Installing and Using *Consolv* 1.0

A tool for prediction of conserved solvation sites across independently-solved crystallographic protein structures.

Introduction

Consolv 1.0 is a tool for predicting whether water molecules bound to the surface of a protein are likely to be conserved or displaced in other, independently-solve crystallographic structures of the same protein. *Consolv* was developed by members of the Protein Structural Analysis and Design Laboratory at Michigan State University, and was supported in part by NSF grant DBI-9600831.

For literature references related to *Consolv*, please see the section on Algorithmic Details.

Back to Table of Contents

Getting and Installing *Consolv*

The latest version of *Consolv*, as well as the most recent version of this documentation, can be found at the home page for the Protein Structural Analysis and Design Laboratory , Department of Biochemistry, Michigan State University. To install *Consolv*, perform the following steps:

- Download consolv.tar.Z <fix link>
- Place this file in the directory into which you wish to install the software, and enter the following command:

  ```
  zcat consolv.tar.Z | tar xvf -
  ```

- Among the files created will be a file called `ahplist.dat`. This file contains the atomic hydrophilicity values for various atom types commonly found in crystallographic protein structures. It is sometimes convenient to move this file to a common directory where all users may edit this file in order to add new atom types. If you wish to move the `ahplist.dat` file, you should do so before the following step.
- To compile the software, use the following command from the directory where *Consolv* was

unpacked:

  - `./Install`

  The install command will ask you for two directories: the location of the `ahplist.dat` file, and the location of the Brookhaven Protein Data Bank (PDB) files. You should use complete path names, and no environment variables or shell expansion characters (i.e. do not use ~username, $HOME, etc.).

- *Consolv* will now be compiled using the file locations specified. If you later wish to change the default file locations, you can re-run `./Install` from the *Consolv* installation directory.
- To run *Consolv*, run the `Consolv` script in the installation directory. For example, if *Consolv* is installed in `/usr/local/biochem/consolv`, then you would use the command: `/usr/local/biochem/consolv/Consolv` to run the software.

Back to Table of Contents

Troubleshooting Shell Scripts

Many of the individual commands that make up *Consolv* are shell scripts. These scripts assume that the C-shell, the Bourne-shell and the Korn-shell are installed in the standard locations. If you have difficulty running one or more of these scripts, you may edit the first line of each script to point to the correct shell location for your system. *Consolv* scripts which are dependent on shell locations include:

Script	Shell
`Consolv`	`/bin/sh`
`Install`	`/bin/sh`
`sumpred`	`/bin/ksh`
`bin/allwats`	`/bin/csh`
`bin/consolv2pdb`	`/bin/sh`
`bin/extract`	`/bin/csh`

If you do not have `ksh` installed on your system, try using `sh` instead. Some versions of `sh` have extended features similar to those found in `ksh`.

Back to Table of Contents

Running *Consolv*

Overview of usage

Consolv can be run in one of three modes. The most common usage of *Consolv* is to take a single crystallographic protein structure, extract environmental information about each of the water molecules in the structure, and then predict whether each water molecule in the first-hydration shell is likely to be conserved or displaced in an independently-solved structure of the same protein. This is accomplished by running *Consolv* in **application mode**.

Test mode can be used to test the predictive accuracy of *Consolv* using a pair of structures of the same protein. This mode requires two independently-solved crystallographic structures of the same protein, superimposed into the same coordinate system. As in application mode, environmental information is extracted for waters in the first hydration shell of the first structure, and the conserved/displaced status of each water molecule is predicted. In test mode, however, an additional accuracy test is done using the second structure. The actual conserved/displaced status of each first shell water with regard to the second crystallographic structure is determined, and the predictions given by *Consolv* are compared to the actual conservation between the two structures. This mode may be useful in testing the accuracy of *Consolv* on a particular class of proteins, verifying that *Consolv* is running correctly, etc.

In **environment mode**, no conserved/displaced predictions are made. The environmental information for all waters from a single protein structure are extracted and placed into output files, and then *Consolv* exits.

Back to Table of Contents

Preparing Coordinate Files for Use with *Consolv*

Application mode

Running *Consolv* in application mode usually requires no preprocessing of PDB files. However, if your PDB file includes explicit hydrogen coordinates, they should be removed prior to running *Consolv*. There is a simple script, called `pdbdehydrogen`, which is included with *Consolv*, and can be used for this purpose. To remove hydrogen coordinates from a PDB file, you can use this script as follows:

```
pdbdehydrogen OldPDBfile > NewPDBfile
```

After this step is completed, you are ready to continue on to the next section: Running *Consolv* in Application Mode.

Test mode

Test mode requires several additional preparation steps. As for application mode, any explicit hydrogen coordinates should be removed from coordinate files before running *Consolv*. Test mode requires two independently-solved structures of the same protein, which must both be in the same coordinate system. Thus, the next step in preparing structures to use with *Consolv* in test mode is to superimpose the two structures into a common coordinate system. In development and testing of *Consolv,* this step was performed using *InsightII* molecular graphics software from Molecular Simulations, Inc. However, any software capable of superimposing the backbone atoms of the two structures and writing a PDB-formatted file will suffice. Since large conformational changes in the protein backbone between the two structures can affect *Consolv*'s identification of conserved waters, a root mean squared deviation of less than 1.0 Å for backbone atoms is desirable.

When running in test mode, *Consolv* tags the active-site atoms of the protein so that active-site results can be analyzed independently of other first-shell water molecules. Identification of the active-site requires that *Consolv* be able to identify the protein atoms and the ligand atoms in the PDB file - this is done using the PDB chain ID. The chain ID is a single-letter identifier found in column 22 of `ATOM` and `HETATM` records of the PDB file. For peptidal ligands, most PDB files will already have independent chain ID's for the protein and the ligand. For ligands composed entirely of HETATMS, however, the ligand may not have a chain ID at all. When *Consolv* is run in test mode, it will prompt you for the chain ID of the protein, and the chain ID of the ligand, and it will use this information to determine the active-site location. If the protein and ligand do not have distinct chain ID's in the PDB files, they must be added using a text editor. When adding chain ID's, it is easiest to use a single chain ID for all protein atoms, and another chain ID for all ligand atoms. The following example shows several `ATOM` and `HETATM` records from a PDB file with the chain ID "P" assigned to protein atoms and the chain ID "I" (for inhibitor) assigned to ligand atoms:

```
ATOM   1485  C   VAL P 186       1.503 -12.869  15.079  1.00 64.80      1DR31641
ATOM   1486  O   VAL P 186       1.111 -14.030  14.731  1.00 67.83      1DR31642
ATOM   1487  CB  VAL P 186       3.742 -11.875  15.901  1.00 63.38      1DR31643
ATOM   1488  CG1 VAL P 186       4.221 -12.801  17.031  1.00 63.87      1DR31644
ATOM   1489  CG2 VAL P 186       4.921 -11.087  15.348  1.00 64.26      1DR31645
TER    1490      VAL P 186                                              1DR31646
HETATM 1491 AP   TAP P 191      21.945   5.238  17.955  1.00 21.66      1DR31647
HETATM 1492 AO1  TAP P 191      21.158   3.995  18.042  1.00 20.05      1DR31648
HETATM 1493 AO2  TAP P 191      21.275   6.284  17.084  1.00 24.33      1DR31649
HETATM 1494 AO5* TAP P 191      22.304   5.950  19.398  1.00 28.20      1DR31650
...
HETATM 1539  NA2 BIO I 198      15.405   4.874   3.681  1.00 50.36      1DR31695
HETATM 1540  C2  BIO I 198      16.061   5.663   4.586  1.00 55.15      1DR31696
HETATM 1541  N1  BIO I 198      15.745   5.471   5.959  1.00 59.31      1DR31697
HETATM 1542  N3  BIO I 198      16.951   6.629   4.244  1.00 55.38      1DR31698
HETATM 1543  C4  BIO I 198      17.664   7.490   5.047  1.00 56.73      1DR31699
```

In this example, the bound Thio-NADP$^+$ (TAP) is considered a part of the protein, and is thus labeled with a chain ID of "P", while the biopterin ligand atoms are labeled with a chain ID of "I". Once the two coordinate files are properly superimposed and chain ID's have been added, they are ready for use by *Consolv* in test mode. Click here to go to the section on Running *Consolv* in Test Mode.

Environment mode

As with application mode, minimal preparation is needed to run *Consolv* in environment mode. Any explicit hydrogen coordinates should be removed from the PDB file as described in the section on preparing to run *Consolv* in application mode. Once any hydrogens have been removed from your PDB file, click here for details on running *Consolv* in environment mode.

Back to Table of Contents

Running *Consolv* in Application Mode

Once a protein coordinate file has been prepared as described above, running *Consolv* in application mode is straightforward, simply execute the following command:

```
Consolv protPDBCode protPDBFile
```

Where `protPDBCode` is the PDB code of the protein to work with, and `protPDBFile` is the full pathname of the file containing the coordinates of the protein. If the protein has not yet been assigned a PDB code, you may use any 4-character combination of digits and letters. The PDB code is included in the output files to identify the water molecules being classified.

If the *Consolv* directory is not in the execution path, you may need to specify the entire path name to the *Consolv* script. Following is an example of running *Consolv* in application mode where the *Consolv* directory is `/usr/local/Consolv`, and the PDB file `pdb3apr.ent` has been copied or linked to the current directory:

```
/usr/local/Consolv/Consolv 3apr pdb3apr.ent
```

You may now use `sumpred` to view the prediction results, as described in the section Output Files Generated by *Consolv*.

Back to Table of Contents

Running *Consolv* in Test Mode

Running *Consolv* in test mode is similar to running in application mode, but there are a few more details which must be included on the command line. Specifically, *Consolv* needs to know the chain ID's of the protein and the ligand. This information is used to find and label the active-site protein atoms and water molecules. For details on assigning chain ID's in the PDB file, see the section on preparing coordinate files for use with *Consolv*.

Consolv can be run in test mode using any two independently-solved structures of the same protein. At least one of the two structures must be a complex between the protein and an inhibitor or other ligand, so that the active-site of the protein may be identified. Only the PDB file for this complex needs to have chain ID's assigned. Once the PDB files are prepared, you can run *Consolv* in test mode using the following command line:

```
Consolv protPDB cplxPDB protFile cplxFile protID ligID
```

Where `Consolv` is the full path to the main *Consolv* script, `protPDB` is the PDB code for the ligand-free protein structure (or the one without chain ID's assigned, if any), `cplxPDB` is the PDB code for the complex structure (which must have chain ID's assigned), `protFile` is the full path name for the protein structure PDB file, `cplxFile` is the full path name for the complex structure PDB file, `protID` is the chain ID for the protein in the complex structure, and `ligID` is the chain ID for the ligand in the complex structure.

The following example should work correctly, assuming that `/usr/local/Consolv` is replaced with the correct pathname for *Consolv*, and the PDB files `pdb2apr.ent` and `pdb3apr.ent` are copied or linked to the current directory:

```
/usr/local/Consolv/Consolv 2apr 3apr pdb2apr.ent pdb3apr.ent E I
```

You may now use `sumpred` to view the prediction results, as described in the section Output Files Generated by *Consolv*.

Back to Table of Contents

Running *Consolv* in Environment Mode

Running *Consolv* in environment mode is identical to running in application mode, except that the keyword **`noscale`** is included as the last command-line parameter. For example, if *Consolv* is installed in `/usr/local/Consolv`, and the PDB file `pdb2apr.ent` is present in the current directory, then you would invoke *Consolv* in environment mode as follows:

```
/usr/local/Consolv/Consolv 2apr pdb2apr.ent noscale
```

You may now use `sumpred` to view the prediction results, as described in the section Output Files Generated by *Consolv*.

Back to Table of Contents

Output Files Generated by *Consolv*

The classification results produced by *Consolv* are split across several output files. To produce a human-readable summary of these results, use the `sumpred` program in the *Consolv* directory. Make sure you are in the same directory as the output files from a *Consolv* run, and then type:

```
/usr/local/Consolv/sumpred PDBcode
```

Replacing `/usr/local/Consolv/` with the directory in which you have installed *Consolv* locally, and `PDBcode` with the PDB code of the protein for which you are classifying water molecules. If you do, for example, `sumpred 2apr`, sumpred will expect the files `2apr.pred` and `2apr.act`, both of which are produced by *Consolv*, to be present in the current directory.

Sumpred will display a summary to your screen. You may redirect this output to a file in the usual way. For example:

```
     sumpred 2apr > 2apr.consolv
```

The details included in the summary include Residue number of each water, its predicted classification (`Disp` = displaced, `Cons` = conserved), and the number of votes that were cast for each class, which can act as a measure of confidence in each prediction. For details on voting and the classification process, see the section on Algorithmic Details. If *Consolv* was run in test mode, then the summary will also include a comparison of *Consolv*'s predictions with the conserved/displaced status of each water as observed by superimposing the two structures and comparing water molecule positions.

During a run, *Consolv* produces various output files. Files which are necessary for `sumpred` include `*.pred`, `*.cons`, `*.env`, and `*.actsite.hits.` Other files contain intermediate results, or auxiliary information about the protein. A summary of *Consolv* output files follows. Filenames for these additional output files are often of the form `PDBcode.*`, where `PDB` is the PDB code of the protein which produced this output file, or `PDBfile.*`, where `PDBfile` is the filename of the PDB file which was used as input to *Consolv.*

Filename	Modes	Contents
`PDBcode.env`	All	The environmental parameters for all waters in the structure.
`PDBfile.hits`	All	The PDB records of all the water molecules in the structure.
`PDBcode.scl`	Application, Test	The environmental parameters for all **first-shell** water molecules in the structure, scaled over the range [1.0 - 10.0].
`PDBcode.pred`	Application, Test	The conserved/displaced predictions for all of the waters in the structure. This file is not formatted for human analysis. Use **`sumpred`** to summarize the prediction results.
`PDBcode.active.hits`	Test	The PDB records of all the active-site water molecules in the structure.
`PDBcode.cons`	Test	The actual conserved/displaced status of each water in the structure, as determined by superimposing the two structures and comparing water positions. 0=displaced, 1=conserved.

Back to Table of Contents

Diagnostics: Warnings and Error Messages

The most common warning message seen when running *Consolv* is similar to the following:

```
WARNING - Unknown atom type!
Can't find AHPhil value for...
Atom: NH2A Residue: ARG

Using average AHP value of 0.351000 for: N
```

This indicates that *Consolv* is unsure of the exact value to use for the atomic hydrophilicity of an atom of type NH2A in the context of an ARG residue. For most protein atoms, this often means that an unusual nomenclature was used for this atom in the PDB file. This warning is more common for HETATMs, since only a small subset of the many types of HETATMs which can appear in a PDB file are known to *Consolv*.

The atomic hydrophilicity values for every atom type known to *Consolv* are stored in the file `ahplist.dat` This file is probably installed in the same directory as the rest of *Consolv*, but it can be installed elsewhere, as described in the section Getting and Installing *Consolv*. If you wish to add additional atom types to this file, you may edit the file as described in the next section, Input Files and Adjustable Parameters. If *Consolv* does not find an exact match for the atom and residue in `ahplist.dat` , it will use the average hydrophilicity value for all atoms of the same type. In the example above, an unknown type of nitrogen atom is assigned a hydrophilicity value equal to the average value for all types of nitrogen atoms. The average values to use are also assigned in the file `ahplist.dat` .

Back to Table of Contents

Input Files and Adjustable Parameters

Other than the PDB files it is currently processing, *Consolv* only references one additional datafile. This file is the `ahplist.dat` file mentioned in the previous section. This file contains the atomic hydrophilicity values for many common protein atom types and HETATMs in the context of various residues. This file may be edited to include new atom types and alternate names for existing atom types. Each entry of this file has the following format:

```
ATOM    RESIDUE    VALUE
```

`ATOM` is the name of the atom, exactly as it appears in the PDB file ATOM or HETATM record. Similarly, `RESIDUE` is the residue name, however the symbol `*` may be used to match **any** residue name. The `VALUE` is the atomic hydrophilicity value to assign to atoms of this type.

Back to Table of Contents

Other Useful Shell Scripts and Tools Provided with *Consolv*

Several additional tools are provided in the `bin` subdirectory of the *Consolv* installation directory.

`extract`

This script can be used to extract a subset of the available features from a `.env` file produced by *Consolv*. Usage is:

```
extract feature-list envfile
```

Where `feature-list` may include any or all of the following tags:

TAG	Feature
`adn`	Atomic Density
`ahp`	Atomic Hydrophilicity
`bval`	B-Value
`hbdp`	# Hydrogen-bonds to protein atoms
`hbdw`	# Hydrogen-bonds to other water molecules
`mob`	Mobility
`nbval`	Net B-value of neighboring protein atoms
`abval`	Average B-value of neighboring protein atoms
`cons`	Conserved/Displaced tag
`as`	Active-site tag

For example, to extract only atomic density and atomic hydrophilicity information from the `.env` file for the aspartic protease 2APR, you would use:

```
extract ahp adn 2apr.env
```

consolv2pdb

This script takes *Consolv*'s conserved/displaced predictions, and produces two PDB-formatted files - one containing all the waters predicted by *Consolv* to be conserved, and the other containing all the waters predicted to be displaced. This can be useful for visualizing *Consolv*'s prediction results via molecular graphics software. The program is executed as follows:

```
consolv2pdb PDBcode
```

The script expects the files `PDBcode.pred`, `PDBcode.env`, and `pdb<PDBcode>.ent` to be found in the current directory. It produces two files: `PDBcode.preddisp`, and `PDBcode.predcons`. Both are PDB-formatted files containing only water molecules, and should be viewable by most molecular graphics software.

Back to Table of Contents

Algorithmic Details of *Consolv*

Several papers have been published describing the classification algorithm used by *Consolv* to predict water site conservation, the data used to train the classifier, and results of cross-validation testing of the accuracy of the classifier. For further details on *Consolv*, please see the following references:

M. L. Raymer, P. C. Sanschagrin, W. F. Punch, S. Venkataraman, E. D. Goodman, and L. A. Kuhn. Predicting conserved water-mediated and polar ligand interactions in proteins using a k-nearest-neighbor genetic algorithm. *J. Mol. Biol.*, 265:445-464, 1997.

M. L. Raymer, W. F. Punch, E. D. Goodman, P. C. Sanschagrin, and L. A. Kuhn. Simultaneous Feature Scaling and Selection Using a Genetic Algorithm. *Proc. Seventh Int. Conf. Genetic Algorithms (ICGA)*. Th. Baeck, ed. Morgan Kaufmann Publishers, San Francisco, 561-567, 1997.

Back to Table of Contents

Contact Information

Inquiries, bug reports, etc. should be directed to Dr. Leslie Kuhn at the following email address:

kuhn@agua.bch.msu.edu