Xh-pi interactions

Emma Rand

22/08/2019

Contents

1 Data used

data files in "../xhpi_data_new/" are in three directories, one for each data-set.

- Neu_43_adv neutron diffraction
- Res_43_adv high resolution
- Temp_43_adv high temperature

Number of structures in each varies

Each protein has 3 files

- 7a3h.pdb
- 7a3h.hpml
- 7a3h.pdb.report .report files are the plain text files containing information about the XH/pi interactions.

Geometric constraints for this data was d<7 and theta<50, although cut-offs of d<4.3 angstroms and theta<25 deg are often used to define XH/pi interactions, so many of the interactions reported are not actually XH/pi interactions.

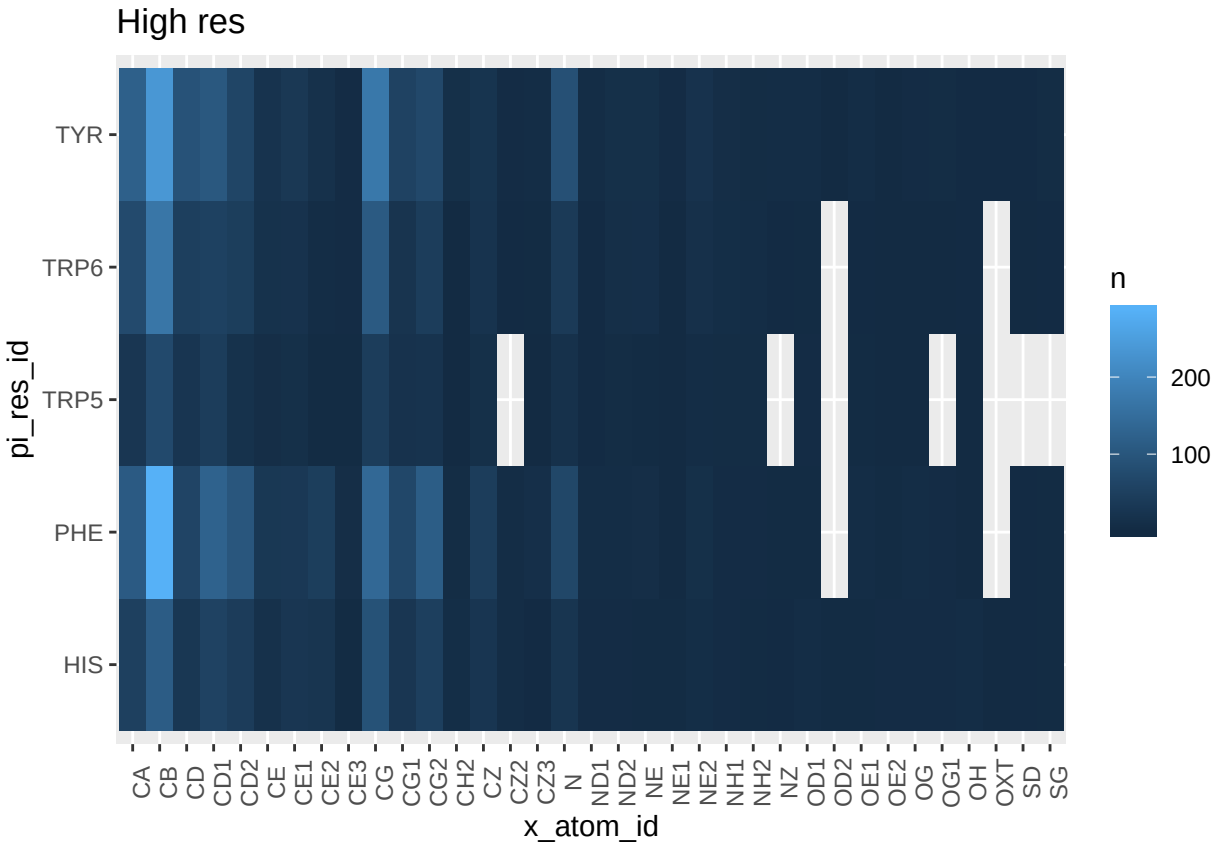
1.1 Number of interactions

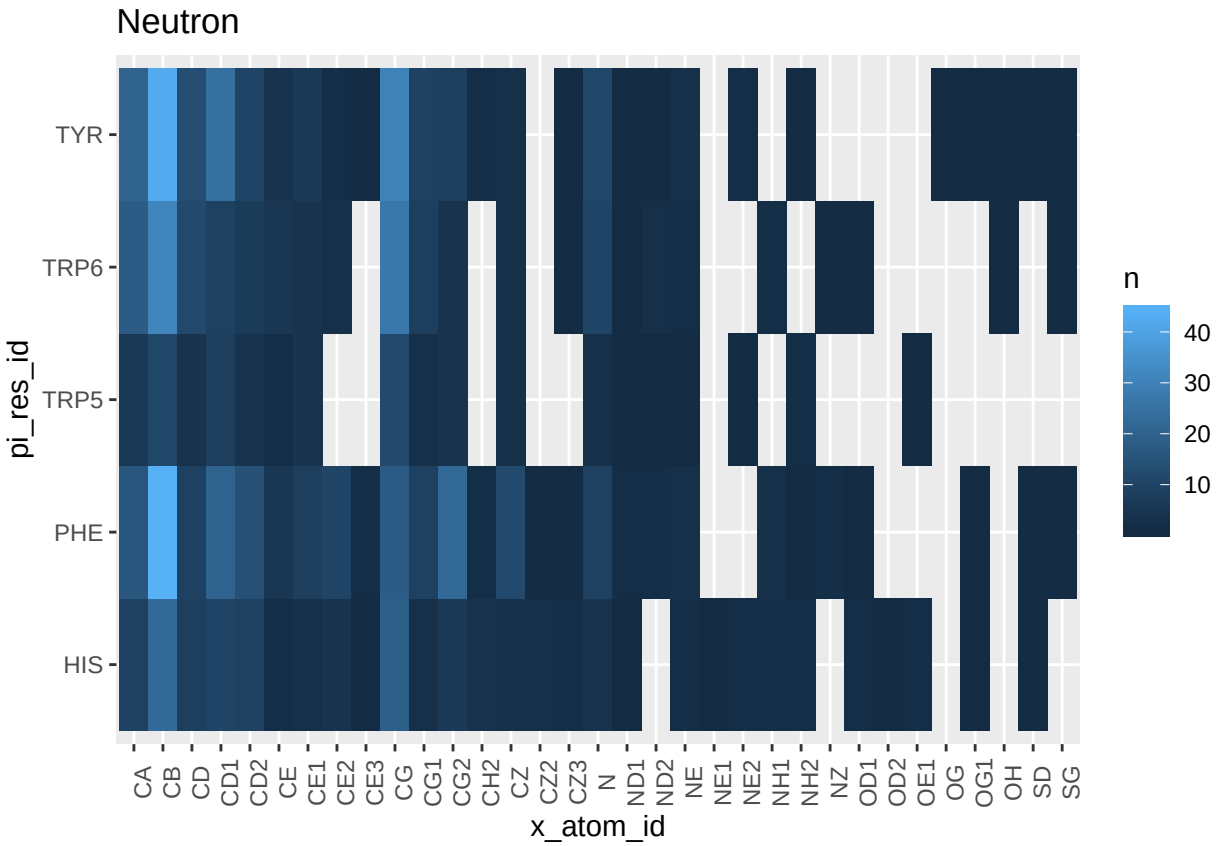
- neutron diffraction: 772
- high resolution: 4305
- high temperature: 1788

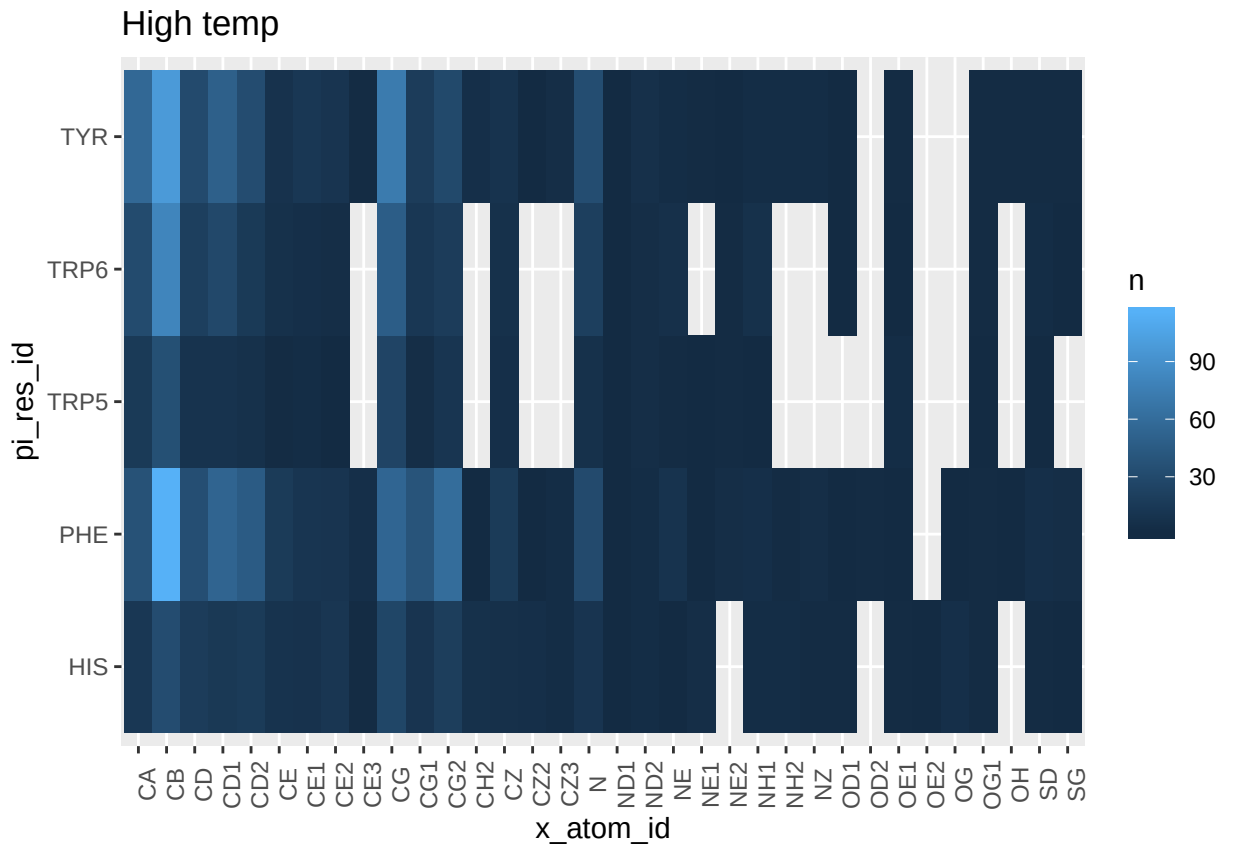
Note, we don't have resolution data for highres or hightemp

1.2 Interaction types

Raw counts, not normalised


High res
pi_res_id
TYR
TRP6
TRP5
PHE
HIS
CA
CB
CD
CD1
CD2
CE
CE1
CE2
CE3
CG
CG1
CG2
CH2
CZ
CZ2
CZ3
N
ND1
ND2
NE
NE1
NE2
NH1
NH2
NZ
OD1
OD2
OE1
OE2
OG
OG1
OH
OXT
SD
SG
x_atom_id
n
200
100


Neutron
pi_res_id
TYR
TRP6
TRP5
PHE
HIS
x_atom_id
CA
CB
CD
CD1
CD2
CE
CE1
CE2
CE3
CG
CG1
CG2
CH2
CZ
CZ2
CZ3
N
ND1
ND2
NE
NE1
NE2
NH1
NH2
NZ
OD1
OD2
OE1
OG
OG1
OH
SD
SG
n
40
30
20
10



2 Clustering mixed data types

Used R (R Core Team 2018) with packages tidyverse (Wickham 2017) and data.table (Dowle and Srinivasan 2019) to perform partitioning around medoids (PAM) (Kaufman and Rousseeuw 1987, @kaufmanl1990clustering) to cluster based on Gower distance (Gower 1971) both implemented in the cluster package (Maechler et al. 2018)

- distance calculation - gower distance
- clustering algorithm - partitioning aroung medoids
- selecting the number of clusters - silhouette width

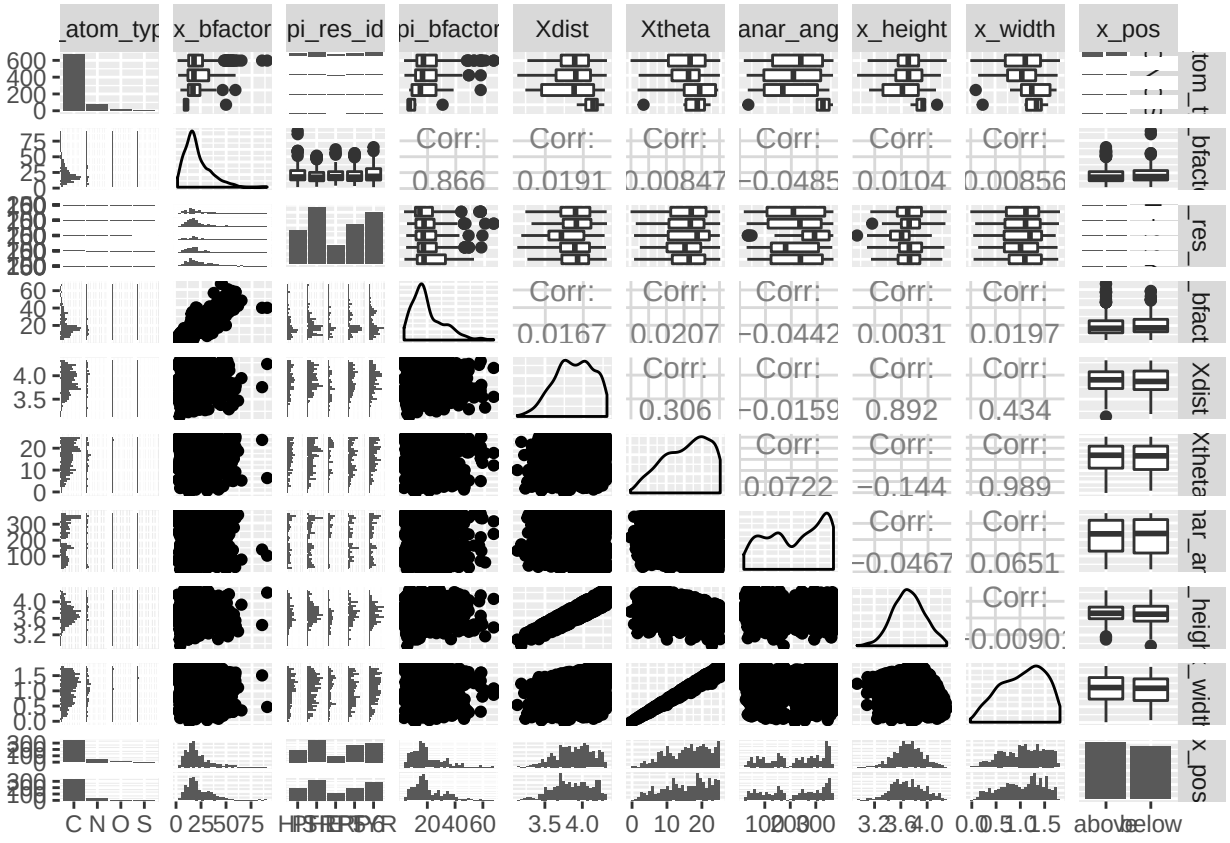
2.1 Neutron

Some variables will not used in clustering:

- x_res_num
- x_atom_num
- x_chain
- pdb
- pi_res_num
- resolution
- pi_chain

Variables used in cluserting are: x_res_id, x_atom_id, x_atom_type, x_bfactor, pi_res_id, pi_bfactor, Xdist, Xtheta, planar_angle, x_height, x_width, x_pos

2.1.1 Relationships between variables



2.1.2 Distance calculation, no variable transformations applied

```
## 297606 dissimilarities, summarized :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01449 0.35781 0.42227 0.42197 0.48595 0.81606
## Metric :  mixed ;  Types = N, N, N, I, N, I, I, I, I, I, I, N
## Number of objects : 772
```

Which observations are least and most similar?

most similar i.e., min disimilarity

```
##     x_res_id x_atom_id x_atom_type x_bfactor pi_res_id pi_bfactor Xdist
## 542      LYS        CG           C     18.23      TRP6      18.04 3.974
## 115      LYS        CG           C     21.62      TRP6      17.38 4.016
##     Xtheta planar_angle x_height x_width x_pos
## 542  18.98        89.05    3.758   1.292 above
## 115  19.14       102.70    3.794   1.317 above
```

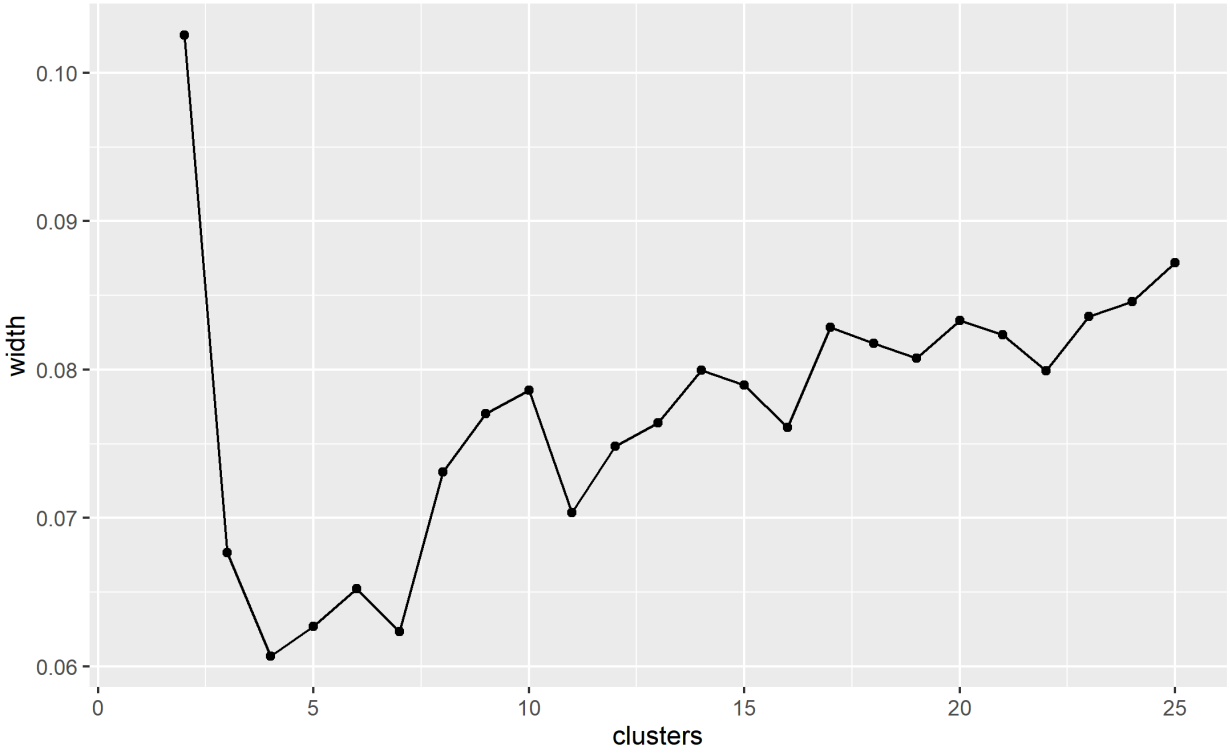
least similar i.e., max disimilarity

```
##     x_res_id x_atom_id x_atom_type x_bfactor pi_res_id pi_bfactor Xdist
## 657      ARG       NH1           N     52.62      TRP6      49.01 3.292
## 25       ALA        CA           C      9.18       TYR      10.12 4.215
##     Xtheta planar_angle x_height x_width x_pos
```

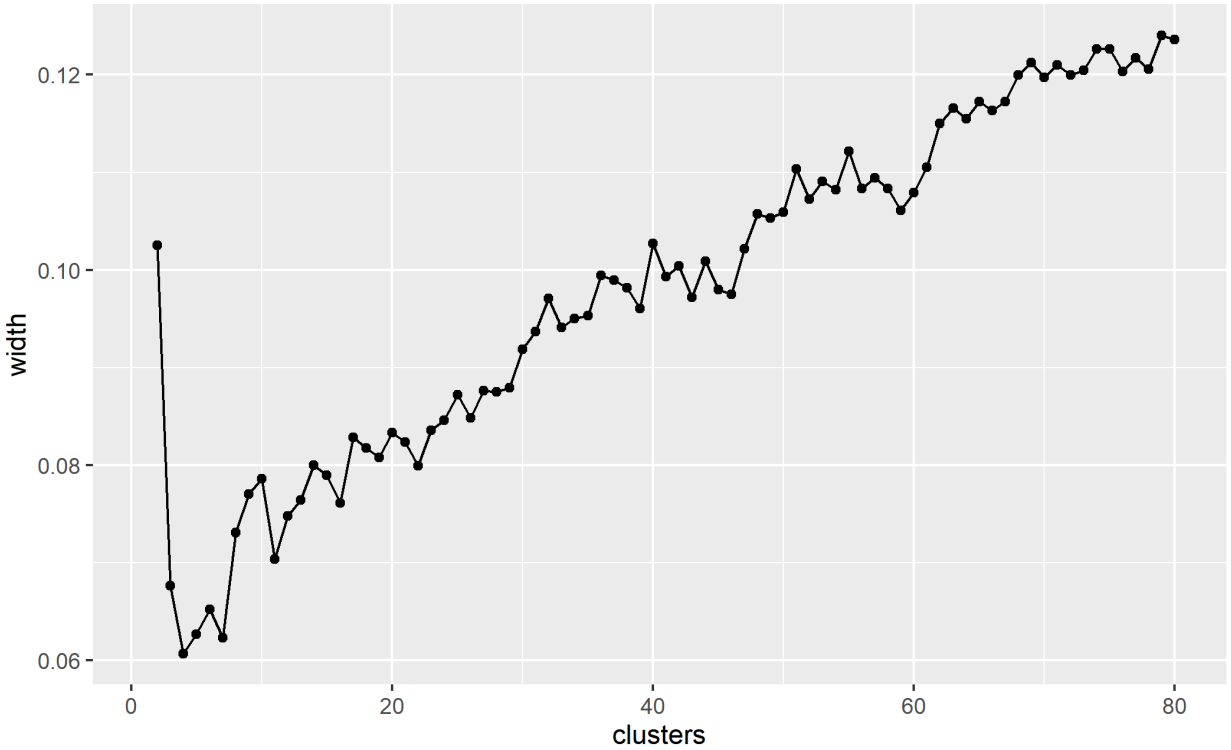
```
## 657   5.61        92.11    3.276  0.3218 above
## 25   24.33       357.30    3.841  1.7370 below
```

2.1.2.1 Cluster with PAM partitioning round medoids

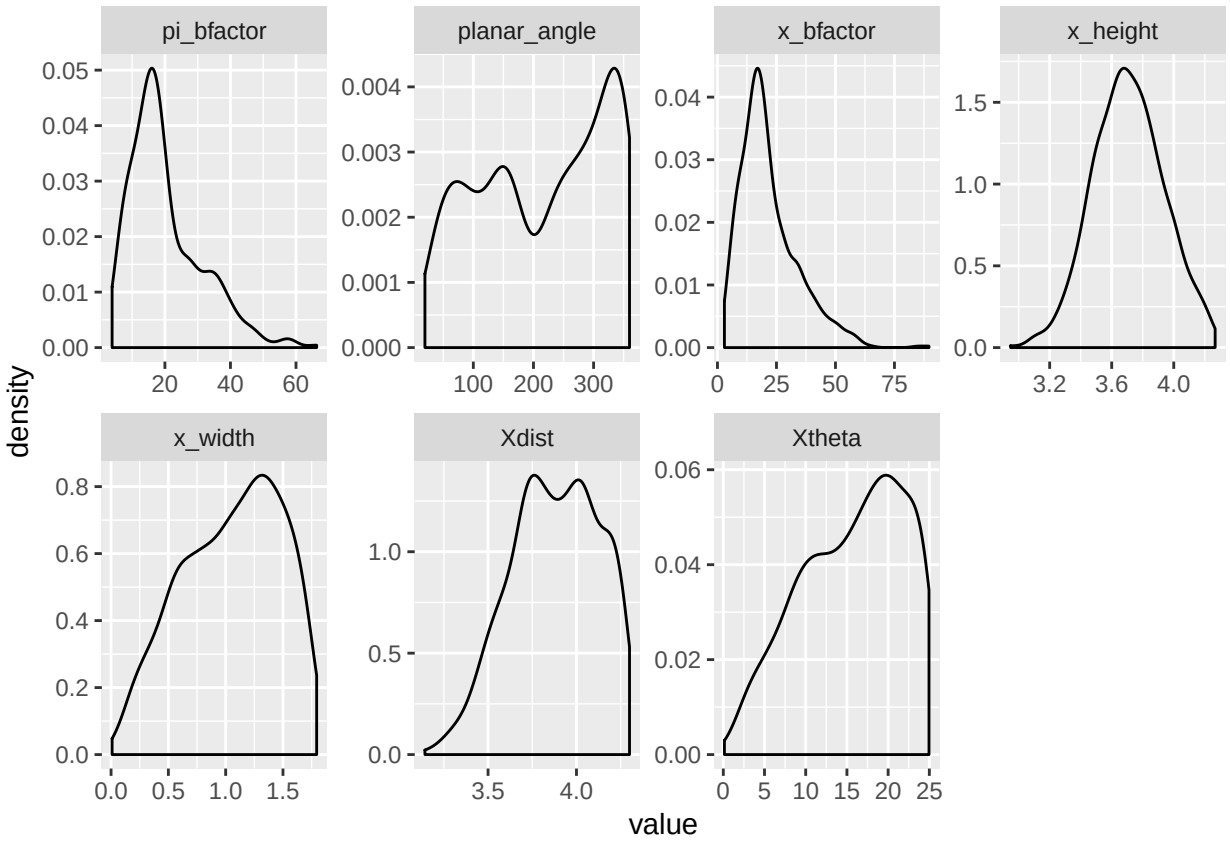
The silhouette width, measure for selecting the number of clusters, was calculated for clusters ranging from 2 to 25 A higher silhouette width is better. Ideally, one sees it peak and fall again



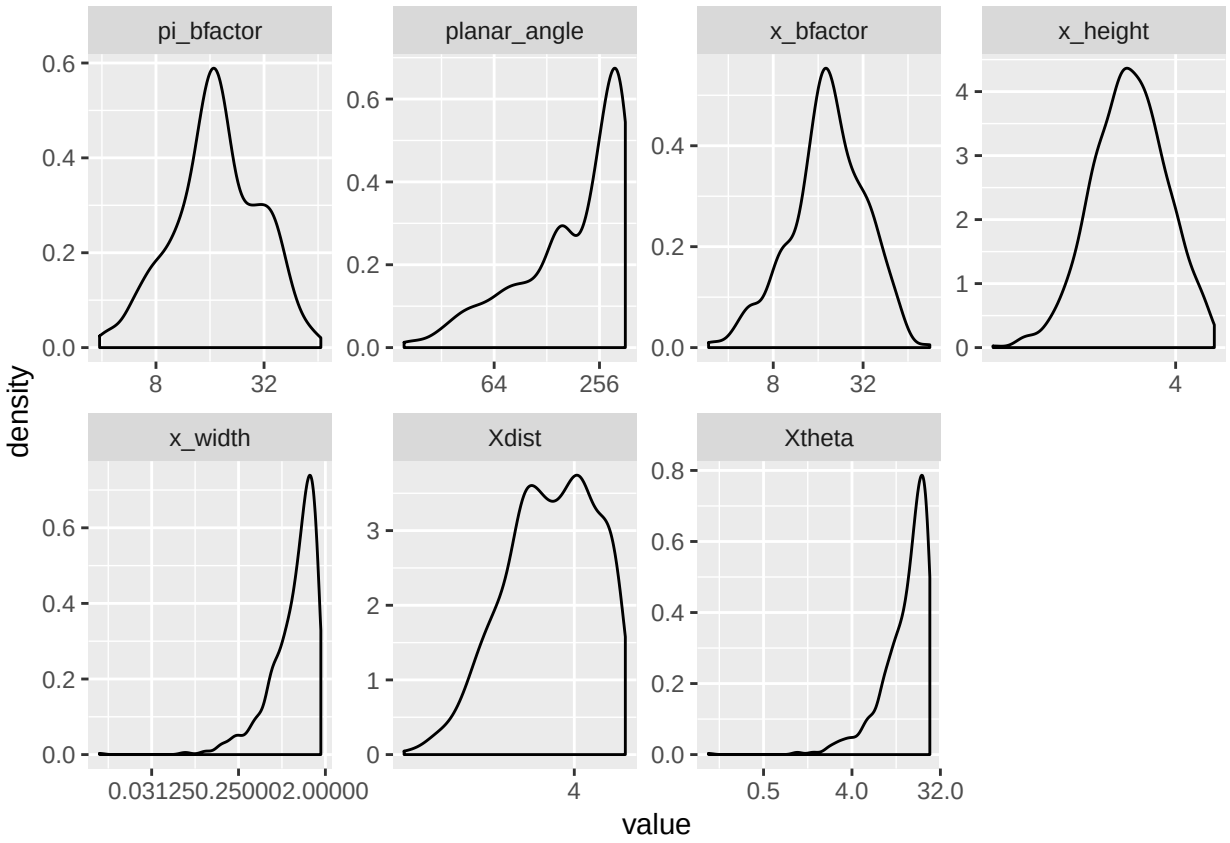
Since the silhouette width was still increasing, I tried raising to 80.



Possibly beginning to flatten but this many clusters is difficult to interpret and may be uninformative Gower distance is sensitive to non-normality and extreme values in the continuous variable and transformations of these might help. The continuous variables are: x_bfactor, pi_bfactor, Xdist, Xtheta, planar_angle, x_height and x_width. Their distributions are as follows:



Log transformed variables have the following distributions



It appears that bfactors would benefit from log transformation when calculating the distances. These are straightfroawrd to apply so will be done first. x_width and xtheta might benefit from some transformation (squaring or ranking maybe).

2.1.3 Distance calculation, transformed b-factors

```
## 297606 dissimilarities, summarized :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01561 0.36353 0.42788 0.42765 0.49168 0.84709
## Metric :  mixed ;  Types = N, N, N, I, N, I, I, I, I, I, I, N
## Number of objects : 772
```

Which observations are least and most similar?

most similar i.e., min disimilarity

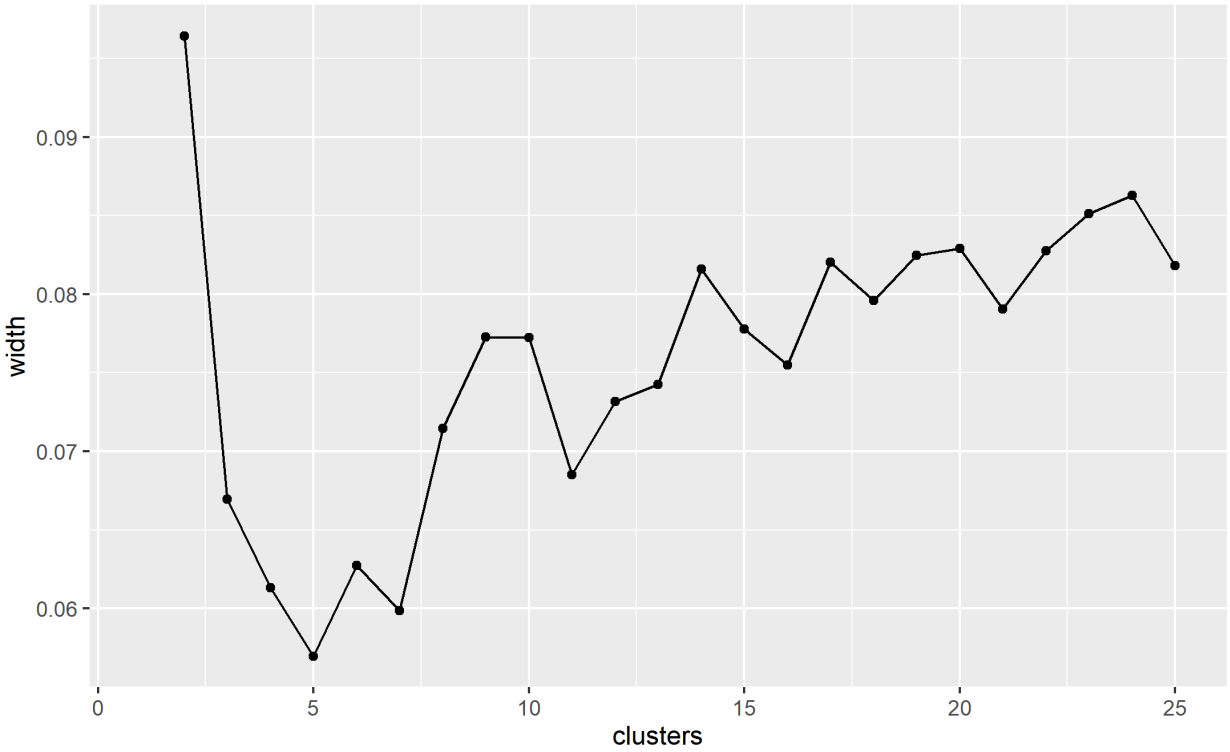
```
##     x_res_id x_atom_id x_atom_type x_bfactor pi_res_id pi_bfactor Xdist
## 542      LYS        CG           C     18.23      TRP6      18.04 3.974
## 115      LYS        CG           C     21.62      TRP6      17.38 4.016
##     Xtheta planar_angle x_height x_width x_pos
## 542  18.98        89.05    3.758   1.292 above
## 115  19.14       102.70    3.794   1.317 above
```

least similar i.e., max disimilarity

```
##     x_res_id x_atom_id x_atom_type x_bfactor pi_res_id pi_bfactor Xdist
## 677      ARG       NH2           N     56.47      TRP5     41.880 3.371
## 267      LYS        CG           C      3.06       HIS      4.795 4.189
```

```
##     Xtheta planar_angle x_height x_width x_pos
## 677  5.727        136.0    3.354  0.3364 above
## 267 24.360        359.1    3.816  1.7280 below
```

2.1.3.1 Cluster with PAM partitioning round medoids



The b-factor transformation has made little impact.

Is this the same for the other structure determination methods?

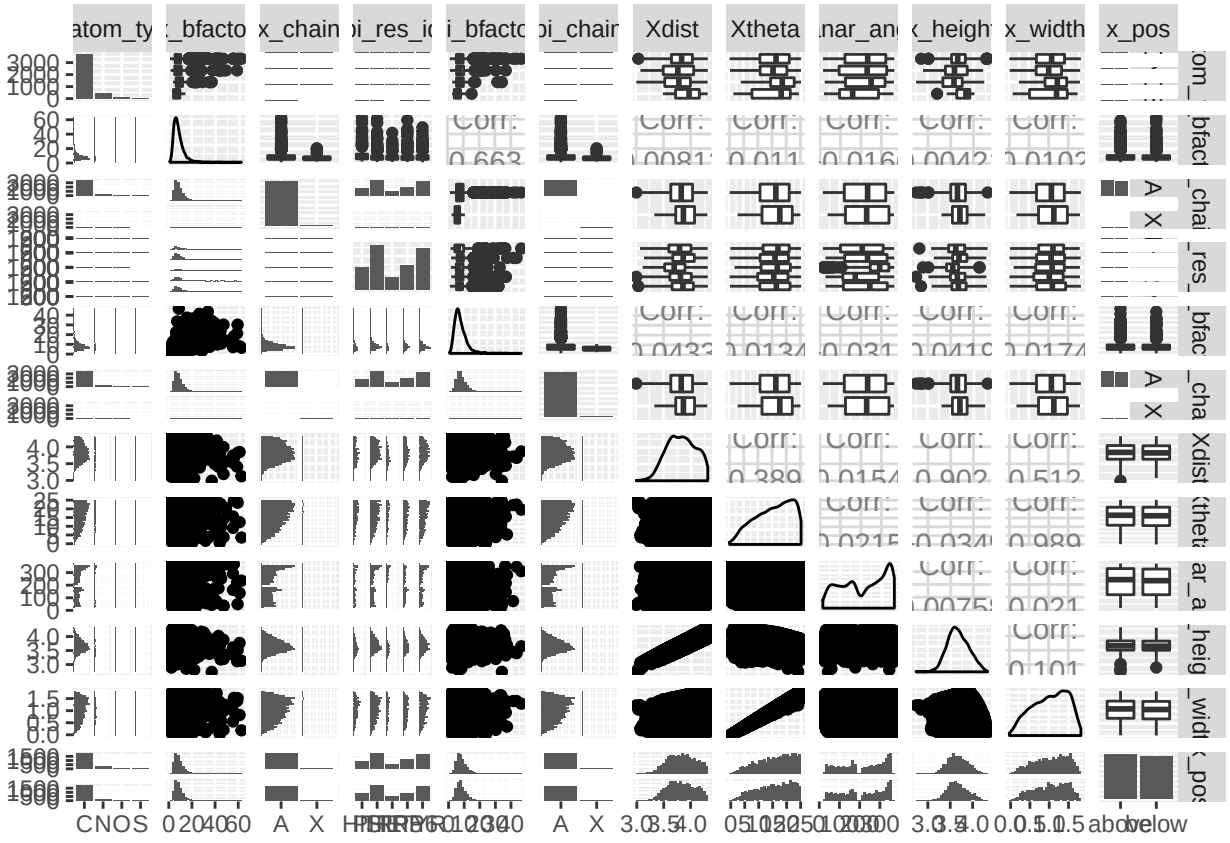
2.2 High resolution

Some variables will not used in clustering:

- x_res_num
- x_atom_num
- pdb
- pi_res_num
- resolution note: pi_chain and x_chain are retained

Variables used in cluserting are: x_res_id, x_atom_id, x_atom_type, x_bfactor, x_chain, pi_res_id, pi_bfactor, pi_chain, Xdist, Xtheta, planar_angle, x_height, x_width, x_pos

2.2.1 Relationships between variables



2.2.2 Distance calculation, no variable transformations applied

```
## 9264360 dissimilarities, summarized :
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.005929 0.293480 0.348650 0.349150 0.403080 0.865050
## Metric :  mixed ;  Types = N, N, N, I, N, N, I, N, I, I, I, I, I, N
## Number of objects : 4305
```

Which observations are least and most similar?

most similar i.e., min disimilarity

```
##      x_res_id x_atom_id x_atom_type x_bfactor x_chain pi_res_id pi_bfactor
## 2106      GLN        CG           C      7.03       A       TYR      7.745
## 46        GLN        CG           C      6.67       A       TYR      6.622
##      pi_chain Xdist Xtheta planar_angle x_height x_width x_pos
## 2106        A 3.962  16.22        290.4    3.804   1.107 below
## 46          A 3.965  16.14        304.1    3.809   1.102 below
```

least similar i.e., max disimilarity

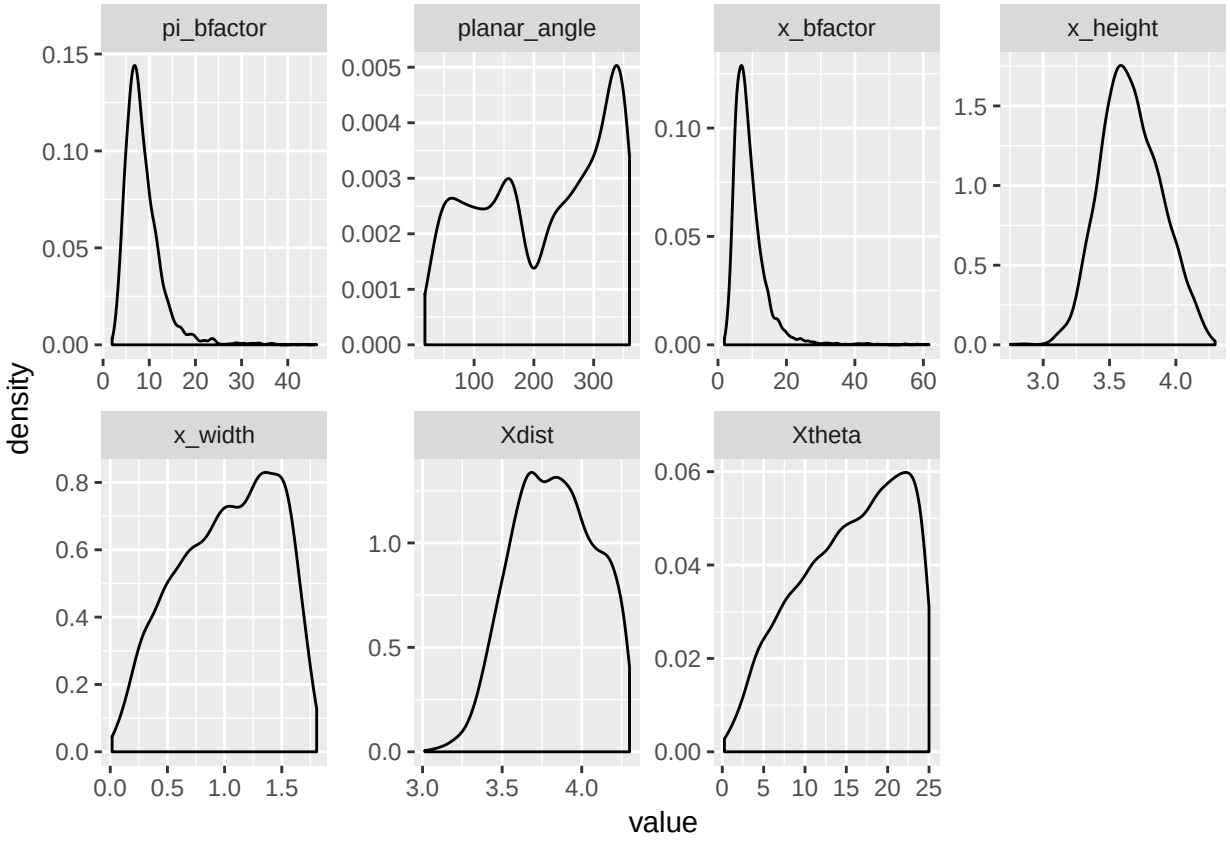
```
##      x_res_id x_atom_id x_atom_type x_bfactor x_chain pi_res_id pi_bfactor
## 2258      LYS        CG           C      4.39       X       TYR      4.697
## 1669      GLU       OE2           O     36.82       A      TRP6     34.230
##      pi_chain Xdist Xtheta planar_angle x_height x_width x_pos
```

```
## 2258        X 4.162  6.294       352.70    4.136  0.4562 below
## 1669        A 3.015 23.990        42.24    2.754  1.2260 above
```

2.2.2.1 Cluster with PAM partitioning round medoids

n.b. many more observations so clustering takes much longer

Distributions of the continuous variables: x_bfactor, pi_bfactor, Xdist, Xtheta, planar_angle, x_height and x_width.



2.2.3 Distance calculation, transformed b-factors

```
## 9264360 dissimilarities, summarized :
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.008299 0.301600 0.356970 0.357490 0.411830 0.863470
## Metric :  mixed ;  Types = N, N, N, I, N, N, I, N, I, I, I, I, I, N
## Number of objects : 4305
```

Which observations are least and most similar?

most similar i.e., min disimilarity

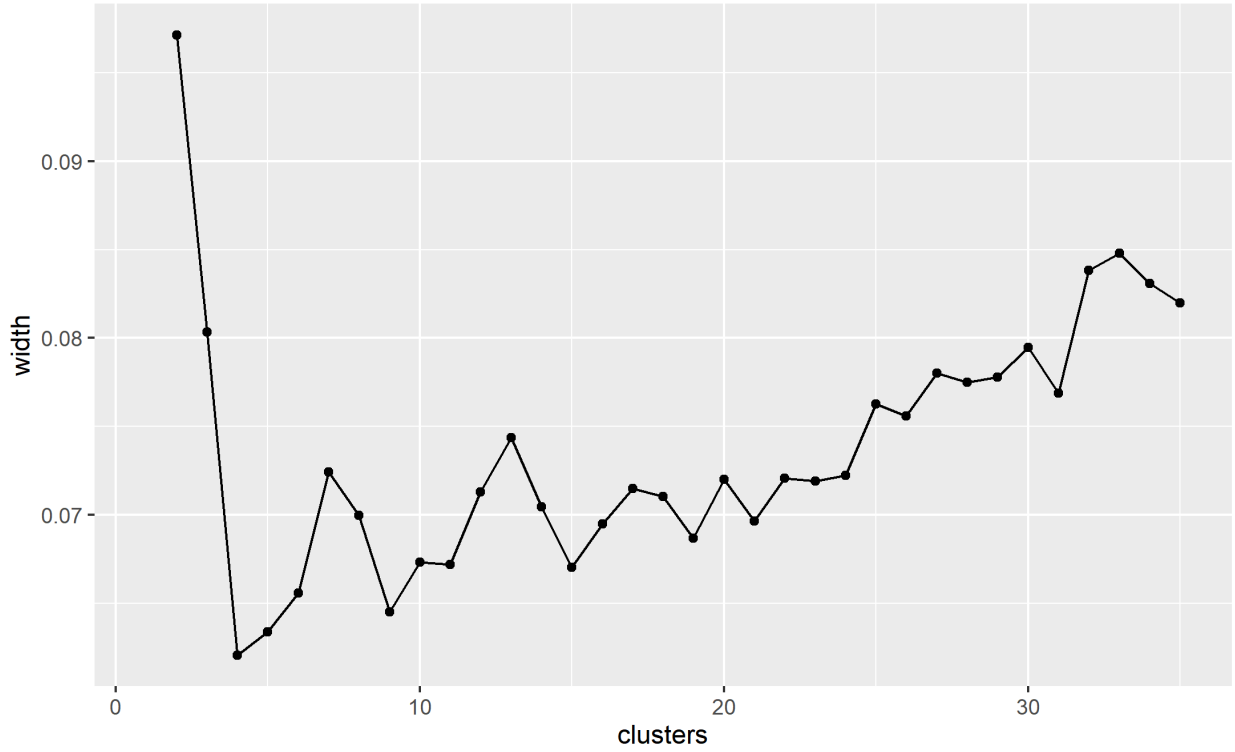
```
##      x_res_id x_atom_id x_atom_type x_bfactor x_chain pi_res_id pi_bfactor
## 2106      GLN        CG           C      7.03       A       TYR      7.745
## 46        GLN        CG           C      6.67       A       TYR      6.622
##      pi_chain Xdist Xtheta planar_angle x_height x_width x_pos
## 2106        A 3.962  16.22        290.4    3.804   1.107 below
```

```
## 46          A 3.965  16.14        304.1    3.809   1.102 below
```

least similar i.e., max disimilarity

```
##      x_res_id x_atom_id x_atom_type x_bfactor x_chain pi_res_id pi_bfactor
## 2258      LYS        CG           C      4.39       X       TYR      4.697
## 1669      GLU       OE2           O     36.82       A      TRP6     34.230
##      pi_chain Xdist Xtheta planar_angle x_height x_width x_pos
## 2258        X 4.162  6.294       352.70    4.136  0.4562 below
## 1669        A 3.015 23.990        42.24    2.754  1.2260 above
```

2.2.3.1 Cluster with PAM partitioning round medoids



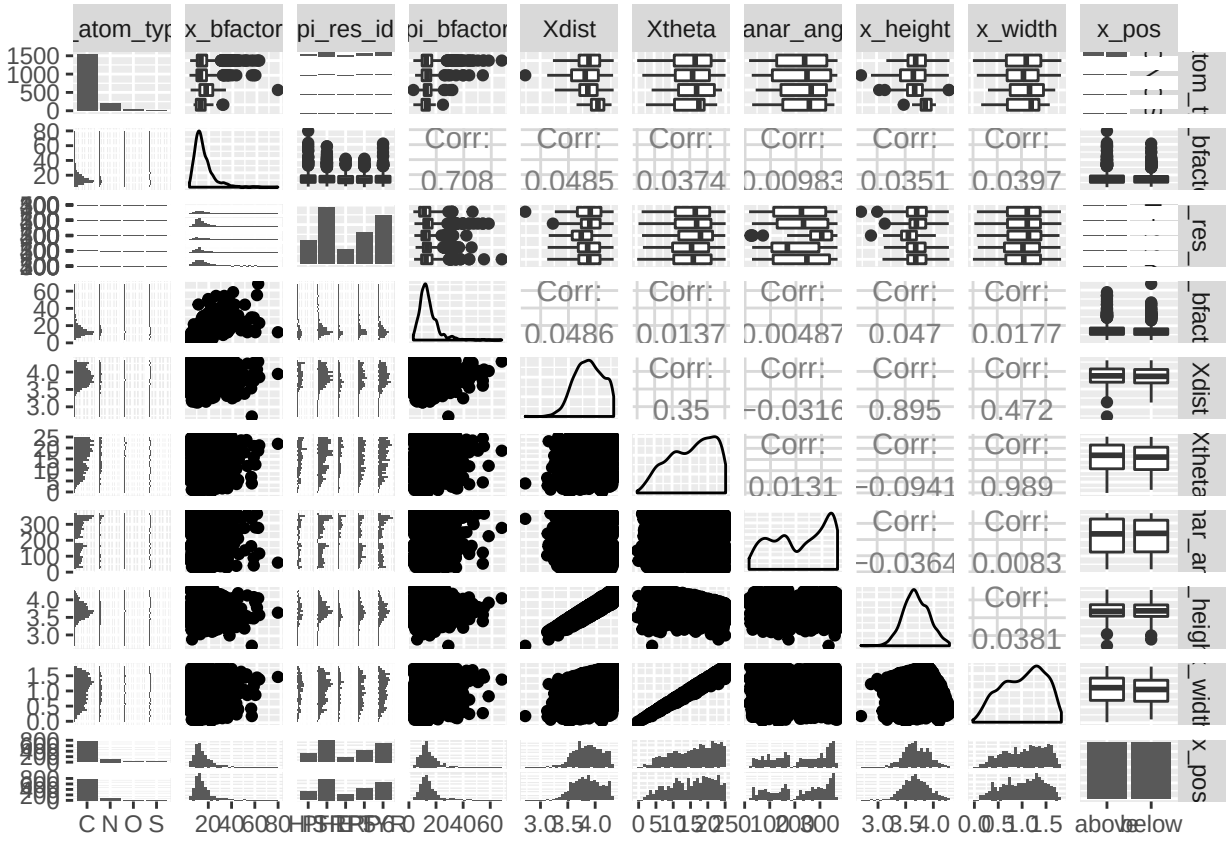
2.3 High temperature

Some variables will not used in clustering:

- x_res_num
- x_atom_num
- pdb
- pi_res_num
- resolution
- pi_chain
- x_chain

Variables used in cluserting are: x_res_id, x_atom_id, x_atom_type, x_bfactor, pi_res_id, pi_bfactor, Xdist, Xtheta, planar_angle, x_height, x_width, x_pos

2.3.1 Relationships between variables



2.3.2 Distance calculation, no variable transformations applied

```
## 1597578 dissimilarities, summarized :
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.009904 0.340570 0.403620 0.402360 0.464560 0.854380
## Metric :  mixed ;  Types = N, N, N, I, N, I, I, I, I, I, I, N
## Number of objects : 1788
```

Which observations are least and most similar?

most similar i.e., min disimilarity

```
##     x_res_id x_atom_id x_atom_type x_bfactor pi_res_id pi_bfactor Xdist
## 522      LEU       CD1           C      9.32       PHE      8.127 4.011
## 6        LEU       CD1           C      8.64       PHE      8.065 3.991
##     Xtheta planar_angle x_height x_width x_pos
## 522  24.76        165.8    3.643   1.680 below
## 6    24.16        150.8    3.641   1.633 below
```

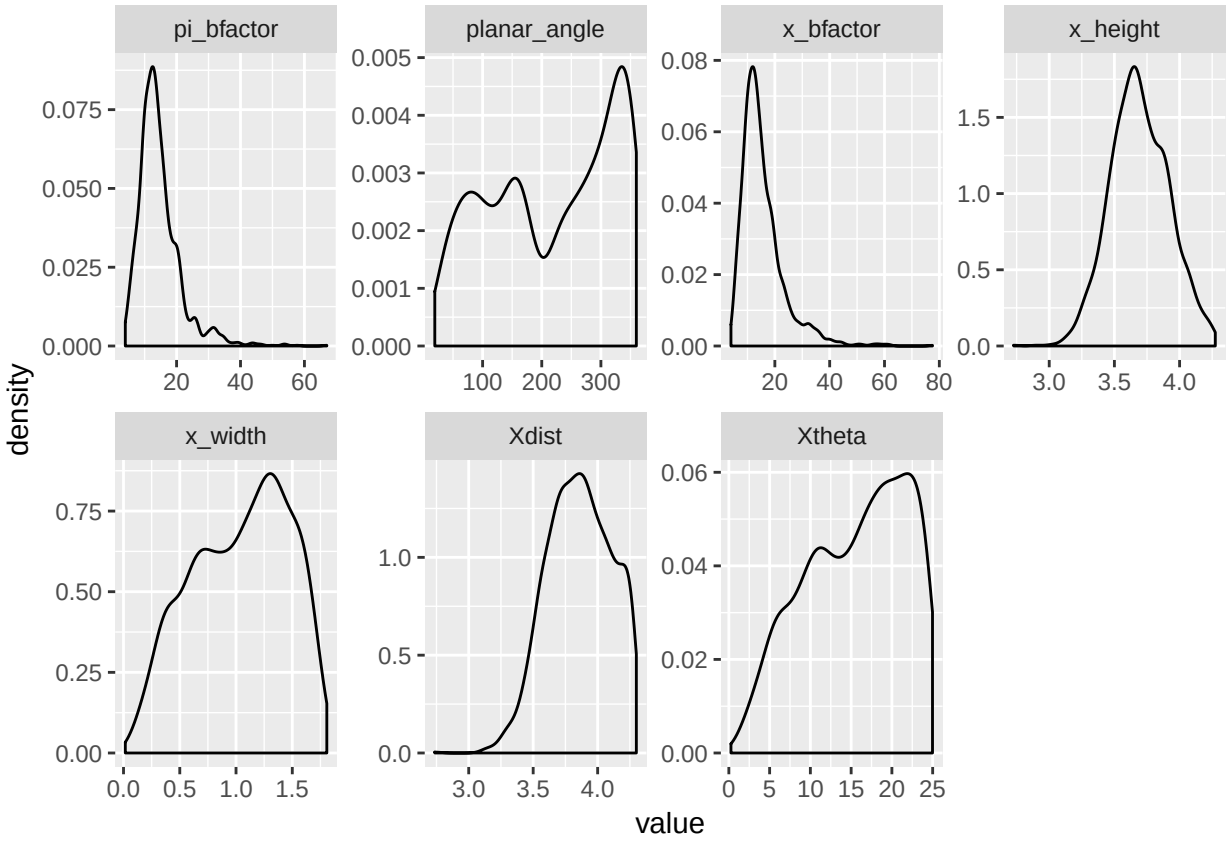
least similar i.e., max disimilarity

```
##      x_res_id x_atom_id x_atom_type x_bfactor pi_res_id pi_bfactor Xdist
## 1027      LYS        NZ           N     55.80       HIS     28.940 2.737
## 529       GLU        CB           C      9.25       TYR      9.952 4.257
##      Xtheta planar_angle x_height x_width x_pos
```

```
## 1027  4.427        325.5    2.728  0.2112 above
## 529  24.130         48.1    3.885  1.7400 below
```

2.3.2.1 Cluster with PAM partitioning round medoids

Distributions of the continuous variables: x_bfactor, pi_bfactor, Xdist, Xtheta, planar_angle, x_height and x_width.



2.3.3 Distance calculation, transformed b-factors

```
## 1597578 dissimilarities, summarized :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0114  0.3499  0.4133  0.4121  0.4748  0.8703
## Metric :  mixed ;  Types = N, N, N, I, N, I, I, I, I, I, I, N
## Number of objects : 1788
```

Which observations are least and most similar?

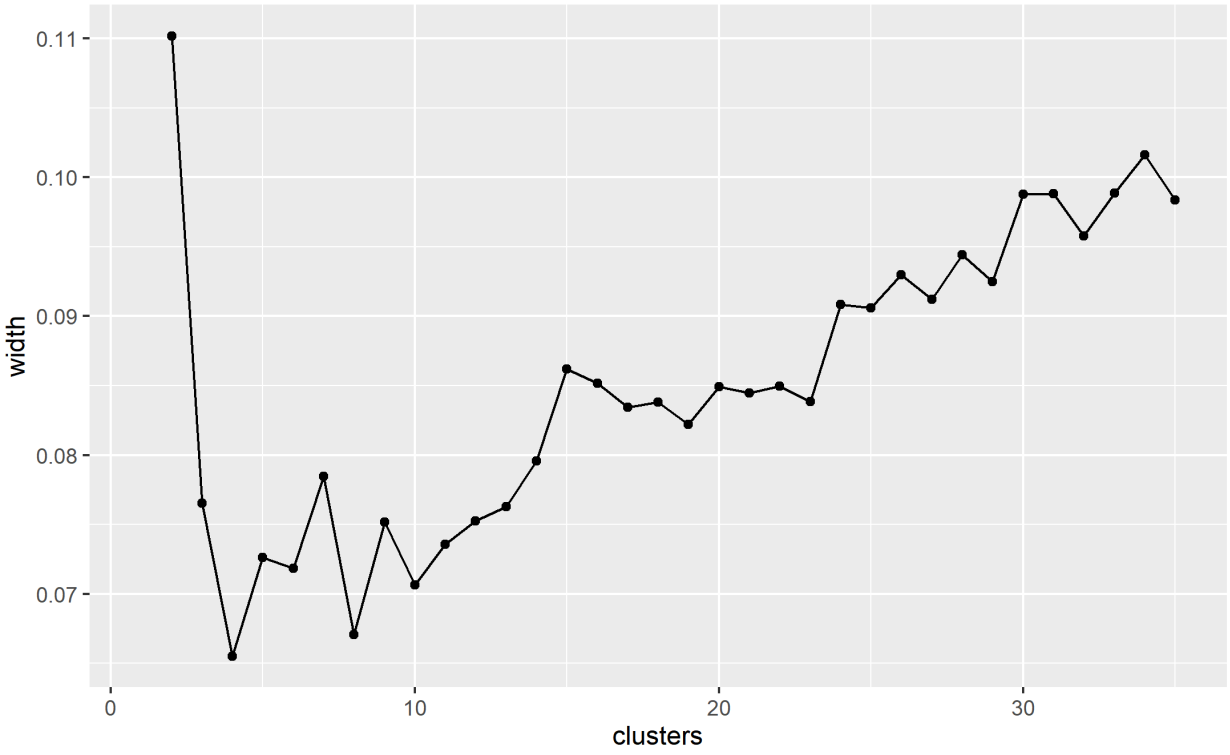
most similar i.e., min disimilarity

```
##     x_res_id x_atom_id x_atom_type x_bfactor pi_res_id pi_bfactor Xdist
## 522      LEU       CD1           C      9.32       PHE      8.127 4.011
## 6        LEU       CD1           C      8.64       PHE      8.065 3.991
##     Xtheta planar_angle x_height x_width x_pos
## 522  24.76        165.8    3.643   1.680 below
## 6    24.16        150.8    3.641   1.633 below
```

least similar i.e., max disimilarity

```
##      x_res_id x_atom_id x_atom_type x_bfactor pi_res_id pi_bfactor Xdist
## 1173      ASN        CG           C      4.67       TYR      4.512 4.043
## 1027      LYS        NZ           N     55.80       HIS     28.940 2.737
##      Xtheta planar_angle x_height x_width x_pos
## 1173 23.360         71.3    3.712  1.6030 below
## 1027  4.427        325.5    2.728  0.2112 above
```

2.3.3.1 Cluster with PAM partitioning round medoids



references

Dowle, Matt, and Arun Srinivasan. 2019. *Data.table: Extension of 'Data.frame'*. https://CRAN.R-project.org/package=data.table.

Gower, J. C. 1971. "A General Coefficient of Similarity and Some of Its Properties." *Biometrics* 27 (4): 857–71. http://www.jstor.org/stable/2528823.

Kaufman, Leonard, and Peter J. Rousseeuw. 1987. "Clustering by Means of Medoids." Edited by In: Dodge Y and editor. Amsterdam: North Holland / Elsevier.

———. 2008. "Partitioning Around Medoids (Program Pam)." In *Finding Groups in Data*, 68–125. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470316801.ch2.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2018. *Cluster: Cluster Analysis Basics and Extensions.*

R Core Team. 2018. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2017. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.