

Getting Started with BioMedCAChE

Molecular Modeling in Drug Design

**CAChE Group
Fujitsu**

Beaverton, Oregon

The Fujitsu logo is displayed in red. It features the word "FUJITSU" in a bold, sans-serif font. Above the letter "I", there is a stylized infinity symbol (∞) that is integrated into the design of the letter.

Copyright statement

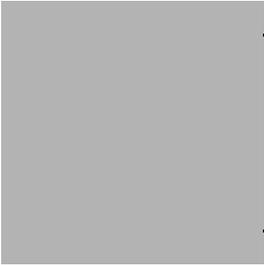
© 2003 **Fujitsu Limited**. All rights reserved.

This document may not be used, sold, transferred, copied or reproduced, in any manner or form, or in any medium, to any person other than with the prior written consent of **Fujitsu Limited**.

Published by **Fujitsu Limited**.

All possible care has been taken in the preparation of this publication but **Fujitsu Limited** does not accept any liability for any inaccuracies that may be found.

Fujitsu Limited reserves the right to make changes without notice both to this publication and to the product it describes.



Foreword

Molecular modeling is a powerful thinking tool that helps us decide which experiments will be the most productive. Each time we experiment with new drug designs and new targets, we discover something new or gain unexpected insight into their properties and interactions.

However, the greatest benefit often comes when we find experimental results that differ from our predictions. Every model we create reflects our understanding about the structure and chemistry of the compounds and is a simplification of the full experimental conditions. Results different from experiment often point out our own misconceptions about and over simplifications of the underlying chemistry. Understanding what we have omitted in our model often creates the breakthroughs needed to solve our research problem.

A decade ago, modeling drugs and proteins on the desktop was a tedious, slow process. The twenty to a hundred-fold increases in computational speed and dramatic improvements in software make it a pleasure to screen properties of virtual libraries and even explore protein-ligand interactions in the quiet of our offices where we can plan the next set of experiments.

The slow rate of adoption for modeling by medicinal and experimental chemists in industry is therefore surprising considering that the minority who use modeling are so productive. Part of the rate of slow adoption can be traced to the barriers that have existed in the past. Because desktop computers were once slow, early modeling systems ran on expensive Unix worksystems that required an expert to operate. Innovative companies, seeking to leverage the power of modeling for their experimental chemists, compromised by setting up modeling facilities for their experimental chemists. Unfortunately, located down the hall or in the next building, the modeling facilities were far from the laboratory and office where the papers and reference materials used on our research project reside. The facilities were inconvenient and the software was difficult to use. As a result, modeling was under used.

Our experience with BioMedCACHe on our desktop machines in our own offices is completely different. Modeling is always accessibly within arms reach without a waiting line. Within minutes we can be exploring a new idea without interrupting our thought process. With our biochemical models on the same computer as our word processing and presentation software, we quickly produce the reports and presentations needed to communicate with our team members and advance the project.

The experiments in this manual focus on tasks that we find key to lead discovery and optimization. We encourage you to work through all exercises

completely. You will be amazed at how quickly you will be integrating what you learn into your current research projects. Even better, you will see a dramatic increase in your productivity and success as you become one of the growing number of experimental chemists who integrate theory and modeling into their decisions about which compounds should be made next.

Table of Contents

Foreword	iii
How to use Getting Started	7
Importing and Cleaning Protein Crystal Structures	9
Background	10
Importing proteins	11
Saving a PDB molecule	13
Simplifying the structure	15
Cleaning protein structure	19
Viewing and Analyzing Proteins, Ligands and their Complexes	23
Overview	23
Viewing the accessible surface	24
Viewing the protein sequence	25
Analyzing the protein sequence	27
Simplifying the protein structure: ribbons	29
Locating the active site	31
Locating the hydrogen bonds	33
Displaying the active site pocket	33
Docking Ligands into Proteins	37
Overview	37
Background	38
Docking by superposition	38
Adjusting the docked position in the active site pocket	41
Evaluating the docking	43
Refining the docking	44
So what?	44
Discovering Active Sites of Homologous Proteins by Sequence Alignment	45
Overview	45
Background	46
Using the crevice surface to scan for potential binding sites	46
Aligning sequences	48
Validating the alignment using multiple sequences	51
Comparing the active site to the crevice surface	52
Docking a bound ligand from one protein into a homologous protein	53
Refining the docking	54
So what?	54

1

How to use Getting Started

This booklet contains step-by-step tutorials that are designed to introduce you to the BioMedCACHe interface and to demonstrate how BioMedCACHe may be used in drug discovery.

Installing BioMedCACHe

Refer to the Installation Notes for a step-by-step description of how to install BioMedCACHe. In this booklet it is assumed that BioMedCACHe is already installed.

Using this booklet

You should work through the examples in this booklet using the programs CACHe Workspace and ProjectLeader.

The booklet is divided up into five experiments that help you get started in drug discovery:

- importing and cleaning a protein from the Protein Data Bank
- analyzing proteins and protein ligand interactions
- importing, modifying and screening potential ligands
- docking ligands into proteins
- structure activity analysis

TIP

Your CACHeUser directory is often C:\CACHeUser.

You can find the example files used in this booklet in the `Getting Started Exercises` directory inside your CACHeUser directory.

After you finish this booklet

Once you are familiar with the interface and concepts of CACHe, you can start working on your own projects with CACHe. Refer to the “Using CACHe” part of the BioMedCACHe User Guide for further information on the database, the graphical interface, and the syntax of commands. There is also a comprehensive on-line Help system accessible from all components of the CACHe Workspace.

Starting BioMedCACHe

- 1.1 To start CACHe, choose **Start | Programs | CACHe | Workspace** from the Windows taskbar.
- Alternatively, move the cursor over the **My Computer** icon and double-click with the left mouse button to display a window that contains more icons. Double-click on the **C** icon, then on the **Program Files** icon, then on the **Fujitsu** icon, then on the **CACHe** icon, then on the **CACHe.exe** icon.

An initializing CACHe splash screen containing CACHe version details is displayed.

After a few seconds, the CACHe application window opens, containing an empty document window, referred to as the 3D Structure window or alternatively as the Workspace.

2

Importing and Cleaning Protein Crystal Structures

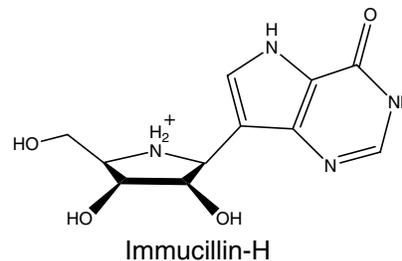
Overview

This tutorial describes how to open Protein Data Bank (PDB) files, view protein structures and sequences and build proteins and peptides from their sequence. You'll learn how to:

- open and view PDB molecules
- inspect and correct hetero groups
- add hydrogen atoms
- balance charges

Background

In this tutorial, we will investigate the 1.75Å resolution crystal structure (1G20) of *Mycobacterium tuberculosis* purine nucleoside phosphorylase (TB-PNP) complexed with the transition-state analogue Immucillin-H (ImmH) and phosphate.¹



TB kills millions each year and it has been estimated that one-third of the world's population is infected with latent TB.

Analysis of the genome sequence of *Mycobacterium tuberculosis* (TB) predicted that it expresses purine nucleoside phosphorylase (PNP) which catalyzes the phosphorolysis of purine nucleotides to purine bases and deoxynucleosides to (deoxy)ribosyl 1-phosphate. PNP recycles purines, a crucial function for organisms that do not synthesize purines. It is thought that inhibition of this enzyme will cause physiological changes in TB that will cause the bacterium to enter a latent state thereby preventing the development of active TB in infected individuals².

One problem with this approach is that humans also express PNP. In humans it is known that inhibitors of human PNP have potential clinical use as immunosuppressants. Since there is a risk that a TB-PNP inhibitor would also inhibit human PNP, leading to unacceptable side-effects during treatment, there is a need to develop TB-PNP inhibitors that do not bind to human PNP. Although TB-PNP is evolutionarily related to human PNP, a recent crystal structure shows a significant difference in the hydrogen bonding to ImmH between the mammalian and TB PNPs. It is therefore possible that molecular modeling can be useful in designing selective inhibitors for TB PNP. The need to develop selective inhibitors is a common problem in discovery and development.

The goal of these exercises is to teach you how to use BioMedCACHe to design your own selective inhibitors by using peer reviewed studies from the current scientific literature applied to the important problem of obtaining new drugs for the treatment of TB.

-
1. 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.
 2. Ojha, A. K., Muckherjee, T. K., and Chatterji, D., "High Intracellular Level of Guanosine Tetraphosphate in *Mycobacterium smegmatis* Changes the Morphology of the Bacterium", *Infect. Immun.*, **2000**, *68*, 4084-4091.

Importing proteins

The Protein Data Bank (PDB) maintained by the Research Collaboratory for Structural Bioinformatics (RCSB)¹ contains tens of thousands x-ray crystal structures of proteins and other biomolecules.

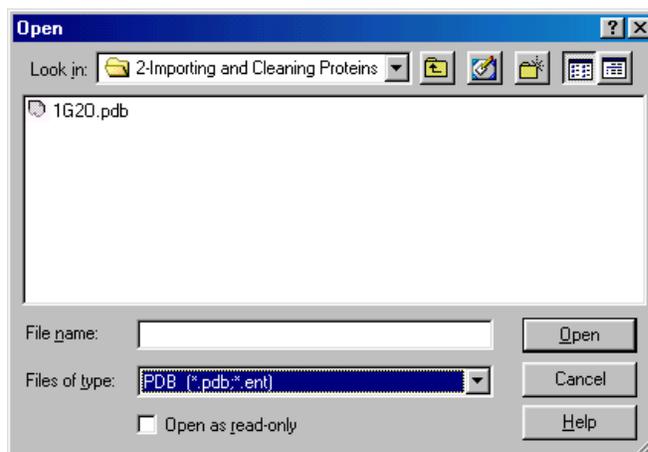
NOTE

The PDB file is 1g2o. The last letter is 'oh', not zero.

The structures of proteins and other biomolecules can be downloaded from <http://www.rcsb.org/pdb/> as PDB files. PDB files have an extension of .pdb or .ent. To view the 3D structure of TB-PNP complexed with ImmH, you first download 1G2O.pdb from the PDB and then open it in the CACHE workspace. In this exercise, find the previously downloaded 1G2O.pdb file in the directory 2-Importing and Cleaning Proteins.

To open the PDB file

1. Choose **File** | **Open** to display the Open dialog box.



2. Select the arrow button in the **Look in** drop down box to display a drop-down list of:
 - folders in the directory structure above the currently open folder
 - available drives from which you can open the file.
3. Select the arrow button in the **Files of type** box and choose **PDB (*.pdb, *.ent)** from the drop-down list.
4. Choose the folder 2-Importing and Cleaning Proteins or drive where 1G20.pdb is located from the drop-down list.

Do one of the following:



- To move up a level of folders to locate the folder containing the file, select the dialog box button shown to the left.
- To locate the file in a subfolder, click and drag the scroll bar in the

1. Berman, H. M., Westbrook, J., Feng, Z., Gilliland G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E., "The Protein Data Bank", *Nucl. Acids. Res.*, **2000**, 28, 235-242.

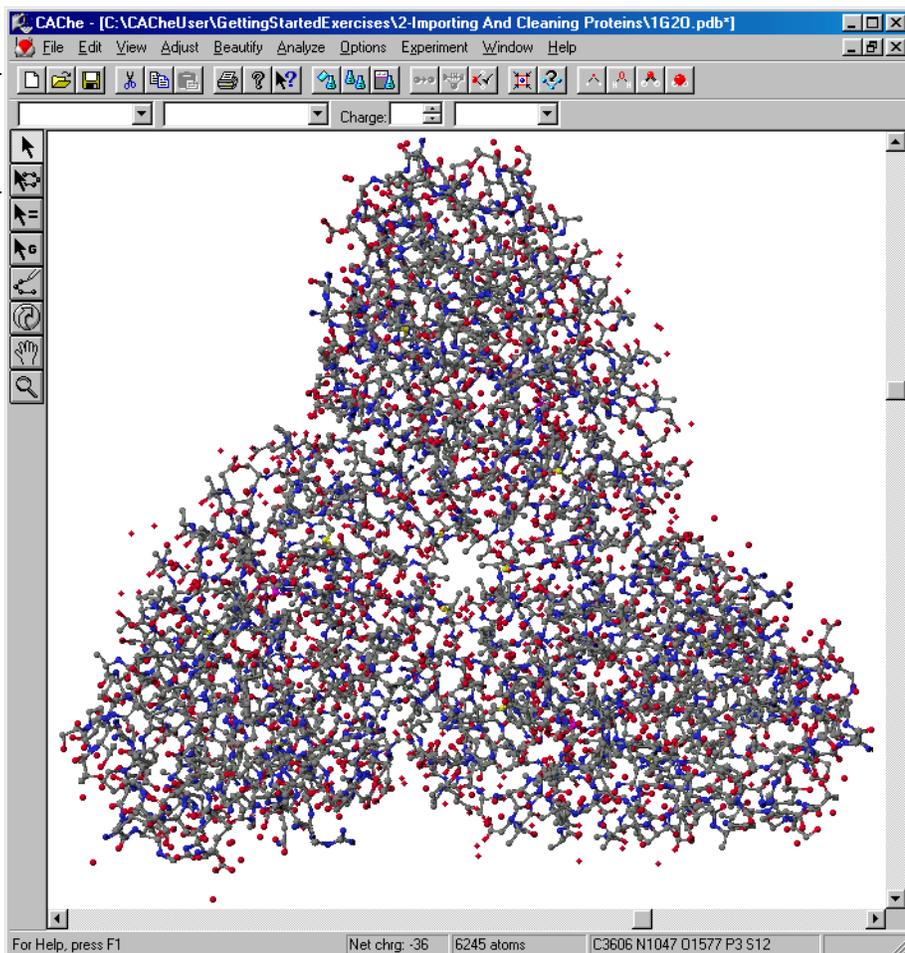
scrolling list to display the folder where the file is located, and double-click on the folder to open it.

5. Do one of the following:
 - Click and drag the scroll bar in the scrolling list to locate 1G20.pdb and select the file by clicking on it.
 - Click in the **File name** text box and type 1G20.pdb.
6. Select **Open** to open the file and to close the Open dialog box.

A new workspace opens and the obviously trimeric structure displays.

 **NOTE**

It may take 30 seconds to load the PDB file, depending on the speed of your computer.

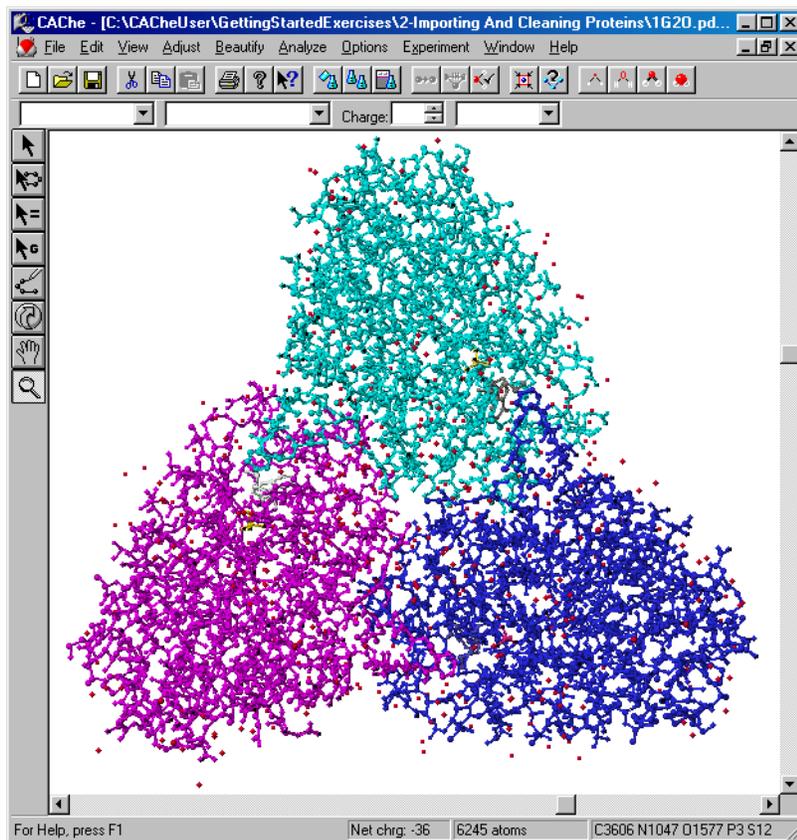
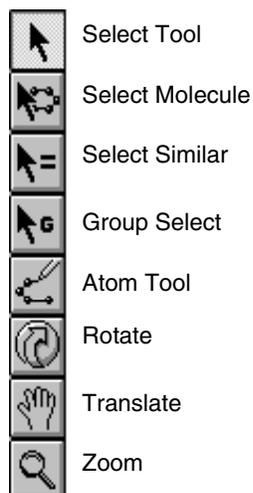


To see clearly the three chains and their ligands in the trimeric structure:

1. choose **View | Color by Molecule**.

Each chain and ligand gets a different color. The oxygen atoms from

water molecules all remain red. Choose the select by molecule tool and



click on each colored component to examine it in more detail.

TIP

Opening csf files is much faster than opening PDB files. 1G20.csf will open in a few seconds.

At this point you should save the file as a chemical sample with the name 1G20.csf. From now on you will work with the CAChe chemical sample file (csf) rather than the PDB file so that you can retain all of the views and information you create in modeling.

Saving a PDB molecule

To save a PDB molecule

1. Select **File | Save**.

When you save this workspace, you can choose to save it as a chemical sample file (*.csf), a PDB file, or other file type. Save the workspace as a chemical sample file (*.csf) to preserve all of the information you have added.

Otherwise, information such as rendering style and computed atom properties will be lost. In CAChe, the term ‘chemical sample’ is used to refer to the proteins, ligands, water molecules and other chemical entities that are contained in the model you develop.

NOTE

You will be informed that the charge on the protein is -36. Click **Save**. You will adjust the charge later.

All of the information in the PDB file has been stored in the CAChe chemical sample format. For example, residues have been stored as Groups and you can use the Select Group tool to select residues.

To view this information, choose **Analyze | Chemical Properties Spreadsheet**. The chemical sample properties window appears.

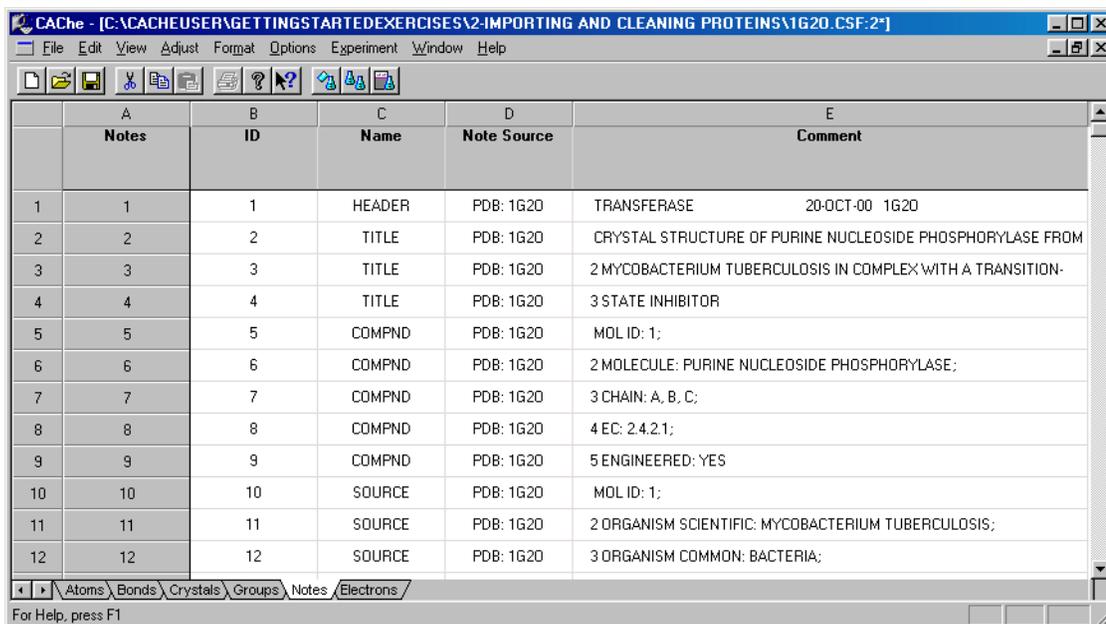
	A	B	C	D	E	F	G	H	I
	Atom List	ID	Formal Charge (charge_au)	X Coordinate (angstrom)	Y Coordinate (angstrom)	Z Coordinate (angstrom)	Partial Charge (charge_au)	Name	Temperature Factor
1	N1	1	1	65.481	35.981	63.049	0.000	N	32.870
2	C2	2	0	66.390	36.448	61.964	0.000	CA	32.600
3	C3	3	0	66.999	35.279	61.191	0.000	C	31.940
4	O4	4	0	67.744	34.469	61.746	0.000	O	31.650
5	C5	5	0	67.511	37.308	62.549	0.000	CB	33.340
6	C6	6	0	68.521	37.729	61.502	0.000	CG	34.010
7	O7	7	0	68.112	38.354	60.500	0.000	OD1	35.120
8	O8	8	-1	69.721	37.436	61.678	0.000	OD2	34.470
9	N9	9	0	66.690	35.184	59.888	0.000	N	31.280
10	C10	10	0	67.188	34.123	59.006	0.000	CA	30.670

Note that the sample properties for a PDB molecule contain several worksheets:

- **Atoms** contains atoms and their properties
- **Bonds** contains bonds and their properties
- **Crystals** contains the space group and other crystallographic information
- **Groups** contains residues and their properties
- **Notes** contains REMARK, NOTE, SOURCE, TITLE, COMPND, HEADER and other PDB records.
- **Electrons** worksheet keeps track of nonbonded electron pairs so that the valency can be checked. Electrons are added when you beautify a structure and do not appear when the PDB file is first opened.

Importing proteins

Select the **Notes** tab and expand the Comment column, click column **E** to select the **Comment** column and choose **Format | Left** so that you see this:



	A	B	C	D	E
	Notes	ID	Name	Note Source	Comment
1	1	1	HEADER	PDB: 1G20	TRANSFERASE 20-OCT-00 1G20
2	2	2	TITLE	PDB: 1G20	CRYSTAL STRUCTURE OF PURINE NUCLEOSIDE PHOSPHORYLASE FROM
3	3	3	TITLE	PDB: 1G20	2 MYCOBACTERIUM TUBERCULOSIS IN COMPLEX WITH A TRANSITION-
4	4	4	TITLE	PDB: 1G20	3 STATE INHIBITOR
5	5	5	COMPND	PDB: 1G20	MOL ID: 1;
6	6	6	COMPND	PDB: 1G20	2 MOLECULE: PURINE NUCLEOSIDE PHOSPHORYLASE;
7	7	7	COMPND	PDB: 1G20	3 CHAIN: A, B, C;
8	8	8	COMPND	PDB: 1G20	4 EC: 2.4.2.1;
9	9	9	COMPND	PDB: 1G20	5 ENGINEERED: YES
10	10	10	SOURCE	PDB: 1G20	MOL ID: 1;
11	11	11	SOURCE	PDB: 1G20	2 ORGANISM SCIENTIFIC: MYCOBACTERIUM TUBERCULOSIS;
12	12	12	SOURCE	PDB: 1G20	3 ORGANISM COMMON: BACTERIA;

TIP

Click the **Comment** column header to highlight the column, then select **Format | Left** to left justify the comments.

Read through the comments to confirm that this is the structure for ImmH complexed with TB-PNP at 1.75Å resolution. Scroll to row 249 and observe that the first six residues (Met-Ala-Asp-Pro-Arg-Pro) in each of the three chains TB-PNP were not resolved and are missing from the structure. Find the HET records and the HETNAM records (scroll to row 415). These records identify groups that are not standard amino acids such as ligands, metals, solvent, or ions. There is one IMH and one PO₄ for each of the three chains in TB-PNP. The names and molecular formula of each HET group is given in the HETNAM and FORMUL records. You will use this information later when cleaning up the HET groups.

Simplifying the structure

Each of the chains in TB-PNP is the same and each has ImmH and PO₄ complexed in its active site. In this tutorial you will analyze the binding of ImmH to the active site in TB-PNP and it is helpful to work with the simplest model first, a single chain and its complexed groups. We will reduce the structure to the monomer by removing chains B and C.

Choose the **Groups** tab in the workbook. The Groups worksheet appears

	A	B	C	D	E	F	G	H	I
	Molgroups	ID	Name	Residue Sequence Number	PDB Residue Name	Molgroup Type	Residue Code	Chain ID	PDB Secondary Structure
1	1	1	ASP1	7	ASP	AA	D	A	helix RA
2	2	2	PRO2	8	PRO	AA	P	A	helix RA
3	3	3	ASP3	9	ASP	AA	D	A	helix RA
4	4	4	GLU4	10	GLU	AA	E	A	helix RA
5	5	5	LEU5	11	LEU	AA	L	A	helix RA
6	6	6	ALA6	12	ALA	AA	A	A	helix RA
7	7	7	ARG7	13	ARG	AA	R	A	helix RA
8	8	8	ARG8	14	ARG	AA	R	A	helix RA
9	9	9	ALA9	15	ALA	AA	A	A	helix RA
10	10	10	ALA10	16	ALA	AA	A	A	helix RA
11	11	11	GLN11	17	GLN	AA	Q	A	helix RA
12	12	12	VAL12	18	VAL	AA	V	A	helix RA

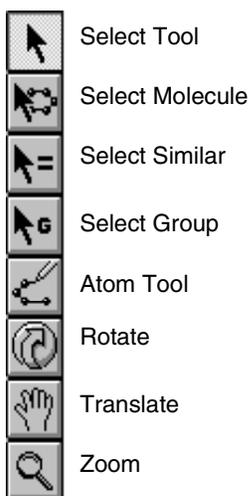
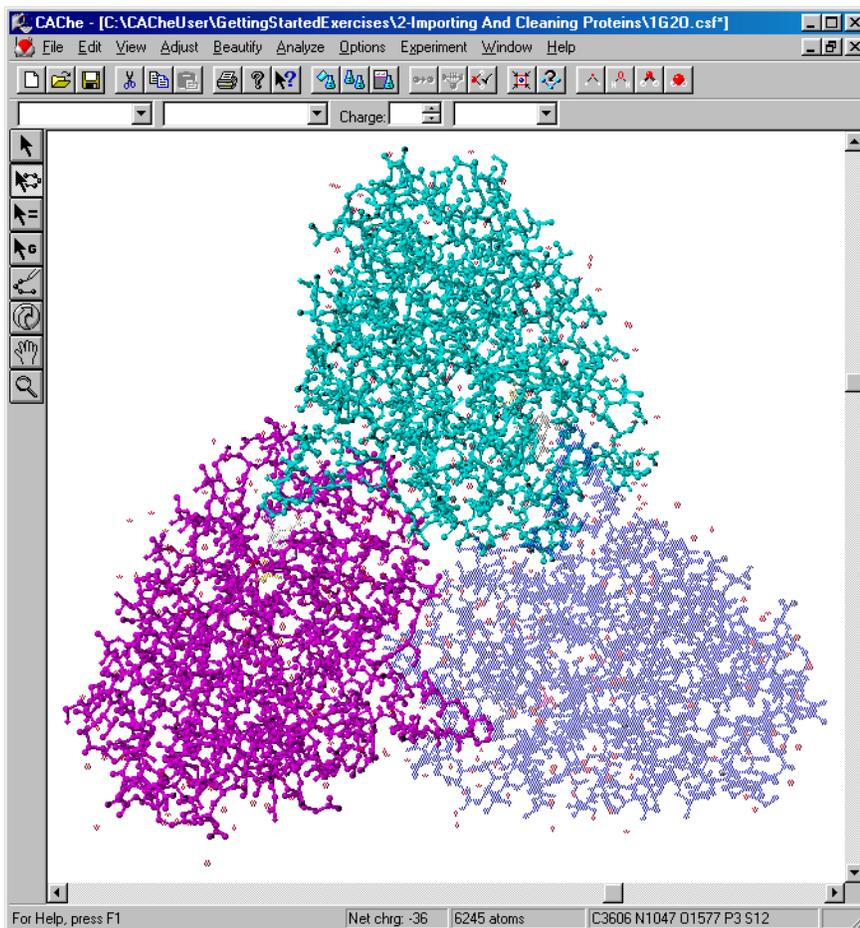
Do the following:

NOTE

This is just one technique for deleting Chain B and C. Alternatively, you could delete these chains from the Sequence View after selecting with the chain tool or from the 3D Structure Window after selecting with the Select Molecule Tool.

1. Scroll the window until you see the start of rows containing **Chain ID B**.
2. Select the first row with **Chain ID B** by clicking the row number on the left. (Row 263)
Row 263 highlights.
3. Scroll until you see the last row containing **Chain ID C**. (Row 786)
4. Shift-click the row number on the left.
All residues in chains B and C are highlighted indicating that they are selected.
5. Close the Sample Properties Window by clicking the close box (x) in the window's upper right-hand corner. This close box is below the close box (x) for the CAChe application. Be careful to choose the correct close box.

The 3D Structure Window comes to the front and you see:



6. Select **Edit | Delete**.

Chains B and C disappear exposing the HET groups and water molecules associated with chains B and C.

7. Using the Select Molecule Tool click on an atom in Chain A.

Chain A is highlighted and the rest is greyed.

8. Select **Edit | Select Neighbors** and choose **Residues, waters, HETs** from the **Select Nearest** pulldown. The **Selection Radius** should be 5 Å.

9. Choose **OK**.

Chain A and all residues, waters and HET groups within 5 Å of any atom in chain A are highlighted.

10. Select **Edit | Invert Selection**.

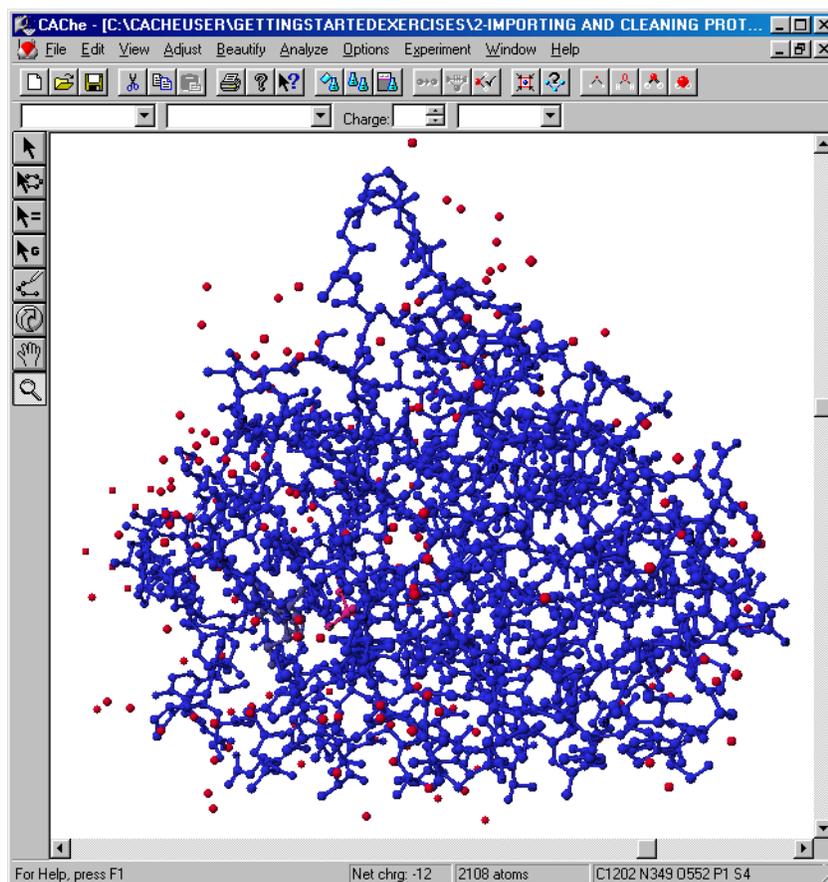
Highlighting reverses.

11. Select **Edit | Delete**.

The excess water and HET groups disappear and the remaining structure is highlighted.

12. With the Select Molecule Tool, click on the ImmH group at the top of Chain A and choose **Edit | Delete**.

The display at the bottom center of the window shows that 2108 atoms remain, and your structure looks like this:



 NOTE

You will be informed that the charge on the protein is -12. Click **Save**. You will adjust the charge later.

13. Choose **File | Save As** and save this file as `TB-PNP-monomer.csf`.
Next you will prepare the structure for molecular modeling and analysis.

Cleaning protein structure

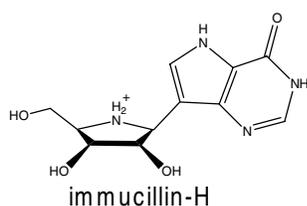
When a PDB file is opened in CAChe, some new information is automatically calculated and added. However, the information required for molecular modeling must still be added such as hydrogen atoms, atom hybridization and correct bond types for HET groups and non standard residues. If residues are missing or incomplete, it may be necessary to correct their structures. Initially, you lock atoms to prevent moving them accidentally from their crystallographic positions.

↪ Lock atoms at their crystallographic positions

1. Choose **Edit | Select All**.
All atoms and bonds are highlighted.
2. Choose **Adjust | Lock**.
Selected atoms are locked at their current position in space.

↪ Check and correct the bonding in HET groups

1. Choose **View | Color by Molecule**.
The protein chain, immucillin, and phosphate groups are displayed in different colors.
2. Choose the Select Molecule Tool and click on the Immucillin-H (ImmH) ligand.
ImmH is highlighted, the rest of the structure is dimmed.
3. Choose **View | Hide Unselected**.
All unselected atoms and bonds disappear.
4. Choose **View | Fit in Window**.
ImmH zooms to fill the window. You may need to rotate ImmH to see all the atoms and bonds.
5. Choose **View | Color by Element**.
The ligand changes so that carbon atoms are grey, oxygen atoms are red, nitrogen atoms are blue and hydrogen atoms are white.

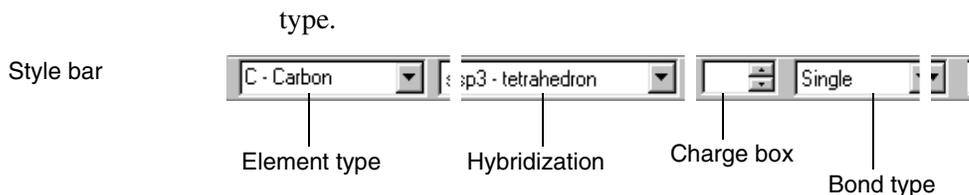


📖 NOTE

CAChe uses the file `het_dictionaryCIF.txt` from the PDB to assign the bonding and charges. You should update this file by downloading it from the PDB when a new version of the PDB is released.

Examine ImmH and change the bonding and charge to agree with immucillin-H structure shown in the figure.

6. To change a bond order, use the Select Tool and click a bond to select it.
7. From the style bar, pull down the Bond type menu and select the new bond



All selected bonds change to the new bond type.

Next examine ImmH and set the charges to agree with the figure

TIP

If you type the charge in the charge box instead of using the scrolling arrows, press the Enter key to apply the new charge.

8. Click the nitrogen atom in the iminoribitol ring to select it
9. Change the charge in the Charge Box to 1 using the up arrow next to the charge box.

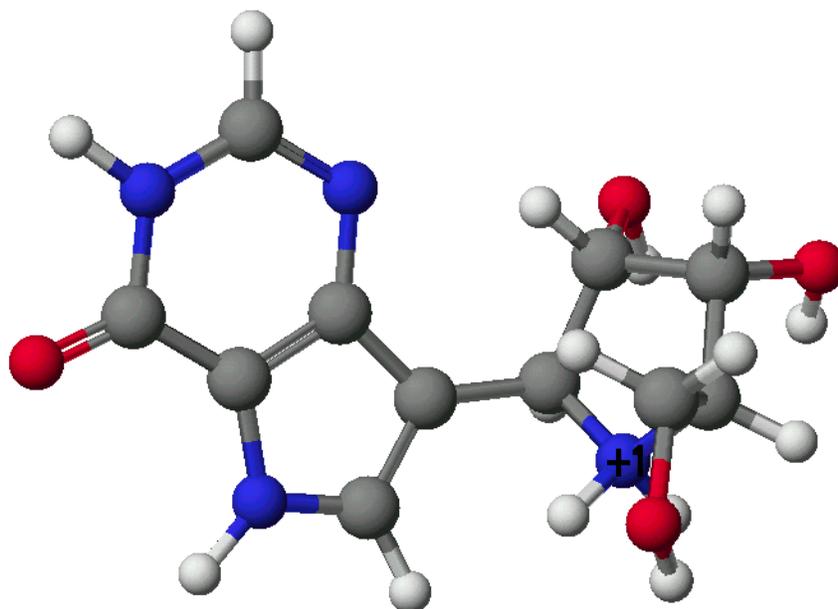
A **+1** appears on the selected nitrogen atom and the total charge on the whole chemical sample decreases to -11.

Add hydrogen atoms and define atom hybridization for ImmH

1. Choose the Select Molecule Tool and click on the ImmH ligand.
ImmH is highlighted.
2. Choose **Beautify | Valence**.

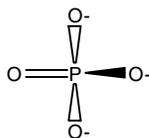
Hydrogen atoms, electrons and the atom hybridization are added.

ImmH should look like this now



Add hydrogen atoms and define atom hybridization for PO₄

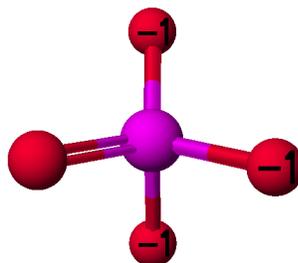
1. Choose **View | Show All**.



The full structure is displayed.

Repeat the steps if needed to clean the PO_4^{3-} ligand.

For this exercise, don't be concerned about which oxygen atoms have the negative charge. The cleaned PO_4^{3-} should look like this:



Balance the charge in the protein

Depending upon its environment in a protein crystal, the histidine residue can be neutral or protonated. When CAChe reads a PDB file, histidine is assumed to be neutral unless a charge has been specified in the PDB file. In TB-PNP, histidine should be protonated. In the following you will protonate all histidine residues.

1. Choose **View | Show All**.

All the atoms and bonds in the chemical sample appear.

2. Choose **Edit | Select All**.

All the atoms and bonds are highlighted.

3. Choose **View | Fit in Window**.

The view zooms out to show all atoms.

4. Choose **Edit | Find**.

The Find dialog opens.

5. Choose **Atoms** from the **Search for** pull down.

6. Choose **Name** from the **Where** pull down.

7. Type ND1 in the text box next to the **Is** item.

8. Choose **Find**.

The 10 nitrogen atoms in the ten histidine residues highlight and the rest of the structure unhighlights.

9. Choose Close.

The Find dialog closes.

10. Choose the Select Tool.

The ribbon at the top of the 3D Structure window shows **N - Nitrogen**.

11. Change the **charge** in the ribbon to 1.

NOTE

This quick trick for setting the charge on histidine depends on the protein containing only histidine residues with an atom named ND1.

The charge on each of the highlighted ND1 atoms changes to +1 and the total charge on TB-PNP drops to -1.

 **Add hydrogen atoms and atom hybridization to the protein**

1. Choose **Edit | Select All**

Everything in the 3D Structure window highlights.

2. Choose **Beautify | Valence**.

In just half a minute, hydrogen atoms, electrons and hybridization are added to the protein and the water molecules. The total number of atoms changes to 4427.

3. Choose **Beautify | H-Bonds**.

In four seconds, hydrogen atoms in -OH and water molecules are rotated to maximize hydrogen bonding.

Only the geometry of hydrogen atoms are changed with these **Beautify** commands.

 **Save your work as TB-PNP-cleaned.csf.**

1. Choose **File | Save As** and save the chemical sample as TB-PNP-cleaned.csf.
2. Choose **File | Close** to end your work with this chemical sample.

At this point you would normally refine the position of the hydrogen atoms using molecular mechanics, but we will not do so at this time. You will learn how to optimize structures in a subsequent exercise.

 **NOTE**

Depending on the speed of your computer, it may take a minute to **Beautify** the full protein and hydrogen bonds.

3

Viewing and Analyzing Proteins, Ligands and their Complexes

Overview

Analyzing the properties of proteins containing thousands of atoms is best accomplished by representing molecular properties with graphics that reduce complexity. Because drug design focuses only on that relatively small portion of the protein structure where the small drug molecule binds, it is important to easily visualize only the active binding sites and their ligands. In addition, finding and visualizing small molecule-protein interactions, such as hydrogen bonding, must be simple, fast and easy.

For example it is important that you are able to:

- Define and manipulate groups of atoms
- Display hydrogen bonds and atoms that bump together
- Manipulate sequences
- Identify active sites using crevice maps and sequence alignment
- Color by properties, like hydrophobicity and hydrophilicity
- Label atoms that bump together
- Extract information from sequence alignments

In this exercise, you will be introduced to BioMedCACHe's unique Sequence View, a powerful analysis tool for navigating and visualizing structures containing tens of thousands of atoms.

In this exercise, you will use CACHe to

- view the protein's accessible surface
- view and analyze the sequence of a protein
- locate the active site
- show and measure hydrogen bonds between the ligand and the protein
- display surface of the binding pocket and look at docking interactions

Viewing the accessible surface

An accessible surface is the surface of a protein that could be touched by a 1.4 Å sphere: a sphere that is approximately the size of a single water molecule.

To create and view the accessible surface

1. Choose **File | Open**, move into the folder 3-Viewing and Analyzing Proteins and Ligands .
2. Set **Files of type:** to **Chemical Sample (*.csf)** and select TB-PNP-All-Refined.csf .
3. Choose **Open**.

TB-PNP-All-Refined.csf opens and displays the cleaned structure after refinement of the hydrogen positions and then subsequent refinement of all atom positions.

4. Using the Select Molecule Tool, click an atom in the protein. The protein highlights.
5. Choose **View | Hide Unselected**.

All except the selected protein disappear. Only visible atoms - both selected and unselected - contribute to the surface. The selected atoms define the limits of the box in which the surface is drawn. Areas of the surface outside a box around the selected atoms are not drawn.

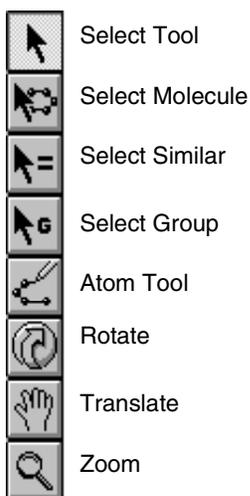
6. Choose **Analyze | Accessible Surface**.

After a few seconds, the accessible surface file is created, read into memory and displayed. The small black text - **asf 0.01**- is the surface label. You select a surface by clicking on the center of the label. You may need to use the Zoom tool to enlarge the label to select it.

Mauve (purple-red) areas and blue areas are hydrophilic while the cream area is hydrophobic. Mauve indicates hydrogen acceptors (*e.g.* C=O) and blue indicates areas of hydrogen donors (*e.g.* -NH or more generally -XH).

As expected since proteins tend to fold so that hydrophobic residues are on the interior and hydrophilic groups are on the exterior, hydrophilic regions dominate the outside of the protein. Large patches of hydrophobic areas (cream colored) on the surface of a protein, suggest that a protein is involved in interactions with other proteins. You should notice that the cream colored portions of the TB-PNP are located where chains B and C of the trimer touched the left and top of this chain.

This is a crystal of TB-PNP with a bound inhibitor - ImmH - in the active site. You should note that there is no tunnel leading to the active site. TB-PNP has folded around the ligand, closing the tunnel and making it difficult to identify an active site from the solvent accessible surface.



The labels (+1 and -1) showing through the surface are the atom charges. Before continuing, hide the accessible surface and redisplay the hidden atoms.

↪ **Hide the accessible surface**

1. Choose **Analyze | Show Surfaces**.
The Show Surfaces dialog opens.
2. Uncheck `accessible1.asf` and choose **OK**.
The surface disappears.

↪ **Show the hidden atoms and bonds**

1. Choose **View | Show All**.
All hidden atoms and bonds are displayed in preparation for the next step in the exercise.

Viewing the protein sequence

To view the sequence data for a protein 3D structure that is open in CAChe you analyze it in the Sequence View window. When you open additional proteins and analyze their sequence, their sequence data are added to the Sequence View window. Only one Sequence View window can be open at one time.

You use the sequence window to:

- view the sequence
- color the sequence by property
- mutate residues
- edit the sequence
- build peptides and proteins from residues
- analyze secondary structure
- adjust the conformation or secondary structure of a protein
- align sequences
- superimpose sequences.

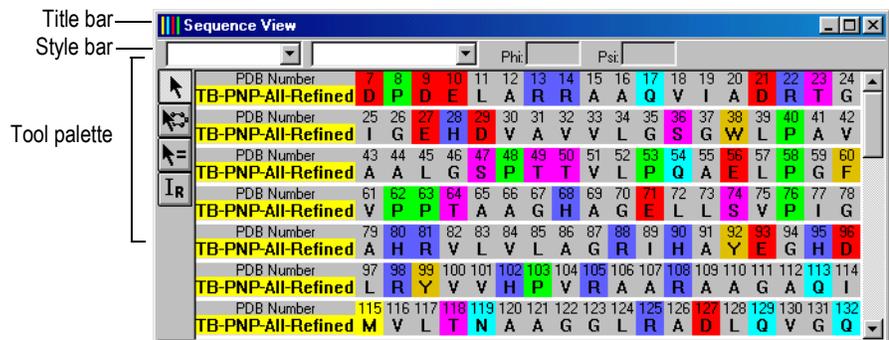
Residue	letter code	
	1	3
Alanine	A	ALA
Arginine	R	ARG
Asparagine	N	ASN
Aspartic acid	D	ASP
Cysteine	C	CYS
Glutamine	Q	GLN
Glutamic acid	E	GLU
Glycine	G	GLY
Histidine	H	HIS
Isoleucine	I	ILE
Leucine	L	LEU
Lysine	K	LYS
Methionine	M	MET
Phenylalanine	F	PHE
Proline	P	PRO
Serine	S	SER
Threonine	T	THR
Tryptophan	W	TRP
Tyrosine	Y	TYR
Valine	V	VAL

↪ **To view the sequence of TB-PNP**

1. From a 3D Structure window, choose **Analyze | Sequence** to display the Sequence View window.
The Sequence View window displays sequences using 1-letter codes in rows that begin with the chemical sample name.

2. Choose **View | 3-Letter-Code**.

3-letter codes are displayed for residues colored by chemical type.



NOTE

The sequence number starts with 7 since the first 6 residues were not observed in the crystallography.

The sequence has a direction, starting from the N terminus of the first residue to the C terminus of the last one. HET groups are always displayed, but water molecules (HOH or DOD) are never displayed in the Sequence View.

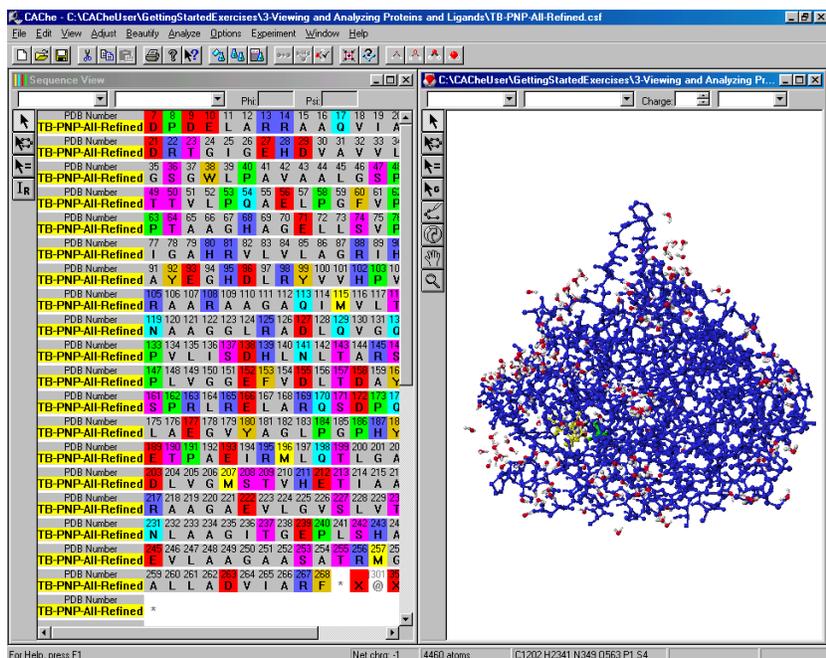
When you view a sequence, the Sequence View window comes to the front and the menus and toolbar change.

3. Choose **Window | Tile**.

The Sequence and 3D Structure Windows are displayed side-by-side.

4. Scroll the Sequence View Window to the bottom.

Your view should look like this



TIP

To orient your TB-PNP as shown, click in the 3D Structure window to activate it, type the letter 'i'. The protein disappears. The letter command 'i' restores the initial viewing transform which in this case placed the protein off the screen. Choose **View | Fit in Window** to bring the protein back into the window.

An asterisk "*" called the "chain terminal" marks the end of each protein chain. Note that the chain terminal is not part of the structure and you cannot select the chain terminal.

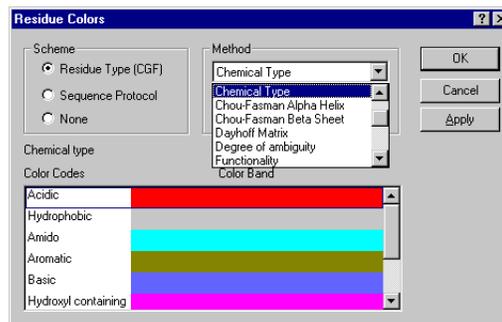
Analyzing the protein sequence

Initially residues in the Sequence View are colored by the chemical type method. You can change the coloring method to display other properties.

To color a sequence by property or residue type

1. Click on the title bar of the Sequence View to activate it and choose **View | Residue Colors...**

The Residue Colors dialog appears.



The Chemical Type method is highlighted.

2. Choose the color **Scheme | Sequence Protocol**.

The list of choices in the **Method** pulldown menu changes.

3. Choose Method **Karplus and Schulz Flexibility** and select **OK**.

The dialog closes and the residue colors change to indicate flexibility based on the local 7-residue sequence. A red residue is predicted to be in a highly flexible portion of the protein. A blue residue is in a relatively inflexible portion of the protein.

Method details are explained in the User Guide Chapter 21, “Understanding Sequence Property Predictions”

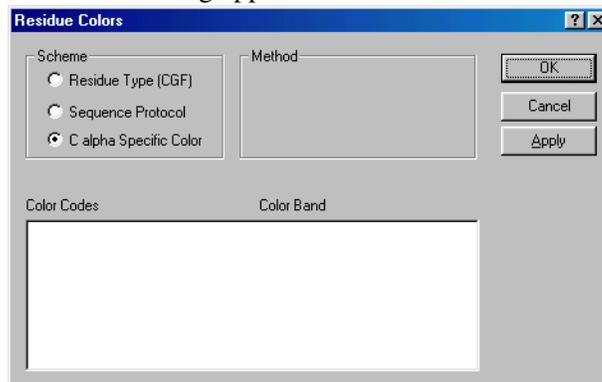
The **Method Window** specifies the number of adjacent residues that are grouped together for analysis. For example, a window of 7 will result in the sequence being analyzed in blocks of 7 residues to determine properties such as the antigenicity or hydrophathy. Note that residues at the beginning and end of the sequence are not colored because they are outside the Method Window.

You may also color the protein by property in the 3D Structure window.

To color the 3D structure by sequence property or residue type

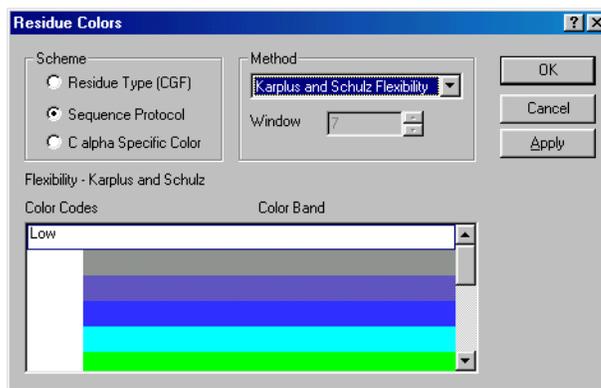
1. Click on the title bar of the 3D Structure window to activate it and choose **View | Color by Residue**

The Residue Colors dialog appears.



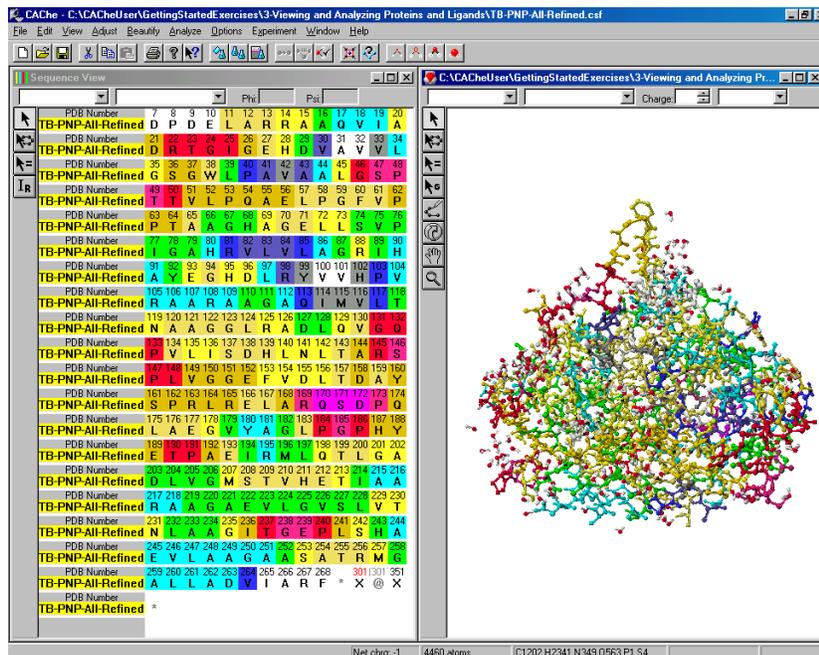
Initially, the colors assigned are the atom specific colors that you have set in the **View | Atom Attributes** dialog.

2. Choose the color **Scheme | Sequence Protocol**.
The list of choices in the **Method** pulldown menu changes.
3. Choose Method **Karplus and Schultz Flexibility** and select **OK**.



The dialog closes and the atom colors change to indicate flexibility based on the local 7 residue sequence. A red residue is predicted to be in a highly flexible portion of the protein. A blue residue is in a relatively inflexible

portion of the protein..

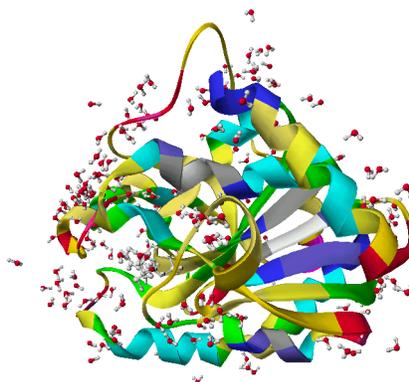


Simplifying the protein structure: ribbons

➤ To display the 3D structure as a ribbon

1. Click on the title bar of the 3D Structure window to activate it and choose **View | Backbone Ribbon | Solid Ribbon**

You see ribbons and tubes colored by the residue flexibility that trace the protein backbone.



You may also color the ribbon by property in the 3D Structure window.

➤ To color the ribbon by sequence property or residue type

1. Click on the title bar of the 3D Structure window to activate it and choose

View | Color by Residue

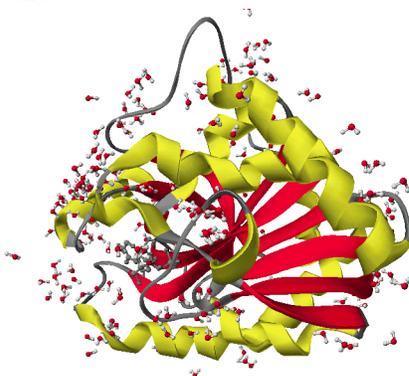
The Residue Colors dialog appears.

2. Choose the color **Scheme | Residue Type**.

The list of choices in the **Method** pulldown menu changes.

3. Choose Method **Secondary Structure** and select **OK**.

4. The dialog closes and the ribbon colors change to indicate secondary structure:



Helices are colored yellow, sheets are colored red and other regions of the protein are colored grey.

To turn off the ribbon

1. Click on the title bar of the 3D Structure window to activate it and choose **View | Backbone Ribbon | None**

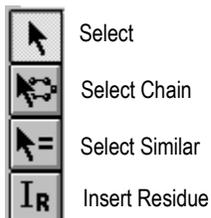
The ribbon disappears and the protein reappears with atoms colored by secondary structure.

2. In the 3D Structure window, click the background with the Select Tool.
All visible objects highlight.

3. Choose **View | Color by Element**.

All atoms are colored by element type.

Locating the active site



To locate residues in the active site

1. In the Sequence View Window, choose the Select tool from the tool palette and click **IMH**.

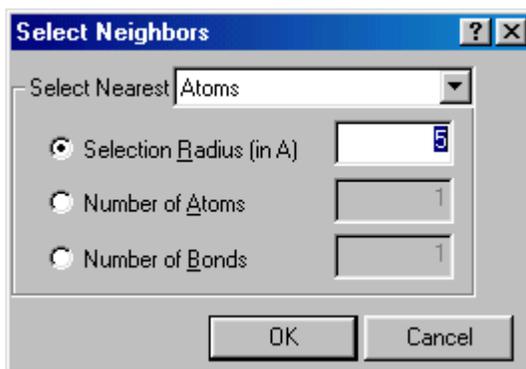
The bound ImmH ligand highlights and the rest of the sample becomes grey. The sequence view provides fast access to structures by residue and group.

2. Click on the title bar of the 3D Structure window to activate it.

The title bar becomes blue indicating the 3D Structure window is active.

3. In the 3D Structure window, choose **Edit | Select Neighbors**.

The **Select Neighbors** dialog appears.



4. Choose **Residues, waters, HETs** from the **Select Nearest** pull down menu in the Select Neighbors dialog, enter a **Selection Radius** of 3 Å and choose **OK**.

The dialog closes and atoms within 3 Å of any atom in ImmH are selected. In addition, residues in Sequence View that contain a selection atom are highlighted and all others are dimmed.

5. Scroll through the sequence in the Sequence View and observe which residues have been highlighted.

Note that residues near the ligand are often far from each other in the sequence. The three dimensional structure is necessary to understand active sites. Also note that no residues have very high flexibility. The residues associated with the active site have low to moderate flexibility.

6. In the Sequence View window, ctrl-click on IMH.

IMH unhighlights in both the Sequence View and 3D Structure windows.

At this point you should have all the residues in the active site selected. The 16 residues are:

Ser36, His90, Tyr92, Ala120, Ala121, Gly122, Tyr188, Glu189, Val205, Gly206, Met207, Ser208, Thr230, Asn231, His243, Val246.

PO_4^{3-} and two water molecules are also selected.

Differences in protein-ligand active sites may be important in determining binding selectivity. In Exercise 6, you will examine the structure of human PNP and determine whether the active site residues in TB-PNP are conserved in human PNP.

 **To name the active site**

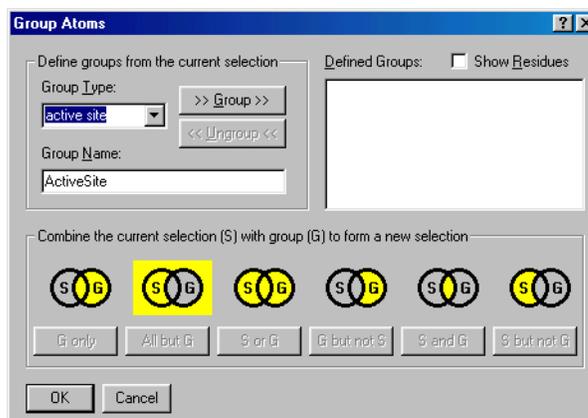
1. Move into the 3D Structure window by clicking its title bar, choose **Edit | Group Atoms**.

The Group Atoms dialog appears.

2. Choose **active site** from the **Group Type** drop down menu.

The group type is used in CAChe to specify whether a group is an amino acid, ligand, nucleic acid, hetero group, active site or none of these.

3. Type the name "ActiveSite" into the **Group Name** text box.



The **>>Group>>** button highlights as you type.

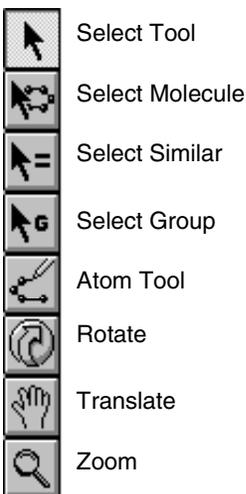
4. Choose **>>Group>>**.

The ActiveSite group is created and added to the **Defined Groups** list on the right.

5. Choose **OK**.

The Group Atoms dialog closes.

In the future, you will be able to pick the active site quickly from **Edit | Group Atoms** or by using the Select Group Tool.



Locating the hydrogen bonds

ImmH has specific hydrogen bond interactions with TB-PNP. Understanding these interactions is important when designing new inhibitors. If other interactions are equal, inhibitors with more specific hydrogen-bond interactions will bind better than those with fewer interactions.

TIP

To turn off the numbers on the H-bond labels, choose the Select Similar Tool, click on an H-bond label to select all H-bond labels, then choose **View | Geometry Label Attributes**. Uncheck **Value** on the **Label** tab and choose **OK**. Reselect ImmH to continue the exercise.

To display hydrogen bonds between ImmH and TB-PNP

1. In the Sequence View window, with the Select Tool click **IMH**.
ImmH highlights in both the 3D Structure and Sequence View windows.
2. Click the title bar in the 3D Structure window to activate it and choose **Analyze | Label H-bonds**.
Blue distance labels appear indicating the distance between hydrogen bond donors and acceptors. Each blue distance label measures the length of a hydrogen bond interaction.
3. Click **View | Fit in Window**.
ImmH fills the window.

To identify the hydrogen bonding residues in TB-PNP

1. In the 3D Structure window, click a protein atom at the end of an H-bond distance label. Then, shift-click each atom at the end of each of the remaining protein H-bonds to ImmH. Ignore the H-bonds to PO_4^{3-} .
The selected atoms highlight in the 3D Structure Window and their residues highlight in the Sequence View Window. Other residues dim.
2. Scroll through the Sequence View and note the residue type and name.
You should identify bonding to at least 5 different residues: Tyr188, Glu189, Met207, Asn231 (2 bonds), and His243. You may identify more. The hydrogen bond to Met207 is relatively long and expected to be weaker than the other hydrogen bonds.

TIP

To delete all the H-bond and other distance labels, choose the Select Similar Tool, click on an H-Bond label to select all H-bond and distance labels, then choose **Edit | Delete**.

Changes in ligand-protein hydrogen bonding can be important in determining binding selectivity. In Exercise 6, you examine the structure of human PNP and determine whether these residues important to hydrogen bonding are conserved.

Displaying the active site pocket

The TB-PNP protein active site is a pocket in the protein that contains the bound ligand. The surface of the protein near a bound ligand maps out the pocket. To generate this surface in BioMedCACHe, the ligand is selected first and then the accessible surface of the protein contained in a box enclosing the ligand is drawn. This is the surface of the protein adjacent to the ligand and the surface that forms a pocket around the ligand.

↪ **To view the surface of the binding site adjacent to the ligand**

1. In the Sequence View Window, choose the Select Tool and then choose residue **IMH**.

ImmH highlights in the Sequence View Window and in the 3D Structure Window.

2. Activate the 3D Structure Window by clicking on its title bar, then choose **View | Show All**.

You do this to ensure that all atoms are included in the surface determination. Hidden atoms do not contribute to accessible surfaces.

3. Next choose **Analyze | Adjacent Surface - Pocket**.

After several seconds, a wireframe surface is drawn around ImmH.

4. Choose **View | Hide Unselected**.

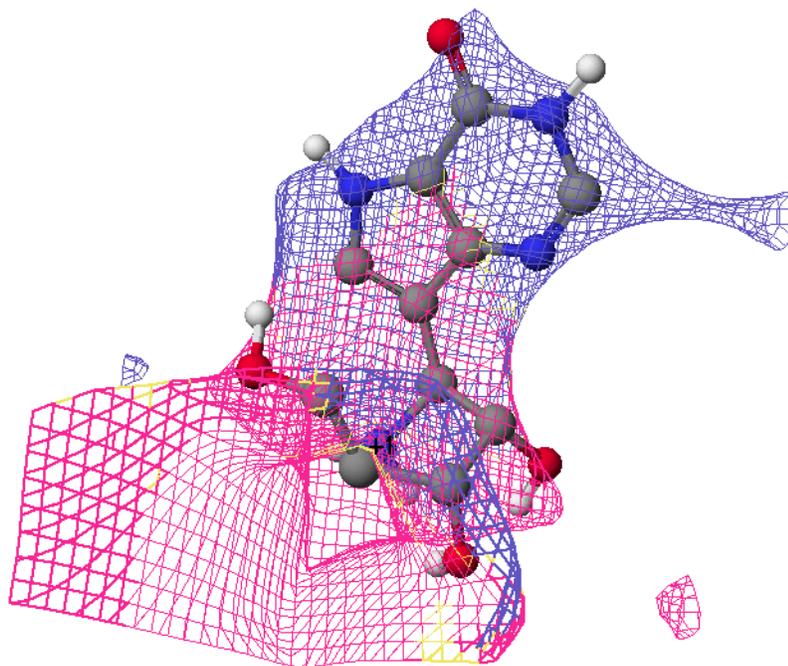
All atoms and bonds except ImmH disappear. The adjacent surface¹ and ImmH remain visible.

5. Choose **View | Fit in Window**.

You see a view similar to this.

📖 **NOTE**

You may have to rotate the structure to orient it like this view.



Red marks a surface where the protein needs H-acceptors (*e.g.* -C=O), blue marks a surface where the protein needs H-donors and cream marks the hydrophobic surface. The pocket surface is colored so that it is easy to design ligands. Ligands that bind well should have H-bond acceptors (*e.g.*

1. Bohacek, R. and McMartin, C., *J Med Chem* **1992**, 35, 1671-1684.

-C=O) groups touching the red surface and N-H groups with the N-H bond poking through the surface. Note ImmH's hydrogen bond donors hook through the surface and ImmH's hydrogen bond acceptors touch the surface. ImmH's hydrophobic atoms are far from the surface.

Notice extra space near the blue -NH-C=N- region in the upper right-hand corner. This empty space suggests that ImmH might be modified to yield compounds of higher affinity by the addition of functional groups to the carbon. The new functional group should donate hydrogens to acceptor groups on the protein which do not currently form hydrogen bonds. You will check for this region in human PNP. If it is not present, then you may be able to exploit this region for selectivity.

This completes the viewing and analyzing proteins and ligands introductory exercise. As you work with BioMedCACHe you will discover many additional views and analysis methods.

4

Docking Ligands into Proteins

Overview

NOTE

“The antithesis of rational methods for drug design is the complete reliance on chance discoveries of active ligands. High-throughput screening (HTS) is a chance-based method. The practical limitations of HTS are twofold: the volume of screen throughput that can be achieved within commercial constraints and the theoretical coverage within the diversity of the compound set being screened. If the number of compounds currently in existence is, for argument’s sake, 10^8 and the number of drug-like molecules that could potentially be made is 10^{30} , the probability of a compound set of a size currently in existence showing useful statistical coverage of chemical space is minuscule.”¹

Structure-guided drug discovery is an established approach to discovering and optimizing lead compounds. Structure-guided methods have resulted in the discovery of entirely novel classes of compounds. By contrast, high-throughput screening methods are often best used to optimize lead molecules that are discovered in a variety of ways.¹

Screening commercially available compounds provides a list of potential lead compounds based upon established and known structures. Unfortunately, these structures are likely to be well-known and therefore may not be patentable. To create a new class of compounds that are patentable, you need to create innovative structures which are also active. A powerful way to design new candidate drugs is to compare the interactions of these molecules, when bound within the protein, with those of known active compounds.

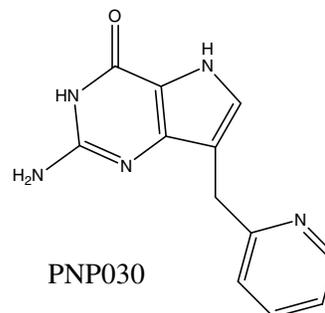
In this exercise, you will use CAChe to

- dock a ligand into a binding site by superposition
- view the docked ligand inside the active site pocket wireframe
- manually refine the position
- evaluate the quality of the binding by counting the number of binding interactions
- refine the docking with molecular mechanics

1. Gane, P. J. and Dean, P. M., “Recent advances in structure-based rational drug design”, *Current Opinion in Structural Biology*, **2000**, *10*, 401-404.

Background

In the previous exercise, you identified the binding site of ImmH in TB-PNP. Here, you take compound 30 (PNP030 . csf), one of the most active inhibitors (IC_{50} 15.3 nM) of calf spleen PNP, from a class of 9-substituted-9-deazaguanine compounds and dock it into TB-PNP.¹

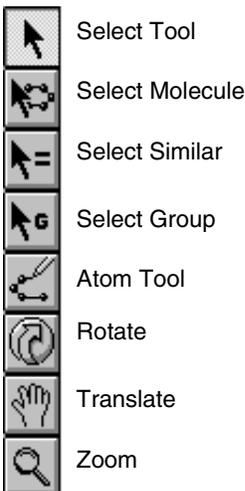


Docking by superposition

The fastest and easiest method for docking a ligand into an active site is to superimpose the ligand onto a bound ligand already in the active site and then delete the bound ligand.

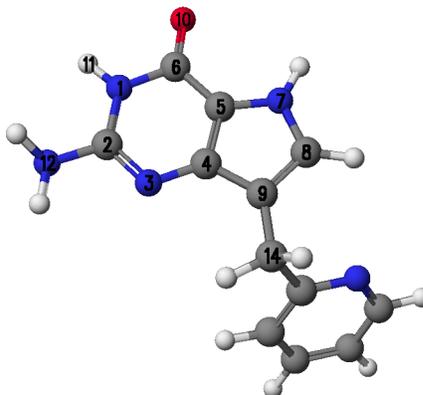
To dock PNP030 into TB-PNP

1. Navigate to the folder 4-Docking Ligands Into Proteins, choose **File | Open** and select TB-PNP-All-Refined.csf.
TB-PNP-All-Refined.csf opens and displays the cleaned structure.
2. Choose **Edit | Group Atoms**, select ImmH from the list of **Defined Groups**, choose **G Only**.
ImmH highlights.
3. Choose **OK**
The Group Atoms dialog closes.
4. Choose **View | Hide Unselected**.
All parts of the chemical sample except ImmH disappear.
5. Choose **View | Color by Element**.
Atoms and bonds in ImmH are colored according to element type.
6. Choose **File | Open** and select PNP030.csf.



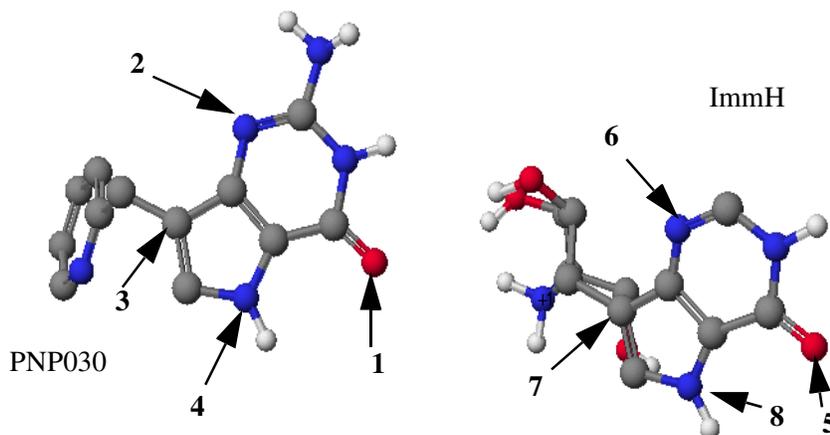
1. Farutin, V., Masterson, L., Andricopulo, A.D., Cheng, J., Riley, B., Hakimi, R., Frazer, J. W., and Cordes, E. H., "Structure-Activity Relationships for a Class of Inhibitors of Purine Nucleoside Phosphorylase", J. Med. Chem., **1999**, 42, 2422-2431.

PNP030 opens in a new 3D sample window.



7. With the Select Molecule tool and click on an atom in PNP030.
PNP030 is highlighted.
8. Choose **Edit | Copy**.
PNP030 is copied to the clipboard.
9. Choose **Window | ...\TB-PNP-All-Refined.csf**.
The 3D Sample Window for TB-PNP comes to the front.
10. Choose **Edit | Paste**.
PNP030 appears in the center of the TB-PNP window and the atom numbers on PNP030 change.
11. Choose **Edit | Group Atoms**.
The Group Atoms dialog appears.
12. Choose **ligand** from the **Group Type** pull down menu.
13. Type PNP030 into the **Group Name** text box and choose **>> Group >>**.
PNP030 is added to the **Defined Groups** list.
14. Choose **OK**.
The Group Atoms dialog closes. PNP030 is still highlighted and ImmH is not highlighted.
15. Choose **Edit | Move Selected** and use the Rotate Tool and Translate Tool to position PNP030 so that the fused rings have the same orientation as ImmH.

The contents of your window looks similar to this. PNP030 is on the left



NOTE

When superimposing molecules, the order of selection is important because, the first atom selected in the first molecule is superimposed on the first atom selected in the second molecule, etc.

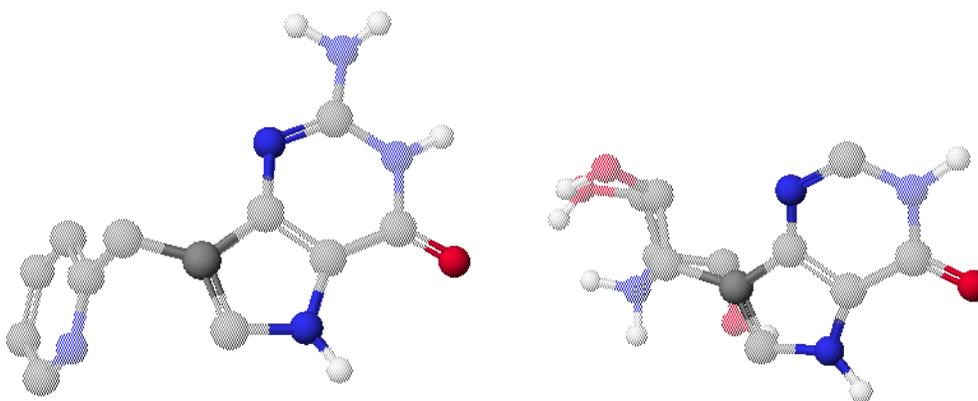
and ImmH is on the right.

- With the Select Tool, click first on the carbonyl oxygen (1) in PNP030, and then hold the shift down and click on the atoms pointed to with arrows 2, 3 and 4.

The selected atoms are highlighted. All others are dimmed.

- With the Select Tool, hold the shift key down and - following the same order - click on the corresponding atoms in ImmH (5, then 6, then 7, then 8).

Your structures should look similar to this



TIP

The molecule containing the first selected atom is the one that moves.

- Choose **Analyze | Superimpose**.

PNP030 moves on top of ImmH so that first selected atom in PNP030 is superimposed on the first selected atom in ImmH, the second selected in PNP030 is superimposed on the second selected in ImmH, and so forth. The RMS error for the four superimposed pairs of atoms is displayed as a text label.

19. Choose **View | Color by Molecule**.

The color of PNP030 and ImmH change.

20. Choose **View | Show All**.

TB-PNP appears with PNP030 docked in the active site and superimposed on ImmH.

21. Choose **File | Save As** and name the file `TB-PNP+PNP030.csf`.

Adjusting the docked position in the active site pocket

The protein active site is a pocket in the protein that contains the bound ligand. The surface of the protein adjacent to a bound ligand maps out the pocket. You use the adjacent surface to dock the ligand.

To view the surface of the binding site adjacent to PNP030

1. Choose **Edit | Group Atoms**, click on PNP030 in the **Defined Groups** list, and choose **G Only**.

PNP030 highlights and the rest of the structure dims.

2. Click on ImmH in the **Defined Groups** list, and choose **S or G**.

PNP030 and ImmH highlight and the rest of the structure is dim.

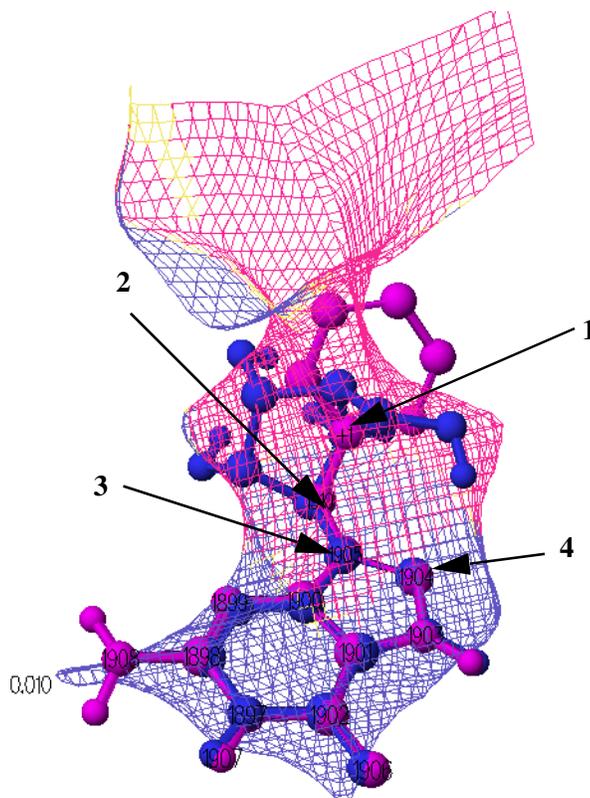
3. Choose **OK**.

4. Next choose **Analyze | Adjacent Surface - Pocket**.

After several seconds, a wireframe surface is drawn around PNP030.

5. Choose **View | Hide Unselected**.

You see a view similar to this.



Blue marks a portion of the surface where the ligand needs H-bonds, red marks a portion of the surface where the ligand needs hydrogen bond donors and cream marks places where it should have a hydrophobic surface to achieve good binding to the protein.

Note that PNP030's hydrogen bond donors (*e.g.* -NH) hook through the blue surface and that PNP030's hydrogen bond acceptors (C=O) touch the red surface instead of sticking through it. PNP030's hydrophobic atoms should be, but are not, buried inside the surface. This suggests that PNP030 in this conformation will not bind as well as ImmH.

Note that the -NH₂ group on PNP030 fills the empty space identified in the previous exercise with H-bond donors.

6. You can set another conformation for PNP030 by rotating around flexible single bonds. With the Select Tool, click on atom **1** in the above figure. Then while holding down the shift key, click on atoms **2**, **3** and **4** in order. Be careful to choose all atoms in PNP030. It is easy to pick atoms from ImmH.

Atoms **1**, **2**, **3** and **4** highlight. The rest of the molecule dims.

7. Choose **Adjust | Dihedral Angle**.

The Set Dihedral Angle dialog appears.

8. Type -85.0 into the Angle text box, check **Define Geometry Label** and choose **OK**.

The pyridine ring moves to fit inside the binding pocket and the rotated angle is labeled -85.00 degrees.

You can further refine the geometry by selecting the whole molecule and then using **Edit | Move Selected** and the Rotate tool and Translate tool to move PNP030 around in the wire frame pocket.

9. With the Select Molecule tool, select ImmH and choose **Edit | Delete**.
ImmH disappears and PNP030 is highlighted.

10. Choose **File | Save**.

TB-PNP+PNP030.csf is updated.

Evaluating the docking

The binding energy depends upon the number of hydrogen bonds between the ligand and the protein and on any overlapping atoms that bump too close together.

To find the number of hydrogen bonds between PNP030 and TB-PNP

1. With the Select Molecule Tool, click on PNP030.
PNP030 highlights.
2. Then choose **View | Show All**.
The rest of the structure appears. It should all be dimmed.
3. Next choose **Analyze | Label H-bonds**.

Distance labels appear for each hydrogen bond between the PNP030 and the protein and for each hydrogen bond internal to PNP030. You see a total of 3 intermolecular hydrogen bonds binding PNP030 and TB-PNP.

To find any bumping atoms between PNP030 and TB-PNP

1. With the Select Molecule Tool, click on PNP030.
PNP030 highlights and the rest of the structure dims.
2. Next choose **Analyze | Label Bumps**.
New distance labels appear for each atom pair that is too close together. Ideally, you should see no new distance labels. However, PNP030 bumps into adjacent atoms and several labels appear especially near the pyridine ring. These bumps suggest that the binding of PNP030 into a rigid receptor is not as good as that of ImmH. Refinement of the manual docking with molecular mechanics is required before further analysis.

NOTE

Here, PNP030 does not bump into PO_4^{3-} . This suggests the possibility that PO_4^{3-} is also bound with PNP030. If you decide that it does not co-bind, then delete PO_4^{3-} before refinement

Refining the docking

The docking can be refined by molecular mechanics which you setup to allow both the receptor and the ligand to relax.

To refine the PNP030 docking in TB-PNP

1. Choose **Experiment | New**.
The Experiment Dialog appears.
2. Choose **Property of: | chemical sample, Property: | optimized geometry** and **Using: | MM geometry (MM3)**.
3. Click **Start**.

TB-PNP+PNP030.csf is saved, the Experiment Status window opens and a molecular mechanics calculation using the MM3 force-field runs. When it completes after about 10 minutes, the 3D Sample window is updated.

So what?

You have seen how the new ligand, PNP030, could interact with TB-PNP. You have counted the hydrogen bonding interactions, noted that there is unexploited space in the active site pocket, and examined whether PNP030 bumps into other parts of the structures.

You should note that the -NH₂ group in PNP030 exactly fills the unexploited space in the active site discovered in the previous exercise.

This analysis has helped you understand the affinity of the ligands for TB-PNP. But a good drug is also selective. Therefore, you need to understand how ligands that bind to TB-PNP bind to related proteins such as human PNP. Selective ligands will have at least one binding interaction that is different between TB-PNP and human PNP.

In the next exercise, you will use the docked TB-PNP+PNP030 structure to locate the active site in the homologous human-PNP, dock PNP030 into human PNP and look for selective interactions.

5

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 effects?

To answer this question, you need to determine how your lead compound might interact with human PNP. Ideally this could be accomplished by obtaining a crystal structure of the lead compound (PNP030) bound to human PNP. Unfortunately, this is often not possible for a variety of technical reasons. In addition, if the ligand is selective it will not bind to human PNP in any case!

Molecular modeling represents a relatively rapid approach to determining the likelihood that your lead compound will exhibit binding selectivity. You use modeling to determine how the ligand binds to human PNP, so you can understand whether the ligand will exhibit high selectivity. Modeling provides the basis for understanding the mechanism for selectivity.

In the case of PNP there is a crystal structure for the human PNP enzyme, but there is no ligand bound to it. 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 1ULA lacks a 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-PNP³ 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, the active site residues and their relative geometries are conserved in evolutionarily related enzymes that catalyze the same transformation. 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 human PNP.

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

In the following exercise, you will use files in the directory 5-Discovering an Active Site from a Homolog.

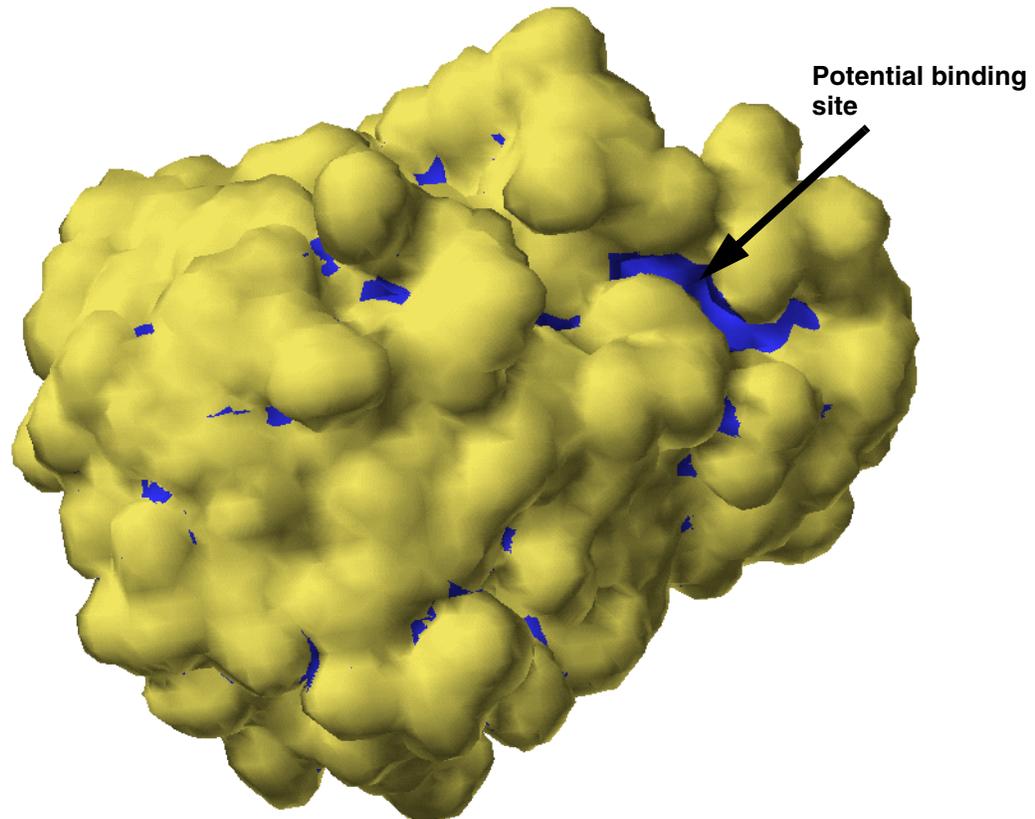
-
1. Cook, W.J., Ealick, S. E., Bugg, C. E., Stoeckler, J. D., Parks, R. E., "Crystallization and Preliminary X-ray Investigation of Human *Erythrocytes* Purine Nucleoside Phosphorylase", *J. Biol. Chem.*, **1990**, 285, 1812.
 2. Journal of Computer-Aided Molecular Design, **2000**, 14, 383-401.
 3. 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.

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 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 more 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. To do so, 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 1ULA (human-PNP) as **Sequence1** and TB-PNP+PNP030 as **Sequence2** and press **OK**.

Gaps appear in human-PNP and in TB-PNP that align the sequences according to the maximum scoring alignment using the BLOSUM50¹ substitution matrix in the Needleham-Wunsch alignment algorithm².

1. Durbin, R. , Mitchison, G. , Eddy, S. , Krogh, A., *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1997.

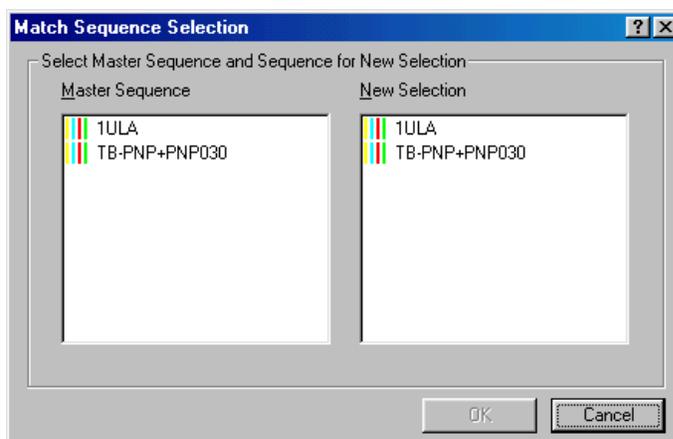
↪ **To select the active site residues in TB-PNP**

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

-
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 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 open 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. Choose **active site** from the **Group Type** pull down menu.
3. 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.

4. Choose **File | Save**.

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

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.

NOTE

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

Validating the alignment using multiple sequences

Clustal-W multiple sequence alignment of purine nucleoside phosphorylase homologs

Y. pestis PNP	-----MATPHIN-AEMGDFADVLLMPGDPLRAKFAIETFLQD	36
E. coli PNP	-----MATPHIN-AEMGDFADVLLMPGDPLRAKYIAETFLQD	36
H. influenzae PNP	-----MTPHIN-APEGAFADVLLMPGDPLRAKYIAETFLQD	35
V. cholerae PNP	-----MATPHIN-AQMGDFADVLLMPGDPLRAKYIAENFLDN	36
S. aureus PNP	-----MKSTPHIKPMNDVEIAETVLLPGDPLRAKFAIETYLDD	38
H. sapiens PNP	-MENGYYTYEDYKNTAEWLLSHTK--HRPQVAICGSG-LGGLTDKLTQAQ	46**
B. taurus PNP	-MQNGYTYEDYQDTAKWLLSHTTE--QRQVAVICGSG-LGGLVNKLTQAQ	46
M. musculus PNP	-MENEFTYEDYETAKWLLQHTE--YRPQVAVICGSG-LGGLTAHLKEAQ	46
B. subtilis PNP	-MK-----DRIERAAAFIKQNLN--ESPKIGLILGSG-LGILADEIENPV	41
M. tuberculosis PNP	MADPRPDPDELARRAAQVIADRTGIGEHDAVAVLGSGLWLPAAALGSPTT	50++
P. aerophilum PNP	-----MVKLTNPKSPKELGFDEFPSIGIIGSG--LYDPGIFENAV	40
Y. pestis PNP	VREVNVRVRLMLG-----FTGTYKGRKISVMG--HGMGIPS----	69
E. coli PNP	AREVNVRVRLMLG-----FTGTYKGRKISVMG--HGMGIPS----	69
H. influenzae PNP	VVEVTNVRNMLG-----FTGTYKGRKISIMG--HGMGIPS----	68
V. cholerae PNP	AVQVCDVRNMF-----YTGTYKGRKISVMG--HGMGIPS----	69
S. aureus PNP	VEQFTNVRNMF-----FTGTYKGRKISVMG--SGMGMP-----	71
H. sapiens PNP	IFDYGEIPNFPRSTVPGHAGRLVFGFLNGRACVMMQ--GRFHM ^Y EGYPLW	94**
B. taurus PNP	TFDYSEIPNFPPESTVPGHAGRLVFGILNGRACVMMQ--GRFHM ^Y EGYPFW	94
M. musculus PNP	IFDYNEIPNFPQSTVQGHAGRLVFGLLNGRCCVMMQ--GRFHM ^Y EGYSLS	94
B. subtilis PNP	KLKYEDIPEFPVSTVEGHAGQLVLGTLEGVSVIAMQ--GRFHF ^Y EGYSME	89
M. tuberculosis PNP	VLPQAEPLPGFVPTAAGHAGELLSVPIGAHRVVLAA--GRIHA ^Y EGHDLR	98++
P. aerophilum PNP	EVQIHTPYGLPSDN-----VIVGRVAGRVVAFLRPHGRG ^{HKY} P ^{PH} KIP	83
Y. pestis PNP	CSIYAKELITDFGVKKIIRVSGCGAVRTDVKLRDVIIGMAGCTDSKVNRI	119
E. coli PNP	CSIYTKELITDFGVKKIIRVSGCGAVLPHVKLRDVIIGMAGCTDSKVNRI	119
H. influenzae PNP	CSIYAKELITEYGVKKIIRVSGCGTVRMDVKVRDVIIGLACTDSKVNRI	118
V. cholerae PNP	CSIYVTELIKIDYGVKKIIRVSGCGAVNEGLKVRDVIIGMAGCTDSKVNRI	119
S. aureus PNP	IGIYSYELIHTFGCKKLRVSGCGAMQENIDLVDVIAAQASTDSNYVQQ	121
H. sapiens PNP	KVTFPVRVRFLLGVETLVVTAAGGLNPNFVVDI ^{ML} IRDHINLPGFSGQ	144**
B. taurus PNP	KVTFPVRVRFLLGVETLVVTAAGGLNPNFVVDI ^{ML} IRDHINLPGFSGE	144
M. musculus PNP	KVTFPVRVRFLLGVETLVVTAAGGLNPNFVVDI ^{ML} IRDHINLPGFCGQ	144
B. subtilis PNP	KVTFPVRVMKALGVEALIVTNAAGGVNTEFRAGDLMIT ^{DH} IN--FMGT	136
M. tuberculosis PNP	YVVH ^{PV} RAARAAGAQMVLVTAAGGLRADLQVGPV ^L ISDHLN--LTAR	145++
P. aerophilum PNP	YRAN-IYSLYMLGVRISIVAVS ^{AV} GLR ^{PDY} APGDFVVPDQFVDMTKGREY	132
Y. pestis PNP	RFKDH-----DYAAIADFEMTRNAVDAAKAKG--VNVRVGNLFS	156
E. coli PNP	RFKDH-----DFAAIADFDMVRNAVDAAKALG--IDARVGNLFS	156
H. influenzae PNP	RFKDN-----DFAAIADFMAQAQVQAARAKG--KVVRVGNLFS	155
V. cholerae PNP	RFKDH-----DFAAIADYKMKVKAEEAAKARG--IDVKVGNLFS	156
S. aureus PNP	YQLPG-----HFAPIASYQLLEKAVETARDKG--VRHHVGNVLS	158
H. sapiens PNP	NPLRGPNDERFGDRFPAMSDAYDR TM RQALSTW ^K QMG ^Q RELQ ^B EGTYVM	194**
B. taurus PNP	NPLRGPNEERFGVRF ^P AMSDAYDR ^{MR} QKAHSTW ^K QMG ^Q RELQ ^B EGTYVM	194
M. musculus PNP	NPLRGPNDERFGVRF ^P AMSDAYDR ^{MR} QKAFTAW ^K QMG ^Q RKLQ ^B EGTYVM	194
B. subtilis PNP	NPLIGPNEADFGARFPDMSAYDKDLSLAEKIAKDLN--IPIQKGVYTA	184
M. tuberculosis PNP	SPLVG-----GEFVDLTDAYS ^{PR} LRELARQSD-----PQLA ^B EGVYAG	182++
P. aerophilum PNP	TFYDGP-----TCHIQIGLEPFTQEI ^{RQ} IL ^I ETAKKYN--RTHDGGCYVC	176
Y. pestis PNP	ADLFYTPDPQMFDVM-EKYGILGVEMEAAGICGVAEFGAKALTI ^{CT} VSD	205
E. coli PNP	ADLFYSPDGMFDVM-EKYGILGVEMEAAGIYGVAEFGAKALTI ^{CT} VSD	205
H. influenzae PNP	ADLFYTPDVMFDVM-EKYGILGVEMEAAGIYGVAEYGA ^K ALTI ^{CT} VSD	204
V. cholerae PNP	AELFYTPDPSMFDVM-DKYGILGVEMEAAGIYGVAEYGA ^K ALAI ^{CT} VSD	205
S. aureus PNP	SDIFYNADTTASERW-MRMGILGVEMESAALYMNAIYAGVEALGVFTVSD	207
H. sapiens PNP	VAGPS ^{FET} V ^A E ^{CR} LL-QKL ^G ADA ^V M ^S T ^V PE ^V I ^V ARH ^C GL ^R V ^F G ^S L ^I TN	243**
B. taurus PNP	LGGPN ^{FET} V ^A E ^{CR} LL-RNL ^G ADA ^V M ^S T ^V PE ^V I ^V ARH ^C GL ^R V ^F G ^S L ^I TN	243
M. musculus PNP	LAGPN ^{FET} V ^A E ^S RLL-KML ^G ADA ^V M ^S T ^V PE ^V I ^V ARH ^C GL ^R V ^F G ^S L ^I TN	243
B. subtilis PNP	VTGPS ^{YET} PAEVRFL-RTMGSDA ^V M ^S T ^V PE ^V I ^V ANHAGMRVLG ^S IS ^I SN	233
M. tuberculosis PNP	LPGPH ^{YET} PAEIRML-QTL ^G ADL ^V M ^S T ^V HET ^I AARAAGA ^E RVL ^G SVL ^V TN	231++
P. aerophilum PNP	IEGPR ^F STKAESRIWREVFCDI ^I GMTL ^V PEINLARELGM ^C YGLIAL ^V TD	226
Y. pestis PNP	HIRTGEQTT---A ^A ERQ ^T TFNDMIEIAL ^S VLLGDNA-----	239
E. coli PNP	HIRTHEQTT---A ^A ERQ ^T TFNDMIKIAL ^S VLLGDKE-----	239
H. influenzae PNP	HIRTHEQTT---A ^E ERQ ^L TFNDMIEIAL ^S VDLIGDAL-----	238
V. cholerae PNP	HIKTGEQTT---S ^E ERQ ^N TFNEMIEIAL ^S VDLIGDQAGY-----	241
S. aureus PNP	HLIHETSTT---P ^E ERERAF ^T DMIEIAL ^S LV-----	235
H. sapiens PNP	KVIMDY ^S LEK ^{AN} HEE ^V LEAGKQAAQKLEQFV ^S ILMASI ^P LPDKAS---	289**
B. taurus PNP	KVIMDY ^S EQK ^{AN} HEE ^V LEAGKQAAQKLEQFV ^S ILMASI ^P VSGHTG---	289
M. musculus PNP	KVIMDYENLEK ^{AN} HEE ^V LDAGKAAQTLERFV ^S ILMESI ^P LPDRGS---	289
B. subtilis PNP	AAAGLLDQP--LS ^H DE ^V MEVTEKVKAGFLKLVKAI ^V AQYE-----	271
M. tuberculosis PNP	LAAGITGEP--LS ^H AE ^V LAAGAASATRMGALLAD ^V IARF-----	268++
P. aerophilum PNP	YDIWVPHQP--VT ^A EA ^V EKMMTEKLGIIK ^V IAEAVPKLPAELPK ^C SETL	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 in *E. coli*, *Y. pestis*, *H. influenzae*, *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

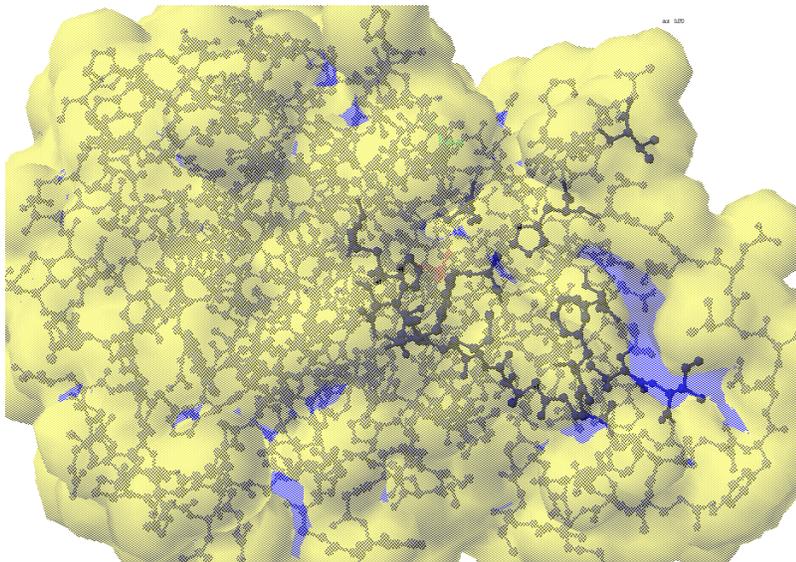
1. From the 1ULA 3D Structure window, choose **Analyze | Show Surfaces**.

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.



3. Choose **Analyze | Surfaces**, uncheck **crevice1.acs** and press **OK**.
The crevice surface disappears.

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. If not, use the **Edit | Match Selection** to reselect the matching residues in 1ULA.
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.
5. Choose **Superimpose Probe onto Target** and click **OK**.
After approximately 10 seconds, the Superimpose Sequences dialog closes and superposition is complete.

NOTE

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

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, PO_4^{3-} 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 4-43.

Refining the docking

The docking can be refined by molecular mechanics which you setup to allow both the receptor and the ligand to relax.

To refine the PNP030 docking in 1ULA

1. Choose **Experiment | New**.

The Experiment Dialog appears.

2. Choose **Property of: | chemical sample, Property: | optimized geometry** and **Using: | MM geometry (MM3)**.

3. Click **Start**.

`1ULA+PNP030.csf` is saved, the Experiment Status window opens and a molecular mechanics calculation using the MM3 force-field runs. When it completes after about 10 minutes, the 3D Sample window is updated.

So what?

You have seen how the new ligand, PNP030, could be docked in human PNP using a superposition method. You have generated an adjacent surface map of the active site pocket and seen that there is a difference between the TB-PNP pocket and the TB-PNP pocket that might be exploited to make a selective drug.