J. Braz. Chem. Soc., Vol. 22, No. 6, S1-S10, 2011.
Printed in Brazil - ©2011 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

Supplementary Information

SI

Purification, Characterization and Structural Determination of UDP-*N*-Acetylglucosamine Pyrophosphorylase Produced by *Moniliophthora perniciosa*

Manoelito C. Santos Junior,[a] Priscila A. Gonçalves,[a] Alex G. Taranto,[b] Maria G. B. Koblitz,[c] Aristóteles Góes-Neto,[d] Carlos P. Pirovani,[e] Júlio C. M. Cascardo,[e] Sandra H. da Cruz,[f] Russolina B. Zingali,[g] Gonçalo A. G. Pereira,[h] Cristiano V. Dias[d] and Sandra A. de Assis*,[a]

[a]Departamento de Saúde, [c]Departamento de Tecnologia and [d]Departamento de Ciências Biológicas, Universidade Estadual de Feira de Santana, 44031-460 Feira de Santana-BA, Brazil

[b]Laboratório de Bioinformática, CCO, Universidade Federal de São João Del-Rei, 35501-296 Divinópolis-MG, Brazil

[e]Laboratório de Genômica e Expressão Gênica, DCB, Universidade Estadual de Santa Cruz, 45650-000 Ilhéus-BA, Brazil

[f]Departamento de Agroindústria, Alimentos e Nutrição, Escola Superior de Agricultura Luiz de Queiroz, ESALQ, Universidade de São Paulo,13418-900 Piracicaba-SP, Brazil

[g]Centro de Ciências da Saúde,Universidade Federal do Rio de Janeiro, Ilha do Fundão, 21941-590 Rio de Janeiro-RJ, Brazil

[h]Instituto de Biologia, Universidade Estadual de Campinas, CP 6109, 13083-970 Campinas-SP, Brazil

Figure S1. Production of *M. perniciosa*.

Primary Structure Determination

Sequence data genomic of M. perniciosa

The strategy adopted by the Genome Project for sequencing the fungus *M. perniciosa* genome was to assemble a library of shotgun genomic DNA (gDNA) and sequence fragments (reads) selected at random (Figure S2), then to compare each read with genes deposited in GenBank using the tBLASTx[1] program from the National Center for Biotechnology Information (NCBI)[2] to find significant similarity between sequenced reads and known genes. This favors the identification and characterization of genes, which can be done during sequencing, without the need to complete assembly of the genome, thus saving time and money.

With the development of the genome project of *M. perniciosa*, a bioinformatics system was constructed at UNICAMP that automated the entire process of acquiring and comparing sequences, creating a friendly interface through which researchers from various fields can explore the database with the aid of simple tools (such as search by keyword). This allowed the location of reads that had some similarity to sequences already characterized, as well as genes that code for the UDP-*N*-Acetylglucosamine pyrophosphorylase.

*e-mail: sandraassis@uefs.br

Search for sequences homologous to UDP-GlcNAc pyrophosphorylase in eukaryotic organisms

Through a search for keywords in the NCBI website (http://www.ncbi.nlm.nih.gov/) the following cDNA sequences were identified: UAP1 *Homo sapiens* (AC: NM_003115), (AC: NM_003115), *Drosophila melanogaster* (AC: NM_164690), *Caenorhabditis elegans* (AC: NM_065376), *Neurospora crassa* (AC: EAA34867), *Schizosaccharomyces pombe* (AC: NP_596832), *Saccharomyces cerevisiae* (AC: NP_010180), *Candida albicans* (AC: AB011003) and *Encephalitozoon cuniculi* (AC: AL590450). These sequences were used to infer the organizational structure of the gene encoding the enzyme UDP-GlcNAc pyrophosphorylase (EC: 2.7.7.23) of *M. perniciosa*. These were also used to identify the locus for each enzyme in the genomes by BLAST search (Altschul *et al.*[1]).

Sequence analyses

In the search for conserved regions, a comparison of the amino acid sequences was achieved using the alignment program CLUSTAL W 1.82 (http://www.ch.embnet.org/software/ClustalW.html)[3] from the European Bioinformatics Institute (EBI). Possible conserved domains were inferred using the programs from the PROSITE database (http://us.expasy.Org/PROSITE/)[4] and Prodomo (http://protein.toulouse.inra.fr/prodom/2002.1/htm/home.php).[5] Pfam (http://pfa m.wustl.edu / hmmsearch.shtml)[6] was used to confirm the protein family to which the contig had been linked.

Results

Searching sequences

The search for sequences similar to UDP-GlcNAc pyrophosphorylase led to identification of six sequences (reads), which allowed the formation of a contig (Table S1).

Table S1. Number of reads identified with possible similarity with the enzyme UDP-GlcNAc pyrophosphorylase. The search for sequences performed through the database of the Genome Project *M. perniciosa*

Reads Selected	Similarity (BLAST)
CP02-PF-000-002-E09-UE.R	Similar to gene Qri1p of *Saccharomyces cerevisiae*
CP02-S2-000-085-C01-UC.F	Similar to gene BcDNA.LD24639 of *Drosophila melanogaster*
CP02-S2-028-248-F09-UE.R	Similar to gene Qri1p of *Saccharomyces cerevisiae*
CP02-S2-033-367-F08-UE.F	Similar to gene Ugp1p of *Saccharomyces cerevisiae*
CP02-S2-000-024-F07-EM.R	Similar to gene AgX1 of *Homo sapiens*
CP02-PF-000-002-E03-UE.F	Plate functional

Alignment

Taking into account the consensus region proposed by Mio *et al.*[7] for all UDP-sugar pyrophosphorylases ([L (X) 2GXGTXM (X) 4PK], where X represents any amino acid), and their probable involvement in the catalytic activity of the enzyme, an alignment with UDP-GlcNAc pyrophosphorylase eukaryotic was already identified (Figure S2), which allowed the region to propose a likely consensus sequence [(A / S) GGQXTRLG (X) 3PKG] for eukaryotic UDP-GlcNAc pyrophosphorylases (Figure S3).

Analysis of contig

The contig generated from the reads, located in the database of the *M. perniciosa* Genome Project, has a length of 1739 base pairs. A tBLASTx against the world bank of genes, the NCBI GenBank, was performed to search for similar sequences. The tBLASTx translates the contig in its six possible amino acid sequences (frames and -1-2-3 +1 +2 +3) and these are then compared with the database to look for similarityies. The tBLASTx search revealed two regions of the contig similar to UDP-GlcNAc pyrophosphorylase *N. crassa*. The two regions were located in different frames, possibly due to the existence of an intron, which probably changed the reading frame of the polypeptide translated from the DNA sequence.

The deduced amino acids of the exons of the contig and the deduced amino acid gene *N. crassa* were aligned (CLUSTAL W 1.82) in an attempt to map the coverage of the likely gene contig (Figure S4). We observed coverage of over 50%, comprising virtually the entire C-terminal region of protein *N. crassa* (Figure S5).

References

1. Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J.; *Nucleic Acids Res.* **1997**, *25*, 3389.
2. http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/index.html. Acessed in April 2007.
3. http://www.ch.embnet.org/software/ClustalW.html. Acessed in September 2008.
4. http://us.expasy.Org/PROSITE. Acessed in October 2009.
5. http://protein.toulouse.inra.fr/prodom/2002.1/htm/home.php. Acessed in April 2007.
6. http://pfa m.wustl.edu / hmmsearch.shtml. Acessed in August 2009.
7. Mio, T.; Yabe, T.; Arisawa, M.; Yamada-Okabe, H.; *J. Biol. Chem.* **1998**, *273*, 14392.

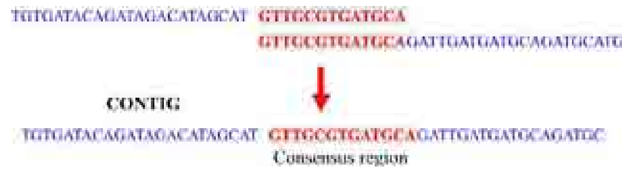

Figure S2. Clustering process. The sequences of reads similar to UDP-GlcNAc pyrophosphorylase were aligned by the algorithm pHeader Pharpar, which seeks areas of consensus (or overlap). This enables a group to form a larger sequence, increasing the coverage area of the gene.

Figure S3. Alignment of amino acid sequences of eukaryotic organisms. The amino acid sequences of UDP-GlcNAc pyrophosphorylase of UAP1 *H. sapiens*, *D. melanogaster*, *C. elegans, N. crassa, S. pombe, S. cerevisiae, C. albicans and E. cuniculi*, were aligned using the program CLUSTAL W 1.82. Identical amino acids are identified by (*) conservative substitutions by (:) and semiconservative substitutions by (..).

```
dme            130 GLLQISNGHVAVLLMA GGQGTRLGFDHPKGMYDVGLQS--RKTLF 173
Homo sapiens    94 GLFQISQNKVAVLLLA GGQGTRLGVAY PKGMYDVGLPS--RKTLF 136
Candida         94 GLKAIGNGEVAVLLMA GGQGTRLGSSAPKGCFNIELPS--QKSLF 137
scc             94 GLEAIGKGEVAVILMA GGQGTRLGSSQPKGCYDIGLPS--KKSLF 137
Neurospora      99 GLQLIADNKVAVVLMA GGQGTRLGSSAPKGCFDIGLPS--HKSLF 142
spo             88 GLREIARGHVAALVLA GGQGTRLGFAGPKGCFRLGLPN--NPSIF 131
Caenorhabditis  92 GMDAIGRGEVCAIVLAGGQATRLGSSQPKGTIPLGINASFGDSLL 137
Encephalitozoon 30 GERLLREKKLGVVILSGGQGTRLGSDEPKGLFKIKGKT------- 68
                    *  :  .: .:::***.****  *** :
```

Consensus

Figure S4. Comparison of the likely domain of UDP-GlcNAc pyrophosphorylase. The amino acid sequences from *H. sapiens, D. melanogaster, C. elegans, N. crassa, S. pombe, S. cerevisiae, C. albicans* and *E. cuniculi* were aligned using the program CLUSTAL W 1.82. Amino acids identical among all proteins are highlighted in red. The likely binding site for N-Acetylglucosamine-1-phosphate is indicated by the arrow, and the consensus sequence [(A / S) GGQXTRLG (X) 3PKG] was defined for all UDP-GlcNAc pyrophosphorylases.

```
Neurospora  MAVDPVPTPEQVSELKDKYTNAGQGQVFTFYDSLSSEEQAQLYKQLAGFDPLYINKIAAK 60

Neurospora  ALTPQSSESEKPTLEPLPDSARASTLDSDKQTQDEWWNRGLQLIADNKVAVVLMAGGQGT 120

Neurospora  RLGSSAPKGCFDIGLPSHKSLFQIQAERIARLQVLASQRREQAGSPVVPWYVMTSGPTRK 180

Contig      ------------------------GTLPCLTMDGKVLLGSPSHVAVAPDGNGGLYAATRS 36
Neurospora  ATEDFFKTNNYFGLSPDQVIIFEQGVLPCISNDGKILLESKSRVAVAPDGNGGIYNALVD 240
                                    *.***:: ***:** * *:**********:* *  .

Contig      PLSPKDKSRTVLSDLAKRKILYVVHAYCVDNCLVRVADPVFLGYSIQKQADCAAKVVPKTR 96
Neurospora  --------AKVLDDMARRGIEHVHAYCVDNCLVKVADPVFIGYCASQNVDIGTKVVRKRN 292
             .**.*:*:* * :***********:******:** .:::* .:*** * .

Contig      PTESVGVVARRGDKFSVVEYSEISXDRQRNEIPDGRVVLRX------------------ 137
Neurospora  ATEPVGLILLKNGKPDVVEYSEIDDAVAAEEDPAQPGVLRFRAANIVNHYYSFRFLKSIP 352
            .**.**:: :..* .*******.    :* *    ***

Contig      --------HIARKKISHIDVETGELVKPSKPNGMKLELFVFDVFPYTER--FAVLEVERK 187

Neurospora  EWASNLPHHIARKKIPYADLESGETVKPEKPNGIKLEQFVFDVFPLIELSKFACMEVKRE 412
                    ******* .: *:*:** ***.****:*** *******  *   ** :**:*:

Contig      EEFSPLKNAPGTGSDDPETSRADLFSQHKRFLEHAGATVKDGVE--IEISPLVSYAGEGL 245
Neurospora  DEFSPLKNARGTGEDDPDTSKHDIMAQGRRWLEAAGAKFAEGAEIXGVEVSPLVSYCGEGL 472
            :******** ***.***:**: *:::* :*:** ***. :*.*  :*:******.****

Contig      ESVKGKTFSKSGLVESIEELDALL 269

Neurospora  QSYADR---KVVAVDRIE------ 487
            :* .:  *  *: **
```

Figure S5. Gene region of *Neurospora crassa* covered by the contig. The alignment of the deduced amino acid sequence from the contig and amino acids of the *N. crassa* UDP-GlcNAc pyrophosphorylase protein, using CLUSTAL W 1.82, helped define what percentage of the gene is covered and what region of the protein was identified. The portion bounded in gray bounding is area identified by Pfam (http://pfam.wustl.edu/hmmsearch.shtml) for UDP-GlcNAc pyrophosphorylase. Identical amino acids are identified by (*), conservative substitutions by (:) and semi conservative substitutions by (.).

Figure S6. Surface plots of activity of *M. perniciosa* as a function of temperature (°C) and pH: A) PyroMp I; B) PyroMp II; C) PyroMp III; D PyroMp IV.

Figure S7. Amino acid sequence of 1JV1 indicating the regions without determined 3D structure.

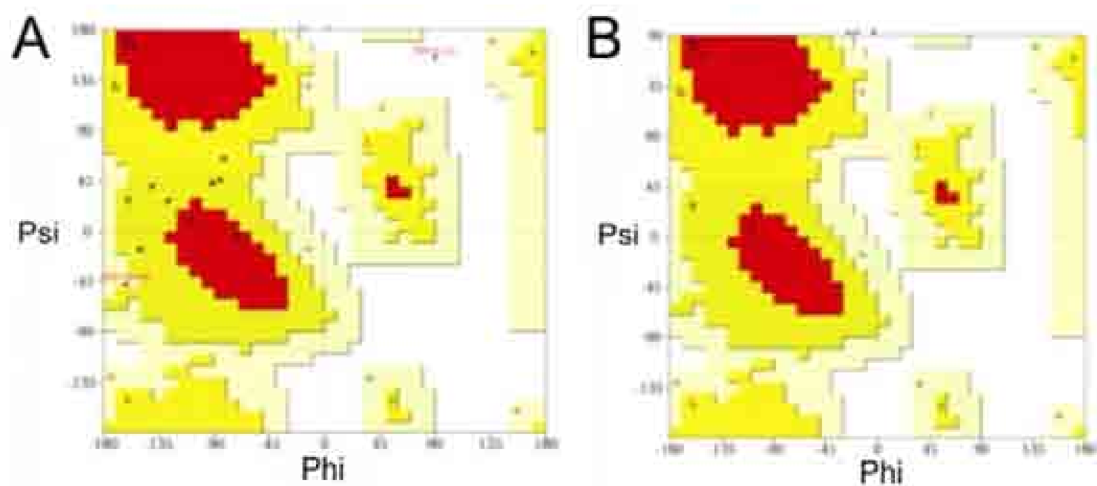

Figure S8. Ramachandran plot of GAP1 (A) and GAP2 (B).

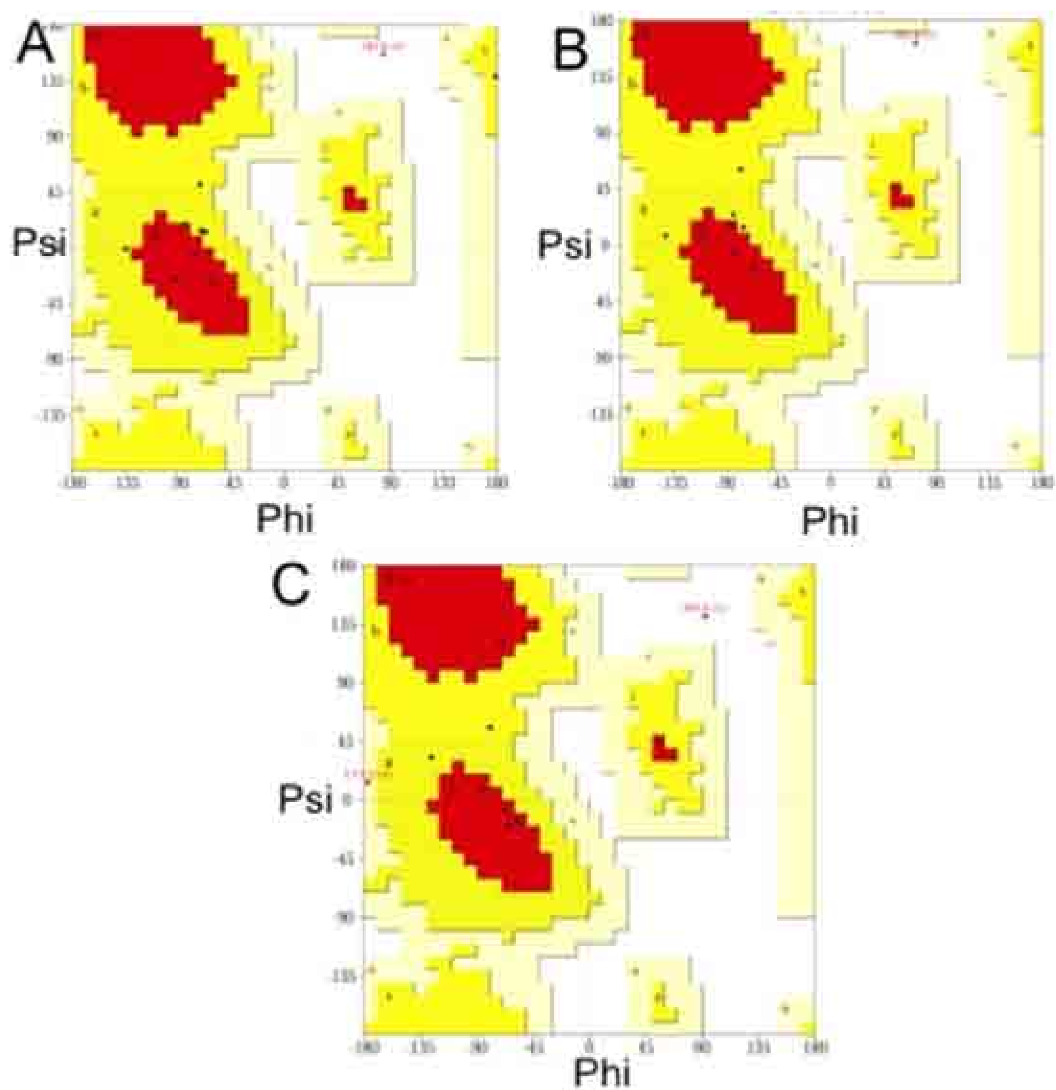

Figure S9. Ramachandran plot of the lowest energy structures structures generated by calculations of molecular dynamics at temperatures of 100 K (A), 200 K (B), and 300 K (C).

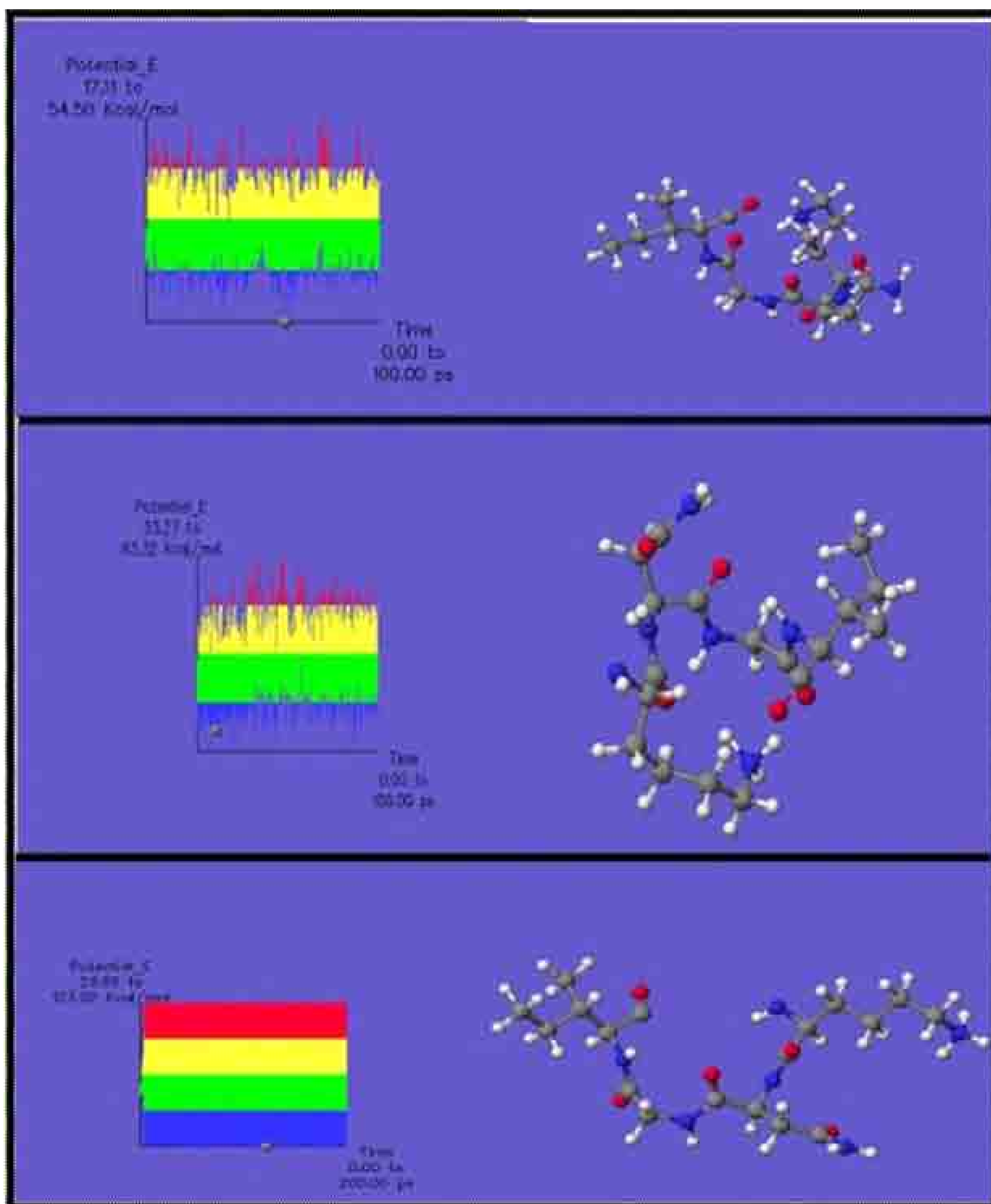

Figure S10. Results from simulations of molecular dynamics using temperatures of 100 K (A), 200 K (B), and 300 K (C), for GAP 2.

Figure S11. Ramachandran plot of 1JV1 after the structural addition of the GAP, corrections and neutralization.

Figure S12. Alignment by Clustal W between the protein 1JV1 and the sequence problem. "*" denotes conserved residues. ":" denotes conserved substitutions. "." denotes half-conserved region,.

Figure S13. Alignment between the 1JV1 template and the sequence of pyrophosphorylase. The rectangles indicate the regions where amino acids have been inserted into the sequence of pyrophosphorylase.

Figure S14. Ramachandran plot generated by PROCHECK for the MCJ4 model.

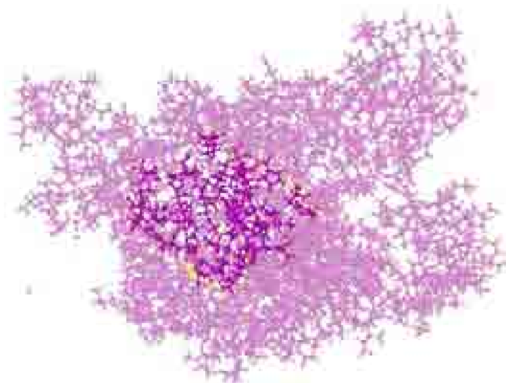

Figure S15. Active site of MCJ4 model.

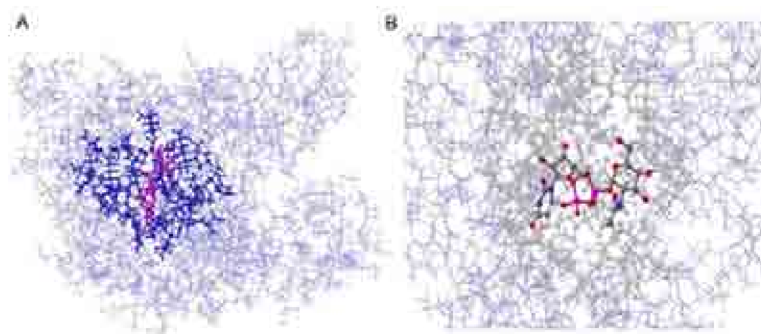

Figure S16. (A) Localization of the ligand in the active site of the MCJ4 model. (B) Detail of the ligand and the amino acids that are part of the active site of the enzyme.

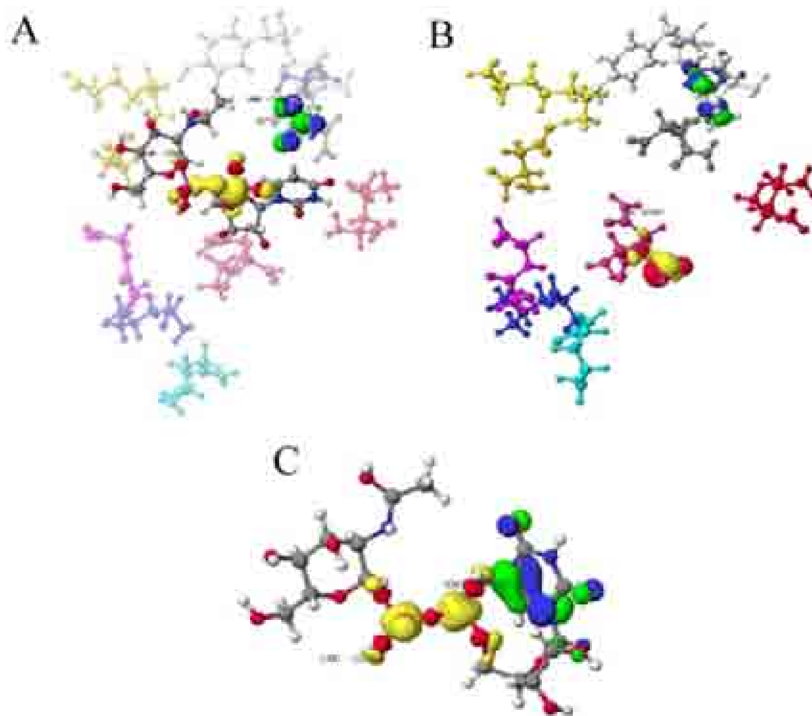

Figure S17. Orbitals of border atoms: (A) Ligand-protein complex. (B) Active site. (C) Ligand.