Guiding Protein Docking Simulations with Chemical Cross-link Data

XLdock = Cross-links + RosettaDock + Xwalk

Abdullah Kahraman, Malmstroem Lab, ETH Zurich

RosettaCON 2011, 6th August 2011

Introduction

- Protein-protein docking troublesome due to large conformational space and imperfect scoring functions
- experimental constraints can be key in producing close native models
  - NMR, FRET, ...

Introduction

- Protein-protein docking troublesome due to large conformational space and imperfect scoring functions
- experimental constraints can be key in producing close native models
  - NMR, FRET, …
- **Chemical cross-linking** coupled to **mass spectrometry** (XL-MS) is another means to obtain distance information.

Introduction

- Protein-protein docking troublesome due to large conformational space and imperfect scoring functions
- experimental constraints can be key in producing close native models
  - NMR, FRET, ...
- **Chemical cross-linking** coupled to **mass spectrometry** (XL-MS) is another means to obtain distance information.
- Often a flexible linear cross-linker molecule with reactive ester sites on both ends is used to cross-link lysine pairs

Introduction

- Protein-protein docking troublesome due to large conformational space and imperfect scoring functions
- experimental constraints can be key in producing close native models
  - NMR, FRET, ...
- **Chemical cross-linking** coupled to **mass spectrometry** (XL-MS) is another means to obtain distance information.

- Often a flexible linear cross-linker molecule with reactive ester sites on both ends is used to cross-link lysine pairs

Introduction

- Protein-protein docking troublesome due to large conformational space and imperfect scoring functions
- experimental constraints can be key in producing close native models
  - NMR, FRET, …
- **Chemical cross-linking** coupled to **mass spectrometry** (XL-MS) is another means to obtain distance information.

- Often a flexible linear cross-linker molecule with reactive ester sites on both ends is used to cross-link lysine pairs

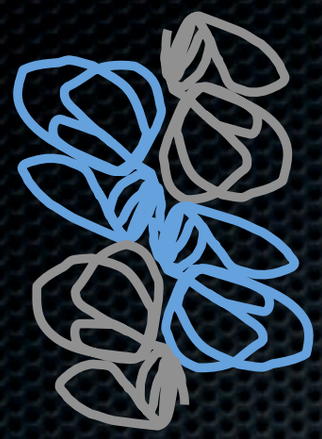
Introduction

- Protein-protein docking troublesome due to large conformational space and imperfect scoring functions
- experimental constraints can be key in producing close native models
  - NMR, FRET, ...
- **Chemical cross-linking** coupled to **mass spectrometry** (XL-MS) is another means to obtain distance information.

- Often a flexible linear cross-linker molecule with reactive ester sites on both ends is used to cross-link lysine pairs

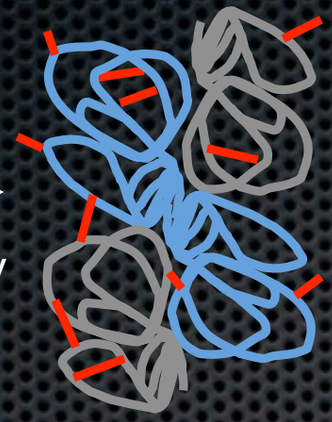
- As the cross-linker has a certain length, finding two lysine residues to be cross-linked yields an **upper bound** on their distance in Cartesian space.

General MS based cross-linking workflow

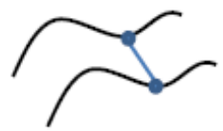


Cross-linking

with isotopically labeled cross-linker



Tryptic digest



cross-link

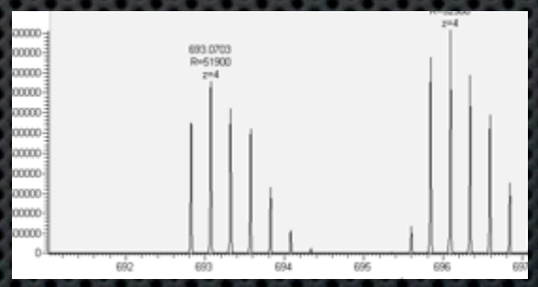
loop-link

mono-link

+ non modified peptides

LC-MS/MS analysis

xQuest



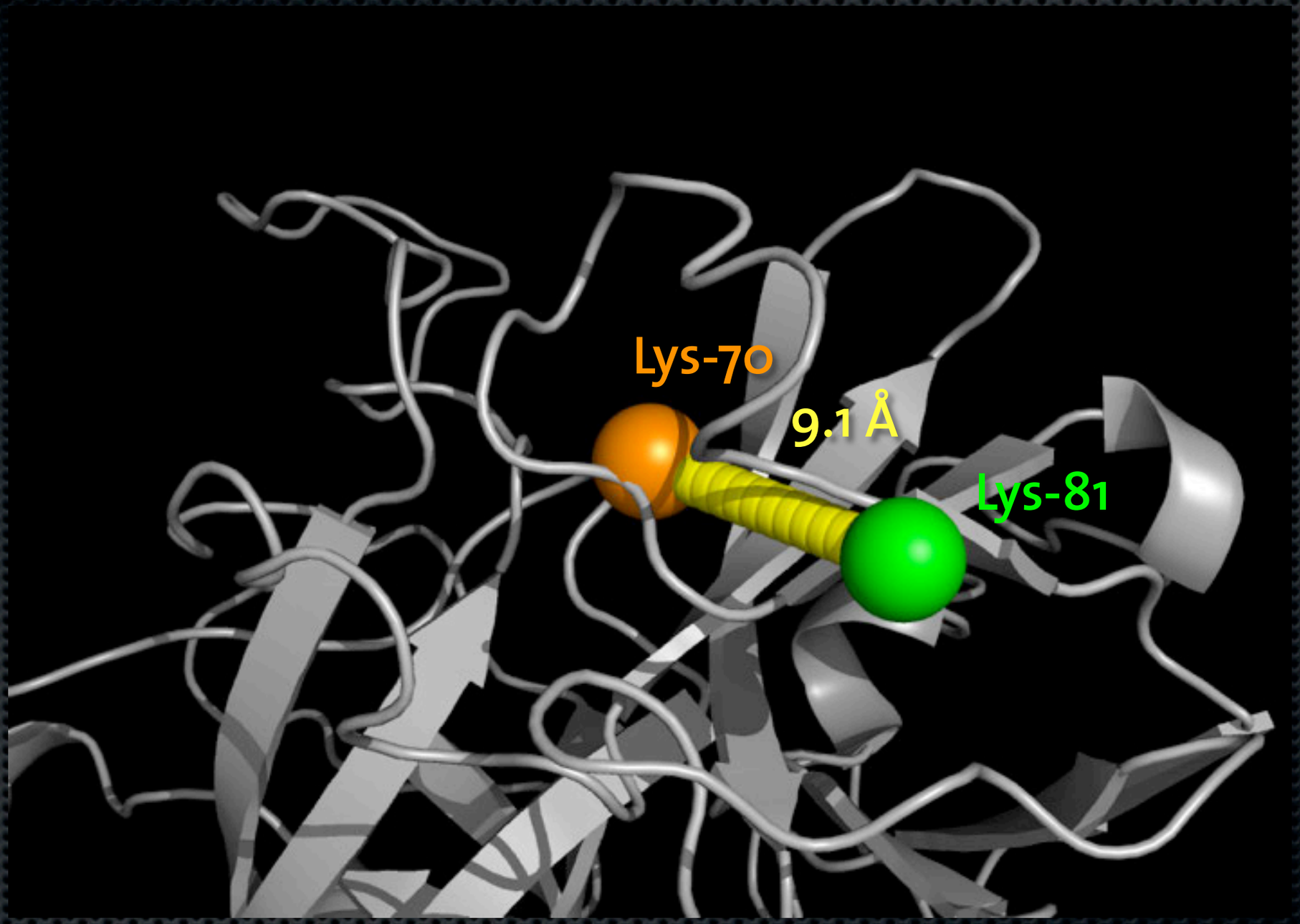
isotopic scan pairs

Data processing
Generation Error Model

Using the constraints:
Interaction, Modelling

Rinner, O. et al. Identification of cross-linked peptides from large sequence databases. Nat Methods 5, 315–318 (2008).

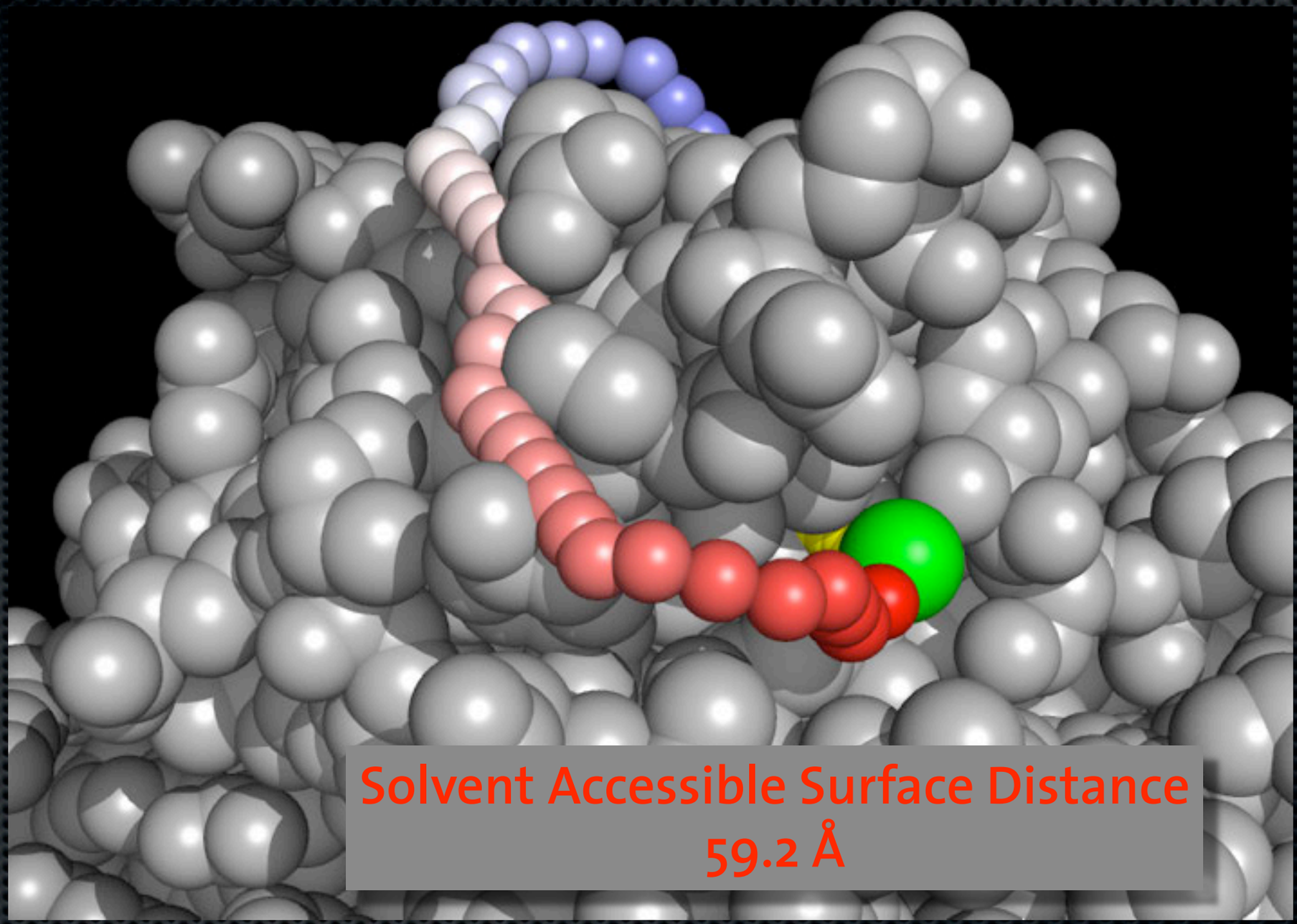
Introduction - Euclidean Measure



Human prothrombin (1dx5-E)

1. Potluri, S. et al. Geometric analysis of cross-linkability for protein fold discrimination. *Pac Symp Biocomput* **9**, 447–458 (2004).
2. Kahraman, A., Malmström, L. & Aebersold, R. Xwalk: Computing and Visualizing Distances in Cross-linking Experiments. *Bioinformatics* (2011).

Introduction - Euclidean Measure

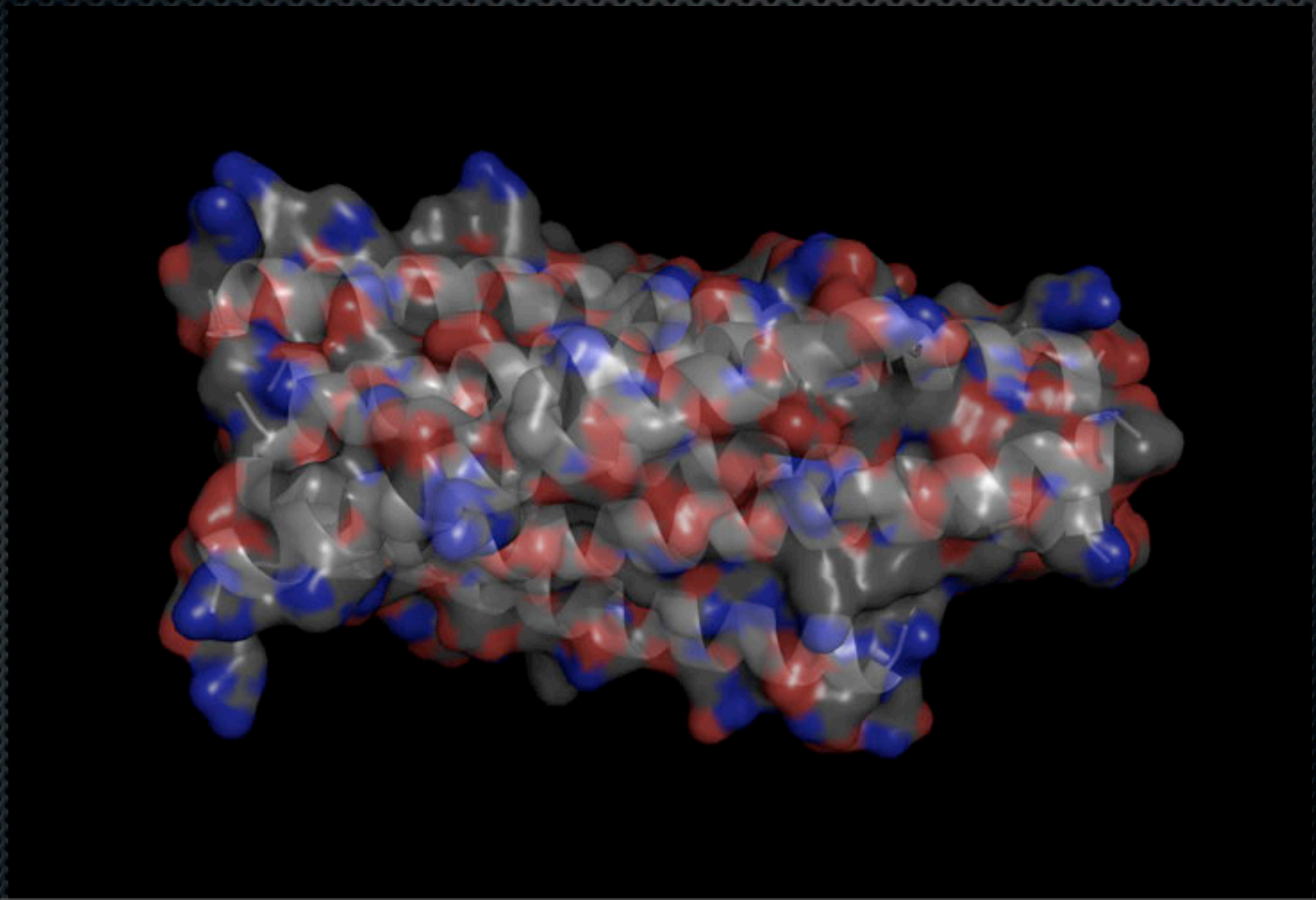


Human prothrombin (1dx5-E)

1. Potluri, S. et al. Geometric analysis of cross-linkability for protein fold discrimination. *Pac Symp Biocomput* **9**, 447–458 (2004).
2. Kahraman, A., Malmström, L. & Aebersold, R. Xwalk: Computing and Visualizing Distances in Cross-linking Experiments. *Bioinformatics* (2011).

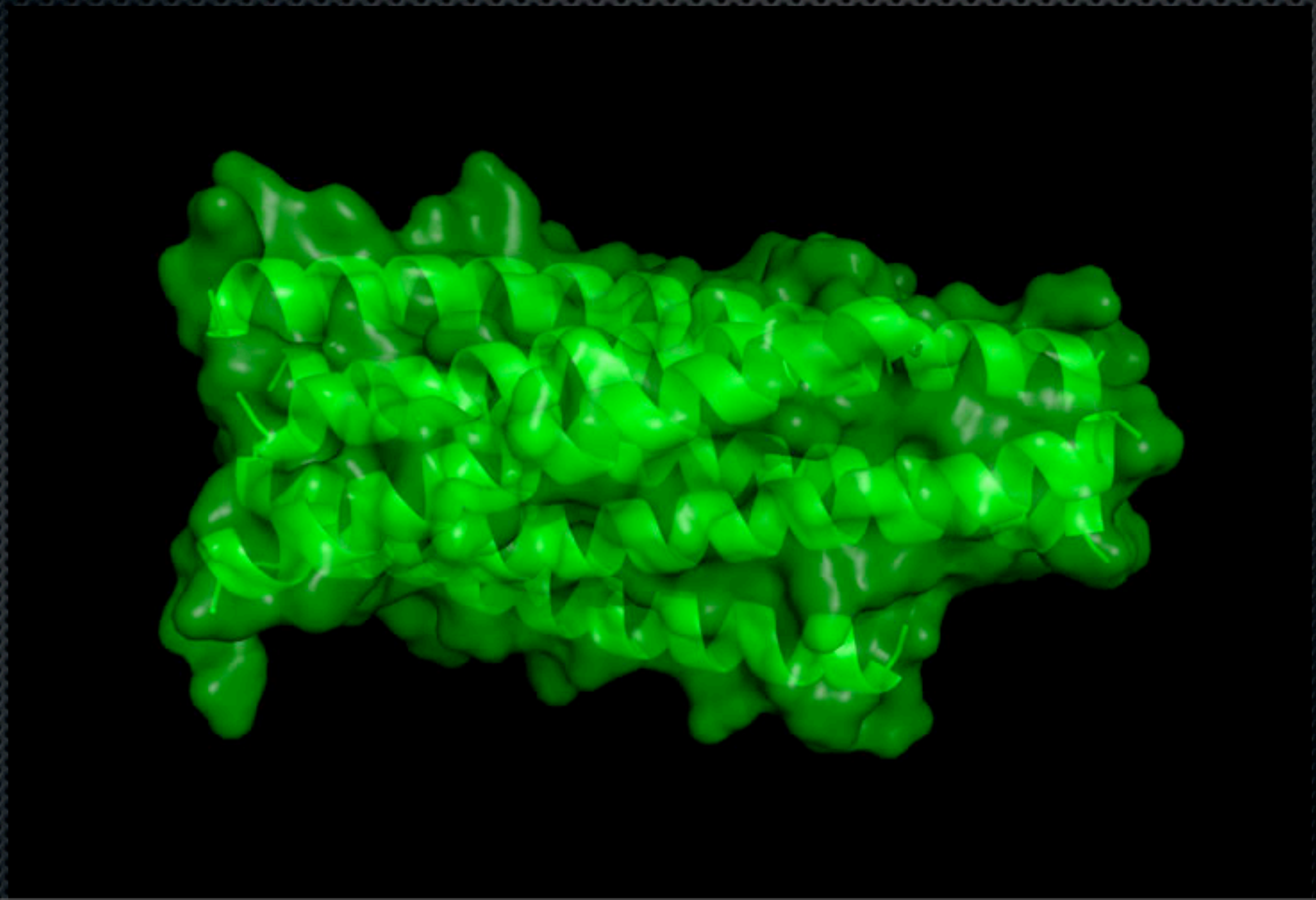
Xwalk - Algorithm I

- PDB Id: 1jek, triple-hairpin motif of Visna virus fusion protein



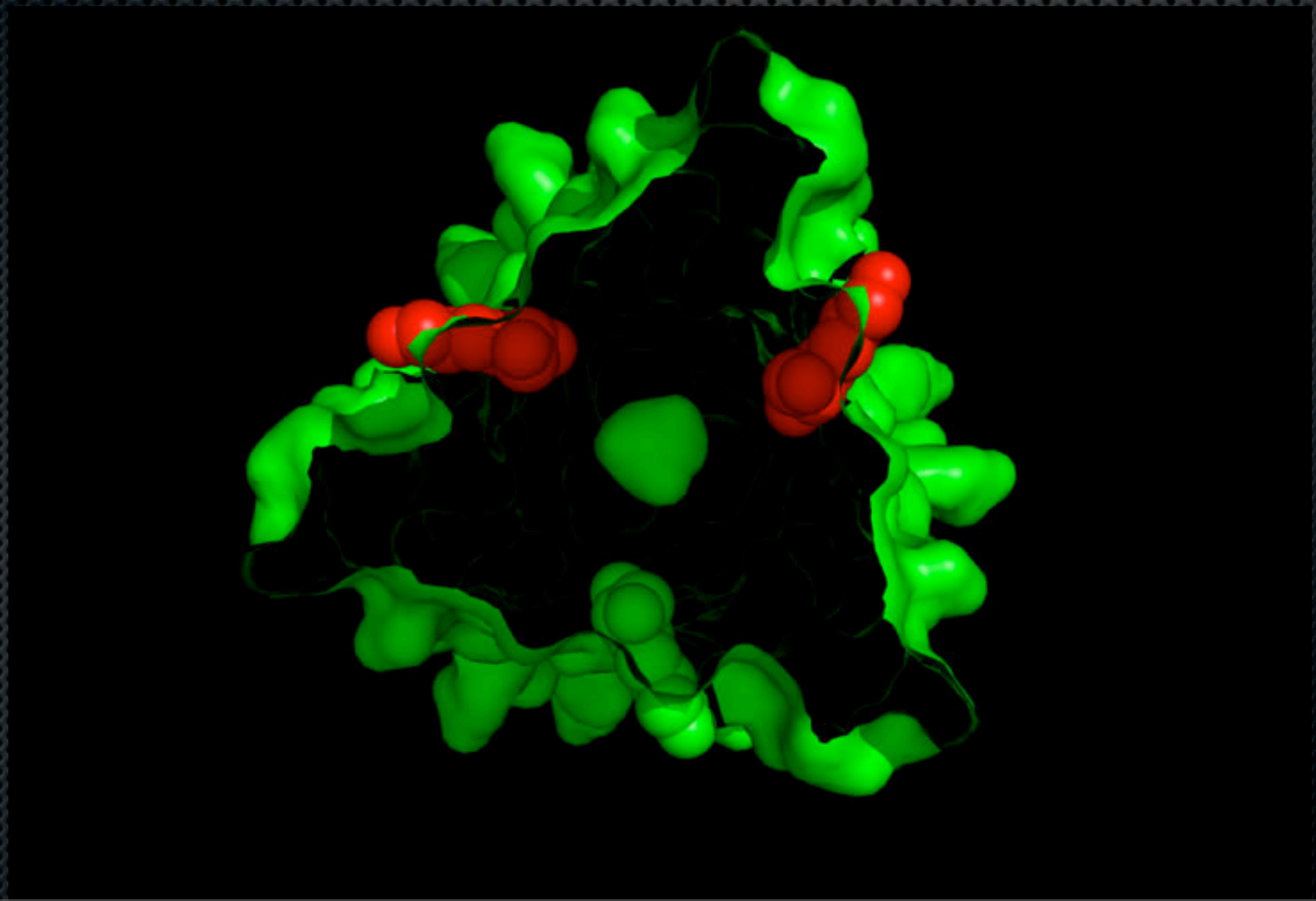
Xwalk - Algorithm II

- Find all **lysine** residues in structure



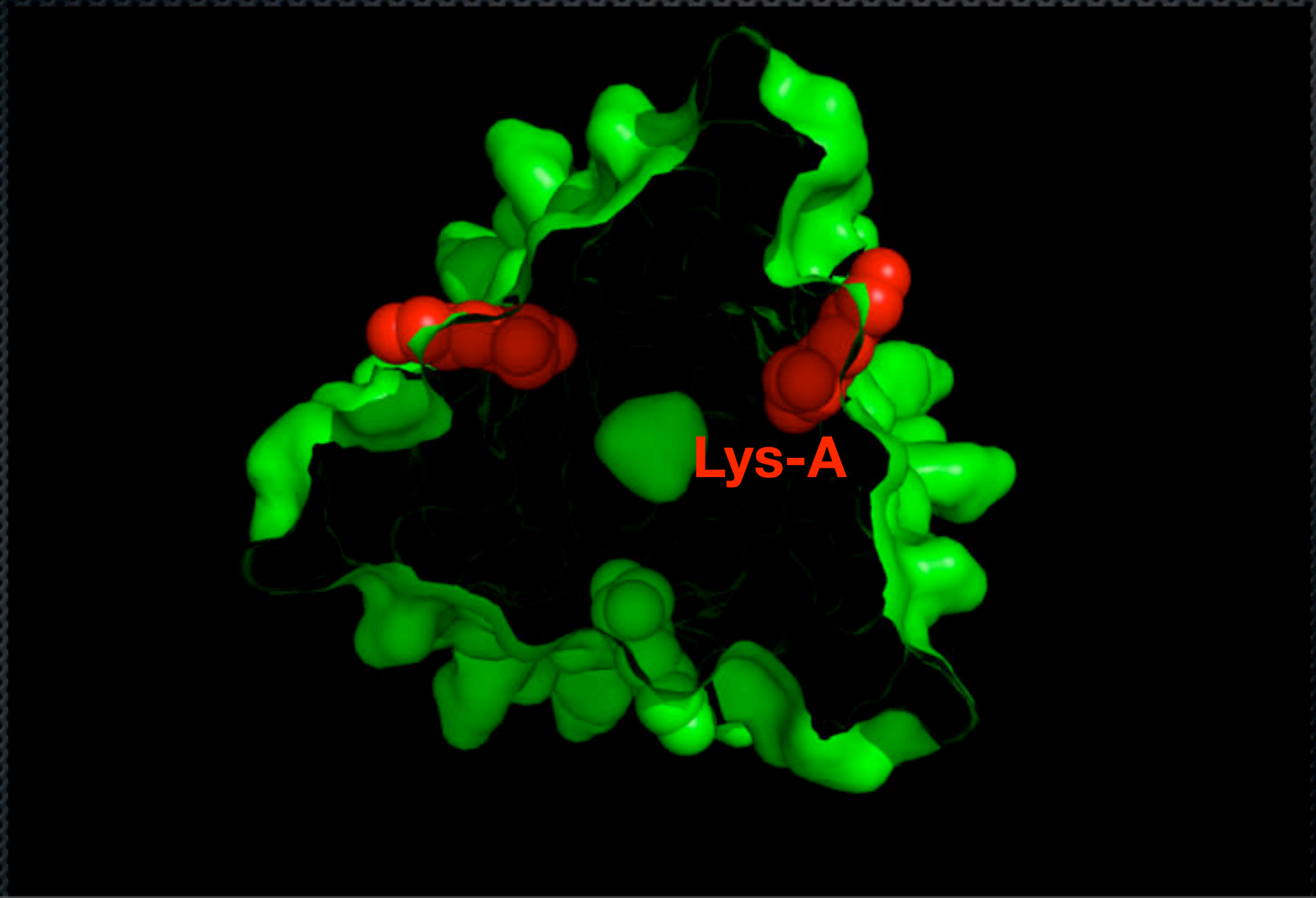
Xwalk - Algorithm II

- Find all **lysine** residues in structure



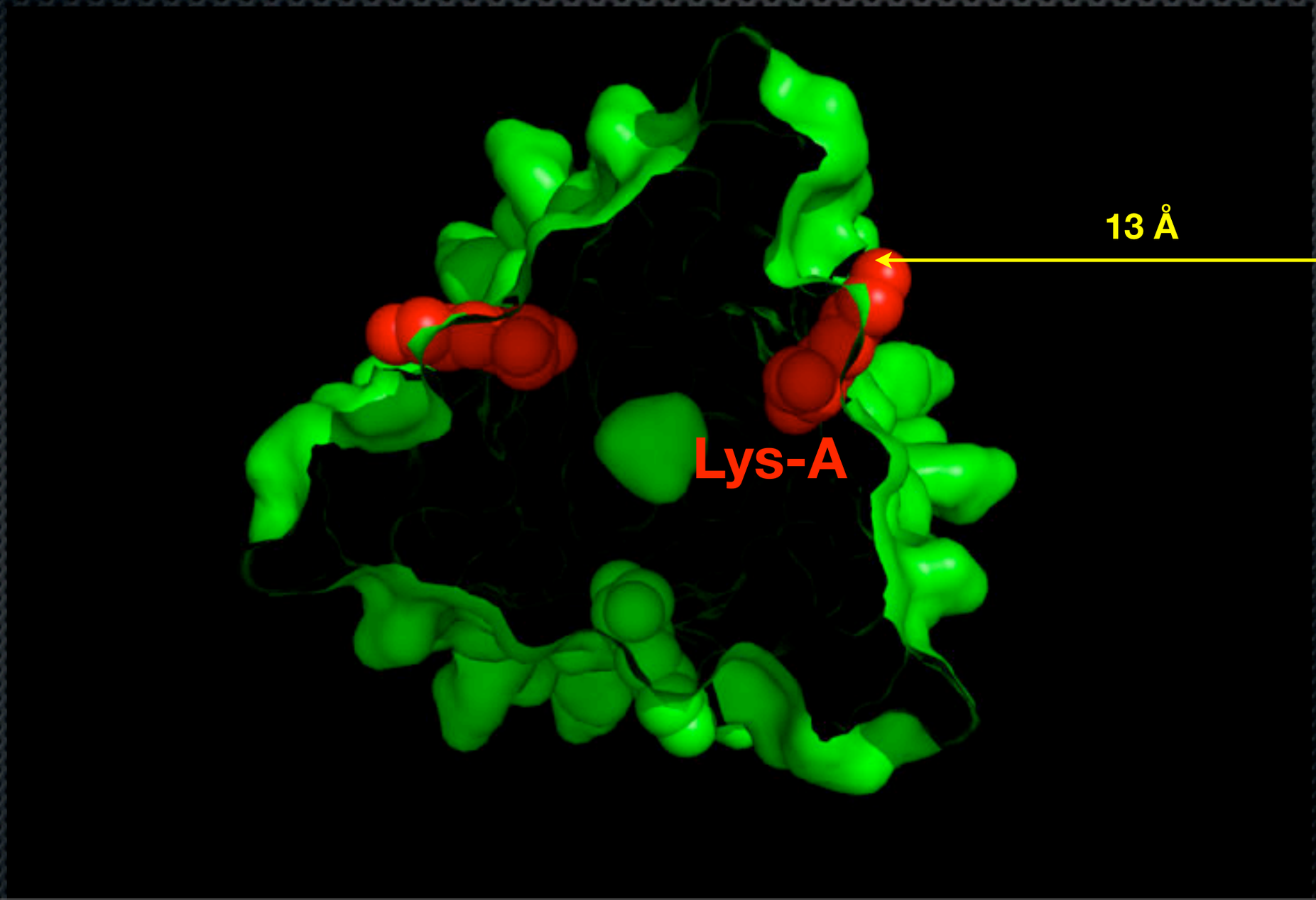
Xwalk - Algorithm III

- Place a grid on **Lys-A**
  - size of the grid corre**sponds to the maximum length of cross-linke**r



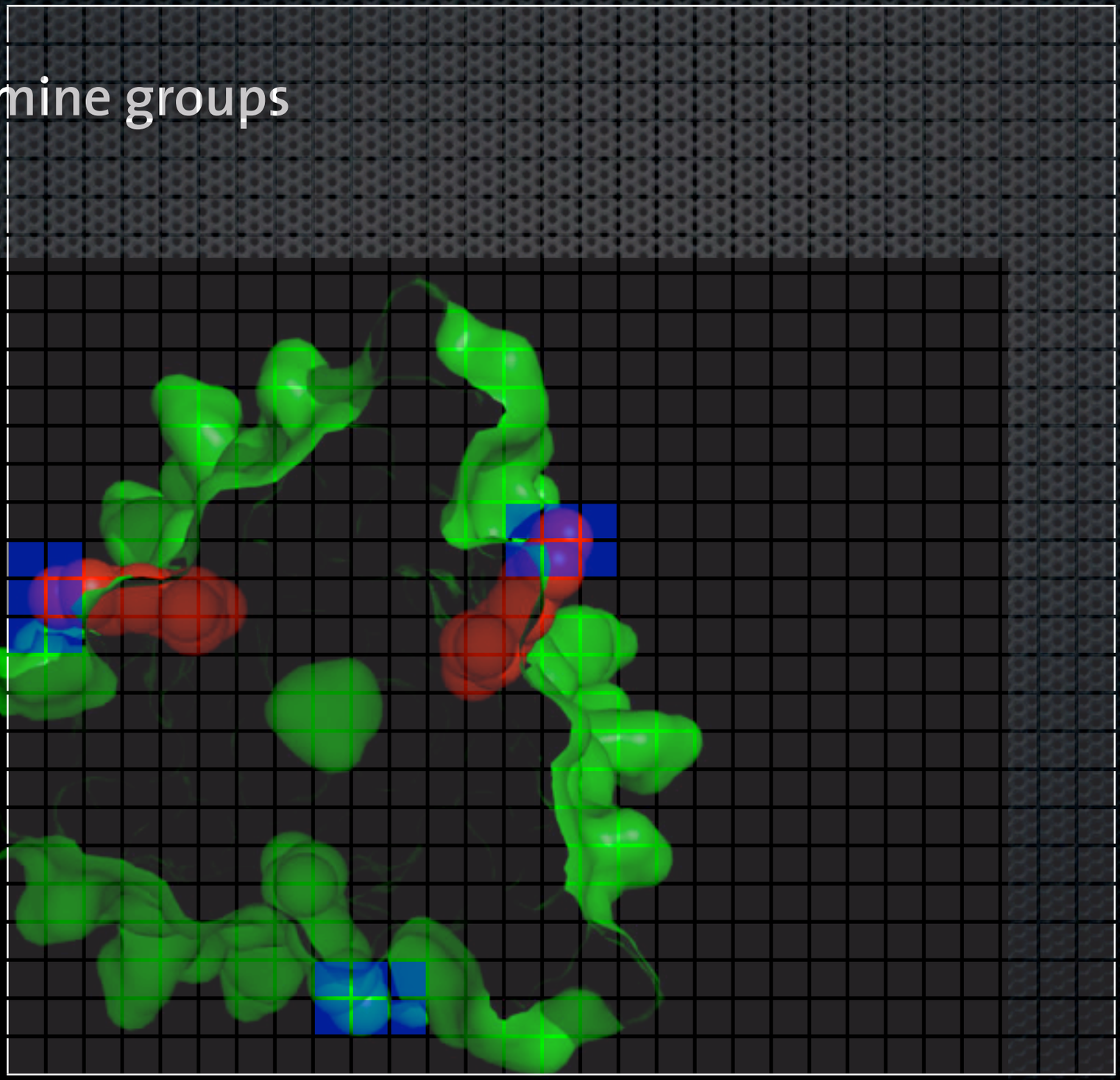
Xwalk - Algorithm III

- Place a grid on **Lys-A**
  - size of the grid corre**sponds to the maximum length of cross-linke**r



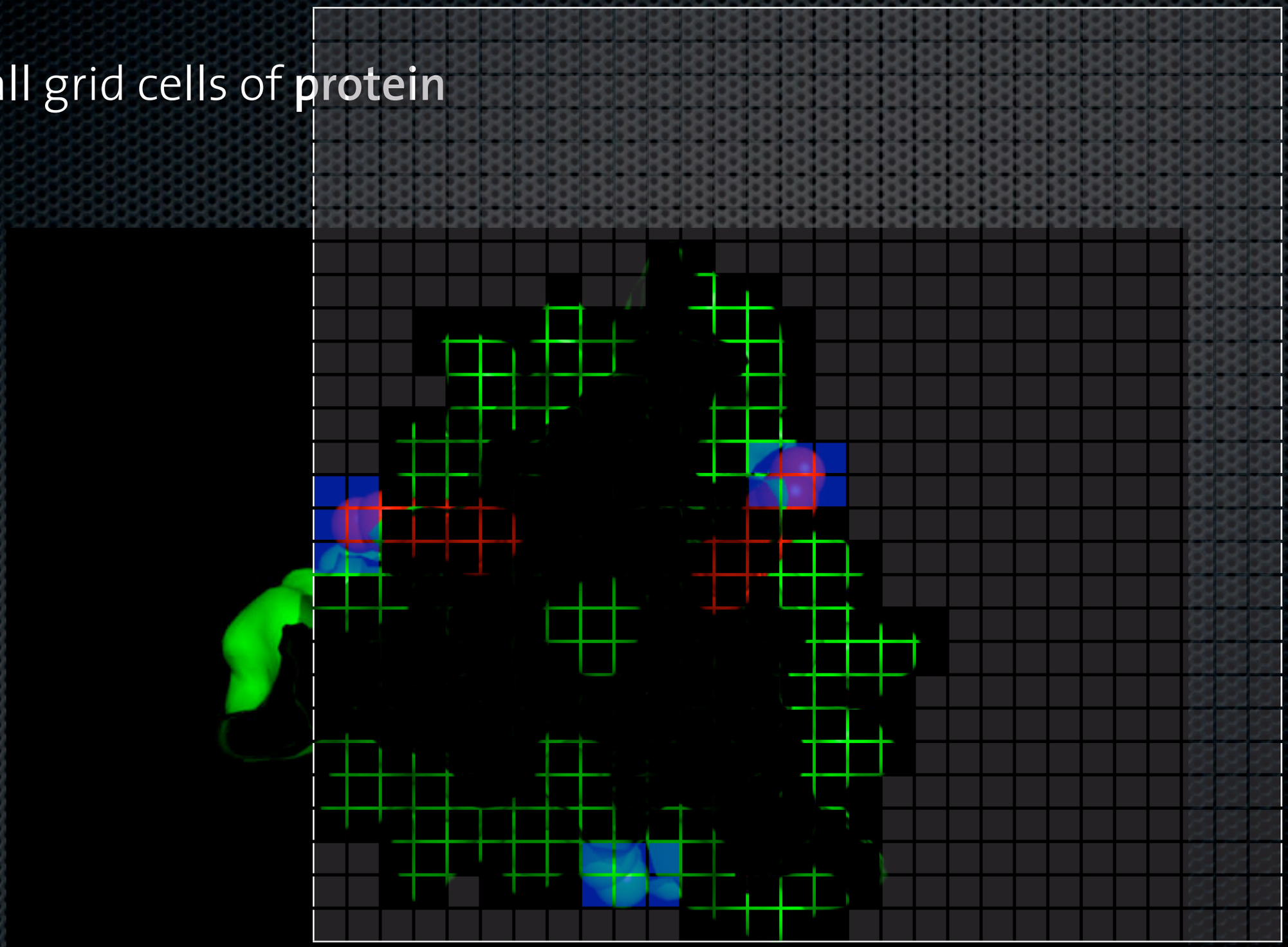
Xwalk - Algorithm IV

- **Label** grid cells of ε-amine groups



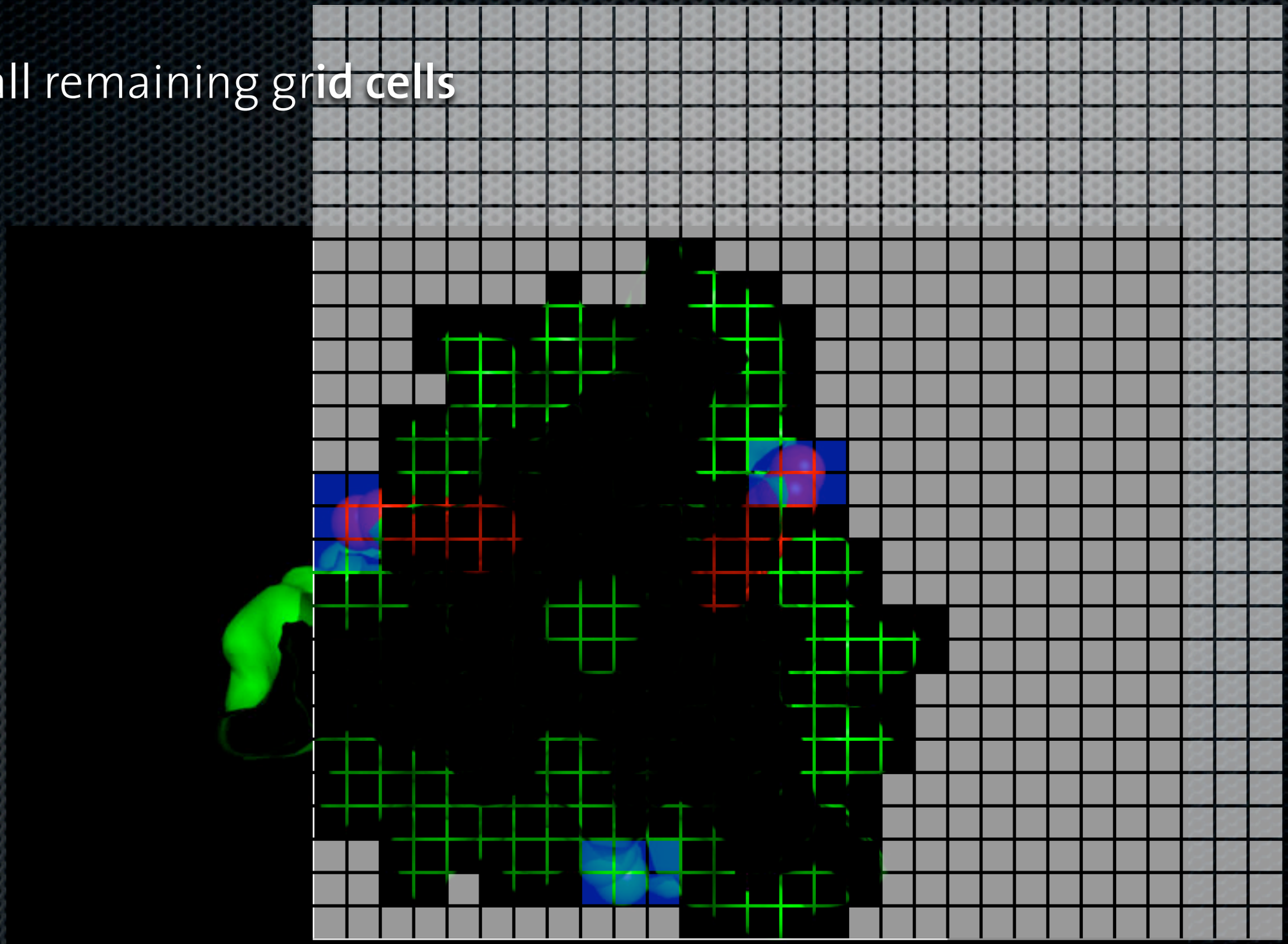
Xwalk - Algorithm V

- Label all grid cells of protein



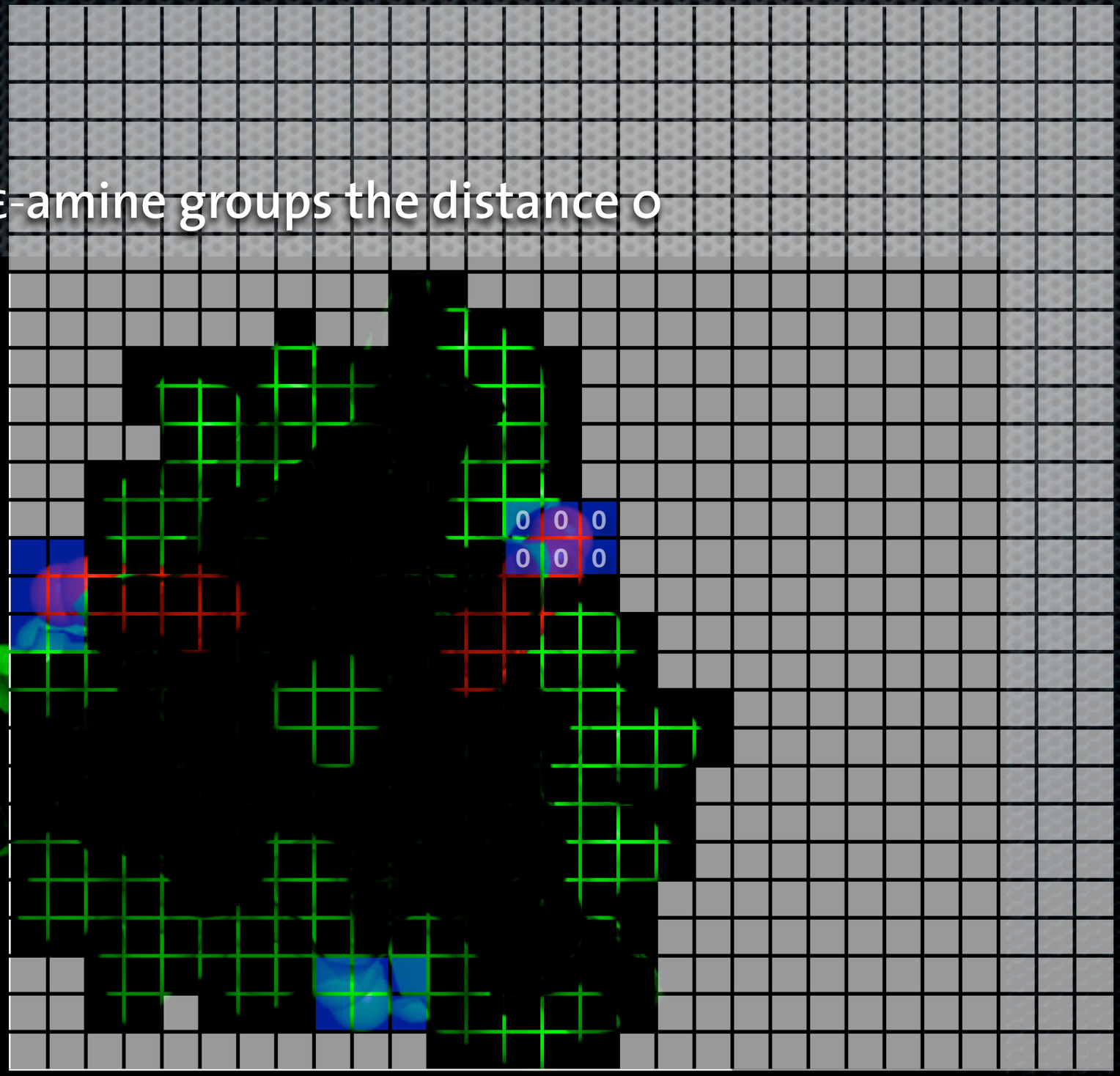
Xwalk - Algorithm VI

- Label all remaining grid cells



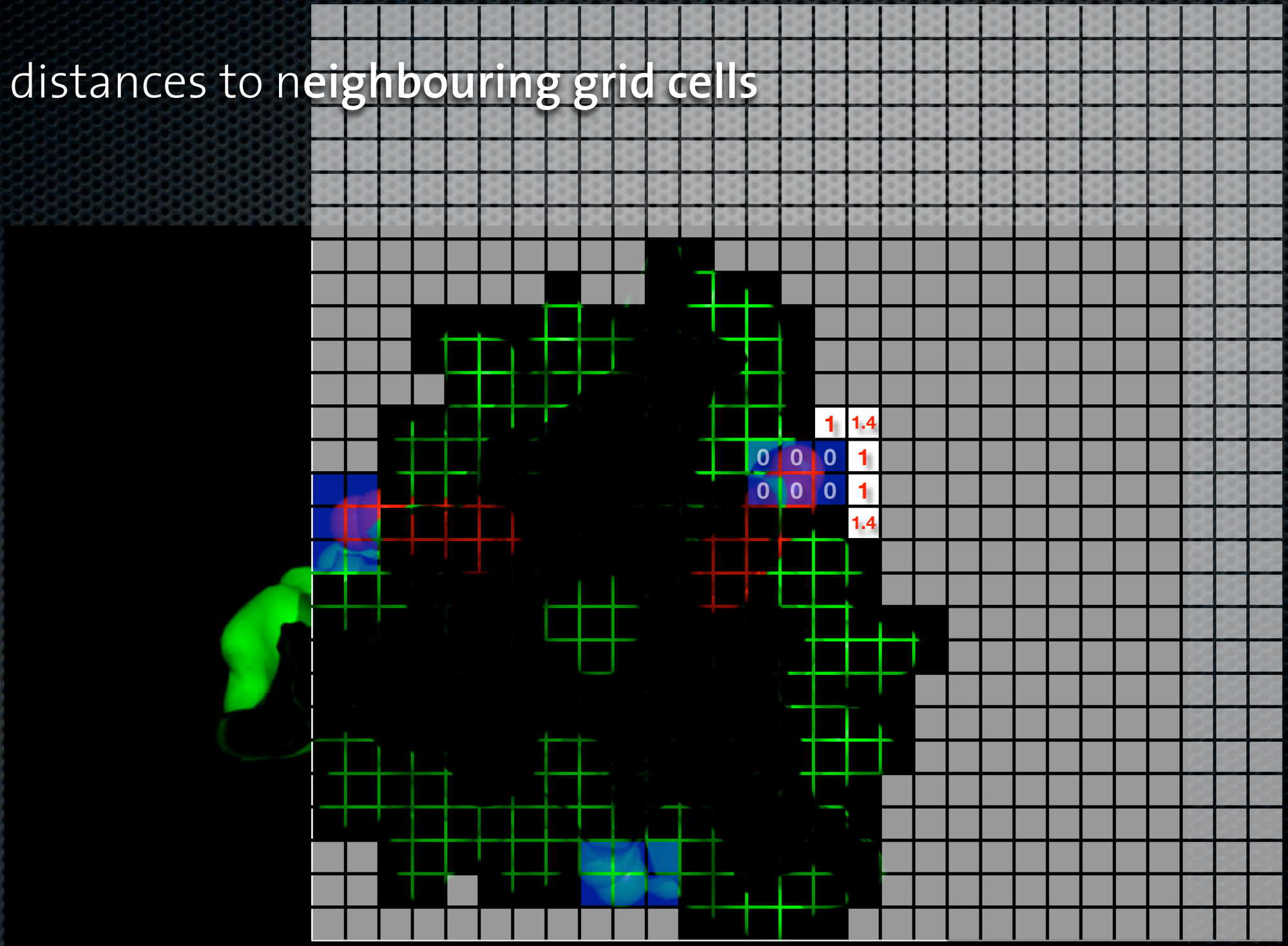
Xwalk - Algorithm VII

- Breadth-First Search
  - Assign grid cells of **ε-amine groups the distance o**



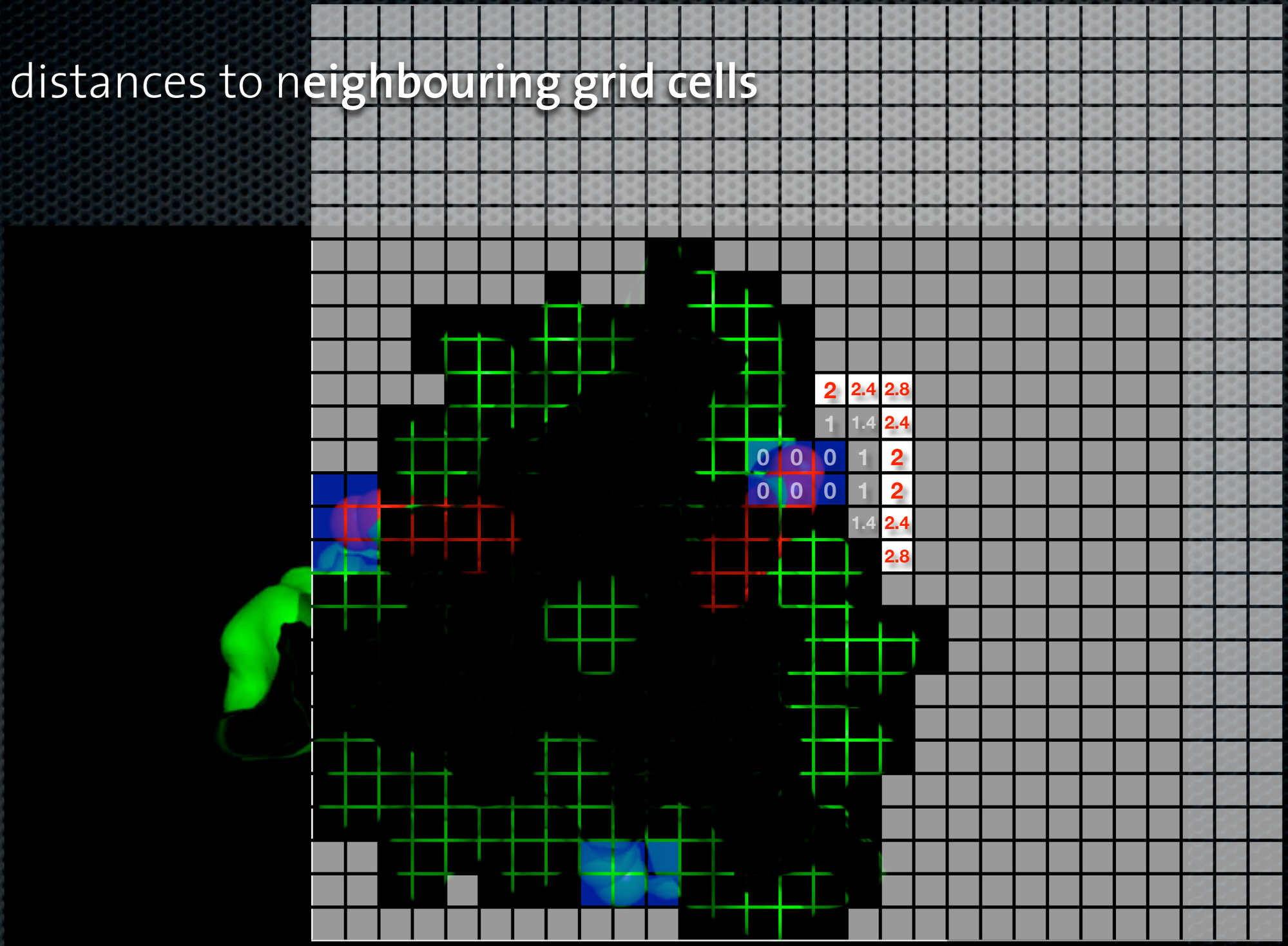
Xwalk - Algorithm VIII

- Assign distances to n**eighbouring grid cells**



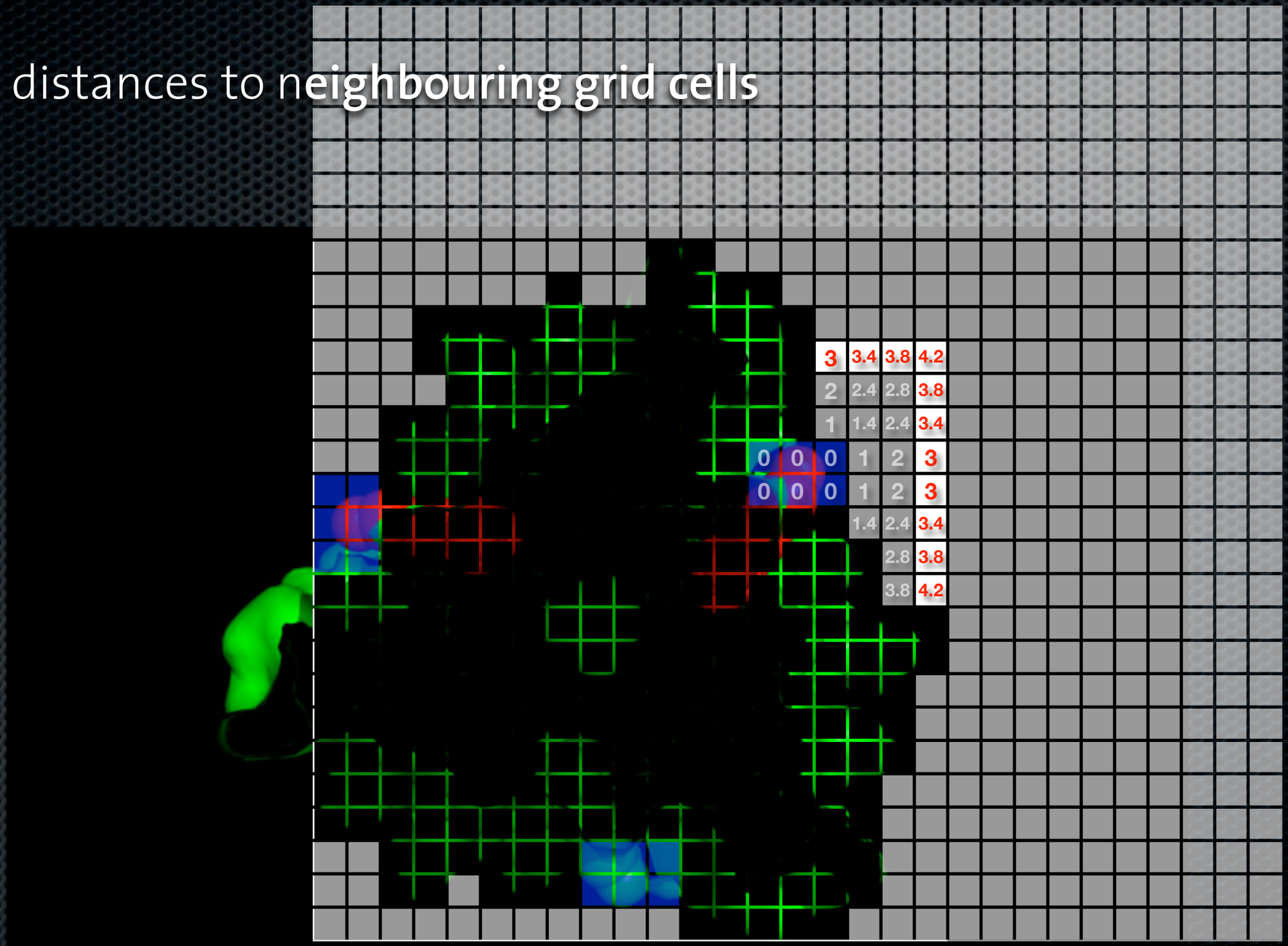
Xwalk - Algorithm IX

- Assign distances to n**eighbouring grid cells**



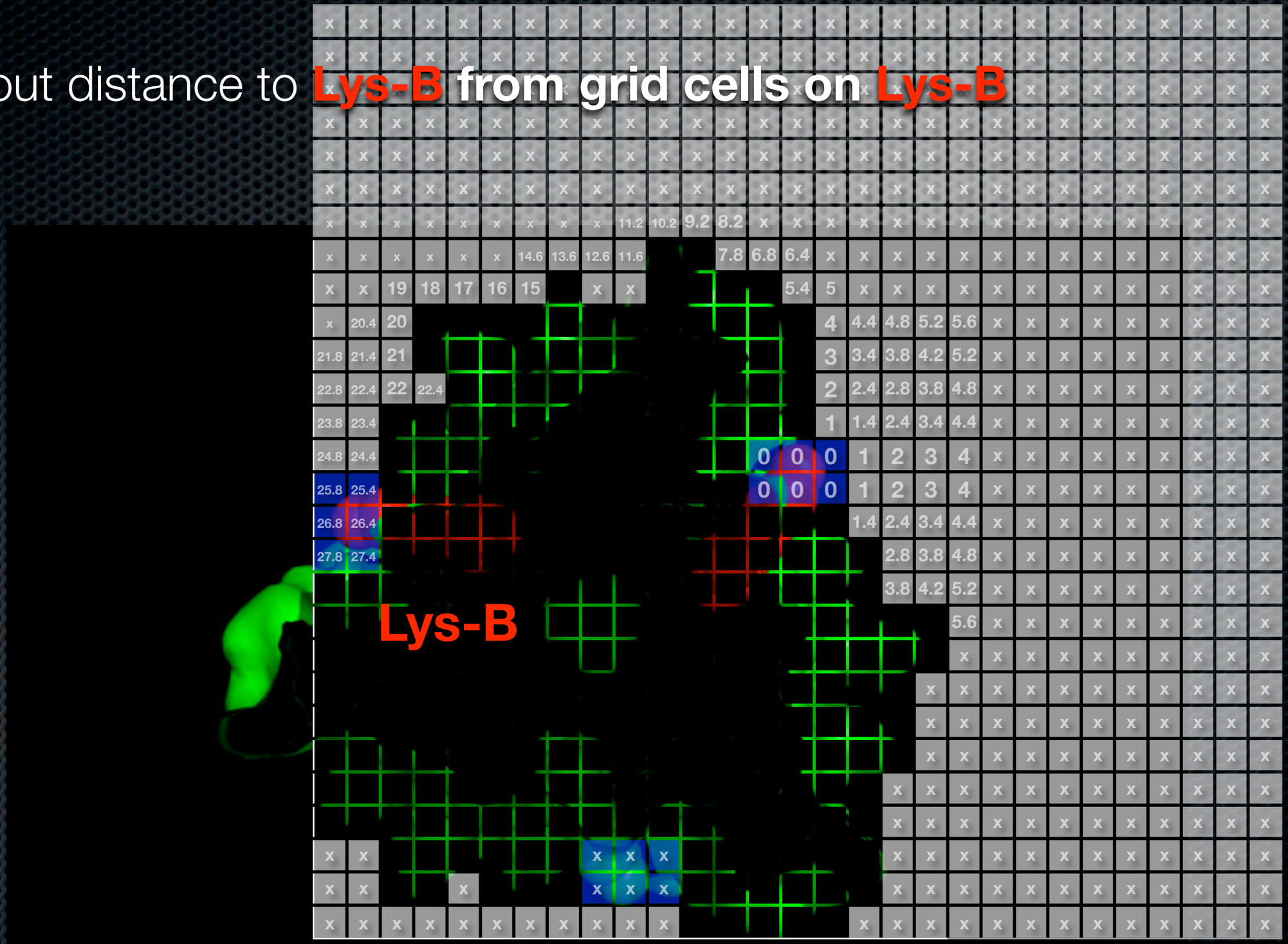
Xwalk - Algorithm X

- Assign distances to n**eighbouring grid cells**



Xwalk - Algorithm XI

- Read out distance to **Lys-B from grid cells on Lys-B**



www.xwalk.org

Beta version

Xwalk

Prediction, Validation and Visualisation of Chemical Cross-Link Data

Home | Download | About | Help | Contact

Example: ALDOA_RABIT

1. Choose your Running Mode:

(•) Validation Mode

Validate measured chemical cross-links on a protein 3D structure.

( ) Production Mode

Predict potential chemical cross-links using a protein 3D structure.

2. Choose your Input File or ID:

Upload PDB file (*max.* 1MB): Choose File no file selected

or

Give protein identifier: PDB ID

3. Set your Cross-Link Parameter:

1st residue in cross-links: Lys

2nd residue in cross-links: Lys

Index	Number of 1st Residue	Number of 2nd Residue

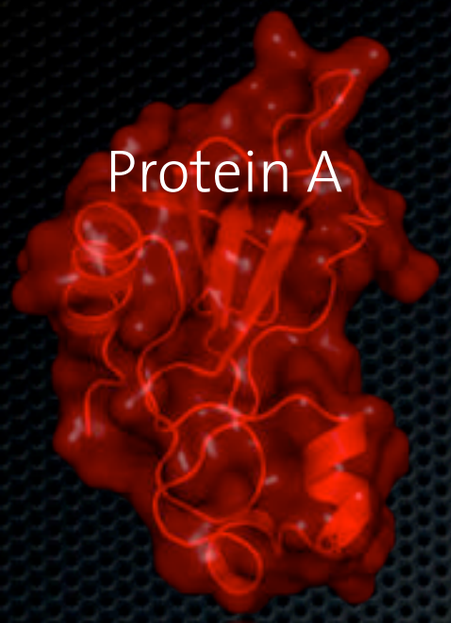
News

- Non-polypeptide molecules are removed from PDB files, which could cause Xwalk to crash. *(25/06/11)*
- A limitation on the maximum number (=150) of SASD calculations for a single protein structure has been placed. This step was necessary to save computer resources on our server as some proteins can have more than 2000 potential vXL. Users still interested in calculating SASD for very large proteins/complexes are advised to download the Xwalk executable.

1. Kahraman, A., Malmström, L. & Aebersold, R. Xwalk: Computing and Visualizing Distances in Cross-linking Experiments. *Bioinformatics* 27, 2163-2164 (2011)

XLdock

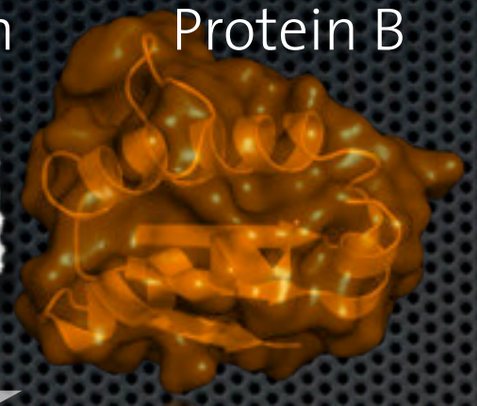
Protein A



Distance information

1	1brs.pdb	LYS-39-A-NZ	LYS-1-D-NZ	72	13.2	15.3
2	1brs.pdb	LYS-62-A-NZ	LYS-21-D-NZ	69	15.3	20.1
3	1brs.pdb	LYS-39-A-NZ	LYS-22-D-NZ	93	18.7	20.8
4	1brs.pdb	LYS-39-A-NZ	LYS-2-D-NZ	73	21.8	24.1
5	1brs.pdb	LYS-39-A-NZ	LYS-78-D-NZ	147	20.6	25.5
6	1brs.pdb	LYS-27-A-NZ	LYS-78-D-NZ	159	17.5	26.0
7	1brs.pdb	LYS-98-A-NZ	LYS-21-D-NZ	33	21.9	26.4

Protein B



Input

XLdock

Global Sampling

Filter: BSA and Xwalk

Clustering

Local Sampling

Filter: BSA and Xwalk

Clustering

Output

Complex: barnase/barstar complex (1brs-AD)

XLdock

Protein A

Distance information

1	1brs.pdb	LYS-39-A-NZ	LYS-1-D-NZ	72	13.2	15.3
2	1brs.pdb	LYS-62-A-NZ	LYS-21-D-NZ	69	15.3	20.1
3	1brs.pdb	LYS-39-A-NZ	LYS-22-D-NZ	93	18.7	20.8
4	1brs.pdb	LYS-39-A-NZ	LYS-2-D-NZ	73	21.8	24.1
5	1brs.pdb	LYS-39-A-NZ	LYS-78-D-NZ	147	20.6	25.5
6	1brs.pdb	LYS-27-A-NZ	LYS-78-D-NZ	159	17.5	26.0
7	1brs.pdb	LYS-98-A-NZ	LYS-21-D-NZ	33	21.9	26.4

Protein B

Input

XLdock

Global Sampling

Filter: BSA and Xwalk

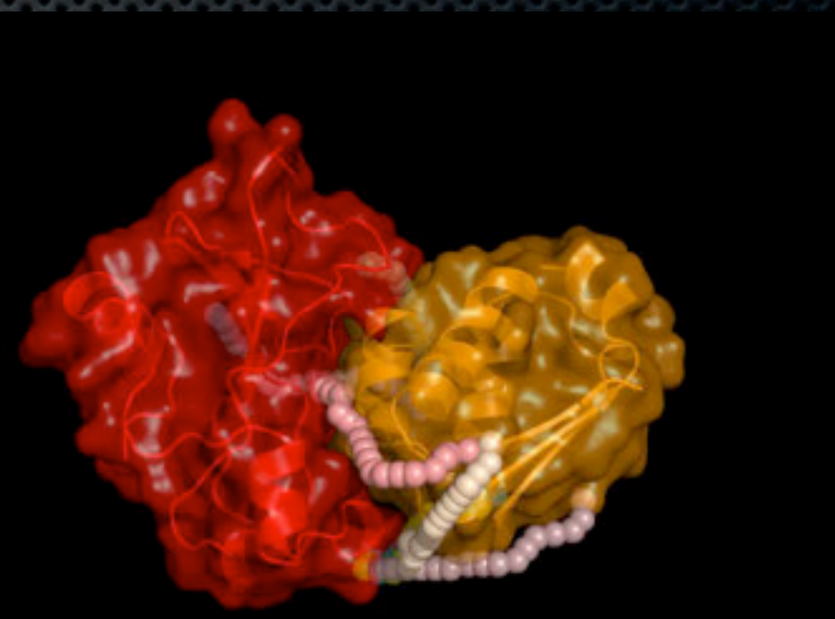
Clustering

Local Sampling

Filter: BSA and Xwalk

Clustering

Output



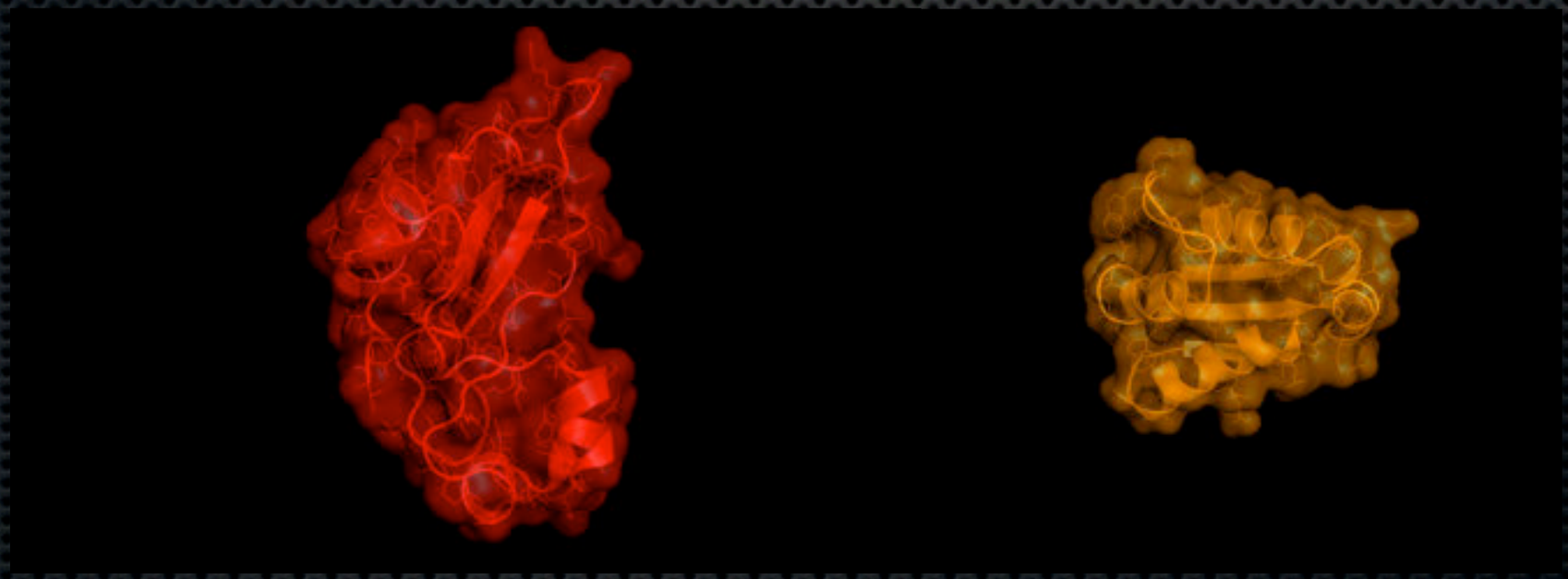
Complex: barnase/barstar complex (1brs-AD)

XLdock - Automated Docking Pipeline

- Test Run (Check for errors in submitted structures)
- **Relax** each protein component of the complex

XLdock - Automated Docking Pipeline

- Test Run (Check for errors in submitted structures)
- **Relax** each protein component of the complex
  - Choose minimum energy structure of 10 relaxation run.



Crystal Structure Intemediate Relaxation Energy minimum

XLdock - Automated Docking Pipeline

- Test Run (Check for errors in submitted structures)
- **Relax** each protein component of the complex
- Execution time estimation for global sampling

$$t_{test} = \sum_{i=1}^{N} t_i$$

$$d_{job}(T, \bar{t}_i) = T/\bar{t}_i * c_t$$

$$n_{job}(M, d_{job}) = M/d_{job}$$

N =10 decoys

T = 8h = 28,000s
c_t = 0.75

M = 100,000 decoys

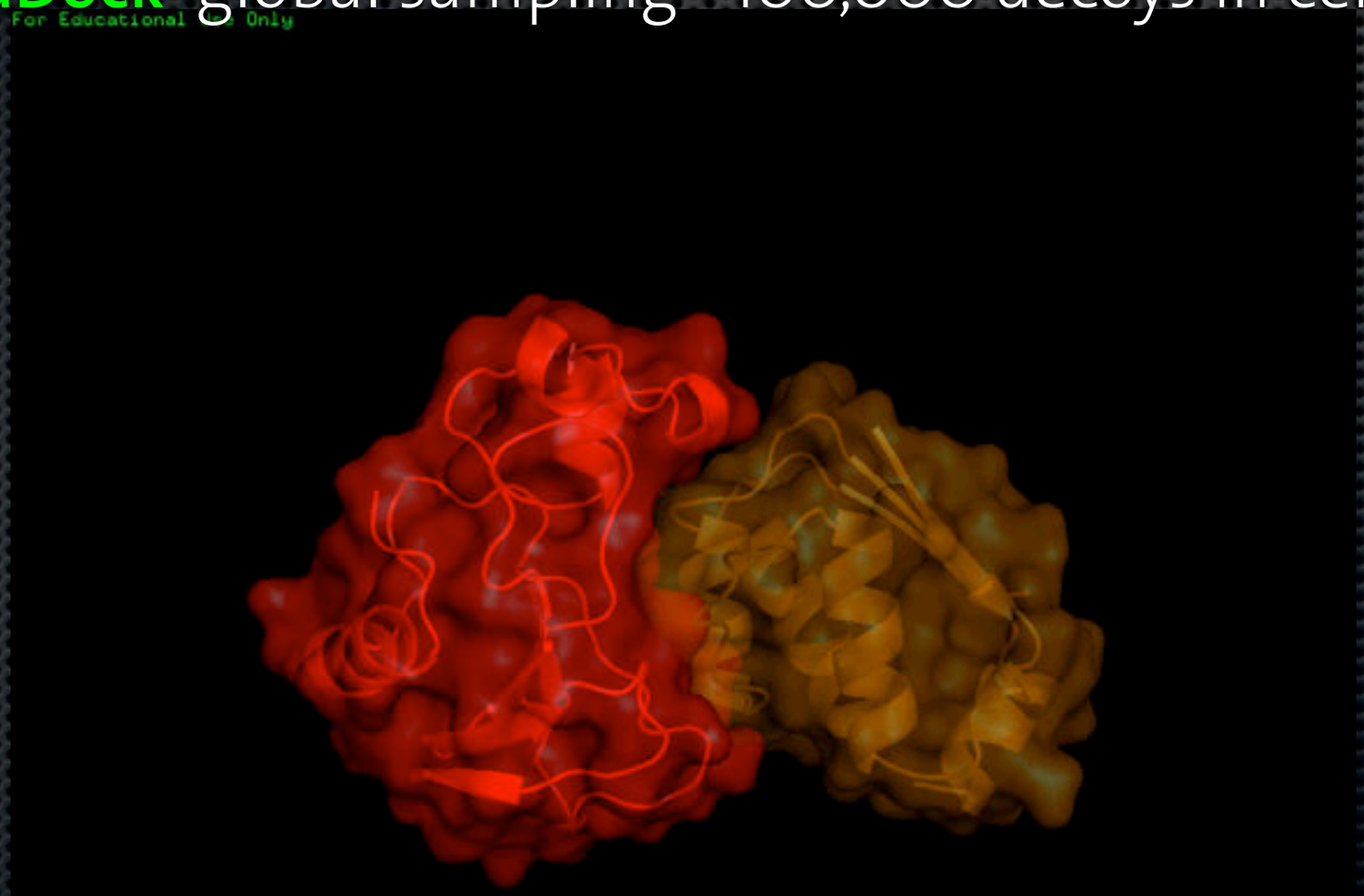
XLdock - Automated Docking Pipeline

- Test Run (Check for errors in submitted structures)
- **Relax** each protein component of the complex
- Execution time estimation for global sampling
- **RosettaDock**[1] global sampling - 100,000 decoys in centroid mode

1. Gray, J. et al. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol **331**, 281–299 (2003).

XLdock - Automated Docking Pipeline

- Test Run (Check for errors in submitted structures)
- **Relax** each protein component of the complex
- Execution time estimation for global sampling
- **RosettaDock**[1] global sampling - 100,000 decoys in centroid mode



= native complex

= decoy

1. Gray, J. et al. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol **331**, 281–299 (2003).

XLdock - Automated Docking Pipeline

- Test Run (Check for errors in submitted structures)
- **Relax** each protein component of the complex
- Execution time estimation for global sampling
- **RosettaDock**[1] global sampling - 100,000 decoys in centroid mode
- Filtering decoys by **SASD** with Xwalk
- Extract top 500 decoys with lowest Rosetta score

1. Gray, J. et al. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol **331**, 281–299 (2003).
Janin, J., Bahadur, R.P. & Chakrabarti, P. Protein-protein interaction and quaternary structure. *Q Rev Biophys* 41, 133–180 (2008).

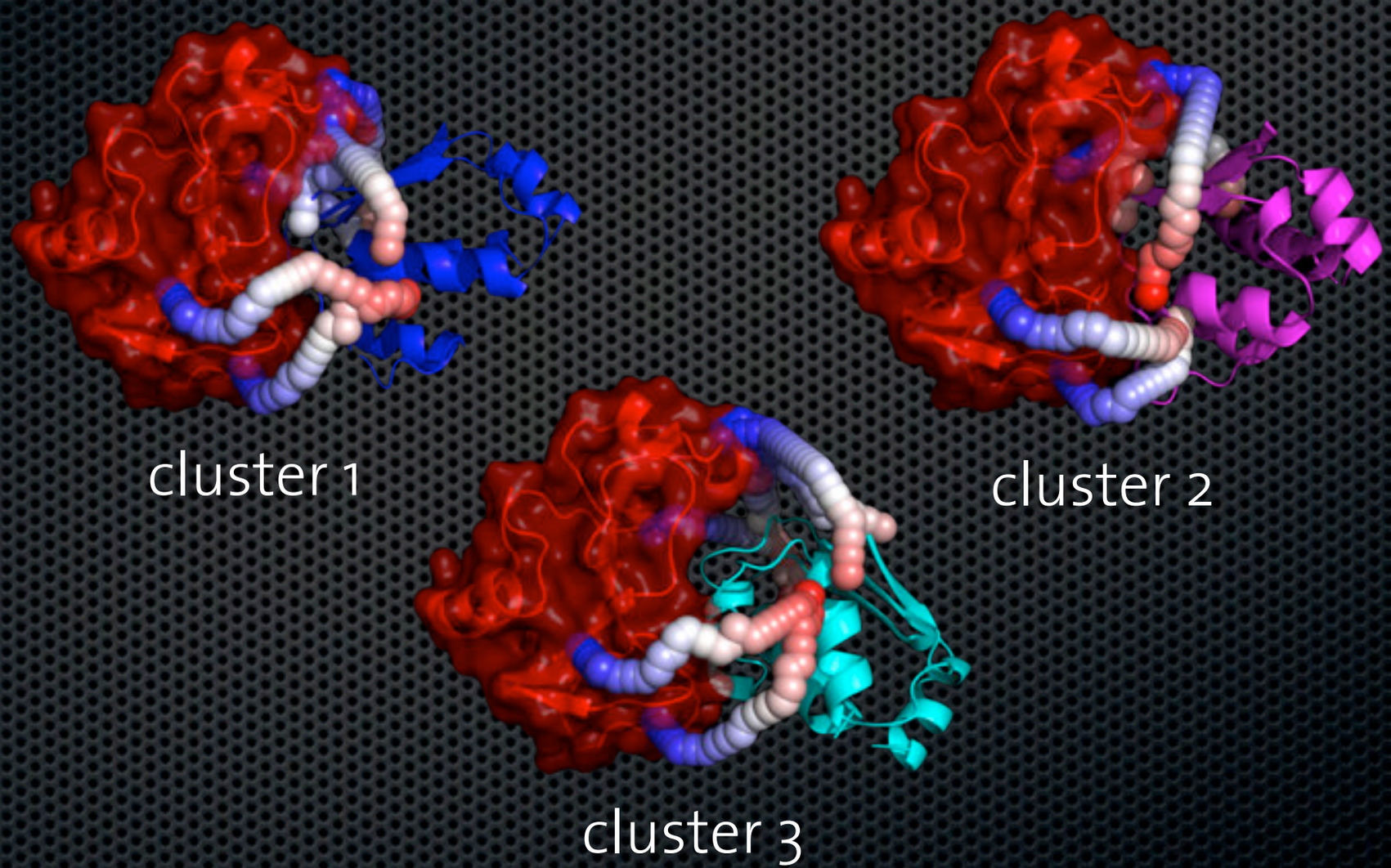
XLdock - Automated Docking Pipeline

- Test Run (Check for errors in submitted structures)
- **Relax** each protein component of the complex
- Execution time estimation for global sampling
- **RosettaDock**[1] global sampling - 100,000 decoys in centroid mode
- Filtering decoys by **SASD** with Xwalk
- Extract top 500 decoys with lowest Rosetta score
- Filter by **BSA** > 900 Å^2 (Janin et al.)

1. Gray, J. et al. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol **331**, 281–299 (2003).
Janin, J., Bahadur, R.P. & Chakrabarti, P. Protein-protein interaction and quaternary structure. *Q Rev Biophys* 41, 133–180 (2008).

XLdock - Automated Docking Pipeline

- Quality threshold **clustering** and choosing largest three clusters



cluster 1

cluster 2

cluster 3

XLdock - Automated Docking Pipeline

- Quality threshold clustering and choosing largest three clusters
- Execution time estimation for local sampling

$$t_{test} = \sum_{i=1}^{N} t_i$$

$$d_{job}(T, \bar{t}_i) = T/\bar{t}_i * c_t$$

$$n_{job}(M, d_{job}) = M/d_{job}$$

N = 5 decoys

T = 8h = 28,000s

c_t = 0.75

M = 5,000 decoys

XLdock - Automated Docking Pipeline

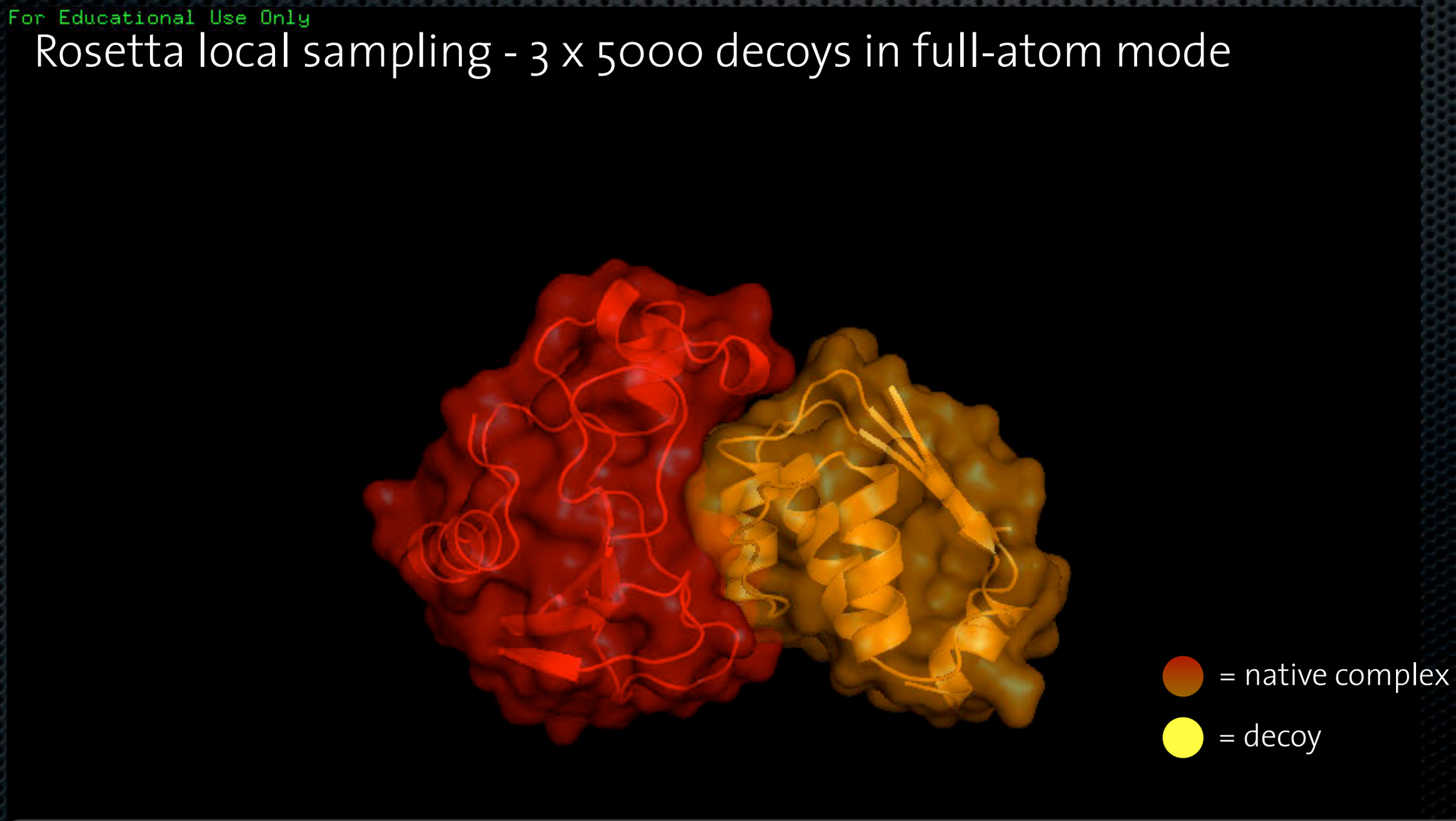
- Quality threshold clustering and choosing largest three clusters
- Execution time estimation for local sampling
- Rosetta local sampling - 3 x 5000 decoys in full-atom mode

XLdock - Automated Docking Pipeline

- Quality threshold clustering and choosing largest three clusters
- Execution time estimation for local sampling

For Educational Use Only

- Rosetta local sampling - 3 x 5000 decoys in full-atom mode

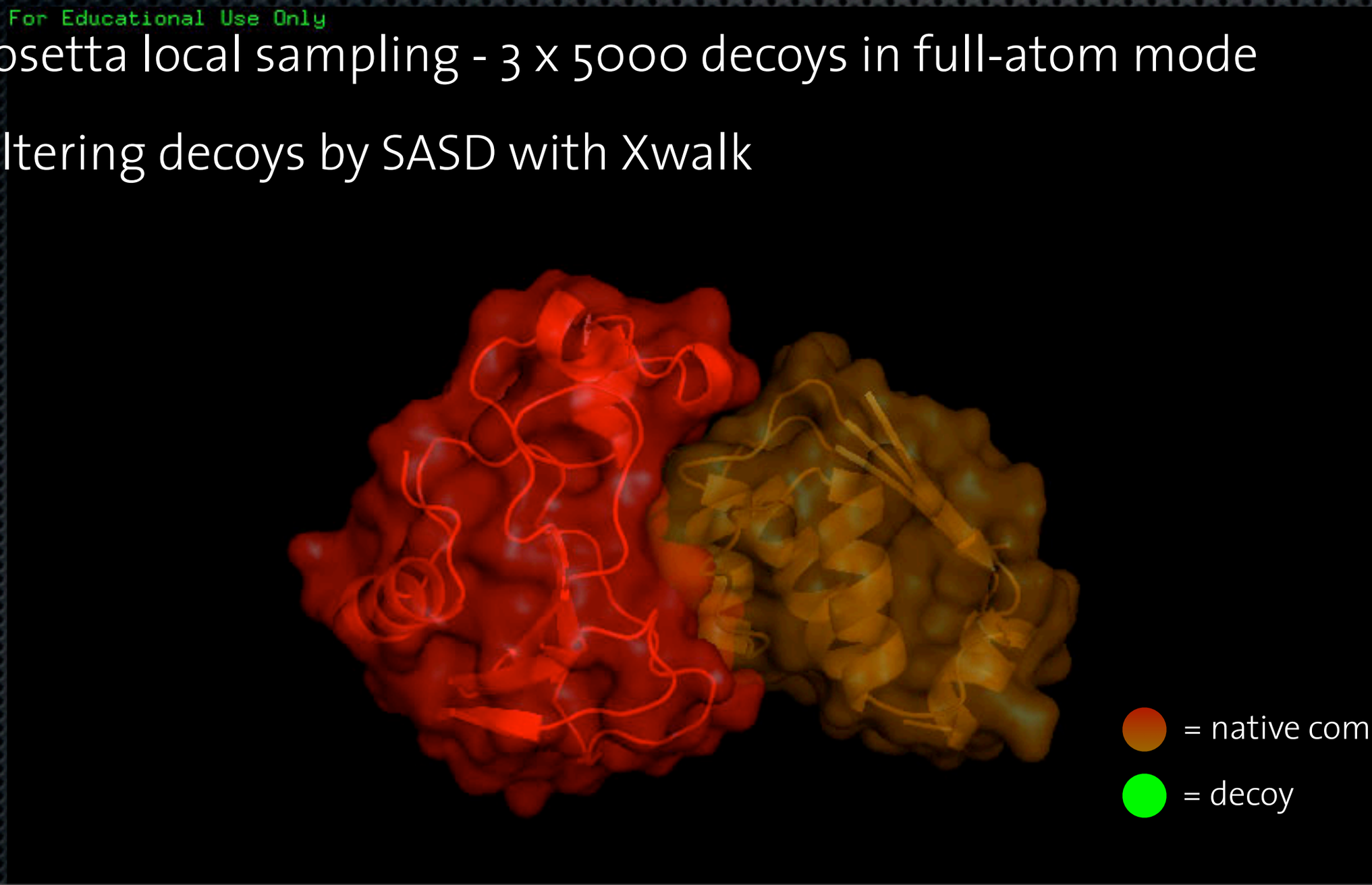


XLdock - Automated Docking Pipeline

- Quality threshold clustering and choosing largest three clusters
- Execution time estimation for local sampling
- Rosetta local sampling - 3 x 5000 decoys in full-atom mode
- Filtering decoys by SASD with Xwalk

XLdock - Automated Docking Pipeline

- Quality threshold clustering and choosing largest three clusters
- Execution time estimation for local sampling
- Rosetta local sampling - 3 x 5000 decoys in full-atom mode
- Filtering decoys by SASD with Xwalk

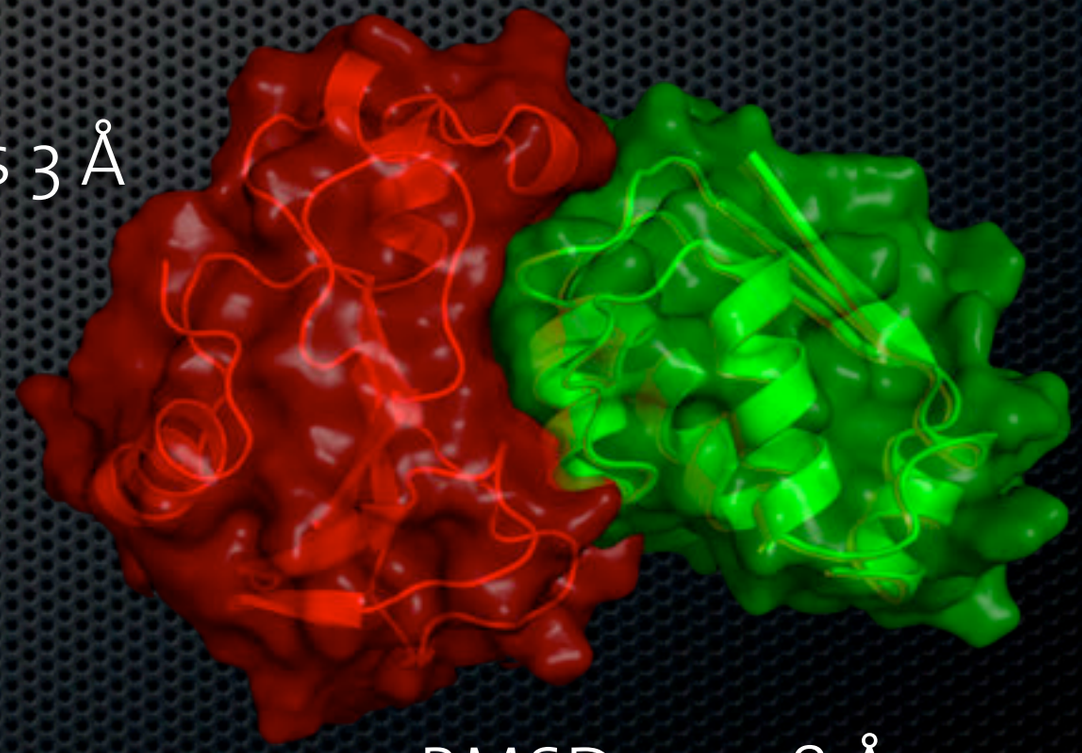


XLdock - Automated Docking Pipeline

- Quality threshold clustering and choosing largest three clusters
- Execution time estimation for local sampling
- Rosetta local sampling - 3 x 5000 decoys in full-atom mode
- Filtering decoys by SASD with Xwalk
- Extract top 500 decoys with lowest Rosetta score and filter by BSA > 900 Å^2

XLdock - Automated Docking Pipeline

- Quality threshold clustering and choosing largest three clusters
- Execution time estimation for local sampling
- Rosetta local sampling - 3 x 5000 decoys in full-atom mode
- Filtering decoys by SASD with Xwalk
- Extract top 500 decoys with lowest Rosetta score and filter by BSA > 900 Å^2
- Hierarchal clustering with cluster radius 3 Å



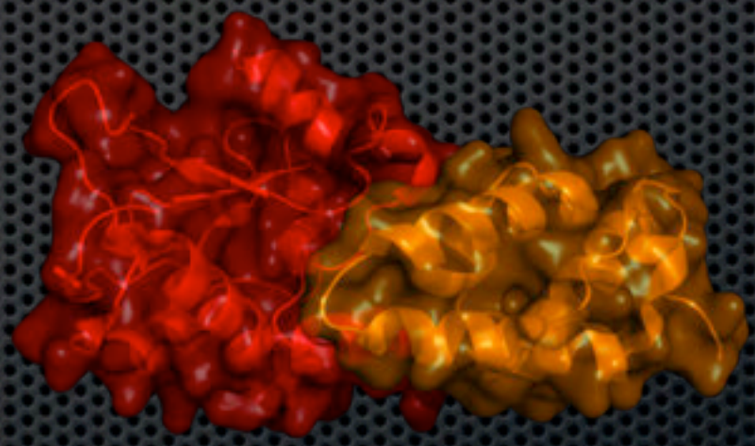
RMSD = 0.58 Å

Test Cases

- Complex 1
  - Barnase, Barstar (PDB: 1brs)
  - 7 interprotein LYS **virtual** XL

Test Cases

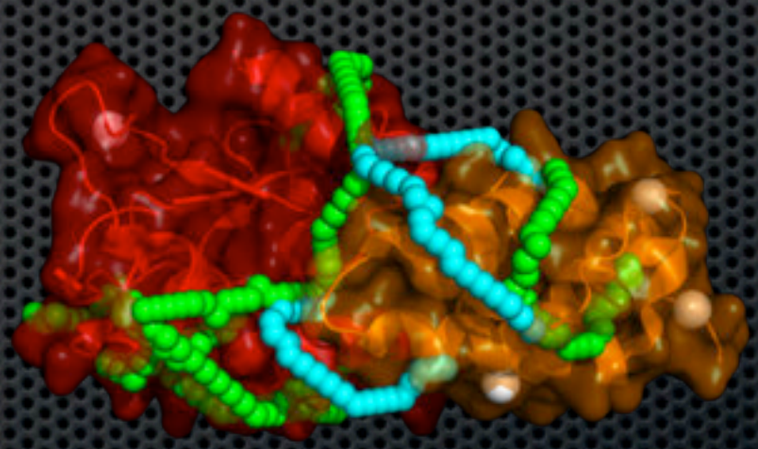
- Complex 1
  - Barnase, Barstar (PDB: 1brs)
  - 7 interprotein LYS **virtual** XL
- Complex 2:
  - Colicin DNAse, Colicin inhibitor (PDB: 1ujz)
  - 3 inter-protein XL, 16 intra-protein XL, 5 mono-links all **experimental** [1]



1. Seebacher, J. et al. Protein cross-linking analysis using mass spectrometry, isotope-coded cross-linkers, and integrated computational data processing. J Proteome Res 5, 2270–2282 (2006).

Test Cases

- Complex 1
  - Barnase, Barstar (PDB: 1brs)
  - 7 interprotein LYS **virtual** XL
- Complex 2:
  - Colicin DNAse, Colicin inhibitor (PDB: 1ujz)
  - 3 inter-protein XL, 16 intra-protein XL, 5 mono-links all **experimental** [1]



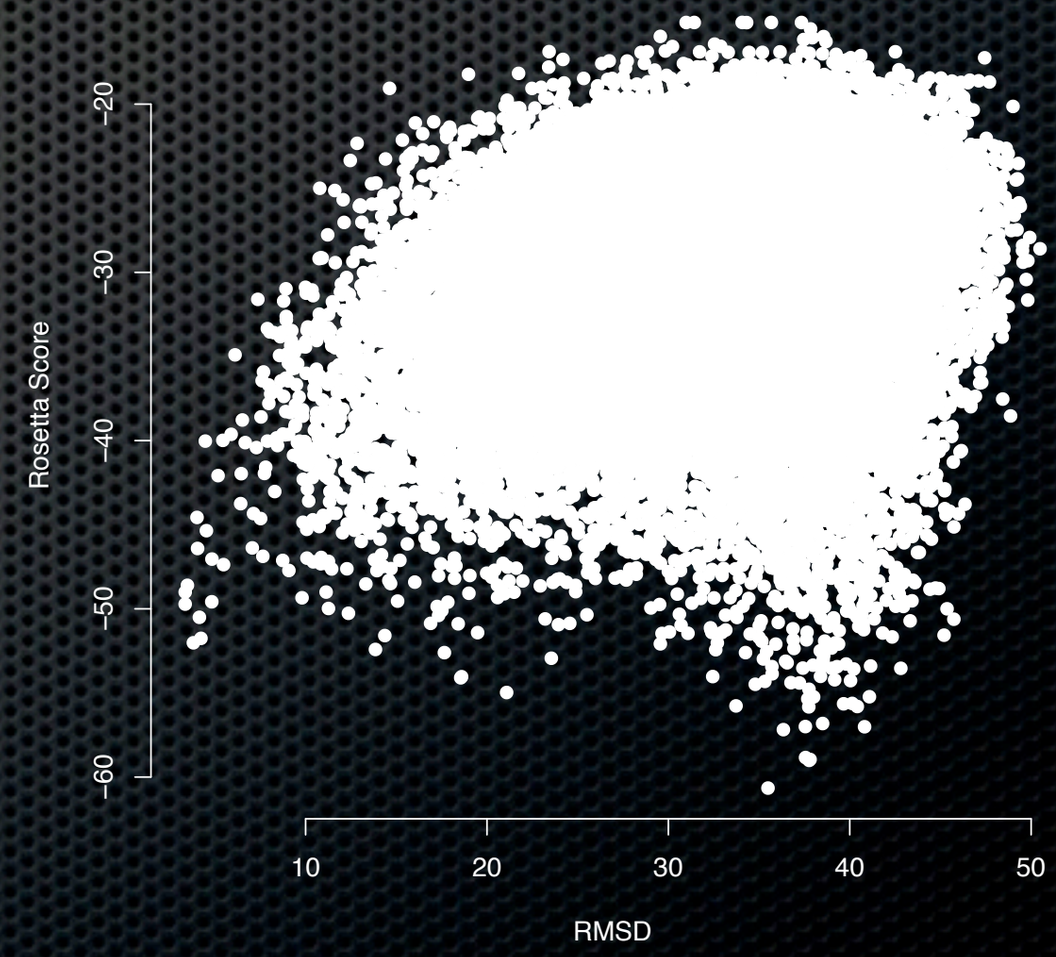
1. Seebacher, J. et al. Protein cross-linking analysis using mass spectrometry, isotope-coded cross-linkers, and integrated computational data processing. J Proteome Res 5, 2270–2282 (2006).

Colicin DNAse - Inhibitor

- Global sampling stage:
  - 100,000 decoys in total

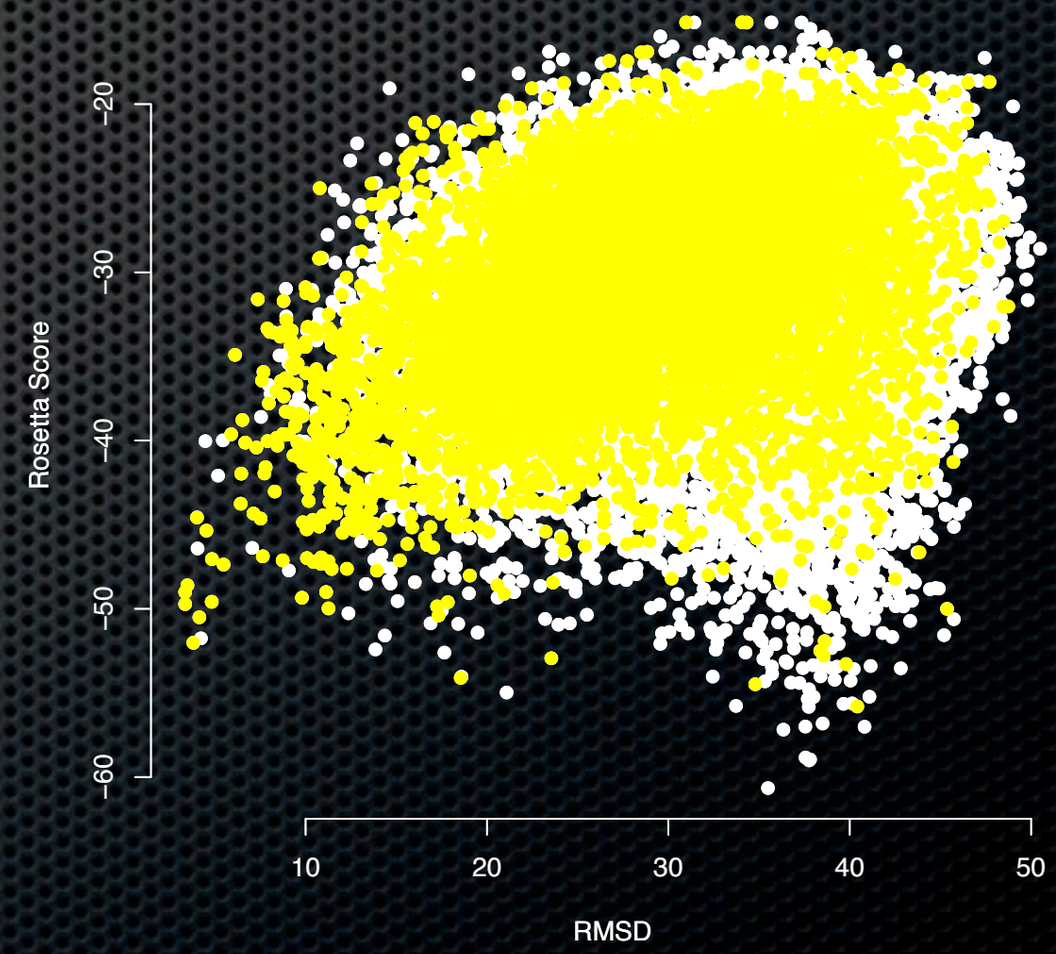
Colicin DNAse - Inhibitor

- Global sampling stage:
  - 100,000 decoys in total
    - **32,539** pass Euclidean distance filter



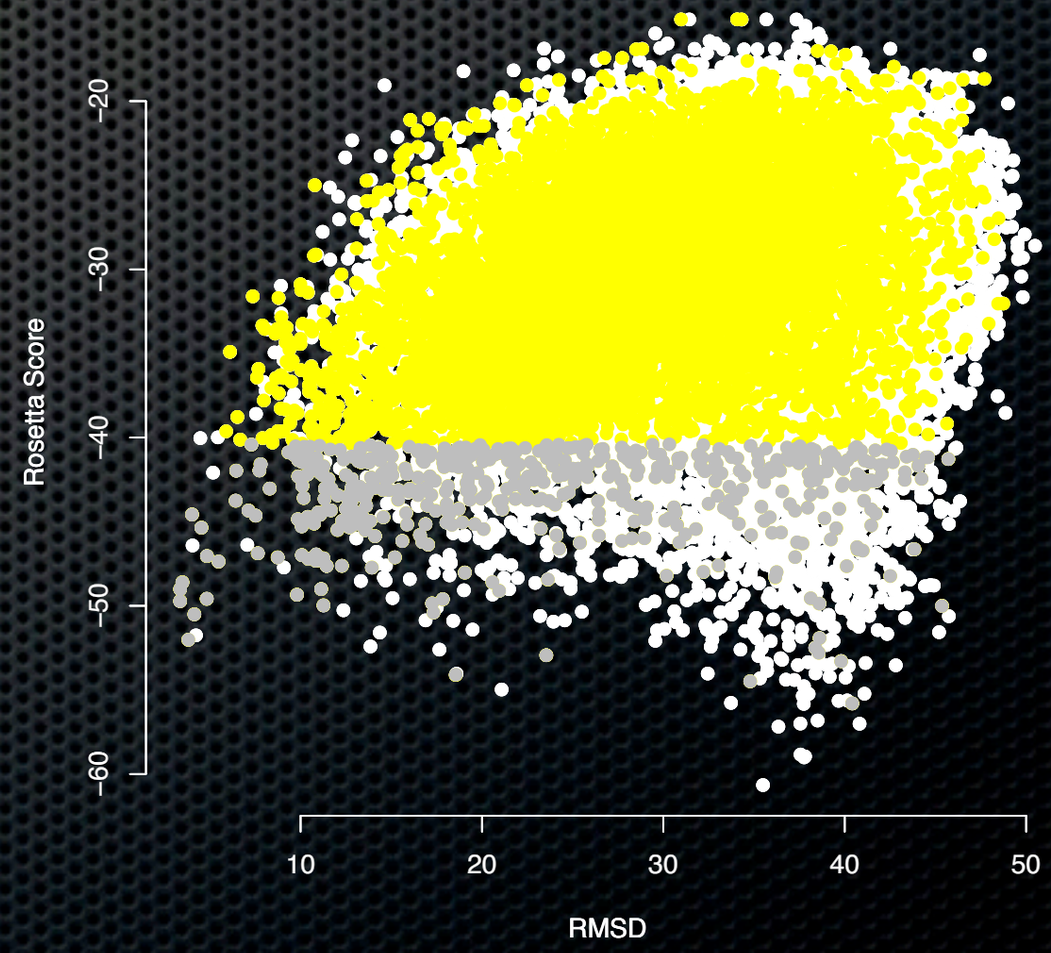
Colicin DNAse - Inhibitor

- Global sampling stage:
  - 100,000 decoys in total
    - **32,539** pass Euclidean distance filter
      - **8865** pass Xwalk filter



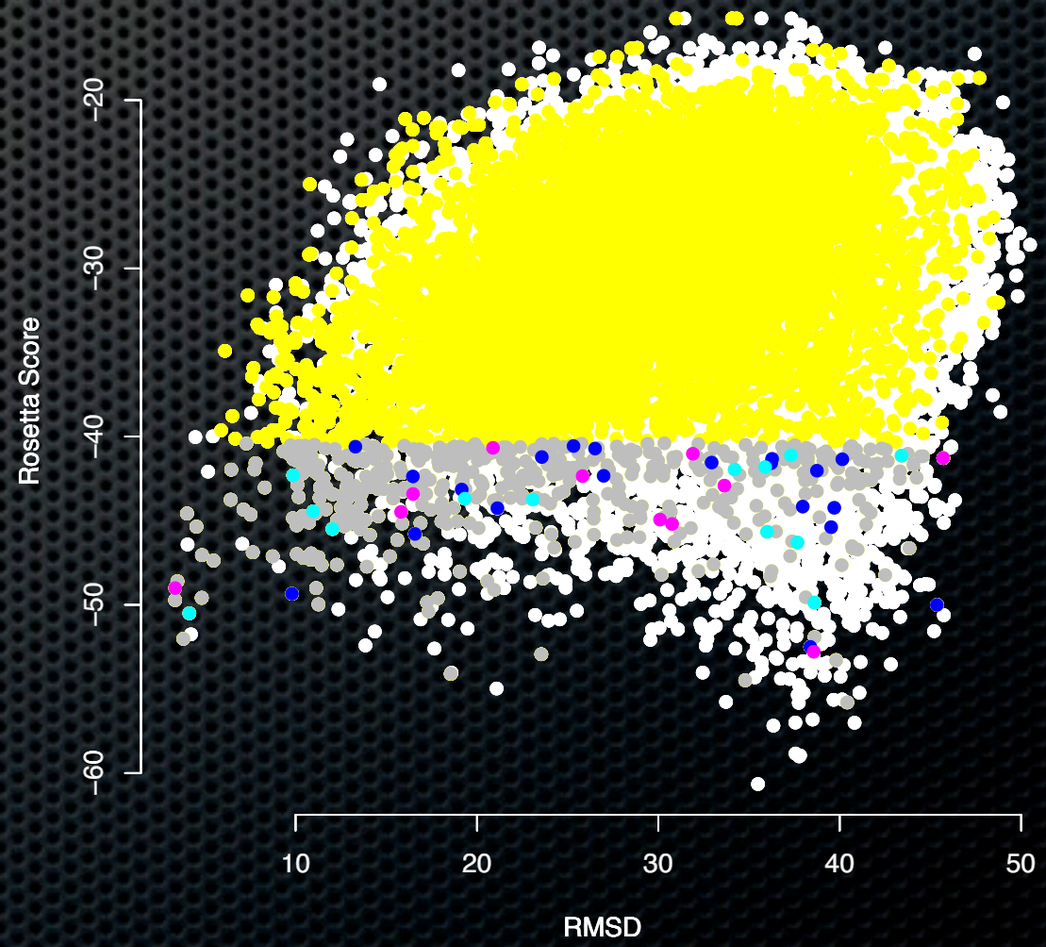
Colicin DNAse - Inhibitor

- Global sampling stage:
  - 100,000 decoys in total
    - **32,539** pass Euclidean distance filter
      - **8865** pass Xwalk filter
        - top 500 decoys

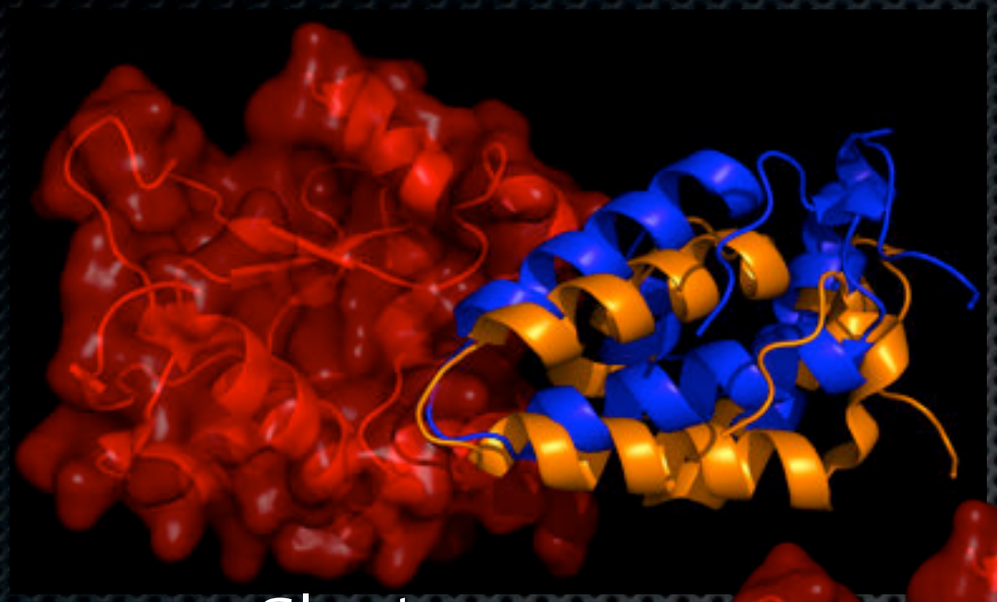


Colicin DNAse - Inhibitor

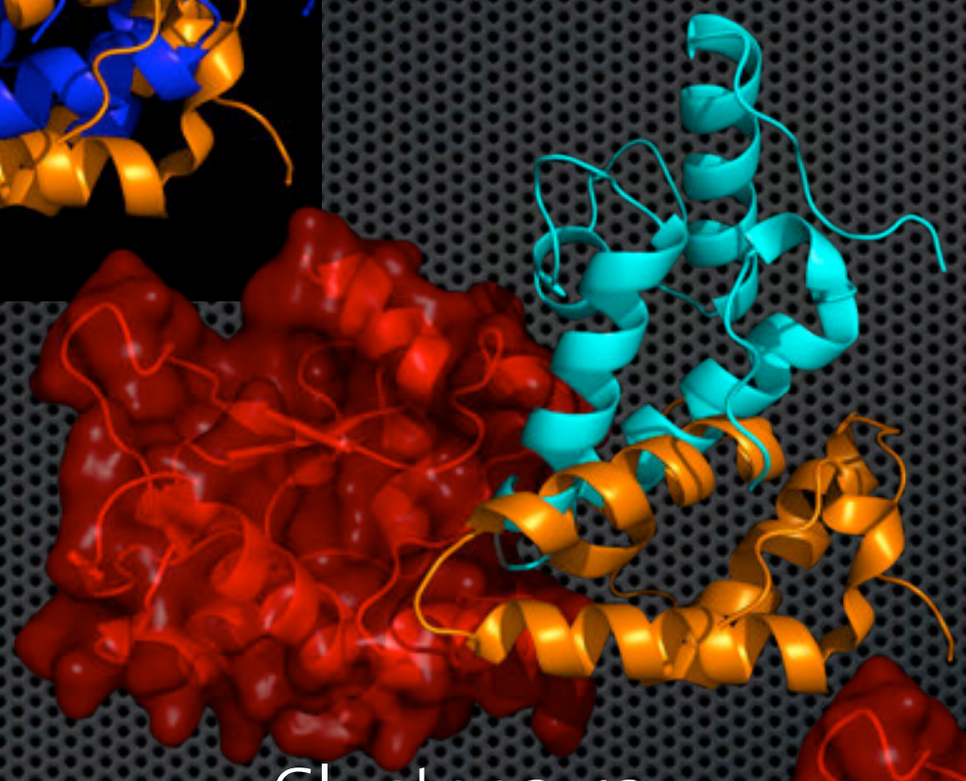
- Global sampling stage:
  - 100,000 decoys in total
    - **32,539** pass Euclidean distance filter
      - **8865** pass Xwalk filter
        - top 500 decoys
        - top 3 cluster sizes: 20, 13, 11



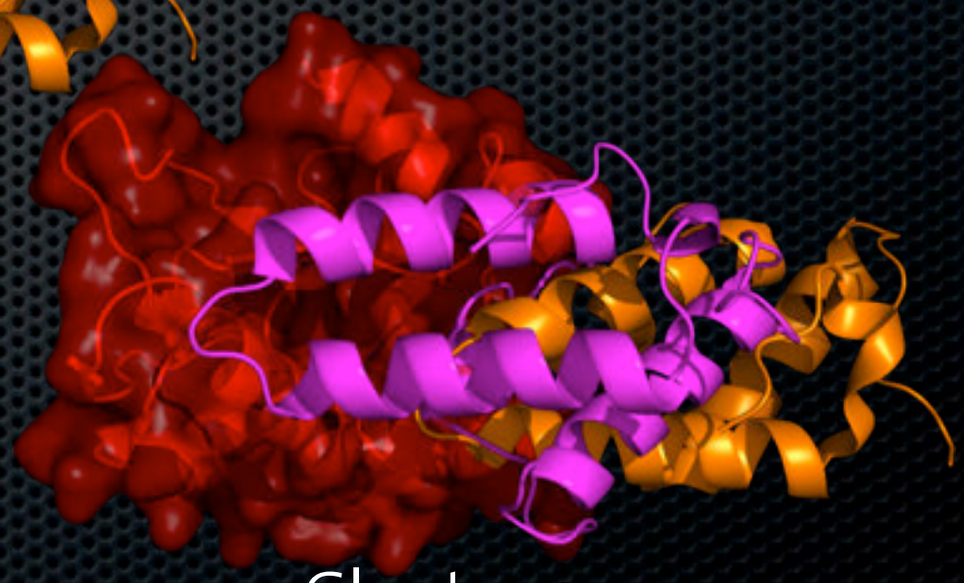
Colicin DNAse - Inhibitor



Cluster 1: 20



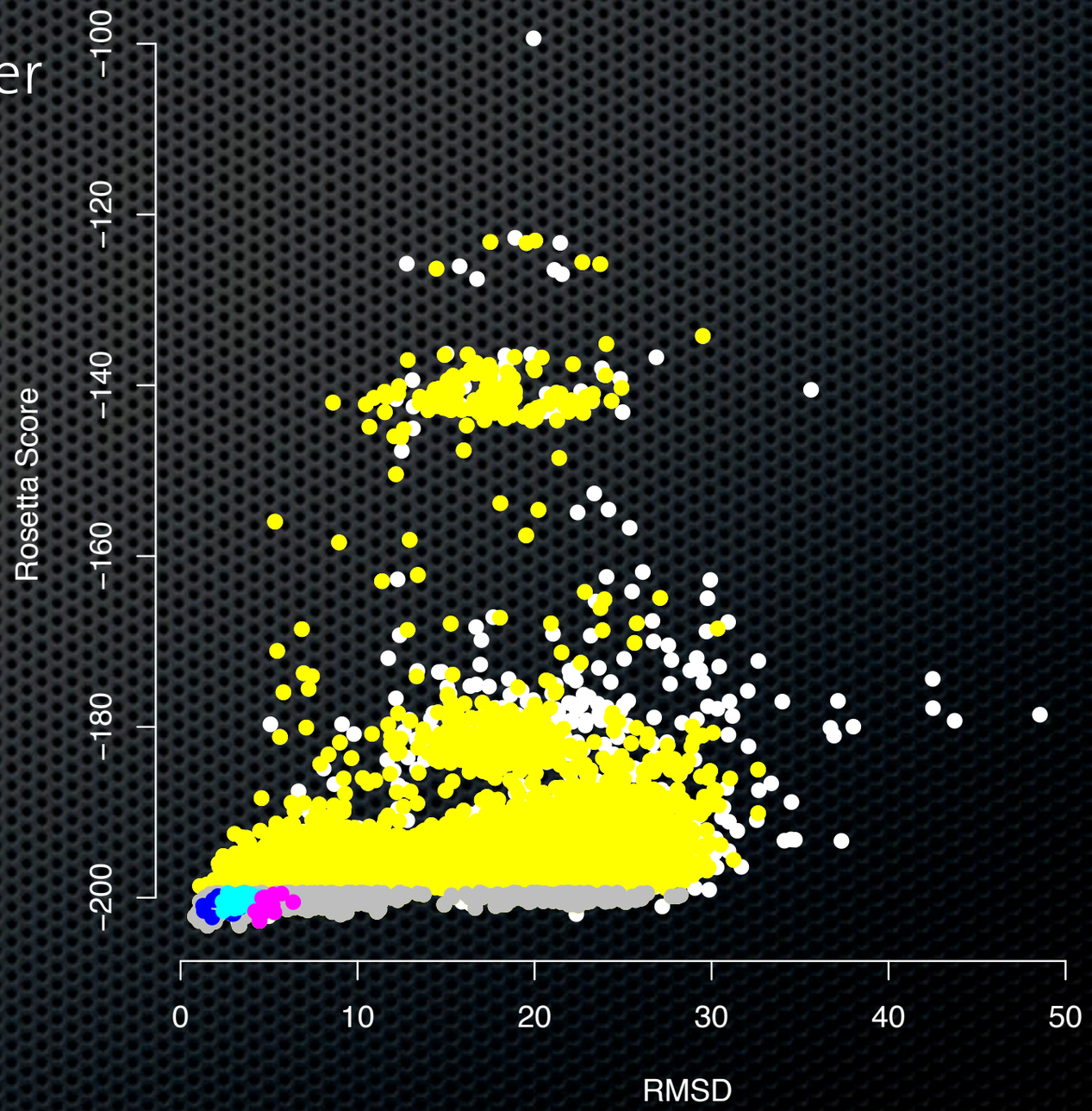
Cluster 2: 13



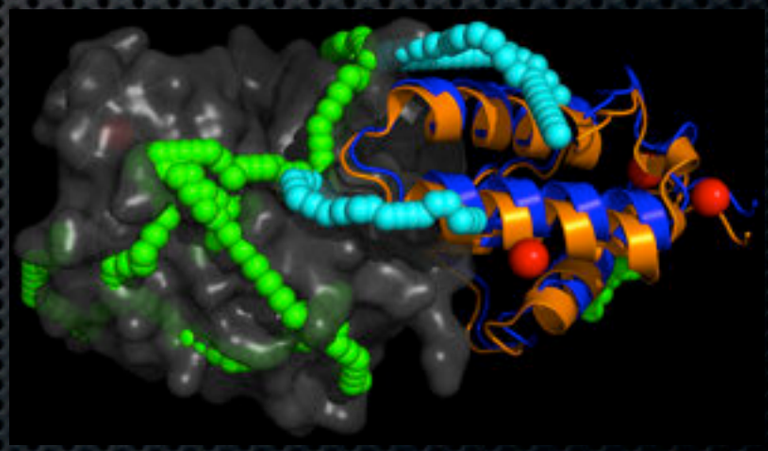
Cluster 3: 11

Colicin DNAse - Inhibitor

- Local sampling stage:
  - 3 x 5,000 decoys in total
  - 14,995 pass Euclidean distance filter
  - 13,230 pass Xwalk filter
  - top 500 decoys
  - top 3 cluster sizes:24, 20, 16

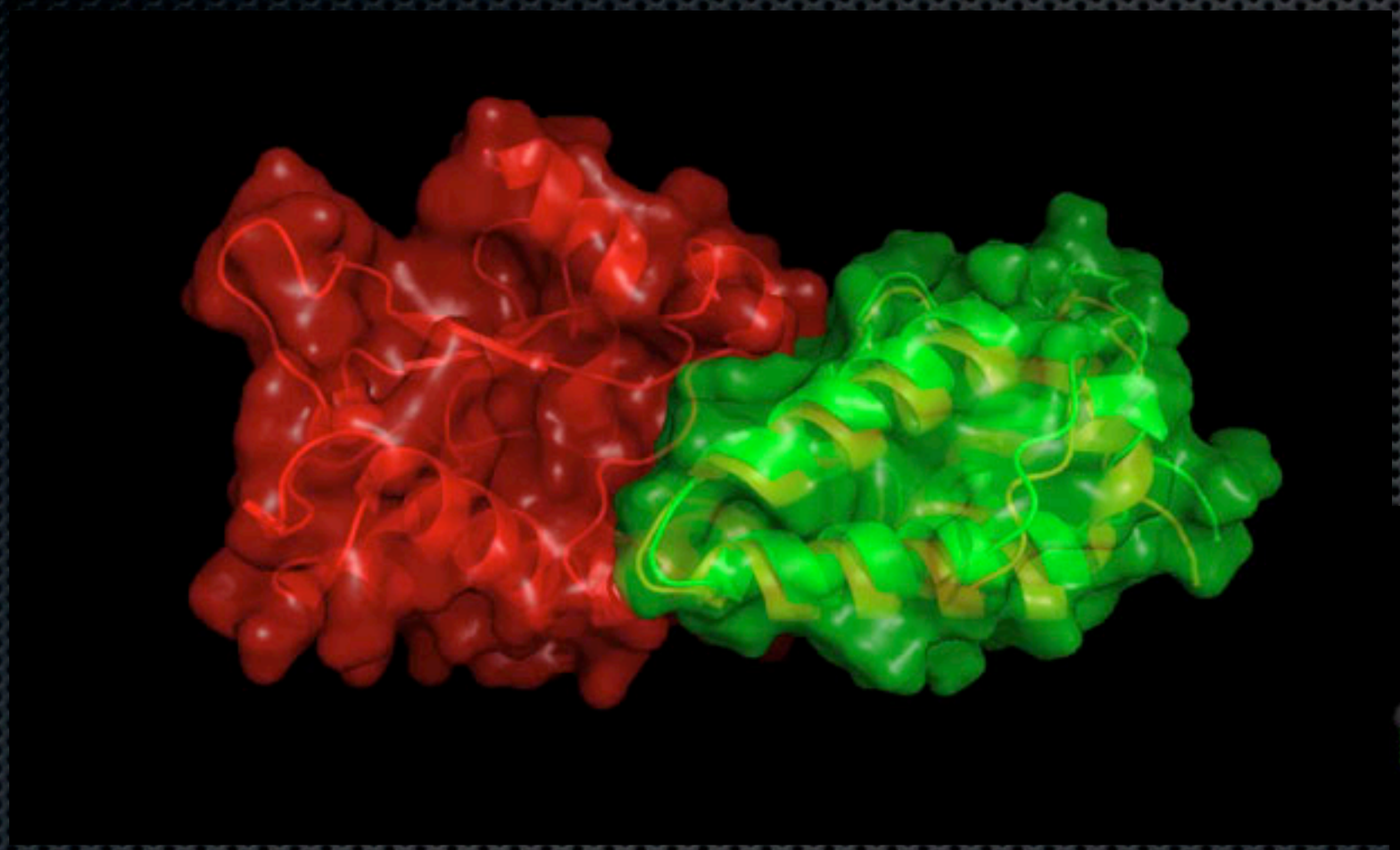


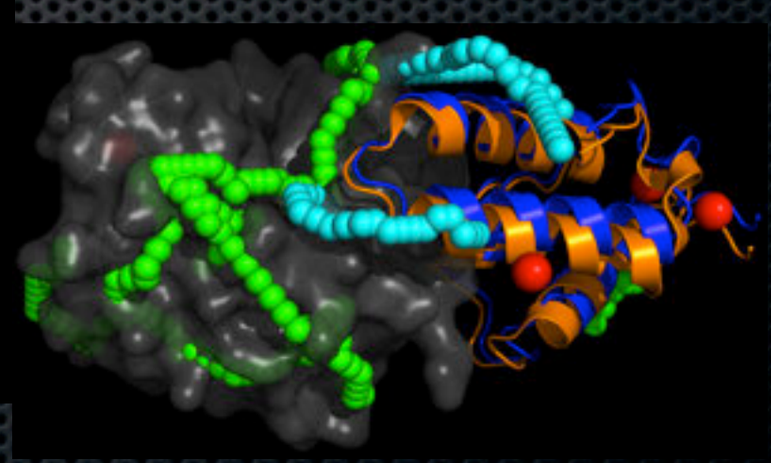
Colicin DNAse - Inhibitor



- Lowest energy decoy in largest hierarchal cluster of local sampled decoys
  - RMSD = 1.78 Å

Colicin DNAse - Inhibitor





- Lowest energy decoy in largest hierarchal cluster of local sampled decoys
  - RMSD = 1.78 Å

Acknowledgment

Lars Malmström

Ruedi Aebersold

www.xwalk.org

Franz Herzog

Funding:

Thomas Walzthöni

Alexander Leitner