4 Discovering Active Sites of Homologous Proteins by Sequence Alignment

Overview

You have developed a model of binding for a lead compound in TB-PNP. Is the lead compound also going to bind to the human PNP leading to a loss of selectivity and potential side effect?

You need to determine how the ligand binds to human PNP so you can understand whether the ligand will exhibit high selectivity. Ideally you want to have a crystal structure of human PNP bound to the lead compound. But if the compound is selective it won’t bind and you can’t get a crystal structure. On the other hand, neither failure to get a crystal structure or failure to get a positive binding assay is not proof of selectivity because there are many reasons that you might fail to get a crystal structure or a positive assay. You need to know why your compound fails to bind. Modeling provides the basis for understanding the mechanism for selectivity.

In the case of PNP there is a crystal structure for the human enzyme, but there is no ligand bound to the protein. You must therefore find the active site in human PNP by comparison with the active site with TB-PNP and model ligands in the human PNP active site in order to understand whether the ligand is selective.

In this exercise, you will use CAChe to

- create a crevice surface of human PNP and quickly scan it for possible binding regions
- align the sequence of human PNP with the sequence of the homologous TB-PNP for which the active site is known
- identify the active site residues in human PNP from the alignment
- create a named atom group for the active site in human PNP
- compare human PNP’s active site with the crevice surface
- dock a ligand bound in the active site of TB-PNP into the active site of human PNP.

Background

The 2.75 Angstrom resolution crystal structure (`1ULA`) of *homo sapiens erythrocytes* purine nucleoside phosphorylase (human-PNP)[1] is available. In a previous exercise, you identified the active site in TB-PNP by locating the residues adjacent to the bound ligand ImmH. This technique cannot be used with the crystal structure for human-PNP (1ULA) because it lacks the bound ligand.

Crevice surfaces are a first quick step for locating regions of a protein that might be good binding sites.[2] The crevice surface colors the protein surface by the depth from an enclosing smooth surface. The result is that deep crevices within the protein where ligands will bind are colored blue and other areas are cream colored.

Human-PNP and TB-PNP[3] almost certainly evolved from a common ancestor. Both enzymes are trimeric, have a similar 3-dimension fold and catalyze the same reaction. Once nature discovers how to catalyze a reaction, she tends to reuse the active site in enzymes of related species. Therefore, active site residues and geometries are usually conserved in evolutionarily related enzymes. Thus, it is reasonable to suppose that the catalytic residues present in the human-PNP active site are the same as those of TB-PNP.

Multiple sequence alignment of human-PNP with other PNP enzymes related to it by evolution can therefore be used to identify the human-PNP active site.

In this exercise, you will identify the residues in human-PNP that align with the active site residues of TB-PNP and use superposition to confirm that the 3D structure of these residues in human-PNP are the same as that in TB-PNP. The alignment and superposition allow you to identify the active site in hyman PNP.

Finally, you will compare the location of your active site with the crevice map to check for consistency in the methods.

Using the crevice surface to scan for potential binding sites

↳ **To create and view the crevice surface**

1. From the CAChe Workspace, choose **File | Open** and open `1ULA.csf`.

   The 3D Structure window for human-PNP (1ULA) opens.

2. From the 1ULA 3D Structure window, choose **Analyze | Crevice**

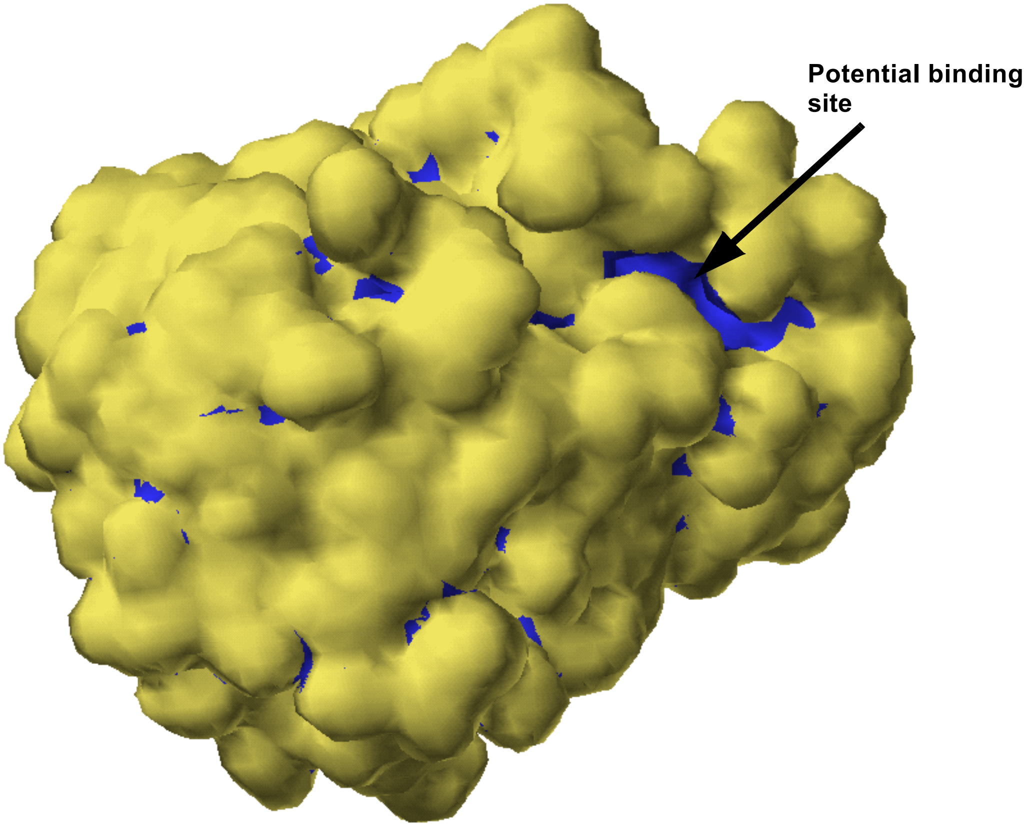
[1] Cook, W.J., Ealick, S. E., Bugg, C. E., Stoeckler, J. D., Parks, R. E., “Crystallization and Preliminary X-ray Investigation of Human *Erythrocytes* Purine Nucleoside Phosphorylase”, *J. Biol. Chem.*, **1990**, *285*, 1812.

[2] Journal of Computer-Aided Molecular Design, **2000**, *14*, 383-401.

[3] Shi, W., Basso, L. A., Santos, D. S., Tyler, P. C., Furneaux, R. H., Blanchard, J. S., Almo, S. C. and Schramm, V. L., “Structures of Purine Nucleoside Phosphorylase from *Mycobacterium tuberculosis* in Complexes with Immucillin-H and Its Pieces,” *Biochemistry*, **2001**, *40*, 8204-8215.

Surface.

After a few seconds, the crevice surface appears.



The crevice surface is an accessible surface colored by the distance from a smoother enclosing accessible surface. Blue is used to identify regions far from an enclosing smoother outer surface.

Notice the largest crevice or “lake” on the surface is in the upper right hand corner. This is a region of the surface that might be a good ligand binding region because of its large complex shape and its depth.

The large irregular shape makes it likely that a site of this shape is unique to the 1ULA protein and does not occur in any other protein. Consequently, a ligand that bound to this entire complex region would be selective.

The depth of the crevice suggests that there may be many binding interactions on the bottom and sides of the crevice leading to strong binding.

At the end of this exercise, you will compare the active site located by homology to the region suggested by the crevice surface.

To close the crevice surface

1. When you have finished analyzing the crevice surface, choose **Analyze | Show Surfaces** and uncheck **crevice1.acs**.

The crevice surface disappears revealing the protein chemical sample.

Aligning sequences

First we will prepare human-PNP and TB-PNP for alignment by analyzing them in the Sequence View.

To prepare human-PNP and TB-PNP for alignment

1. Choose **File | Open** and select `TB-PNP+PNP030.csf`.

   `TB-PNP+PNP030.csf` opens and displays the cleaned structure.
2. Choose **Analyze | Sequence**.

   The Sequence View window opens and the sequence of TB-PNP is displayed.
3. Choose **Window | ...\1ULA.csf**.

   `The 1ULA.csf` 3D Structure Window comes to the front.
4. Choose **Analyze | Sequence**.

   The Sequence View window comes to the front and the sequence of human-PNP (1ULA) is displayed below that of TB-PNP.

Next align the sequences either automatically or manually do one of the following:

To automatically align human-PNP and TB-PNP

1. In the Sequence View, choose **Edit | Align**

   The Align Sequence dialog appears.
2. Select human-PNP as the Target Sequence and TB-PNP as the Probe Sequence and press **OK**.

   Gaps appear in human-PNP and in TB-PNP that align the sequences according to the maximum scoring alignment using the BLOSUM50[1] substitution matrix in the Needleham-Wunsch alignment algorithm[2].

To select the active site residues in TB-PNP

1. Choose **Window |...\TB-PNP+PNP030.csf**.

   The 3D Structure Window for TB-PNP+PNP030 comes to the front.
2. Choose **Edit | Group Atoms**.

[1] REFERENCE MISSING

[2] Needleman, S. B. and Wunsch, C. D., “A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins”, *J. Mol. Biol.*, **1970**, *48*, 443-453; Gotoh, O., “An Improved Algorithm for Matching Biological Sequences”, *J. Mol. Biol.*, **1982**, *162*, 705-708.

The Group Atoms dialog appears.

3. Choose **ActiveSite** from the **Defined Groups:** list.

   ActiveSite appears in the **Group Name:** text box.

4. Choose **G Only** and then **OK**.

   The Group Atoms dialog closes and active site residues are selected. The rest of the chemical sample dims.

5. Choose **Window | Sequence View**.

   The Sequence View window comes to the front.

6. In the Sequence View window, choose the Select Tool

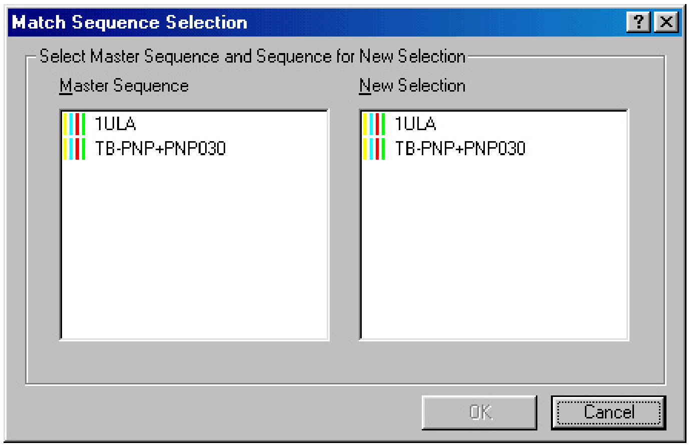
   The active site residues for TB-PNP+PNP030 are highlighted and outlined with black boxes.

Next you will select the corresponding residues in human PNP.

⇘ To select matching residues in human-PNP

1. In the Sequence View, choose **Edit | Match Selection**

   The Match Sequence Selection dialog appears.



2. Choose TB-PNP+PNP030 as the **Master Sequence** and 1ULA as the **New Selection**.

   The OK button highlights.

3. Choose **OK**.

   Residues in 1ULA are selected and a black border appears around the newly selected residues. Note that the selected residues in 1ULA and TB-PNP+PNP030 are identical with two exceptions. Examine the selected residues in the 3D Structure Window for 1ULA.

4. Choose **Window | ...\1ULA.csf**.

   The `1ULA.csf` 3D Structure Window comes to the front. Notice that the selected residues form an <u>open</u> pocket ready to receive a ligand. This is different from the pocket in TB-PNP+PNP030 which has already closed around the PNP030 ligand. Apparently as a ligand binds, the PNP active site pocket closes to wrap the extended residues around the ligand.

To create an active site group in human-PNP

1. In the 3D Structure Window for human-PNP, choose **Edit | Group Atoms**.

   The Group Atoms dialog opens.

2. Type `ActiveSite-human` into the **Group Name** text box, choose **>>Group>>** and then **OK**.

   A named group is created for the active site in `1ULA.csf`.

3. Choose **File | Save**.

   `1ULA.csf` is saved. The gaps you entered in the sequence and the named active site are saved.

NOTE

You may see warning dialogs about hybridization and charge. Press the **Save** button to dismiss each dialog and continue saving the file.

You identified the active site residues by alignment of two proteins. Would you have reached the same conclusions if multiple sequences had been used? In the next section we examine the results of a multiple sequence alignment to answer this question.

Validating the alignment using multiple sequences

Clustal-W multiple sequence alignment of purine nucleoside phosphorylase homologs

```
Y. pestis PNP                  -------------MATPHIN-AEMGDFADVVLMPGDPLRAKFIAETFLQD 36
E. coli PNP                    -------------MATPHIN-AEMGDFADVVLMPGDPLRAKYIAETFLED 36
H. influenzae PNP              --------------MTPHIN-APEGAFADVVLMPGDPLRAKYIAETFLQD 35
V. cholerae PNP                -------------MATPHIN-AQMGDFADVVLMPGDPLRAKYIAENFLDN 36
S. aureus PNP                  ------------MKSTPHIKPMNDVEIAETVLLPGDPLRAKFIAETYLDD 38
H. sapiens PNP                 -MENGYTYEDYKNTAEWLLSHTK--HRPQVAIICGSG-LGGLTDKLTQAQ 46**
B. taurus PNP                  -MQNGYTYEDYQDTAKWLLSHTE--QRPQVAVICGSG-LGGLVNKLTQAQ 46
M. musculus PNP                -MENEFTYEDYETTAKWLLQHTE--YRPQVAVICGSG-LGGLTAHLKEAQ 46
B. subtilis PNP|               -MK-----DRIERAAAFIKQNLP--ESPKIGLILGSG-LGILADEIENPV 41
M. tuberculosis PNP            MADPRPDPDELARRAAQVIADRTGIGEHDVAVVLGSGWLPAVAALGSPTT 50++
P. aerophilum PNP              --------MVKLTNPPKSPKELGFDEFPSIGIIGGSG--LYDPGIFENAV 40

Y. pestis PNP                  VREVNNVRGMLG----------FTGTYKGRKISVMG--HGMGIPS----- 69
E. coli PNP                    AREVNNVRGMLG----------FTGTYKGRKISVMG--HGMGIPS----- 69
H. influenzae PNP              VVEVTNVRNMLG----------FTGTYKGRKISIMG--HGMGIPS----- 68
V. cholerae PNP                AVQVCDVRNMFG----------YTGTYKGRKISVMG--HGMGIPS----- 69
S. aureus PNP                  VEQFNTVRNMFG----------FTGTYKGKKVSVMG--SGMGMPS----- 71
H. sapiens PNP                 IFDYGEIPNFPRSTVPGHAGRLVFGFLNGRACVMMQ--GRFHMYEGYPLW 94**
B. taurus PNP                  TFDYSEIPNFPESTVPGHAGRLVFGILNGRACVMMQ--GRFHMYEGYPFW 94
M. musculus PNP                IFDYNEIPNFPQSTVQGHAGRLVFGLLNGRCCVMMQ--GRFHMYEGYSLS 94
B. subtilis PNP|               KLKYEDIPEFPVSTVEGHAGQLVLGTLEGVSVIAMQ--GRFHFYEGYSME 89
M. tuberculosis PNP            VLPQAELPGFVPPTAAGHAGELLSVPIGAHRVLVLA--GRIHAYEGHDLR 98++
P. aerophilum PNP              EVQIHTPYGLPSDN-------VIVGRVAGRVVAFLPRHGRGHKYPPHKIP 83

Y. pestis PNP                  CSIYAKELITDFGVKKIIRVGSCGAVRTDVKLRDVVIGMGACTDSKVNRM 119
E. coli PNP                    CSIYTKELITDFGVKKIIRVGSCGAVLPHVKLRDVVIGMGACTDSKVNRI 119
H. influenzae PNP              CSIYAKELITEYGVKKIIRVGSCGTVRMDVKVRDVIIGLGACTDSKVNRI 118
V. cholerae PNP                CSIYVTELIKDYGVKKIIRVGSCGAVNEGIKVRDVVIGMGACTDSKVNRI 119
S. aureus PNP                  IGIYSYELIHTFGCKKLIRVGSCGAMQENIDLYDVIIAQGASTDSNYVQQ 121
H. sapiens PNP                 KVTFPVRVFHLLGVDTLVVTNAAGGLNPKFEVGDIMLIRDHINLPGFSGQ 144**
B. taurus PNP                  KVTFPVRVFRLLGVETLVVTNAAGGLNPNFEVGDIMLIRDHINLPGFSGE 144
M. musculus PNP                KVTFPVRVFHLLGVETLVVTNAAGGLNPNFEVGDIMLIRDHINLPGFCGQ 144
B. subtilis PNP|               KVTFPVRVMKALGVEALIVTNAAGGVNTEFRAGDLMIITDHIN---FMGT 136
M. tuberculosis PNP            YVVHPVRAARAAGAQIMVLTNAAGGLRADLQVGQPVLISDHLN---LTAR 145++
P. aerophilum PNP              YRAN-IYSLYMLGVRSIVAVSAVGSLRPDYAPGDFVVPDQFVDMTKGREY 132

Y. pestis PNP                  RFKDH-----------DYAAIADFEMTRNAVDAAKAKG--VNVRVGNLFS 156
E. coli PNP                    RFKDH-----------DFAAIADFDMVRNAVDAAKALG--IDARVGNLFS 156
H. influenzae PNP              RFKDN-----------DFAAIADFDMAQAAVQAAKAKG--KVVRVGNLFS 155
V. cholerae PNP                RFKDH-----------DFAAIADYKMVKAAEEAAKARG--IDVKVGNLFS 156
S. aureus PNP                  YQLPG-----------HFAPIASYQLLEKAVETARDKG--VRHHVGNVLS 158
H. sapiens PNP                 NPLRGPNDERFGDRFPAMSDAYDRTMRQRALSTWKQMGEQRELQEGTYVM 194**
B. taurus PNP                  NPLRGPNEERFGVRFPAMSDAYDRDMRQKAHSTWKQMGEQRELQEGTYVM 194
M. musculus PNP                NPLRGPNDERFGVRFPAMSDAYDRDMRQKAFTAWKQMGEQRKLQEGTYVM 194
B. subtilis PNP|               NPLIGPNEADFGARFPDMSSAYDKDLSSLAEKIAKDLN--IPIQKGVYTA 184
M. tuberculosis PNP            SPLVG-------GEFVDLTDAYSPRLRELARQSD------PQLAEGVYAG 182++
P. aerophilum PNP              TFYDGPR----TCHIQIGLEPFTQEIRQILIETAKKYN--RTHDGGCYVC 176

Y. pestis PNP                  ADLFYTPDPQMFDVM-EKYGILGVEMEAAGICGVAAEFGAKALTICTVSD 205
E. coli PNP                    ADLFYSPDGEMFDVM-EKYGILGVEMEAAGIYGVAAEFGAKALTICTVSD 205
H. influenzae PNP              ADLFYTPDVEMFDVM-EKYGILGVEMEAAGIYGVAAEYGAKALTICTVSD 204
V. cholerae PNP                AELFYTPDPSMFDVM-DKYGIVGVEMEAAGIYGVAAEYGAKALAICTVSD 205
S. aureus PNP                  SDIFYNADTTASERW-MRMGILGVEMESAALYMNAIYAGVEALGVFTVSD 207
H. sapiens PNP                 VAGPSFETVAECRVL-QKLGADAVGMSTVPEVIVARHCGLRVFGFSLITN 243**
B. taurus PNP                  LGGPNFETVAECRLL-RNLGADAVGMSTVPEVIVARHCGLRVFGFSLITN 243
M. musculus PNP                LAGPNFETVAESRLL-KMLGADAVGMSTVPEVIVARHCGLRVFGFSLITN 243
B. subtilis PNP|               VTGPSYETPAEVRFL-RTMGSDAVGMSTVPEVIVANHAGMRVLGISCISN 233
M. tuberculosis PNP            LPGPHYETPAEIRML-QTLGADLVGMSTVHETIAARAAGAEVLGVSLVTN 231++
P. aerophilum PNP              IEGPRFSTKAESRIWREVFGCDIIGMTLVPEINLARELGMCYGLIALVTD 226

Y. pestis PNP                  HIRTGEQTT---AAERQTTFNDMIEIALESVLLGDNA------------- 239
E. coli PNP                    HIRTHEQTT---AAERQTTFNDMIKIALESVLLGDKE------------- 239
H. influenzae PNP              HIRTHEQTT---AEERQLTFNDMIEIALDSVLIGDAL------------- 238
V. cholerae PNP                HIKTGEQTT---SEERQNTFNEMIEIALDSVLIGDQAGY----------- 241
S. aureus PNP                  HLIHETSTT---PEERERAFTDMIEIALSLV------------------ 235
H. sapiens PNP                 KVIMDYESLEKANHEEVLAAGKQAAQKLEQFVSILMASIPLPDKAS---- 289**
B. taurus PNP                  KVIMDYESQGKANHEEVLEAGKQAAQKLEQFVSLLMASIPVSGHTG---- 289
M. musculus PNP                KVVMDYENLEKANHMEVLDAGKAAAQTLERFVSILMESIPLPDRGS---- 289
B. subtilis PNP|               AAAGILDQP--LSHDEVMEVTEKVKAGFLKLVKAIVAQYE---------- 271
M. tuberculosis PNP            LAAGITGEP--LSHAEVLAAGAASATRMGALLADVIARF----------- 268++
P. aerophilum PNP              YDIWVPHQP--VTAEAVEKMMTEKLGIIKKVIAEAVPKLPAELPKCSETL 274
```

The table above shows the Clustal-W multiple sequence alignment of purine nucleoside phosphorylase (PNP) homologs (proteins related by evolution).The “**” line is the sequence of human PNP and the “++” line is the sequence of TB-PNP. A multiple sequence alignment is made to discover portions of evolutionarily related proteins (homologs) that are unchanged by evolution. These unchanged or conserved residues are likely to be important to the protein’s function. In particular the active sites of enzyme homologs tend to be highly conserved.

The multiple sequence alignment shows that the active site of TB-PNP is conserved in human, mouse, bovine and soil bacteria. The TB-PNP active site residues do not appear to be conserved E. coli, Y. pestis, H. influensae, V. cholerae or S. aureus PNP. This suggests that these enzymes have a different active site and can be classified into two sub-families by active site.

Within the TB-PNP sub-family, it appears that our analysis is correct and we have found the conserved active site residues. Only two of sixteen residues in the active site of TB-PNP are different in human PNP. A strategy in the design of selective inhibitors would focus on the interactions of these two residues with ligands.

Comparing the active site to the crevice surface

You validate the crevice surface by checking that the active site is located near the crevice that you identified as a potential binding site at the beginning of this exercise.

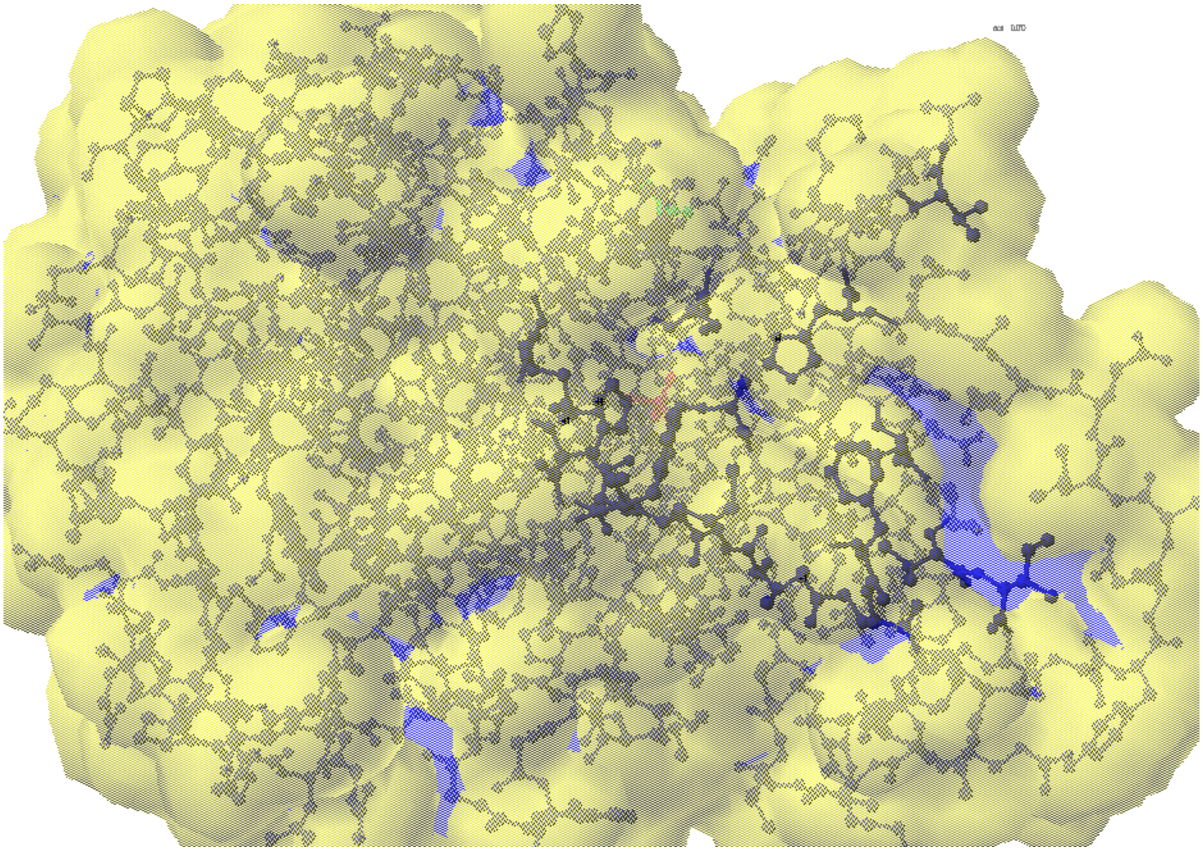
To view the crevice surface

1. From the 1ULA 3D Structure window, choose **Analyze | Show Surfaces**.

   The Show Surfaces dialog opens.

2. Choose the surface **crevice1.acs** and click **OK**.

   After a few seconds, the crevice surface is drawn. Notice that largest lake on the crevice surface is near the active residues.



Docking a bound ligand from one protein into a homologous protein

You will dock PNP030 into human-PNP by superimposing human-PNP onto TB-PNP+PNP030 and then deleting TB-PNP.

To superimpose human-PNP and TB-PNP

1. Choose **Window | Sequence View**.

   The Sequence View window comes to the front.

2. From the Sequence View window, verify that both TB-PNP+PNP030 and 1ULA sequences are displayed, that their sequences are aligned and that only the active site residues are highlighted and outlined with black boxes.

3. From the Sequence View window, choose **Edit | Superimpose Sequences**.

   The Superimpose Sequences dialog opens.

4. Choose the **Probe Sequence** to be 1ULA and the **Target Sequence** to be TB-PNP+PNP030.

🕮 NOTE

Only protein chains and ligands containing selected residues are superimposed onto the target. Thus, the SO_4^{2-} in human-PNP are not superimposed in the target window.

5. Choose **Superimpose Probe onto Target** and click **OK**.

   After approximately 1 minute, the Superimpose Sequences dialog closes.

6. Choose **Window | ...\TB-PNP+PNP030.csf**.

   The TB-PNP+PNP030 3D Structure window comes to the front with both TB-PNP and 1ULA superimposed.

7. In the 3D Structure window, choose the Select Molecule Tool and click on TB-PNP (TB-PNP is the unselected and dimmed protein).

   TB-PNP highlights. 1ULA, PNP030, PO4 and the water molecules dim.

8. Choose **Edit | Delete**.

   TB-PNP disappears and the remaining portions of the structure highlight.

9. Choose **File | Save As** and name the new file `1ULA+PNP030.csf`.

The result is PNP030 docked in human-PNP. At this point, you should analyze the ligand in the active site by counting the number of hydrogen bonds, checking for bumps, and viewing the adjacent surface as you did in the earlier exercise *“Evaluating the docking”*, p 3-7.

You’ve probably noticed that trying to evaluate ligand selectivity and contact surface is an uncertain process. As a result, there have been significant research efforts to develop automated and quantitative measurements of protein-ligand interactions from a three-dimensional structure of a protein-ligand complex. You will apply one of these methods in the next exercise.