Getting Started with BioMedCACHe

Molecular Modeling in Drug Design

CAChe Group
Fujitsu

Beaverton, Oregon

Copyright statement

Published by **Fujitsu Limited**, September, 2003.

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

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

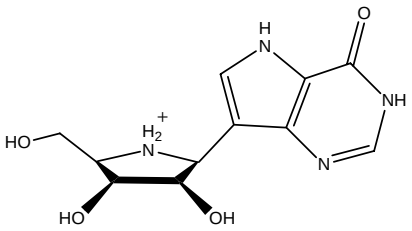
Overview

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 (1G2O) of *Mycobacterium tuberculosis* purine nucleoside phosphorylase (TB-PNP) complexed with the transition-state analogue Immucillin-H (ImmH) and phosphate.[1]



Immucillin-H

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 individuals[2].

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.

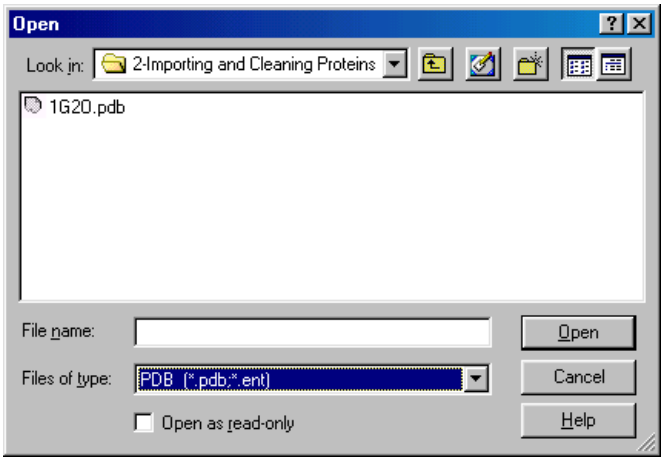
The structures of proteins and other biomolecules can be downloaded from `http://www.rcsb.org/pdb/` as PDB files. PDB files have an extension of `.pdb` or `.ent`. To view the 3D structure of TB-PNP complexed with ImmH, you first download `1G2O.pdb` from the PDB and then open it in the CACHe workspace.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

📖 NOTE

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

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.

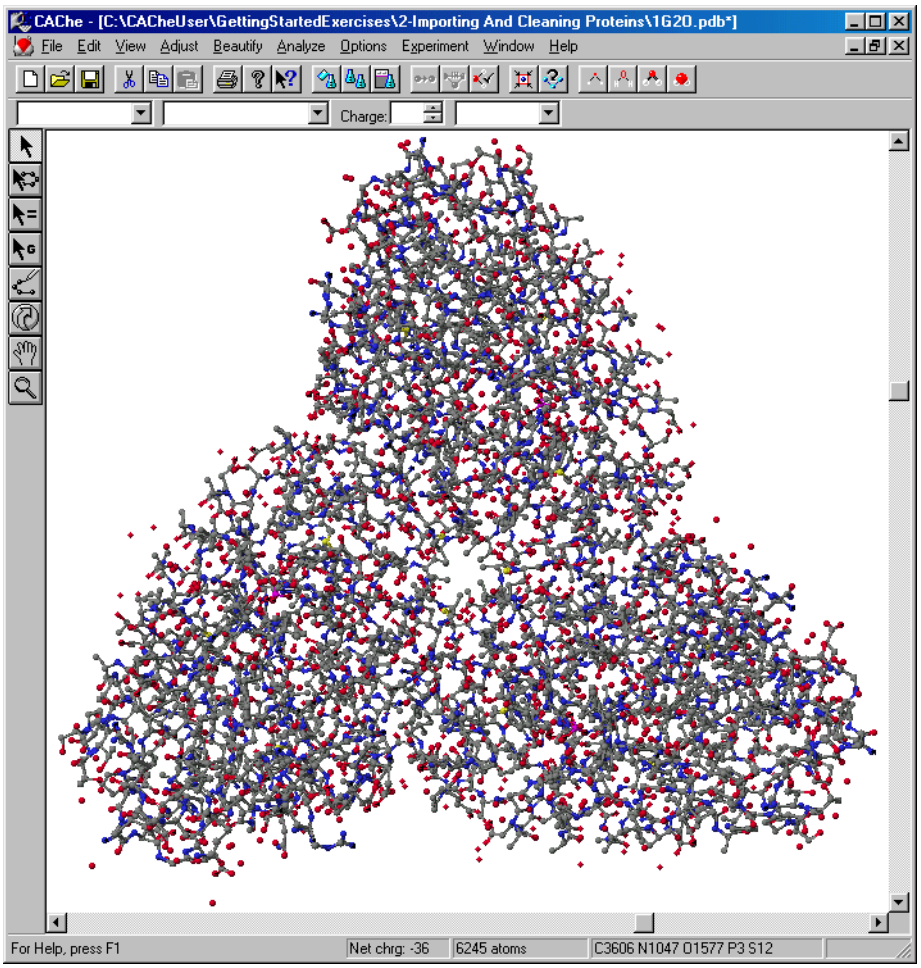
scrolling list to display the folder where the file is located, and double-click on the folder to open it.

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.

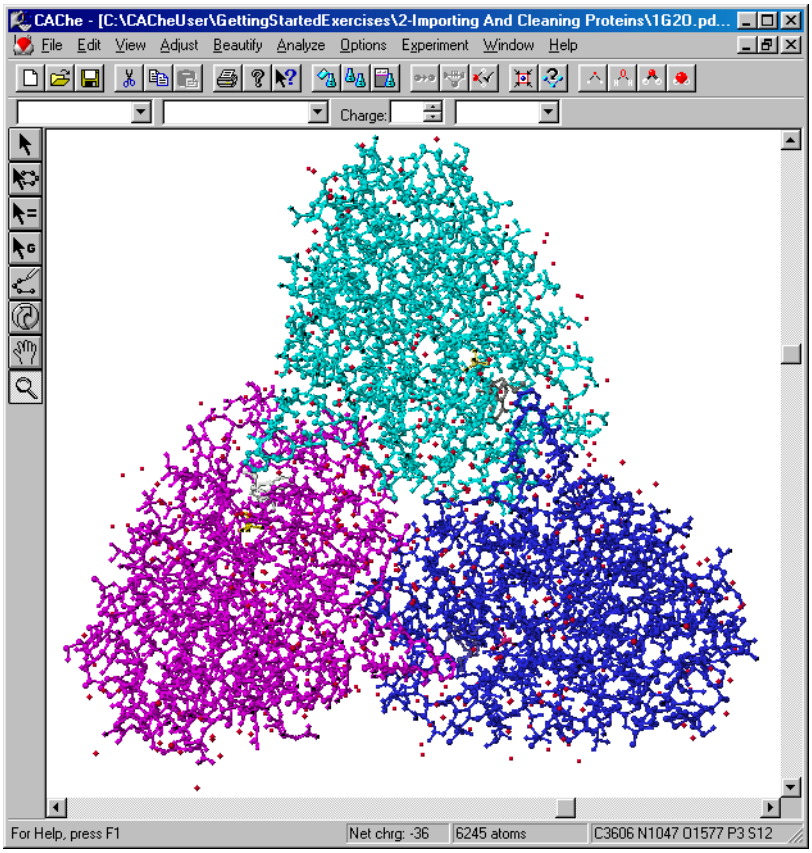


To see clearly the three chains and their ligands in the trimeric structure:

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 1G2O.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**.

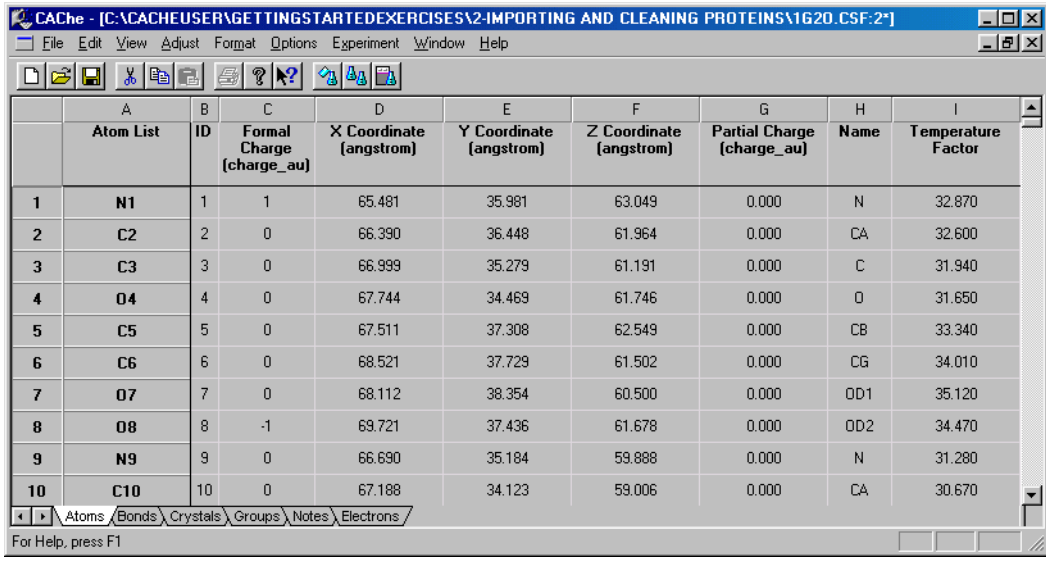
When you save this workspace, you can chose to save it as a chemical sample file (`*.csf`), a PDB file, or other file type. Save the workspace as a chemical sample file (`*.csf`) to preserve all of the information you have added. Otherwise, information such as rendering style and computed atom properties 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.

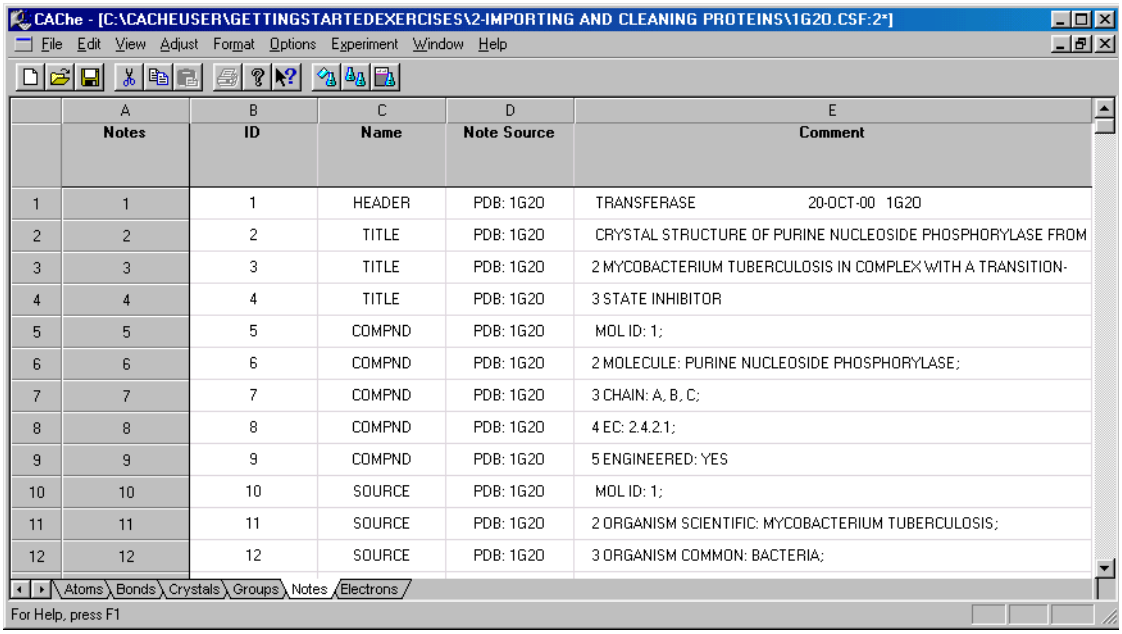


	A Atom List	B ID	C Formal Charge (charge_au)	D X Coordinate (angstrom)	E Y Coordinate (angstrom)	F Z Coordinate (angstrom)	G Partial Charge (charge_au)	H Name	I Temperature Factor
1	N1	1	1	65.481	35.981	63.049	0.000	N	32.870
2	C2	2	0	66.390	36.448	61.964	0.000	CA	32.600
3	C3	3	0	66.999	35.279	61.191	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 informaiton
- **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.

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 Notes	B ID	C Name	D Note Source	E 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

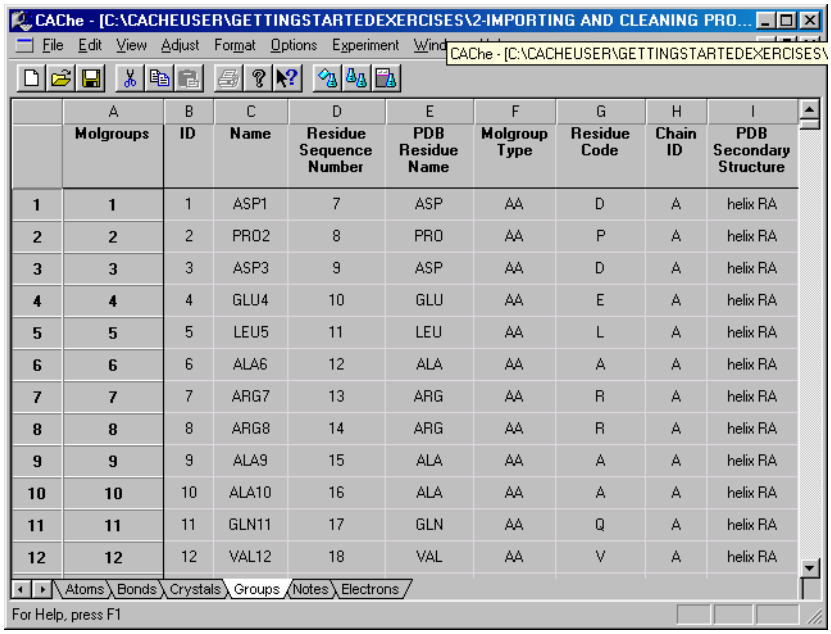
Click the **Comment** column header to highlight the column, then select **Format | Left** to left justify the comments.

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 PO_4 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 PO_4 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 Molgroups	B ID	C Name	D Residue Sequence Number	E PDB Residue Name	F Molgroup Type	G Residue Code	H Chain ID	I 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:

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)

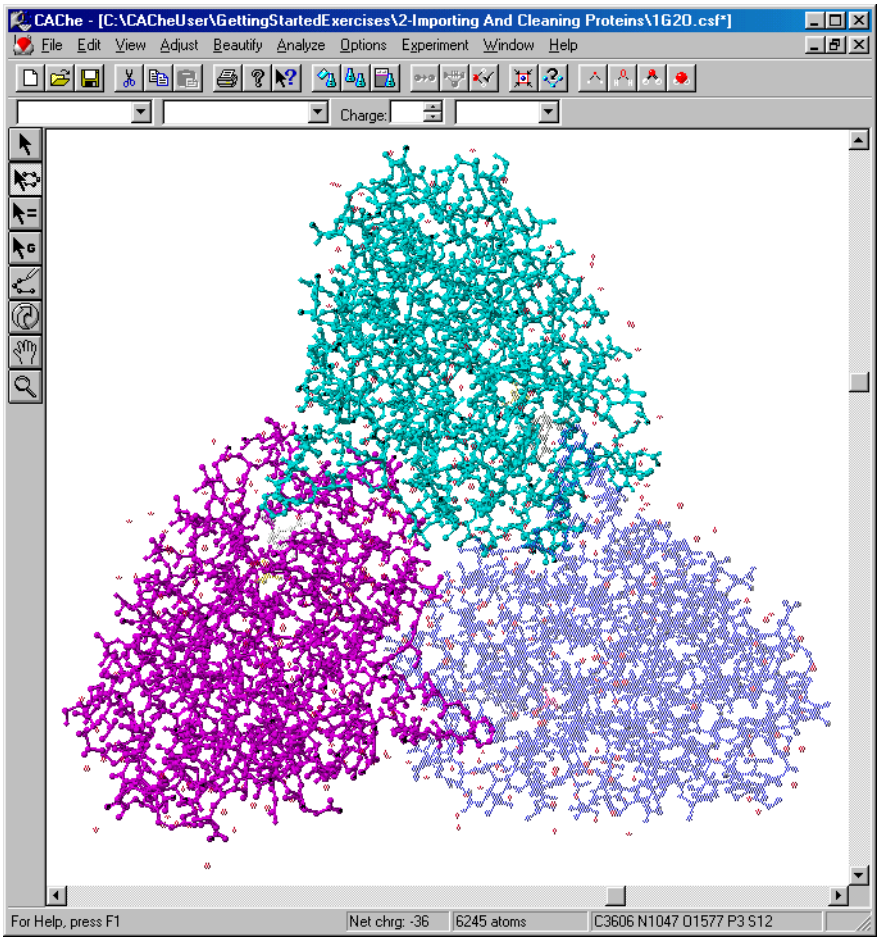
   Row 263 highlights.
3. Scroll until you see the last row containing **Chain ID** C. (Row 786)
4. Shift-click the row number on the left.

   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.

📖 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.

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.

Select Tool

Select Molecule

Select Similar

Select Group

Atom Tool

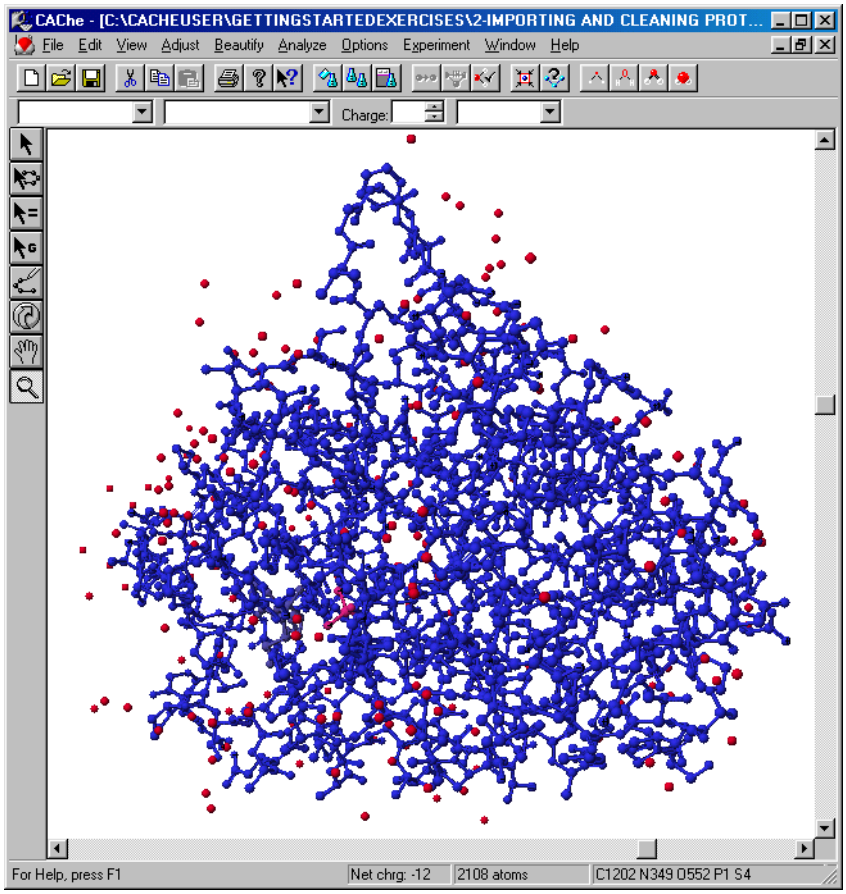
Rotate

Translate

Zoom

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:



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.

📖 NOTE

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

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.

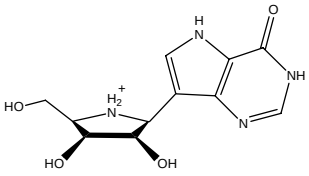
Check and correct the bonding in HET groups

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.



immucillin-H

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.

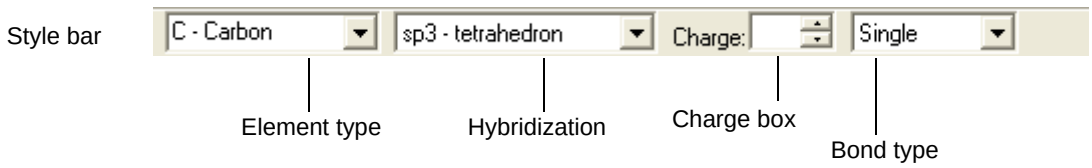
Examine ImmH and change the bonding and charge to agree with immucillin-H structure shown in the figure.

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

📖 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.

type.



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.

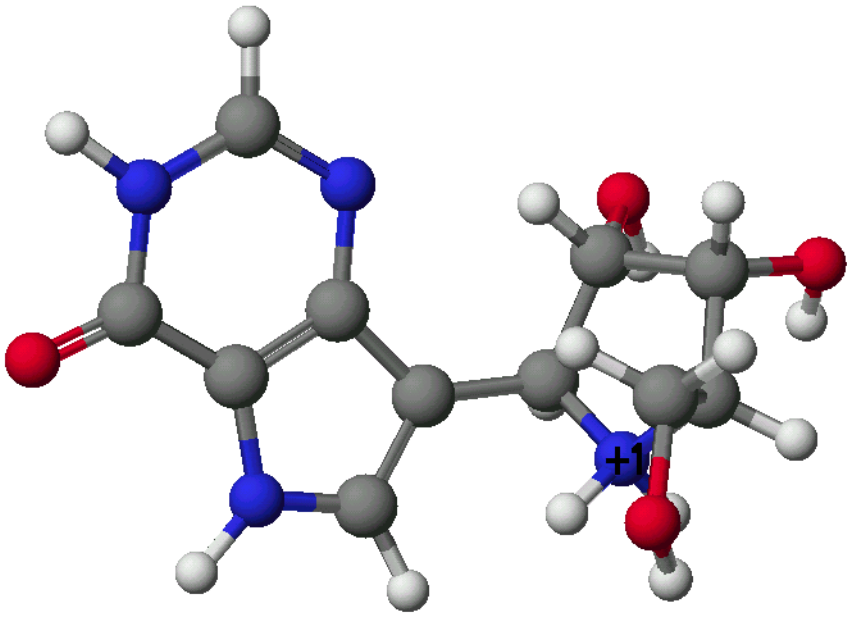
Add hydrogen atoms and define atom hybridization for ImmH

1. Choose the Select Molecule Tool and click on the ImmH ligand.

   ImmH is highlighted.
2. Choose **Beautify | Valence**.

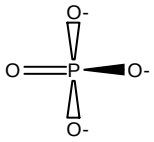
   Hydrogen atoms, electrons and the atom hybridization are added.

ImmH should look like this now



Add hydrogen atoms and define atom hybridization for PO_4

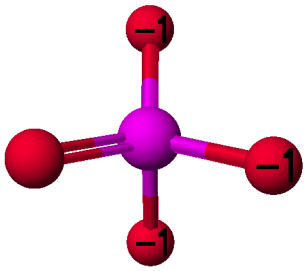
1. Choose **View | Show All**.



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.

📖 NOTE

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

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.

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 neet 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. TTB-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.

	letter code	
Residue	**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

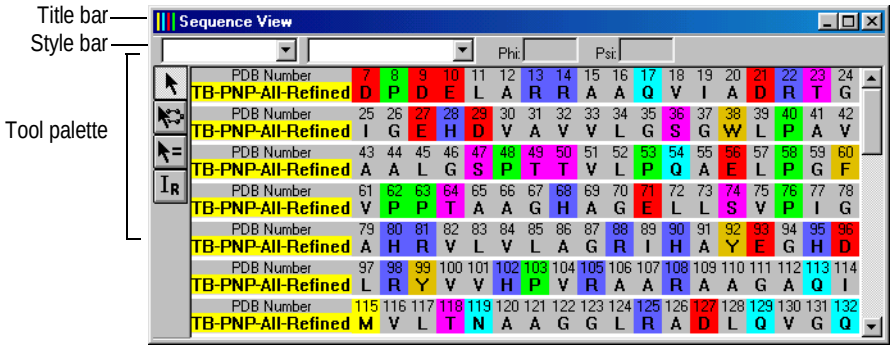
To view the sequence of TB-PNP

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.

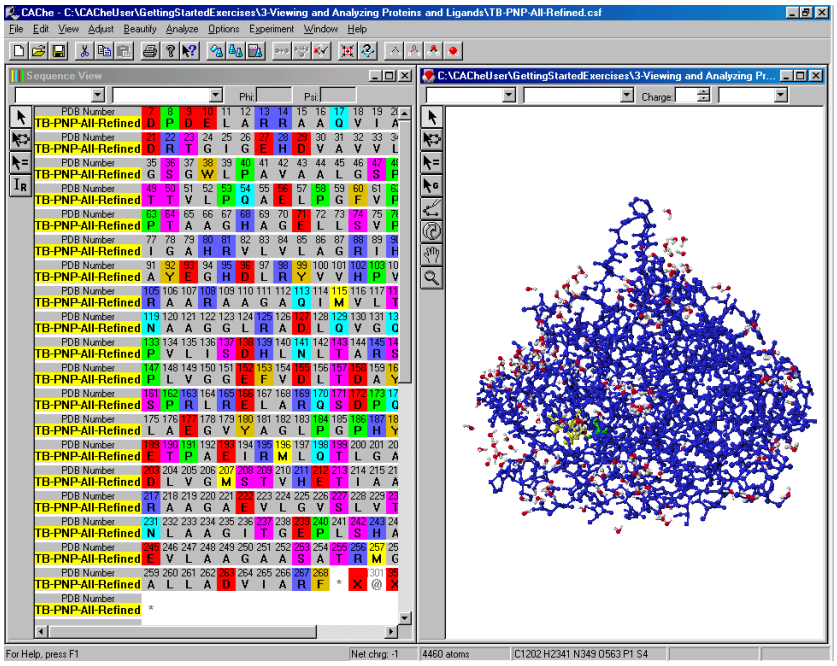
When you view a sequence, the Sequence View window comes to the front and the menus and toolbar change.

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



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.

📖 TIP

o orient your TB-PNP as hown, click in the 3D tructure window to ctivate it, type the letter i'. The protein disappears. he letter command 'i' estores the initial viewing ransform which in this case laced the protein off the creen. Choose **View | Fit in indow** to bring the rotein back into the indow.

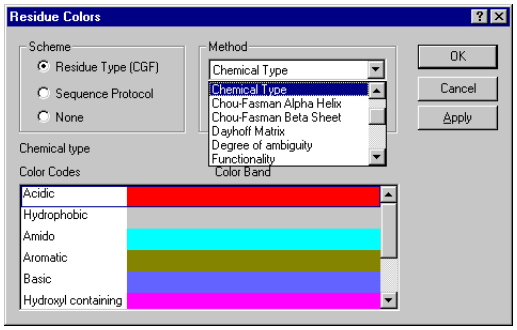
Analyzing the protein sequence

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"*

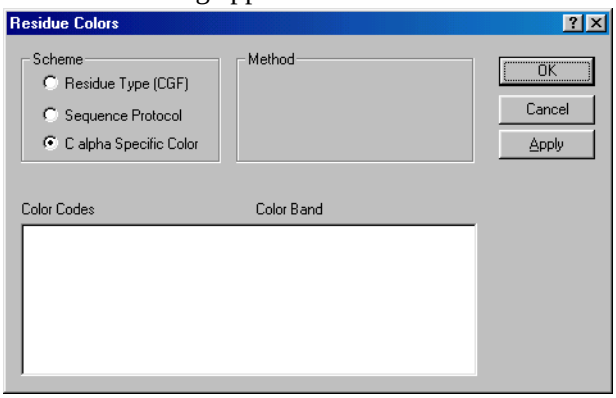
The **Method Window** specifies the number of adjacent residues that are grouped together for analysis. For example, a window of 7 will result in the sequence being analyzed in blocks of 7 residues to determine properties such as the antigenicity or hydropathy. Note that residues at the beginning and 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.

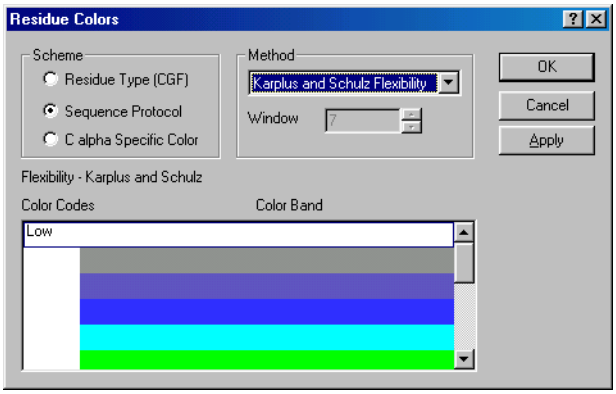


Initially, the colors assigned are the atom specific colors that you have set in the **View | Atom Attributes** dialog.

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

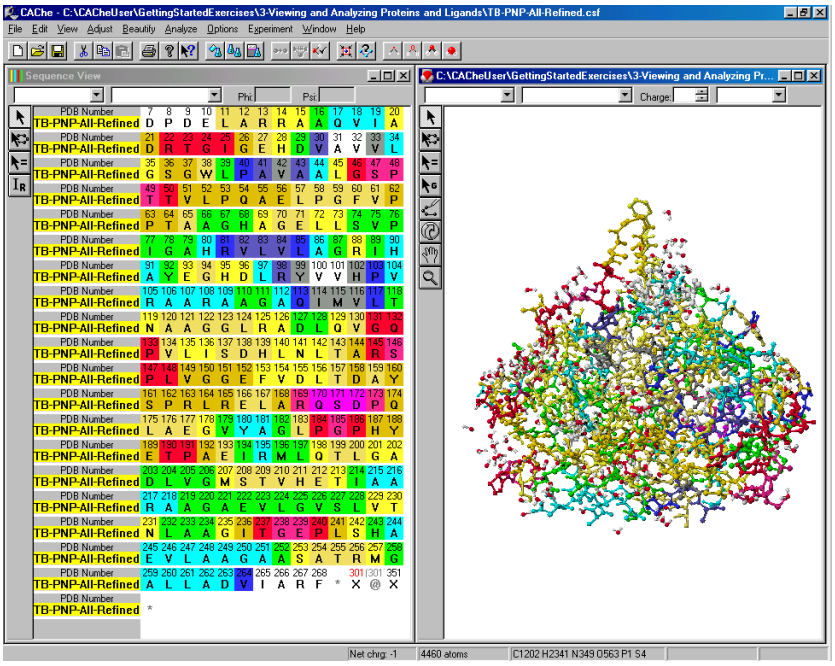
   The list of choices in the **Method** pulldown menu changes.

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..

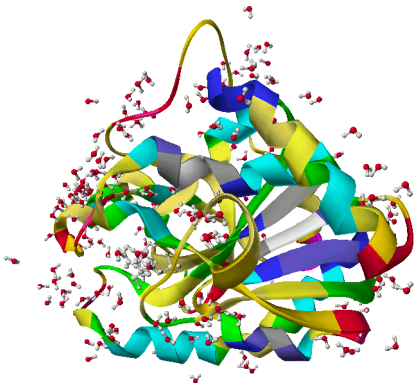


Simplifying the protein structure: ribbons

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

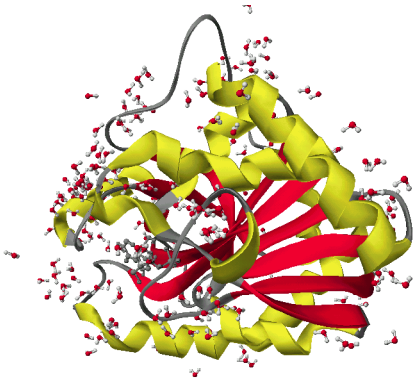
The Residue Colors dialog appears.

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.

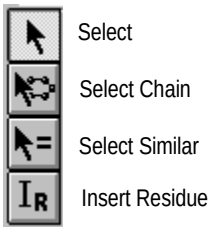
2. In the 3D Structure window, click the background with the Select Tool.

   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**.

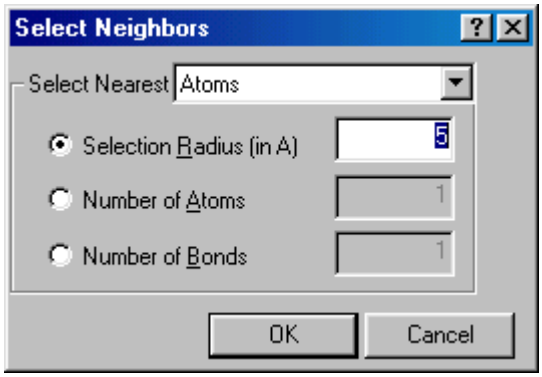
   The bound ImmH ligand highlights and the rest of the sample becomes grey. The sequence view provides fast access to structures by residue and group.

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 moleucles 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

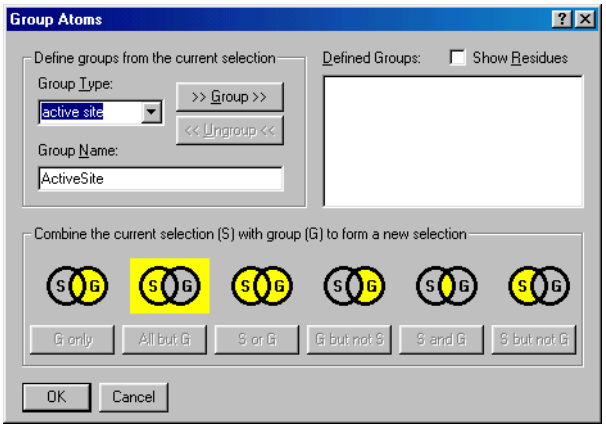
1. Move into the 3D Structure window by clicking its title bar, choose **Edit | Group Atoms**.

   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.

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.

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 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**.

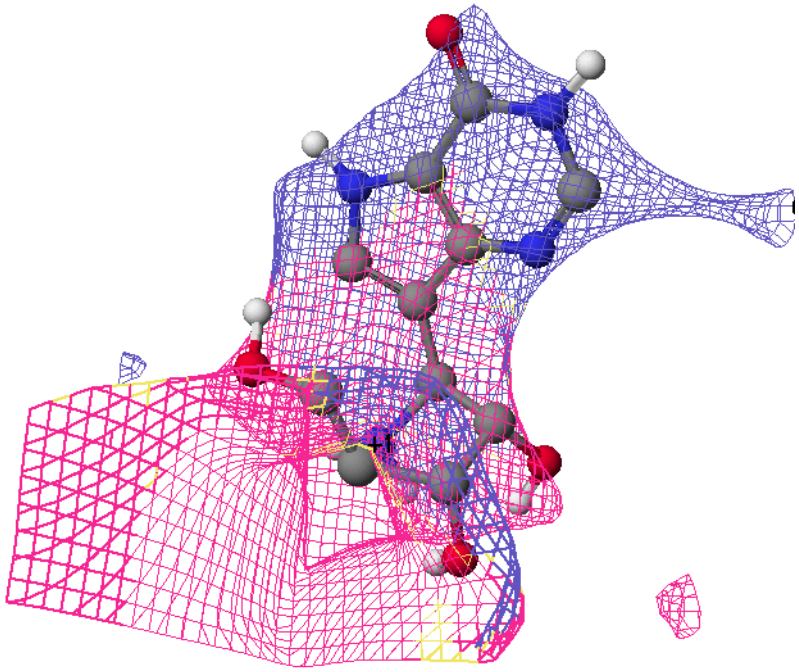
   All atoms and bonds except ImmH disappear. The adjacent surface[1] and ImmH remain visible.

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

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

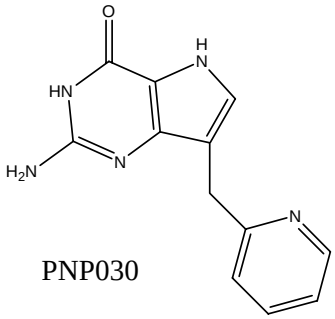
📖 NOTE

"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]

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]



PNP030

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**

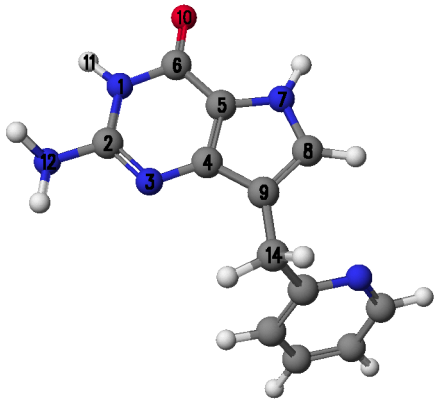
   The Group Atoms dialog closes.
4. Choose **View | Hide Unselected**.

   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 >>**.

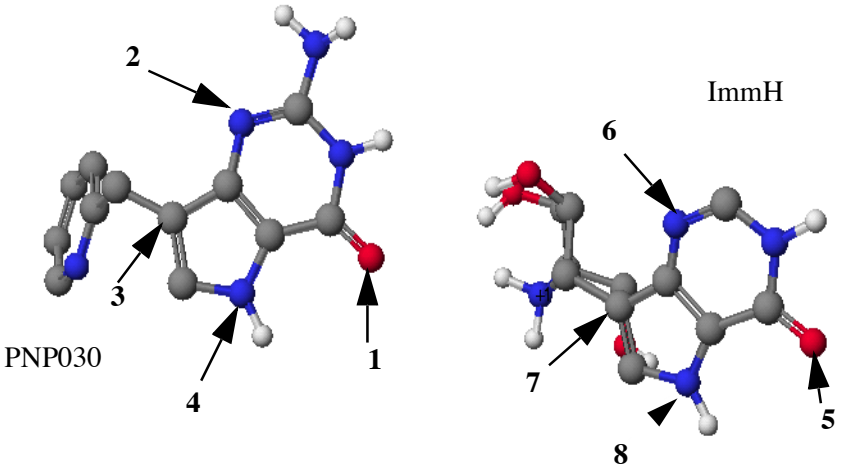
    PNP030 is added to the **Defined Groups** list.

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



and ImmH is on the right.

NOTE

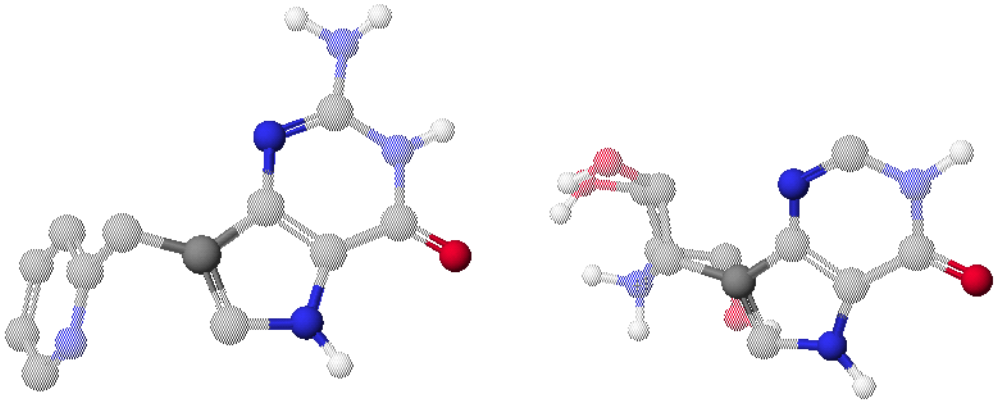
When superimposing molecules, the order of selection is important because, the first atom selected in the first molecule is superimposed on the first atom selected in the second molecule, etc.

16. 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.

17. 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.

18. 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.

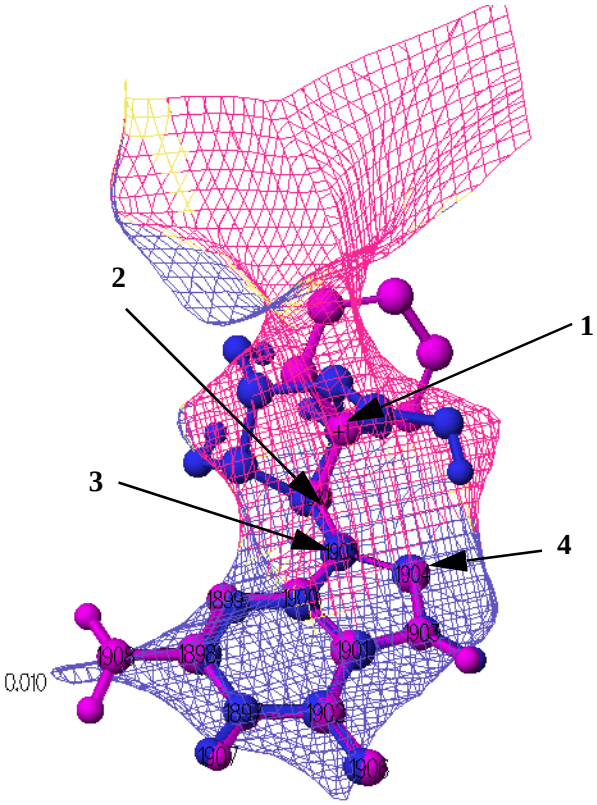
3. Choose **OK**.

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 $-NH_2$ 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 $-NH_2$ 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-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, 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

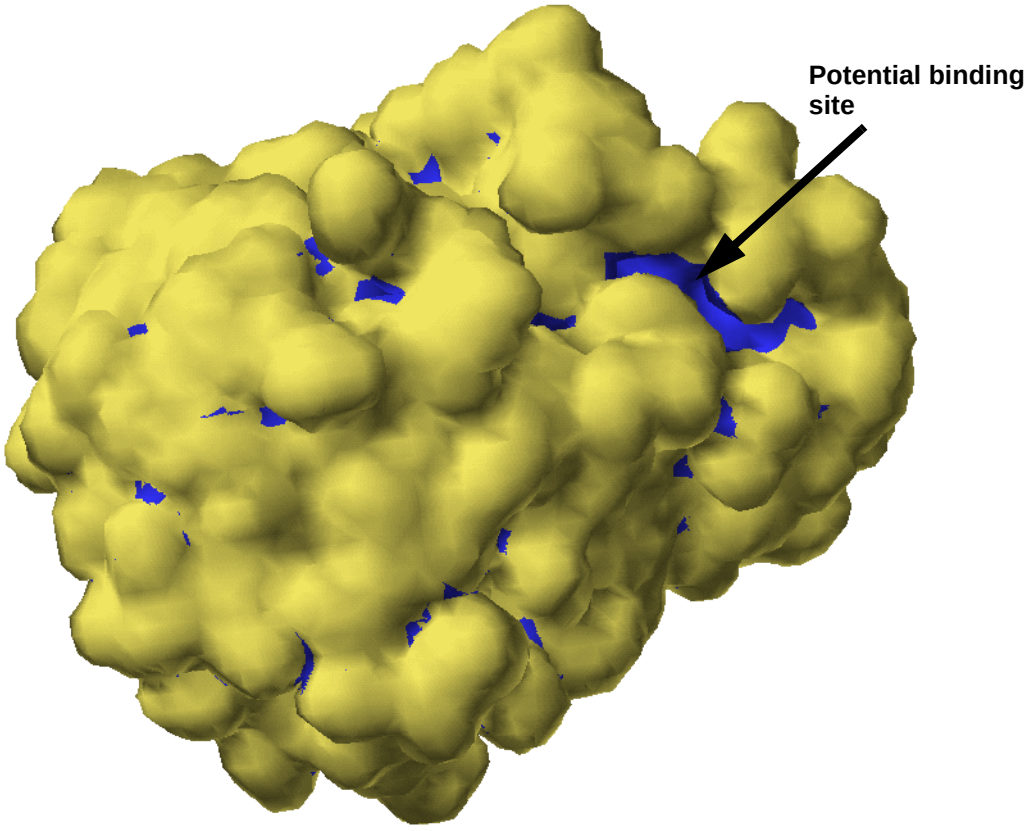
To create and view the crevice surface

1. From the CACheWorkspace, 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 BLOSUM50[1] substitution matrix in the Needleham-Wunsch alignment algorithm[2].

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

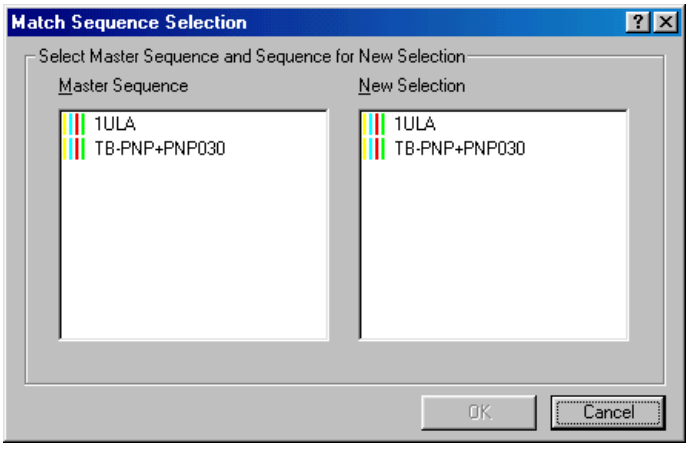
   The active site residues for TB-PNP+PNP030 are highlighted and outlined with black boxes.

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.

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 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

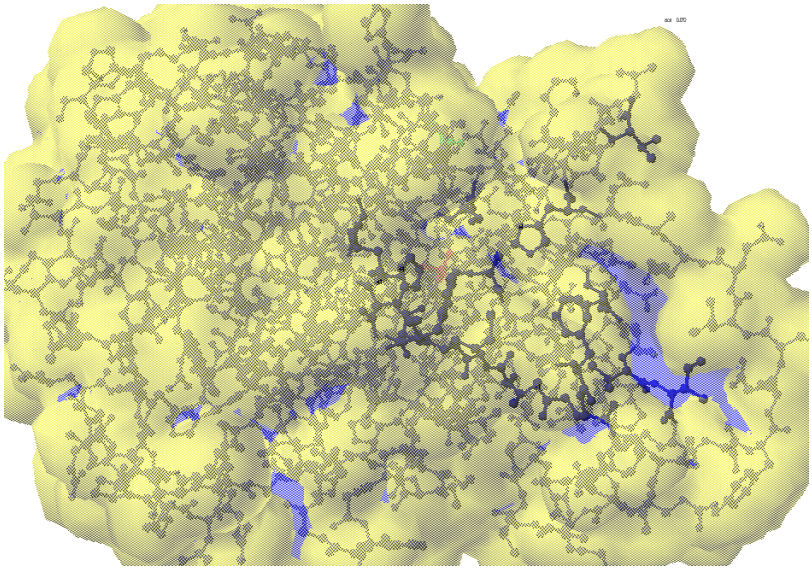
1. From the 1ULA 3D Structure window, choose **Analyze | Show Surfaces**.

   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.

6 Automated Docking and Scoring

Overview

📖 NOTE

To work this exercise, you must have the CAChe ActiveSite add-on to BioCAChe, BioMedCAChe or CAChe WorkSystem Pro. Automated docking is only available with the ActiveSite add-on.

This tutorial describes how to automatically dock and score a ligand in an active site from the Workspace using a potential of mean force (PMF) with a genetic docking algorithm. First, you will validate the docking method by redocking immucillin into its crystal structure in TB-PNP. Then you will dock PNP030 into TB-PNP and compare its score to immucillin.

In this section you will:

- create a copy of immucillin (ImmH) in the TB-PNP chemical sample file
- automatically dock the immucillin copy into the TB-PNP active site
- calculate the RMS distance difference between the docked structure and the original x-ray crystal structure
- dock PNP030 into the active site and compare its score with that of immucillin

Background

CAChe automates the docking of ligands into active sites by using a genetic algorithm with a fast, simplified potential of mean force (PMF).[1]

The potential of mean force is a knowledge-based approach that extracts pairwise atomic potentials from structure information of known protein-ligand complexes contained in the Protein Data Bank. PMF has been demonstrated to show a significant correlation between experimental binding affinities and its computed score for diverse protein-ligand complexes.[2]

Similar to empirical force-fields use in Mechanics and Dynamics, PMF uses atom types. PMF has 34 ligand atom types and 16 protein atom types resulting in 544 unique pair potentials. CAChe represents each potential as a numerical table at 0.2 Å intervals in the `PMF.prm` data file. Docking a ligand into an active site is sensible when both contain only PMF atom types. For ease of use, when CAChe encounters an atom that is not a standard PMF atom type, it assigns it the atom type PMFnone and uses the parameters for aliphatic carbon (PMF atom type CF).

The numerical pair potentials contain information for all ligand-protein atom pair distances, including the short distances that are typical of covalent bonding. Docking models used in CAChe assume that the protein and ligand dock non covalently. It is assumed that bonds are not formed between the ligand and protein because bond formation would cause changes in the atom types and possible substantial changes in the shape of the ligand and active site. Therefore, non bonded terms are added to the potential to keep atoms at typical non bonded distances. The standard PMF implementation uses Amber[3] van der Waals potentials for this purpose. The standard CAChe implementation uses specific 6-12 Lennard-Jones potentials for each pairwise interaction.[4] CAChe's specific pairwise potentials permit hydrogen bonding and electrostatic interactions which are sometimes prohibited by the van der Waals potentials.

PMF offers these advantages: (1) unlike mechanics based methods, there is no need to add hydrogen atoms to the protein or to worry about the charges in the

1. Muegge, I., and Martin, Y., "A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach", *J. Med. Chem.*, **1999**, *42*, 791-804.
2. Muegge, I., "The Effect of Small changes in Protein Structure on Predicted Binding Modes of Known Inhibitors of Influenza Virus Neuraminidase: PMF-scoring in Dock4", *Med. Chem. Res.*, **1999**, *9*, 490-500; Meugge, I., Martin, Y., Hajduk, P. J., and Fesik, S. W., "Evaluation of PMF Scoring in Docking Weak Ligands to the FK506 Binding Protein", *J. Med. Chem.*, **1999**, *42*, 2498-2503; Ha, S., Andreani, R., Robbins, A., and Meugge, I., "Evaluation of docking/scoring approaches: A comparative study based on MMP3 inhibitors", *J. Comp.-Aided Mol. Design*, **2000**, *14*, 435-448; Muegge, I. and Rarey, M., "Small Molecule Docking and Scoring", *Rev. in Comp. Chem.*, **2001**, *17*, 1-60.
3. Cornell, W.D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, Jr., K.M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P.A., "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules", *J. Am. Chem. Soc.*, **1995**, *117*, 5179-5197.
4. U. S. patent pending.

protein; (2) it is based on experimental data and can be refined as more data becomes available; (3) unlike molecular mechanics, there is no partitioning of the energy into many large and mostly cancelling terms e.g. van der Waals and electrostatics; (4) it is fast and simple; (5) potentials implicitly include solvent and entropic effects because they are derived from crystal data[1]; and (6) no knowledge of the binding affinities is required to derive the potentials.

Genetic algorithms (GA) are a standard mathematical technique for optimizing functions of many variables. They are inspired by evolution. These algorithms encode a possible docking solution in a simple sequence of numbers called a chromosome. Here, the chromosome for a flexible ligand contains translation, rotation and dihedral angle numbers that describe the position, orientation and conformation of the ligand in the active site.

The genetic algorithm begins by creating a random population of chromosomes (i.e. random positions, orientations and conformations of the ligand) and then applying recombination operators to these chromosomes so as to preserve those which are the best or most fit. It preserves the best solutions by evaluating the PMF score of each chromosome and allocating reproductive opportunities in such a way that those chromosomes which have a lower score are given more chances to "reproduce" than those chromosomes which are poorer solutions. The process continues for a specified number of generations (usually 1,000 - 40,000).

The use of a Lamarckian genetic algorithm for ligand-protein docking was pioneered by Art Olson's group at Scripps in the program AutoDock.[2]

CAChe uses the GA for docking because it offers the advantage of relatively unbiased searching and it does not require derivatives of the potential. In CAChe, the docking is embedded in the FastDock compute engine. Similar to AutoDock, FastDock uses a Lamarckian GA in which a minimization is performed so that individuals adapt to the environment in which they are placed.

The goal of this exercises is to teach you how to use BioMedCAChe to design your own selective inhibitors by docking flexible ligands into a rigid protein active site.

Of course, proteins are not rigid and often flex to accommodate ligands. You can overcome this limitation by allowing the protein active site to flex during docking, but this increases the time and complexity required for the calculations dramatically. It also complicates the analysis and interpretation of the structure-activity relationships. With CAChe you can apply either strategy; but in this exercise, we will focus only on the fast one.

---

1. Protein crystals typically contain approximately 50% water.
2. Morris, G. M., Goodsell, D. S., Halliday, R.S., Huey, R., Hart, W. E., Belew, R. K. and Olson, A. J., "Automated Docking Using a Lamarckian Genetic Algorithm and Empirical Binding Free Energy Function", *J. Computational Chemistry*, **1998**, *19*, 1639-1662.

Validating the PMF method

Before using any new scientific method, you should first validate that the method works in situations where you know the result. Here we will dock immucillin-H into TB-PNP and compare the docked result to the crystal structure of the bound ligand-protein complex.

To dock a copy of immucillin into TB-PNP

1. Navigate to the folder `6-Automated Docking`, 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 **Edit | Copy**.

   A copy of ImmH is placed on the clipboard.

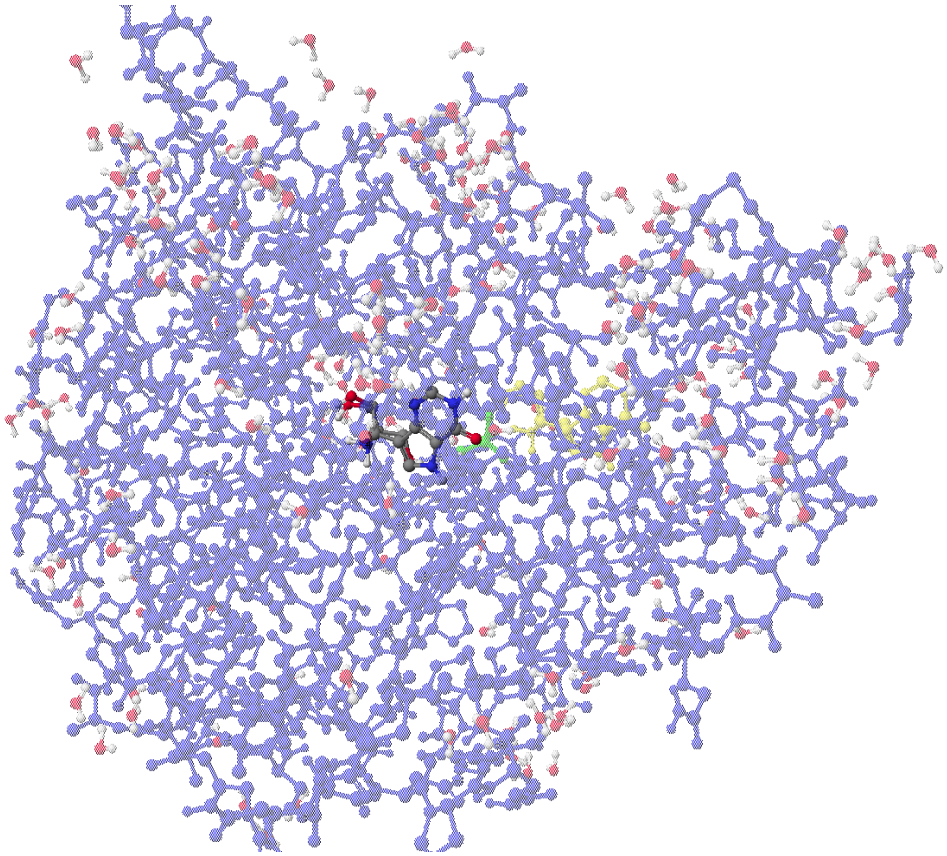
5. Choose **Edit | Paste**.

   A copy of ImmH is placed in the 3D-structure window.

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

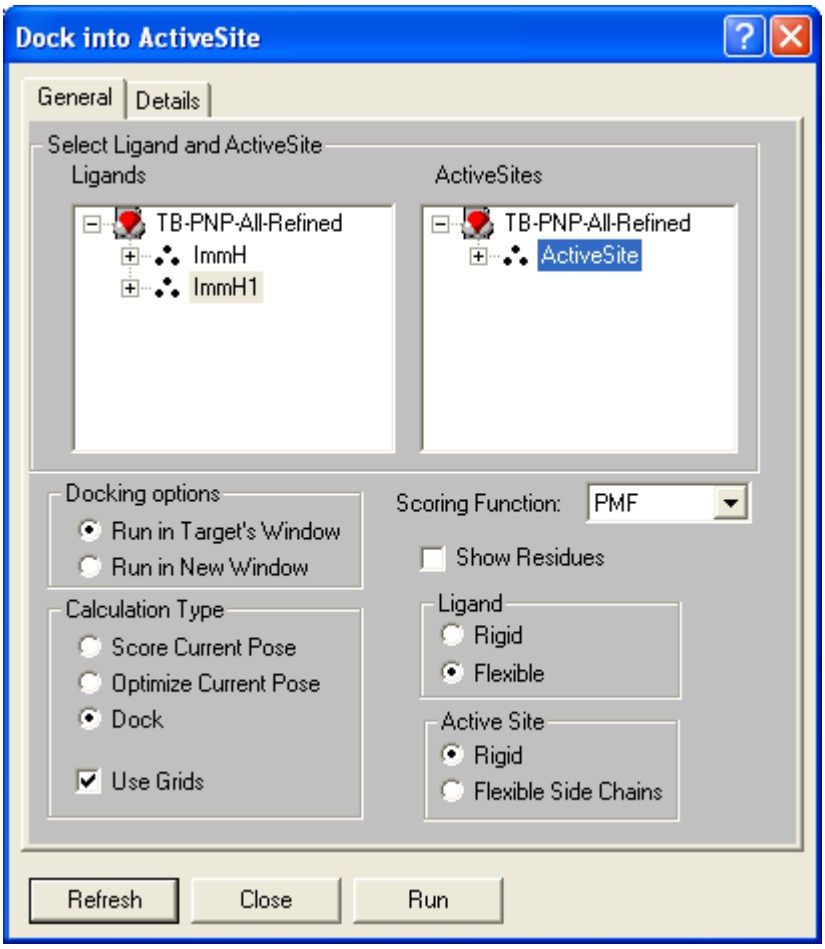
   The copy of ImmH is colored by element so that it is easy to see which is the copy and which is the original ImmH.

Atoms and bonds in ImmH are colored according to element type.



7. Choose **Analyze | Dock into Active Site**.

The Dock into Active Site dialog appears.



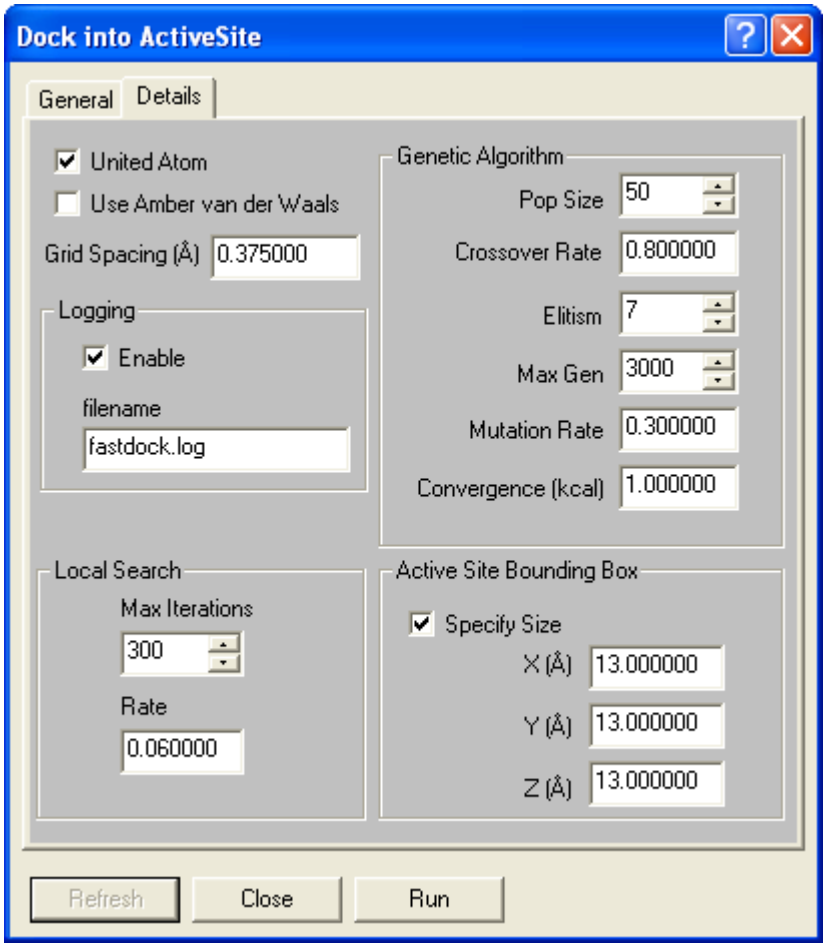
The list on the left contains the names of the ligand type atom groups in each open 3D structure window. When you pasted ImmH, CACheautomatically gave the copy the new group name `ImmH1`. The list on the right contains the names of the active site groups in each open 3D structure window.

8. Set your dialog as shown above.

   These settings will dock the flexible ligand `ImmH1` into the rigid active site group `ActiveSite` in the protein `TB-PNP-All-Refined` in the `Target's Window` using the fast grid-based method.

   When docking a ligand into a target, all other groups of group type **ligand** in that target are ignored. For example, in this case, `ImmH` will be ignored when you dock `ImmH1`. This allows you to have many ligands overlaid in an active site.

9. Click the `Details` tab and set it as shown here.



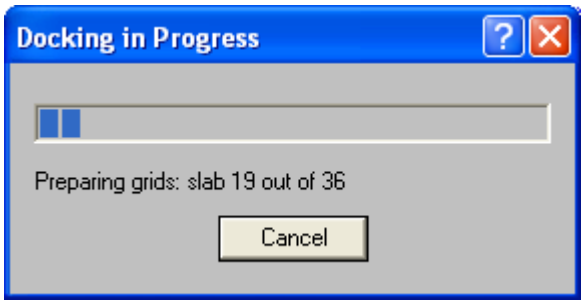
📖 NOTE

If your active site is an open site on the surface of the protein, you may need to add water molecules to simulate the water outside of the protein so that the ligand is bounded on all sides.

These settings are for docking a ligand into a 13x13x13 Å boxl located at the center of the target active site using a united atom (explicit hydrogen atoms are not considered) potential of mean force with a docking algorithm that has a population of 50 chromosomes and runs for 3,000 generations.

10. Choose **Run** and then **Close**.

The **Docking in Progress** dialog appears. It displays the status of the

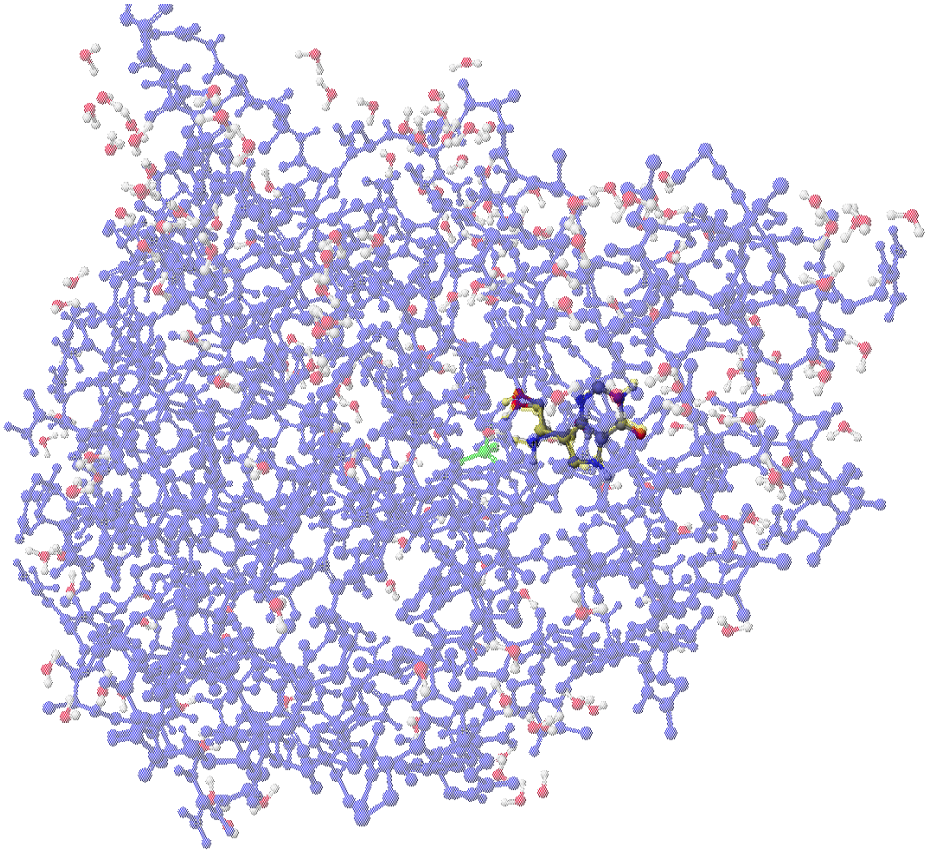


📖 NOTE

Wait until the docking is complete before making changes to your structure.

docking calculation. Initially, the protein potential for interaction with each atom type is evaluated on a grid of 0.375 Å in the 13x13x13 Å box. After that, the best score at every 100th generation is reported and the position of the ligand in the 3D structure window is updated..

At the 3,000th generation, the progress dialog closes and you see the element-colored ImmH1 docked in the active site as shown below.



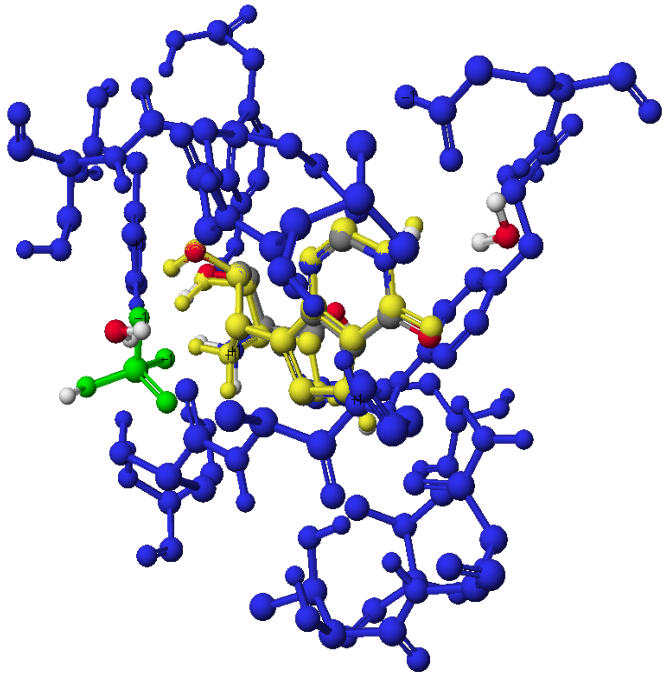
11. Choose **Edit | Group Atoms**, select `ActiveSite` from the list of **Defined Groups**, choose **S or G**. Then select `ImmH` from the list of **Defined Groups**, choose **S or G.** Finally, Choose **OK.**

    The Group Atoms dialog closes and the active site, ImmH and its docked copy are highlighted in the 3D structure window.

12. Choose **View | Hide Unselected** and then **View | Fit in Window**.

    The transparent portion of the protein disappears and the active site and

ligands zoom to fill the center of the 3D structure window so you see this.



To view the final score

1. Choose **Analyze | Chemical Properties Spreadsheet**, then **Window | Tile**.
2. Next with the scroll button in the lower left hand corner, scroll the workbook tabs until the Calculations tab is visible and click it to bring it to the front.
3. Finally, with the scroll bar at the bottom of the window, scroll until you see the column **Docking Score**.

   The final docking score in kcal/mole for each docking experiment is displayed in this column. Your score should be close to -96 but because

	A	F	G
	Calculation History	Date/Time of Modification	Docking Score (kcal/mole)
1	1	Thu Aug 28 23:41:12 2003	-96.180

   the genetic algorithm is a random search algorithm, your score may be a few kcal/mole different. As with all calculation entries, the **Date/Time of Modification** column displays the timestamp for the entry into the chemical sample and the **Comment** column displays the calculation settings from the control panel.

   The final docking score differs from the score reported in the progress

📖 NOTE

The PMF docking score can be related to the binding affinity, but it is not a binding free energy and is usually many times larger than the binding free energy. The next exercise shows how to discover a relationship between PMF score and binding affinity.

dialog, because van der Waals terms for the ligand are added to the score during the docking process. These terms are not included during the final scoring.

Keep both windows open while you perform the remainder of this chapter.

To measure the RMS distance error

1. Click the title bar of the 3D structure window to bring it to the front, then choose **View | Atom Attributes**, and finally choose the **Label** tab.

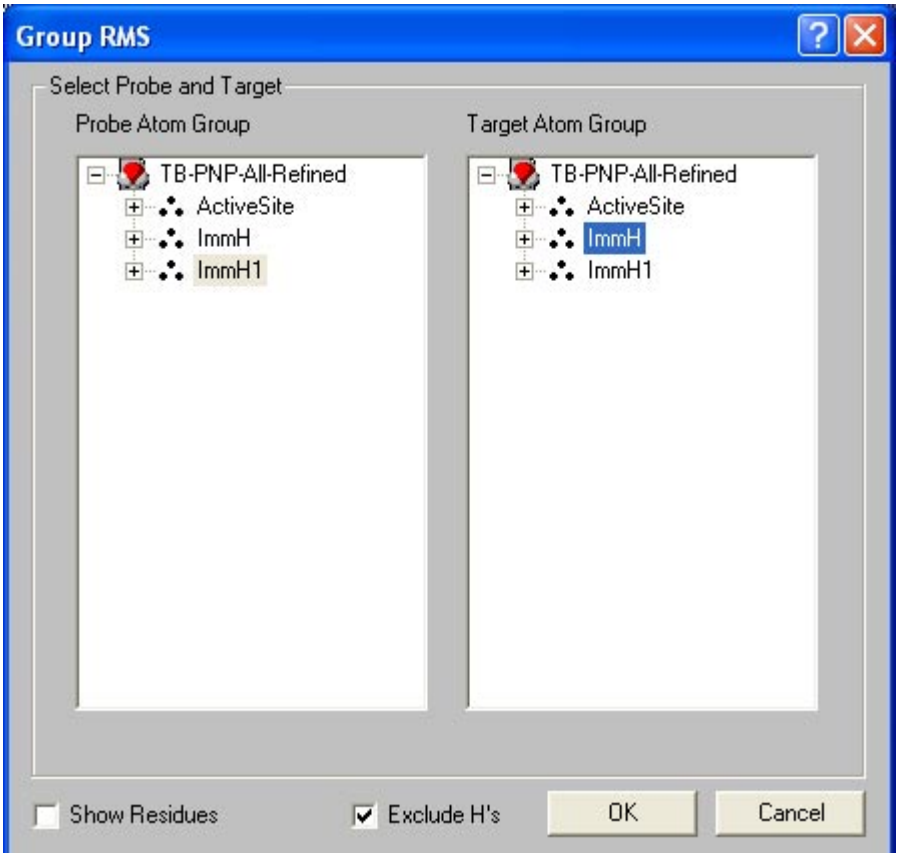
   The Atom Attributes dialog opens and the label tab is at the front.

2. Check **Name** and choose **OK**.

   The dialog closes and the PDB name for each selected atom appears on the atom.

3. Click the title bar of the 3D structure Window to bring it to the front and then choose **Analyze | Group RMS Error**.

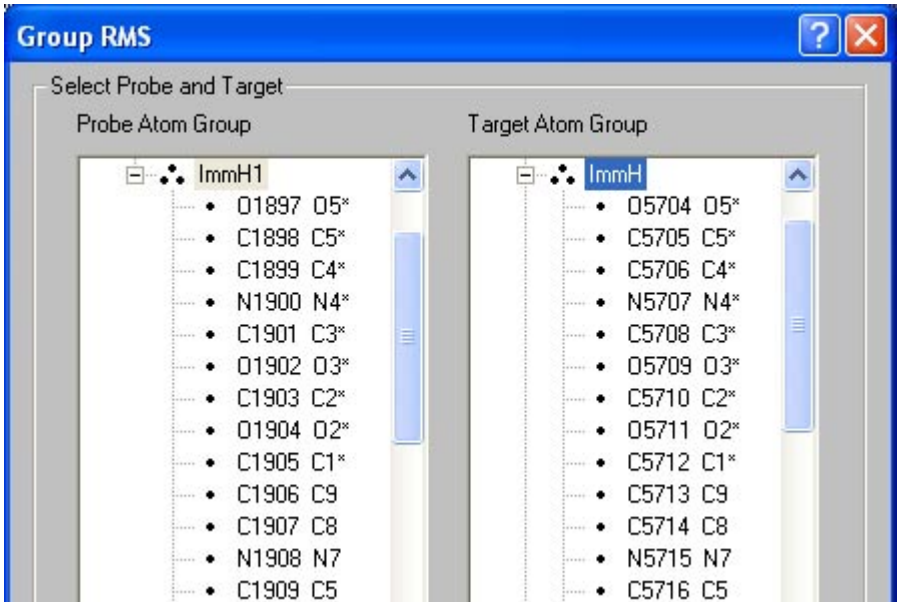
   The **Group RMS** dialog appears. Similar to the **Superimpose Groups**



   dialog, the **Group RMS** dialog lists the groups in each open 3D window.

4. Click the plus next to `ImmH1` in the **Probe Atom Group** and the plus next to `ImmH` in the **Target Atom Group** list.

An ordered list of atoms appears below each ligand. The list identifies



each atom by element type (e.g. O), unique ID number (e.g. 1897) and by Name (e.g. O5*). When the RMS distance is calculated, the distance between each pair of probe and target atoms is calculated in the order defined in the list. That is, the first atom in the probe group (O1897 O5*) is paired with the first atom in the target group (O5704 O5*), the second with the second and so forth.

📖 NOTE

You change the order of atoms in their list by dragging and dropping the atom on the atom it follows.

5. Using the atom names, compare the atom pairs with their positions in the structures. You should see, for example, that C5 in ImmH1 is closest to C5 in ImmH. ImmH1 docks on ImmH very well with the exception of O5.

📖 NOTE

As we will see in the next exercise, our model of the PNP active site is incomplete. There is a phenylalanine from the adjacent chain that caps the site. When this residue is kept in the active site model, O5 in the docked ImmH1 falls on top of O5 in ImmH and the RMS error decreases.

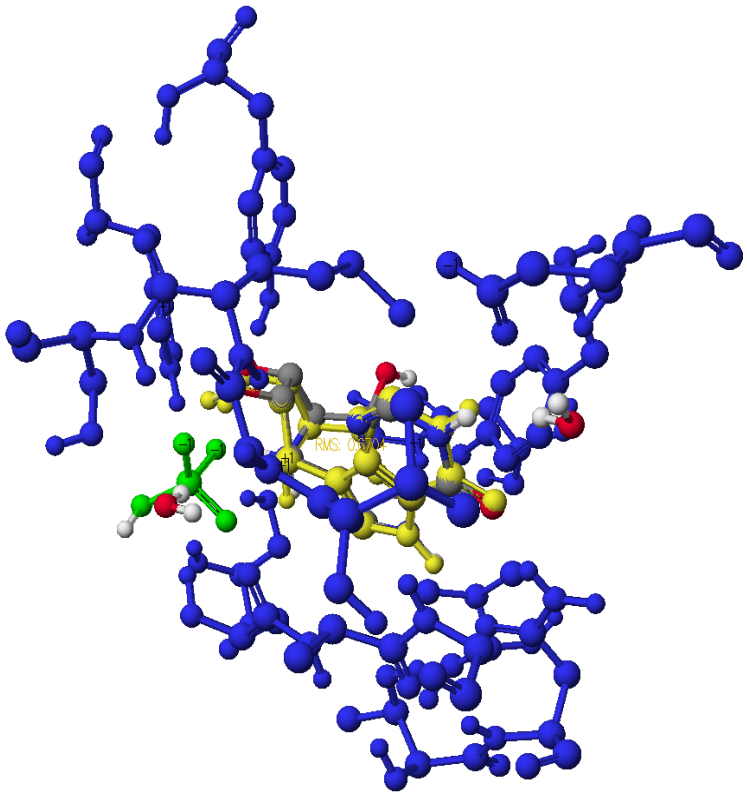
6. Choose `ImmH1` from the **Probe Atom Group** list and `ImmH` from the **Target Atom Group**.

   `ImmH1`, `ImmH` and the **OK** button highlight. You exclude the hydrogen atoms because they were not present in the original crystal structure.

7. Finally, check **Exclued H's** and choose **OK**.

   A RMS distance label with a value of 0.7 appears centered between the

ImmH and ImmH1 ligands. In addition, a new tab for RMS labels is



created in the chemical sample properties workbook.

8. From the Workbook window, choose the **RMS measurements** tab.

	A	B	C	D	E	F
	RMS measurements	**ID**	**X Coordinate (angstrom)**	**Y Coordinate (angstrom)**	**Z Coordinate (angstrom)**	**Distance RMS (angstrom)**
1	**1**	1	58.039	14.780	41.832	0.670

The RMS measurements sheet comes to the front. The root mean square difference in positions between atoms in the measured groups appears in the column **Distance RMS**. The x-, y-, and z-coordinates are the position of the RMS label in the 3D window.

9. From the 3D structure window, choose the **Select Groups** tool from the tool palette and click the center of the RMS label.

   The RMS label and the atoms measured highlight. This is a useful way to determine which atoms are participating in the measurement.

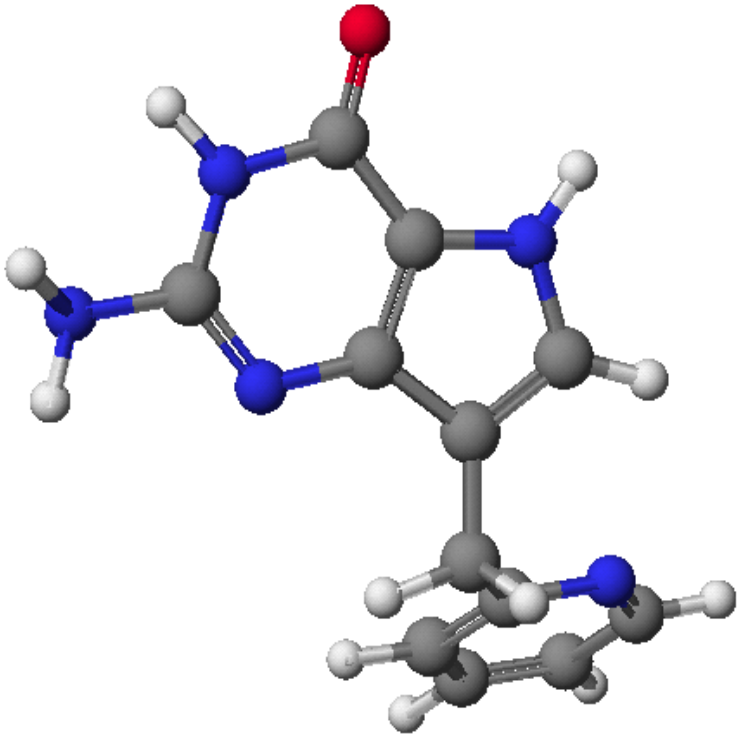
Docking a novel ligand into TB-PNP

The result of re-docking ImmH into the active site of the protein crystal structure in which it was originally bound are in good agreement. To fully validate the docking implementation, you would need to perform dozens of similar docking experiments of other ligands into their ligand-protein crystal structures, as were done in developing the CAChe software. Based on this encouraging single validation, we move to docking another ligand into this active site.

To dock PNP030 into TB-PNP

1. Choose **File | Open** and select `PNP030.csf`.

   PNP030 opens in a new 3D sample window



2. Choose **Edit | Group Atoms**.

   The Group Atoms dialog appears.

3. Choose ligand from the **Group Type** pull down list, type PNP030 into **Group Name** text box, and finally choose **>>Group>>** button.

   The PNP030 group appears in the **Defined Groups** list.

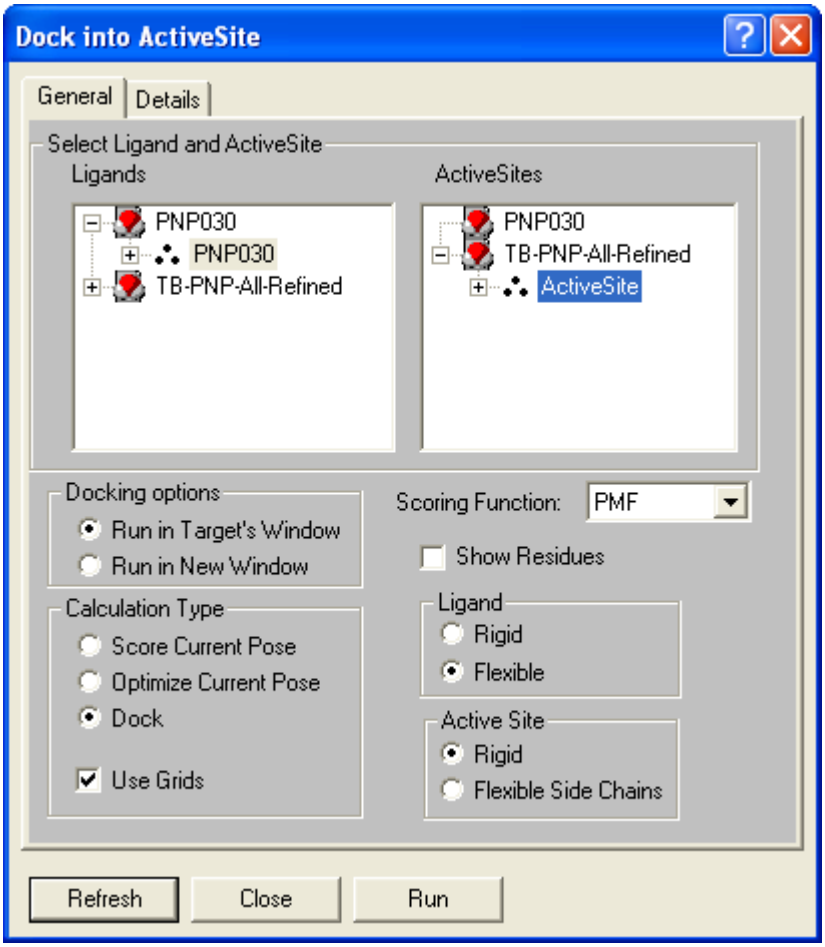
4. Choose **OK**.

   The Group Atoms dialog closes.

5. Choose **Analyze | Dock into Active Site**.

The Dock into Active Site dialog opens.

6. Set the dialog to match this



Note that two different 3D structure windows are listed: PNP030 and TB-PNP-All-Refined. PNP030 does not have an active site type group, but TB-PNP-All-Refined does have one. The dialog is set to dock flexible PNP030 from the PNP030 3D structure window into the rigid group called ActiveSite in the TB-PNP-All-Refined 3D structure window using the fast grid based method

7. Choose **Run** and press **Close**.

   The **Docking Progress** dialog appears and the **Dock into Active Site** dialog disappears. After 3,000 generations, the progress dialog disappears, the chemical sample is updated to the best docked structure and a map file is created containing all docked structures within 100 kcal of the the best docked structure.

8. Close the PNP030 window by clicking the close box in its upper right-hand corner. Do not save the changes.

   The window closes and the tiled workbook and TB-PNP windows are

visible.

9. Click the **Calculations** tab in the workbook and scroll to the **Docking Score** column.

   The final docking score for PNP030 into TB-PNP was -59 kcal/mole. A good score, but not as good as the -96 kcal/mol for immucillin.

10. In the 3D structure window, visually compare the PNP030 and ImmH structures.

    Notice how the inosine rings are oriented the same and the phenyl ring sensibly overlaps the ribose portion of ImmH. The docking of PNP030 satisfies our chemical intuition.

So what?

In this exercise, you demonstrated that CAChe fast docking gives good agreement with one experimental structure. You also demonstrated that it docks a similar inhibitor in a sensible pose. And finally, you showed that the calculated score orders the relative inhibition of PNP030 and ImmH sensibly.

Next, you will learn how to dock a library of ligands into one or more proteins using ProjectLeader. You will also create a relationship between PMF scores and -log(IC50).

7 Discovering Inhibitors By Automatically Scoring and Docking Libraries

Overview

You can use CAChe to screen virtual compound libraries so that you can select those compounds that are most likely to have high binding affinities.

This tutorial describes how to build a model of known PNP inhibitor within the active site of bovine PNP (1A9T) and then rank order a library of compounds by docking the compounds in bovine PNP and scoring their optimal pose.

A useful docking method

- correctly reproduces known crystal structures;
- scores the ligands in order of activity; and
- runs quickly on libraries of compounds.

In this section you will:

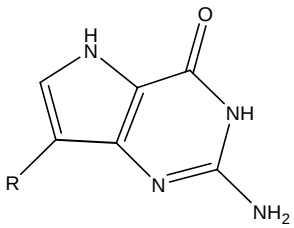
- import 2D structures of a library of substituted 9-deazaguanines, automatically converting it to 3D.
- automatically dock each member of the library into bovine PNP and score the optimal pose.
- examine the docked structures.
- use the scores to obtain a structure activity relationship for compounds in the library and evaluate the relationship of measured activity and docking score.

📖 NOTE

To work this exercise, you must have the CAChe ActiveSite add-on to BioMedCAChe or CAChe WorkSystem Pro. Automated docking is only available with the ActiveSite add-on and ProjectLeader and bio-features are available only in BioMedCAChe or CAChe WorkSystem Pro. In addition, this exercise requires approximately 0.5 GB of free disk space.

Background

One class of compounds that has been extensively characterized as PNP inhibitors is the substituted 9-deazaguanines. Data includes measured binding constants, IC_{50} and EC_{50} values as well as high resolution x-ray crystal structures of substrates, products and inhibitors bound to bovine and human PNP.[1]



9-deazaguanine

In this exercise you will automatically dock a library of 19 compounds into the active site from the bovine PNP crystal structure 1A9T.pdb.[2] Of the several crystal structures available, this crystal structure was chosen because it contained the product substrate that requires the largest pocket. As you will see, the active site model here differs from the single chain active site model you generated for TB-PNP in that it uses the full trimeric chain structure and contains a capping phenylalanine residue from an adjacent chain.

The availability of this experimental data provides a firm foundation for evaluating computational strategies in deriving structured-guided structure activity relationships that are a necessary part of any drug discovery program. This area has been revolutionized by the development of fast, automated methods for obtaining docked ligands inside proteins.

The goal of this chapter is to teach you how to use BioMedCAChe to design your own selective inhibitors by docking a library of flexible ligands into a rigid protein active site. One of the consequences of using a rigid protein is that you must be careful to be sure that the library you screen contains similar sized structures.

Of course, you can overcome this limitation by allowing the protein active site to flex during docking, but this increases the time and complexity required for the calculations dramatically. It also complicates the analysis and interpretation of the structure activity relationships.

With CAChe you can apply either strategy, but here you will focus on docking a flexible ligand into a rigid active site. In previous sections, you have worked one molecule at a time. Here you will use automated methods to work with a library of compounds.

1. Rarutin, V., 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.
   Montgomery, J. A., Niwas, S., Rose, J. D., Secrist III, J. A., Babu, Y. S., Bugg, C. E., Erion, M. D., Guida, W. C., and Ealick, S. E., "Structure-Based Design of Inhibitors of Purine Nucleoside Phosphorylase. 1. 9-(Arylmethyl) Derivatives of 9-Deazaguanine. *J. Med. Chem.*, **1993**, *36*, 55-69.
2. Mao, C., Cook, W.J., Zhou, M., Federov, A.A., Almo, S. C., and Ealick, S.E., "Calf Spleen Purine Nucleoside Phosphorylase Complexed with Substrates and Substrate Analogues", *Biochemistry*, **1998**, *37*, 7135-7146.

Setting up the library

You will often find 2D structures for libraries in chemical data bases. Chemical databases allow you to export the structures and related data in single SD format file. You read these files directly into CAChe ProjectLeader and it automatically converts the structures to 3D, creates chemical sample files, and places the associated activity data into a table.

Importing the SD file

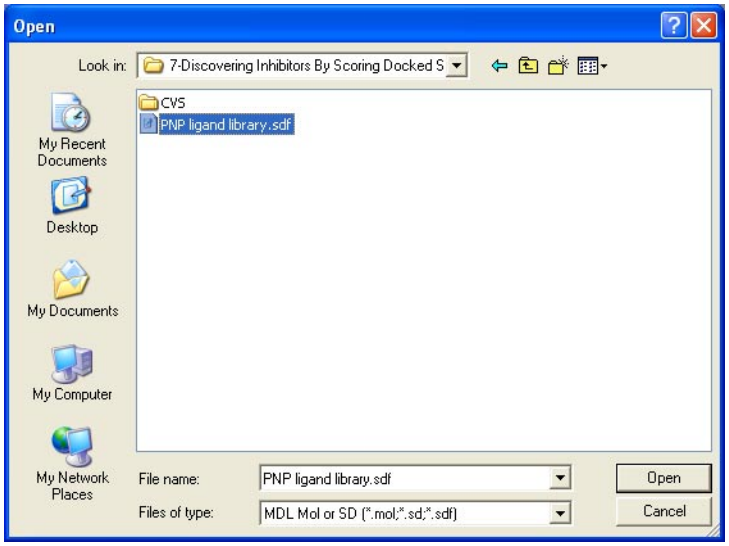
To open the SD file

1. From the Windows Start menu, choose ProjectLeader.

   ProjectLeader opens with a new empty table with a single **Chemical Sample** column.

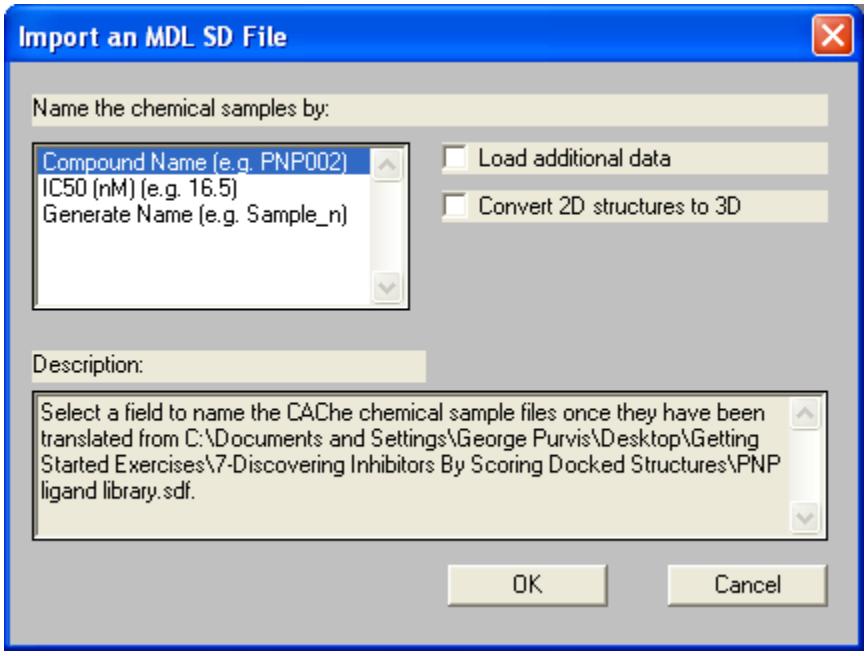
2. Double-click on the empty cell in row 1 of the **Chemical Sample** column.

   The Open dialog appears



3. Browse to the `7-Discovering Inhibitors By Scoring Docked Structures` folder,
4. Set the file type to **MDL Mol or SD (*.mol, *.sd, *.sdf)**.
5. Select the arrow button in the **Files of type** box and choose **MDL Mol or SD (*.mol, *.sd, *.sdf)** from the drop-down list.
6. Choose `PNP ligand library.sdf` and then choose **Open**.

The Open dialog closes and the Import an MDL SD File dialog appears



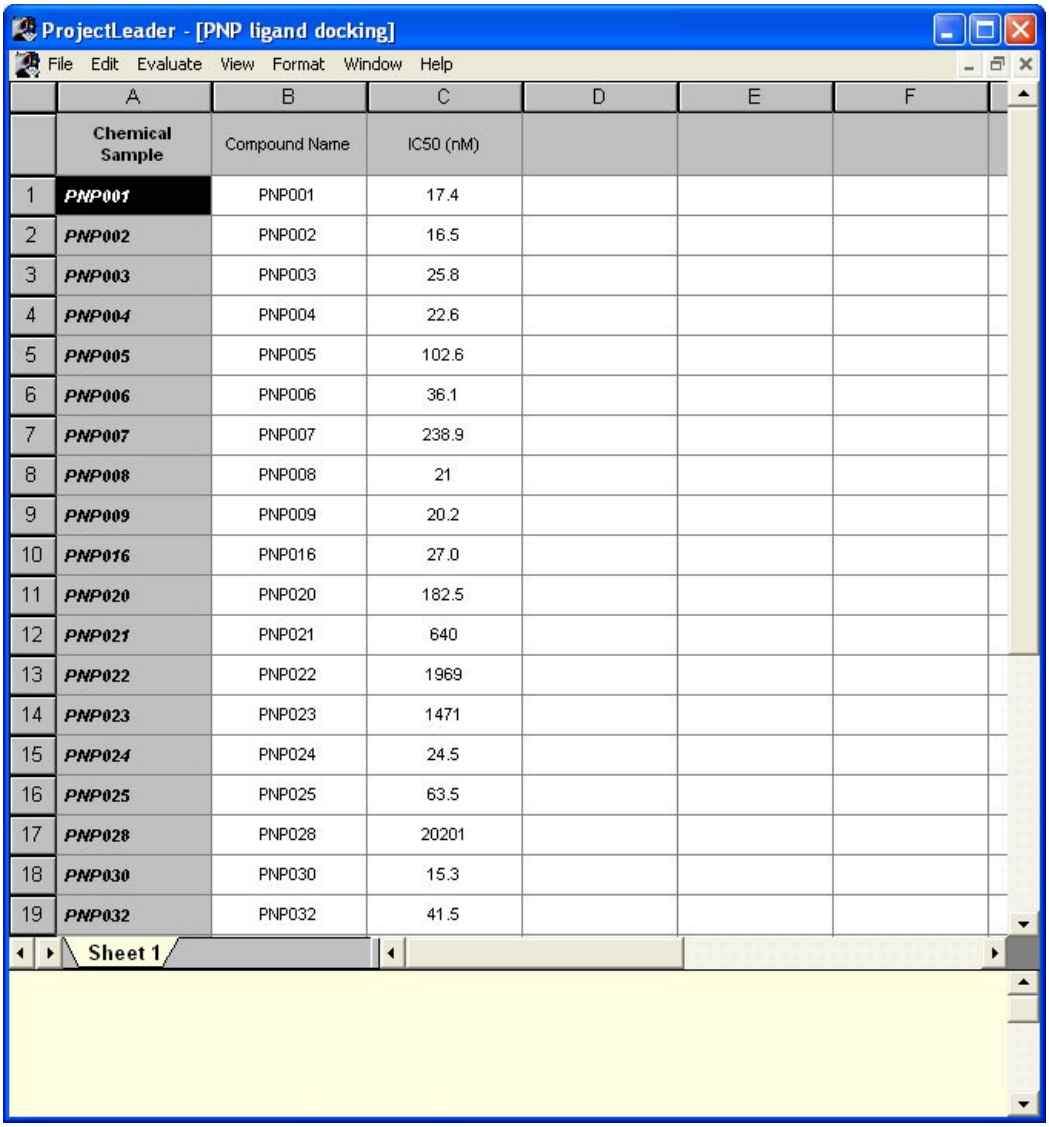
7. Choose **Compound Name** from the list, check **Load additional data** and finally choose **OK**.

   The dialog closes and the **Chemical Sample** column is populated with the 19 structures from the file and new columns containing the compound names and the IC50s are added to the table. These are the 19 structures from the library of 34 PNP inhibitors described in Reference 1. We have excluded ligands that are too big to fit into the rigid active site and those that have chiral centers. When chiral centers are present, it is necessary to investigate all diastereoisomeric forms when docking and that investigation is beyond the scope of this exercise.

   Since the structures in this SD file are 3D structures optimized with MM3, we did not check **Convert 2D structures to 3D**. We checked **Load Additional Data** because this SD file contained IC50 activity data for the compounds and it was convenient to have this added to the project automatically from the information we exported from the database.

8. Choose **File | Save** and save the file as `PNP ligand docking.plp`.

The project should look like this.



	A	B	C	D	E	F
	Chemical Sample	Compound Name	IC50 (nM)			
1	PNP001	PNP001	17.4			
2	PNP002	PNP002	16.5			
3	PNP003	PNP003	25.8			
4	PNP004	PNP004	22.6			
5	PNP005	PNP005	102.6			
6	PNP006	PNP006	36.1			
7	PNP007	PNP007	238.9			
8	PNP008	PNP008	21			
9	PNP009	PNP009	20.2			
10	PNP016	PNP016	27.0			
11	PNP020	PNP020	182.5			
12	PNP021	PNP021	640			
13	PNP022	PNP022	1969			
14	PNP023	PNP023	1471			
15	PNP024	PNP024	24.5			
16	PNP025	PNP025	63.5			
17	PNP028	PNP028	20201			
18	PNP030	PNP030	15.3			
19	PNP032	PNP032	41.5			

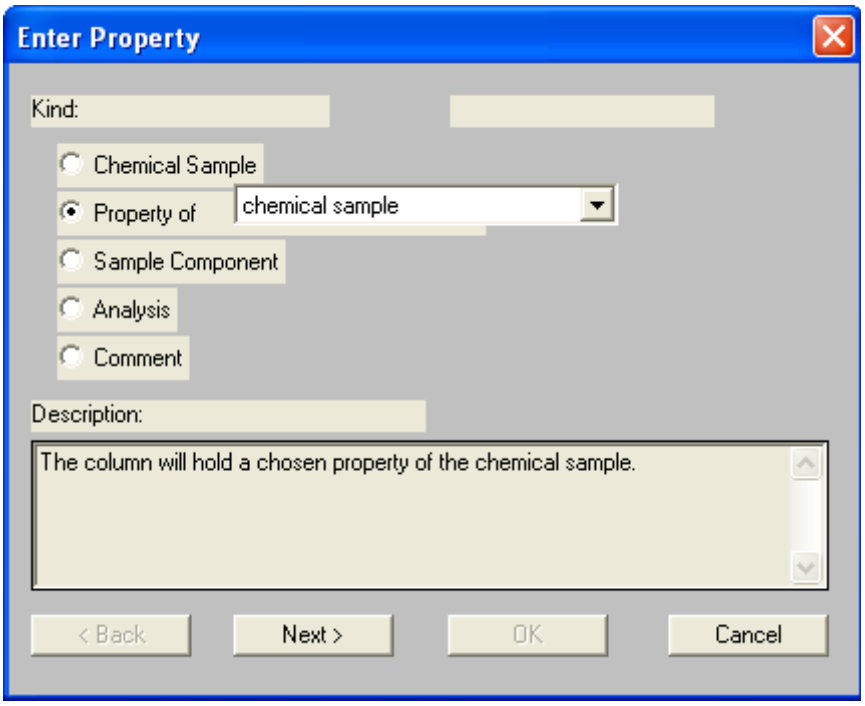
Creating log(IC50) data

To compare the PMF score in kcal/mol with the IC_{50} concentrations, we need to create a column containing the log of the IC_{50}.:

To create a column containing log(IC50)

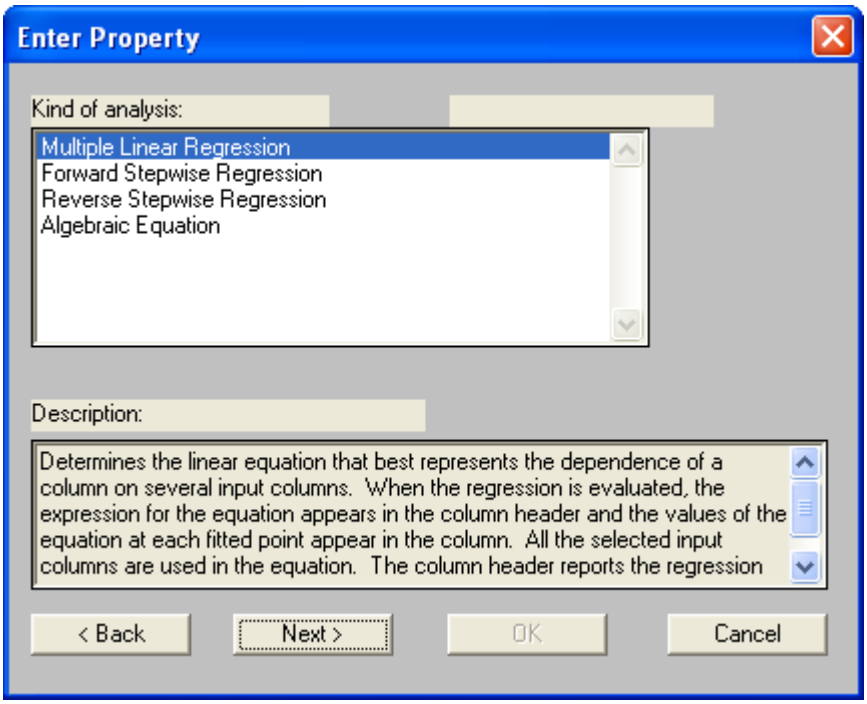
1. Double-click the empty **D** column header.

The Enter Property wizard appears.



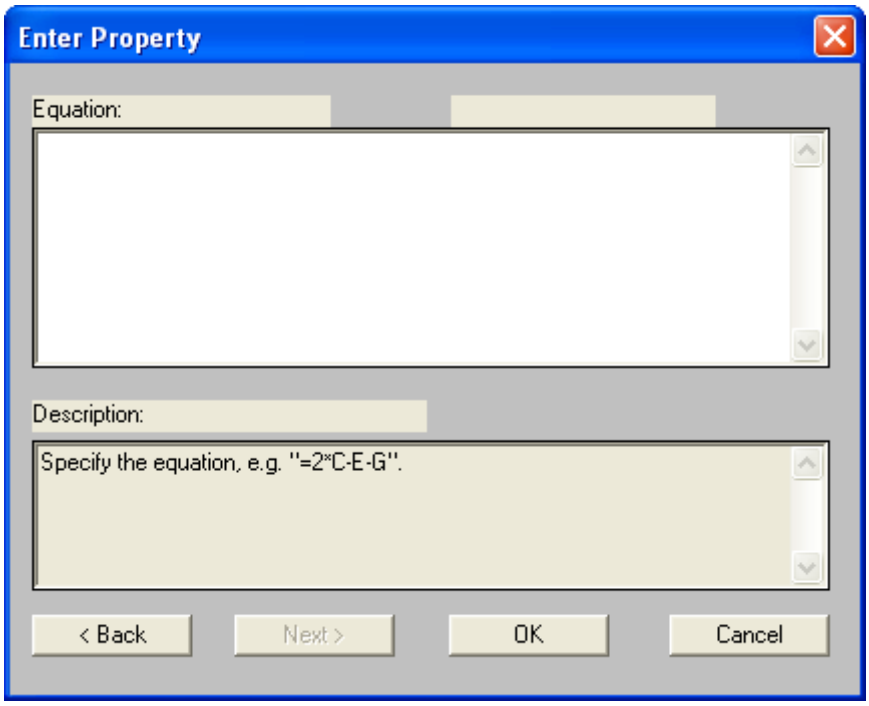
2. Choose **Analysis** and then **Next>**.

   The next step in the wizard appears.



3. Choose **Algebraic Equation** from the list and then choose **Next>**.

The step for entering the algebraic equation appears.



4. Type the equation `log(IC50) = log(C*0.000000001)` into the Equation text box and choose **OK**.

   You multiply for 10^{-9} to account for the nanomolar values of the IC_{50}. The Enter Property dialog disappears and the column header fills with the equation you typed.

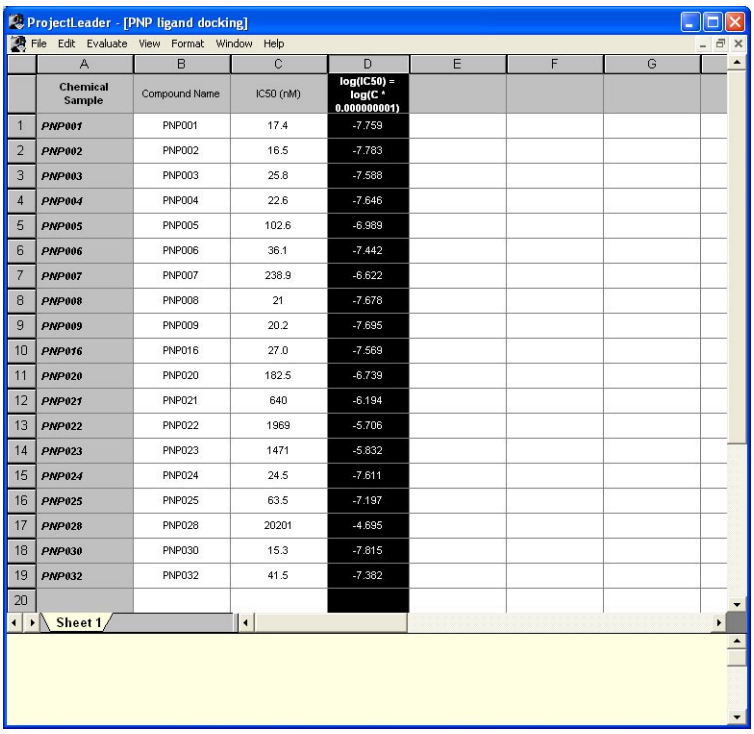
5. Choose the **D** column button above the column header.

   The column is selected and it highlights with a black background.

6. Choose **Evaluate | Cell**.

   The cells in column D fill with the values of log(IC_{50}) and the project

appears like this. You are now ready to dock into the bovine PNP enzyme



ProjectLeader - [PNP ligand docking]

File Edit Evaluate View Format Window Help

	A	B	C	D	E	F	G
	Chemical Sample	Compound Name	IC50 (nM)	log(IC50) = log(C * 0.000000001)			
1	PNP001	PNP001	17.4	-7.759			
2	PNP002	PNP002	16.5	-7.783			
3	PNP003	PNP003	25.8	-7.588			
4	PNP004	PNP004	22.6	-7.646			
5	PNP005	PNP005	102.6	-6.989			
6	PNP006	PNP006	36.1	-7.442			
7	PNP007	PNP007	238.9	-6.622			
8	PNP008	PNP008	21	-7.678			
9	PNP009	PNP009	20.2	-7.695			
10	PNP016	PNP016	27.0	-7.569			
11	PNP020	PNP020	182.5	-6.739			
12	PNP021	PNP021	640	-6.194			
13	PNP022	PNP022	1969	-5.706			
14	PNP023	PNP023	1471	-5.832			
15	PNP024	PNP024	24.5	-7.611			
16	PNP025	PNP025	63.5	-7.197			
17	PNP028	PNP028	20201	-4.695			
18	PNP030	PNP030	15.3	-7.815			
19	PNP032	PNP032	41.5	-7.382			
20							

Sheet 1

in the next step.

Docking the library into bovine PNP (1A9T)

The next step is to dock all of the ligands into bovine PNP (1A9T.csf). 1A9T.csf is a chemical sample file generated from the 1A9T PDB file by applying symmetry operations to create the three chains in the trimeric biologically active form of bovine PNP and adding hydrogen atoms.

Setting up the docking column

First we create a column header that uses a procedure which takes the chemical sample from column **A**, docks it into a specified protein target and reports the docking score in the column cell.:

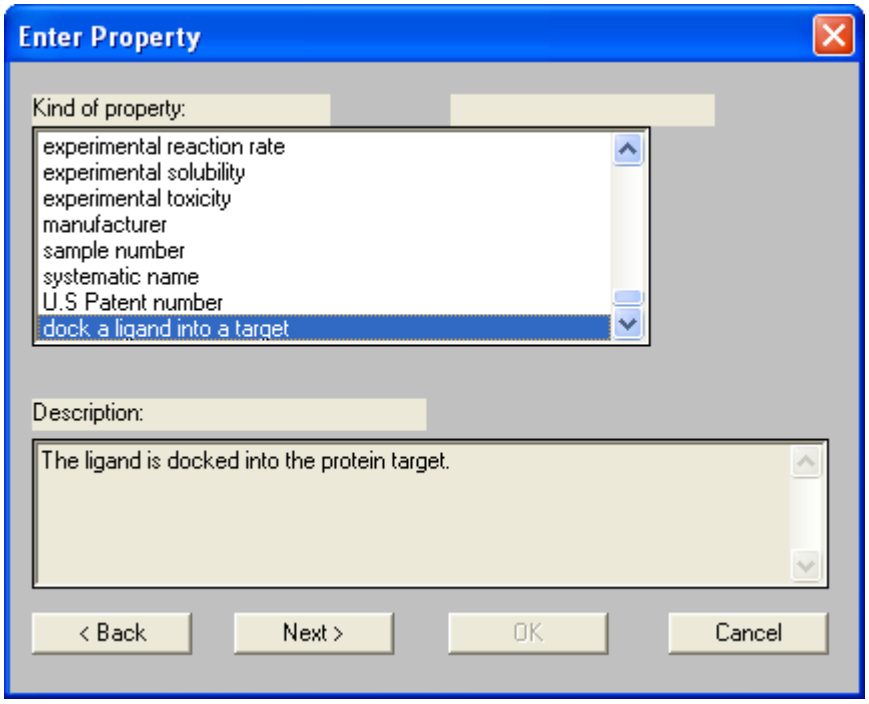
To create a column containing scores from docking

1. Double-click the empty **E** column header.

   The **Enter Property** wizard appears.

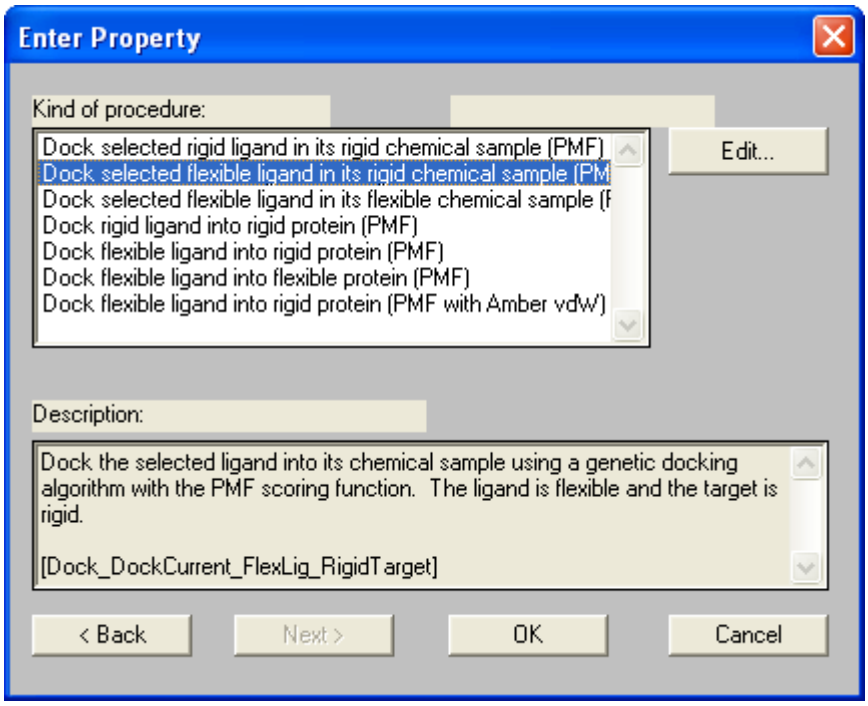
2. Choose **chemical sample** from the **Property of** pull down menu and choose **Next>**.

   The step asking for the property appears.



3. Scroll the property list, choose the property **dock a ligand into a target**, and choose **Next>**.

The step requesting the procedure for docking appears.

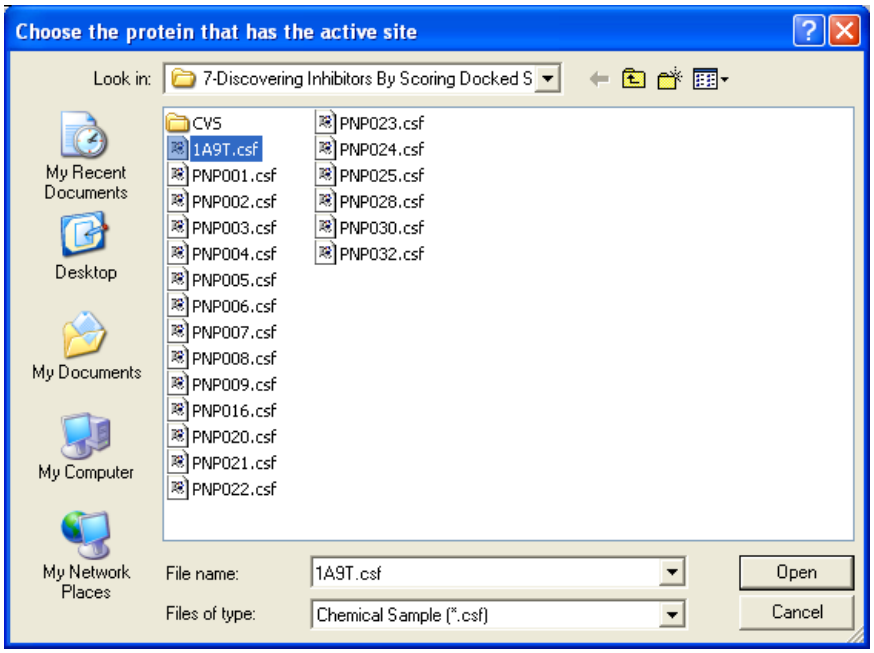


NOTE

You can choose other procedures to dock a selected ligand into the active site in its own chemical sample.

4. Choose **Dock flexible ligand into rigid protein (PMF)** and choose **OK**.

   The Enter Property wizard closes and the **Choose the protein that has the active site** open dialog appears. Ligands will be docked into the first



group of type active site. The protein should be prepared for docking. It

must contain an active site type group that you have created. If the active site lies on the surface of a protein, you may also need to add water molecules to simulate the water on the surface of the protein and the solvent, otherwise the ligands may prefer to "dock" into the vacuum outside of the protein.

5. Choose `1A9T.csf` and choose **Open**.

   The column header fills with the title: **PMF score flexible ligand into a rigid target**.

Docking the library

The project is now ready to dock the library of compounds.

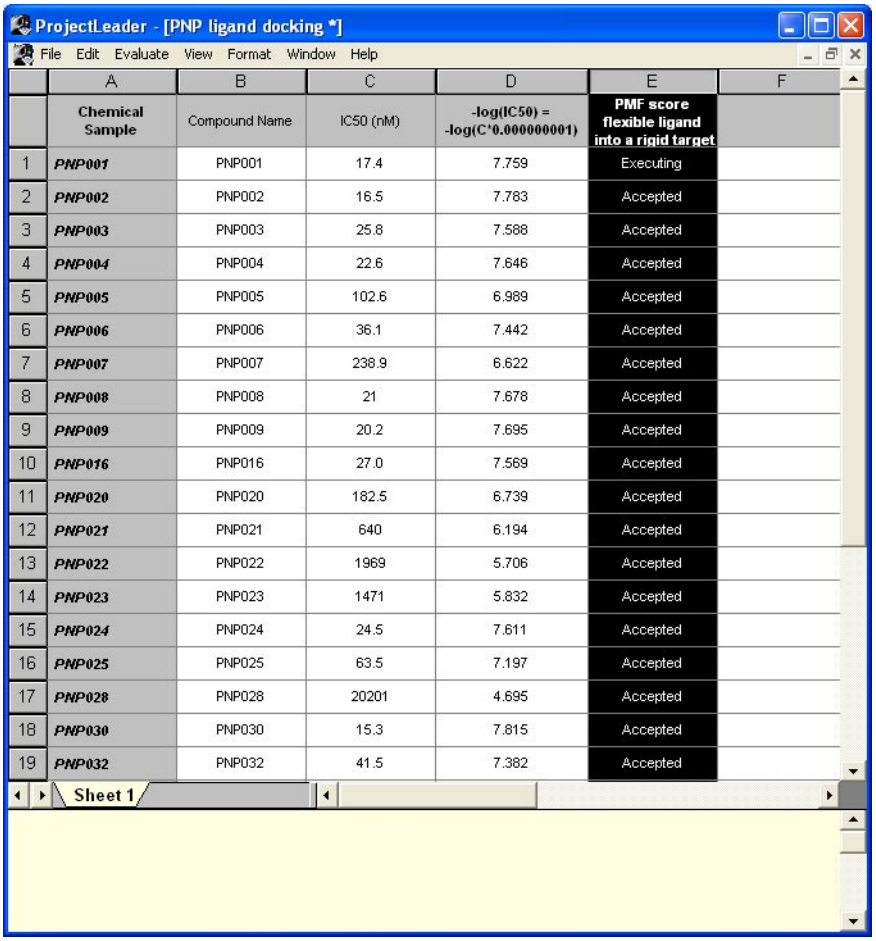
To dock the library

1. Choose **E** at the top of the column.

   Column **E** highlights.

2. Choose **Evaluate | Cell**.

Nineteen docking calculations using the FastDock compute engine are



	A	B	C	D	E	F
	Chemical Sample	Compound Name	IC50 (nM)	-log(IC50) = -log(C*0.000000001)	PMF score flexible ligand into a rigid target	
1	PNP001	PNP001	17.4	7.759	Executing	
2	PNP002	PNP002	16.5	7.783	Accepted	
3	PNP003	PNP003	25.8	7.588	Accepted	
4	PNP004	PNP004	22.6	7.646	Accepted	
5	PNP005	PNP005	102.6	6.989	Accepted	
6	PNP006	PNP006	36.1	7.442	Accepted	
7	PNP007	PNP007	238.9	6.622	Accepted	
8	PNP008	PNP008	21	7.678	Accepted	
9	PNP009	PNP009	20.2	7.695	Accepted	
10	PNP016	PNP016	27.0	7.569	Accepted	
11	PNP020	PNP020	182.5	6.739	Accepted	
12	PNP021	PNP021	640	6.194	Accepted	
13	PNP022	PNP022	1969	5.706	Accepted	
14	PNP023	PNP023	1471	5.832	Accepted	
15	PNP024	PNP024	24.5	7.611	Accepted	
16	PNP025	PNP025	63.5	7.197	Accepted	
17	PNP028	PNP028	20201	4.695	Accepted	
18	PNP030	PNP030	15.3	7.815	Accepted	
19	PNP032	PNP032	41.5	7.382	Accepted	

queued. As each calculation completes,

- the PMF score appears in its cell,
- the chemical sample in column **A** is updated to have the conformation and position of the docked structure in the protein,
- a new chemical sample is created that contains the ligand docked in the protein,
- a new map file is created that contains the protein and those docked ligand structure within 100 kcal of the best structure, and
- the output of the FastDock calculation is returned in the `.io` folder.

Each docking is set to run 10,000 generations and requires about 5 minutes on a 2 GHz Pentium 4 for a total of 95 minutes.

The time required varies depending upon the number of rotatable bonds in the ligand. The time required for docking can be reduced by decreasing the number of generations and by docking rigid ligands into rigid proteins. It is our experience that 1,000 generations of ligands with a few rotatable

📖 TIP

You may continue working on your computer while docking is taking place, but you should periodically save the project to avoid loss of data should your system need restarting.

bonds will give docked poses of rough accuracy, 3,000 generations yield poses of useful accuracy in agreement with chemical intuition, 10,000 generations, poses of good accuracy usually reproducible to a couple of kcal/mole while 40,000 generations are recommended for publication quality work.

Examining docking settings

While the ligands are docking into bovine PNP, you can look at the FastDock settings used in this procedure.

To examine docking settings

1. Double-click the **E** column header.

   The Enter Property wizard appears.

2. Choose **Next>**.

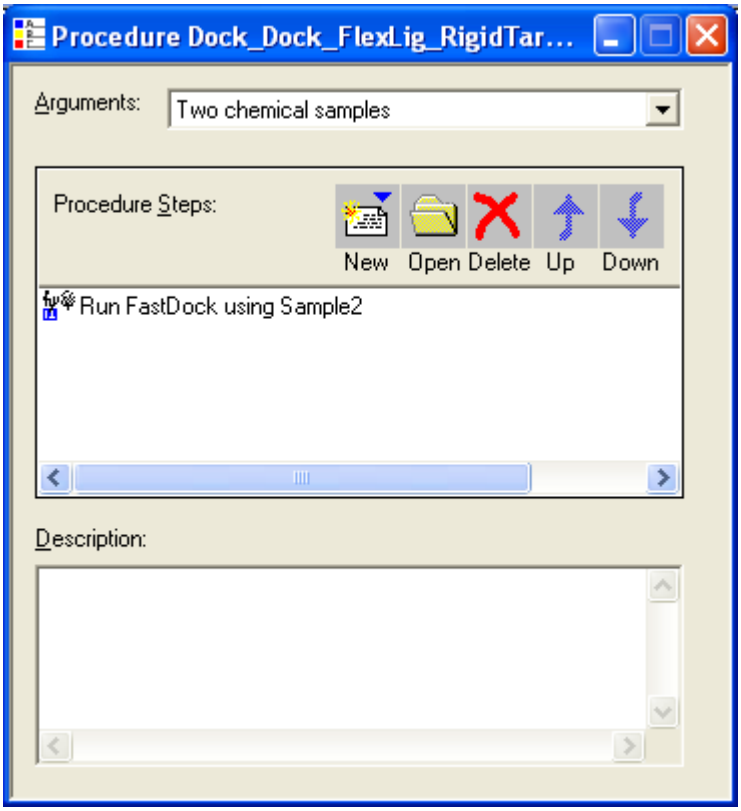
   The step for choosing the **Kind of Property** appears.

3. Choose **Next>**.

   The step for choosing the **Procedure** appears.

4. Choose **Edit**.

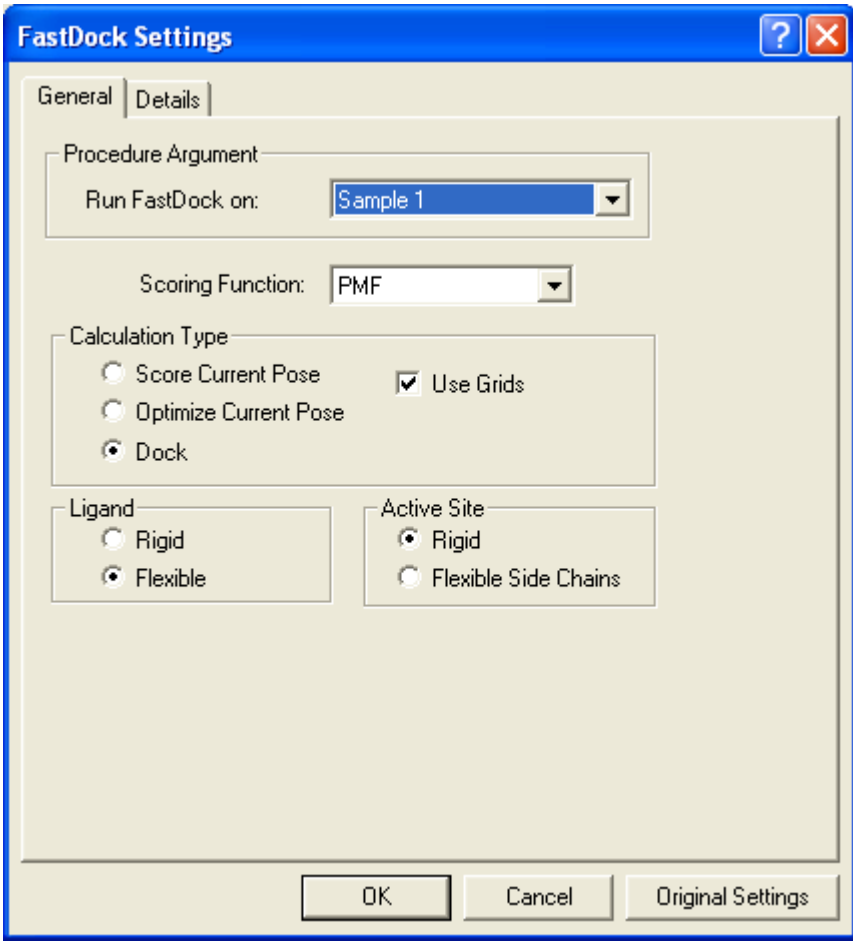
The Procedure Editor opens and displays the dialog Procedure



Dock_Dock_FlexLig_Rigid_Target_Protein_GETTARGET(). This procedure uses two chemical samples for input: one is the ligand to be docked and the other is the protein containing the active site group.

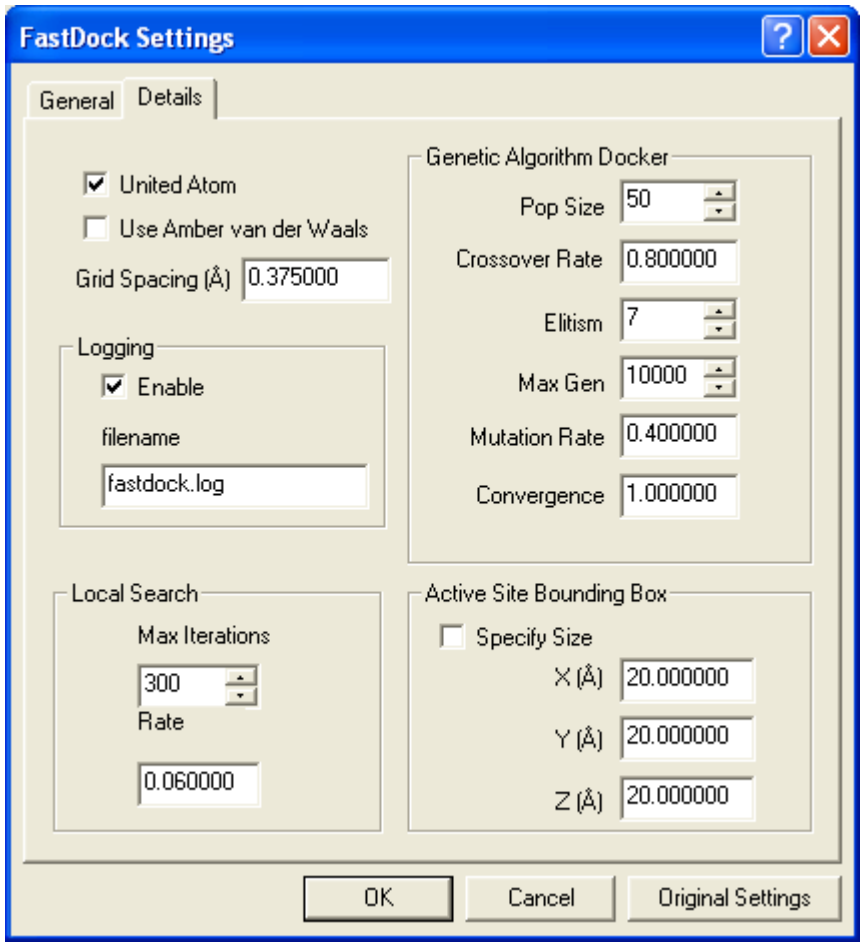
5. Double-click **Run FastDock using Sample2**.

The FastDock control dialog appears. Changes on this tab would create a



procedure that is fundamentally different from its name and should not be made.

6. Choose the **Details** tab.



This shows that the procedure is set to run 10,000 generations on a population of 50 chromosomes keeping the seven best scoring members from one generation to the next. The active site bounding box is not specified and the bounding box is therefore automatically determined from the size of the active site.

7. After examining the settings, choose **Cancel** and **File | Exit** to close the Procedure Editor without making changes.

   When you do need to make changes it is best to create a new procedure and an new experiment with different names so that the old experiments and settings remain available.

📖 TIP

If a ligand docks outside of the active site or protein, then the bounding box may be set too large. In a protein crystal, the space outside of the protein is filled with water that keeps the ligand close to the protein. In CAChe the space is empty and the ligand is free to move into it.

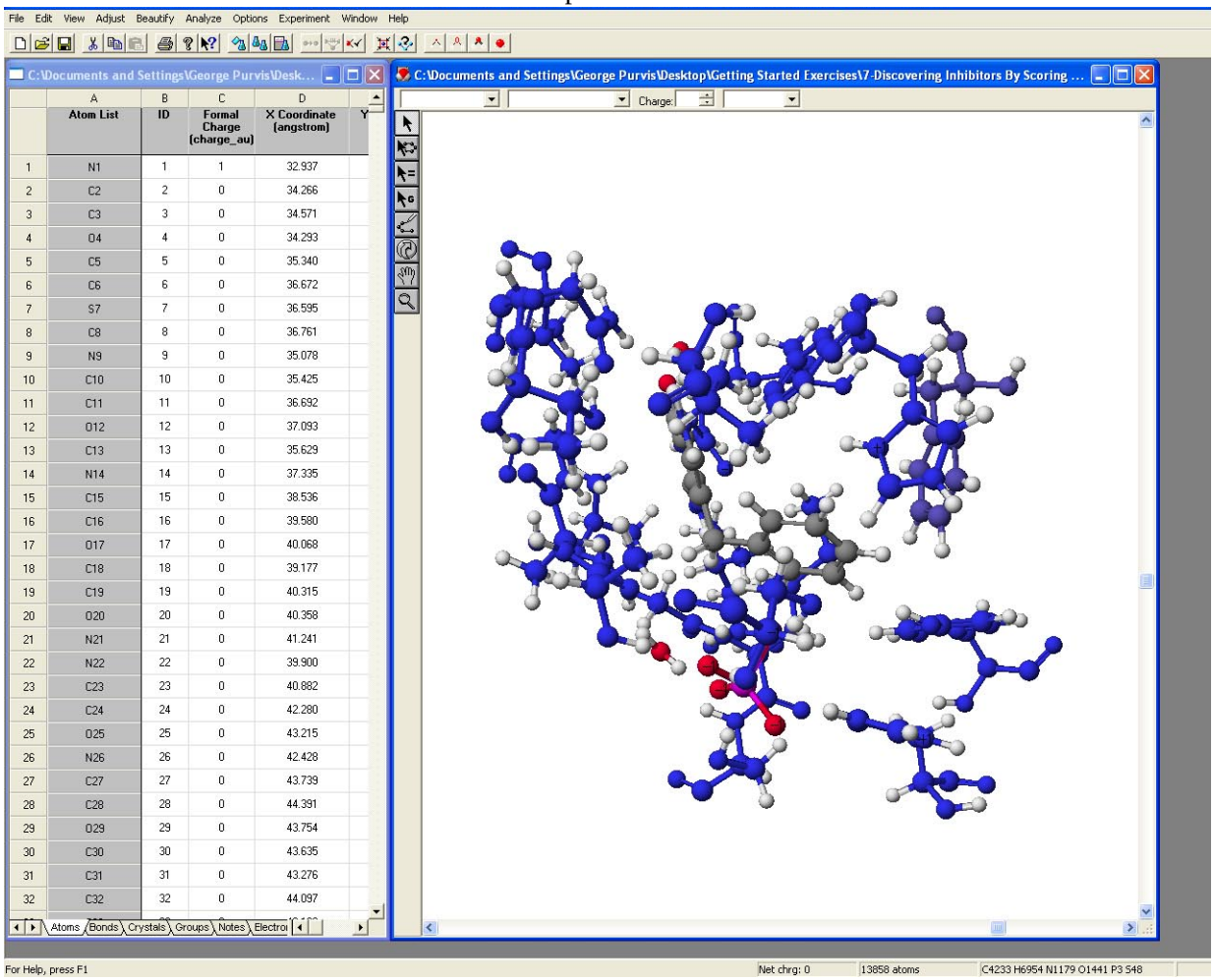
Examining docked structures

You can examine the docked structures as they are generated.

To examine a docked structure

1. Open the Workspace from the Windows Start menu.
2. After the cell for PNP002 fills with the docking score, open PNP002-1A9T.csf in the Workspace.

   PNP002-1A9T opens with a 3D structure window and a Chemical



   Properties Workbook window opens.

   Notice that there are 13,000 atoms present in this protein. This is a trimeric structure generated from the monomer in 1A9T.pdb by applying crystal symmetry operations. Also, note that active site heavy atoms are colored bright blue, except for a dark blue phenyl alanine residue that caps the entrance to the pocket. The phenyl alanine residue comes from an adjacent chain and would be missing in a single chain model of an active site. It is important here in that it closes the active site.

   To learn more, examine the docked structure as you have learned how to do in previous exercises. Answer these questions:

   - How many hydrogen bonds are there between the ligand and the active site?

- Is the ligand oriented as you would expect it to be based upon the 1G2O structure of immucillin?
- What is the final docked score from the workbook?
- Does PNP002 fit in the active site pocket formed by the adjacent surface?

Continue exploring docked structures of other ligands as the project table fills with results.

Evaluating docking scores

Once all docking has completed, you can determine the structure-activity relationship between the docking score and log(IC_{50}).

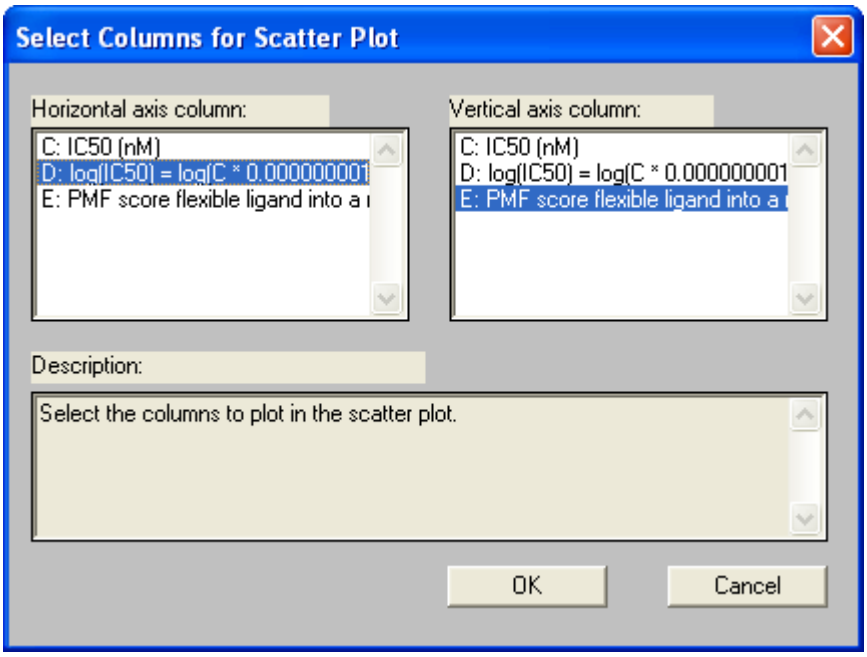
To determine the score-activity relationship

1. Choose the **E** button at the top of column **E**.

   Column **E** highlights and all cells in the column are selected.

2. Choose **View | Scatter** plot.

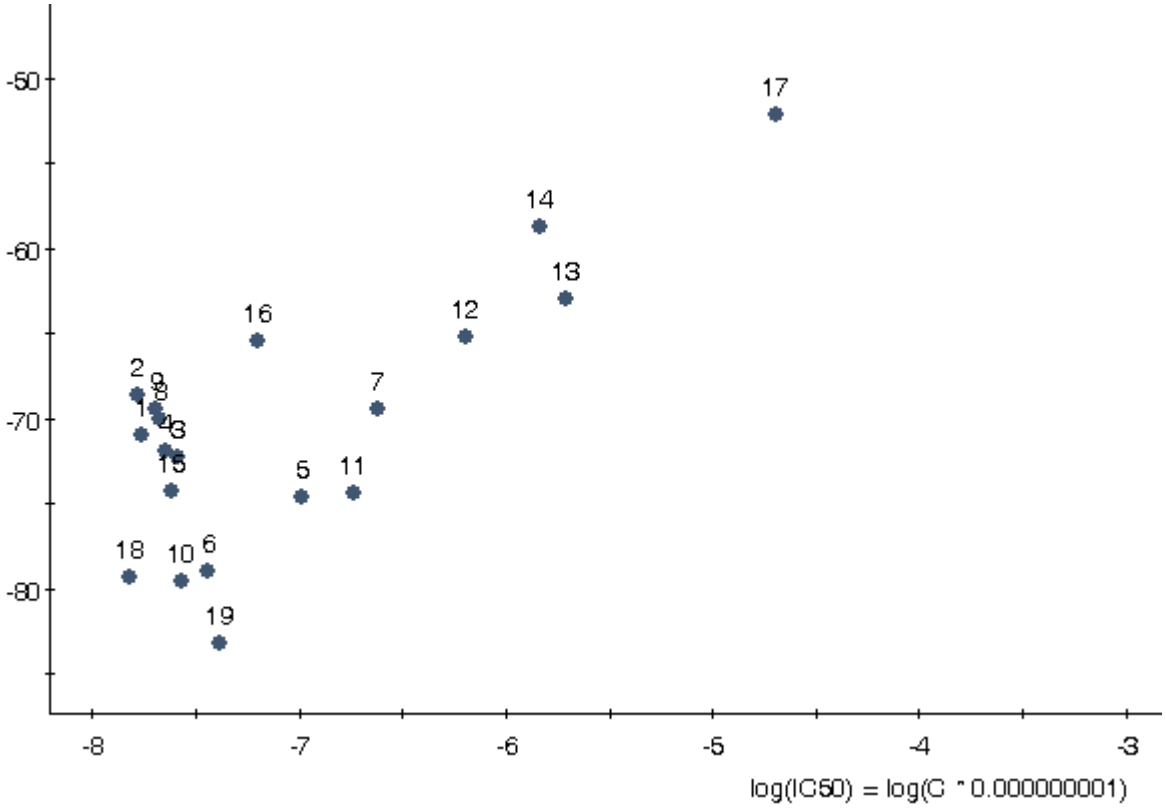
   The Select Columns for Scatter Plot dialog appears.



3. Choose **D: log(IC50) = log(C * 0.000000001)** from the **Horizontal axis column** list and **E: PMF score flexible ligand into a rigid target** from the **Vertical axis column** list.

4. Choose **OK**.

The dialog disappear and a scatter graph appears.that shows the PMF



score for each row in the table vs the log(IC_{50}).

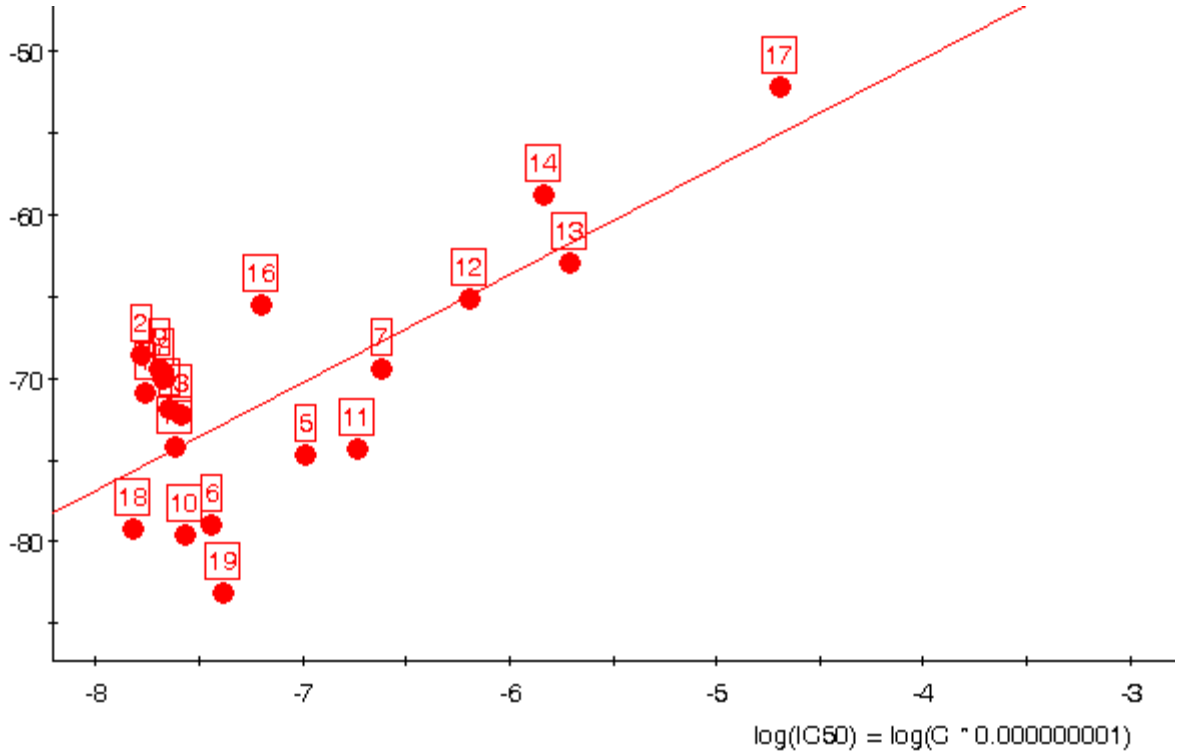
5. To quantify the apparent strong linear relationship between the PMF score and log(IC_{50}), choose the black circle in the graph legend.

   The circle and all of the points in the scatter graph highlight in red.

6. Choose **Evaluate | Regression**.

   The best linear fit to the data is drawn on the graph and the regression

equation appears at the bottom of the graph. The calculated R^2 value of



0.59 and R_{cv}^2 of 0.51 are good for relationships between docking scores and activities.[1]

1. Muegge, I. and Rarey, M., "Small Molecule Docking and Scoring", *Reviews in Computational Chemistry"*, Libkowitz, K. B., and Boyd, D. B. eds., **2001**, *17*, 1-60.

So What?

The development of scoring algorithms for docking is a current area of active research. The definitive scoring function has not yet been developed. Published R^2 values for scoring methods range from 0 to 0.13 to 0.37. In a few special cases for specific classes of ligands and active sites, R^2 values of greater than 0.9 have been reported, but these are the exceptions and the same scoring functions applied to other classes give poor results.

Because proteins flex to adapt to the ligand and because different ligands can incorporate a different number of water molecules into the active site, it is unlikely that simple scoring functions for docking into rigid proteins will ever routinely achieve R^2 values of 0.9 and greater.

However, current docking and scoring methods provide useful models for guiding the design of ligands by suggesting intuitively correct poses for ligands in docked structures and by providing an approximate rational ordering of compound activity prior to synthesis.