direct and reversed oligonucleotides is even more significant (table 12). Thus the data clearly show that only the direct orientation is possible upon CSD-oligonucleotide binding, and this orientation corresponds to those found for the bacterial CSPs [Max *et al.* 2006, Max *et al.* 2007].

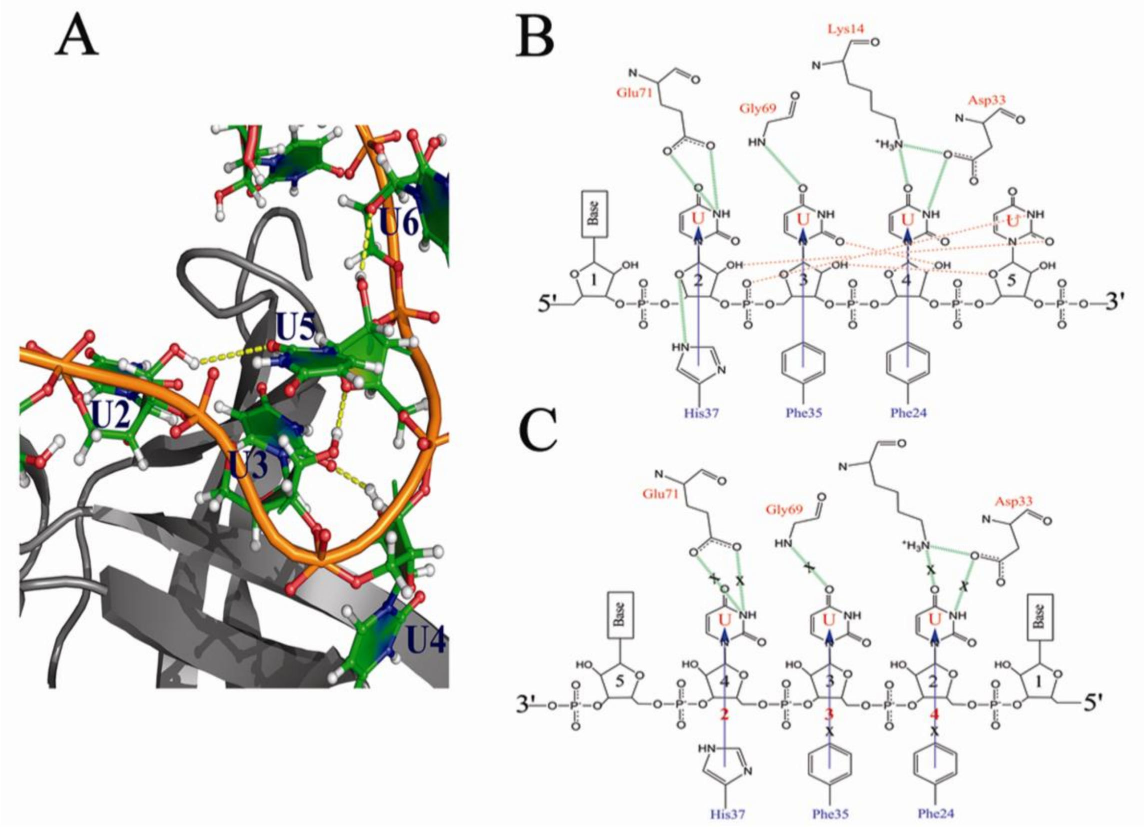

Figure 35. **A** - RNA-binding on the surface of CSD. Inner H-bonds are shown with yellow dotted lines. These H-bonds stabilise the local conformation of RNA in appropriate conformation. They are missed in the case of DNA sequences. **B, C** - Orientation of oligonucleotide influences the interactions with CSD: direct orientation (**B**) and reversed orientation (**C**). Non-bonded interactions upon binding in reversed orientation are significantly weakened. Breakage of the complexes in reverse orientation starts from the area close to nucleotide binding site 4.

This finding has fundamental importance for understanding of YB-1:mRNA interactions. When CSD presumably interacts with 5'-capped terminus of mRNAs [Evdokimova *et al.* 2001, Evdokimova *et al.* 2006, Bader *et al.* 2003] the cap structure should be placed nearby binding site 2 rather than binding site 4. The structural analysis also shows that in this case the A/P-rich domain of YB-1 cannot directly interact with the cap structure as it should be located far away from possible positions of the cap. To check a possible localization of the cap-binding site on the CSD surface and evaluate stability of some proposed interactions between the cap-structure of mRNA and the protein we have performed a manual docking followed by the MD-simulation. Several criteria should be taken into account from the common cap-binding mode observed for other proteins [Mousheng *et al.* 2009] and which includes three kinds of interactions:

1) H-bonds formed by three phosphate groups with positive charged amino acid residues;

2) Stacking interactions between 7N-methylated G-ring and commonly Trp or Tyr residues (the role of such unusual methylation is to increase this stacking due to additional positive charge on the guanine ring);

3) Additional anchoring of the G-ring by H-bonds via Glu/Asp side chain [Ueda *et al.* 1991].

Based on these data we suppose that Trp15 or Tyr49 could be the most probable candidates for stacking with the G-ring. The nearest negatively charged residues Asp55 and Glu57 could be also considered as a probable anchors for m7N-G. Furthermore the three phosphate group comes to a very favorable environment formed by the Asn17, Arg19, Asn20 and Arg51 residues. All this together allows us to propose a quite reasonable model of a cap-binding site.

4.2.2.1. A probable cap-binding site

Several models of binding of the cap on the CSD surface were checked by performance of the short MD simulations. One of the most favorable models is shown on fig. 36. In this model the stable interactions between the cap and CSD are formed by Trp15, Arg51 and Arg19. The side chains of Trp15 and Arg19 are stacked with the bases of the nucleotides in positions 0 and 1, correspondingly, whereas polar atoms of the Arg51 and Arg19 side chains make hydrogen bonds with the phosphate moiety of the cap. This configuration shows a tolerable stability during the MD-simulation that can indicate quite favorable interactions. This model still requires future experimental validations, but we can speculate that although YB-1 did not show strong interactions with the cap these interactions can increase the affinity of YB-1 to capped mRNA as compared with uncapped one. However, if we compare this cap-binding site to those found for human initiation factor 4E [Tomoo *et al.* 2003] we can predict that YB-1 is a very weak competitor of 4E for binding of the cap. Hence the interplay between YB-1 and 4E [Evdokimova *et al.* 2001] should have an indirect character and should be provided with more complicated mechanism rather than simple direct interactions, probably including the participation of other partners.

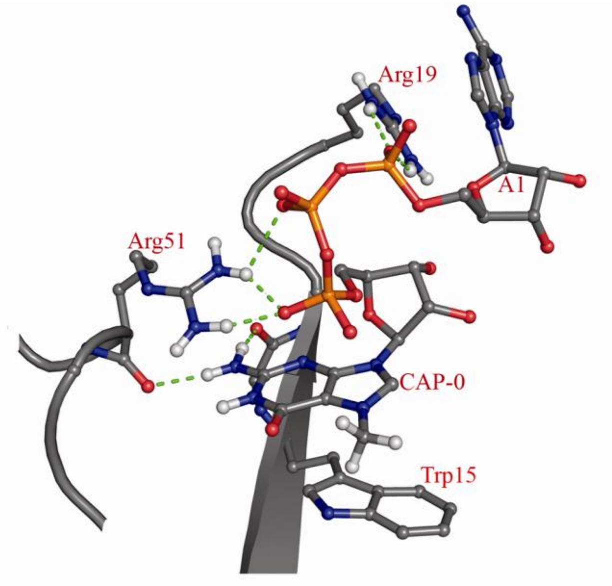

Figure 36. CSD can form some stable non-bonded contacts with the cap-structure. Stacking interactions are accompanied by strong H-bonding with three phosphate groups.

Annexe II

Summary

The crystal structures of the isolated domain I of L1 from *T. thermophilus* (TthL1_d1) and its complex with mRNA were determined at 2.3 Å resolution. It was shown that the conformation of protein:RNA interface contacts for intact TthL1 and its domain I are identical in the isolated state and in the complexes with mRNA. Domain I forms almost the same protein:RNA contacts as TthL1 does. This allows to suggest that the role of the domain II is to increase the affinity of L1 protein to the ribosomal RNA. This increased affinity is necessary for the feedback regulation of translation by L1 protein and allows to maintain the L1/rRNA ratio at an optimal level in cells. As soon as all the targeted ribosomal RNA is saturated by L1, free L1 appears in cells. This free protein then finds its mRNA target and specifically binds to it in order to prevent the further synthesis of the ribosomal proteins encoded by this mRNA (including L1 protein itself). When assembling new ribosome particles, L1 protein, possessing a higher affinity for its rRNA site, releases from the mRNA to bind to the rRNA. This launches the synthesis of new ribosomal protein molecules encoded by this polycistronic mRNA. Thus, the domain I provides the strong and specific binding with its targets on both RNAs.

To further analyze the contacts formed between domain I of L1 and the RNAs we used the method of point mutations, substituting the key residues in domain I which form the RNA-protein contacts. These substitutions are summarized in the table 13.

Table 13. The mutant L1 proteins solved in this work.

Source	Protein residue	RNA-protein interactions	Substituted by	Influence of the substitution
MjaL1	E27	An H-bond inaccessible to solvent molecules	A	Changing the relief of the protein surface; Loss of the RNA-protein H-bond
	T204	An H-bond inaccessible to solvent molecules	A	Changing the relief of the protein surface; Loss of the RNA-protein H-bond
	T204	An H-bond inaccessible to solvent molecules	F	Changing the relief of the protein surface
	M205	Van-der-Waals contacts	D	Changing the relief of the protein surface
	E27 T204	Two H-bond inaccessible to solvent	A A	Changing the relief of the protein surface; Loss of the two RNA-protein H-bond

		molecules		
TthL1	T217	An H-bond inaccessible to solvent molecules	A	Changing the relief of the protein surface; Loss of the RNA-protein H-bond
	F37	Van-der-Waals contacts	I	Changing the relief of the protein surface
	S211	An H-bond accessible to solvent molecules	A	Changing the relief of the protein surface; Loss of the RNA-protein H-bond accessible to solvent molecules
	M218	Van-der-Waals contacts	L	Changing the relief of the protein surface
	T217*	An H-bond inaccessible to solvent molecules	V	Changing the relief of the protein surface; Loss of the RNA-protein H-bond

* The substitution was made for the domain I of TthL1.

The data obtained in this work show that the substitution of the conserved residues forming inaccessible to the solvent RNA-protein H-bonds has dramatic effect on the stability of the complexes with both RNAs. For example, the substitution E27A or T204A in MjaL1 (or T217A in TthL1) leads to the loss of the H-bond inaccessible to the solvent molecules. In fact, in the interface of the complex of the mutant L1 proteins there should appear the dehydrated RNA polar atoms when compared to the wild-type complexes. These polar atoms are not able to make H-bond neither with protein atoms nor with water molecules, as they do in the isolated RNA. As a result, the RNA-protein complexes are dramatically destabilized. This effect is even stronger in the case of double substitution as it was shown for the mutant form E27A/T204A of MjaL1. Such a mutant fails to bind to both mRNA and rRNA.

At the same time the substitution S211A in TthL1 had almost no influence on the complexes. The serine residue is located on the surface of the protein. It makes the H-bond which is accessible to the solvent molecules. Therefore when substituted the missed RNA-protein H-bond can be easily replaced by the H-bonds between RNA and water molecules. This observation confirms a key role of the water molecules in stabilization of the structure of RNA-protein complexes.

The other important residue interacting with both the RNAs is M205 in MjaL1 (or M218 in TthL1). This residue is buried into the RNA-protein interface. Therefore the contacts (mainly Van-der-Waals) formed by this residue are also inaccessible to the solvent. Moreover, this methionine creates a local relief of the protein that is perfectly fitted to interact with the surface of RNA (the two surfaces are mutually complementary). Our data shows that substitution of this residue changes the relief of the protein surface and disturbs the complementary of the interacting surfaces. This strongly destabilizes the complexes with both RNAs. The similar effect was observed in the case of substitution F37I. Although the isoleucine side chain was shown to adopt the location taken by the phenylalanine side chain in the wild-type protein, nevertheless the local relief of the protein surface was changed as a result of this replacement. The local cavity observed in the structure of TthL1_F37I mutant allows the water molecules to penetrate to the RNA-protein interface and destabilize the complexes. Thus, the data obtained indicate the complementary of the interacting surfaces as one of the most important factor determining stability of the RNA-protein complexes.

Similar factors have been revealed to play an important role for the other system, the complexes between Cold Shock Domain of YB-1 protein (CSD) and 9-nt length oligonucleotides, studied in this work with molecular dynamics (MD) simulation method. To that end, we simulated the complexes of CSD with oligoG, oligoA, oligoC and oligoT (in the case of DNA) as well as oligoG, oligoA, oligoC and oligoU (in the case of RNA) in a water environment during 10 ns (see section 3.2). To define relative stability of the different CSD:oligonucletide complexes we evaluated the H-bonds and stacking interactions at the positions 2 to 4 (fig. 31).

The results obtained indicate a sequence dependent affinity of CSD for both DNA and RNA. OligoG appeared as the most preferable ligand followed by oligoU and oligoA. The complexes CSD:oligoT and CSD:oligoC were broken very soon during the simulation indicating unfavorable intermolecular contacts. The difference between the less stable complexes and more stable ones can be evaluated from our simulation as at least two orders of magnitude. In the case of oligoG we detected some specific H-bonds with CSD which were inaccessible to the solvent molecules. Moreover, we detected the specific H-bonds between G2 and G3 (fig. 32B). The latter was not observed for other pairs of nucleotides. This can mean that the nucleotides at positions 2 and 3 are sensitive to each other, and binding of a nucleotide at position 2 should influence the binding of a nucleotide at position 3 and *vice versa*. Thus, in agreement with the data obtained for L1 protein and its complexes with RNAs, we revealed H-bonds inaccessible to solvent molecules as one of the most important factors providing specificity and stability of the complexes CSD :oligonuleotides.

When comparing similar RNA and DNA sequences we can see that OH-groups of riboses participates in intramolecular H-bonds (fig 35A). These contacts help to keep an appropriate local conformation of RNA required for the binding on the surface of CSD. On the other hand we also detected some RNA-protein H-bonds provided by OH-groups at the positions 1, 7 and 8. Altogether, this indicates that CSD should possess a higher affinity for RNA sequences when compared with similar DNA sequences. From our simulation the difference in affinity for RNA should be at least one order of magnitude higher than that for DNA. Again, this specificity is mainly provided by H-bonds inaccessible to the solvent molecules. Extrapolation of the obtained data to the whole YB-1 protein allows us to conclude that CSD seems to determine specificity of YB-1 to nucleic acids whereas the two other domains are necessary for strong interactions.

In the present work we also proposed a probable cap binding site on the surface of CSD and analyzed the proposed contacts by MD simulations. The model of cap binding by CSD is shown on figure 36. This model includes all the commonly observed protein:cap interactions. Among them follows:

1) H-bonds formed by three phosphate groups with R19 and R51 residues;

2) Stacking interactions between 7N-methylated G-ring and W15;

3) H-bonds between the polar atoms of the G-ring and the main chain of R51.

These contacts should contribute to an increased affinity of YB-1 for the capped mRNA as compared to uncapped ones. However, these contacts do not look very stable when compared with the contacts formed between the eukaryotic initiation factor 4E and the cap [Tomoo *et al.* 2003]. Moreover, these interactions are located on the surface of CSD (and probably it is true for the whole YB-1). The latter makes YB-1 a weak competitor of 4E factor for binding of the cap.

To summarize the results obtained here with two different systems and performed with two different methods reveal that a central factor may determine the stability of a complex between a protein and DNA/RNA. This factor is the complementary of the interacting surfaces which includes not only the shape and relief but also the distribution of the polar atoms (donors and acceptors of H-bonds) and their accessibility to the solvent molecules. This principle should have more global character and be applicable for all kinds of interactions in a water environment including intermolecular (protein-protein, protein-ligands etc...) and intramolecular (for example, interaction of different domains or subdomains in a protein).

Next perspectives

In the present work we studied some of the principles of macromolecular recognition in biological systems. To that end, we evaluated the influence of point mutations at the binding site of L1 protein on the interactions L1:RNA by X-ray method. The influence of nucleotide sequence on the interactions CSD:oligonucleotide was studied with molecular dynamics simulation. The results obtained indicate H-bonds which are inaccessible to solvent molecules and complementary of the interacting surfaces as main factors that contribute a lot to the overall stability of these complexes. These factors should also play an important role for other types of intermolecular interactions in water environment.

The tasks for the nearest future are to simulate the complexes of L1 mutants with both mRNA and rRNA to detect the paths and mechanisms leading to the destabilization of these complexes. These simulations should be also performed in water environment as the water molecules are expected to play an important role in the intermolecular interactions.

Regarding YB-1 CSD and its complexes with oligonucleotides, we plan to evaluate the influence of phosphorylation upon Ser102 on the CSD:oligonucleotide interactions. It would be also interesting to further analyze the cap-binding site and the influence of phosphorylation on cap-binding, because it could enlighten significantly molecular mechanism of YB-1 functions. This information would be also used next to develop new anticancer drugs targeted YB-1.

Acknowledgments

I would like to thank the two Institutes where this work was performed: Institute of Protein Research, Russia (Drs. Nikonov and Ovchinnikov) and University of Evry, France (Drs. Curmi and Manivet) and all the people from the labs where I worked.

Thanks a lot to the reviewers Drs. Rak and Mauffret for the reading my manuscript and for the valuable comments and remarks they made.

I also thank the French Embassy in Moscow that financially supported this PhD.

References

Adams PD, Grosse-Kunstleve RW, Hung LW, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK, Terwilliger TC (2002). PHENIX: building new software for automated crystallographic structure determination. *Acta Cryst.*, **D58**, 1948-1954.

Adcock SA, McCammon JA (2006). Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.*, **106**, 1589-1615.

Agarwal RC (1978). A new least-square refinement technique based on the fast Fourier transforn algorithm. *Acta Cryst.*, **A34,** 791-809.

Amadei A, Linssen ABM, Berendsen HJC (1993). Essential dynamics of proteins. *Prot. Struct. Funct. Genet.*, **17**, 412–425.

Andersen HC (1983). Rattle — a velocity version of the shake algorithm for molecular-dynamics calculations. *J. Comput. Phys.*, **52**, 24-34.

Andricioaei I, Karplus M (2001). On the calculation of entropy from covariance matrices of the atomic fluctuations. *J. Chem Phys.*, **115**, 6289–6292.

Ansari SA, Safak M, Gallia GL, Sawaya BE, Amini S, Khalili K. (1999). Interaction of YB-1 with human immunodeficiency virus type 1 Tat and TAR RNA modulates viral promoter activity. *J. Gen. Virol*., **80 (Pt 10)**, 2629-2638.

Bader AG, Felts KA, Jiang N, Chang HW, Vogt PK (2003). Y box-binding protein 1 induces resistance to oncogenic transformation by the phosphatidylinositol 3-kinase pathway. *Proc. Natl. Acad. Sci. USA*, **100**, 12384–12389.

Baier G, Hohenwarter O, Hofbauer C, Hummel H, Stoffler-Meilicke M, Stoffler G (1989). Structure and functional equivalence between ribosomal proteins of *Escherichia coli* and *Methanococcus vannielii* L6. *Syst. Appl. Microbiol.*, **12**, 119–126.

Baier G, Piendl W, Redl B, Stoffler G (1990). Structure, organization and evolution of L1 equivalent ribosomal protein gene of the archaebacterium *Methanococcus vannielii*. *Nucleic Acids Res.*, **18**, 719–724.

Beeman D (1976). Some multistep methods for use in molecular dynamics calculations. *J. Comput. Phys.*, **20**, 130-139.

Berendsen HJ, Hayward S (2000). Collective protein dynamics in relation to function. *Curr Opin Struct Biol*, **10**, 165-169.

Berendsen H, Postma J, DiNola A, Haak JR (1984). Molecular dynamics with coupling to an external bath. *J. Chem.Phys*., **81**, 3684-3690.

Berendsen HJC, Postma JPM, Van Gunsteren WF, Dinola A, Haak JR (1984). Molecular-Dynamics with Coupling to an External Bath. *Journal of Chemical Physics*, **81**, 3684–3690.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000). The Protein Data Bank. *Nucleic Acids Research*, **28,** 235-242.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000). The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235-242.

Blake CCF, Phillips DC (1962). Biological Effects of Ionizing Radiation at the Molecular Level. *International Atomic Energy Agency Symposium, Brno, Czechoslovakia*, 183–191.

Blake CCF, Koenig DF, Mair GA, North ACT, Phillips DC, Sarma VR (1965). Structure of Hen Egg-White Lysozyme: A Three-dimensional Fourier Synthesis at 2 Å Resolution. *Nature*, **206**, 757-761.

Blobel G (1972). Protein tightly bound to globin mRNA. *Biochem. Biophys. Res. Commun.*, **47**, 88-95.

Bond PJ, Sansom MSP (2004). A recent review discussing simulations of outer-membrane proteins. *Mol Membr Biol.*, **21**, 151-161.

Born M, Oppenheimer JR (1927). On the quantum theory of molecules. *Ann. Phys.*, **84**, 457-484.

Bouvet P, Matsumoto K, Wolffe AP (1995). Sequence-specific RNA recognition by the Xenopus Y-box proteins. An essential role for the cold shock domain. *J. Biol. Chem.*, **270**, 28297-28303.

Branlant C, Krol A, Machatt A, Ebel JP (1981). The secondary structure of the protein L1 binding region of ribosomal 23S RNA. Homologies with putative secondary structures of the L11 mRNA and of a region of mitochondrial 16S rRNA. *Nucleic Acids Res.*, **9**, 293-307.

Bragg WL (1913). The Diffraction of Short Electromagnetic Waves by a Crystal. *Proceedings of the Cambridge Philosophical Society*, **17**, 43–57.

Briinden CI, Jones TA (1990). Between objectivity and subjectivity. *Nature*, **343**, 687-689.

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983). CHARMM. A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, **4**, 187-217.

Brooijmans N, Kuntz ID (2003). Molecular recognition and docking algorithms. *Ann. ReV. Biophys. Biomol. Struct.*, **32**, 335-373.

Brunger AT (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, **355**, 472-475.

Brunger AT (1993). Assessment of phase accuracy by cross validation: the free R value. Methods and applications. *Acta Cryst.*, **D49,** 24-36.

Brunger AT (1997). Patterson Correlation Searches and Refinement. *Methods in Enzymology*, **276**, 558-580.

Brunger AT, Adams PD (2002). Molecular dynamics applied to X-ray structure refinement. *Acc Chem Res.*, **35**, 404-412.

Brunger AT, Adams PD, Rice LM (1997). New applications of simulated annealing in X-ray crystallography and solution NMR. *Structure*, **5**, 325-336.

Brunger AT, Karplus M, Petsko GA (1989). Crystallographic refinement' by simulated annealing: application to crambin. *Acta Cryst.*, **A45,** 51-61.

Brunger AT, Krukovski A, Erickson JW (1990). Slow-cooling protocol for crystallographic refinement by simulated annealing. *Acta Cryst.*, **A46**, 585-593.

Brunger AT, Kuriyn J, Karplus M (1987). Crystallographic R factor refinement by molecular dynamics. *Science*, **235,**. 458-460.

Burke K, Werschnik J, Gross EKU (2005). Time-dependent density functional theory: Past, present, and future. *J. Chem. Phys.*, **123**, 062206.

Bussi G, Donadio D, Parrinello M (2007). Canonical sampling through velocity rescaling. *J. Chem. Phys*., **126**, 014101.

Capowski EE, Esnault S, Bhattacharya S, Malter JS (2001). Y box-binding factor promotes eosinophil survival by stabilizing granulocyte-macrophage colony-stimulating factor mRNA. *J. Immunol.*, **167**, 5970-5976.

Chen CY, Gherzi R, Andersen JS, Gaietta G, Jürchott K, Royer HD, Mann M, Karin M (2000). Nucleolin and YB-1 are required for JNK-mediated interleukin-2 mRNA stabilization during T-cell activation. *Genes Dev.*, **14**, 1236-1248.

Chernov KG, Mechulam A, Popova NV, Pastre D, Nadezhdina ES, Skabkina OV, Shanina NA, Vasiliev VD, Tarrade A, Melki J, Joshi V, Baconnais S, Toma F, Ovchinnikov LP, Curmi PA (2008). YB-1 promotes microtubule assembly in vitro through interaction with tubulin and microtubules. *BMC Biochem.*, **9**, 23.

Cohen A, Ellis P, Kresge N, Soltis SM (2001). MAD phasing with krypton. *Acta Cryst.*, **D57**, 233-238.

Coles LS, Bartley MA, Bert A, Hunter J, Polyak S, Diamond P, Vadas MA, Goodall GJ (2004). A multi-protein complex containing cold shock domain (Y-box) and polypyrimidine tract binding proteins forms on the vascular endothelial growth factor mRNA. Potential role in mRNA stabilization. *Eur. J. Biochem.*, **271**, 648-660.

Coles LS, Diamond P, Lambrusco L, Hunter J, Burrows J, Vadas MA, Goodall GJ (2002). A novel mechanism of repression of the vascular endothelial growth factor promoter, by single strand DNA binding cold shock domain (Y-box) proteins in normoxic fibroblasts. *Nucleic Acids Res.*, **30**, 4845-4854.

Coles LS, Lambrusco L, Burrows J, Hunter J, Diamond P, Bert AG, Vadas MA, Goodall GJ (2005). Phosphorylation of cold shock domain/Y-box proteins by ERK2 and GSK3beta and repression of the human VEGF promoter. *FEBS Lett.*, **579**, 5372-5378.

Collaborative Computational Project, Number 4 (1994). The CCP4 Suite: Programs for Protein Crystallography. *Acta Cryst.*, **D50**, 760-763.

Daggett V (2006). Protein Folding−Simulation. *Chem. ReV.*, **106**, 1898-1916.

Davydova EK, Evdokimova VM, Ovchinnikov LP, Hershey JW (1997). Overexpression in COS cells of p50, the major core protein associated with mRNA, results in translation inhibition. *Nucleic Acids Res.*, **25**, 2911-2916.

Day R, Daggett V (2003). All-atom simulations of protein folding and unfolding. *Adv. Protein Chem.*, **66**, 373-403.

Diederichs K, Karplus PA (1997). Improved R-factors for diffraction data analysis in macromolecular crystallography. *Nature Struct. Biol.*, **4**, 269-275.

de Groot BL, Hayward S, van Aalten DMF, Amadei A, Berendsen HJC (1998). Domain motions in Bacteriophage T4 Lysozyme; a comparison between molecular dynamics and crystallographic data. *PROTEINS:Struct., Func. and Gen.*, **31**, 116-127.

De Lano W (2002). The PyMOL Molecular Graphics System. *San Carlos, CA, USA.*

Dephoure N, Zhou C, Villén J, Beausoleil SA, Bakalarski CE, Elledge SJ, Gygi SP (2008). A quantitative atlas of mitotic phosphorylation. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 10762-10767.

Diamond R (1971). A real-space refinement procedure for proteins. *Acta Cryst.*, **A27**, 436-452.

Dodson E (1995). Report of workshop on validation of macromolecular structures solved by X-ray analysis. *Joint CCP4 and ESF-EACBM newsletter on protein crystallography*, **31**, 51-57.

Dong J, Akcakanat A, Stivers DN, Zhang J, Kim D, Meric-Bernstam F (2009). RNA-binding specificity of Y-box protein 1. *RNA Biol*, **6**, 59-64.

Dooley S, Said HM, Gressner AM, Floege J, En-Nia A, Mertens PR (2006). Y-box protein-1 is the crucial mediator of antifibrotic interferon-gamma effects. *J. Biol. Chem.*, **281**, 1784-1795.

Doreleijers JF, Vranken WF, Schulte C, Lin J, Wedell JR, Penkett CJ, Vuister GW, Vriend G, Markley JL, Ulrich EL (2009). The NMR restraints grid at BMRB for 5,266 protein and nucleic acid PDB entries. *J. Biomol. NMR*, **45**, 389–396.

Douglas TL, Rached MM, Deborah SW (2003). Nucleic Acid Recognition by OB-fold proteins. *Annu. Rev. Boiphys. Biomol. Structure*, **32**, 115-133.

Draper DE (1989). How do proteins recognize specific RNA sites? New clues from autogenously regulated ribosomal proteins. *Trends Biochem. Sci.*, **14**, 335-338.

Efimov AV (1994). Common structural motifs in small proteins and domains. *FEBS Lett.*, **355**, 213-219.

Elcock AH (2004). Molecular simulations of diffusion and association in multimacromolecular systems. *Numerical Computer Methods,* **D383**, 166-198.

Eliseeva IA, Kim ER, Guryanov SG, Ovchinnikov LP, Lyabin DN (2011). Y-box-binding protein 1 (YB-1) and its functions. *Biochemistry (Moscow)*, **76**, 1402-1433.

Evdokimova VM, Kovrigina EA, Nashchekin DV, Davydova EK, Hershey JW, Ovchinnikov LP (1998). The major core protein of messenger ribonucleoprotein particles (p50) promotes initiation of protein biosynthesis in vitro. *J. Biol. Chem.*, **273**, 3574-3581.

Evdokimova,V., Ruzanov,P., Anglesio,M.S., Sorokin,A.V., Ovchinnikov,L.P., Buckley,J., Triche,T.J., Sonenberg,N., Sorensen,P. (2006) Akt-Mediated YB-1 Phosphorylation Activates Translation of Silent mRNA Species. *Mol. Cell. Biol.*, **26**, 277–292.

Evdokimova V, Ruzanov P, Imataka H, Raught B, Svitkin Y, Ovchinnikov LP, Sonenberg N (2001). The major mRNA-associated protein YB-1 is a potent 5' cap-dependent mRNA stabilizer. *EMBO J.*, **20**, 5491-5502.

Evdokimova V, Tognon C, Ng T, Ruzanov P, Melnyk N, Fink D, Sorokin A, Ovchinnikov LP, Davicioni E, Triche TJ, Sorensen PH (2009). Translational activation of snail1 and other developmentally regulated transcription factors by YB-1 promotes an epithelial-mesenchymal transition. *Cancer Cell*, **15**, 402-415.

Evdokimova VM, Wei CL, Sitikov AS, Simonenko PN, Lazarev OA, Vasilenko KS, Ustinov VA, Hershey JW, Ovchinnikov LP (1995). The major protein of messenger ribonucleoprotein particles in somatic cells is a member of the Y-box binding transcription factor family. *J. Biol. Chem.*, **270**, 3186-3192.

Ewald P (1921). Die Berechnung optischer und elektrostatischer Gitterpotentiale, *Ann. Phys.*, **369**, 253–287.

Fan H, Mark AE (2004). Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein Sci.*, **13**, 211-220.

Fedoryuk MV (2001). Method of steepest descent. *Encyclopaedia of Mathematics*, *Springer*.

Feig M, Brooks CL (2004). Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.*, **14**, 217-24.

Finkbeiner MR, Astanehe A, To K, Fotovati A, Davies AH, Zhao Y, Jiang H, Stratford

AL, Shadeo A, Boccaccio C, Comoglio P, Mertens PR, Eirew P, Raouf A, Eaves CJ, Dunn SE (2009). Profiling YB-1 target genes uncovers a new mechanism for MET receptor regulation in normal and malignant human mammary cells. *Oncogene*, **28**, 1421-1431.

Fukada T, Tonks NK (2003). Identification of YB-1 as a regulator of PTP1B expression: implications for regulation of insulin and cytokine signaling. *EMBO J.*, **22**, 479-493.

Garcia AE (1992). Large-amplitude nonlinear motions in proteins. *Phys. ReV. Lett.*, **68**, 2696-2699.

Gaudreault I, Guay D, Lebel M (2004). YB-1 promotes strand separation in vitro of duplex DNA containing either mispaired bases or cisplatin modifications, exhibits endonucleolytic activities and binds several DNA repair proteins. *Nucleic Acids Res.*, **32**, 316-327.

Gear CW (1971). Numerical Initial Value Problems in Ordinary Differential Equations. *Prentice: New York.*

Giorgini F, Davies HG, Braun RE (2001). MSY2 and MSY4 bind a conserved sequence in the 3' untranslated region of protamine 1 mRNA in vitro and in vivo. *Mol. Cell. Biol.*, **21**, 7010-7019.

Gourse RL, Thurlow DL, Gerbi SA, Zimmermann RA (1981). Specific binding of a prokaryotic ribosomal protein to a eukaryotic ribosomal RNA: implications for evolution and autoregulation. *Proc. Natl Acad. Sci. USA*, **78**, 2722-2726.

Grant CE, Deeley RG (1993). Cloning and characterization of chicken YB-1: regulation of expression in the liver. *Mol. Cell. Biol.*, **13**, 4186-4196.

Green DW, Ingram VM, Perutz MF (1954). The Structure of Haemoglobin. IV. Sign Determination by the Isomorphous Replacement Method. *Proc. R. Soc.*, **A225**, 287-307.

Gross JD, Moerke NJ, von der Haar T, Lugovskoy AA, Sachs AB, McCarthy JE, Wagner G (2003). Ribosome loading onto the mRNA cap is driven by conformational coupling between eIF4G and eIF4E. *Cell*, **115**, 739-750.

Guay D, Garand C, Reddy S, Schmutte C, Lebel M (2008). The human endonuclease III enzyme is a relevant target to potentiate cisplatin cytotoxicity in Y-box-binding protein-1 overexpressing tumor cells. *Cancer Sci.*, **99**, 762-769.

Hanner M, Mayer C, Kohrer C, Golderer G, Grobner P, Piendl W (1994). Autogenous translational regulation of the ribosomal MvaL1 operon in the archaebacterium *Methanococcus vannielii*. *J. Bacteriol.*, **176**, 409–418.

Harker D (1956). The determination of the phases of the structure factors of non-centrosymmetric crystals by the method of double isomorphous replacement. *Acta Cryst.*, **D9**, 1-9.

Hartmuth K, Urlaub H, Vornlocher HP, Will CL, Gentzel M, Wilm M, Lührmann R

(2002). Protein composition of human prespliceosomes isolated by a tobramycin affinity-selection method. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 16719-16724.

Hasegawa SL, Doetsch PW, Hamilton KK, Martin AM, Okenquist SA, Lenz J, Boss JM (1991). DNA binding properties of YB-1 and dbpA: binding to double-stranded, single-stranded, and abasic site containing DNAs. *Nucleic Acids Res.*, **19**, 4915-4920.

Hendrickson WA (1976). Radiation damage in protein crystallography. *J. Mol. Biol.*, **106**, 889–893.

Hendrickson WA, Konnert JH (1980). Stereochemically restrained crystallographic least-squares refinement of macromolecule structures. *Biomolecular structure, function, conformation and evolution*, **1**, 43-57.

Hendrickson WA, Smith JL, Phizackerley RP, Merritt EA (1988). Crystallographic structure analysis of lamprey hemoglobin from anomalous dispersion of synchrotron radiation. *Proteins,* **4**, 77-88.

Hess B, Bekker J, Berendsen HJC, Fraaije JGEM (1997). LINCS: A linear constraint solver for molecular simulations. *J. Comp. Chem.*, **18**, 1463-1472.

Hess B, Kutzner C, van der Spoel D, Lindahl E (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J. *Chem. Theory Comput.*, **4**, 435-447.

Hestenes MR, Stiefel E (1952). Methods of Conjugate Gradients for Solving Linear Systems. *Journal of Research of the National Bureau of Standards*, **49**, 409-436.

Hooft RWW, Vriend G, Sander C, Abola EE (1996). Errors in protein structures. *Nature*, **381**, 272-272.

Humphrey W, Dalke A, Schulten K (1996). VMD–Visual molecular dynamics. *J. Mol. Graph. Model.*, **14**, 33–38.

Hurlbut CS, Klein C (1985). *Manual of Mineralogy, 20th ed*.

Ise T, Nagatani G, Imamura T, Kato K, Takano H, Nomoto M, Izumi H, Ohmori H, Okamoto T, Ohga T, Uchiumi T, Kuwano M, Kohno K (1999). Transcription factor Y-box binding protein 1 binds preferentially to cisplatin-modified DNA and interacts with proliferating cell nuclear antigen. *Cancer Res.*, **59**, 342-346.

Izumi H, Imamura T, Nagatani G, Ise T, Murakami T, Uramoto H, Torigoe T, Ishiguchi H, Yoshida Y, Nomoto M, Okamoto T, Uchiumi T, Kuwano M, Funa K, Kohno K (2001). Y box-binding protein-1 binds preferentially to single-stranded nucleic acids and exhibits 3'-->5' exonuclease activity. *Nucleic Acids Res.*, **29**, 1200-1207.

Jack GA, Levitt M (1978). Refinement of large structures by simultaneous minimization of energy and R factor. *Acta Cryst.*, **A34,** 931-935.

Jenkins RH, Bennagi R, Martin J, Phillips AO, Redman JE, Fraser DJ (2010). A conserved stem loop motif in the 5'untranslated region regulates transforming growth factor-β(1) translation. *PLoS ONE*, **5**, e12283.

Jolliffe IT (1986). Principal Component Analysis. *Springer-Verlag*, 487.

Jones TA, Zou JY, Cowan SW, Kjeldgaard M (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Cryst.*, **A47**, 110-119.

Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983). Comparison of simple functions for simulating liquid water. *J. Chem. Phys*., **79**, 926–935.

Jurchott K, Bergmann S, Stein U, Walther W, Janz M, Manni I, Piaggio G, Fietze E, Dietel M, Royer H. (2003). YB-1 as a cell cycle-regulated transcription factor facilitating cyclin A and cyclin B1 gene expression. *J. Biol. Chem.*, **278**, 27988-27996.

Kabsch W (2001). Integration, scaling, space-group assignment, postrefinement. *International Tables for Crystallography*, **F**.

Kahvejian A, Svitkin YV, Sukarieh R, M'Boutchou MN, Sonenberg N (2005). Mammalian poly(A)-binding protein is a eukaryotic translation initiation factor, which acts via multiple mechanisms. *Genes Dev.*, **19**, 104-113.

Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Philips DC (1958). A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature*, **181**, 662-666.

Khandelwal P, Padala MK, Cox J, Guntaka RV (2009). The N-terminal domain of y-box binding protein-1 induces cell cycle arrest in g2/m phase by binding to cyclin d1. *Int J Cell Biol*, **2009**, 243532.

Kim HW, Shen TJ, Sun DP, Ho NT, Madrid M, Tam MF, Zou M, Cottam PF, Ho C (1994). Restoring allosterism with compensatory mutations in hemoglobin. *Proc. Natl. Acad. Sci.*, **91**, 11547-11551.

Kirillov SV, Wower J, Hixson SS, Zimmermann RA (2002). Transit of tRNA through the *Escherichia coli* ribosome: cross-linking of the 3’ end of tRNA to ribosomal proteins at the P and E sites. *FEBS Lett.*, **514**, 60-66.

Kleywegt GJ, Jones TA (1995). Braille for pugilists. *Proceedings of the CCP4 Study Weekend*, 11-23.

Kloks CP, Spronk CA, Lasonder E, Hoffmann A, Vuister GW, Grzesiek S, Hilbers CW (2002). The solution structure and DNA-binding properties of the cold-shock domain of the human Y-box protein YB-1. *J.Mol.Biol.*, **316**, 317-326.

Koehl P, Levitt M (1999). De novo protein design. II. Plasticity of protein sequence, *J. Mol. Biol.*, **293**, 1183-1193.

Kohrer C, Mayer C, Neumair O, Grobner P, Piendl W (1998). Interaction of ribosomal L1 proteins from mesophilic and thermophilic Archaea and Bacteria with specific L1-binding sites on 23SrRNA and mRNA. *Eur. J. Biochem.*, **256**, 97–105.

Kojic S, Medeot E, Guccione E, Krmac H, Zara I, Martinelli V, Valle G, Faulkner G (2004). The Ankrd2 protein, a link between the sarcomere and the nucleus in skeletal muscle. *J. Mol. Biol.*, **339**, 313-325.

Konnert JH (1976). A restrained-parameter structure-factor least-squares refinement procedure for large asymmetric units. *Acta Cryst.*, **A32,** 614-617.

Kuriyn J, Brunger AT, Karplus M (1989). X-ray refinement of protein structures by simulated annealing: test of method on myohemerythrin. *Acta Cryst.*, **A45**, 396-409.

Ladomery M, Sommerville J (1994). Binding of Y-box proteins to RNA: involvement of different protein domains. *Nucleic Acids Res.*, **22**, 5582-5589.

Landsman D (1992). RNP-1, an RNA-binding motif is conserved in the DNA-binding cold shock domain. *Nucleic Acids Res.*, **20**, 2861-2864.

Lasham A, Moloney S, Hale T, Homer C, Zhang YF, Murison JG, Braithwaite AW, Watson J. (2003). Regulation of the human fas promoter by YB-1, Puralpha and AP-1 transcription factors. *J. Biol. Chem.*, **278**, 35516-35523.

Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, **26**, 283-291.

Lenz J, Okenquist SA, LoSardo JE, Hamilton KK, Doetsch PW (1990). Identification of a mammalian nuclear factor and human cDNA-encoded proteins that recognize DNA containing apurinic sites. *Proc. Natl. Acad. Sci. U.S.A.*, **87**, 3396-3400.

Leslie AGW (1992). Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4 + ESF-EAMCB Newsletter on Protein Crystallography*, **26**.

Li WW, Hsiung Y, Wong V, Galvin K, Zhou Y, Shi Y, Lee AS (1997). Suppression of grp78 core promoter element-mediated stress induction by the dbpA and dbpB (YB-1) cold shock domain proteins. *Mol. Cell. Biol.*, **17**, 61-68.

Lim V, Kljashtorny V (2006). Kinetic, energetic and stereochemical factors determing molecular recognition of proteins and nucleic acids. *Molecular biology*, **40**, 507-513.

Linge JP, Williams MA, Spronk CAEM, Bonvin AMJJ, Nilges M (2003). Refinement of protein structures in explicit solvent. *Proteins: Struct, Funct, Genet.*, **50**, 496–506.

Lutz M, Wempe F, Bahr I, Zopf D, von Melchner H (2006). Proteasomal degradation of the multifunctional regulator YB-1 is mediated by an F-Box protein induced during programmed cell death. *FEBS Lett.*, **580**, 3921-3930.

MacDonald GH, Itoh-Lindstrom Y, Ting JP (1995). The transcriptional regulatory protein, YB-1, promotes single-stranded regions in the DRA promoter. *J. Biol. Chem.*, **270**, 3527-3533.

Mackerell ADJr (2004). Empirical force fields for biological macromolecules: overview and issues. *J Comput Chem.*, **25**, 1584-604.

MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiórkiewicz-Kuczera J, Yin D, Karplus M (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.*, **102**, 3586-3616.

MacKerell JA, Feig M, Brooks CL (2004). Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comp. Chem.*, **25**, 1400-1415.

Manival X, Ghisolfi-Nieto L, Joseph G, Bouvet P, Erard M (2001). RNA-binding strategies common to cold-shock domain- and RNA recognition motif-containing proteins. *Nucleic Acids Res.*, **29**, 2223-2233.

Mayer C, Kohrer C, Grobner P, Piendl W (1998). MvaL1 autoregulates the synthesis of the three ribosomal proteins encoded on the MvaL1 operon of the archaeon *Methanococcus vannielii* by inhibiting its own translation before or at the formation of the fist peptide bond. *Mol. Microbiol.*, **27**, 455–468.

Matsumoto K, Meric F, Wolffe AP (1996). Translational repression dependent on the interaction of the Xenopus Y-box protein FRGY2 with mRNA. Role of the cold shock domain, tail domain, and selective RNA sequence recognition. *J. Biol. Chem.*, **271**, 22706-22712.

Matsumoto K, Tanaka KJ, Tsujimoto M (2005). An acidic protein, YBAP1, mediates the release of YB-1 from mRNA and relieves the translational repression activity of YB-1. *Mol. Cell. Biol.*, **25**, 1779-1792.

Max KE, Zeeb M, Bienert R, Balbach J, Heinemann U (2006). T-rich DNA single strands bind to a performed site on the bacterial cold-shock protein BS-CspB. *J.Mol.Biol.*, **360**, 702-714.

Max KE, Zeeb M, Bienert R, Balbach J, Heinemann U (2007). Common mode of DNA binding to cold shock domains. Crystal structure of hexathymidine bound to the domain-swapped form of a major cold shock protein from *Bacillus caldolyticus*. *FEBS Journal*, **274**, 1265-1279.

McCoy AJ (2004). Liking Likelihood. *Acta Cryst.*, **D60**, 2169-2183.

McCoy AJ, Grosse-Kunstleve RW, Storoni LC, Read RJ (2005). Likelihood-enhanced fast translation functions. *Acta Cryst.*, **D61**, 458-464.

McCoy AJ, Storoni LC, Read RJ (2004). Simple algorithm for a maximum-likelihood SAD function. *Acta Cryst.*, **D60**, 1220-1228.

Mchaourab HS, Oh KJ, Fang CJ, Hubbell WL (1997). Conformation of T4 Lysozyme in Solution. Hinge-Bending Motion and the Substrate-Induced Conformational Transition Studied by Site-Directed Spin Labeling. *Biochemistry*, **36**, 307-316.

Minich WB, Korneyeva NL, Berezin YV, Ovchinnikov LP (1989). A special repressor/activator system controls distribution of mRNA between translationally active and inactive mRNPs in rabbit reticulocytes. *FEBS Lett.*, **258**, 227-229.

Minich WB, Maidebura IP, Ovchinnikov LP (1993). Purification and characterization of the major 50-kDa repressor protein from cytoplasmic mRNP of rabbit reticulocytes. *Eur. J. Biochem.*, **212**, 633-638.

Minich WB, Ovchinnikov LP (1992). Role of cytoplasmic mRNP proteins in translation. *Biochimie*, **74**, 477-483.

Minich WB, Volyanik EV, Korneyeva NL, Berezin YV, Ovchinnikov LP (1990). Cytoplasmic mRNP proteins affect mRNA translation. *Mol. Biol. Rep.*, **14**, 65-67.

Møller C, Plesset MS (1934). Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.*, **46**, 618–622.

Mongan J (2004). Interactive essential dynamics. *J. Comput. Aided Mol. Des.*, **18**, 433-436.

Moraes KC, Quaresma AJ, Maehnss K, Kobarg J (2003). Identification and characterization of proteins that selectively interact with isoforms of the mRNA binding protein AUF1 (hnRNP D). *Biol. Chem.*, **384**, 25-37.

Molina H, Horn DM, Tang N, Mathivanan S, Pandey A (2007). Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 2199-2204.

Morel C, Gander ES, Herzberg M, Dubochet J, Scherrer K (1973). The duck-globin messenger-ribonucleoprotein complex. Resistance to high ionic strength, particle gel electrophoresis, composition and visualisation by dark-field electron microscopy. *Eur. J. Biochem.*, **36**, 455-464.

Morel C, Kayibanda B, Scherrer K (1971). Proteins associated with globin messenger RNA in avian erythroblasts: Isolation and comparison with the proteins bound to nuclear messenger-likie RNA. *FEBS Lett.*, **18**, 84-88.

Mousheng W, Nilsson P, Henriksson N, Niedzwiecka A, Lim KM, Cheng Z, Kokkoris K, Virtanen A, Song H (2009). Structural Basis of m7GpppG Binding to Poly(A)-Specific Ribonuclease. *Structure*, **17**, 276–286.

Murray MT (1994). Nucleic acid-binding properties of the Xenopus oocyte Y box protein mRNP3+4. *Biochemistry*, **33**, 13910-13917.

Murshudov GN, Vagin AA, Dodson EJ (1997). Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Cryst.*, **D53**, 240-255.

Murthy HM, Hendrickson WA, Orme-Johnson WH, Merritt EA, Phizackerley RP (1988). Crystal structure of Clostridium acidi-urici ferredoxin at 5 Å resolution based on measurements of anomalous X-ray scattering at multiple wavelengths. *J Biol Chem.*, **263**, 18430-18436.

Navaza J (2001). Implementation of molecular replacement in AMoRe. *Acta Crystallogr.*, **D57**, 1367-72.

Nekrasov MP, Ivshina MP, Chernov KG, Kovrigina EA, Evdokimova VM, Thomas AA, Hershey JW, Ovchinnikov LP (2003). The mRNA-binding protein YB-1 (p50) prevents association of the eukaryotic initiation factor eIF4G with mRNA and inhibits protein synthesis at the initiation stage. *J Biol Chem.*, **278**, 13936-13943.

Nevskaya N, Tischenko S, Fedorov R, Al-Karadaghi S, Liljas A, Kraft A, Piendl W, Garber M, Nikonov S (2000). Archaeal ribosomal protein L1: the structure provides new insights into RNA binding of the L1 protein family. *Structure Fold Des*, **8**, 363-371.

Nevskaya N, Tishchenko S, Gabdoulkhakov A, Nikonova E, Nikonov O, Nikulin A, Platonova O, Garber M, Nikonov S, Piendl W (2005). Ribosomal protein L1 recognizes the same specific structural motif in its target sites on the autoregulatory mRNA and 23S rRNA. *Nucleic Acids Res.*, **33**, 478-485.

Nevskaya N, Tishchenko S, Paveliev M, Smolinskaya Y, Fedorov R, Piendl W, Nakamura Y, Toyoda T, Garber M, Nikonov S (2002). Structure of ribosomal protein L1 from *Methanococcus thermolithotrophicus*. Functionally important structural invariants on the L1 surface. *Acta Cryst.*, **D58**, 1023-1059.

Nevskaya N, Tishchenko S, Volchkov S, Kljashtorny V, Nikonova E, Nikonov O, Nikulin A, Köhrer C, Piendle W, Zimmermann R, Stockley P, Garber M, Nikonov S (2006). New Insights into the Interaction of Ribosomal Protein L1 with RNA. *J. Mol. Biol*, **355**, 747-759.

Newkirk K, Feng W, Jiang W, Tejero R, Emerson SD, Inouye M, Montelione GT (1994). Solution NMR structure of the major cold shock protein (CspA) from Escherichia coli: identification of a binding epitope for DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 5114-5118.

Nikonov S, Nevskaya N, Eliseikina I, Fomenkova N, Nikulin A, Ossina N, Garber M, Jonsson B-H, Briand C, Al-Karadaghi S, Svensson A, Evarsson A, Liljas A (1996). Crystal structure of the RNA binding protein L1 from *Thermus thermophilus*. *EMBO*, **15**, 1350-1359.

Nikulin A, Eliseikina I, Tishchenko S, Nevskaya N, Davydova N, Platonova O, Piendl W, Selmer M, Liljas A, Drygin D, Zimmermann R, Garber M, Nikonov S (2003). Structure of the L1 protuberance in the ribosome. *Nature structural biology*, **10**, 104-108.

Ohba H, Chiyoda T, Endo E, Yano M, Hayakawa Y, Sakaguchi M, Darnell RB, Okano HJ, Okano H (2004). Sox21 is a repressor of neuronal differentiation and is antagonized by YB-1. *Neurosci. Lett.,* **358**, 157-160.

Ohga T, Koike K, Ono M, Makino Y, Itagaki Y, Tanimoto M, Kuwano M, Kohno K (1996). Role of the human Y box-binding protein YB-1 in cellular sensitivity to the DNA-damaging agents cisplatin, mitomycin C, and ultraviolet light. *Cancer Res.*, **56**, 4224-4228.

Ohga T, Uchiumi T, Makino Y, Koike K, Wada M, Kuwano M, Kohno K (1998). Direct involvement of the Y-box binding protein YB-1 in genotoxic stress-induced activation of the human multidrug resistance 1 gene. *J. Biol. Chem.*, **273**, 5997-6000.

Okamoto T, Izumi H, Imamura T, Takano H, Ise T, Uchiumi T, Kuwano M, Kohno K (2000). Direct interaction of p53 with the Y-box binding protein, YB-1: a mechanism for regulation of human gene expression. *Oncogene*, **19**, 6194-6202.

Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M (2006). Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, **127**, 635-648.

Orengo A, Thornton JM (1993). Alpha plus beta folds revisited: some favored motifs. *Structure*, **1**, 105-120.

Otwinowsk Z, Minor W (1997). Processing of X-ray Diffraction Data Collected in Oscillation Mode. *Methods in Enzymology*, **276**A, 307-326.

Ovchinnikov LP, Belitsina NV, Avanesov ATs, Spirin AS (1969). Postribosomal RNA-containing particles of cytoplasm of animal cells according to CsCl density gradient centrifugation data. *Dokl Akad Nauk SSSR*, **186**, 1202-1205.

Ovchinnikov LP, Skabkin MA, Ruzanov PV, Evdokimova VM (2001). Major mRNP proteins in the structural organization and function of mRNA in eukaryotic cells. *Mol Biol (Mosk)*, 35, 548-58.

Pannu NS, Read RJ (1996). Improved structure refinement through maximum likelihood. *Acta Cryst.*, **A52**, 659-668.

Palmer RG (1982). Broken Ergodicityю. *Adv. Phys.*, **31**, 669-735.

Paranjape SM, Harris E (2007). Y box-binding protein-1 binds to the dengue virus 3'-untranslated region and mediates antiviral effects. *J. Biol. Chem.*, **282**, 30497-30508.

Parker R, Song H (2004). The enzymes and control of eukaryotic mRNA turnover. *Nat. Struct. Mol. Biol.*, **11**, 121-127.

Patterson AL (1935). A direct method for the determination of the components of interatomic distances in crystals. *Z. Krist.*, **A90**, 517-542.

Pearson K (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arised from Random Sampling. *Philosophical Magazine Series 5*, **50**, 157-175.

Perry RP, Kelley DE (1968). Messenger RNA-protein complexes and newly synthesized ribosomal subunits: analysis of free particles and components of polyribosomes. *J. Mol. Biol.*, **35**, 37-59.

Pettitt BM, Makarov VA, Andrews BK (1998). Protein hydration density: theory, simulations and crystallography. *Curr Opin Struct Biol.*, **8**, 218–221.

Pisarev AV, Skabkin MA, Thomas AA, Merrick WC, Ovchinnikov LP, Shatsky IN (2002). Positive and negative effects of the major mammalian messenger ribonucleoprotein p50 on binding of 40 S ribosomal subunits to the initiation codon of beta-globin mRNA. *J. Biol. Chem.*, **277**, 15445-15451.

Ponder JW, Case DA (2003). Force fields for protein simulations. *Adv Protein Chem.* **66**, 27-85.

Quiocho FA, Mcmurray CH, Lipscomb WN (1972). Similarities Between the Conformation of Arsanilazotyrosine 248 of Carboxypeptidase Aα in the Crystalline State and in Solution. *Proc Natl Acad Sci USA*, **69**, 2850–2854.

Raffetseder U, Frye B, Rauen T, Jürchott K, Royer HD, Jansen PL, Mertens PR (2003). Splicing factor SRp30c interaction with Y-box protein-1 confers nuclear YB-1 shuttling and alternative splice site selection. *J. Biol. Chem.*, **278**, 18241-18248.

Raffetseder U, Rauen T, Djudjaj S, Kretzler M, En-Nia A, Tacke F, Zimmermann HW, Nelson PJ, Frye BC, Floege J, Stefanidis I, Weber C, Mertens PR (2009). Differential regulation of chemokine CCL5 expression in monocytes/macrophages and renal cells by Y-box protein-1. *Kidney Int.*, **75**, 185-196.

Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, **7**, 95-99.

Ravelli RB, Leiros HK, Pan B, Caffrey M, McSweeney S (2003). Specific radiation damage can be used to solve macromolecular crystal structures. *Structure*, **11**, 217-24.

Read RJ (2001). Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Cryst.*, **D57**, 1373-1382.

Roux B (2002). Computational studies of the gramicidin channel. *Acc. Chem. Res.*, **35**, 366-375.

Ruzanov PV, Evdokimova VM, Korneeva NL, Hershey JW, Ovchinnikov LP (1999).

Interaction of the universal mRNA-binding protein, p50, with actin: a possible link between mRNA and microfilaments. *J. Cell. Sci.*, **112 (Pt 20)**, 3487-3496.

Ryckaert JP, Ciccotti G, Berendsen HJC (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comput. Phys.*, **23**, 27-341.

Scheraga HA, Khalili M, Liwo A (2007). Protein-Folding Dynamics: Overview of Molecular Simulation Techniques. *Annual Review of Physical Chemistry*, **58**, 57-83.

Schindelin H, Jiang W, Inouye M, Heinemann U (1994). Crystal structure of CspA, the major cold shock protein of Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 5119-5123.

Sengupta S, Mantha AK, Mitra S, Bhakat KK (2011). Human AP endonuclease (APE1/Ref-1) and its acetylation regulate YB-1-p300 recruitment and RNA polymerase II loading in the drug-induced activation of multidrug resistance gene MDR1. *Oncogene*, **30**, 482-493.

Senn HM, Thiel W (2009). QM/MM Methods for Biomolecular Systems. *Angew. Chem.*, **48**, 1198 – 1229.

Shen Y, Bax A (2007). Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J. Biomol. NMR*, **38**, 289-302.

Shnyreva M, Schullery DS, Suzuki H, Higaki Y, Bomsztyk K (2000). Interaction of two multifunctional proteins. Heterogeneous nuclear ribonucleoprotein K and Y-box-binding protein. *J. Biol. Chem.*, **275**, 15498-15503.

Siegbahn PEM, Himo F (2009). Recent developments of the quantum chemical cluster approach for modeling enzyme reactions. *J Biol Inorg Chem*, **14**, 643–651.

Simonson T (2001). Macromolecular electrostatics: continuum models and their growing pains. *Curr. Opin. Struct. Biol.*, **11**, 243-252.

Skabkin MA, Evdokimova V, Thomas AA, Ovchinnikov LP (2001). The major messenger ribonucleoprotein particle protein p50 (YB-1) promotes nucleic acid strand annealing. *J. Biol. Chem.*, **276**, 44841-44847.

Skabkin MA, Kiselyova OI, Chernov KG, Sorokin AV, Dubrovin EV, Yaminsky IV, Vasiliev VD, Ovchinnikov LP (2004). Structural organization of mRNA complexes with major core mRNP protein YB-1. *Nucleic Acids Res.*, **32**, 5621-5635.

Skabkin MA, Lyabin DN, Ovchinnikov LP (2006). Nonspecific and specific interactions of Y-box-binding protein 1 (YB-1) with mRNA and posttranscriptional regulation of protein synthesis in animal cells. *Molecular Biology*, **40**, 551-563.

Skabkina OV, Lyabin DN, Skabkin MA, Ovchinnikov LP (2005). YB-1 autoregulates translation of its own mRNA at or prior to the step of 40S ribosomal subunit joining. *Mol. Cell.*

Biol., **25**, 3317-3323.

Skalweit A, Doller A, Huth A, Kähne T, Persson PB, Thiele BJ (2003). Posttranscriptional control of renin synthesis: identification of proteins interacting with renin mRNA 3'-untranslated region. *Circ. Res.*, **92**, 419-427.

Soop T, Nashchekin D, Zhao J, Sun X, Alzhanova-Ericsson AT, Björkroth B, Ovchinnikov L, Daneholt B (2003). A p50-like Y-box protein with a putative translational role becomes associated with pre-mRNA concomitant with transcription. *J. Cell. Sci.*, **116**, 1493-1503.

Sor F, Nomura M (1987). Cloning and DNA sequence determination of the L11 ribosomal protein operon of *Serratia marcescens* and *Proteus vulgaris*: translational feedback regulation of the *Escherichia coli* L11 operon by heterologous L1 proteins. *Mol. Gen. Genet.*, **210**, 52-59.

Sorokin AV, Selyutina AA, Skabkin MA, Guryanov SG, Nazimov IV, Richard C, Thong J, Yau J, Sorensen PH, Ovchinnikov LP, Evdokimova V (2005). Proteasome-mediated cleavage of the Y-box-binding protein 1 is linked to DNA-damage stress response. *EMBO J.*, **24**, 3602-3612.

Spirin AS, Belitsina NV, Aitkhozhin MA (1964). Messenger RNA in early embryogenesis. *Zh Obshch Biol.*, **25**, 321-38.

Stanley J, Sloof P, Ebel JP (1978). The binding site of ribosomal protein L1 from *Escherichia coli* on the 23S ribosomal RNA from *Bacillus stearothermophilus*. A possible base-pairing scheme differing from that proposed for *Escherichia coli*. *Eur. J. Biochem.*, **85**, 309-316.

Stein U, Bergmann S, Scheffer GL, Scheper RJ, Royer H, Schlag PM, Walther W (2005). YB-1 facilitates basal and 5-fluorouracil-inducible expression of the human major vault protein (MVP) gene. *Oncogene*, **24**, 3606-3618.

Stein U, Jürchott K, Walther W, Bergmann S, Schlag PM, Royer HD (2001). Hyperthermia-induced nuclear translocation of transcription factor YB-1 leads to enhanced expression of multidrug resistance-related ABC transporters. *J. Biol. Chem.*, **276**, 28562-28569.

Stenina OI, Shaneyfelt KM, DiCorleto PE (2001). Thrombin induces the release of the Y-box protein dbpB from mRNA: a mechanism of transcriptional activation. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 7277-7282.

Storoni LC, McCoy AJ, Read RJ (2004). Likelihood-enhanced fast rotation functions. *Acta Cryst.*, **D6**0, 432-438.

Stratford AL, Fry CJ, Desilets C, Davies AH, Cho YY, Li Y, Dong Z, Berquin IM, Roux PP, Dunn SE (2008). Y-box binding protein-1 serine 102 is a downstream target of p90 ribosomal S6 kinase in basal-like breast cancer cells. *Breast Cancer Res.*, **10**, R99.

Stratford AL, Habibi G, Astanehe A, Jiang H, Hu K, Park E, Shadeo A, Buys TPH, Lam W, Pugh T, Marra M, Nielsen TO, Klinge U, Mertens PR, Aparicio S, Dunn SE (2007). Epidermal growth factor receptor (EGFR) is transcriptionally induced by the Y-box binding protein-1 (YB-1) and can be inhibited with Iressa in basal-like breast cancer, providing a potential target for therapy. *Breast Cancer Res.*, **9**, R61.

Subramanian AR, Dabbs ER (1980). Functional studies on ribosomes lacking protein L1 from mutant *Escherichia coli*. *Eur. J. Biochem.*, **112**, 425-430.

Sussman JL (1985). Constrained-resntained least-squares (CORELS) refinement of proteins and nucleic acids. *Methods in Enzymology,* **115,** 271-303.

Sussman JL, Holbrook SR, Church GM, Kim SH (1977). A structure-factor least-square refinement procedure for macromolecular structures using constrained and restrained parameters. *Acta Cryst.*, **A33,** 800-804.

Sutherland BW, Kucab J, Wu J, Lee C, Cheang MC, Yorida E, Turbin D, Dedhar S, Nelson C, Pollak M, Leighton Grimes H, Miller K, Badve S, Huntsman D, Blake-Gilks C, Chen M, Pallen CJ, Dunn SE (2005). Akt phosphorylates the Y-box binding protein 1 at Ser102 located in the cold shock domain and affects the anchorage-independent growth of breast cancer cells. *Oncogene,* **24**, 4281–4292.

Svitkin YV, Evdokimova VM, Brasey A, Pestova TV, Fantus D, Yanagiya A, Imataka H, Skabkin MA, Ovchinnikov LP, Merrick WC, Sonenberg N (2009). General RNA-binding proteins have a function in poly(A)-binding protein-dependent translation. *EMBO J.*, **28**, 58-68.

Svitkin YV, Ovchinnikov LP, Dreyfuss G, Sonenberg N (1996). General RNA binding proteins render translation cap dependent. *EMBO J.*, **15**, 7147-7155.

Swamynathan SK, Nambiar A, Guntaka RV (2000). Chicken Y-box proteins chk-YB-1b and chk-YB-2 repress translation by sequence-specific interaction with single-stranded RNA. *Biochem. J.*, **348 Pt 2**, 297-305.

Swope WC, Andersen HC, Berens PH, Wilson KRJ (1982). A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *Chem. Phys.*, **76**, 637-649.

Tafuri SR, Wolffe AP (1992). DNA binding, multimerization, and transcription stimulation by the Xenopus Y box proteins in vitro. *New Biol.*, **4**, 349-359.

Thiel W (2009). QM/MM methodology: fundamentals, scope, and limitations. *Multiscale Simulation Methods in Molecular Sciences,* **42**, 203-214.

Thorpe I, Brooks CL (2004). The coupling of structural fluctuations to hydride transfer in dihydrofolate reductase. *Proteins: Struct. Funct. Bioinformatics,* **57**, 444–457.

Tiana G, Simona F, De Mori GMS, Broglia RA, Colombo G (2004). Understanding the determinants of stability and folding of small globular proteins from their energetic. *Protein Sci.*, **13**, 113–124.

Tishchenko S, Nikonova E, Kljashtorny V, Kostareva O, Nevskaya N, Piendl W, Davydova N, Streltsov V, Garber M, Nikonov S (2007). Domain I of ribosomal protein L1 is sufficient for specific RNA binding. *Nucl. Acids Res.*, **35**, 7389-7395.

Tomoo K, Shen X, Okabe K, Nozoe Y, Fukuhara S, Morino S, Sasaki M, Taniguchi T, Miyagawa H, Kitamura K, Miura K, Ishida T (2003). Structural features of human initiatian factor 4E, studied by X-ray crystal analyses and molecular dynamics simulations. *J. Mol. Biol*., **328**, 365-383.

Ueda H, Iyo H, Do M, Inoue M, Ishida T, Morioka H, Tanaka T, Nishikawa S, Uesugi S (1991)**.** Combination of Trp and Glu residues for recognition of mRNA cap structure Analysis of m7G base recognition site of human cap binding protein (IF-4E) by site-directed mutagenesis. *FEBS*, **280**, 207-210.

Vagin A, Teplyakov A (1997). MOLREP: an automated program for molecular replacement. J. *Appl. Cryst.*, **30**, 1022-1025.

van Gunsteren WF, Berendsen HJC (1977). Algorithms for macromolecular dynamics and constraint dynamics. *Mol. Phys.*, **34**, 1311–1327.

van Roeyen CR, Eitner F, Martinkus S, Thieltges SR, Ostendorf T, Bokemeyer D, Lüscher B, Lüscher-Firzlaff JM, Floege J, Mertens PR (2005). Y-box protein 1 mediates PDGF-B effects in mesangioproliferative glomerular disease. *J. Am. Soc. Nephrol.*, **16**, 2985-2996.

Verlet J (1967). Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev*., **159**, 98-103.

Wang W, Donini O, Reyes CM, Kollman PA (2001). Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Ann. ReV. Biophys. Biomol. Struct.*, **30**, 211-243.

Warshel A (2003). Computer simulations of enzyme catalysis: methods, progress, and insights. *Ann. ReV. Biophys. Biomol. Struct.*, **32**, 425-443.

Weiss MS, Hilgenfeld R (1997). On the use of the merging R-factor as a quality indicator for X-ray data. *J. Appl. Crystallogr*., **30**, 203-205.

Wong CF, McCammon JA (2003). Protein Simulation and Drug Design. *Adv. Protein Chem.*, **66**, 87-121.

Wower J, Kirillov SV, Wower IK, Guven S, Hixson SS, Zimmermann RA (2000). Transit of tRNA through the *Escherichia coli* ribosome. *J. Biol. Chem.*, **275**, 37887-37894.

Wutrich K (1986). NMR of Proteins and Nucleic Acids. Wiley, New York.

Zasedateleva OA, Krylov AS, Prokopenko DV, Skabkin MA, Ovchinnikov LP, Kolchinsky A, Mirzabekov AD (2002). Specificity of mammalian Y-box binding protein p50 in interaction with ss and ds DNA analyzed with generic oligonucleotide microchip. *J Mol Biol.*, **324**, 73-87.

Ypma TJ (1995). Historical development of the Newton-Raphson method. *SIAM Review*, **37**, 531–551.

Zimmermann RA (1980). Interactions among protein and RNA components of the ribosome. *Structure, Function and Genetics* of *Ribosomes*, *Baltimore University Park Press*, 135–169.

Zou Y, Evans S, Chen J, Kuo HC, Harvey RP, Chien KR (1997). CARP, a cardiac ankyrin repeat protein, is downstream in the Nkx2-5 homeobox gene pathway. *Development*, **124**, 793-804.