CREVASSE USER MANUAL

If you use Crevasse in your work, please cite the following two papers:

E.E. Thompson, A.P. Kornev, N. Kannan, C. Kim, L.F. Ten Eyck, S.S. Taylor, Comparative surface geometry of the protein kinase family, *Protein Science*, 18(10): 2016-2026, 2009.

J.C. Mitchell, R. Kerr and L.F. Ten Eyck, Rapid atomic density measures for molecular shape characterization, *J. Mol. Graph. Model.*, 19(3): 324-329, 2001.

Compiling Crevasse

Crevasse is written in C++ has been tested on linux systems with GCC, Macintosh computers with Darwin, and Windows systems with Cygwin and the Cygwin implementation of GCC. It should compile on systems that can compile and run FADE.

Crevasse uses the NIST Template Numerical Toolkit v1.26 and the JAMA C/C++ linear algebra packages available at http://math.nist.gov/tnt/index.html. A copy of the library has been included with the distribution for your convenience.

The command "tar –xvzf crevasse_1.0.tgz" will extract the download. Typing the command "make" in the root directory of crevasse will compile the executable.

It is best to increase your stack size limit before running Crevasse. Crevasse runs an STL sort and can run out of stack space with the default settings on some systems. The command limit will show the settings in csh or tcsh, or ulimit –s in bash. If the stacksize is below 96000, set it with limit stacksize 96000 in csh or tcsh, or ulimit –s 96000 in bash.

Other Software

You must have FADE compiled and installed on your computer. Crevasse is only compatible with versions of FADE released after October 8, 2009. FADE is available at http://mitchell-lab.org/mitchell-lab/FADE.php.

Pocket-finding requires that hydrogen atoms be added to the protein. The command-line version of Reduce from the Kinemage website is a simple, accurate way to add them to crystal structures. http://kinemage.biochem.duke.edu/software/reduce.php

---

Finding Pockets

The basic pocket-finding process is simple.

1) **Clean up the PDB structure.**
2) **Add all hydrogen atoms.**
3) **Run FADE.**
4) **Run Crevasse.**

PDB Structure Cleanup

Cleaning up PDB structures includes removing alternate atom locations, converting selenocysteine and selenomethionine residues to cysteine and methionine if necessary, removing water, and choosing whether to include or remove co-crystallized ions and ligands.

A small set of scripts to do common cleanup tasks has been included with Crevasse in the /scripts directory. These scripts are from the CCMS DOT distribution. There are instructions in the "Scripts" section of this manual.

FADE reads both ATOM and HETATM records from PDB files, so it will not automatically remove co-crystallized molecules like glycerol, phosphate ions, or detergent. Generally these molecules should be removed before studying a protein. Other co-crystallized molecules such as NAD, FAD, Ca++, or Mg++ can be either left in the protein or removed. Removing bound ligands will usually result in a pocket computed at the ligand site.

Missing side chain atoms change the shape of the protein surface and will give misleading results. Make note of areas with missing side chains and avoid interpreting your results in those areas on the protein surface. Attempting to rebuild side chains is not recommended because of the difficulty of choosing rotamers.

Add all hydrogen atoms.

FADE works by counting atoms, so the structure must include all of them. Exact protonation states are not important, as long as most of the side chains are reasonable. Hydrogen atoms can be added with Reduce at the command line, by uploading to Kinemage, or with other common tools like CHARMm.

Run FADE.

Recommended FADE parameters are to set the grid spacing 0.5 angstroms, return only FADE scores three grid points from the molecular surface, and to use atomic density cutoffs of 4.3 to 6. These settings have been tested on a wide variety of proteins. (Refer to the FADE manual for more information on the parameters.) The FADE command will be:

FADE –r =2 –o=myfile 3 3 4.3 6

FADE will output a file with grid information, a set of points, and the associated exponential density scores with a file extension .fad. The returned points should line concavities on the protein surface. Points that are interior to the protein and points far from the surface are discarded with these settings, which is extremely important.

Run Crevasse.

Crevasse also runs at the command line. The usage is "crevasse filename" and Crevasse will output a series of files containing information about the pockets on the protein surface. Crevasse reads the FADE output back onto the grid FADE computed and segments the FADE output into individual pockets by looking for connected points. Then it computes a bounding box

around the points (based on the eigenvectors of the covariance matrix) to find the size of the pocket for filtering. The center and length of the longest axis through the box is part of the output.

Options are as follows:

-o output base filename (extensions will be automatically added). Default is the input filename.

-s size of clusters to keep, in grid points. The default is 80, which is small enough to return small pockets, but large enough to discard noise.

-l minimum length of longest box axis in Å. The default is 0, since noise is being discarded with size parameter. Use this parameter in combination with the size in grid points to "tune" the pockets the Crevasse software outputs. Small settings like 6Å are good for finding small, potentially druggable sites, or sites of interaction with one or two amino acid side chains. In a protein where the ligand binding site is the main pocket, settings more like 20Å can filter noise.

A typical Crevasse run is started with:

crevasse –o Myoutfile –l 6 myfile

Output.

Crevasse writes out a series of files once the computation is finished. If you gave Crevasse the name "Myoutfile" the following files would be created.

Myoutfile.sum is a summary of the crevasse computation for importing into a spreadsheet or further processing. It has one line for each pocket found, reporting the number of points, a rough volume based on the number of points, the volume of the bounding box, the dimensions of the bounding box, and the xyz coordinates for the box center.

Myoutfile.xyzc is a file containing all the individual points and which cluster they belong to. The format is "x y z crevasse#". This file can be use to find nearby atoms and flag them as belonging to a pocket. It has been used with some success to weight DOT docking runs.

Myoutfile.pdb is a PDB formatted file that can be read into a viewer such as VMD or PyMol for visualization. All of the points are represented as hydrogen atoms. Each cluster has been a different residue number, so coloring by residue will show the points in different colors. In PyMol, showing a semi-transparent surface over the pockets is a useful visualization.

Advanced Options

The suggested FADE cutoffs and Crevasse settings have been tested on a wide variety of proteins and generally work reasonably well. However, FADE and Crevasse are both very flexible.

FADE filters its output, and Crevasse is written to assume the FADE output has been filtered in such a way as to return a shallow shell of points at the molecular surface. The recommended 3 gridpoint distance returns points very close to the VDW surface of the protein (within grid error). If pockets at the solvent-excluded surface are desired, FADE should be set to return points farther than 3 gridpoints away. When the distance from the protein is changed, the 4.3-6 FADE score cutoff should also be adjusted. FADE scores change distribution depending on the cutoff, increasing with increasing distance from the protein surface.

Crevasse also has two advanced tuning options that change the way it finds clusters of points. The first controls the way Crevasse searches, and the second controls the degree of connectedness required to add a point to a cluster.

-6point (no value for this parameter) By adding -6point to the command line, Crevasse can be shifted to a mode where the search does not explore points on the grid that are diagonal to the point in question. The setting is sometimes useful if the FADE parameters have been changed to return a deeper shell of points around the protein.

-n minimum neighbors required to add a point to the growing pocket, defaults to 2. Raising the required number of neighbors can divide the FADE output into more small pockets.

Scripts

Some scripts are provided with Crevasse for to make bulk protein processing easier. The simpler script, run_crevasse.sh will run Crevasse on a single file downloaded from the PDB. The more complex script will bulk process a series of pdb files.

Both scripts require some environment variables.

For FADE, set FADE_EXEC according to the FADE manual. The default location is

FADE_EXEC /usr/local/apps/fadepadre/bin/exec

For Crevasse, set CREVASSE_DIR to the directory with the crevasse executable. (This will be wherever you put the crevasse_1.0 directory, possibly /usr/local/apps/crevasse_1.0.

These scripts use Reduce (mentioned above) to add hydrogen atoms to PDB files. Reduce must be available on your path for them to work. Try "which reduce" to see if your system can find it.

Find pockets on one pdb file from a script.

The script will remove alternate atom locations (do this by hand if you care which conformer is selected), convert MSE and CSE to MET and CYS respectively, strip water molecules, and remove bound ligands and other non-protein atoms designated as HETATM. Type:

./run_crevasse.sh pdbid

The script assumes your file came from the PDB, so if your pdbid is 1RGS, the filename would be pdb1rgs.ent and you would type the following. ./run_crevasse.sh 1rgs The script will automatically name all your files starting with 1rgs.

Find pockets on a large series of proteins.

You will do the following:

1) **Organize your PDB files (this is done for you if you bulk download from RCSB)**
2) **Make a list of your PDB IDs to run**
3) **Make a custom master script**
4) **Run the master script to process your PDB files and find pockets**

The make_crevasse_script assumes your files are stored in the RCSB directory heirarchy of /pdb/middle 2 characters. For example, 1ATP would be stored as pdb/at/pdb1atp.ent.

Make a list of all the PDB ids in a text editor like vi or gedit. Put one four-letter pdb ID, like "1atp", on each line of the file. Match the capitalization to your filename convention, lowercase if you did a bulk download from RCSB.

Next, create a custom master script, using make_crevasse_script.pl. Create your custom script with the following two commands:

```
perl ~$CREVASSE_DIR/scripts/make_crevasse_script.pl < pdbids.txt > crevasse.sh
```

```
chmod crevasse.sh 755
```

Go one directory above your "pdb" directory so from our 1atp example pdb/at is a valid path. Run your custom script and find pockets on a large group of proteins at once by running:

```
crevasse.sh | tee crevasse_log.txt
```

Reduce, FADE, and Crevasse output files will be in the directory with the PDB file and a log of the run will be where you started the script.