PRIME 2010 Summer

Kai En Tang, Osaka University Japan

APPLICATION OF WATERMARKING TECHNIQUES TO BIOMEDICAL RESEARCH DATA

PRIME Program

- Pacific RIM Undergraduate Experiences
- Allows undergraduates from UCSD to experience both culture and research in a foreign hosting country.

PRIME

PACIFIC RIM UNDERGRADUATE EXPERIENCES

Purpose of the Project

This project focuses on the development of watermarking algorithms to secure virtual screening research data from illegal distribution.

Our goals is to develop from perl scripting and apply a watermarking technique that satisfies robustness, imperceptibility, and security to protect biological research data.

Virtual Screening/Docking

Virtual screening is a computational method to discover new chemicals that bind with high affinity to a specific protein target by screening millions of chemicals from a database. (Very useful in drug discovery)

Comparing energy score docking (mol2 file) and AMBER score docking (pdb file*) the affinity/specificity (how well it binds) of the chemical to the target is determined.

Chimera is a visualization software that reads these files and outputs interactive 3D images.

*Protein data base file

Chimera Background

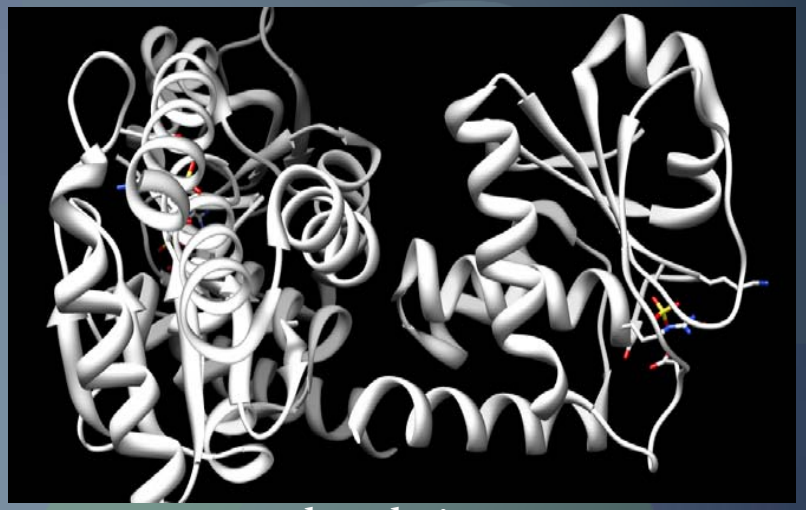
Chimera created by UCSF, is a multiple application program for 3-D interactive visualization and analysis of molecular structures.

It reads pdb or mol2 files to generate high-quality images and animations.

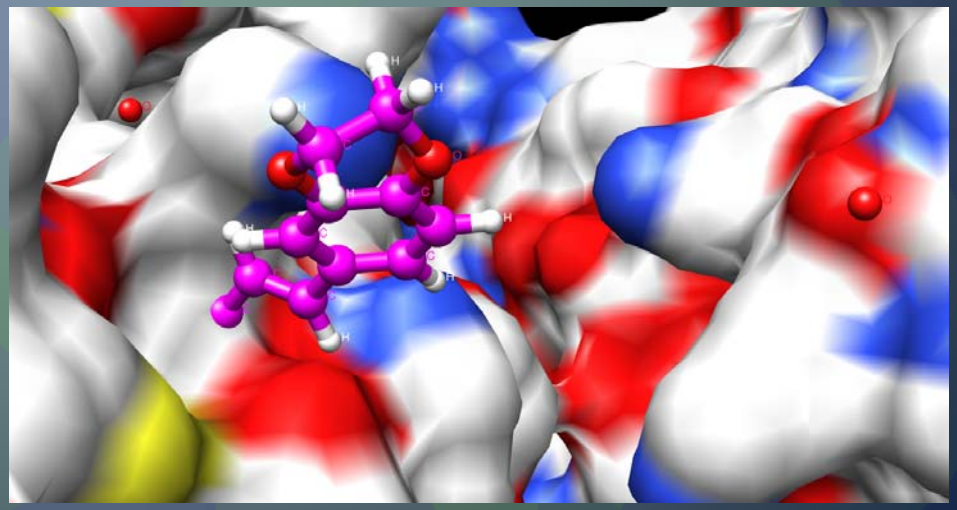
Available functions such as:

- density maps
- supramolecular assemblies
- sequence alignments
- docking results, trajectories
- conformational ensembles

It is used to confirm whether docking results are false positives or not.



molecule images



What is a Mol2 File?

Tripos Mol2 file (.mol2) is an ASCII file which contains all the information needed to reconstruct a *SYBYL molecule (contains information about its ligands and molecule).

Mol2 files are written in a ***free format*** to avoid the restrictions created by fixed format files.

Program Chimera can read mol2 file through ViewDock (Lau, et al., Bioinformatics, 2010).

*computational chemistry and molecular modeling

@<TRIPOS>ATOM

Data record associated with atoms in the molecule; SYBYL (a specification for describing the structure of chemical molecules using short ASCII strings).

Format:

atom_id atom_name x y z atom_type [subst_id [subst_name [charge [status_bit]]]]

```
@<TRIPOS>ATOM
1       C1      1.207   2.091   0.000   C.ar    1       BENZENE0.000
2       C2      2.414   1.394   0.000   C.ar    1       BENZENE0.000
3       C3      2.414   0.000   0.000   C.ar    1       BENZENE0.000
4       C4      1.207   -0.697  0.000   C.ar    1       BENZENE0.000
5       C5      0.000   0.000   0.000   C.ar    1       BENZENE0.000
6       C6      0.000   1.394   0.000   C.ar    1       BENZENE0.000
7       H1      1.207   3.175   0.000   H       1       BENZENE0.000
8       H2      3.353   1.936   0.000   H       1       BENZENE0.000
9       H3      3.353   -0.542  0.000   H       1       BENZENE0.000
10      H4      1.207   -1.781  0.000   H       1       BENZENE0.000
11      H5      -0.939  -0.542  0.000   H       1       BENZENE0.000
12      H6      -0.939  1.936   0.000   H       1       BENZENE0.000
```

SYBYL Atom Types:

C.3 = sp3 carbon

Cl = chlorine

N.p13 = trigonal planar nitrogen

LP = lone pair

- White space: inserting white spaces or tabs between within the text file.

1 C1 17.1604 43.3438 8.6969 C.3 1 <0> −0.1513 Added white space
2 C2 17.5991 43.2946 6.2377 C.3 1 <0> −0.1662 Added tab
3 C3 18.9924 44.7383 7.7265 C.3 1 <0> −0.1586 Original

- Adding additional trailing numbers after 4th decimal place.

6 O1 17.44015123 40.99701231 5.7424 O.2 1 <0> −0.4322
7 C6 15.40784123 41.90371236 6.5150 C.2 1 <0> −0.2455

- Line swap: by swapping every two lines encrypt 1 bit of information.

4 C4 17.5991 43.2946 6.2377 C.3 1 <0> −0.1662
3 C3 18.9924 44.7383 7.7265 C.3 1 <0> −0.1586
} Encrypt 1 bit

However, above methods are easily attacked by simply resaving as a new file in Chimera.

Possible Watermarking Approaches (cont.)

- Because these possible approaches are easily attacked, another approach that is resistant to reset after reading by Chimera was needed.
- Although not ideal, altering the x y z coordinates is the most effective way of protecting the data.
- This alteration of x y z coordinates will be based on “Correlation Watermarking", which is a robust embedding scheme for these data files.

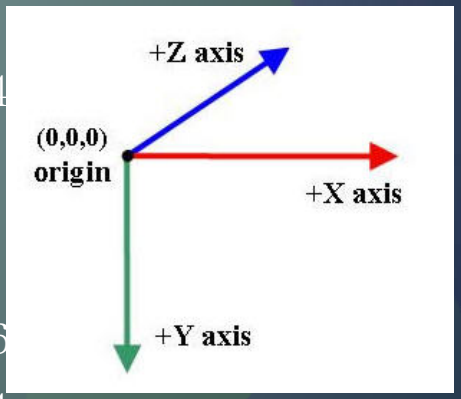
Altering mol2 File

The method to alter all the xyz data slightly in the mol2 file is almost invisible because it changes only slightly entries in the @TRIPOS<ATOM> section of mol2 files.

8 C7 15.0494 43.0395 7.4436 C.2 1 <0> 0.3936
9 O2 15.1228 44.2233 7.1925 O.2 1 <0> −0.4044

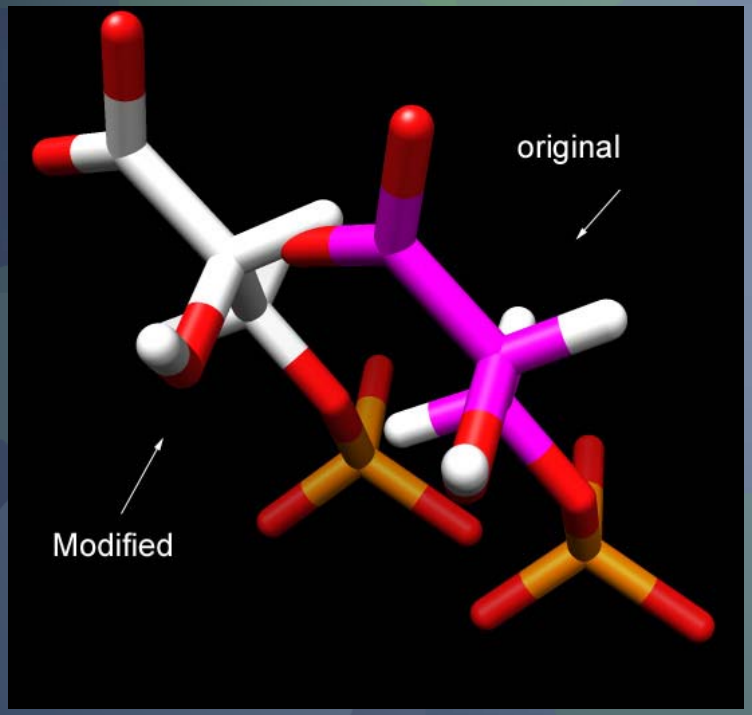
Add + 0.0001 to coordinates x and y

8 C7 15.0495 43.0396 7.4436 C.2 1 <0> 0.3936
9 O2 15.1229 44.2234 7.1925 O.2 1 <0> −0.4044

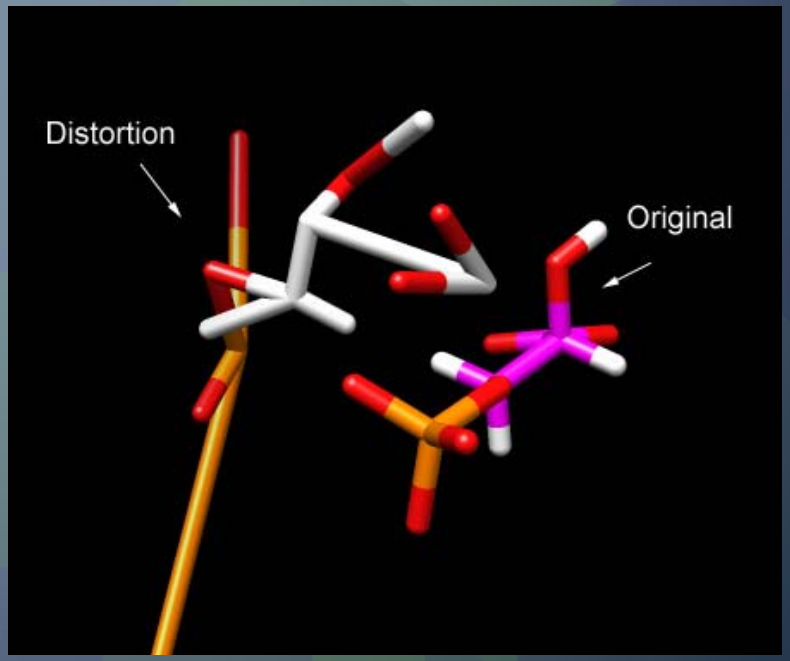


Possible modifications

- Slight alteration versus distortion.
- Distortion becomes very obvious.



Maintain the same molecular shape

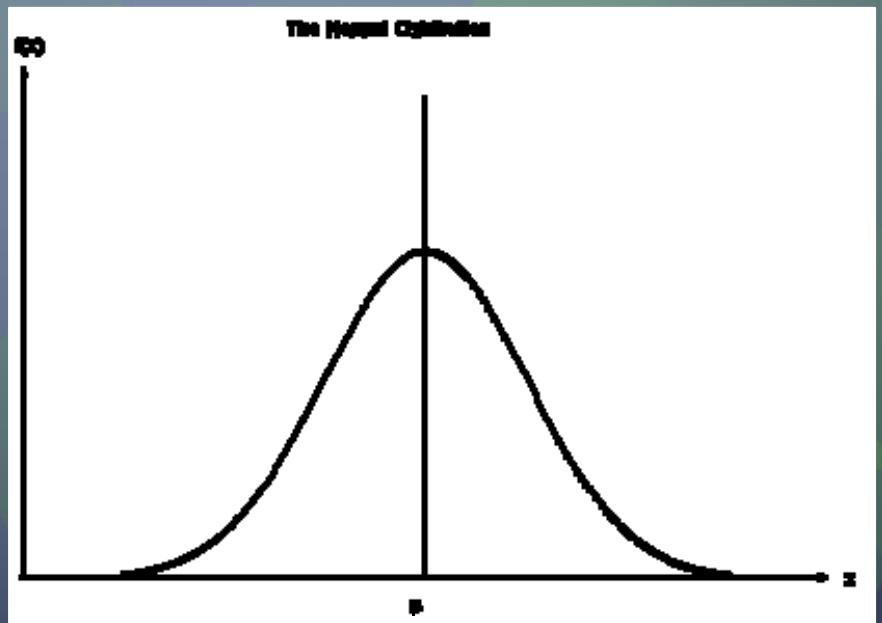


Molecular shape lost

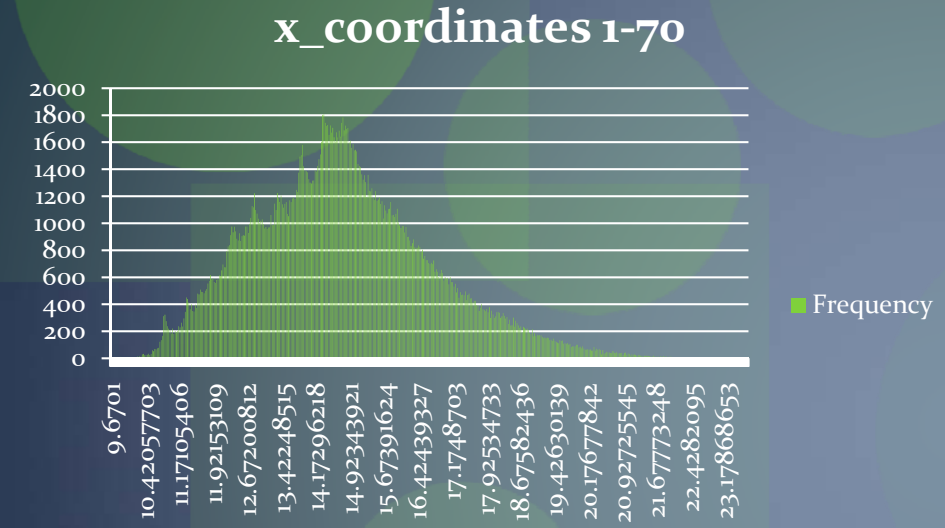
Evaluating xyz Coordinates

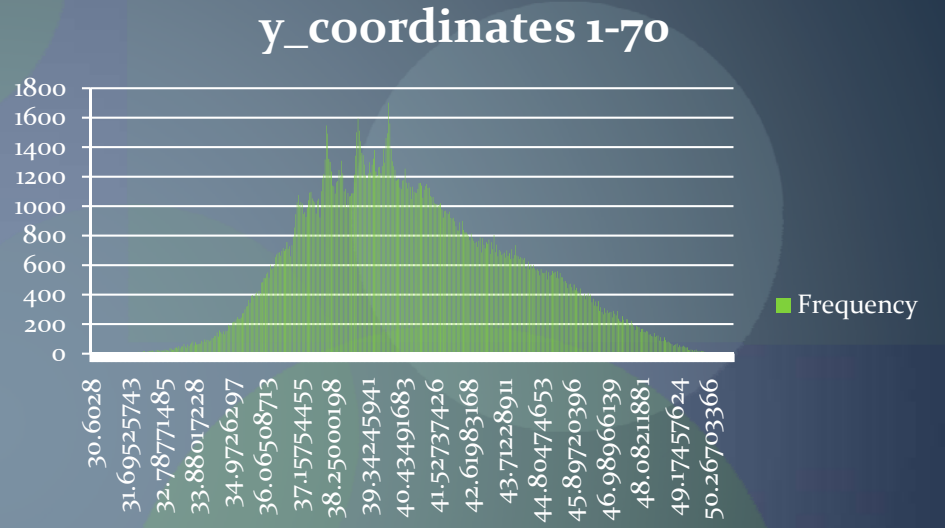
To detect the embedded watermark correctly, the best location to embed watermark data must be determined.

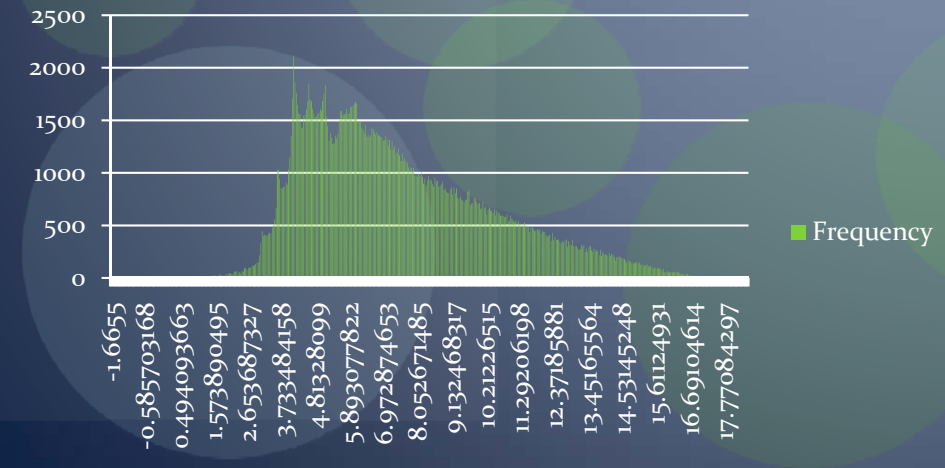
The correlation watermarking method that we have chosen require data points to have a normal distribution.



Analysis of the Distribution of xyz Coordinate Data







Base on analyzing 70 mol2 files we have eliminated z coordinates as an embedding location.

Watermark Mechanism

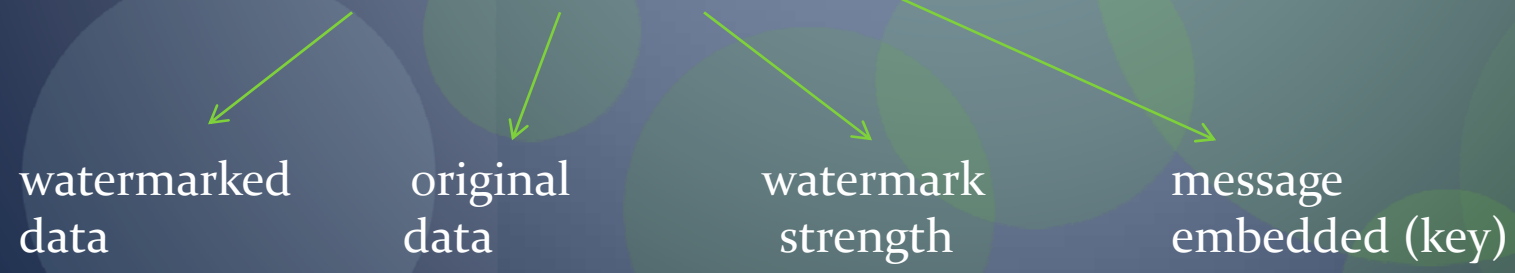
- Embedding watermark; by using appropriate watermark strength chosen, and chosen message to create the watermark.
- Detecting watermark and analyze its correlation; detecting watermark by analyzing correlation between original file and watermarked file.
- Decoding watermark; retrieve original data points by subtracting watermark from watermarked data.

Basic Algorithm of Embedding Watermark

- Transform message to random bits string. s[i]
- Then multiple by the watermark strength labeled as a[i]. Exhibits the watermark s[i].
- Add watermark to original data x[i]-> output watermarked data x'[i]

Emb (x_coordinates) x[i]

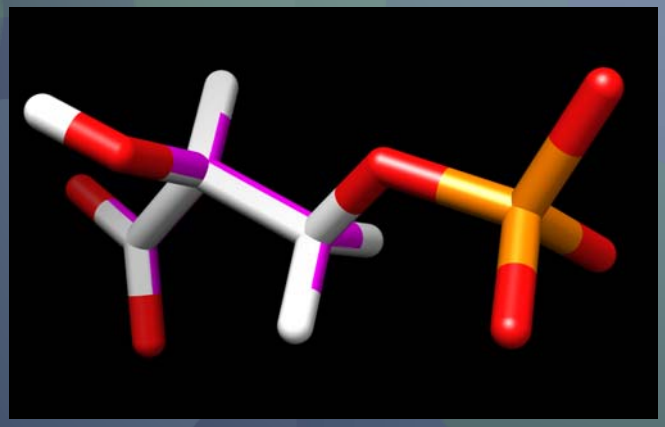
$$x'[I]=x[i]+a[i]*s[i]$$



Watermark Strength (a[i])

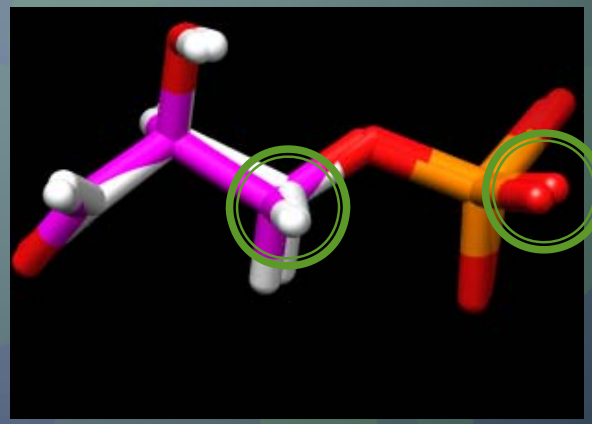
- A scalar value that must be large enough for embedding a watermark, but small enough to satisfy imperceptibility (invisible to human eye).
- If the xyz coordinate values are altered by more than ±0.001 the watermarking becomes visible when compared with the original orientation.

a[i]=0.001



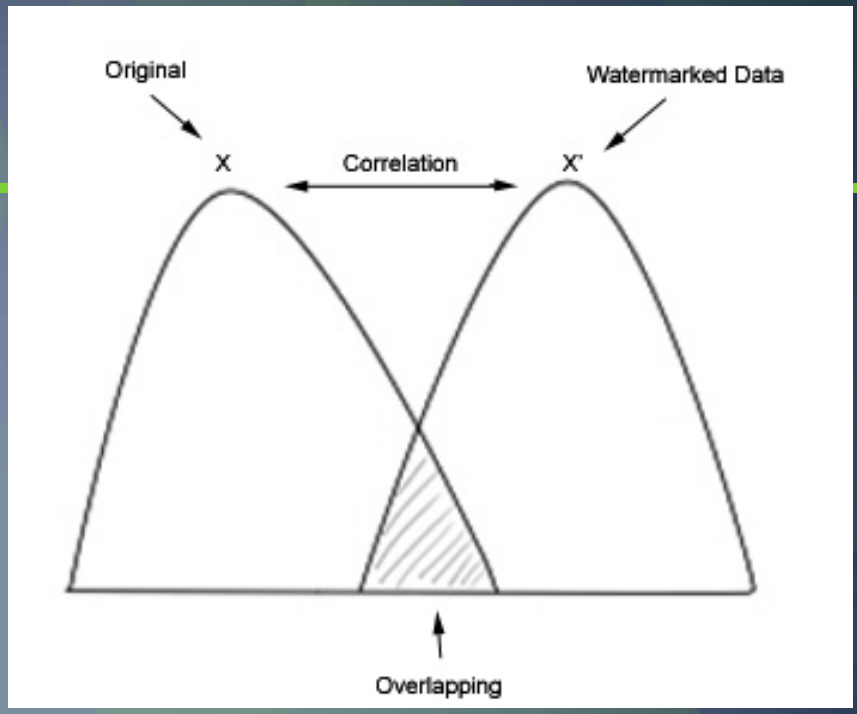
Not visible

a[i]=0.07



Visible

The correlation between the original data and the watermarked data is used to determine whether the data file is embedded with a watermark.



By equation:

$$\ell = \sum(x'[i] - \mu[i]) a[i] * s[i]$$

- correlation
- watermarked data
- mean of watermarked data
- watermarking strength
- watermarking message

All data must be normally distributed.

Detection by calculating correlation (x and y coordinates)

Larger ℓ(correlation) is the lower of error probability, proportional to the difference between x[i] ≠x'[i].

Watermarked data ≠ Original Data

(a_i=0.01)	45.mol2	53.mol2	0.mol2
X correlation	-5.09343	-0.09666	-4.57364
Y correlation	13.0723	23.1487	-8.7671

Results of watermarking testing

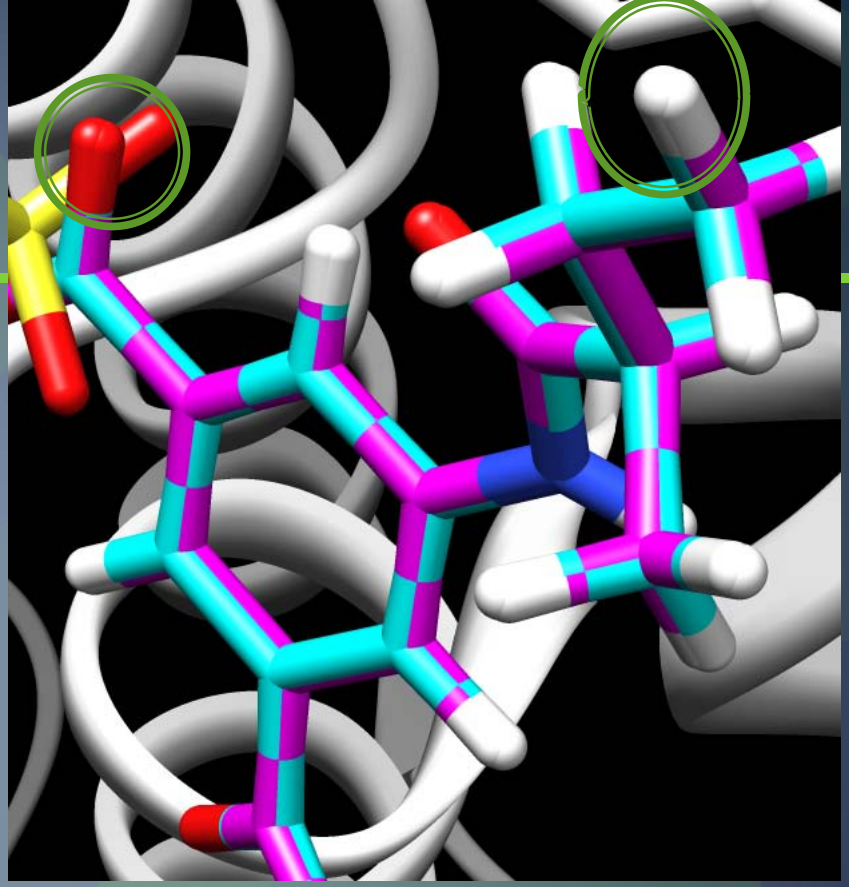
Then attempt to apply watermark only to x coordinates and evaluate its correlation. Obtain the following:

Also note: more data points results in a higher correlation.

ai=0.01	Standard deviation	Mean	Variance	correlation	Number of Data Points
45.mol2(not detected)	1.9925	14.8289	3.9696	0.4072	4072
53.mol2(not detected)	1.9152	14.6617	3.6654	0.5331	5331
0.mol2 (not detected)	2.05194	14.9496	4.2089	0.6997	6997

Results of watermarking testing (cont.)

The a[i] chosen should be small but large enough for detection (watermarking strength). After testing a[i] is best at 0.06.



ai=0.06	Standard deviation	Mean	Variance	correlation	Number of Data Points
45.mol2	1.9925	14.8285	3.9702	7.8496	4072
53.mol2	1.9152	14.6617	3.6679	15.910	5331
0.mol2	2.05194	14.9492	4.2105	17.0693	6997

Probability of Error

- A certain probability of error is involved when information is extracted from watermarked data.
- This measurement can be used for determining the watermarking performance.
- Pfp = false positive error probability
  - Probability of yielding a positive result in watermark detection test when xyz coordinates do not contain a watermark generated by message s[i].
- Pfn = false negative error probability
  - Probability of failing to detect a watermark when there actually is one embedded within the data file.

Relationship Between Pfp and Pfn

m = mean of correlation for watermarked data over possible key message embedded (s[i])

v^2 = variance of correlation for watermarked data over possible key message embedded (s[i]).

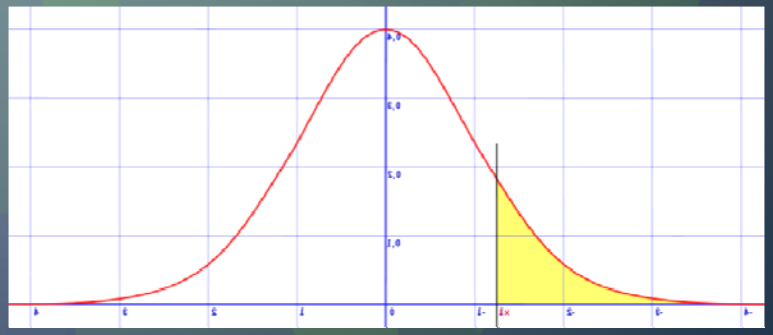
$$Pfn = 1 - Q\left(Q^{-1}(Pfp) - \left(m / \sqrt{v^2} \right) \right)$$

Pfn: Positive false negative probability

Pfp: Positive false positive probability

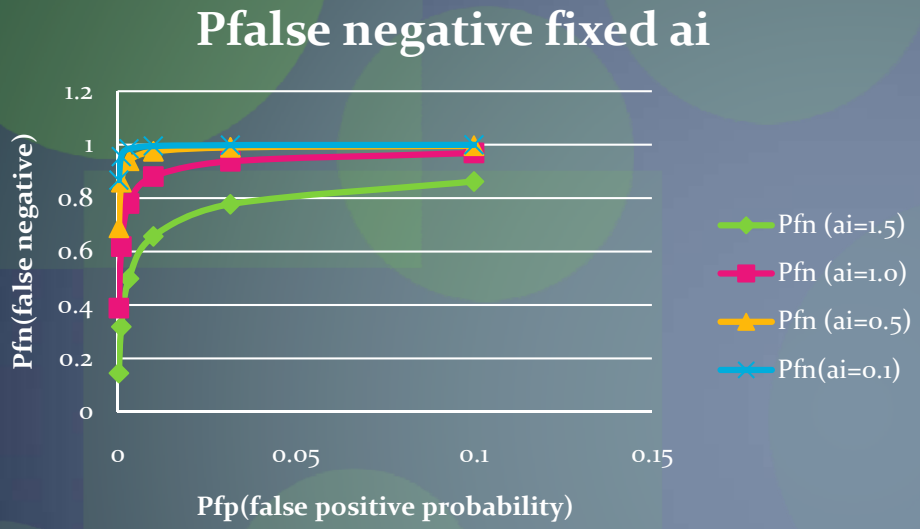
m: mean

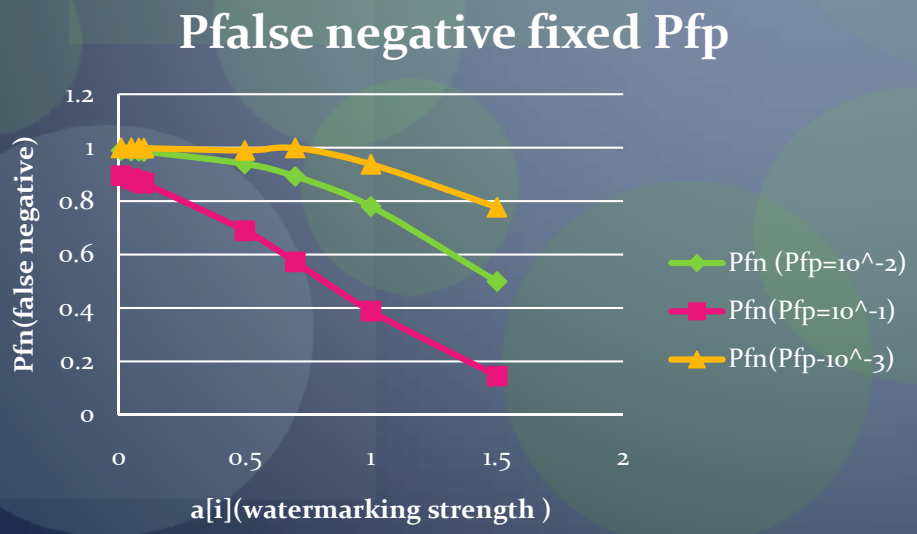
v^2: variance



*Q will be solved from using the function (error function)erf:
$Q(x)=1/2-1/2\,erf(x/\sqrt{2})$

Trade-offs Between Error Probability and Watermarking Strength





Concluding from this graph, a[i] is best around 1.5, but watermarking becomes extremely visible in Chimera. Therefore we need to develop a better algorithm.

Retrieving and Decoding Watermark

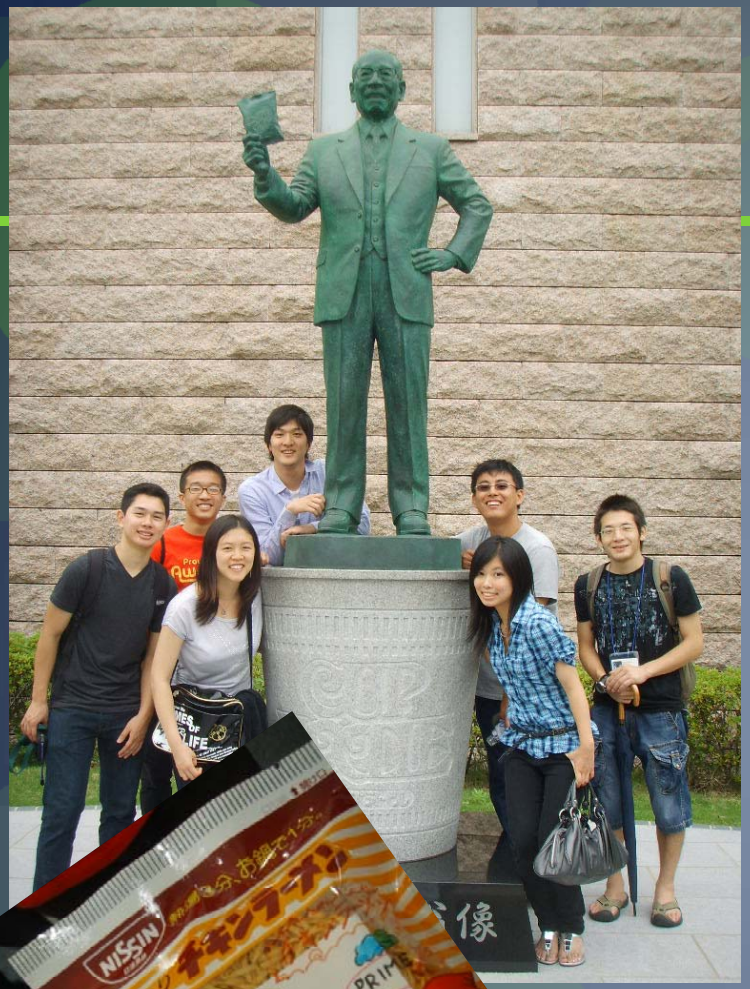
- In order to retrieve original data, we subtract watermark from watermarked data by using the message to generate random bits string. s[i].
- Multiple by a[i] to create the same watermark.

$$x[i]=x'[i]-a[i]*s[i]$$

- Only user with the message can decode the watermarked data and obtain the correct mol2 file.

Future Goals

- Further strengthen watermarking algorithm by developing algorithms that reduces the error probability.
- Improve embedding perl script by changing parameters separately for x and y coordinates (such as individually based on two different watermarking strength a[i])
- Apply watermarking algorithms to pdb files as well.

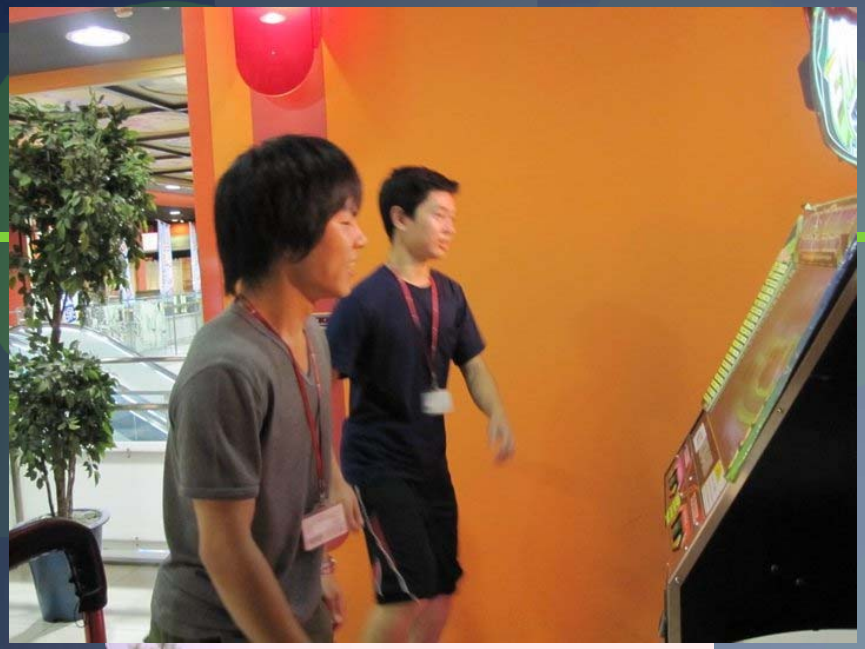



NISSIN
チキンラーメン


NISSIN
チキンラーメン


お社
焼かま


POLO JEANS



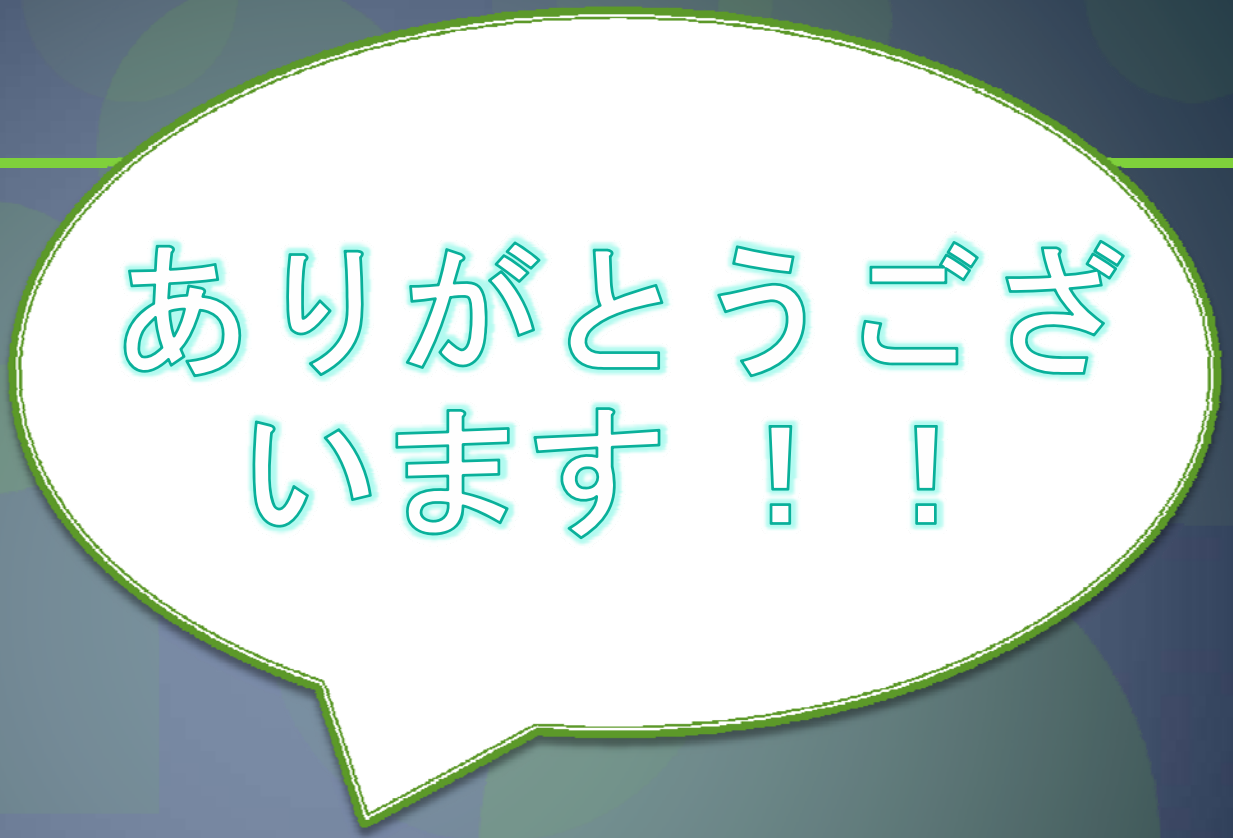
Acknowledgements

Programs

- UCSD PRIME, Chancellor's Research Scholarship
- HHMI, NSF, NIH

Laboratories and People

- Dr. Susumu Date, Dr. Toru Fujiwara and Dr. Maki Yoshida (Osaka University)
- Dr. Gabriele Wienhausen, Dr. Peter Arzberger, Teri Simas and Dr. Jason Haga, Christopher Lau (UCSD)



Thank you! Questions?