vanddraabe: Thrombin Conserved Waters

Emilio Xavier Esposito

2017-08-10

Introduction

The ability to identify conserved waters from a collection of related protein structues is important for gaining a better understanding of the ligand binding environment. The `vanddraabe` package is based on the work of Sanschagrin and Kuhn (*Protein Science*, 1998, **7** (*10*), pp 2054-2064. DOI: 10.1002/pro.5560071002 (http://doi.org/10.1002/pro.5560071002)) and Patel, Gruning, Gunther, and Merfort (*Bioinformatics*, 2014, **30** (*20*), pp 2978-2980. DOI: 10.1093/bioinformatics/btu424 (http://doi.org/10.1093/bioinformatics/btu424)). Expanding on WatCH and PyWATER, `vanddraabe` and returns statistical parameters for each water cluster, informative graphs, and a PyMOL session file to visually explore the conserved waters and protein along with intermediate information.

This vignette demonstrates the steps and thought process to analyze the conserved waters of ten thrombin crystallographically determined structures and is based on the work of Sanschagrin and Kuhn (*Protein Science*, 1998, **7** (*10*), pp 2054-2064). The results presented herein are part of the original vanddraabe article. There are six main steps to determine conserved water within a collection of protein structures:

- Download PDB structures from the RCSB (http://www.rcsb.org)
- Determine quality of structures and retain those meeting specified requirements
- Clean PDB structures
- Align structures
- Determine conserved waters
- Analyze and visualize results

Before identifying and analyzing conserved waters, several R packages need to be loaded in addition to `vanddraabe`. To aid consistency, the `filename` prefix needs to be defined. For this example `"thrombin10"` is being used. The files (PDB, Excel workbook, and PyMOL session files) and directories (folders) generated during this analysis will start with thrombin10 while the files will also include a date-time stamp to differentiate results.

```
library(vanddraabe)
library(bio3d)
library(reshape2)
library(ggplot2)
library(cowplot)

thrombin10.filename <- "thrombin10"
```

Download PDB structures

```
thrombin10.PDBids <- c("1hai", "1abj", "1ppb", "1tmb", "1hah",
                       "1tmt", "1abi", "1thr", "1ths", "1ihs")

thrombin10.PDBs <- get.pdb(ids=thrombin10.PDBids, split=FALSE, path="thrombin10_rawPDBs")
```

Determine structure quality

The quality of the structures impacts the results of the conserved water analysis. Often the resolution is used to define the quality of the structure but $R_{observed}$ and R_{free} should also be taken into consideration. Smaller resolution values indicate a greater confidence in the location of atoms. Protein structures with reported resolution values greater than or equal to 3.0 Angstroms illustrate the basic contours of the protein chain and thus the atomic structure of the backbone and sidechains is inferred. The $R_{observed}$ – also known as R-value Observed – value indicates how well the “modeled” atoms of the protein structure match the electron density maps with values of 0.20 or less being typical. The corresponding R_{free} value is how well a held-out collection of 5-10% of the atoms were fit; values of 0.26 or less are considered acceptable. In `vanddraabe` structures are evaluated using any combinate of the resolution, $R_{observed}$, and R_{free}. Not all structures have $R_{observed}$ and R_{free} values reported. Only the resolution values are provided for the thrombin example presented here.

```
thrombin10.rcsbCLEANING <- getRCSBdata(prefix = "./thrombin10_rawPDBs",
                                       resolution = 3.0,
                                       rFree = NULL,
                                       rObserved = NULL,
                                       filename = thrombin10.filename)
## Please be patient... Getting PDB information from www.rcsb.org
## The R-observed cutoff is set to "NULL" and is not a factor in evaluating structures for removal.
## The R-free cutoff is set to "NULL" and is not a factor in evaluating structures for removal.
##
## ----- getRCSBdata SUMMARY _____
## getRCSBdata is DONE!
## RCSB information for each PDB structure was written to the Excel workbook: thrombin10_DATA_RESULTS.xlsx
## All structures (10) PASSED the structure evaluation requirements and were copied to the "thrombin10_RCSB_passe
d" folder.
```

PDB structures with values greater than those provided are removed from further analysis. To remove a structural evaluation from the RCSB cleaning provide `NULL`. If no `resolution` value is provided the `CleanRCSBdataset` will automatically use `3.0`. The following information is returned:

- the RCSB (http://www.rcsb.org) information for each PDB structure (along with being provided in an Excel workbook)
- the RCSB (http://www.rcsb.org) information for the PDB structures *passing* the quality requirements
- the RCSB (http://www.rcsb.org) information for the PDB structures ***not*** *passing* the quality requirements

All ten thrombin structures passed the 3.0 Angstrom resolution requirement. The following table contains some structural information for the thrombin structures.

Portion of the RCSB Structural Information

	chainId	resolution	experimentalTechnique	source	citation	depositionDate
1ABI	H,I,L	2.3	X-RAY DIFFRACTION	Homo sapiens	Qiu et al. Biochemistry (1992)	1992-08-24
1ABJ	H,L	2.4	X-RAY DIFFRACTION	Homo sapiens	Qiu et al. Biochemistry (1992)	1992-08-24
1HAH	H,I,L	2.3	X-RAY DIFFRACTION	Homo sapiens	Vijayalakshmi et al. Protein Sci. (1994)	1994-06-27
1HAI	H,L	2.4	X-RAY DIFFRACTION	Homo sapiens	Vijayalakshmi et al. Protein Sci. (1994)	1994-06-27
1IHS	H,I,L	2.0	X-RAY DIFFRACTION	Homo sapiens	Zdanov et al. Proteins (1993)	1993-08-04
1PPB	H,L	1.92	X-RAY DIFFRACTION	Homo sapiens	Bode et al. EMBO J. (1989)	1991-10-24
1THR	H,I,L	2.3	X-RAY DIFFRACTION	Homo sapiens	Qiu et al. J.Biol.Chem. (1993)	1993-06-16
1THS	H,I,L	2.2	X-RAY DIFFRACTION	Homo sapiens	Qiu et al. J.Biol.Chem. (1993)	1993-06-16
1TMB	H,I,L,T	2.3	X-RAY DIFFRACTION	Homo sapiens	Maryanoff et al. Proc.Natl.Acad.Sci.USA (1993)	1993-05-27
1TMT	H,I,J,L	2.2	X-RAY DIFFRACTION	Homo sapiens	Priestle et al. Protein Sci. (1993)	1994-05-26

Clean PDB structures

Protein structures from the RCSB (www.rcsb.org) commonly do ***not*** contain hydrogen atoms but there are the rare occurrence of hydrogen atoms being added by the depositing authors. Often atoms will be modeled – added by the crystallographer and/or crystallographic software – when there is not enough electron density to resolve the atom. This is common when a portion of the amino acid residue is resolved and based on the protein's sequence the "missing" portion of the residue is know. Atoms are also removed when they are assigned a B-value or occupancy value outside the normal range. B-values have a range of 0 - 100 (0 is no variation in position and 100 being diffuse; values less than 40 are considered optimal) and occupancy values have a range of 0 to 1.0 (0 being no occupancy and 1.0 being present in all reflections; values greater than 0.90 are considered optimal).

PDB files obtained from the PDB conform to a specific set of formatting standards but this does not mean the data within the PDB files is always correct. This function *cleans* the PDB file and summaries the atom evaluations. This function does the following (in this order):

- Reads in the PDB file
- Adds/updates the element symbol using the atom type
- Removes hydrogen atoms
- Removes atoms with occupancy values determined to be out of range

- Removes atoms with B-values determined to be out of range
- Bins (counts) the occupancy values
- Bins (counts) the B-values
- Bins (counts) the normalized B-values
- Bins (counts) the mobility values
- Removes modeled atoms
- Removes water oxygen atoms greater than user defined value `cutoff.prot.h2o.dist` from the protein
- Writes cleaned protein structure to a PDB file

```
thrombin10.CLEANED <- CleanProteinStructures(prefix = "./thrombin10_RCSB_passed",
                                             CleanHydrogenAtoms = TRUE,
                                             CleanModeledAtoms = TRUE,
                                             cutoff.prot.h2o.dist = 6.0,
                                             cleanDir = thrombin10.filename,
                                             filename = thrombin10.filename)
## Cleaning 1abi...
##   HEADER    HYDROLASE/HYDROLASE INHIBITOR           24-AUG-92   1ABI
##  - 241 of the 246 water oxygen atoms are within 6 Angstroms of the protein
##  - Wrote 1abi_cleaned.pdb to ./thrombin10_CLEANED
## Cleaning 1abj...
##   HEADER    HYDROLASE/HYDROLASE INHIBITOR           24-AUG-92   1ABJ
##  - 192 of the 196 water oxygen atoms are within 6 Angstroms of the protein
##  - Wrote 1abj_cleaned.pdb to ./thrombin10_CLEANED
## Cleaning 1hah...
##   HEADER    COMPLEX(SERINE PROTEINASE/INHIBITOR)    27-JUN-94   1HAH
##    PDB has ALT records, taking A only, rm.alt=TRUE
##  - Removed modeled atoms
##  - 204 of the 204 water oxygen atoms are within 6 Angstroms of the protein
##  - Wrote 1hah_cleaned.pdb to ./thrombin10_CLEANED
## Cleaning 1hai...
##   HEADER    HYDROLASE/HYDROLASE INHIBITOR           27-JUN-94   1HAI
##    PDB has ALT records, taking A only, rm.alt=TRUE
##  - Removed modeled atoms
##  - 194 of the 194 water oxygen atoms are within 6 Angstroms of the protein
##  - Wrote 1hai_cleaned.pdb to ./thrombin10_CLEANED
## Cleaning 1ihs...
##   HEADER    HYDROLASE/HYDROLASE INHIBITOR           04-AUG-93   1IHS
##  - 146 of the 146 water oxygen atoms are within 6 Angstroms of the protein
##  - Wrote 1ihs_cleaned.pdb to ./thrombin10_CLEANED
## Cleaning 1ppb...
##   HEADER    HYDROLASE/HYDROLASE INHIBITOR           24-OCT-91   1PPB
##  - 333 of the 402 water oxygen atoms are within 6 Angstroms of the protein
##  - Wrote 1ppb_cleaned.pdb to ./thrombin10_CLEANED
## Cleaning 1thr...
##   HEADER    HYDROLASE(SERINE PROTEINASE)            16-JUN-93   1THR
##  - Removed modeled atoms
##  - 190 of the 190 water oxygen atoms are within 6 Angstroms of the protein
##  - Wrote 1thr_cleaned.pdb to ./thrombin10_CLEANED
## Cleaning 1ths...
##   HEADER    HYDROLASE/HYDROLASE INHIBITOR           16-JUN-93   1THS
##  - 140 of the 140 water oxygen atoms are within 6 Angstroms of the protein
##  - Wrote 1ths_cleaned.pdb to ./thrombin10_CLEANED
## Cleaning 1tmb...
##   HEADER    HYDROLASE/HYDROLASE INHIBITOR           27-MAY-93   1TMB
##  - 229 of the 239 water oxygen atoms are within 6 Angstroms of the protein
##  - Wrote 1tmb_cleaned.pdb to ./thrombin10_CLEANED
## Cleaning 1tmt...
##   HEADER    HYDROLASE/HYDROLASE INHIBITOR           26-MAY-94   1TMT
##  - 111 of the 111 water oxygen atoms are within 6 Angstroms of the protein
##  - Wrote 1tmt_cleaned.pdb to ./thrombin10_CLEANED
## ----- Results written to Excel workbook _____
```

The information returned from cleaning the protein structure are:

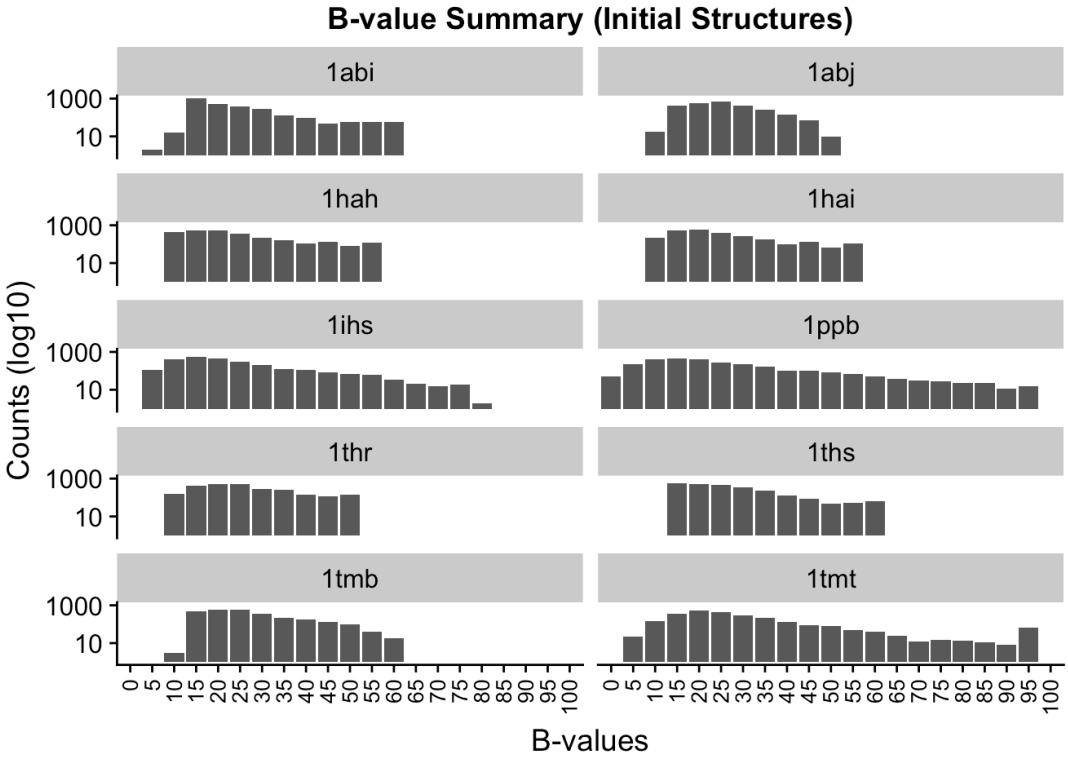
- **cleaning.summary**: summary indicating

    - if hydrogen atoms were removed `TRUE/FALSE`
    - number of out of range atoms for B-values and occupancy values
    - number of modeled (and thus removed)
    - number of atoms ***NOT*** modeled (and thus retained)
    - number of water oxygen atoms beyond the user defined cutoff
    - the number of water oxygen atoms within the user defined cutoff.
- **Bvalue.counts**: binned B-value values with binwidths = 5 (0 to 100)
- **normBvalue.counts**: binned normalized B-value values with binwidths = 0.1 (-4 to 6)
- **occupancy.counts**: binned occupancy values with binwidths = 0.1 (0 to 1)
- **mobility.counts**: binned mobility values with binwidths = 0.1 (0 to 6)
- **call**: parameters provided by the user

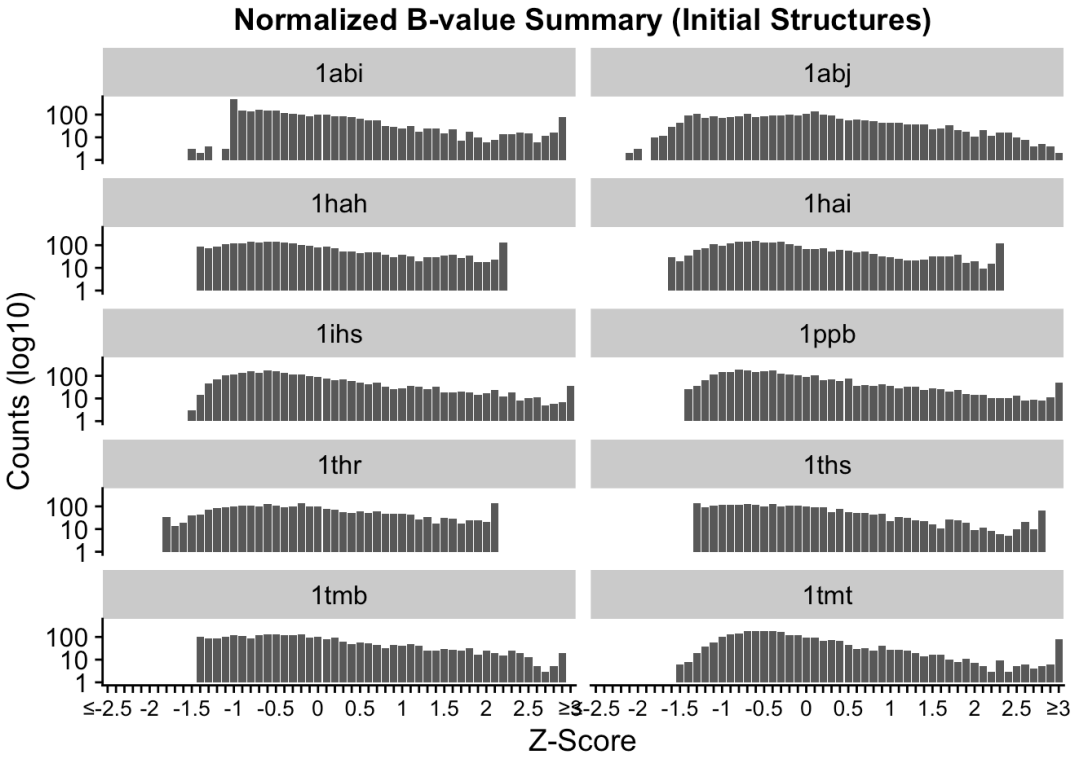
Cleaning Summary

	removedHydrogens	num.o.OoR	num.b.OoR	num.Modeled	num.notModeled	num.WatersDistantRemoved	num.WatersRetained
1abi	FALSE	0	0	0	2703	5	241
1abj	FALSE	0	0	0	2531	4	192
1hah	FALSE	0	0	63	2590	0	204
1hai	FALSE	0	0	54	2524	0	194
1ihs	FALSE	0	0	0	2561	0	146
1ppb	FALSE	0	48	0	2771	69	333
1thr	FALSE	0	0	15	2526	0	190
1ths	FALSE	0	0	0	2529	0	140
1tmb	FALSE	0	0	0	2644	10	229
1tmt	FALSE	0	0	0	2526	0	111

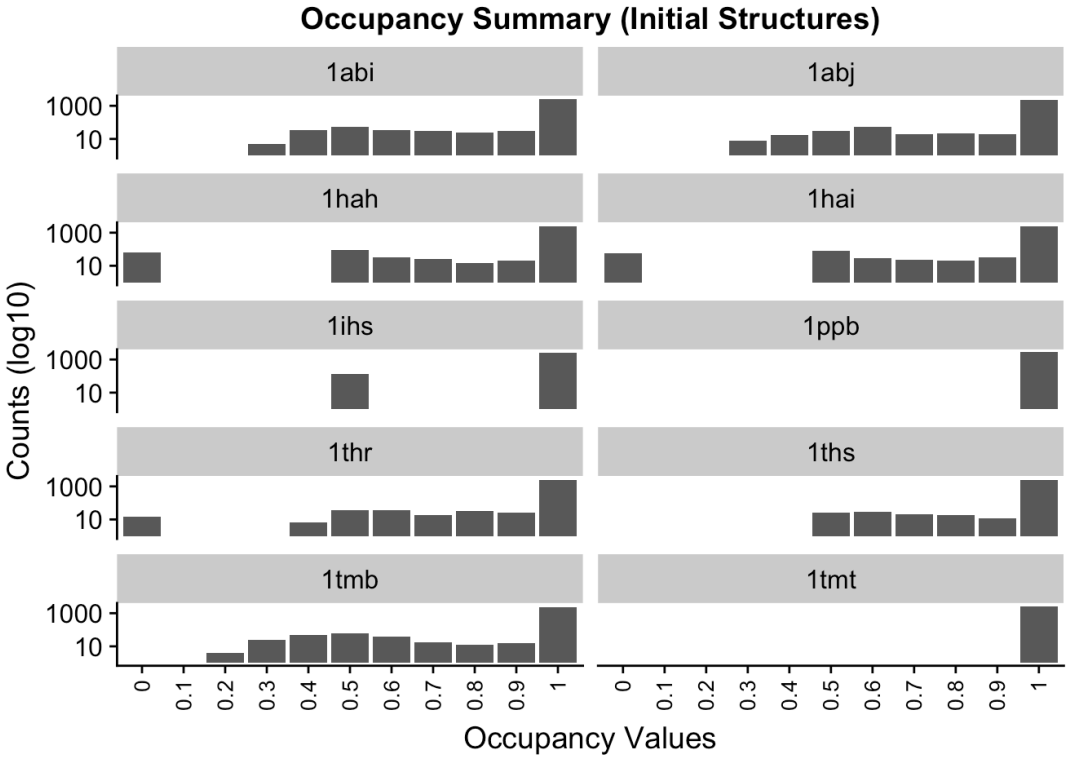
The `cleaning.summary` is a table shows there were no thrombin structures with occupancy values outside the normal range of 0 to 1 but there was a single structure, 1ppb, with 48 atoms assigned B-values outside the normal range of 0 to 100. These 48 atoms were removed from 1ppb. Three structures had atoms with occupancy values of 0.01 or less and these atoms were also removed. Four structures had water oxygen atoms beyond 6 Angstroms from a protein atom and thus these “distant” waters were removed.



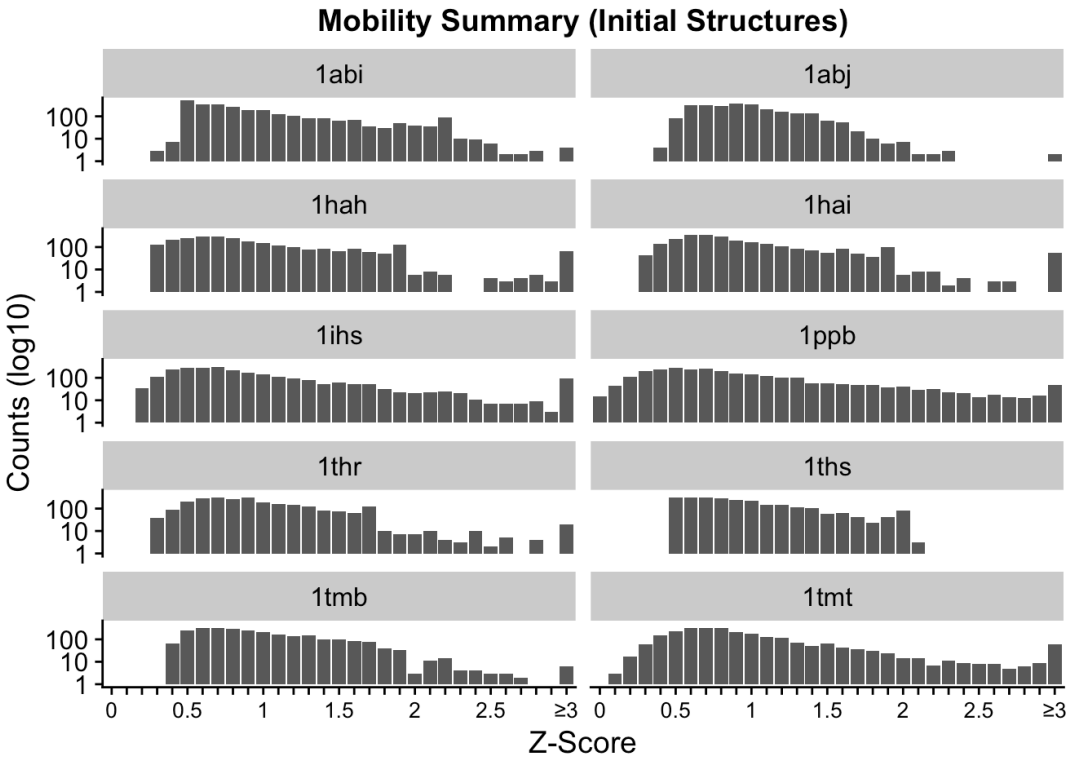
The B-value barplots – before the structure is cleaned – illustrates the difference in quality of structures based on B-values. Three structures (PDBids: 1ihs, 1ppb, and 1tmt) in the Thrombin dataset have atoms with B-values of 65 or greater. Atoms with B-values greater than 60 are considered lower quality because of their greater variance.



Normalizing the B-values provides a way to compare B-values across a collection of protein structures; Z-score values less then 0 indicate atoms with B-values less than the mean B-value for a structure while Z-score values greater than 0 indicates and atom with B-values greater than the mean B-value. Within vanddraabe, water atoms with normalized B-values greater than 1.0 are removed from analysis. The inclusion of all atoms within the protein structures indicates the overall quality of the atoms within the structure. Normalized B-values are calculated for protein, non-protein (ligands), and water atoms separately during the evaluation of atoms prior to determining conserved waters.



While a majority of the atoms in the ten protein structures have occupancy values of 1.0, only two structures (PDBids: 1ppb and 1tmt) possess atoms present in ***all*** copies of the protein. Interestly, these two protein structures have a wider range of atom position variance. Three structures (PDBids: 1hah, 1hai, and 1thr) have modeled atoms – those with an occupancy values of 0.01 or less. The percentage of atoms with less then perfect occupancy for these seven structures is less than XX%.



Mobility values take into consideration the occupancy and B-values to evaluate the quality of individual atoms. Atoms with mobility values greater than 2.0 are considered lower quality. Like the normalized B-values, the mobility values are calculated for protein, non-protein (ligands), and water atoms independently during the evaluation of atoms prior to determining conserved waters.

Align structures

The alignment of the proteins' primary sequence needs to be performed before the protein structures can be aligned. Using the "cleaned" structures, align the sequences.

```
thrombin10.PDBs.passed <- paste("./thrombin10_CLEANED", list.files(path="./thrombin10_CLEANED/", pattern="*.pdb")
, sep="/")

thrombin10.PDBs.align <- pdbaln(thrombin10.PDBs.passed, fit=TRUE)
```

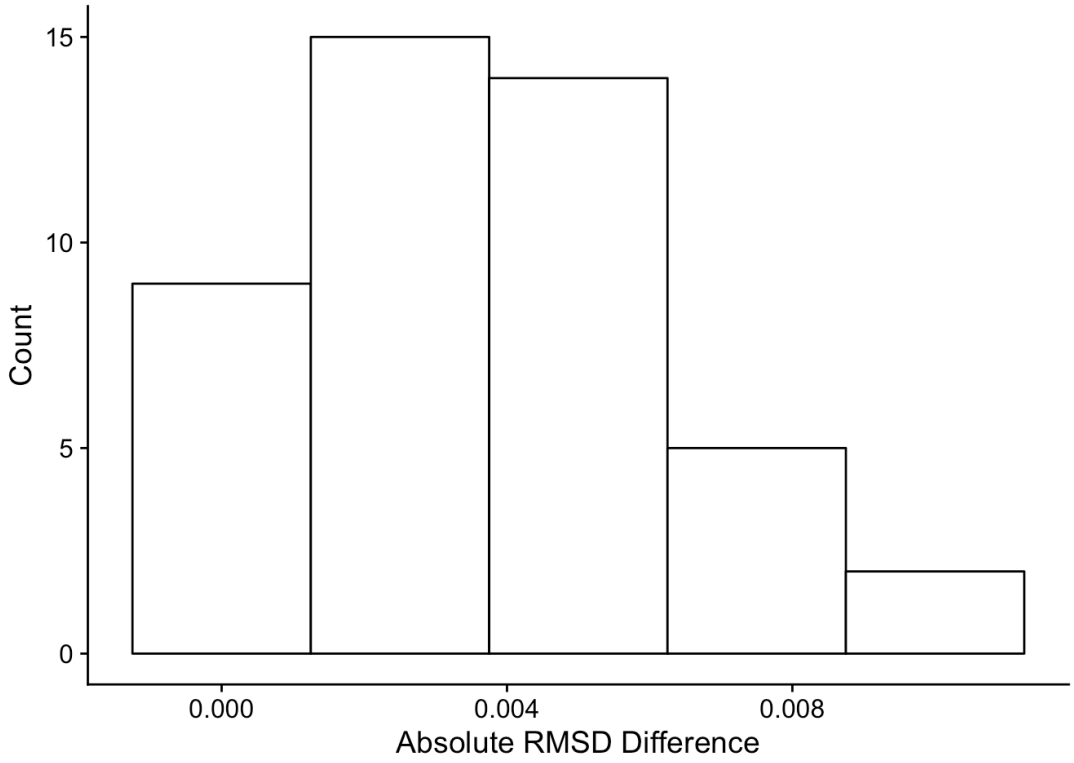
Once the sequences are aligned, align the protein structures. The `stop.vol` is the "minimal core volume" to stop the structural alignment.

```
thrombin10.core <- core.find(thrombin10.PDBs.align, write.pdbs=TRUE, rm.island=TRUE,
                             stop.vol=0.01, outpath="thrombin10_coreFind")
thrombin10.core.2.0ang.idc <- print(thrombin10.core, vol=2.0)
thrombin10.core.1.5ang.idc <- print(thrombin10.core, vol=1.5)
thrombin10.core.1.0ang.idc <- print(thrombin10.core, vol=1.0)
thrombin10.core.0.5ang.idc <- print(thrombin10.core, vol=0.5)

thrombin10.xyz.1.0 <- fit.xyz(fixed=thrombin10.PDBs.align$xyz[7, ],
                              mobile=thrombin10.PDBs.align,
                              fixed.inds=thrombin10.core.1.0ang.idc$xyz[res2xyz(c(50:59,61:96,98:128,130:196,206:
220,222:256))],
                              mobile.inds=thrombin10.core.1.0ang.idc$xyz[res2xyz(c(50:59,61:96,98:128,130:196,206
:220,222:256))],
                              outpath="thrombin10_fitlsq_1.0ang",
                              full.pdbs=TRUE, verbose=TRUE)
```

Determine the quality of the alignment

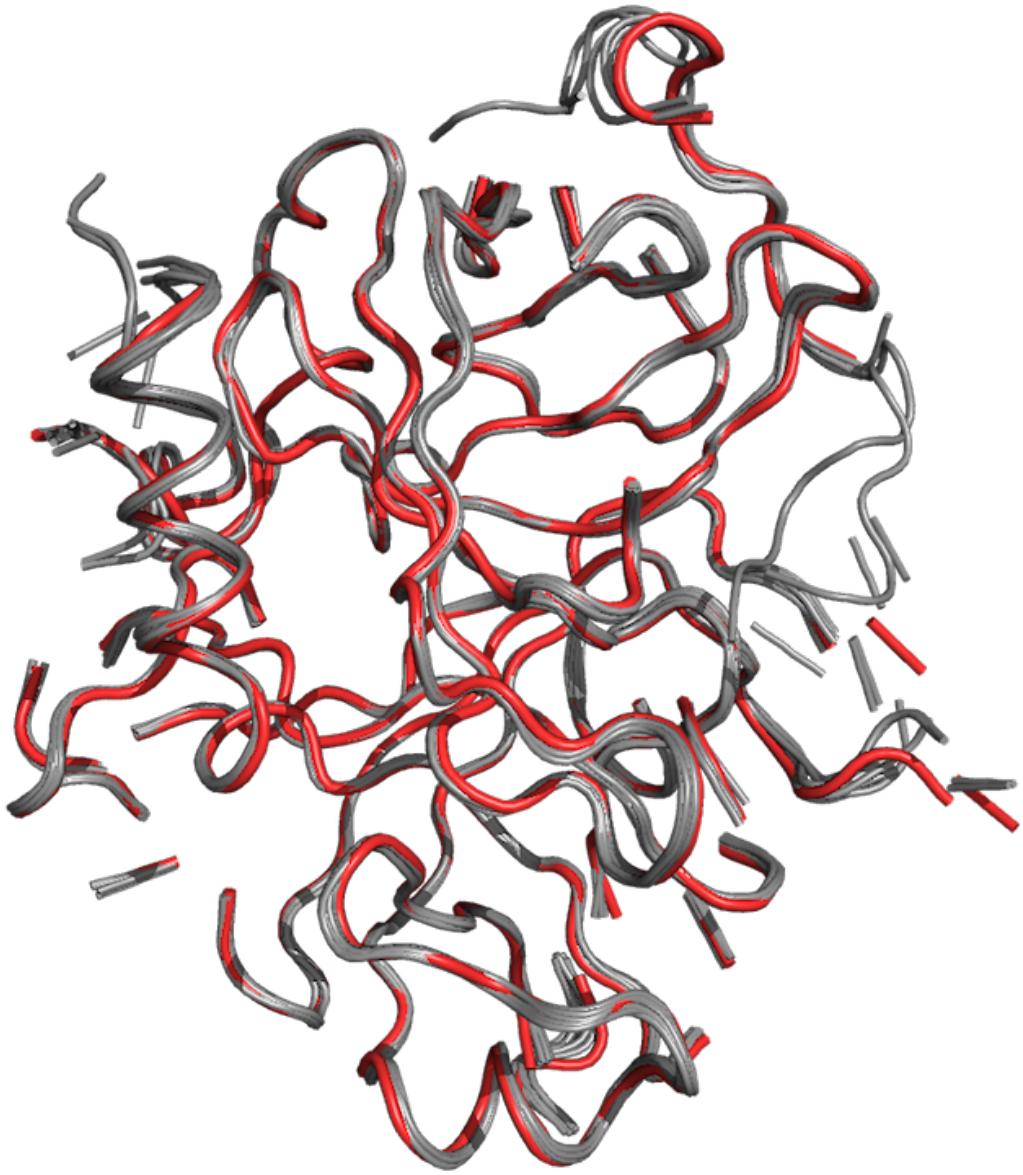
```
rmsd.fit <- rmsd(thrombin10.xyz.1.0, fit=TRUE)
rownames(rmsd.fit) <- colnames(rmsd.fit) <- sort(thrombin10.PDBids)
rmsd.fit.upper <- rmsd.fit[upper.tri(rmsd.fit)]

rmsd.NOfit <- rmsd(thrombin10.xyz.1.0, fit=FALSE)
rownames(rmsd.NOfit) <- colnames(rmsd.NOfit) <- sort(thrombin10.PDBids)
rmsd.NOfit.upper <- rmsd.NOfit[upper.tri(rmsd.NOfit)]


rmsd.diff <- rmsd.fit - rmsd.NOfit
rmsd.diff.upper <- abs(rmsd.diff[upper.tri(rmsd.diff)])


df.rmsd <- data.frame(fit=rmsd.fit.upper,
                      NOfit=rmsd.NOfit.upper,
                      diff=rmsd.diff.upper)
df.rmsd.melt <- melt(df.rmsd)
```

The alignment of the ten thrombin structures was inspected via RMSD. Two different methods of RMSD calculation were used: (*i*) *with* coordinate superposition prior to the RMSD calculation and (*ii*) *without*. The resulting RMSD values have a correlation (r) of 1 and the mean difference between the RMSD values is 0.0037 ($\pm$ 0.0024). The small difference between the fitted and non-fitted RMSD values indicates the 1.0 Angstrom alignment volume of all ten thrombin structures is optimal to determine the conserved waters.



Histogram of Absolute RMSD Differences (RMSD fit - RMSD No Fit)

The RMSD between all the structures is 0.6 with the excpetion of PDBid 1tmt (http://www.rcsb.org/pdb/explore/explore.do?structureId=1tmt) with RMSD values ranging from 1.79 to 2.13 Angstroms. Visualizing the backbone of these ten thrombin structures shows 1tmt (red tube structure) does not vary significantly from the other nine thrombin structures.



Aligned Thrombin Structures

Pairwise RMSD Values (No Fit)

	1abj	1hah	1hai	1ihs	1ppb	1thr	1ths	1tmb	1tmt
1abi	0.6	0.23	0.3	0.27	0.6	0.22	0.26	0.29	1.79
1abj		0.55	0.5	0.54	0.28	0.6	0.6	0.57	2.13
1hah			0.27	0.23	0.54	0.19	0.24	0.32	1.83
1hai				0.26	0.49	0.3	0.34	0.29	1.85
1ihs					0.52	0.27	0.34	0.26	1.84
1ppb						0.59	0.6	0.56	2.12
1thr							0.25	0.34	1.79
1ths								0.35	1.83
1tmb									1.84
1tmt									

Retain only a single aligned structure

Often more than one biological assembly is contained within the PDB files. To reduce the number of waters analyzed, the chains of each aligned protein are compared in 3D space to the chains of a reference structure and chains sharing at least 60% overlap – based on Cα atoms within 3.0 Angstroms of the reference structure – are retained. It is crucial the reference structure is high quality and contains all the structural features of interest (chains) because the other protein structures will be compared and evaluated based on the reference structure.

```
thrombin10.alignOverlap <- AlignOverlap(aligned.dir = "thrombin10_fitlsq_1.0ang",
                                        out.dir = thrombin10.filename,
                                        ref.PDBid = "1thr",
                                        overlap = 0.6, CA.dist = 3.0,
                                        filename = thrombin10.filename)
## thrombin10_alignedGood already exists. Contents will be overwritten or added to.
## thrombin10_alignedPoor already exists. Contents will be overwritten or added to.
##   PASSED -->> 1abi Overlap: 100%  PASSED -->> 1abi Overlap: 52.9%  PASSED -->> 1abi Overlap: 78.8%
##   PASSED -->> 1abj Overlap: 96.5%  PASSED -->> 1abj Overlap: 96.3%
##   PASSED -->> 1hah Overlap: 99.6%  PASSED -->> 1hah Overlap: 88.9%  PASSED -->> 1hah Overlap: 76.5%
##   PASSED -->> 1hai Overlap: 100%  PASSED -->> 1hai Overlap: 78.8%
##   PASSED -->> 1ihs Overlap: 97.3%  PASSED -->> 1ihs Overlap: 47.1%  PASSED -->> 1ihs Overlap: 89.7%
##   PASSED -->> 1ppb Overlap: 96.5%  PASSED -->> 1ppb Overlap: 79.4%
##   PASSED -->> 1thr Overlap: 100%  PASSED -->> 1thr Overlap: 100%  PASSED -->> 1thr Overlap: 100%
##   PASSED -->> 1ths Overlap: 100%  PASSED -->> 1ths Overlap: 100%  PASSED -->> 1ths Overlap: 81.8%
##   PASSED -->> 1tmb Overlap: 100%  PASSED -->> 1tmb Overlap: 88.9%  PASSED -->> 1tmb Overlap: 89.7%
##   PASSED -->> 1tmt Overlap: 97.3%  PASSED -->> 1tmt Overlap: 80%  PASSED -->> 1tmt Overlap: 100%
## ----- Results written to Excel workbook _____
## done!!
```

Aligned Protein Structure Overlap

PDBids	reference	chains.initial	chains.retained	passed	pct.overlap	pct.overlap.min	pct.overlap.max
1abi	FALSE	L, H, I	L, H, I	TRUE	100, 52.941, 78.788	52.9	100
1abj	FALSE	L, H	L, H	TRUE	96.512, 96.296	96.3	96.5
1hah	FALSE	L, H, I	L, H, I	TRUE	99.603, 88.889, 76.471	76.5	99.6
1hai	FALSE	L, H	L, H	TRUE	100, 78.788	78.8	100
1ihs	FALSE	L, H, I	L, H, I	TRUE	97.287, 47.059, 89.655	47.1	97.3
1ppb	FALSE	L, H	L, H	TRUE	96.525, 79.412	79.4	96.5
1thr	TRUE	L, H, I	L, H, I	TRUE	100, 100, 100	100	100
1ths	FALSE	L, H, I	L, H, I	TRUE	100, 100, 81.818	81.8	100
1tmb	FALSE	L, H, I, T	L, H, I	TRUE	100, 88.889, 89.655	88.9	100
1tmt	FALSE	L, H, I, J	L, H, J	TRUE	97.276, 80, 100	80	100

The thrombin protein system is atypical because it has a heavy chain (denoted as chain H), a light chain (denoted as chain L), and an inhibitor chain (denoted at chain I) compared to the standard A, B, and C (the chains can be lettered alphabetically through Z) chain designations. The Aligned Protein Structure Overlap table above indicates if the structure was the reference structure, the initial and retained chains, if the structure passed, the percent overlap, and the minimum and maximum overlap for a chain. Within the reference structure, 1thr, there are three chains (L, H, I) and the chains of the other structures were compared to these chains. Thus if a structure did not have a chain overlapping with chains L, H, or I of 1thr then the non-overlapping chain was removed from the structure.

Determine conserved waters

```
thrombin10.conservedWaters <- ConservedWaters(prefix = "thrombin10_alignedGood/",
                                              cluster = 2.4,
                                              mobility = 2.0, nBvalue = 1.0,
                                              prot.h2o.dist.min = 6.0,
                                              chain = "all",
                                              cluster.method = "complete",
```

```
                                        PDBinfo = thrombin10.rcsbCLEANING$PDB.info.passed,
                                        filename = thrombin10.filename)
## ----- Reading in the aligned structures _____
## Reading structure 1abi...
##  - 241 of the 241 water oxygen atoms are within 6.001 Angstroms of the protein
##  - Structure 1 of 10 -->> 1abi has 241 waters.
## Reading structure 1abj...
##  - 192 of the 192 water oxygen atoms are within 6.001 Angstroms of the protein
##  - Structure 2 of 10 -->> 1abj has 192 waters.
## Reading structure 1hah...
##  - 204 of the 204 water oxygen atoms are within 6.001 Angstroms of the protein
##  - Structure 3 of 10 -->> 1hah has 204 waters.
## Reading structure 1hai...
##  - 194 of the 194 water oxygen atoms are within 6.001 Angstroms of the protein
##  - Structure 4 of 10 -->> 1hai has 194 waters.
## Reading structure 1ihs...
##  - 146 of the 146 water oxygen atoms are within 6.001 Angstroms of the protein
##  - Structure 5 of 10 -->> 1ihs has 146 waters.
## Reading structure 1ppb...
##  - 333 of the 333 water oxygen atoms are within 6.001 Angstroms of the protein
##  - Structure 6 of 10 -->> 1ppb has 333 waters.
## Reading structure 1thr...
##  - 190 of the 190 water oxygen atoms are within 6.001 Angstroms of the protein
##  - Structure 7 of 10 -->> 1thr has 190 waters.
## Reading structure 1ths...
##  - 140 of the 140 water oxygen atoms are within 6.001 Angstroms of the protein
##  - Structure 8 of 10 -->> 1ths has 140 waters.
## Reading structure 1tmb...
##  - 224 of the 224 water oxygen atoms are within 6.001 Angstroms of the protein
##  - Structure 9 of 10 -->> 1tmb has 224 waters.
## Reading structure 1tmt...
##  - 108 of the 108 water oxygen atoms are within 6.001 Angstroms of the protein
##  - Structure 10 of 10 -->> 1tmt has 108 waters.
##
## ----- Summary about imported structures _____
## 10 structures were read for water cluster analysis containing 1972 water molecules.
##   - 10 Strucutures have water molecules.
##   - 0 Strucuture(s) do NOT have water molecules.
##
## ----- Checking for high quality structures using B-value Normalization and Mobility _____
## Water molecules from all imported structures will be used in the water conservation analysis.
##
## ----- Clustering ALL waters from the provided structures _____
## Calculating 1,943,406 pairwise distances for 1,972 water molecules.
## The pairwise distance calculations for 1,943,406 water molecules took 0.03025 seconds and is 15 Mb in size.
## Clustering the individual waters...
## Constructing clusters...
## Constructing table of structures with cluster present...
## Constructing data.frame with cluster information...
## Calculating the mean distance between waters AND between the centroid within each cluster...
## ----- Clustering waters that PASSED the B-value Normalization and Mobility _____
## Calculating 1,357,128 pairwise distances for 1,648 water molecules.
## The pairwise distance calculations for 1,357,128 water molecules took 0.01127 seconds and is 10.5 Mb in size.
## Clustering the individual waters...
## Constructing clusters...
## Constructing table of structures with cluster present...
## Constructing data.frame with cluster information...
## Calculating the mean distance between waters AND between the centroid within each cluster...
## ----- Constructing the summary table _____
## ----- Summary table written _____
## ----- Writing the conserved waters to PDB files _____
## ----- PDB files written _____
## ----- Writing results to Excel workbook _____
## ----- Results written to Excel workbook _____
## ----- Done! _____
```

Analyze results

The conserved water analysis is performed twice: (*i*) all waters present in the provided protein structures and (*ii*) the waters passing the mobility and normalized B-value cutoffs provided by the user. This allows the user to explore the differences in conserved waters based on the quality – and quantity – of the available waters.

The results are available within the R session and are written to the `thrombin10_DATA_RESULTS.xlsx` Excel workbook as a collection of sheets. The name of each sheet for the performed analysis has the date and time (*e.g.*, aug042017_1620 indicates the analysis was performed on August 4, 2017 at 4:20pm). The analysis creates six sheets within the noted workbook. The sheets are:

- **PDBsInfo_all**: The RCSB information for each PDB structure considered for the conserved water analysis.
- **PDBsInfo_pass**: The RCSB information for each PDB structure *passing* the user defined mobility and normalized B-value parameters for the conserved water analysis.
- **PDB_cleanSumm**: The results from cleaning the protein structures with `CleanProteinStructures()`. The number of atoms with "Out of Range" occupancy and B-values, number of modeled (and not modeled) atoms, and number of water molecules (atoms) beyond (and within) the user defined distance are indicated along with removed from the structures. See the Cleaning Summary table above.
- **BvalueBins**: The number of atoms with B-values within specified bins; between 0 and 100 in bins of 5.
- **nBvalueBins**: The number of atoms with normalized B-values within specified bins; between -7 and 7 in bins of 0.10. Values less than -7 are counted in the -7-bin and values greater than 7 are counted in the 7-bin.
- **OccBins**: The number of atoms with occupancy values within specified bins; between 0 and 1.0 in bins of 0.10.
- **MobilityBins**: The number of atoms with mobility values within specified bins; between 0 and 6 in bins of 0.10. Values great than 6 are counted in the 6-bin.
- **PDBid_AliOver**: The results from the `AlignOverlap()` analysis. The reference structure is the PDBid in the sheet name and is indicated within the results. The chains meeting the user defined parameters are retained and written to a new PDB file for each structure.
- **ClusterStats**: The information provided in the Conserved Water Clustering Statistics Summary table below.
- **all_ClustSumm**: Statistical information for each conserved water along with information regarding the conserved waters average environment. This summary based on the conserved water analysis when all available waters are analyzed.
- **all_OccurSumm**: The occurrence summary for waters that passed the user defined mobility and normalized B-value parameters. For each protein structure provided the mean and standard deviation for the experimental occupancy, experimental B-value, mobility, and normalized B-values are included along with the number of waters, number of waters passing the mobility test, number of waters passing the normalized B-value test, the number and percentage of waters passing the user defined cutoffs, and the number of clusters the waters from the protein structure participate. Also included is a heatmap indicating if the protein structure contributed to a conserved water cluster.
- **pass_ClustSumm**: Statistical information for each conserved water along with information regarding the conserved waters average environment. This summary based on the conserved water analysis when all available waters are analyzed.
- **pass_OccurSumm**: The occurrence summary for waters that passed the user defined mobility and normalized B-value parameters. For each protein structure provided the mean and standard deviation for the experimental occupancy, experimental B-value, mobility, and normalized B-values are included along with the number of waters, number of waters passing the mobility test, number of waters passing the normalized B-value test, the number and percentage of waters passing the user defined cutoffs, and the number of clusters the waters from the protein structure participate. Also included is a heatmap indicating if the protein structure contributed to a conserved water cluster.
- **InitWaterData**: The initial data for the waters provided from the conserved water analysis along with calculated parameters and the conserved cluster the water contributes to.

Waters with 80% conservation or greater can be considered highly important.

The PDB file containing the conserved waters is written to visualize the results in your favorite molecular visualization package or with the automatically generated PyMOL script (below). The X, Y, and Z coordinates are the mean of the waters' X, Y, and Z atomic positions. The B-values for each conserved water are calculated using the amount of fluctation within the conserved water cluster – determined using the [bio3d::rmsf()] function – and indicate how far the waters are from the cluster's centriod. The occupancy values are the percent conservation for the conserved water and represents the fraction of initial protein structures contributing a water to that cluster. A conserved water with a low B-value (5) and a low occupancy value (0.40) indicates a conserved water composed of waters from 40% of the protein structures that are very close together in 3D space. The opposite is possible of a conserved water with a large B-value (70) and a high occupancy value (1.00) indicated all the protein structures contributed to this conserved water but the waters are more dispersed in 3D space.

Conserved Water Clustering Statistics Summary

	values	percentages	values	percentages
Number of structures	1.000e+01		1.000e+01	
Number of initial waters	1.972e+03		1.972e+03	

Number of used waters	1.972e+03		1.648e+03	
Number of water clusters	7.000e+02		5.970e+02	
Average conservation	2.817e+00		2.760e+00	
<50%	5.710e+02	81.5714285714286	4.860e+02	81.4070351758794
50-69%	5.200e+01	7.42857142857143	4.600e+01	7.70519262981574
70-79%	1.200e+01	1.71428571428571	1.200e+01	2.01005025125628
80-89%	2.200e+01	3.14285714285714	1.400e+01	2.34505862646566
90-99%	1.500e+01	2.14285714285714	1.600e+01	2.68006700167504
100%	2.800e+01	4	2.300e+01	3.85259631490787
Number of pairwise distances	1.943e+06		1.357e+06	
Memory size of pairwise distances (Mb)	1.500e+01		1.050e+01	
Pairwise distance calc time (s)	3.020e-02		1.130e-02	
Cluster centroid distance calcs time (s)	2.137e-01		1.784e-01	

Visualize results

To aid in the analysis of the conserved waters the ability to generate informative plots and PyMOL sessions are provided.

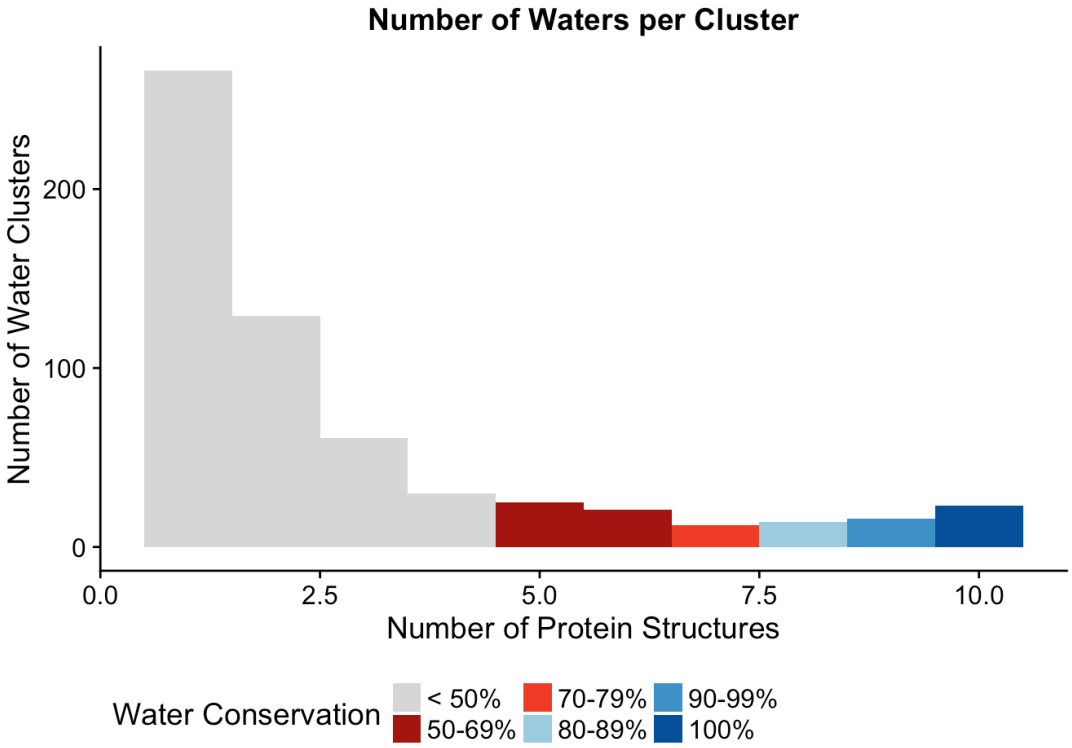
Plots

Being able to visualize the statistical values and the location of the conserved waters in relation to a representative protein structure greatly improves the comprehension of the conserved waters. The plots related to percent conservation share the same color scheme. A water within a cluster with less than 50% conservation are colored light grey, clusters with 50 to 69% water conseration are dark red, clusters with 70 to 79% conservation are red, 80 to 89% conservation are light blue, 90 to 99% conservation are blue, and 100% conservation (waters from all structures) are dark blue. Only some of the plots include conserved waters with less than 50% conservation for clarity. The occupancy, mobility, B-value, and normalized B-value plots are a barplot layered on top of a density plot. The bars provide an exact count while the densities indicate an approximate trend of the data. Ignore the y-axis for the barplot plus density plots.

A collection of plots can be automatically generated to illustrate the:

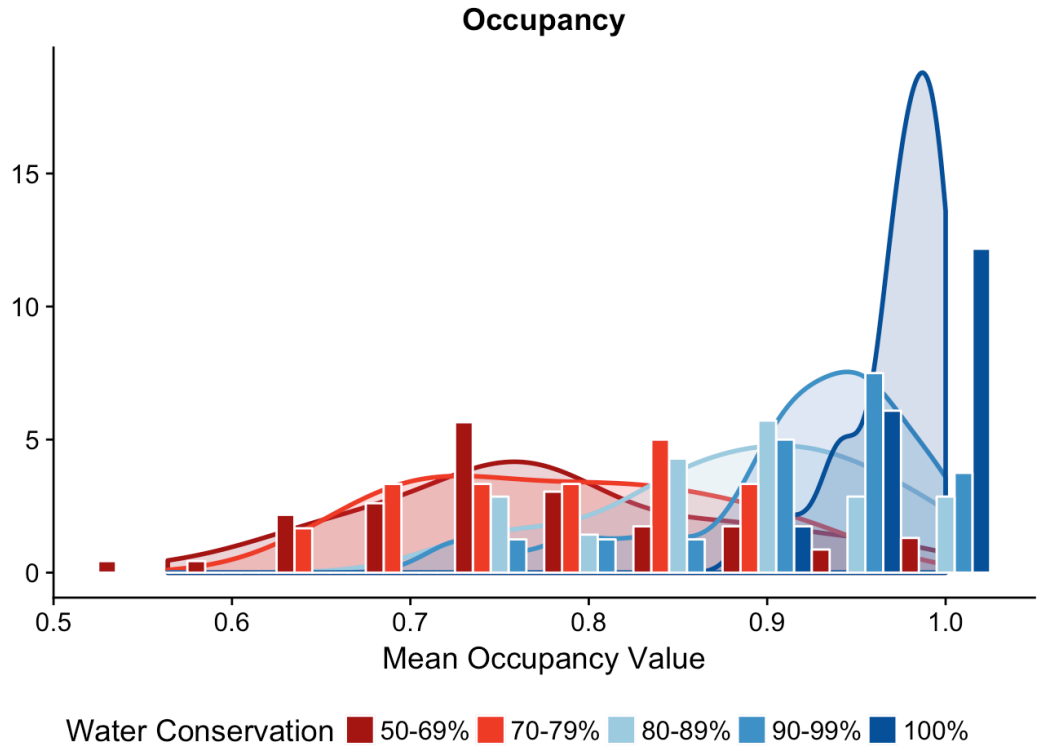
- **Number of waters per cluster**: A barplot displaying the number of water clusters for the number of waters in each water cluster. This histogram indicates there are 23 conserved waters with 100% conservation; all ten protein structures contribute a water to 23 conserved water clusters. There are 470 conserved waters with contributions from less than half of the provided structures. It is expected for there to be more conserved water clusters with less than 50% conservation than conserved waters with a water from all structures.

```
ConservationPlot(data = thrombin10.conservedWaters, passed.waters = TRUE)
```



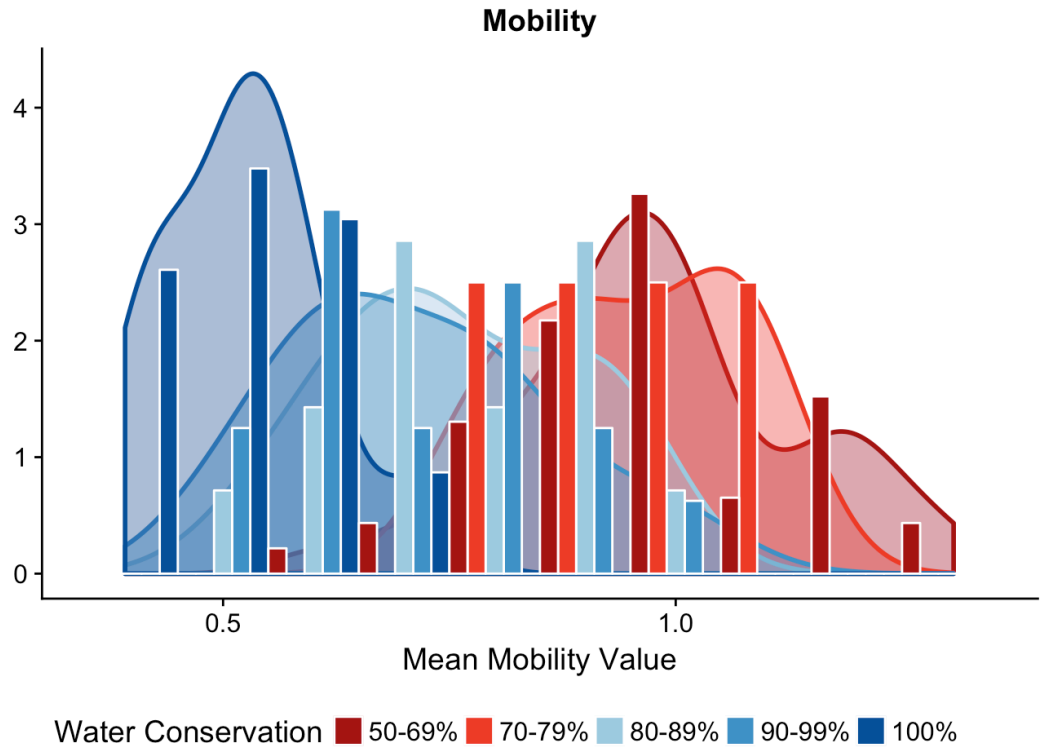
- **Occupancy Barplots**: This plot depicts the amount and distribution of water occupancy values for all the structures within the analysis and are binned based on the percent conservation. Only waters with 50% conservation or greater are displayed. Waters with 80% or greater conservation have a mean occupancy value of 0.7 meaning the waters comprising a conserved water are present in 70% or greater of the reflections used to construct the 3D structures of the proteins.

```
OccupancyBarplot(data = thrombin10.conservedWaters, passed.waters = TRUE)
```



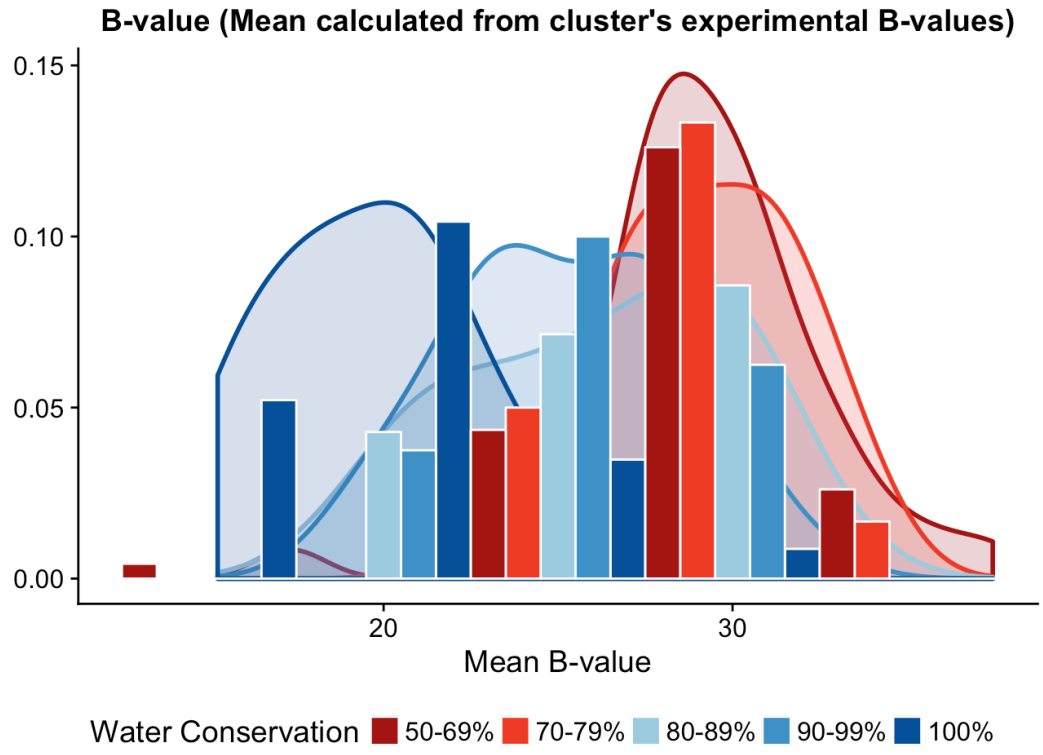
- **Mobility Barplots**: The mobility – calculated for each water in the analysis – is a way to provide a single measure of a water's quality using the occupancy and B-value. Values closer to zero are considered ideal while waters with mobility values greater than 2.0 are often removed from analysis. The mobility plot below illustrates how the mobility and percent conservation are inversely related; highly conserved waters have low mean mobility values and the mean mobility values increase as percent conservation decreases.

```
MobilityBarplot(data = thrombin10.conservedWaters, passed.waters = TRUE)
```

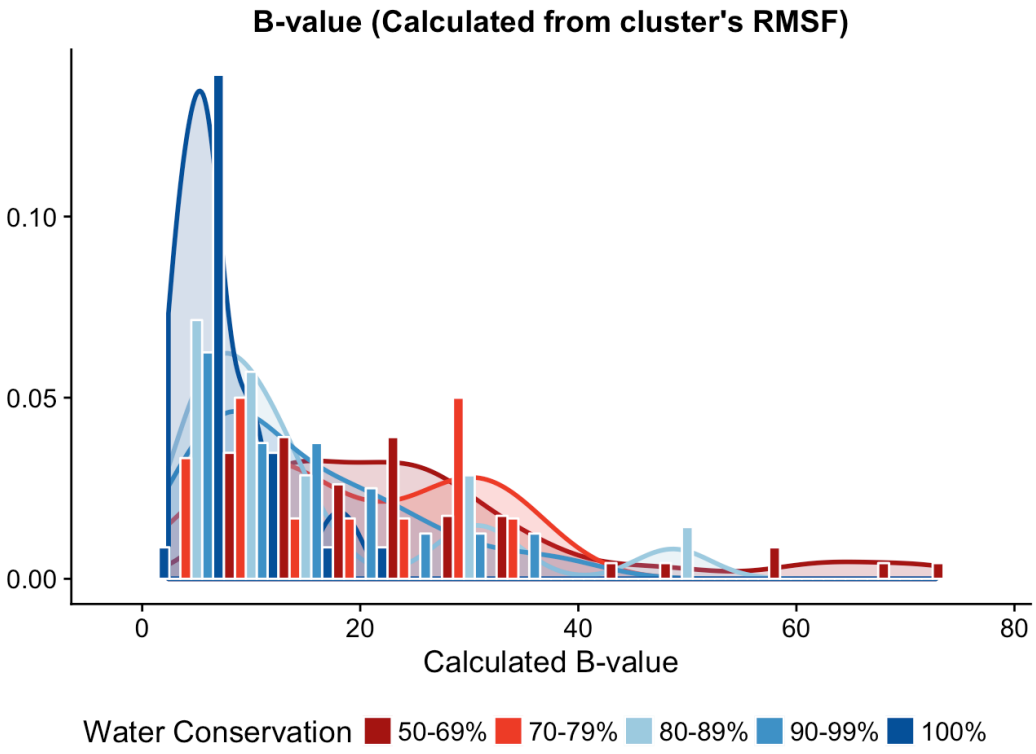


- **B-value Barplots**: A B-value less than 40 is considered ideal while B-values between 60 and 100 indicate atoms with significant variance in their location. The B-value plot below demonstrates the inverse relationship between low B-value and highly conserved waters. None of the conserved waters – those with 50% or greater – in this analysis have a B-value greater than 40.

```
BvalueBarplot(data = thrombin10.conservedWaters, passed.waters = TRUE, calc.values = FALSE)
```

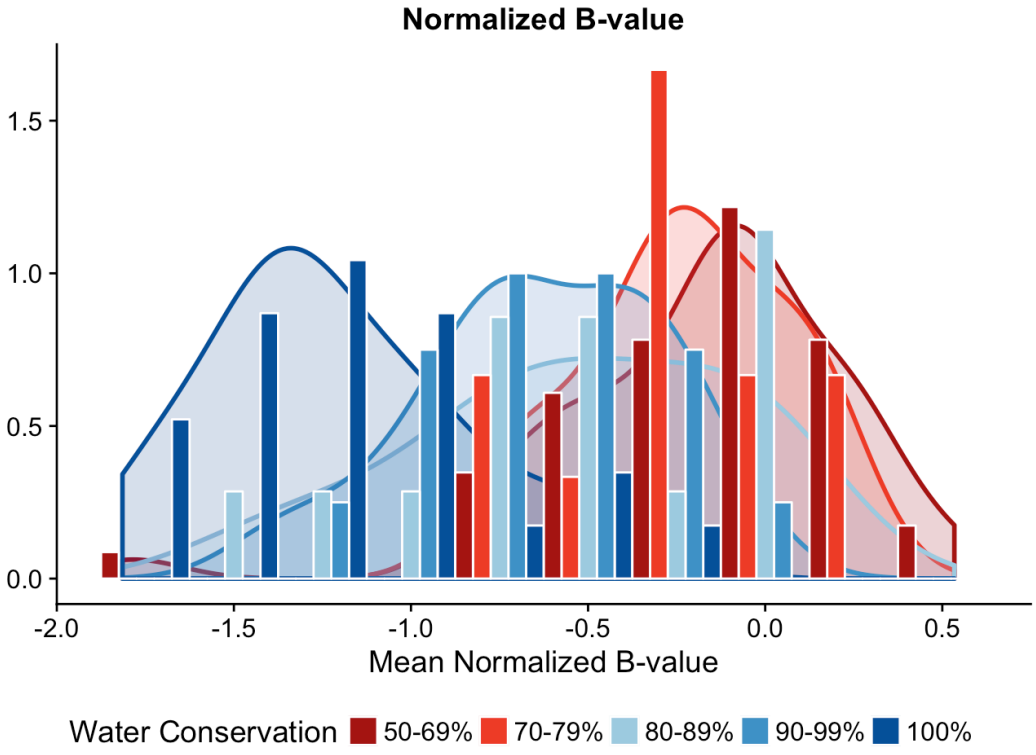



```
BvalueBarplot(data = thrombin10.conservedWaters, passed.waters = TRUE, calc.values = TRUE)
```



- **Normalized B-value Barplots**: The normalized B-values for this analysis range from approximately -2 to about 0.5 with highly conserved waters having a more negative value than less conserved waters; again an inverse relationship. Individual waters with a normalized B-value 1.0 or less are considered sound and included in the analysis.

```
nBvalueBarplot(data = thrombin10.conservedWaters, passed.waters = TRUE)
```

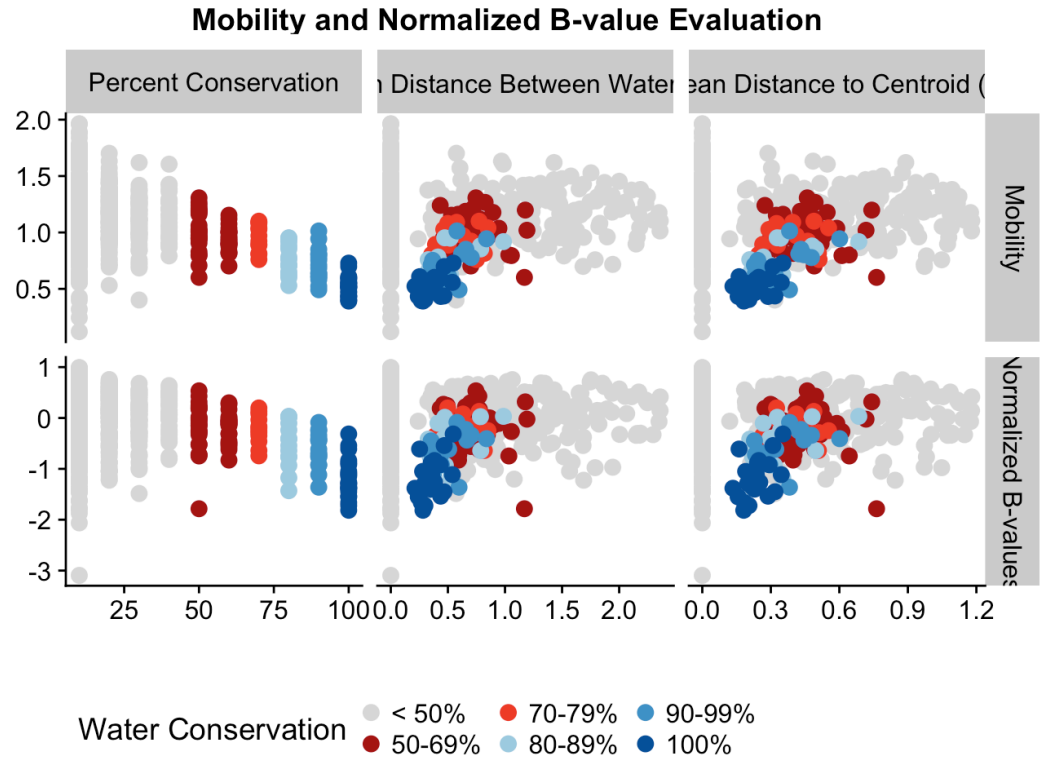


- **Cluster Summary Plots**: The five conserved water quality plots (Number of waters per cluster, Occupancy, Mobility, B-value, and Normalized B-value plots) from above can be created in a single plot object using the `ClusterSummaryPlots(data = thrombin10.conservedWaters, passed.waters = TRUE, plot.labels = NULL)` command. The `plot.labels` option provides the ability to include upper-case letters next to each plot (*e.g.*, A, B, C, D, E; use `"AUTO"`), lower-case letters (*e.g.*, a, b, c, d, e; use `"auto"`), or no letters (use `NULL`). This plot is not displayed.

- **Mobility and Normalized B-values Evaluation Plots**: This three-panel plot illustrates the relationships between mobility and normalized B-values and the percent conservation, the mean distance between waters within the cluster (in Angstroms), and the mean distance between waters within a cluster and the cluster's centroid. A majority 53 conserved water clusters with 80% conservation or greater have a mobility value are less than 1.0 while the normalized B-value ranges between approximately -1.8 and 0.04.

The light-grey points in the second and third panels at the 0.0 Angstrom X-axis represents conserved water clusters with a single water. Overall, waters with a conservation value of 50% or greater have a mean value distance between cluster waters less than approximately 1.25 Angstroms and less than 0.9 Angstroms between the waters and the cluster's centroid. The distances between waters – and between the centriod – is reduced as the amount of conservation increases. This indicates highly conserved water clusters have less positional variability compared to lesser conserved water clusters.

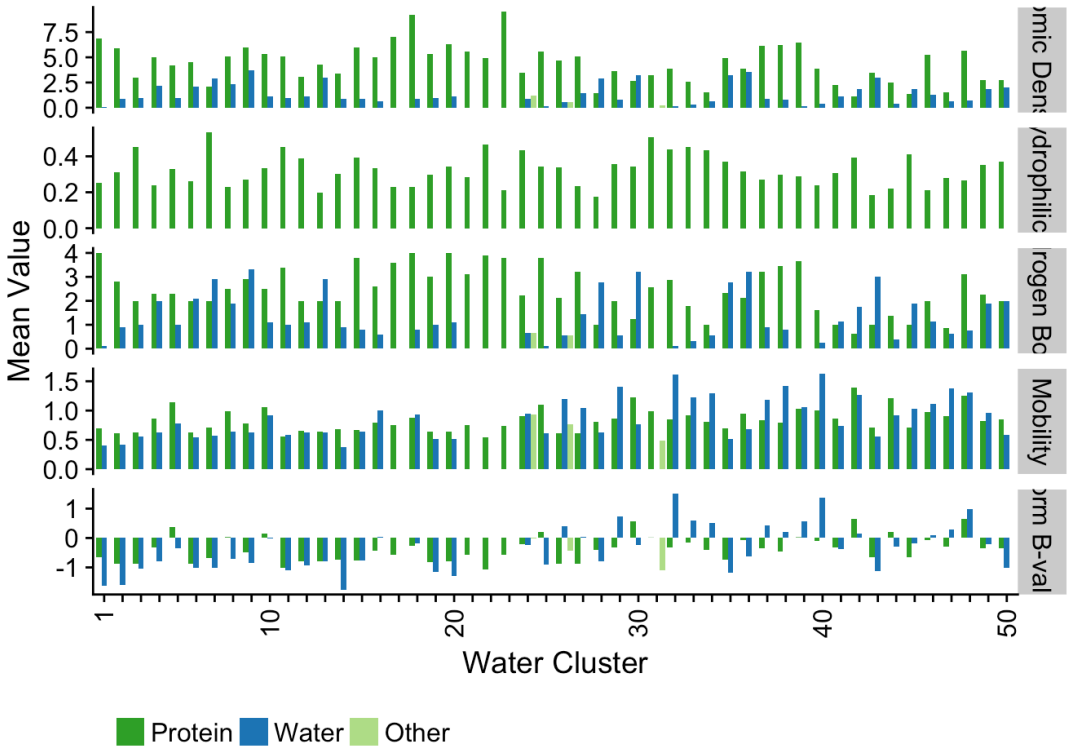
```
MobNormBvalEvalPlots(data = thrombin10.conservedWaters, passed.waters = TRUE,
                     title = "Mobility and Normalized B-value Evaluation")
```



- **Bound Water Environment Barplots**: A collection of five plots illustrating the bound environment for the top 50 conserved waters; the number of conserved waters is user defined.
  - ***Atomic Density***: The first plot indicates the number of protein, water, and other (HETATM) heavy atoms within 3.6 Angstroms of the conserved water. Conserved waters with a large number of protein atoms are likely buried while those with nearby waters likely contribute to water networks.
  - ***Hydrophilicity***: The next plot is the mean hydrophilicity value for nearby protein atoms based on the recalculated hydrophilicity values. Mean values less than 0.3 typically represents hydrophobic regions, values 0.3 are 0.4 are BLAH, and values greater than 0.4 are considered hydrophilic. Because
  - ***Hydrogen Bonds***: The average number of potential hydrogen bonds between the water cluster and the protein, waters, and other (HETATM) heavy atoms. The more potential hydrogen bonds between the conserved water and protein atoms indicates the water is likely highly stable while more potential hydrogen bonds between the water and other waters can indicate a water-based hydrogen bond network. Typically, a water can have a maximum of four hydrogen bonds but because the presented values are based on a collection of structures – and each water can experience slightly different environments – the total number of potential hydrogen bonds might be greater than four.
  - ***Mobility***: Based on the experimentally obtained occupancy and B-values, the mean mobility is a measure of the quality of atoms within a crystallographically derived structure and values closer to zero (0) indicate better quality heavy atoms. This plot illustrates the quality of atoms surrounding the conserved water and is an average of the structures contributing to the conserved water cluster.
  - ***Normalized B-values***: The average normalized B-values are another measure of quality for the heavy atoms surrounding the conserved water. Values less than zero indicate higher quality heavy atoms. A majority of the protein, water, and other heavy atoms surrounding the top 50 conserved waters are considered high quality.

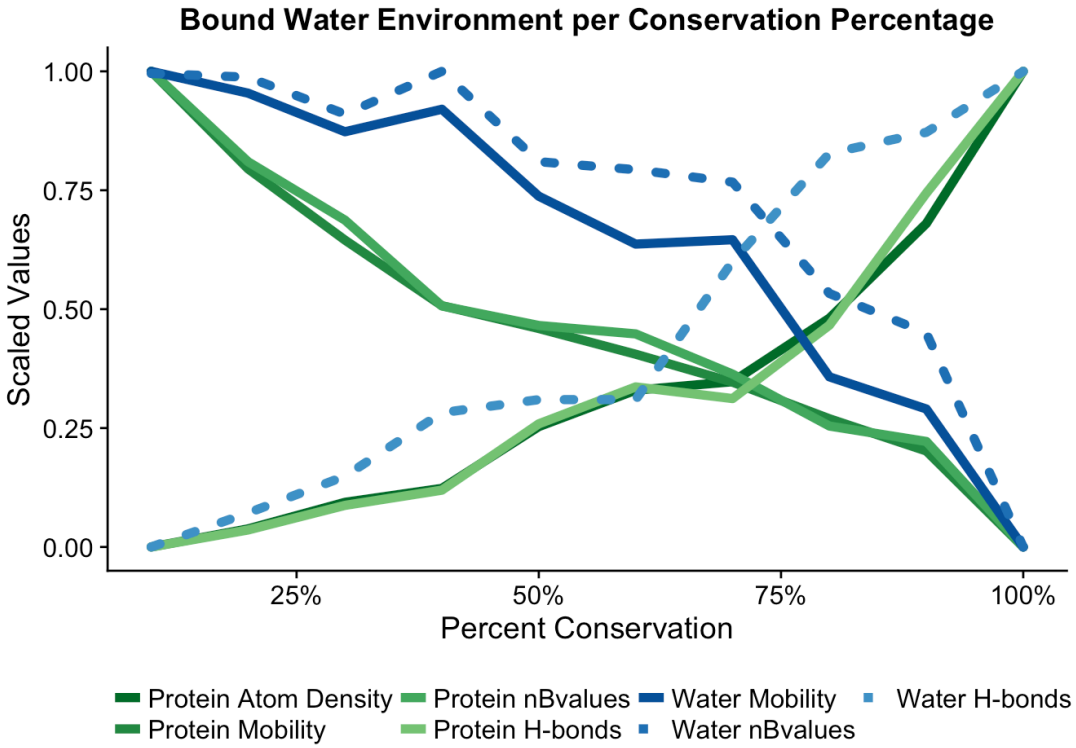
All values are based on environment of each cluster's waters in their original protein structures.

```
BoundWaterEnvPlots(data = thrombin10.conservedWaters, passed.waters = TRUE,
                   pct.conserved.gte = 50, num.clusters = 50)
```



- **Bound Water Environment Summary Plot**: This plot is designed to illustrate how various features of the protein heavy atoms surrounding the conserved water cluster change as the percent conservation (precentage of protein structures contributing to the conserved water cluster) changes. The average and scaled values for neighboring protein and water atoms related to the conserved water clusters at each conservation percentage are plotted. Obvious trends include, the protein atomic density and potential hydrogen bonds between the protein and between waters increases with percent conservation indicating that highly conserved waters interact with the protein and other waters. There is an inverse relationship between the quality of protein and water heavy atoms within 3.6 Angstroms of a conserved cluster and the percent conservation.

```
BoundWaterEnvSummaryPlot(data = thrombin10.conservedWaters, passed.waters = TRUE,
                         title = "Bound Water Environment per Conservation Percentage")
```



Molecular Visualization with PyMOL

A PyMOL script file is generate to graphically display the conserved waters in relation to a representative protein. Two versions of the PyMOL script file are generated: one with a black background and the other with a white background. The color of the pocket residues is changed based on the background color; the pocket residues are colored light-grey for the black background and dark-grey for the white background. The ligand is assigned the user-defined color for both representations. Pocket residues – and associated molecular surface – are defined as those within 5 Angstroms of the conserved waters. The depicted cartoon representation is for residues within 15 Angstroms of the ligand(s). This initial graphical rendering should be considered a starting point to generate a final image. The PyMOL representation depicts:

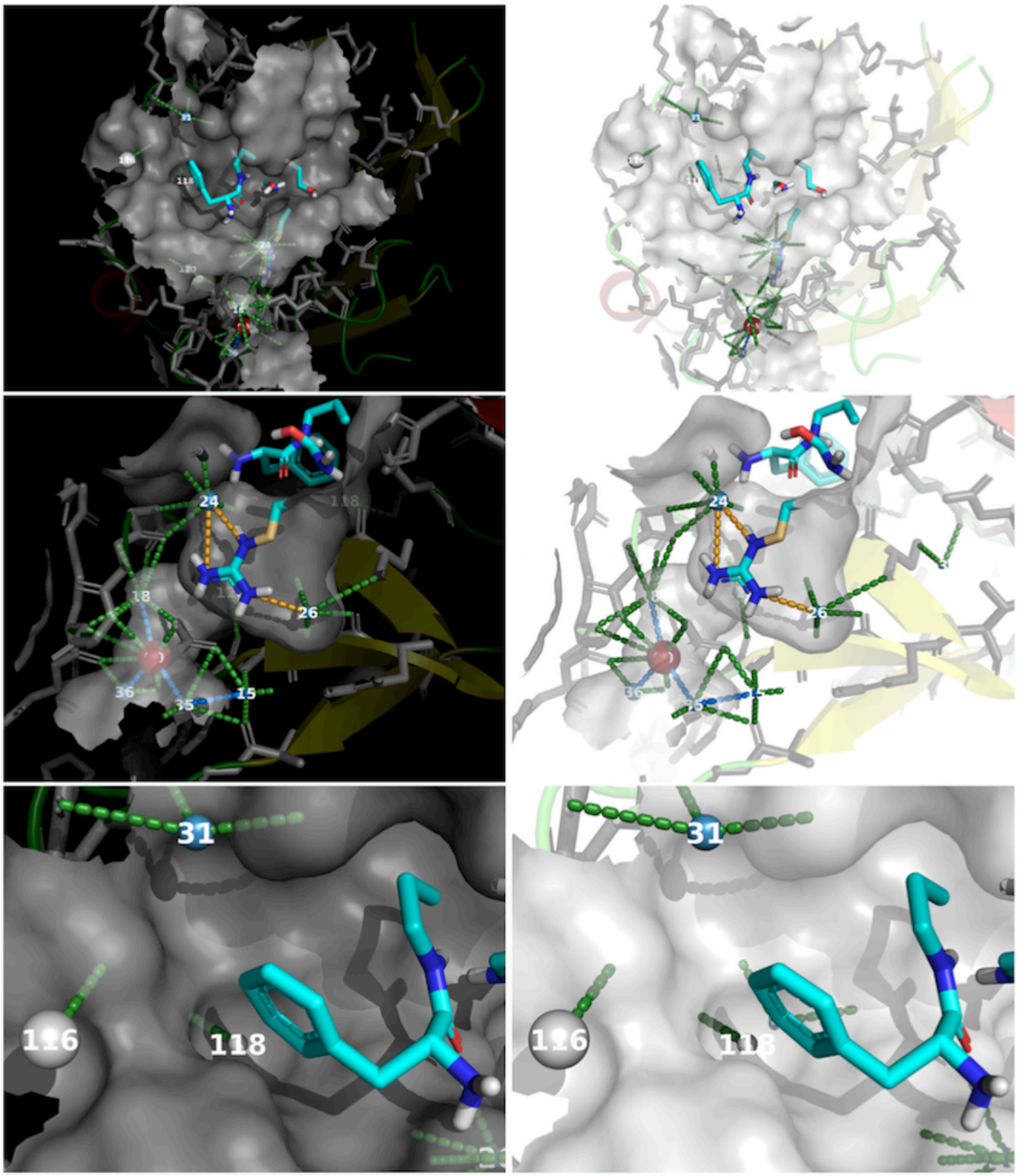
The conserved waters - waters are colored using the colors defined for percent conservation - the radius of the waters are scaled based on the calculated B-values; smaller waters have smaller calculated B-values - waters are numbered based on percent conservation and calculated B-value - only waters within the binding site are visualized

The potential hydrogen bonds are depicted based on the interactions between: - conserved waters and ligand: orange dashed line - conserved waters and protein: green dashed line - conserved waters: blue dashed line

```
CreatePyMOLscript(conservedWaters.data = thrombin10.conservedWaters,
                  passed.waters = TRUE,
                  PDBid.ref = "1hai",
                  LigResname.ref = "0g6",
                  hbond = 3.75,
                  lig.carbon.color = "cyan",
                  filename = thrombin10.filename)
## The PyMOL script files (thrombin10_ConservedWaters_PASSED_PyMOL_black_background_aug102017_1601.pml and thromb
in10_ConservedWaters_PASSED_PyMOL_white_background_aug102017_1601.pml) to easily view the conservedwaters are in
the /Users/emilio/GitHub/vanddraabe/vignettesLongForm directory (folder).
```

This series of three views of Thrombin's binding site were created using PyMOL (version 1.8.1.0) from the PyMOL scripts created using the above `CreatePyMOLscript` command. In the middle panel, conserved water #24 has five potential hydrogen bonds with protein atoms (green dashed lines) and two potential hydrogen bonds (orange dashed lines) with the PPACK ligand. The slightly obscured, comparatively large, red sphere is conserved water #90 with three potential hydrogen bonds to neighboring waters and five potential hydrogen bonds to protein atoms. Conserved water #24 has 90% conservation, a calculated B-value of approximately 22 (an average distance between contributing waters of 0.67 Angstroms and an average distance of 0.44 Angstroms from the contributing waters to the conserved water's centroid). Conserved water #90 has a conservation percentage of 50% and a calculated B-value of 61.6 (an average distance between contributing waters of 1.06 Angstroms and an average distance of 0.61 Angstroms from the contributing waters to the conserved water's centroid). An average of six protein heavy atoms surround conserved water #90 with an average of 3 potential hydrogen bonds between the contributing waters and their corresponding protein and approximately 3.4 hydrogen bonds to neighboring waters. Further inspection of conserved water #90 and its surroundings indicates it is

within an extended gorge with several conserved waters of higher quality – such as conserved waters #15, 18, 35, and 36. Taking this information into consideration indicates conserved water #90 might not be highly conserved and has the potential for numerous interactions with the protein and other water atoms within the gorge. The third panel focuses on an exposed region of the binding site and illustrations the locations of conserved waters #31, 116, and 118. Based on the size and color of the represented oxygen atoms, conserved water #31 is highly conserved (90%), has a low calculated B-value (18.8), and has three potential hydrogen bonds with the protein (green dashed lines). Notice the size difference between conserved waters #116 and 118 in the third panel. Both conserved waters – #116 and 118 – are composed of waters from four protein structures; conserved water 116 has a calculated B-value of 46.3 while 118 has a calculated B-value of 7.3 indicating a tighter cluster of waters comprising 118 and this is cooborated by an average distance between waters of 1.02 Angstroms compared to 0.41, respectively, and average distances to the centroid of conserved waters of 0.60 and 0.26 Angstroms, respectively. Another possible contributing factor to #118's tighter collection of contributing waters is the number protein heavy atoms within 3.6 Angstroms of the waters in the original structures; 116 has seven total protein heavy atoms for the four structures and 118 has 19 total protein heavy atoms for the four structures. The mean hydrophilicity values for the protein heavy atoms near these two conserved waters is low – 0.27 for #116 and 0.14 for #118 – indicating a hydrophobic environment but both have a potential hydrogen bond with a carboxyl backbone oxygen atom (GLU 97 for #116 and ASN 98 for #118). It is likely the less solvent exposed pocket containing conserved waters comprising #118 contrains these waters to a specific location more than the environment of the waters of conserved water #116's cluster.



Thrombin 10 Results

Session information for this vignette

```
sessionInfo()
```

```
## R version 3.3.3 (2017-03-06)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X Yosemite 10.10.5
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] cowplot_0.8.0    ggplot2_2.2.1    reshape2_1.4.2   bio3d_2.3-3
## [5] vanddraabe_1.0.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.12       knitr_1.16         magrittr_1.5
##  [4] munsell_0.4.3      colorspace_1.3-2   rlang_0.1.1
##  [7] highr_0.6          stringr_1.2.0      fastcluster_1.1.22
## [10] plyr_1.8.4         tools_3.3.3        parallel_3.3.3
## [13] grid_3.3.3         gtable_0.2.0       htmltools_0.3.6
## [16] yaml_2.1.14        lazyeval_0.2.0     rprojroot_1.2
## [19] digest_0.6.12      tibble_1.3.3       bitops_1.0-6
## [22] RCurl_1.95-4.8     evaluate_0.10.1    rmarkdown_1.6
## [25] labeling_0.3       openxlsx_4.0.17    stringi_1.1.5
## [28] scales_0.4.1       backports_1.1.0    XML_3.98-1.9
```