Practical on Comparative Binding Energy (COMBINE) Analysis

Ting Wang

European Media Laboratory, Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg, Germany

24-05-2002

1 Introduction

In this practical, we will show how to use the 3D structures of influenza neuraminidase (NA)-inhibitor complexes to derive quantitative structure-activity relationships (QSARs) by Comparative Binding Energy (COMBINE) Analysis. Basically, this practical will follow the procedure in the paper Wang, T. & Wade, R.C., *J. Med. Chem*. 2001, 44, 961-971, but with much simplification and automation of the steps. Both the paper and an A4-page poster are provided with this practical.

The following programs will be used:
WHATIF (http://www.smbi.kun.nl/whatif/),
AMBER7 (http://sigyn.compchem.ucsf.edu/amber/) and
GOLPE4.5 (http://www.miasrl.com/golpe.html).

2 Overview of the procedure

1. **Obtaining coordinates of NA-inhibitor complexes**
   Table 1 lists 45 complexes. 24 of which are crystal structures from the PDB, 8 were provided by Dr. Babu of BioCryst Pharmaceuticals, Inc., 9 were docked using AUTODOCK, and 4 were modeled by superposition. This modelling step will be skipped in this practical and the later steps will be based on crystal structures of complexes. As you can see in Table 1, the proteins include type A subtypes N2 and N9 and an active site mutant of the N9 subtype. The inhibitors include sialic acid and benzoic acid analogues with diverse frameworks and substitution groups (Scheme 1).

2. **Modeling a NA-inhibitor complex to obtain interaction energy components: Tutorial 1**
   - Modify the complex to retain only protein, ligand, one Ca^{2+} ion and ordered water sites.
   - Add hydrogen atoms to the ligand coordinates
   - Add hydrogen atoms to the protein and water coordinates
   - Derive force field parameters for the ligand, and the Ca^{2+} ion.
   - Generate topology and coordinate files of the complex
   - Energy-minimize the complex
   - Process the water molecules
   - Delete residues in the insertion sequences in N2 and N9 subtypes
   - Generate topology and coordinate files of the complex without the insertion residues
   - Calculate Lennard-Jones and electrostatic interaction energies between each residue (including the inhibitor, Ca^{2+}ion and designated water molecules).
   - Input energy descriptors and activity values into the Golpe program.

3. **Chemometric Analysis: Tutorial 2**

Table 1 Influenza Neuraminidase-Inhibitor Complexes

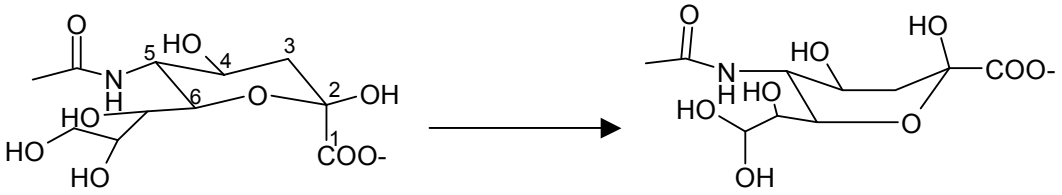
no.	inhibitor:NA [a]	code[b] of complex	mutation	inhibitor name	inhibitor charge (e)	complex source	pIC_{50}^{exp} [c]	pIC_{50}^{pred}
1	**5**:N9mutant	2qwc	Arg292Lys	Neu5Ac2en	-1	PDB	3.39[30]	2.67
2	**6**:N9mutant	2qwd	Arg292Lys	4AM	0	PDB	4.0[30]	4.58
3	**7**:N9mutant	2qwe	Arg292Lys	GNA	0	PDB	6.96[30]	5.02
4	**8**:N9mutant	2qwf	Arg292Lys	G20	0	PDB	5.28[30]	7.07
5	**9**:N9mutant	2qwg	Arg292Lys	G28	0	PDB	3.64[30]	4.97
6	**11**:N9mutant	2qwh	Arg292Lys	G39	0	PDB	4.89[30]	4.77
7	**8**:N9	2qwi		G20	0	PDB	7.70[30]	7.26
8	**9**:N9	2qwj		G28	0	PDB	6.64[30]	6.74
9	**11**:N9	2qwk		G39	0	PDB	8.70[30]	8.06
10	**5**:N9	1nnb		Neu5Ac2en	-1	PDB	4.70[30]	3.94
11	**7**:N9	1nnc		GNA	0	PDB	8.70[30]	8.45
12	**10**:N9	1bji		G21	0	PDB	8.70[10]	7.42
13	**3**:N9	1iny	Ser370Leu	ePANA	-2	PDB	3.16[11]	4.27
14	**12**:N2	1ivd		ST1	-1	PDB	3.12[13]	1.65
15	**13**:N2	1ivc		ST2	0	PDB	1.7[13]	1.48
16	**14**:N2	1ive		ST3	0	PDB	1.4[13]	1.44
17	**15**:N2	1ing		ST5	-1	PDB	2.40[13]	4.15
18	**16**:N2	1inh		ST6	0	PDB	2.30[13]	2.45
19	**5**:N2	1ivf		Neu5Ac2en	-1	PDB	4.82[13]	3.90
20	**3**:N2	1inx		ePANA	-1	PDB	4.7[11]	4.32
21	**18**:N9	bc1	Gly336Asn	bcx-140	0	Babu	5.30[14]	4.24
22	**19**:N9	bc2	Gly336Asn	bcx-384	-1	Babu	1.96[14]	3.31
23	**20**:N9	bc3	Gly336Asn	bcx-167	0	Babu	3.70[14]	4.56
24	**21**:N9	bc4	Gly336Asn	bcx-352	0	Babu	5.00[14]	4.18
25	**22**:N9	bc5	Gly336Asn	bcx-141	+1	Babu	2.52[14]	3.46
26	**23**:N9	bc6	Gly336Asn	bcx-448	0	Babu	3.00[14]	3.18
27	**24**:N9	bc7	Gly336Asn	bcx-1023	+1	Babu	4.74[14]	3.30
28	**25**:N9	bc8	Gly336Asn	bcx-869	0	Babu	3.55[14]	4.75
29	**26**:N9	lma	Gly336Asn	LMA	-1	autodock	3.60[12]	4.59
30	**27**:N9	lmb	Gly336Asn	LMB	-1	autodock	4.70[12]	4.30
31	**28**:N9	lmc	Gly336Asn	LMC	0	autodock	3.12[12]	4.07
32	**29**:N9	qwm	Gly336Asn	QWM	0	autodock	5.30[12]	5.55
33	**30**:N9	qwl		QWL	-1	autodock	7.32[12]	6.89

Table 1 (continued)

34	**17**:N9	bc9	Gly336Asn	ST8	+1	autodock	4.15[15]	5.01
35	**17**:N2	dk3		ST8	+1	autodock	4.15[15]	4.27
36	**6**:N9	wdd		4AM	0	superposition	6.50[10]	5.84
37	**7**:N2	dk1		GNA	0	superposition	8.81[8]	7.45
38	**18**:N2	dk2		bcx-140	0	superposition	5.00[13]	3.63
39	**6**:N2	dam		4AM	0	superposition	6.41[8]	6.68
40	**1**:N9mutant	2qwb	Arg292Lys	α-Neu5Ac	-1	PDB	0.70[41]	2.63[d]
41	**1**:N9	1mwe		α-Neu5Ac	-1	PDB	1.70[41]	6.17[d]
42	**1**:N2	2bat	Asp338Asn	α-Neu5Ac	-1	PDB	2.7[13]	6.94[d]
43	**4**:N2	1inw		aPANA	-1	PDB	2.7[11]	4.88[d]
44	**31:**N9	rw9		bcx-1812	0	autodock	e)[18]	8.36
45	**32:**N2	rw2		bcx-1812	0	autodock	e)[18]	7.13

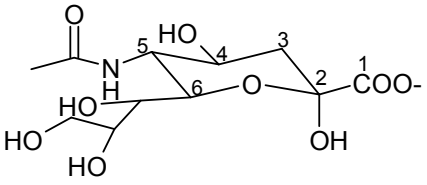
a): Inhibitors are shown in Scheme 1. The N9 subtype is referred to as "N9mutant" if there is a mutation in the active site.
b): PDB identifier if with 4 letters
c): Experimental data were taken from the references noted in parentheses
d): External predicted activities for the N2+N9 model derived with 39 complexes.
e): The pIC_{50} range against 15 different strains of type A NA is 8.85-10.

Scheme 1.

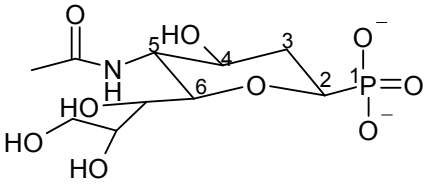


chair form boat form

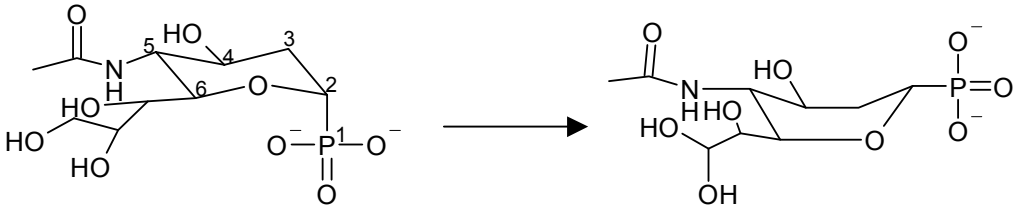
1 (α-sialic acid)



2 (β-sialic acid)*

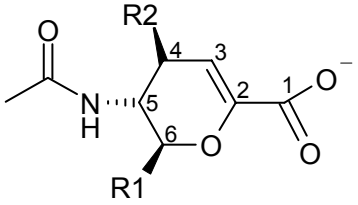


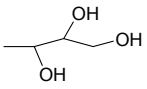
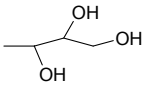
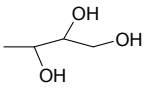
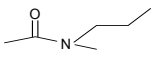
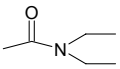
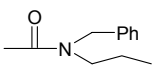
3 (equatorial-PANA)



chair form boat form

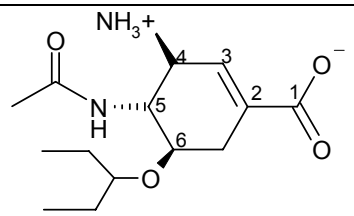
4 (axial-PANA)



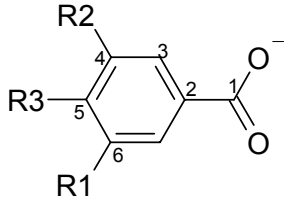
	R1	R2
5		–OH
6		$-NH_3^+$
7		$-NH^+=C(NH_2)_2$
8		$-NH^+=C(NH_2)_2$
9		$-NH_3^+$
10		$-NH_3^+$

* compound **2** (β-sialic acid) does not form a complex with NA.

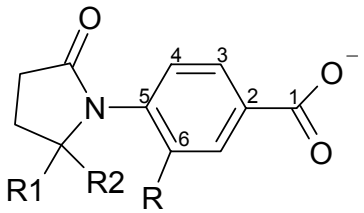
Scheme 1 (continued)



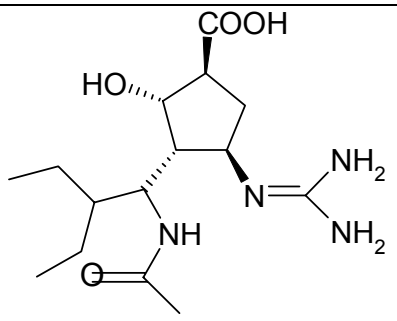
11



	R1	R2	R3
12	$-NO_2$	–OH	$-NHCOCH_3$
13	–OH	$-NH_3^+$	$-NHCOCH_3$
14	$-NH_3^+$	–H	$-NHCOCH_3$
15	$-NHCOCH_2OH$	–H	$-NHCOCH_3$
16	$-NHCOCH_2NH_3^+$	–H	$-NHCOCH_3$
17	$-NH^+=C(NH_2)_2$	$-NH^+=C(NH_2)_2$	$-NHCOCH_3$
18	$-NH^+=C(NH_2)_2$	–H	$-NHCOCH_3$
19	–H	–CH=NOH	$-NHCOCH_3$
20	$-NH^+=C(NH_2)_2$	–H	$-NHSO_2CH3$
21	$-NH^+=C(NH_2)_2$	–H	$-CONHCH_3$
22	$-NH^+=C(NH_2)_2$	$-NH^+=C(NH_2)_2$	–H
23	$-NH^+=C(NH_2)_2$	–H	–H
24	$-NH^+=C(NH_2)_2$	–H	$-SO_2NH_3^+$
25	$-NH^+=C(NH_2)_2$	–H	$-CH_2SOCH_3$



	R	R1	R2
26	$-NH^+=C(NH_2)_2$	–H	–H
27	$-NH^+=C(NH_2)_2$	–H	$-CH_2OH$
28	–H	$-CH_2OH$	$-CH_2OH$
29	$-NH^+=C(NH_2)_2$	$-CH_2OH$	$-CH_2OH$
30	$-NHCH(CH_2CH_3)_2$	$-CH_2OH$	$-CH_2OH$



31

3 Tutorial 1: Modelling a neuraminidase-inhibitor complex to obtain interaction energy components

In this tutorial, we will model the complex of an N9 neuraminidase mutant with compound 7 (GG167/Relenza), which is given the residue name GNA in the PDB file 2qwe.brk. This is complex number 3 in Table 1.

3.1 Let's first have a look at the original PDB file 2qwe.brk

```
% cp $COMBINE/2qwe.brk .  /* copy the file to your working directory*/
% nedit 2qwe.brk &
```

This file contains:
Protein: NA, from residue 82 to 468, 388 residues in total, without hydrogen atoms.
Inhibitor: GNA, residue 800, without hydrogen atoms.
Two Ca^{2+} ions: CA, residue 989 (far from the binding site) and 999 (near the binding site).
Sugar molecules: NAG and MAN.
Lots of bound water molecules.

3.2 Visualize and modify the complex

For COMBINE analysis, 2qwe.brk has to be modified by removing all sugars, the Ca2+ ion residue 989, and water molecules surrounding the sugars. The modified complex is saved as 2qwe.pure.pdb. Here we use Rasmol to visualize the structure of 2qwe.pure.pdb (Figure 1)

```
% cp $COMBINE/2qwe.pure.pdb .
```

```
Setup Rasmol:
Rasmol>
PDB file name: 2qwe.pure.pdb
Rasmol> Select GNA
23 atoms selected
Display: Sticks
Rasmol> select 82-468
3154 atoms selected
Display: Ribbons
Rasmol> select 999
1 atom selected
Display: Spacefill
Rasmol> select HOH
380 atoms selected
Display: Ball&Stick
```

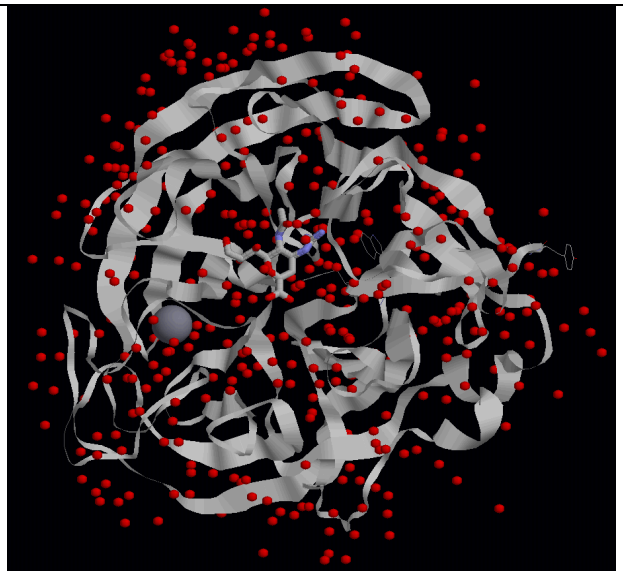


Figure 1. Visualization of 2qwe.pure.pdb in Rasmol

As you can see, 2qwe.pure.pdb contains only heavy atoms. In the following steps, we will add hydrogen atoms to the inhibitor, the protein and the water molecules.

3.3 Add hydrogen atoms to inhibitor GNA

Hydrogens can be easily added to the inhibitor coordinates by InsightII or other software. Here we skip this step, and you should just copy the file with hydrogens already added, $COMBINE/2qwe.inhibitor.H.pdb, to your working directory.

```
% cp $COMBINE/2qwe.inhibitor.H.pdb .
```

3.4 Add polor hydrogen atoms to protein and waters by WHATIF

```
% cp $COMBINE/2qwe.inhibitor.H.pdb u.pdb /* Rename the file for use in
    WHATIF scripts */
% $WHATIF/DO_WHATIF.COM < $COMBINE/wiaddh_in > ! whatif.log
 /* This command outputs the file uH.pdb with polar hydrogens added */
```

Have a look at the WHATIF log file whatif.log and pay attention to residues HIS and ASN.

Questions: What protonation states are assigned to the 7 HIS residues?
Is there flipping on any ASN?

To run next script `whatif2uhbd`, you need to edit the output file uH.pdb to remove the inhibitor and the Ca2+ ion.

```
% nedit uH.pdb &
```

Copy the part containing the inhibitor and the Ca^{2+} ion as temp.pdb, then delete it from uH.pdb and save the file as uH.whatif.pdb

```
% $COMBINE/whatif2uhbd < uH.whatif.pdb > uu.pdb /* convert pdb formats */
```

Now, we put the inhibitor and the Ca2+ ion back to uu.pdb

```
% nedit uu.pdb &
```

Paste temp.pdb to uu.pdb, placing the Ca^{2+} and the inhibitor coordinates after residue 468.
Save uu.pdb as uu.whatif.pdb

```
% $COMBINE/whatiftopdb uu.whatif.pdb 2qwe.whatif.pdb
/*convert the format of waters*/
```

NOTE: if you want to try Step3.4 again, the files generated automatically by WHATIF must be deleted. These are WHATIF.FIG, pdbout.txt, pdbout.tex, tmp.pdb, u.pdb, uH.pdb, uu.pdb.

3.5 Edit 2qwe.whatif.pdb for use in Xleap of AMBER7

In the later steps, we will use the program AMBER7 to further model the complex and the file 2qwe.whatif.pdb will be input into the Xleap module of AMBER7. To fit the format in Xleap, the file 2qwe.whatif.pdb needs some modifications.

```
% nedit 2qwe.whatif.pdb &
```

Rename the N-terminal residue 82 as ARG.
Rename the first 3 H atoms as H1, H2 and H3.
Rename the last residue 468 as Leu.

Delete the last line of residue 468 "ATOM 3817 O2 LEU 468" /* This atom will be added later by Xleap */
Rename the Ca^{2+} residue as CAA.
Rename 18 CYS residues as CYX. /* The 18 cystine residues in NA form disulfide bonds and thus they have to be renamed as CYX */

Save as 2qwe.modif.pdb.

Now the file is ready for input to Xleap with full atoms in the inhibitor and the water molecules and heavy atoms and polar hydrogen atoms in the protein. Non-polar hydrogen atoms will be added to the protein later by Xleap.

3.6 Derive force field parameters for the inhibitor GNA

For proteins and water molecules, the Xleap module of AMBER7 can assign atomic partial charges and other force field parameters automatically, but for the inhibitor and the Ca2+ ion, some preparation work needs to be done.

```
% grep GNA < 2qwe.modif.pdb > GNA.pdb /* extract ligand from the complex file */
```

Setup AMBER7:

```
% antechamber –i GNA.pdb –fi pdb –o GNA.prep –fo prepi –c bcc –rn GNA
/* generate AM1-BCC atomic charges of the inhibitor and save them in the last column of GNA.prep*/
% parmchk –i GNA.prep –f prepi –o GNA.parm –p $AMBERHOME/dat/leap/parm/gaff.dat
/* generate force field parameters missing in the general force field (gaff.dat) for the inhibitor */
```

Take a look at the atomic charges in GNA.prep and other force field parameters in GNA.parm.

3.7 Derive force field parameters for Ca2+

Same as for the inhibitor, some preparation work needs to be done for the Ca2+ ion.

```
% grep CAA < 2qwe.modif.pdb > CAA.pdb /* extract Ca2+ from the complex file */
% cp $AMBERHOME/dat/leap/parm/parm94.dat my_parm94.dat
% nedit my_parm94.dat &
```

Go to the last part of the file and put the line
" C0 1.74 0.0465 Ca2+ JCC 12 (1991) 1125-1128"
behind the line
"EP 0.0 ……"
/* This is to define a new atom type C0 for Ca2+ with vdw radius 1.74Å and epsilon value 0.0465kcal/mol */

Start xleap:

```
% xleap &
```

In xleap:

>CAA = loadpdb CAA.pdb

>edit CAA /* create a unit called CAA for Ca^{2+} */

You get a new window, double click the green dot in the window and pull down the `Edit` menu -> `Edit selected atoms`, you will get a table window, fill out the table as follows:

Set TYPE = C0, CHARGE = 2.0000

Now, go to the `Table` menu, choose `Save and quit`, back to the unit editor window, then go to the `Unit` menu, choose `quit`, back to the main window of Xleap.

>saveoff CAA CAA.lib /* Save Ca2+ information in CAA.lib */

3.8 Load 2qwe.modif.pdb into Xleap and generate topology and coordinate files.

We have prepared force field parameters for the inhibitor GNA and the Ca2+ ion, now we can use Xleap to get the topology and coordinate files for the complex.

In xleap:

>source leaprc.ff94 /* load libraries*/

You will get the following messages in the Xleap window:

>loadamberparams my_parm94.dat /* load modified amber94 force field */

>loadamberparams gaff.dat /*load the general force field for organic compounds */

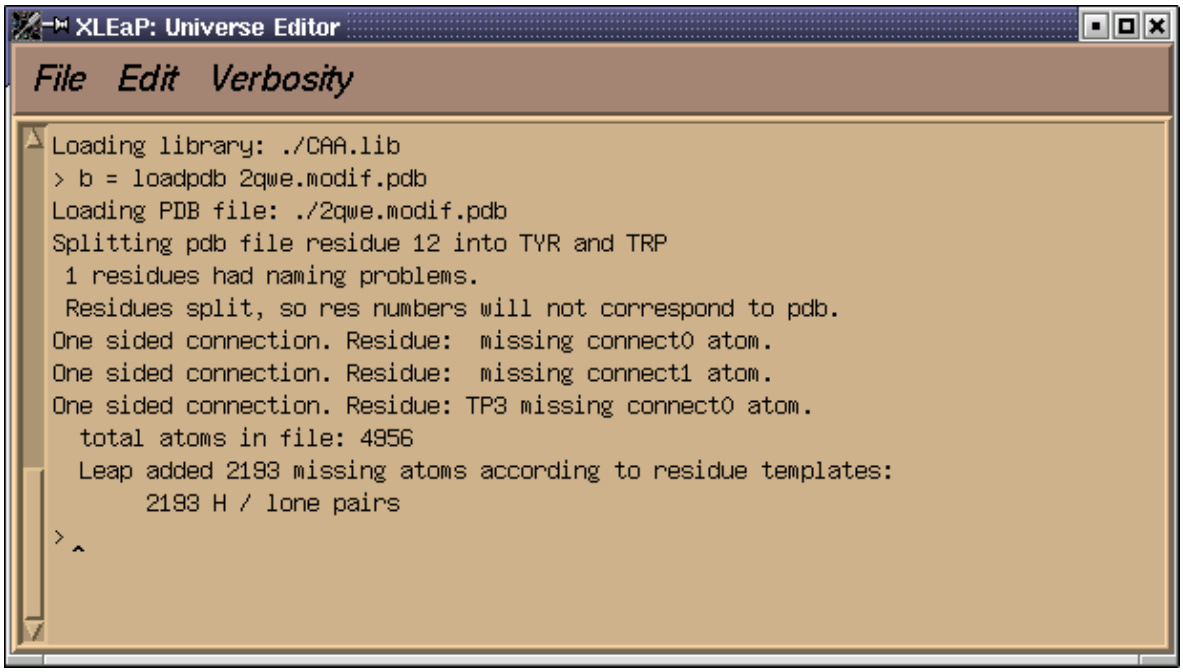
>loadamberprep GNA.prep /* load inhibitor */

>loadamberparams GNA.parm /* load parameters of inhibitor GNA*/

>loadoff CAA.lib /* load library of Ca2+*/

>b = loadpdb 2qwe.modif.pdb /* load complex to get non-polar Hs */

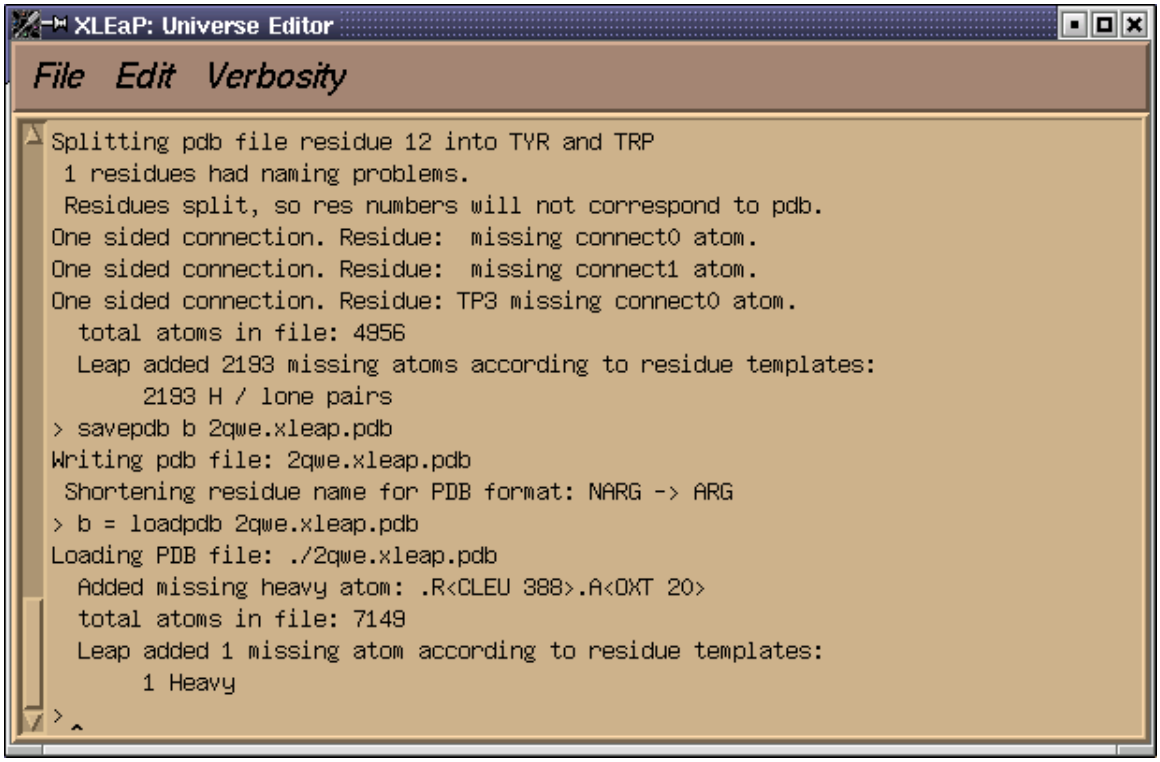
You should get the following messages in the Xleap window:



>savepdb b 2qwe.xleap.pdb /* save complex with all Hs */

>b = loadpdb 2qwe.xleap.pdb /* load complex with all Hs */

This time you got only one message saying the C-terminal oxygen atom was added:



```
>bond b.11.SG b.337.SG          /* add 9 disulfide bonds */
>bond b.43.SG b.48.SG
>bond b.95.SG b.113.SG
>bond b.103.SG b.150.SG
>bond b.152.SG b.157.SG
>bond b.198.SG b.211.SG
>bond b.200.SG b.209.SG
>bond b.238.SG b.256.SG
>bond b.341.SG b.367.SG
> saveamberparm b 2qwe.tpp 2qwe.rst /* generate topology file and coordinate file of
  the complex */
>quit
```

You can save the above commands into a file. eg. leap_script, and use tleap to run it:

```
% tleap -f leap_script
```

3.9 Minimize the energy of the complex.

Given the topology and coordinate files (2qwe.tpp and 2qwe.rst), We can use the Sander module in AMBER7 to do energy minimization. The main purpose is to optimize the positions of the modeled hydrogen atoms, the inhibitor and the water molecules.

```
% cp $COMBINE/min.in .       /* copy the input file */
```

Have a look at min.in:

```
% nedit min.in &
```

The protein non-hydrogen atoms are restrained to their crystallographic positions by a harmonic potential with a force constant of 32kcal/(mol.Å)^2 while the hydrogen atoms, the inhibitor and water molecules are unrestrained. A non-bonded cutoff of 10Å and a distance-dependent dielectric constant ($\varepsilon=r_{ij}$) are used. To save time, we carry out only 200 steps.

```
% sander –O –i min.in –o 2qwe.min.out –p 2qwe.tpp –c 2qwe.rst –r 2qwe.min.xyz –ref 2qwe.rst
```

/* This step will take about 5 min */

Take a look at the output file 2qwe.min.out.

Generate a minimized coordinate file in PDB format:

```
% ambpdb  -p 2qwe.tpp < 2qwe.min.xyz > 2qwe.min.pdb
```

The minimized structure is in 2qwe.min.pdb

3.10 Process water molecules

After energy minimization, all bound water molecules, except one (WAT478) involved in interactions between the inhibitors and the protein, are discarded. The retained water molecule is located between the C5 and C6 pockets. It accepts a hydrogen-bond from the C5 position amide hydrogen of the inhibitor and donates hydrogen-bonds toGlu197 (Glu277 in 2qwe.brk) and Glu147 (Glu227 in 2qwe.brk).

You can do this with Nedit or just copy the file from $COMBINE

```
% cp $COMBINE/2qwe.w1.pdb .
```

3.11 Delete gap residues

Although both N2 and N9 subtypes contain 388 amino acid residues and conserved active site residues, they differ in most positions in their sequence alignment. The alignment of N2 and N9 subtypes revealed 10 residue insertions: 5 in N2 and 5 in N9 (**Figure 2**).

The insertion residues (where there are gaps in the other sequence) should be deleted.

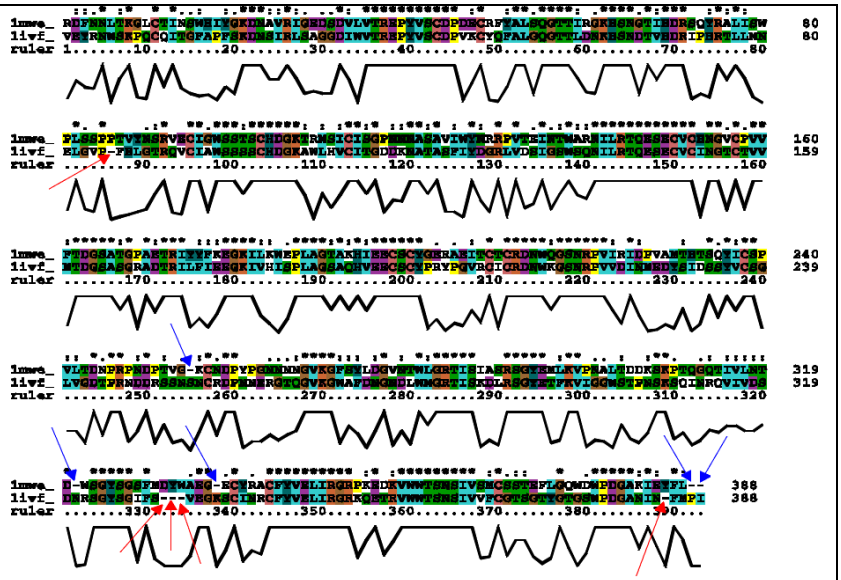


Figure 2. Sequence alignment of N2 and N9 subtypes done with ClustalX1.8. The sequence identity is ca. 50% and there are 10 residues where there are insertions/gaps (gaps shown by arrows; N9: blue arrows and N2: red arrows)

```
% $WHATIF/n9align 2qwe.w1.pdb 2qwe.w1.alig.pdb
/* delete 5 insertion residues */
```

Now the complex contains 383 protein residues, one inhibitor molecule, one Ca2+ ion, one bound water molecule, 386 residues in total.

3.12 Generate topology and coordinate files of the complex without gap residues

```
% xleap
```

In xleap:

```
>source leaprc.ff94          /* load libraries*/
>loadamberparams my_parm94.dat   /* load modified amber94 force field */
>loadamberparams gaff.dat /*load the general force field for organic compounds
*/
>loadamberprep GNA.prep           /* load inhibitor */
>loadamberparams GNA.parm       /*  load parameters of inhibitor GNA*/
>loadoff CAA.lib                  /* load library of Ca2+*/
>b = loadpdb 2qwe.w1.alig.pdb     /* load complex */
> saveamberparm b 2qwew1lg.tpp 2qwew1lg.rst /* generate topology file and
coordinate file of the complex  without gap residues */
>quit
```

3.13 Calculate Lennard-Jones and electrostatic interaction energies between each residue (1-386).

We will use the Analysis module in AMBER7 to calculate the interaction energies between each residue.

```
% cp $COMBINE/anal.in .       /* copy the input file */

% anal -O -i anal.in -o 2qwew1lg.aout -p 2qwew1lg.tpp -c
2qwew1lg.rst
```

Take a look at the output file 2qwew1lg.aout, how big is this file?

3.14 Input energies from the AMBER analysis module into Golpe45 and extract Lennard-Jones and electrostatic interaction energies between the inhibitor GNA and each protein residue and the bound water molecule.

The procedure described in steps 3.1-3.13 should be repeated for each complex that will be used for generating the COMBINE QSAR model. We will only model one complex in this practical and will use coordinates and energy components generated previously for other complexes to build a COMBINE model. The steps necessary to process a set of complexes so that chemometric analysis can be carried out with GOLPE are **as follows (do not do in this practical).**

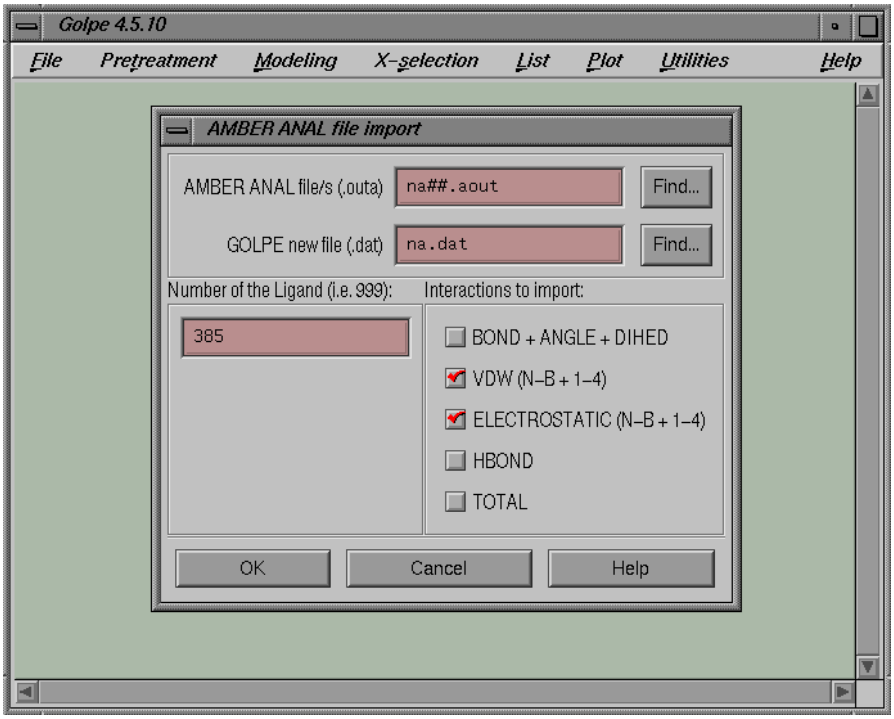
Rename all Anal output files (.aout) in sequence, to give them names such as: na01.aout, na02.aout, …, na39.aout.

Start the Golpe4.5 program.

```
% golpe
```

Input all aout files into Golpe to obtain a single dat file for all complexes using the “file” pull-down menu:

```
File -> Input interactions -> AMBER
```



Fill out the above “AMBER ANAL file import” window, press OK.

Back to the main window of Golpe and pull down the File menu again:

```
File -> Open data file -> Select a file: na.dat
```

The file na.dat will have 770 energy descriptors for each complex, and you need to add the activity (pIC50 value) as the 771th variable:

```
Utilities -> Add new variables -> How many new variables: 1,
OK.  /* give the pIC50 value  for each object (complex) */
```

Take a look at the file `na.dat`, which has been modified to have the 771th variable.

Now, you are ready for chemometric analysis by Golpe, which will be next tutorial (Tutorial 2).

4 Tutorial 2

In this tutorial, we will show how to use GOLPE to do chemometric analysis and derive relationships between energy descriptors and activities for 39 NA-inhibitor complexes. The results shown in the paper *J. Med. Chem*. 2001, 44, 961-971 will be reproduced here.

```
% cp $COMBINE/na39com.dat . /* we use the Golpe dat file of 39 complexes*/
```

In this data file (`na39com.dat`), each complex has 770 energy descriptors (x-variables) and 1 activity value (y-variable, pIC50). The first 385 and the second 385 x-variables, respectively, are the Lennard-Jones and the electrostatic interaction energies between the inhibitor and each protein residue and the bound water molecule.

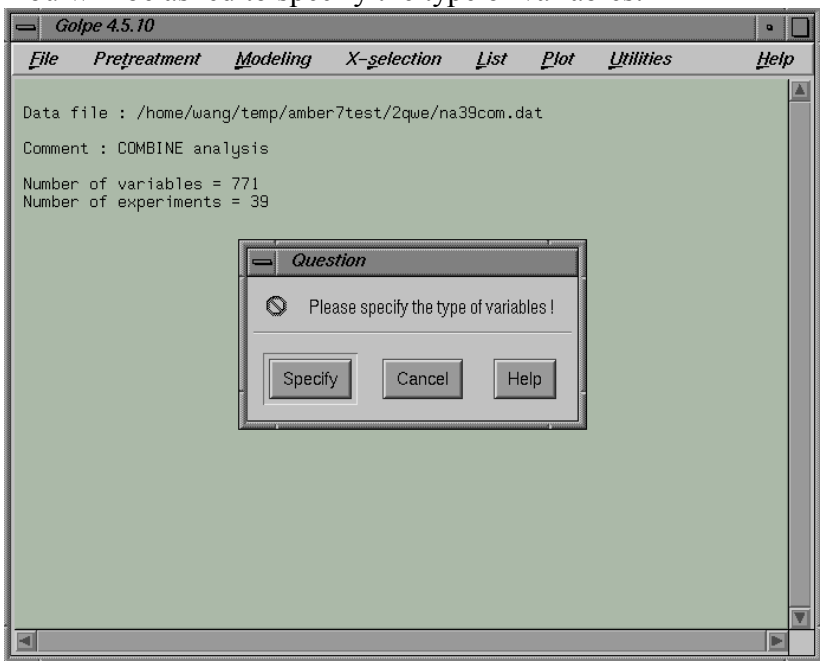
Start Golpe4.5.10

```
% golpe
```

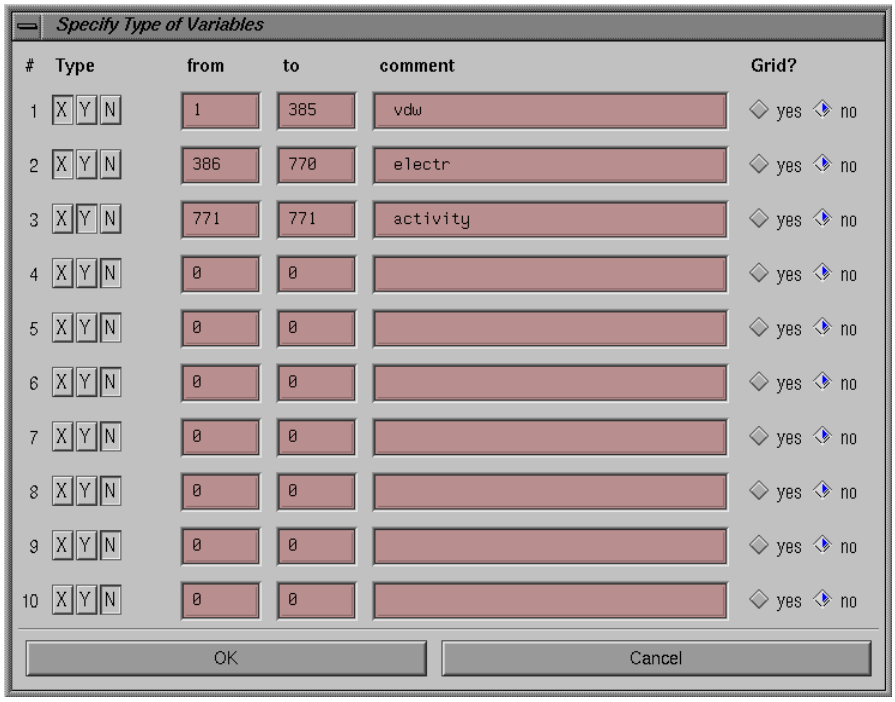
4.1 Input the data and specify the variables

```
File-> Open data file -> Select a file: na39com.dat
```

You will be asked to specify the type of variables:



Click `Specify` to get a table window: set type “X ” for the 1-385th variables with the comment “vdw”, type “X” for the 386-770th with the comment “electr ” and type “Y” for the 771th variable with the comment “ activity ”. In this way, the X-variables are divided into two blocks: a vdw block and an electrostatic block.



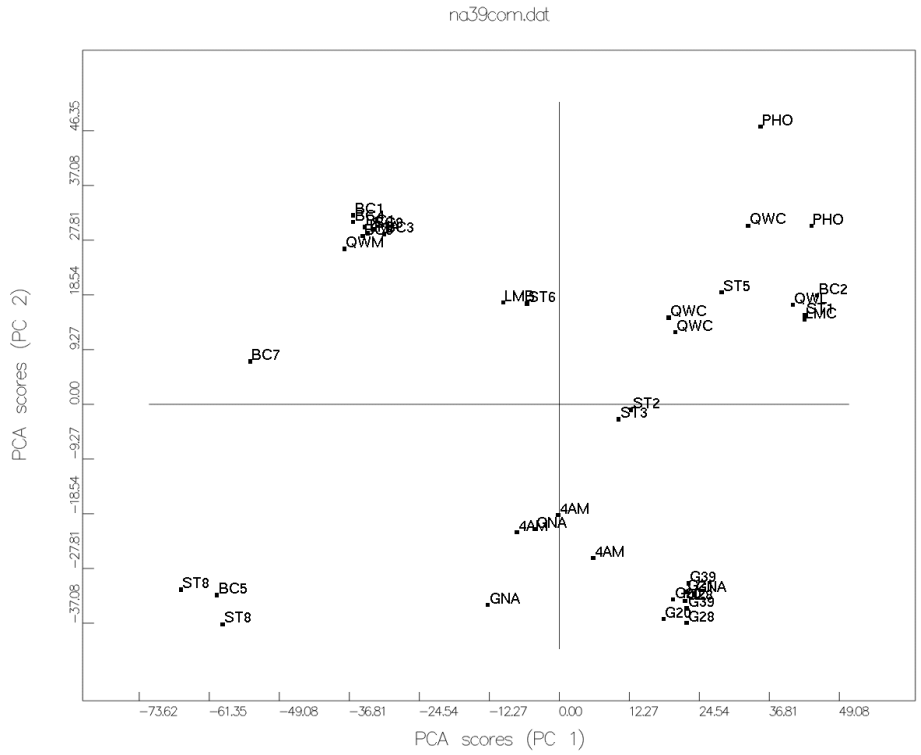
4.2 Principal Component Analysis (PCA) in the energy space.

You might be interested to know how the 39 complexes are distributed in the space defined by their interaction energies. To get the answers, you can run principal component analysis (PCA) on the dataset.

`Modeling -> Generate PCA model -> dimensionality: 5, OK.`

`Plot -> 2D Plot-> PCA-scores-> x axis:1, y axis:2, OK.`

And click the `right button` of your mouse to get the following plot:



Try the 3D plot yourself.

Question: Which complexes cluster together? What do they have in common?

4.3 Partial Least Square (PLS) analysis on the original data

The main purpose of our chemometric analysis is to obtain relationships between the energy descriptors and the activities, namely the following equation:

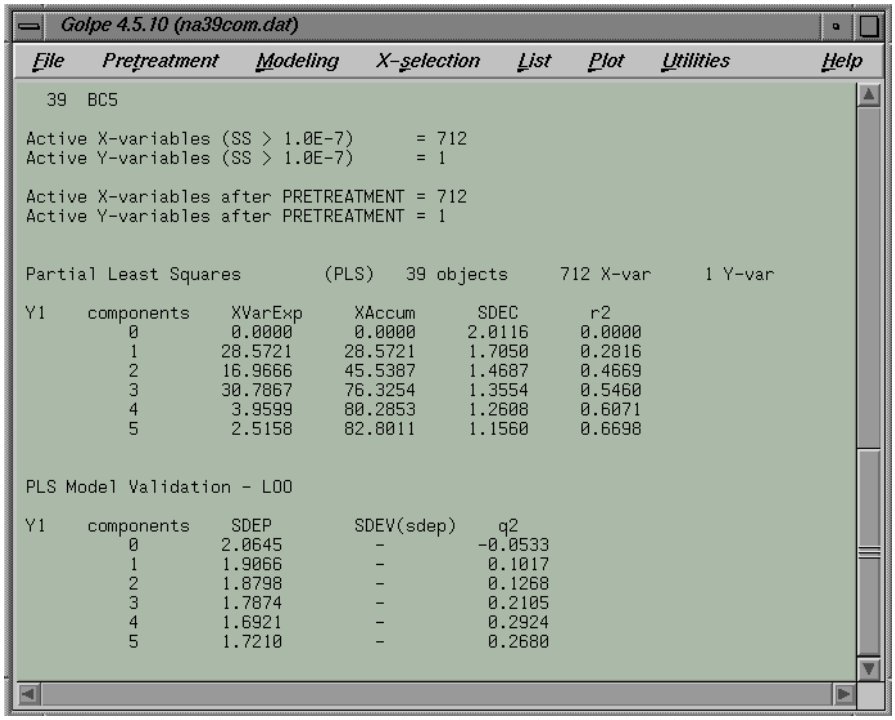
$$pIC_{50} = \sum_i w_i^{vdw} u_i^{vdw} + \sum_i w_i^{ele} u_i^{ele} + C$$

As the number (770) of x-variables is much more than the number (39) of the objects (complexes), we will use the Partial Least Square (PLS) analysis to build models.

```
Modeling -> Generate PLS model -> dimensionality: 5, OK.

Modeling -> Validate PLS model -> Max.dimensionality: 5,
validation mode: Leave One Out, OK.
```

You will see the models have very low R^2 and Q^2 values:

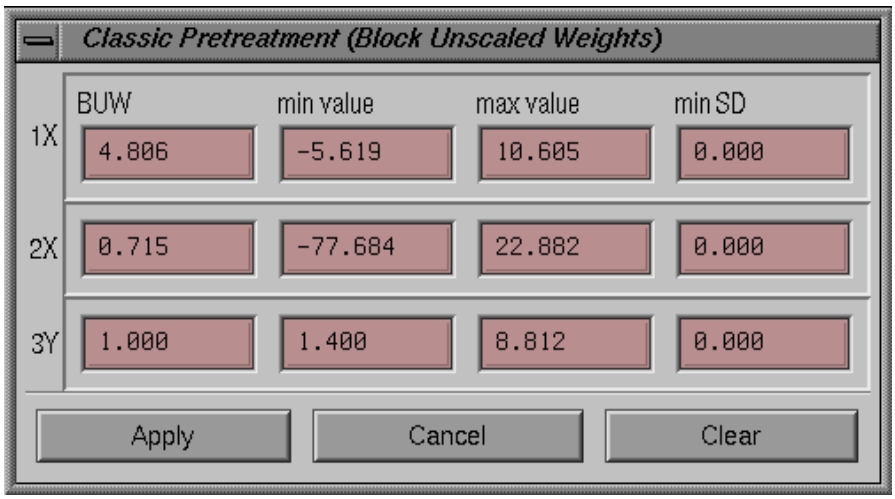


Think about what can be done to improve the models.

4.4 Block Unscaled Weights (BUW) pre-treatment on the original data.

It appears some pre-treatment work has to be done on the original data. Here we use the Block Unscaled Weights (BUW) pre-treatment. That means scaling the x-variables in each block (vdw block and electrostatic block) to give them the same importance:

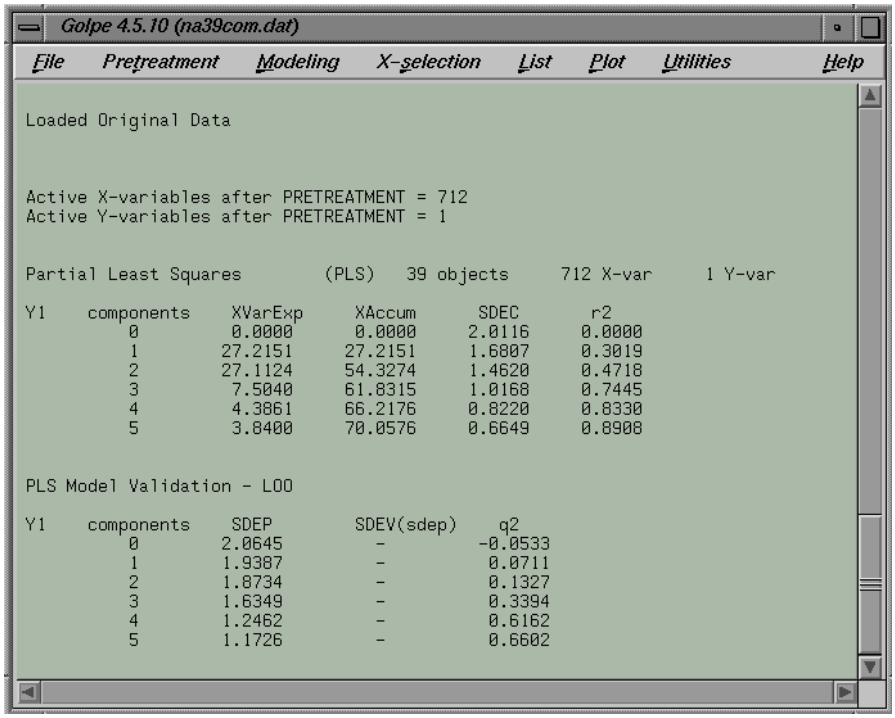
`Pretreatment -> Classic Pretreatment -> Setup pre-treatment (BUW)`



Click Apply to implement the pre-treatment.

4.5 Partial Least Square (PLS) analysis after BUW pretreatment.

Re-build PLS models, we will see the models become much better:



4.6 Variable selection

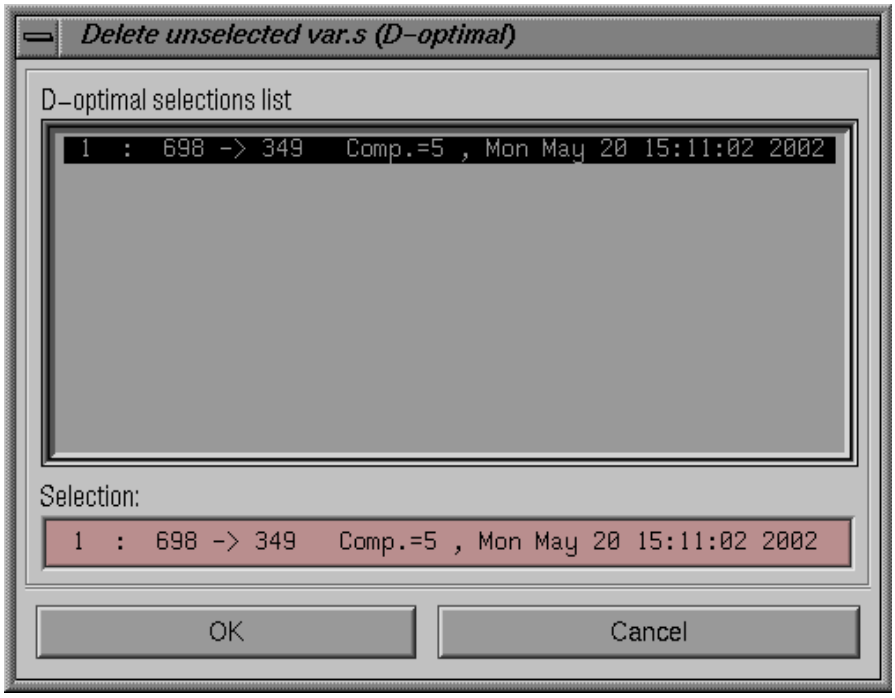
Models can be further improved by applying variable selection methods. Here we will use two variable selection procedures (D-optimal and Fractional Factorial Design) to get rid of noise variables and further improve the predictive ability.

`x-Selection -> D-optimal preselection -> Max.dimensionality: 5, OK.`

Wait until a blue window appears, then:

`Pretreatment -> delete unselected var.s (D-optimal).`

You will obtain the following window:



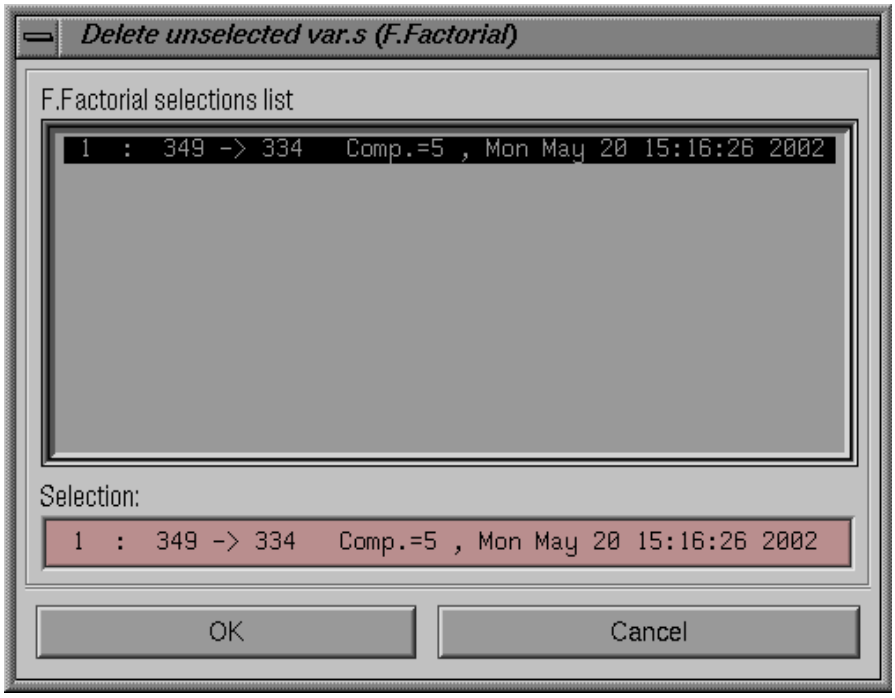
Click the list and press OK, the x-variable will be reduced to 349 (less than half of the original).

Question: Build PLS models yourself, see any difference?

```
x-Selection-> F.Factorial selection -> Max.dimensionality: 5,
execution: window, OK.
```

The calculation will take ca. 2 min, then:

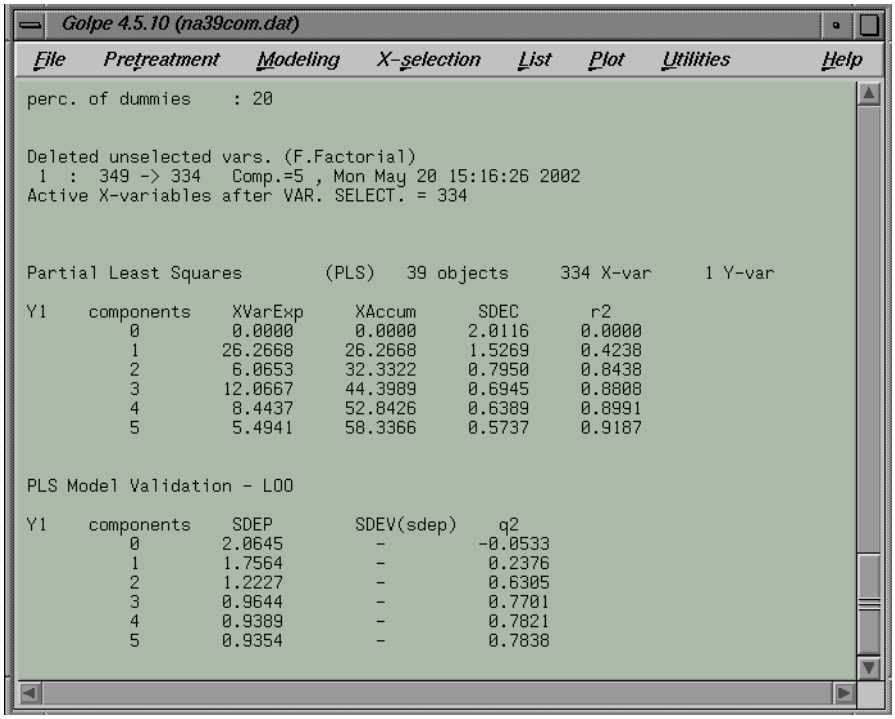
```
Pretreatment -> delete unselected var.s (F.Factorial).
```



Click the list, the x-variables are reduced to from 349 to 334.

4.7 Final PLS models

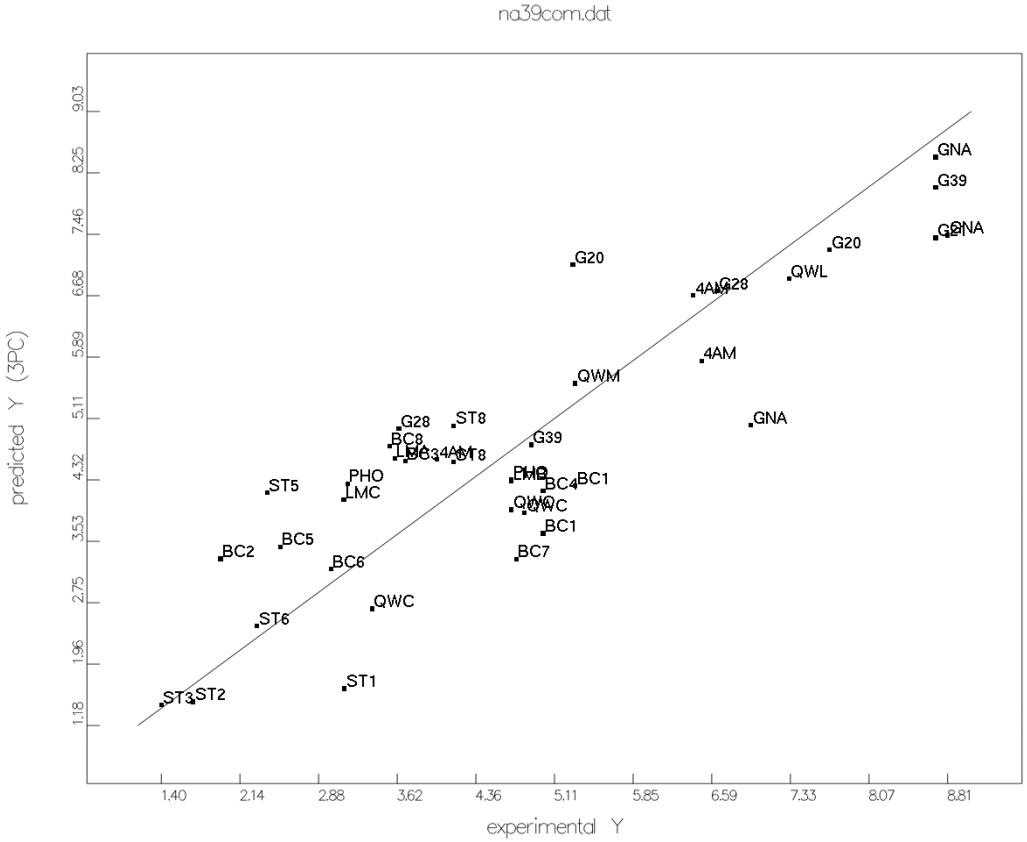
4.7.1 Now let's build our final PLS models:



Compare the R^2 and Q^2 values with those before variable selection.

4.7.2 Generate a prediction plot:

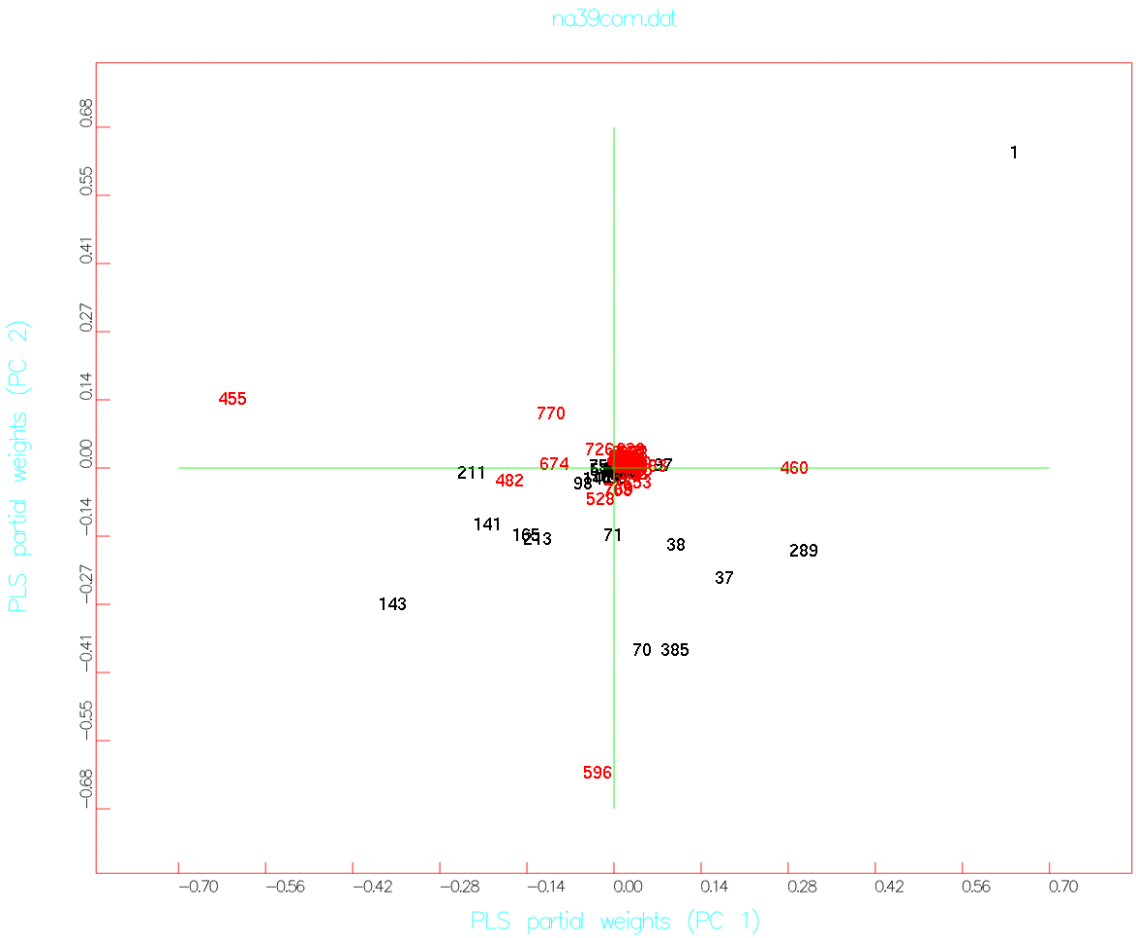
Plot -> 2D plot -> Pred. vs Exper. -> dimensionality: 3, OK.



This plot corresponds to Figure 2 in the paper.

4.7.3 Investigate the contributions of the x-variables to the activity (pIC50):

`Plot -> 2D plot -> PLS partial weights -> X axis: 1; Y axis: 2; mark block: 2; OK.`



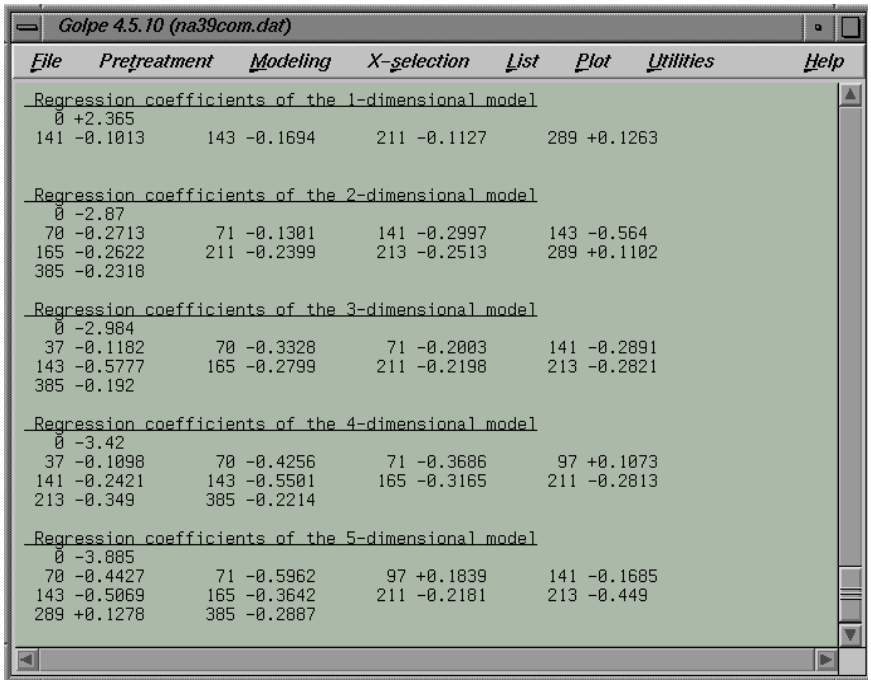
This plot corresponds to Figure 4a in the paper.

Try to reproduce Figure 4b yourself.

Questions: Which variables predominantly define the first two latent variables?
Which residues are they from?
How many latent variables are required to describe the models?

4.7.4 List the real PLS coefficients of significant x-variables:

```
List -> PLS coefficients -> real -> Enter the minimal value: 0.1
```

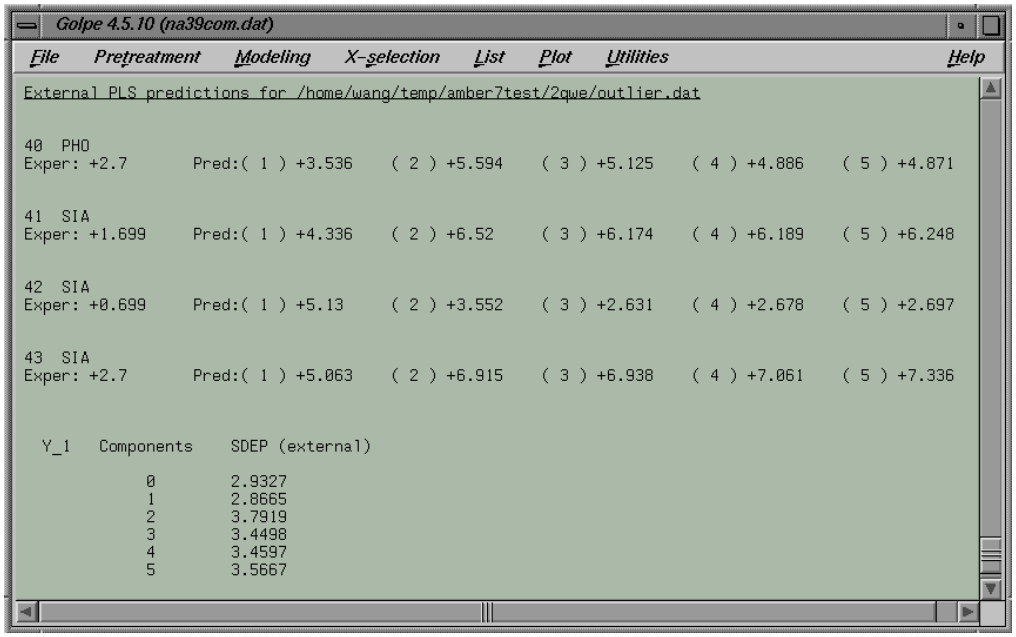


You will see all variables with real PLS coefficients of > 0.1 are vdw energies (in the first 385). Try the minimal value of 0.01 to list electrostatic variables.

4.8 Outliers

Initially, we prepared 43 complexes (the first 43 in Table 1), but only a dataset containing the first 39 provides good models. The remaining 4 complexes are *outliers*. In 3 of these neuraminidase binds to sialic acid, and in 1, it binds to axial-PANA.

```
% cp $COMBINE/outlier.dat . /* copy the Golpe dat file of the 4 outliers */
Utilities -> PLS-predictions -> Select a file: outlier.dat
```

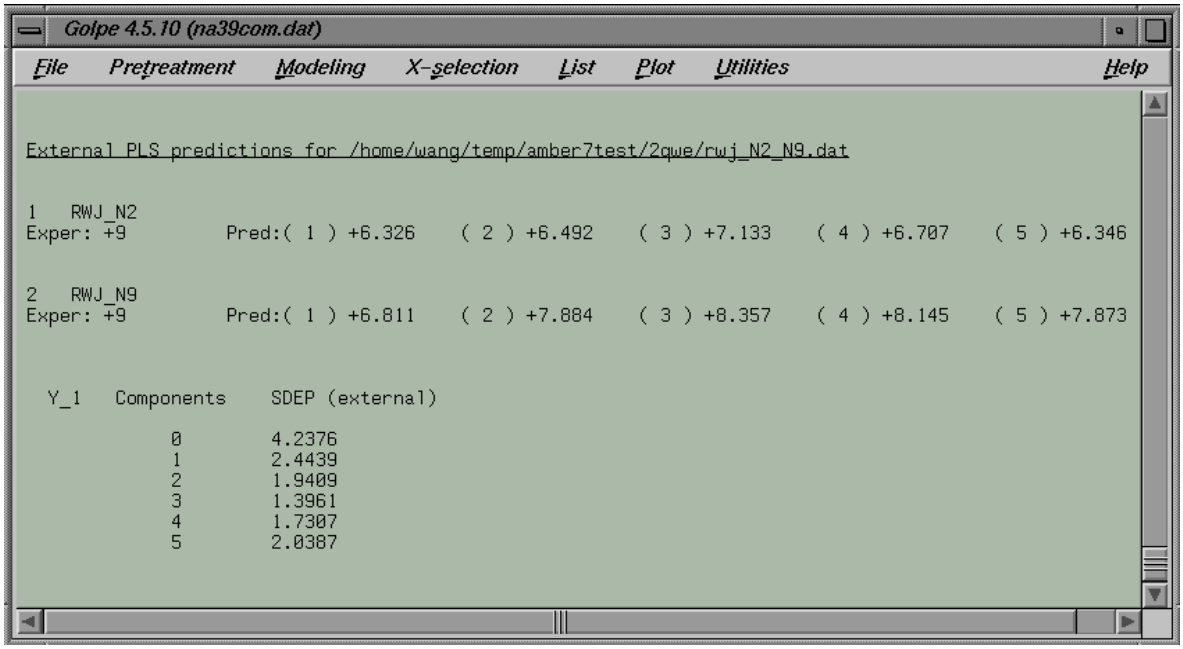


You can see that the activities of these 4 complexes are overpredicted by the models. At the optimal dimensionality of 3 latent variables, the prediction errors are 2.4 for aPANA:N2, 4.4 for sialic aid: N9, 2.0 for sialic aid: N9 mutant, and 4.2 for sialic aid:N2. The main reason is that the two compounds in these complexes undergo conformational change upon binding, from the low energy chair form to the high energy boat form. (see Scheme 1). Intramolecular energy changes were not accounted for in the COMBINE analysis.

4.9 External prediction for a new inhibitor bcx-1812(rwj-270201)

The new inhibitor has a five-member ring framework (see the last structure in Scheme 1) and is currently in clinical trials. The N2 and N9 complexes were constructed by using the docking program AUTODOCK.

```
% cp $COMBINE/rwj_N2_N9.dat . /* copy the Golpe file of the new inhibitor*/
Utilities -> PLS-predictions -> Select a file: rwj_N2_N9.dat
```



You can see that at the optimal dimensionality of 3 latent variables, the activity for the N9 subtype was predicted very well by the COMBINE model but a bit worse for the N2 subtype.