Article

J. Mex. Chem. Soc. **2012**, *56(1),* 23-31
© 2012, Sociedad Química de México
ISSN 1870-249X

Metal Binding Sites in Plant Soluble Inorganic Pyrophosphatases. An Example of the Use of ROSETTA Design and Hidden Markov Models to Guide the Homology Modeling of Proteins

Luis Rosales-León,[1] Eric Edmundo Hernández-Domínguez,[2] Samantha Gaytán-Mondragón,[2] and Rogelio Rodríguez-Sotres*[2]

[1] Facultad de Medicina. Universidad Nacional Autónoma de México. Circuito Escolar, Ciudad Universitaria, México, D.F., C.P. 04510 México.

[2] Facultad de Química Departamento de Bioquímica. Universidad Nacional Autónoma de México. Circuito Escolar, Ciudad Universitaria, México, D.F., C.P. 04510 México. (+52)5556225285 sotres@servidor.unam.mx.

Dedicated to Dr. Estela Sánchez de Jiménez for her invaluable contributions to plant biochemistry

Received January 4, 2011; accepted April 28, 2011

Abstract. In contrast to their counterparts in bacteria and animals the soluble inorganic pyrophosphatases from plant cells are active as monomers. The isoforms 1 and 4 from *Arabidopsis thaliana* have been characterized with more detail, but their three-dimensional structure is unavailable. Here, a recently published protocol (ROSETTA design-HMMer), is used to guide well-known techniques for homology-modeling, in the production of reliable models for the three-dimensional structure of these two arabidopsis isoforms. Their interaction with magnesium ions and pyrophosphate is analyzed *in silico*.

Key words: Protein 3D structure, comparative structural modeling, pyrophosphatase EC 3.6.1.1.

Resumen. En contraste con sus contrapartes bacterianas y animales, la pirofosfatasas inorgánicas solubles de células vegetales son activas como monómeros. La isoformas 1 y 4 de *Arabidopsis thaliana* se han caracterizado con mayor detalle, pero su estructura tridimensional no está disponible. Aquí, se emplea un protocolo recientemente publicado (ROSETTA design-HMMer), para guiar el modelado por homología en la obtención de modelos confiables de la estructura tridimensional de estas dos proteínas de arabidopsis. Su interacción con iones de magnesio y pirofosfato se analiza *in silico*.

Palabras clave: Estructura 3D de proteínas, modelado comparativo estructural, pirofosfatasas EC 3.6.1.1.

Abreviations: PPi, pyrophosphate; siPPaI, soluble inorganic pyrophosphatase of the family I; AtPPa1, *Arabidopsis thaliana* soluble inorganic pyrophosphatase isoenzime 1; AtPPa4, *Arabidopsis thaliana* soluble inorganic pyrophosphatase isoenzyme 4; MD, molecular dynamics; Rd.HMM, ROSETTA design- HMMer protocol.

Introduction

Soluble inorganic pyrophosphatases (E.C. 3.6.1.1) are ubiquitous enzymes in living cells [1]. They fulfill an essential role because their activity recycles the pyrophosphate produced by many anabolic reactions [2, 3]. In general, these enzymes are highly specific for pyrophosphate and use divalent metal cations as essential activators [4].

Judging from the information available in current sequence databases, the soluble inorganic pyrophosphatases from all Eukaryotes and many bacteria belong to the family I (siPPaI), which are Mg^{2+}-dependent enzymes [1, 5]. The soluble inorganic pyrophosphatases from *Escherichia coli* [6] and *Saccharomyces cereviceae* [7] are the best studied enzymes of this group. These two enzymes are nearly perfect catalysts [4], but differ in their quaternary structure, because the siPPaI from bacterial sources studied to date are obligate-homohexamers [8], while the *Saccharomyces cereviceae* enzyme, considered as the prototype of Eukaryotic siPPaI, is an obligate homodimer [9].

In plants [10, 11] and some protists [12, 13, 14], the pyrophosphate is known to play additional roles related to the regulations of primary metabolism, sulphur metabolism and growth, although, the details of these roles are understood poorly [11, 12, 14].

In contrast to the fungal and bacterial enzymes, the siPPaI from *Arabidopsis thaliana* [5, 15], *Chlamydomonas reinhardtii* [5], and *Leshmania major* [16, 17] are active monomers. Of these last group of proteins. the kinetics of the isoforms 1 and 4 from *Arabidopsis thaliana* (AtPPa1 and AtPPa4, respectively) have been studied in more detail. These two isozymes exhibit a reduced catalytic efficiency and some other kinetic differences respect to the bacterial and yeast enzymes; in addition, they are similar to each other, but were found to differ from in their affinity for Magnesium [15].

The three-dimensional structures of siPPaI from many bacteria and from yeast have been determined [1, 7, 8, 9], but the 3D-structure form none of the monomeric siPPaI enzymes is available. The amino-acid-sequences of the monomeric siPPaI show similarities slightly over 40 % with some oligomeric enzymes with known 3D-structure-, and the isoforms 1 to 5 from arabidopsis are related to the bacterial siPPaI [1], therefore, these proteins have enough sequence similarity to support homology-modeling. Respectively, AtPPa1 and AtPPa2 were 46 and 44% similar (similarity matrix PAM40) to *Pyrococcus horikoshii* siPPaI (PDB 1UDE).

Here, we present three-dimensional reliable models of high quality of the soluble inorganic pyrophosphatase isoforms AtPPa1 and AtPPa4 obtained through a novel strategy based on a recently published protocol designated as Rd.HMM [18]. the

Rd.HMM protocol starts by removing the natural amino acid sequence from the 3D-structure and attempts to reconstruct many different amino acid sequences energetically compatible with this 3D-backbone coordinates, using ROSETTA design [19]. The resulting set of sequences (usually over 100) has now a sample of the amino acids that can be accommodated at each position, without destabilizing the 3D-structure under analysis. In this step, the function-related information is lost, because ROSETTA design has no information to retain the catalytic, binding and allosteric properties of the protein. The set of sequences is melded into an statistical device called hidden Markov model which is used to compare sequences from a protein sequence database to this set, finding all compatible sequences and producing a score and an expectancy value (E-value). When used to search the international protein sequence databases, the Rd.HMM protocol from PDB structures obtained by X-ray crystallography was able to find the natural amino acid sequence belonging to the protein of the corresponding PDB file, with a high score and low expectancy (high statistical significance). The score depends on the number of amino acid positions in the sequence of the database matching the Rd:HMM with high likelihood and, in consequence, longer proteins will produce higher scores. X-ray crystallography structures usually give scores around of 0.6 times the length of their amino acid sequence. NMR solved structures give values in the range of 0.3 to 0.5 times their sequence length. Acceptable modeled structures can give scores around 0.3 times their sequence length, or higher. Poor models give values close to zero, or negative scores [18]. Wrong models may score a different amino acid sequence (usually with close to zero or negative scores), *i.e.* not the one intended to represent, or none at all.

Based on the structural models presented here, with the aid of molecular dynamics simulations and molecular docking, we identified *in silico* the amino-acids possibly participating in the binding of this proteins to divalent metal cations and pyrophosphate. From these information the position of the putative active sites is deduced and compared to the available experimental data from the enzymes with known structures. Previous reports, based on the analysis of three-dimensional models of unknown quality for the AtPPa enzymes indicted that the AtPPa enzymes could be oligomers with very similar properties to the bacterial type, family I pyrophosphtases [1]. In contrast, the data present here indicate that the structure of these monomeric enzymes presents important differences with the bacterial and fungal enzymes.

Results

Obtention and quality evaluation of the three-dimensional models for the AtPPa1, AtPPa4 soluble inorganic pyrophosphatases

Starting 3D-models for the AtPPa1, and AtPPa4 soluble inorganic pyrophosphatases were obtained from the SAM-T08 server [20, 21, 22] and minimized using Hyperchem [23] under the Amber99 forcefield, as described in the methods section. These models were scored using Rd.HMM [18], and used to score the entire NCBI RefSeq database [24]. The starting Rd.HMM scores are included in table 1.

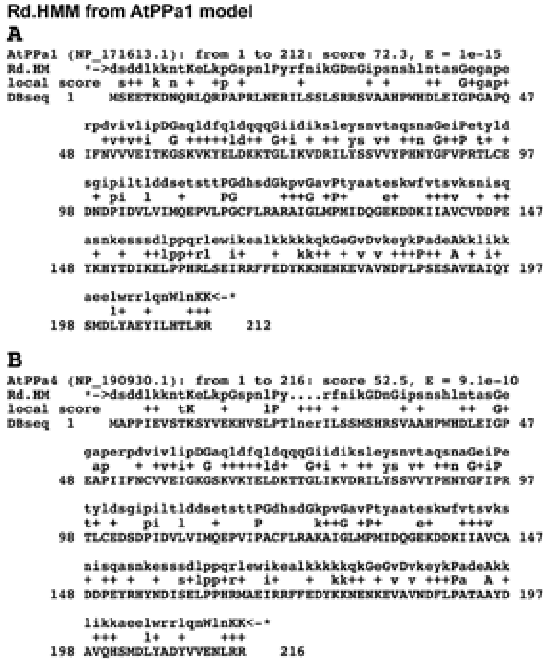
The first of these two models did recover the amino-acid sequence corresponding to the AtPPa1 with the highest score (table 1), and the score equals the length of the AtPPa1 amino-acid sequence times 0.34. Its alignment (Fig. 1A) was in frame with the Rd.HMM and free of gaps. Therefore, this first model should be considered as an acceptable approximation to the AtPPa1 three-dimensional structure [18].

Table 1. Scores for several siPPiaI with the Rd.HMM corresponding to different starting three-dimensional structures taken from the PDB (1UDE and 1E9G) or produced by homology-modeling, with or without further refinements, as described in the methods section.

Rd. HMM	hit #	Database ID[c]	Score[d]	Log E-value[e]	Biological source	Description
1UDE	2	PDB_1UDE	77.7	−16.49	*Pyrococcus furiosus*	siPPiasa, bacterial
1UDE	104	NP_182209.1	47.7	−7.49	*Arabidopsis thaliana*	siPPiasa; AtPPa3
1UDE	134	NP_171613.1	45.9	−6.92	*Arabidopsis thaliana*	siPPiasa; AtPPa1
1UDE	232	NP_179415.1	41.6	−5.64	*Arabidopsis thaliana*	siPPiasa; AtPPa2
1UDE	340	NP_190930.1	39.5	−5	*Arabidopsis thaliana*	siPPiasa; AtPPa4
1UDE	475	NP_192057.1	36.7	−4.17	*Arabidopsis thaliana*	siPPiasa; AtPPa5
AtPPa1[a]	1	NP_171613.1	72.3	−15	*Arabidopsis thaliana*	siPPiasa; AtPPa1
AtPP1[a]	34	NP_190930.1	52.5	−9.04	*Arabidopsis thaliana*	siPPiasa; AtPPa4
AtPPa4[a]	5	NP_190930.1	12.3	0.67	*Arabidopsis thaliana*	siPPiasa; AtPPa4
AtPPa1[b]	1	NP_171613.1	222.2	−60.05	*Arabidopsis thaliana*	siPPiasa; AtPPa1
AtPPa4[b]	1	NP_190930.1	153.5	−39.21	*Arabidopsis thaliana*	siPPiasa; AtPPa4

[a]Initial model after energy minimization. [b]Final model after molecular dynamics refinement and relaxation with Rosetta 3.1 relax-fast algorithm (see methods). [c]Codes starting with NP are Reference sequence accessions from NCBI, the code starting with PDB is a Protein Data Bank entry. [d]Rd.HMM alignment score. [e]The E-values are sequence and database size dependent, the ones presented here consider the current RefSeq NCBI database, roughly 5 million sequences. A positive Log(E-value) indicates the lack of statistical significance.

The model for AtPPa1 was improved using 5 ns molecular dynamics simulations (MD) at 313 °K in a water periodical box, in the presence of 0.15 M NaCl, at constant pressure, as described in methods. The trajectory was analyzed and the radius of gyration was observed to be close to a plateau after 4 ns. The frames in the last ns of the trajectory were used to isolate the most representative structure and subjected to structural minimization in GROMACS [25, 26]. This improved three-dimensional model was scored again with Rd.HMM, and the MD simulation was repeated once more. At the end of the second round of simulation, clustering and energy-minimization, the Rd.HMM score was in fact slightly smaller than after the first round. Therefore, the energy-minimized three-dimensional model after the first round of MD simulation was further relaxed using the ROSETTA 3.1 relax-fast protocol [27]. This last protocol was chosen because amongst the programs for the prediction of the three-dimensional structure of proteins, ROSETTA is the one that can render structures with the closest geometry to those found in the files form the Protein Data Bank [19].



Rd.HMM from AtPPa1 model

A

```
AtPPa1 (NP_171613.1): from 1 to 212: score 72.3, E = 1e-15
Rd.HM   *->dsddlkkntKeLkpGspnlPyrfnikGDnGipsnshlntasGegape
local score s++ k n +  +p +      +       + +    ++  G+gap+
DBseq  1   MSEETKDNQRLQRPAPRLNERILSSLSRRSVAAHPWHDLEIGPGAPQ 47

        rpdvivlipDGaqldfqldqqqGiidiksleysnvtaqsnaGeiPetyld
          +v+v+i  G +++++ld++ G+i + ++ ys v+ ++n G++P t+ +
     48 IFNVVVEITKGSKVKYELDKKTGLIKVDRILYSSVVYPHNYGFVPRTLCE 97

        sgipiltlddsetsttPGdhsdGkpvGavPtyaateskwfvtsvksnisq
         + pi  l   +    PG     +++G +P+    e+   +++v  + ++
     98 DNDPIDVLVIMQEPVLPGCFLRARAIGLMPMIDQGEKDDKIIAVCVDDPE 147

        asnkesssdlppqrlewikealkkkkkqkGeGvDvkeykPadeAkklikk
         +  +  ++lpp+rl  i+   +  kk++ + v v +++P++ A + i+
    148 YKHYTDIKELPPHRLSEIRRFFEDYKKNENKEVAVNDFLPSESAVEAIQY 197

        aeelwrrlqnWlnKK<-*
           l+  +    +++
    198 SMDLYAEYILHTLRR    212
```

B

```
AtPPa4 (NP_190930.1): from 1 to 216: score 52.5, E = 9.1e-10
Rd.HM   *->dsddlkkntKeLkpGspnlPy....rfnikGDnGipsnshlntasGe
local score    ++  tK   +    lP +++  +       + +    ++  G+
DBseq  1   MAPPIEVSTKSYVEKHVSLPTlnerILSSMSHRSVAAHPWHDLEIGP 47

        gaperpdvivlipDGaqldfqldqqqGiidiksleysnvtaqsnaGeiPe
         ap   + +v+i+ G +++++ld+  G+i + ++ ys v+ ++n G+iP
     48 EAPIIFNCVVEIGKGSKVKYELDKTTGLIKVDRILYSSVVYPHNYGFIPR 97

        tyldsgipiltlddsetsttPGdhsdGkpvGavPtyaateskwfvtsvks
        t+ +   pi  l   +    P      k++G +P+    e+   +++v
     98 TLCEDSDPIDVLVIMQEPVIPACFLRAKAIGLMPMIDQGEKDDKIIAVCA 147

        nisqasnkesssdlppqrlewikealkkkkkqkGeGvDvkeykPadeAkk
        + ++ +  +  s+lpp+r+  i+   +  kk++ + v v +++Pa  A +
    148 DDPETRHYNDISELPPHRMAEIRRFFEDYKKNENKEVAVNDFLPATAAYD 197

        likkaeelwrrlqnWlnKK<-*
         +++   l+  +    +++
    198 AVQHSMDLYADYVVENLRR    216
```

Fig. 1. Structurally aware sequence alignments of the AtPPa1 Rd.HMM consensus against the AtPPa1 (A) and AtPPa4 (B) amino acid sequences. The upper line is the consensus sequence. Uppercase letters indicate invariant positions [18]. The central line gives a local score, a blank space indicates no coincidence, a + symbol indicates a coincidence with low score, a lowercase letter indicates a coincidence with high probability, and a uppercase letters indicate a coincidence in an invariant position. The lower line is the sequence under consideration found by the Rd.HMM in the RefSeq [25] database.

The final model of the AtPPa1 protein is presented in figure 2A. This three-dimensional structure was analyzed with the Rd.HMM protocol and the score was now 1.05 times the amino-acid sequence length (Table 1). This score is above the 0.6-0.8 value found for the three-dimensional structure of proteins solved by X-ray crystallography, which reveals a bias in the Rd.HMM protocol for structures produced by the ROSETTA geometry-relaxation algorithm. This is not surprising, since Rd.HMM and the ROSETTA relax-fast algorithm share the same sidechain rotamer-database and the same energy forcefield to locate the energy minimum.

Nevertheless, the use of the ROSETTA rotamer-database and the ROSETTA energy score is not enough to increase the Rd.HMM score and give the impression of a biologically meaningful model, as revealed by the scores of the Rd.HMM from several ROBETTA models for the AtPPa1 sequence (Table 2). The ROBETTA server uses ROSETTA and homology-modeling to produce three-dimensional structures starting form an amino-acid sequence. The 5 models produced by the RO-

Table 2. Rd.HMM scores corresponding to different starting three-dimensional structures models of the siPPiaI AtPPa1 produce by the ROBETTA server [33].

Rd. HMM	hit #	Database ID[a]	Score[b]	Log E-value[c]	Biological source, description
1UDE	2	pdb|1UDE	77.7	−16.49	*Pyrococcus furiosus*, siPPiasa
1UDE	134	NP_171613.1	45.9	−6.92	*Arabidopsis thaliana*, siPPiasa 1
ROBETTA_1	1	pdb|1UDE	51.0	−12.38	*Pyrococcus furiosus*, siPPiasa
ROBETTA_1	25	NP_171613.1	30.4	−6.18	*Arabidopsis thaliana*, siPPiasa 1
ROBETTA_2	1	pdb|1UDE	47.8	−11.43	*Pyrococcus furiosus*, siPPiasa
ROBETTA_2	43	NP_171613.1	27.6	−5.47	*Arabidopsis thaliana*, siPPiasa 1
ROBETTA_3	1	pdb|1UDE	41.9	−9.64	*Pyrococcus furiosus*, siPPiasa
ROBETTA_3	34	NP_171613.1	17.6	−4.17	*Arabidopsis thaliana*, siPPiasa 1
ROBETTA_4	1	pdb|1UDE	39.9	−9.05	*Pyrococcus furiosus*, siPPiasa
ROBETTA_4	26	NP_171613.1	23.9	−5.17	*Arabidopsis thaliana*, siPPiasa 1
ROBETTA_5	1	pdb|1UDE	49.2	−11.82	*Pyrococcus furiosus*, siPPiasa
ROBETTA_5	25	NP_171613.1	30.4	−6.18	*Arabidopsis thaliana*, siPPiasa 1

[a]PDB code or RefSeq accession (see Table 1). [b]Rd.HMM alignment score. [c]Large negative Log(E-value) indicate high statistical significance (see Table 1).

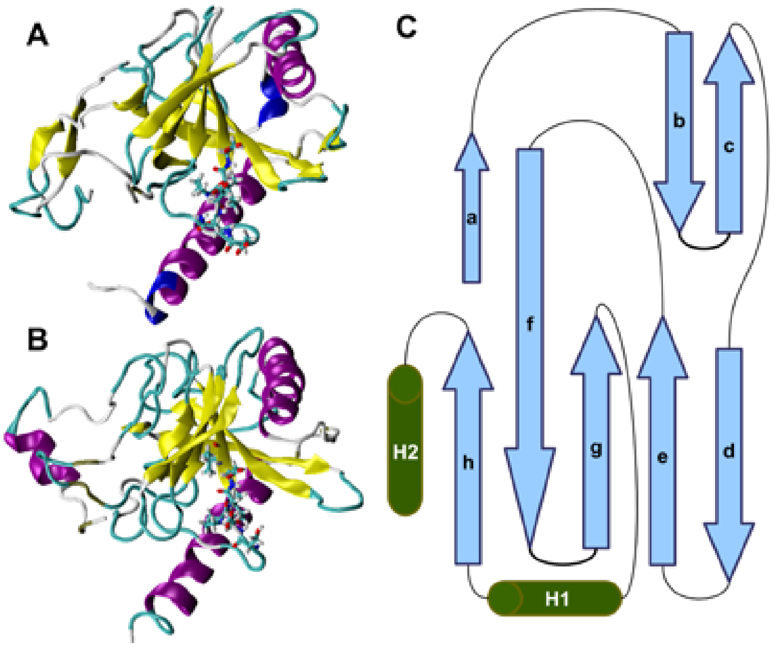


Fig. 2. Three-dimensional models of the *Arabidopsis thaliana* soluble inorganic pyrophospatases in an schematic representation. Both models are shown in a similar orientation. A) AtPPa1 model. B) AtPPa4 model. The classic DxPDxD conserved motif is shown as licorices. These images were prepared using VMD [28]. C) general topology of the core of the soluble inorganic pyrophosphatases from the family I.

BETTA server for the AtPPa1 amino-acid sequence rendered structures scoring better the amino-acid sequence of the PDB template chosen by the server (1UDE), than for the aminoacid sequence intended to model (AtPPa1). The scores for these last models were considerably smaller than the one from our final model, and were nearly half the score given to those sequences by the Rd.HMM for the 1UDE crystal itself. This means that the new ROBETTA models are not a better description of the AtPPa1 protein than was the structure of the bacterial protein than the 1UDE crystal structure.

In contrast to the above, the Rd.HMM from the AtPPa4 starting model did recover the amino-acid sequence of the corresponding amino acid sequence with a score of only 0.057 times its sequence length, marginal statistical significance (Log E-value close to 1, see table 1), and the alignment presented several gaps (not shown). In fact, the AtPPa1 model was a better approximation to the three-dimensional structure of the AtPPa4 protein, because amongst the sequences recovered from the RefSeq database, by the AtPPa1 Rd.HMM, the sequence for the AtPPa4 protein was found with a score of 0.24, a low E-value (see Table 1), and the corresponding alignment only showed a gap near the amino-terminal region (Fig. 1B).

Because the Rd.HMM alignments were found to bear a strong relationship with the structure [18], and taking advantage of the good quality of the model for AtPPa1 produced, we guided the homology modeling of AtPPa4 with Rd.HMM for AtPPa1 and used MODELLER 9v4 [30] to generate a structural model for this last sequence. To allow enough variability, 50 three-dimensional models were produced. Each model was scored with Rd.HMM, and the model with the highest Rd.HMM score and the best alignment was selected for a second MODELLER/Rd.HMM round, but now using the model generated in the first MODELLER try as input. After this second round the model for the AtPPa4 with the highest Rd.HMM score was better, than the best model in the first round, so a third round of MODELLER/Rd.HMM was performed. The best model in the third MODELLER/Rd.HMM round was as good as the best model in the previous round. Therefore this last model was relaxed by MD simulations and ROSETTA fast-relax, as described before for the AtPPa1 model. Results of the Rd.HMM scores for the starting and final models for AtPPa4 in table 1. The final model for the AtPPa4 protein has a score of 0.71 times its sequence length and the Rd.HMM alignment was now in-frame and free form gaps (not shown).

In this work, we selected the models produced by the SAM-T08 server, which are known to present structural defects, such as very long bonds and unreal bond angles at some positions in the model. There other structure prediction servers, such as I-TASSER [30, 31], and ROBETTA [32, 33]. However, we have scored several models for different proteins from these three servers, using the Rd.HMM protocol. In our test the rate of success in the prediction of biologically meaningful three-dimensional folding patterns for protein sequences, I-TASSER and ROBETTA were only marginally better than the energy minimized SAM-T08 models. On the other hand, I-TASSER and ROBETTA produce structures of higher quality, but may take several weeks to give a result, while SAM-T08 will answer after one or two days. In many cases, all three servers are able to produce biologically meaningful models, judging from the Rd.HMM scores of the resulting three-dimensional models, but the SAM-T08 server has a higher rate of success in the generation of models for amino acid sequences of membrane proteins (unpublished data).

Now, because in addition to an score of the appropriateness of a three-dimensional model, the Rd.HMM protocol provides a guide to improve an starting model, the SAM-T08 server may be a better choice for most cases. Other servers may be equally good, but we have not tested them.

At this point, it is worth mentioning that common schemes in homology modeling rely on the alignment of amino acid sequences between the target sequence and the amino acid sequence corresponding to the possible three-dimensional templates, and the alignment is based on the sequence conservation in the natural protein sequences. Such conservation is high at function-related residues because it derives from natural selection forces acting simultaneously on the active site, protein-ligand binding sites, protein-protein interaction regions, and most other functional features. Therefore, those function related residues become the key positions to guide the sequence alignment between the target sequence and the template sequence, and they will occupy equivalent positions in the final three-dimensional model.

In contrast, the Rd.HMM protocol [18] starts by removing the natural amino acid sequence and reconstructs many possible amino acids sequences using ROSETTA design [19]. In this step the function-related information is lost. Now the Rd.HMM alignments can be used to decide if a target amino acid sequence will fit into the template, because they include position-specific information for each amino-acid that contributes to increase the score, and this information can be used to guide the homology modeling. However, the function-related amino acid conservation has no influence on the alignment [18], and the amino acids participating of functional sites will not necessarily occupy equivalent positions in the model and the template, unless structural constraints so require.

Comparison of the three-dimensional models for the *Arabidopsis thaliana* soluble inorganic pyrophosphatases 1 and 4, with X-ray solved structures from other sources

The overall structure of the AtPPa1 and AtPPa4 three-dimensional models is shown in figure 2A and 2B, respectively. Both structures share a distorted 5-stranded ß-barrel with Greek-Key topology (8; Fig. 2C). In these proteins the loop between strands **d** and **e** is where the classic active site signature of this group resides (DXDPXD; residues 98-103 in AtPPa1, and 103-108 in AtPPa4; see licorices in figure 2A and 2B). In the enzyme form yeast, this loop presents an insertion and is nearly twice as long (not shown). As already mentioned, given the method followed, the similarities in the organization of the active site found are not forced by the sequence alignment given to the homology modeling program, but are a consequence of following the structural constraints of the amino acid sequence.

The two plant enzymes resemble the yeast enzyme, in that they show an extension in their N-terminal side, but differ in that they lack the extension on the C-terminal side. In addition, comparison of the model in figures 2A and 2B shows important differences in the folding of the N-terminal extension. To date, the procaryotic type enzymes with experimentally determined structure are all hexamers from bacterial sources, while the yeast enzyme is a dimer. In both enzymes the dissociation of the oligomers leads to a loss of activity. Thus, it could be argued that the extended amino acid sequences are associated to the need for increased stability in the dimeric enzyme. However, the plant enzymes only show the extension at their N-termini, but are active as monomers [15]. Thus, the N-terminal extension appears to be enough to make the monomers active, and, as data in figure 2 suggest, there seem to be more than one solution to this problem. In both enzymes, the extensions fold over the region equivalent to the subunit-subunit interface of the bacterial 1UDE protein.

The AtPPa genes 1 to 5 encode for proteins with putative cytoplasmic localization and very similar in sequence, except for the N-termini, where important differences appear. This variations in the N-terminal side are possibly related to the differences in the regulatory properties of these enzymes.

As already mentioned the AtPPa6 gene is more closely related to the yeast enzyme, and it is interesting that the only plant pyrophosphatase with higher similarity to the Eukaryotic enzyme is a chloroplastic protein [35], nevertheless, this enzyme is also active as a monomer [5].

Identification of the metal and substrate binding sites in the three-dimensional models for the *Arabidopsis thaliana* soluble inorganic pyrophosphatases 1 and 4

The basic kinetics of the AtPPa1 and AtPPa4 enzymes has been studied with the pure recombinant proteins [15]. These proteins have an absolute requirement for magnesium, and only manganese II was found to replace magnesium, but the activity is reduced roughly ten times. AtPPa1 and AtPPa4 were reported to differ in their affinity for free magnesium. In addition, calcium or other divalent cations have been shown to inhibit the enzyme [15]. AtPPa6 was reported to have an absolute requirement for magnesium [5, 35], but its kinetic mechanism, the number of metal binding sites and their relative affinities have not been analyzed in detail. These proteins also show important differences in the kinetic behavior, because they have a reduced k_{CAT}/K_M value, and they saturation kinetics with pyrophosphate shows high substrate inhibition [15]. In this report, the high substrate inhibition could not be explained by the formation of a metal chelate by the substrate, but was ascribed to a novel kinetic mechanism. Though very similar in their kinetic behavior, the AtPPa1 showed higher affinity for Mg^{2+} than the AtPPa4 [15].

Therefore it is of interest to locate and compare the magnesium binding sites in these models. To this aim, molecular dynamics simulations of these models were performed in the presence of NaCl and $MgCl_2$, as described in the methods section. The contacts with the ions were followed throughout the simulation trajectory, and those sites were the residence time of the metal was long (several ns). were considered as putative Mg binding sites. Due to its larger size, the exploration of the interaction of pyrophoshate with the AtPPa models though MD would take too long. We used Autodock 4.0 [41] to explore these interactions. These studies were performed in the free protein, then the Mg^{2+} ions were added to the molecule at the sites previously identified during the MD simulations, and the overall geometry was minimized. With those complexes that fell close to the expected active sites a 5 ns MD simulation was performed to analyze the permanence of the complex. Both AtPPa1 and AtPPa4 showed at least 2 sites for the binding of Mg^{2+}. These sites were in the vecinity of the DXDPXD active site motif. It must be said that the whole region bears overall negative charge, due to the abundance of negatively charged residues.

Figure 3 shows a superposition of the locations identified by Autodock 4.0 as possible pyrophosphate binding sites for AtPPa1 and AtPPa4. Despite the significant similarities in the three-dimensional structure of these two proteins, Autodock 4.0 found different binding conformations of pyrophosphate to the AtPPa4 protein (Fig. 3B, site 1), but all fell around a region where the putative active site residues are present. In contrast, in AtPPa1, the binding of pyrophosphate comprised three sites (Fig. 3A), one roughly corresponding to the same region found for AtPPa4 (site 1), a second site in between the active site and the C-terminal α-helix (site 2), and a third site in a region between the N-terminal extension and the external part of the core barrel (site 3). This last site is not only absent in AtPPa4, but it is also the one with the lowest energy of all of the sites. We expected the binding of pyrophosphate to occur with moderate energies, because Mg^{2+} was not present in this docking experiments. However, binding of pyrophosphate at site 3 in AtPPa1 seem not to require Mg^{2+}, since at this site, the interaction did occur with high (negative) energy. This site coincides with the location of a putative phosphorylation site at residue Ser 24. It is possible that the binding of the pyrophosphate reflects the ability of the site to accept the phosphorylated residue, or a regulatory site for the binding of pyrophosphate. This is one hypothesis that can be tested experimentally, thanks to the insight gained trough the present modeling exercise.

The complexes corresponding to pyrophosphate bound to site 1 in both AtPPa1 and AtPPa4 were complemented with the addition of Mg^{2+} ions at the positions were MD simulations have indicated as possible Mg^{2+} binding sites. In overall, up to 4 Mg^{2+} ions could be docked into this putative active site. The structure was then minimized using AMBER99 forcefield, and the resulting complex was compared to the structure of the yeast soluble inorganic pyrophosphatase in complex with phosphate and Mn^{2+} (1E9G). The comparison is shown in figure 4. In this image, both proteins were superimposed and the crystal structure of the yeast enzyme with its active site in complex with the product.

The docking of pyrophosphate to AtPPa4 was in a conformation much closer to the one observed for the product and the metal ions in the yeast soluble inorganic pyrophosphatase (Fig. 4A). However, the position of the ions did show some differences. In the case of AtPPa4 (Fig. 4B) the position of pyrophosphate was in a different binding pocket, and the Mg^{2+} ions were also located in the vicinity of this second binding site. This different binding modes reveal important differences in the overall organization of the two active sites. In particular, this second model presents a deep hydrophobic cavity in one of the active site walls. This cavity is where the pyrophosphate sits, and correspond to the pyrophosphate binding site with lowest population in the autodock trials, that appears below the main binding site.

While the resolution of the present complexes is not fine enough to allow for an estimation of interaction energies, the data indicate clear differences between the two models. Further studies are on their way to offer a more detailed picture of the differences between these two enzymes.

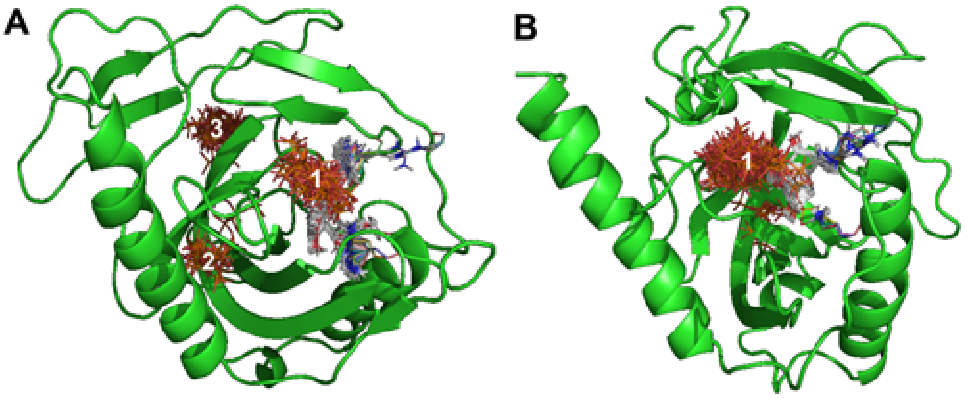


Fig. 3. Pyrophosphate possible binding sites identified by Autodock 4.0 in the three-dimensional model of the AtPPa1 (A) and AtPPa4 (B) proteins. Pyrophosphate is shown as licorice in orange (phosphorus) and red (oxygen). The residues belonging to the putative active site were allowed to move during the docking procedure (K62 R76 Y88 Y172 K173 for AtPPa1, and K6, R80, I92, Y176, K177 for AtPPa4). These sites are selected by the automatic annotation protocols of the NCBI web site [23] and are shown as licorices mostly in gray (carbon) and blue (nitrogen). The figure was prepared with PyMOL [43].

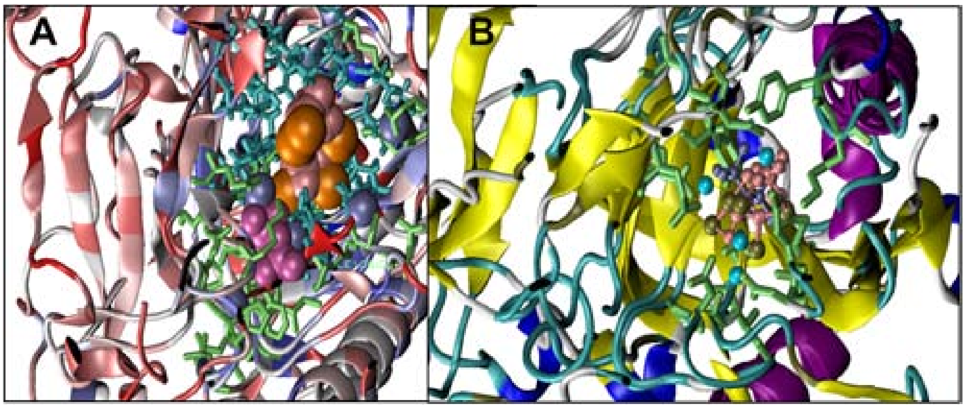


Fig. 4. Comparison of the putative Pyrophosphate and Mg binding sites in AtPPa1 and AtPPa4 models. For comparison. the structure of the crystal for the yeast siPPaI bound to phosphate and Mn was superimposed to the AtPPa three-dimensional models. A) Superimposed view of AtPPa1 three-dimensional model and yeast siPPa. Pink and brown spheres indicate the phosphate and manganese at the yeast actives site. Dark blue and cyan spheres indicate the pyrophoshate and magnesium ions (respectively) docked at site 1. Lime residues are the AtPPa1 sidechains in contact with the substrate. B) Superimposed view of active site from yeast isPPaI with 2 phosphates (pink) and manganese (orange) in the active site to AtPPa1 three-dimensional . In mauve pyrophosphate and dark blue magnesium in site 1 is shown. Cyan residues are the yeast sidechains in contact with the phosphate and manganese, while lime residues are AtPPa1 residues en contact with pyrophosphate and Mg^{2+}.

Methods

Obtention of the models for the three dimensional structure of the Arabidopsis inorganic pyrophosphatase proteins, isoforms 1 and 4

As a first step the sequences of the siPPaI isoforms 1 and 4 from *Arabidopsis thaliana* (AtPPa1 and AtPPa4, respectively) were blasted against the sequences in the PDB database. Those PDB files with the closer similarity were the siPPaI from *Pyrococcus horikosii* (PDB 1UDE) and the siPPaI from *Sulfolobus acidocaldarius* (PDB 1QEZ), with similarities above 40%. Starting models for the three-dimensional fold of AtPPa1 and AtPPa4 were obtained from the SAM-T08 server [20, 21, 22], this server was preferred over I-TASSER [31, 32], and the ROBETTA [33, 34] servers, because the response of the first takes two or three days at most, while the other two may take from weeks to months. However, SAM-T08, I-TASSER and ROBETTA are amongst the top programs in the CASP protein structure prediction contests [36]. The models send by the SAM-T08 improved through energy minimization with molecular mechanics, under the Amber 99 force-field using Hyperchem 7.5 (Hypercube, Inc.). This program was selected because it uses internal topologies to assign the correct atom connectivity to the model's atoms, instead of atom-atom distances. The imported models were minimized using a slow procedure to introduce the smaller distortion possible to the fold. In step A, the side-chains were fixed and the backbone minimized for 50 cycles or until the gradient was smaller than 0.1 kcal mol^{-1} $Å^{-2}$. Then in step B the backbone was fixed and the side-chains were minimized for 25 cycles or until the gradient was 0.25 kcal mol^{-1} $Å^{-2}$. Steps A and B were repeated until both steps reached the target gradient. Then all atoms were released and a minimization was performed until the gradient was below 0.1 kcal mol^{-1} $Å^{-2}$.

The models were scored using the Rosetta design-HMMer protocol (Rd.HMM) published elsewhere [18]. In this protocol, a three-dimensional model of a protein is considered to be very close to an equilibrium structure, similar to those found in crystals, if it retrieves for the database the sequence intended to model, with an score close to 0.6 times the length of its aminoacid sequence, and the alignment produced by the Rd.HMM neither does show gaps, nor a frame-shift. The score for the amino acid sequence intended to model should be amongst the top scores (ideally the first), and the sequences in this group should present high sequence similarity amongst them, (usually above 90% identity). Rd.HMM was performed using 13 intermediates with randomized sequences and each was reconstructed 11 times. The searches were done against the RefSeq-protein sequence database at NCBI [23].

An Rd.HMM score of 0.3 times the amino-acid sequence length of the model is considered acceptable, but improvements should be attempted, if the score is lower than this last number then the model must be improved. Even with a good score, if the HMM alignment has gaps or a frame-shift the model requires further work. Completely wrong models will fail to retrieve the sequence intended to model and some may retrieve nothing.

Further improvement of the models was obtained with three different strategies:

I. When gaps or a frame shift were detected in the Rd.HMM alignments, the structurally aware sequence alignments produced by the Rd.HMM from the AtPPa1 model were converted to PIR format and fed into MODELLER 9v4 [29, 30] to produce several new models. Each model was then scored Rd.HMM and the model with highest score and better alignment was selected.

II. The alignment of the Rd.HMM protocol can be free of a frame shift and lack gaps, but some sections may show regions of poor local score (absence of coincidences in the local score line of the alignment, see figure 1A, positions 1 to 47). In this case, the model was relaxed using molecular dynamics simulations, as described below. The central conformer from the most populated cluster was selected for energy minimization and scored using the Rd.HMM protocol. When required, the simulations were extended, and clustering repeated, until no further improvements in the Rd.HMM score were observed.

III. Models with Rd.HMM scores in the range of 0.25 to 0.5 times the corresponding sequence length, the alignments were free of gaps or a frame shift, and the local score line of the alignment did not show long blank sections, the overall geometry was improved using ROSETTA relax protocol. For this task the fast relax protocol implemented in ROSETTA version 3.1 [27] was used. After relaxation, the model was scored again with Rd.HMM.

Molecular dynamics simulations

Molecular dynamics simulations were performed using GROMACS [37, 38], and the GROMOS 53a6 forcefield [38]. Simulations were performed with an integration interval of 2 fs, in explicit water, with 0.15 M NaCl, at 313 K, and constant pressure. Electrostatics was accounted using particle-mesh Ewald, and pressure was controlled using a Berendsen barostat. After 5 to 10 ns simulation, the conformers in the last ns of the trajectory were clustered using the Jarvis-Patrick algorithm [40] as implemented in the GROMACS package [37, 38].

Molecular dynamics simulations were also performed in the presence of $MgCl_2$, alone or in combination with NaCl, at 303 K. Other conditions were as above. The resulting trajectories were analyzed to determine the amino acids making contact with the ions along the simulation and the stability of such contacts. It is worth noting that, the MD here employed used only molecular mechanics, and in the case of Mg^{2+} the formation of covalent coordination bonds (true metal chelates) is not considered. Yet, the simulations revealed sites where the divalent metal ions had long residence times, and these sites were considered as possible metal-binding sites. The amino acid residues participating in such contacts were recorded along the trajectories.

Molecular docking of pyrophosphate

Possible molecular docking sites for pyrophosphate into the AtPPa models were explored with the use of Autodock 4.0 and Autotools [41].

The resulting protein-pyrophoshate complexes were edited to add the bound Mg^{2+} at the sites previously identified and then MD simulations were performed in the presence of additional free Mg^{2+} ions. The topology for the fully ionized pyrophosphate was prepared using the automated topology builder [42], manually curated, and validated with the calculation of the solvation free energy using a classic free energy perturbation method, as recommended for the GROMOS G53a6 force field [39].

Acknowledgments

We thank Dr. Martín González-Andrade for his technical help. We thank the three anonymous reviewers for their insightfull comments and sugestions. This work was funded by CONACyT, PAPIIT-DGAPA-UNAM, and UNAM Supercomputing-DGTIC-UNAM grants. E. E. Hernández-Domínguez is recipient of a Ph.D. scholarship from CONOCyT, S. Gaythan-Mondragón is recipient of an M. Sc. scholarship from CONACyT, and L. Rosales-León was recipient of a postdoctoral fellowship form DGAPA, UNAM.

References

1. Sivula, T.; Salminen, A.; Parfenyev, A. N.; Pohjanjoki, P.; Goldman, A.; Cooperman, B. S.; Baykov, A. A.; Lahti, R. *FEBS Lett.* **1999**, *454*, 75-80.
2. Imsade, J.; Handler, P. in: *The Enzymes*, Vol. 4, Boyer, P.D.; Lardy, H.; Myrbäck, K., Eds., Academic Press, New York, 2a ed. **1961**, pp. 281-304.
3. Korngberg, A. in: *Horizons in Biochemistry* Kasha M, Pullman B, Eds. New York, Academic Press **1962**, pp. 251-254.
4. Cooperman, B. S.; Baykov, A. A.; Lathi, R. *Trends Biochem. Sci.* **1992**, *17*, 262-266.
5. Gómez-García, M. R.; Losada, M.; Serrano, A. *Biochem. J.* **2006**, *395*,211-221.
6. Hyytiä, T.; Halonen, P.; Salminen, A.; Goldman, A.; Lahti, R.; Cooperman, B. S. *Biochemistry* **2001**, *40*, 4645-4653.
7. Oksanen, E.; Ahonen, A. -K.; Tuominen, H.; Tuominen, V.; Lahti, R.; Goldman, A; Heikinheimo, P. *Biochemistry* **2007**, *46*, 1228-1239.
8. Teplyakov, A.; Obmolova, G.; Wilson, K. S.; Ishii, K.; Kaji, H.; Samejima, T.; Kuranova, I. *Protein Sci.* **1994**, *3*, 1098-1107.
9. Harutyunyan, E. H.; Kuranova I P; Lamzin V S.; Dauter Z.; Wilson K S. *Eur. J. Biochem.* **1996**, *239*, 220-228.
10. Farré, E. M.; Geigenberger, P.; Willmitzer, L.; Trethewey, N. *Plant Physiol.* **2000**, *123*, 681-688.
11. Farré, E. M.; Tech, S.; Trethewey, R.N.; Fernie, A. R.; Willmitzer, L. *Plant Mol. Biol.* **2006**, *62*, 165-179.
12. Pérez-Castiñeira, J. R.; Gómez-García, R., Lopez-Marqués, R. L.; Losada, M.; Serrano, A. *Int. Microbiol.* **2001**, *4*, 135-142.
13. Pérez-Castiñeira, J. R.; Alvar, J.; Ruiz-Pérez, L. M.; Serrano, A. *Biochem. Biophys. Res. Commun.* **2002**, *294*, 567-573.
14. Mi-ichi F, Abu Yousuf M, Nakada-Tsukui K, Nozaki T. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 21731-21736.
15. Navarro-De La Sancha, E.; Coello-Coutiño, M. P.; Valencia-Turcotte, L. G.; Hernández-Domínguez, E. E.; Trejo-Yepes, G.; Rodríguez-Sotres, R. *Plant Sci. (Shannon, Irel.)* **2007,** *172*, 796-807.
16. Gómez-García, M.R.; Ruiz-Pérez, L.M.; González-Pacanowska, D.; Serrano, A. *FEBS Lett.* **2004**, *560*, 158-166.
17. Gómez-García, M. R.; Losada, M.; Serrano, A. *FEBS J.* **2007**, *274,* 3948-3959.
18. Martínez-Castilla, L. P.; Rodríguez-Sotres, R. *Plos One* **2010,** *5*, e12483.
19. Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. *Science* (Washington, DC, U. S.) **2003,** *302*, 1364-1368.
20. http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html; accessed in January 2010.
21. Karplus, K.; Karchin, R.; Draper, J.; Casper, J.; Mandel-Gutfreund, Y.; Diekhans, M.; Hughey, R. *Proteins* **2003,** *53* Suppl 6, 491-496.

22. Karplus K.; Barrett C.; Hughey, R. *Bioinformatics* **1998**, *14*, 846-856.
23. Hypercube, Inc. *Hyperchem professional 7.5. Mollecular Modelling Software.* Gainesville, FL, USA. 2007.
24. Sayers, E. W.; Barrett, T.; Benson, D. A.; Bolton, E.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; Dicuccio, M.; Federhen, S.; Feolo, M.; Geer, L. Y.; Helmberg, W.; Kapustin, Y.; Landsman, D.; Lipman, D. J.; Lu, Z.; Madden, T. L.; Madej, T.; Maglott, D. R.; Marchler-Bauer, A.; Miller, V.; Mizrachi, I.; Ostell, J.; Panchenko, A.; Pruitt, K. D.; Schuler, G. D.; Sequeira, E.; Sherry, S. T.; Shumway, M.; Sirotkin, K.; Slotta, D.; Souvorov, A.; Starchenko, G.; Tatusova, T. A.; Wagner, L.; Wang, Y.; John Wilbur, W.; Yaschenko, E.; Y. E. J. *Nucleic Acids Res.* **2010,** *38*, D5-16.
25. Lindahl, E.; Hess, B.; Van Der Spoel, D. *J. Mol. Model.* **2001,** *7*, 306-317.
26. Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435-447.
27. Raman, S.; Vernon, R.; Thompson, J.; Tyka, M.; Sadreyev, R.; Pei, J.; Kim, D.; Kellogg, E.; Dimaio, F.; Lange, O.; Kinch, L.; Sheffler, W.; Kim, B.-H.; Das, R.; Grishin, N. V.; Baker, D. *Proteins* **2009,** *77* Suppl 9, 89-99.
28. Humphrey, W.; Dalke, A.; Schulten, K. *J Mol. Graph.* **1996,** *14*, 33-8, 27-8.
29. http://salilab.org/modeller/; accessed in February 2010.
30. Fiser, A.; Sali, A. *Methods Enzymol.* **2003,** *374*, 461-491.
31. http://zhang.bioinformatics.ku.edu/I-TASSER; accessed in June 2009.
32. Zhang, Y. *BMC Bioinf.* **2008,** *9*, 40.
33. http://robetta.bakerlab.org/ accessed in June 2009.
34. Chivian, D.; Baker, D. *Nucleic Acids Res.* **2006**, *34*, e112.
35. Schulze, S.; Mant, A.; Kossmann, J.; Lloyd, J. R. *FEBS lett.* **2004,** *565*, 101-105
36. Shi, S.; Pei, J.; Sadreyev, R. I.; Kinch, L. N.; Majumdar, I.; Tong, J.; Cheng, H.; Kim, B.; Grishin, N. V. *Database (Oxford)* **2009**, 2009, bap003.
37. Lindahl, E.; Hess, B.; Van Der Spoel, D. *J. Mol. Model.* **2001,** *7*, 306-317.
38. Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008,** *4*, 435-447.
39. Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. *J. Comput. Chem.* **2004,** *25*, 1656-1676.
40. Jarvis R A.; Patrick, E. A. *IEEE Trans. Comput.* **1973**, C22, 1025-1034.
41. Morris, G. M.; Huey, R.; Olson, A. J. in: *Current Protocols in Bioinformatics* Chapter 8, Baxevanis, A. D., Ed., Wiley Online Library, ebook, **2008**, Unit 8.14.
42. http://compbio.biosci.uq.edu.au/atb; accessed in October 2010.
43. DeLano WL. *PyMOL 1.3. Molecular Graphics System,* Available at: http://www.pymol.org. 2002. Accessed September 13, 2009.