Structure Pattern Discovery

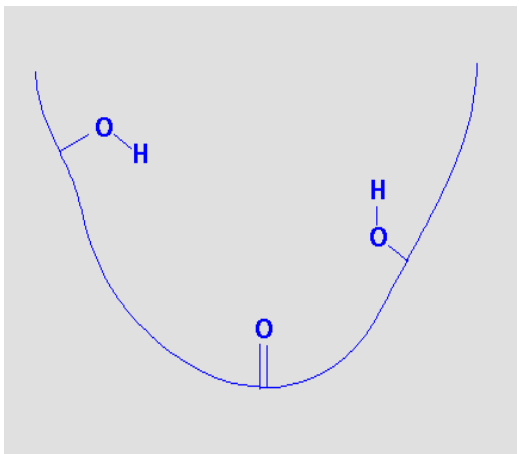
COT 6936

Tom Milledge

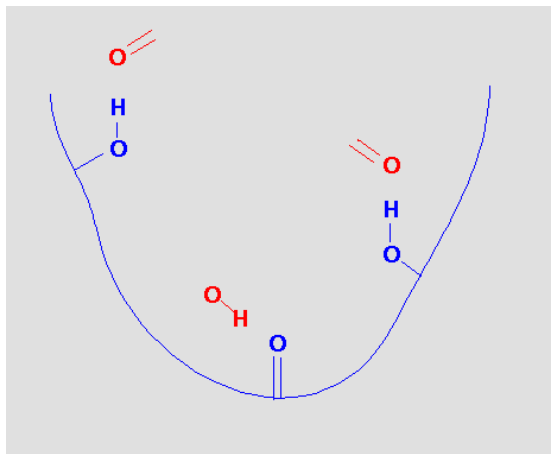
- Background on Structure Patterns
- DNA Binding Motifs
- Motif Pattern Discovery Tool
- Structure Patterns in HTH and Winged HTH proteins

Structure Patterns: Pharmacophores

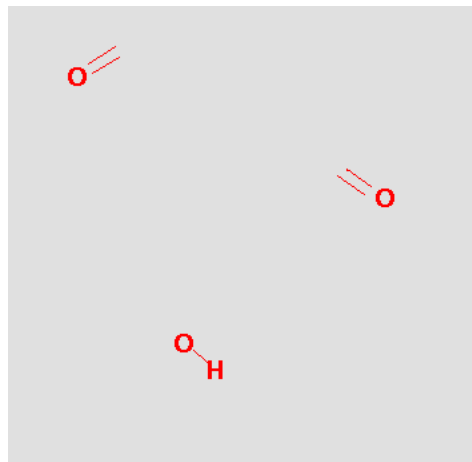
Binding site



Functional groups

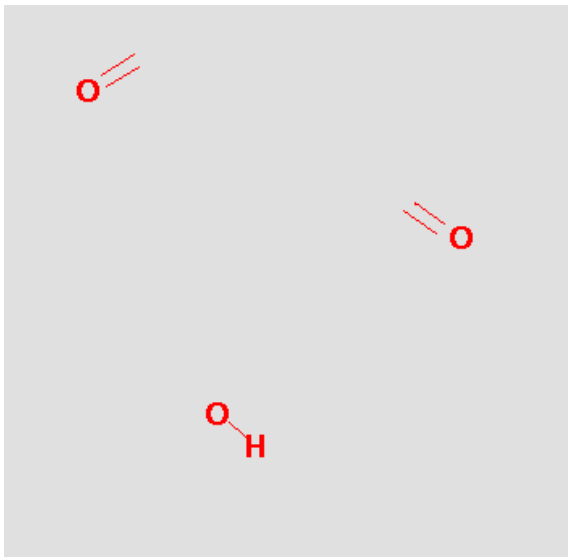


Pharmacophore

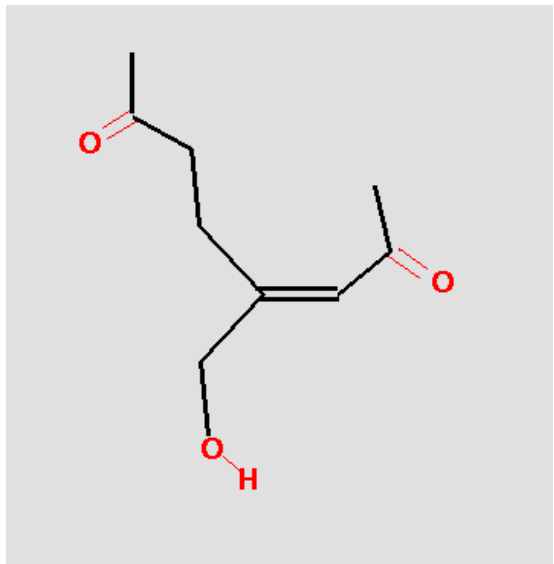


Structure Patterns: Pharmacophores

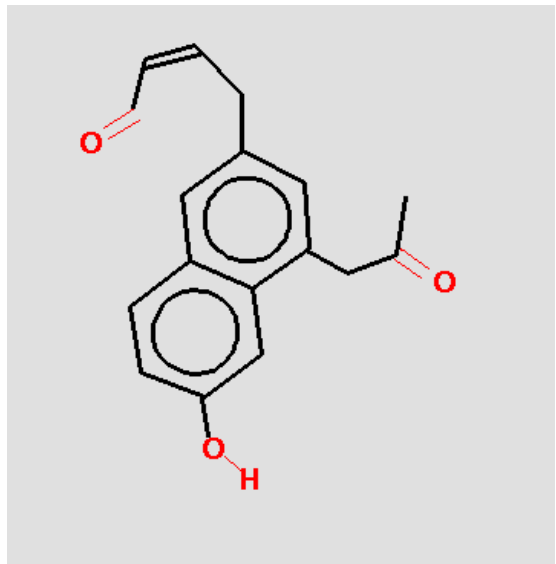
Pharmacophore



Database hit 1



Database hit 2



Structure Patterns: Jonassen

Table 3. The matches to the pattern represented by the sequence motif V-x(9,223)-L-x(2)-A-x(3)-A

Protein	Pattern			
	V	L	A	A
2dbm	V11	L130	A133	A137
1fua	V91	L164	A167	A171
1gdoA	V131	L141	A144	A148
1dciA	V105	L253	A256	A260
1iow	V143	L163	A166	A170
1qusA	V127	L351	A354	A358
4pgaA	V95	L140	A143	A147
1bw9A	V26	L49	A52	A56
1lam	V248	L339	A342	A346

Structure Patterns: Jonassen

Table 2. The matches to the pattern represented by the sequence motif C-x(4,19)-C-x(5,9)-C-x(4,17)-C

Protein	Pattern			
	C	C	C	C
1clvI	C508	C517	C523	C531
1fleI	C32	C38	C44	C53
1bx7	C6	C11	C17	C22
3ebx	C3	C17	C24	C41
1bteA	C11	C31	C41	C59
9wgaA	C3	C12	C18	C24
9wgaA	C46	C55	C61	C67
9wgaA	C89	C98	C104	C110
9wgaA	C132	C141	C147	C153

DNA Binding Motifs

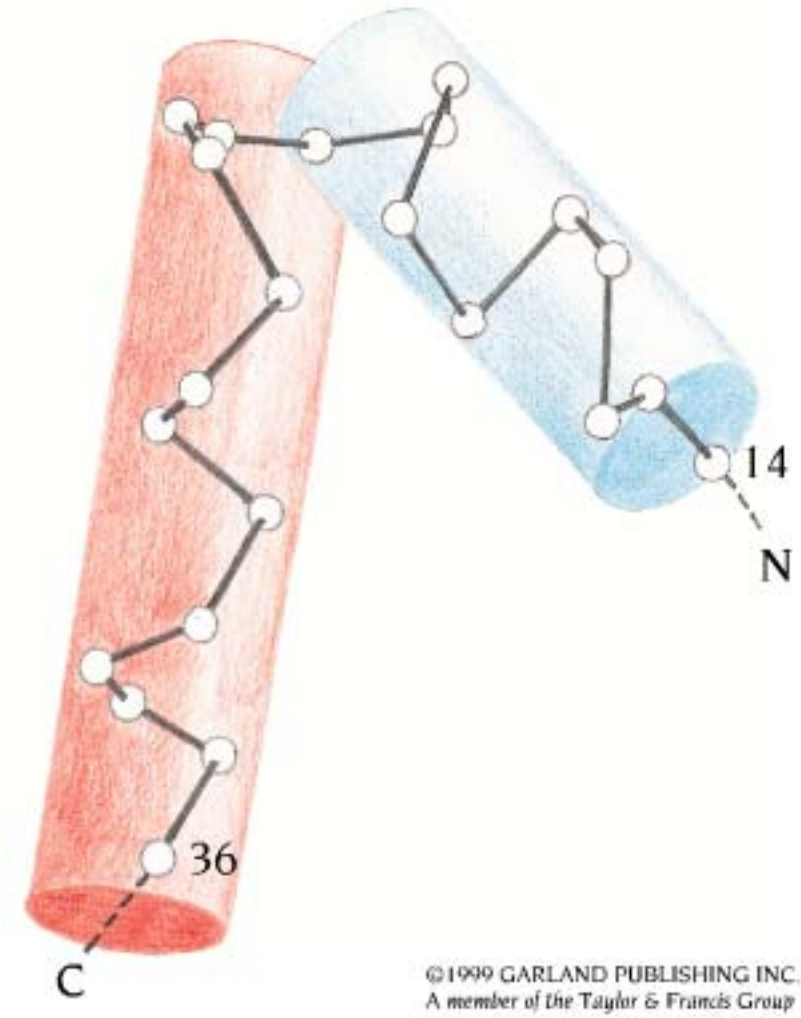
- Crucial feature is atomic contacts between protein residues and the bases and sugar-phosphate backbone of DNA
- Most contacts are in the major groove of DNA
- 80% of regulatory proteins can be assigned to one of three classes:
  - helix-turn-helix (HTH)
  - zinc finger
  - leucine zipper
- In addition to DNA-binding domains, these proteins usually possess domains that interact with other proteins

Alpha Helices and DNA

- Recurring feature of DNA-binding proteins: Presence of α-helical segment(s) that fit directly into the major groove of B-form DNA
- Diameter of helix is 1.2 nm (12 Angstroms)
- Major groove of DNA is about 1.2 nm wide and 0.6 to 0.8 nM deep
- Proteins can recognize specific sites in DNA

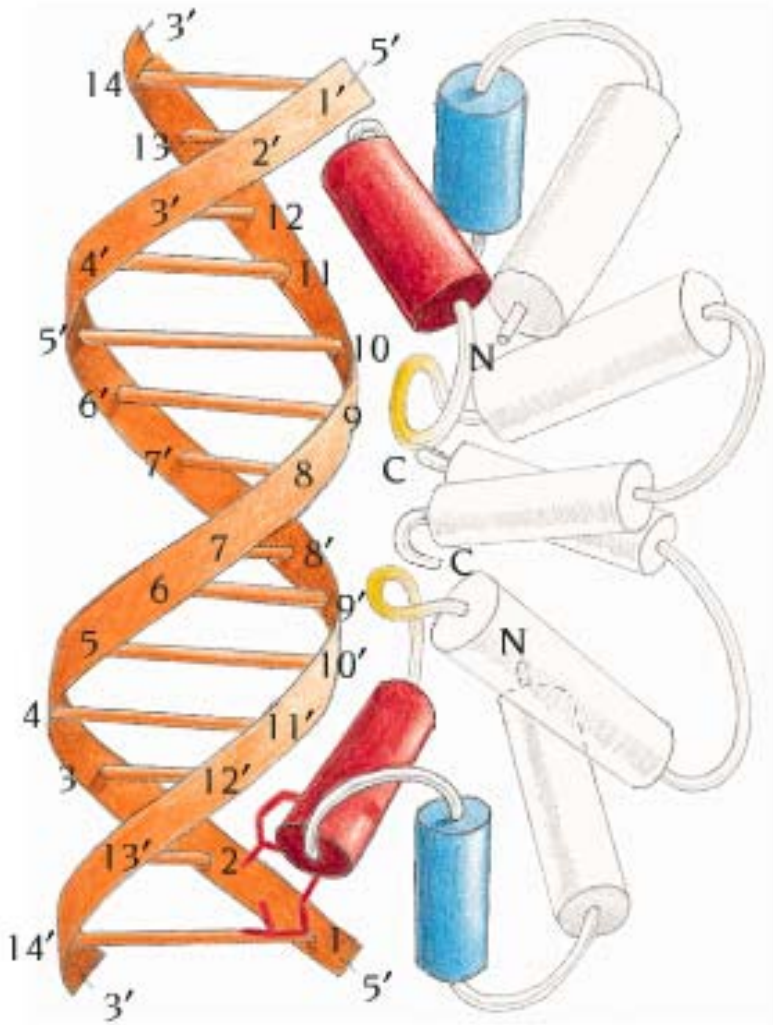
The Helix-Turn-Helix Motif

- Generally bind as dimers to dyad-symmetric sites on DNA
- All contain two alpha helices separated by a loop with a beta turn
- The C-terminal helix fits in major groove of DNA
- N-terminal helix stabilized by hydrophobic interactions with C-terminal helix

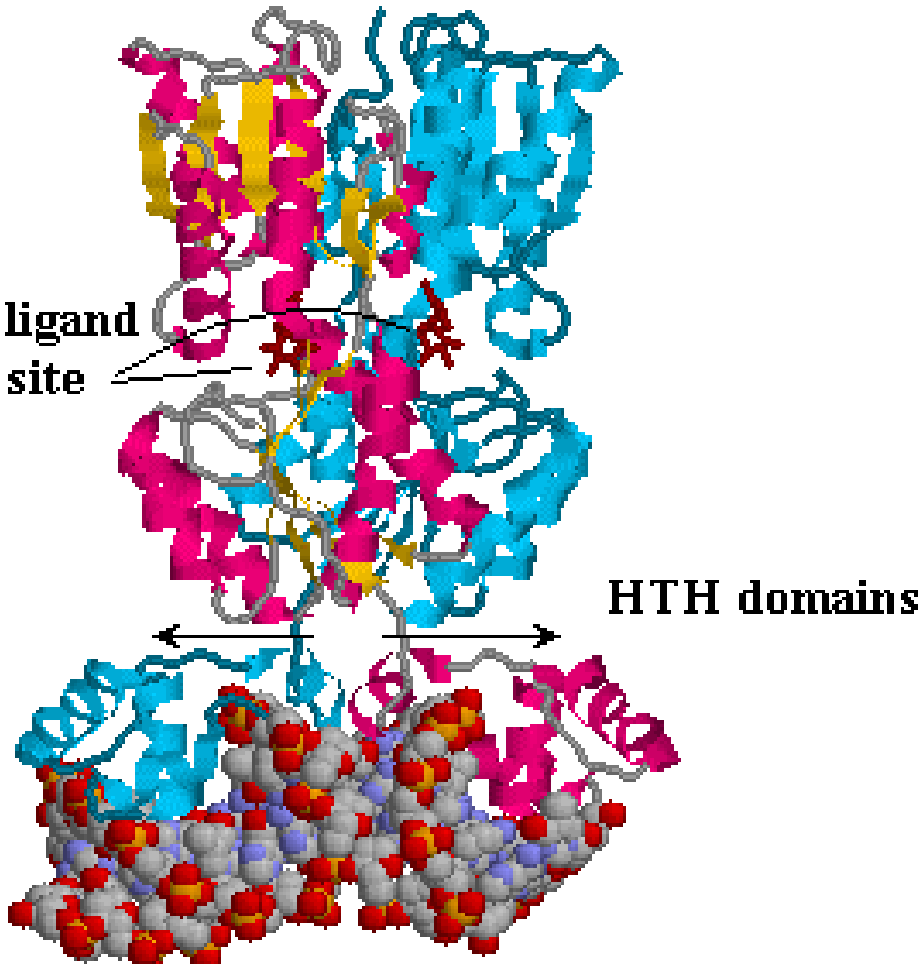


The Dimeric Protein & DNA

- Dimer symmetry requires palindromic DNA sequence
- A Glycine between the helices makes the turn
- Recognition is in the major groove



Protein with HTH domains



MPAT (motif pattern) BioPerl script

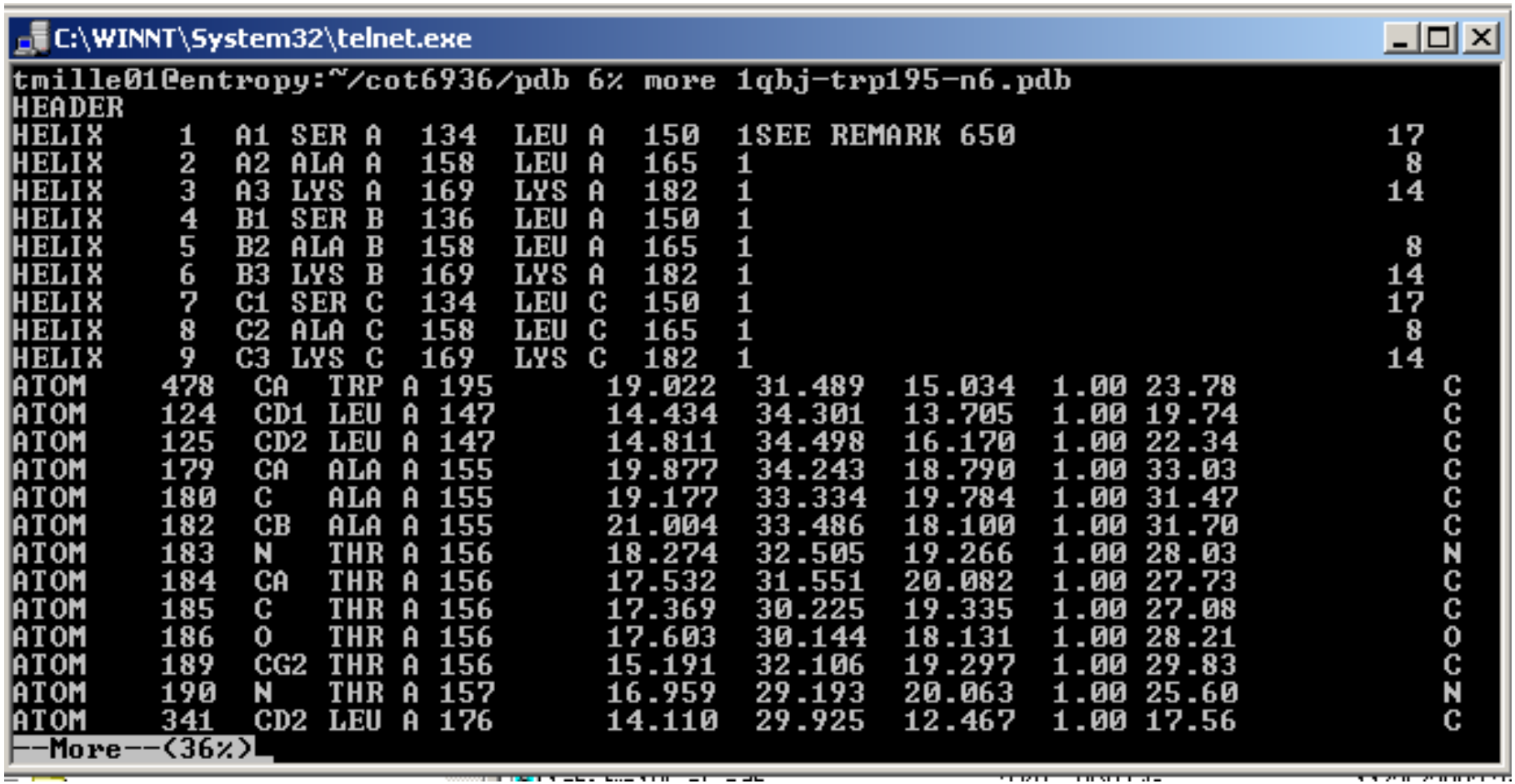
```
tmille01@entropy:~/cot6936/pdb 7% perl mpat
usage: perl mpat [-bhnprtv] [filename]

-b      base atom (default: CA)
-h      show water molecules
-n      radius (number) (default: 5)
-p      PDB file format output
-r      residue (e.g. GLY-28)
-t      target atom
-v      verbose output

tmille01@entropy:~/cot6936/pdb 8%
```

EX: >perl mpat –p –r TRP-195 –n 6 1QBJ.pdb > out.pdb

MPAT – PDB output



MPAT - Functions

1. **Process command line parameters**
2. **Print status information**
3. **Input PDB file into BioPerl structure**
4. **Process text lines of the PDB file for output in PDB format**
5. **Calculate the distance between atoms from the XYZ coordinates**
6. **Construct and output structure pattern**

Header and parameter processing

```
#!/usr/bin/perl -w
# Tom Milledge
# MPAT-BioPerl motif pattern discovery tool
use Bio::Structure::IO;
use Getopt::Std;
if(!@ARGV) {die "usage:…} …
getopts("b:hn:pr:t:v");
if ($opt_b) {
  $baseatom = $opt_b;
  print "Base atom: $opt_b ";} …
```

Inputting PDB file into BioPerl structure

```
while(<>) {}
my $pdb_file= $ARGV ;

my $structio = Bio::Structure::IO->new(
  -file => $pdb_file,  -format => 'pdb');

my $struc = $structio->next_structure;
```

Processing text lines of PDB file

```
open(INFILE, "< $pdb_file") or die "Can't
  open $pdb_file for reading: $!\n";
@lines = <INFILE>;
@pdbatoms = grep { /^ATOM / } @lines;
@ss = grep { /^HELIX|^TURN/ } @lines;
if($opt_h) {
  @h2os = grep { /^HETATM/ } @lines;
}
```

Calculate distance between atoms

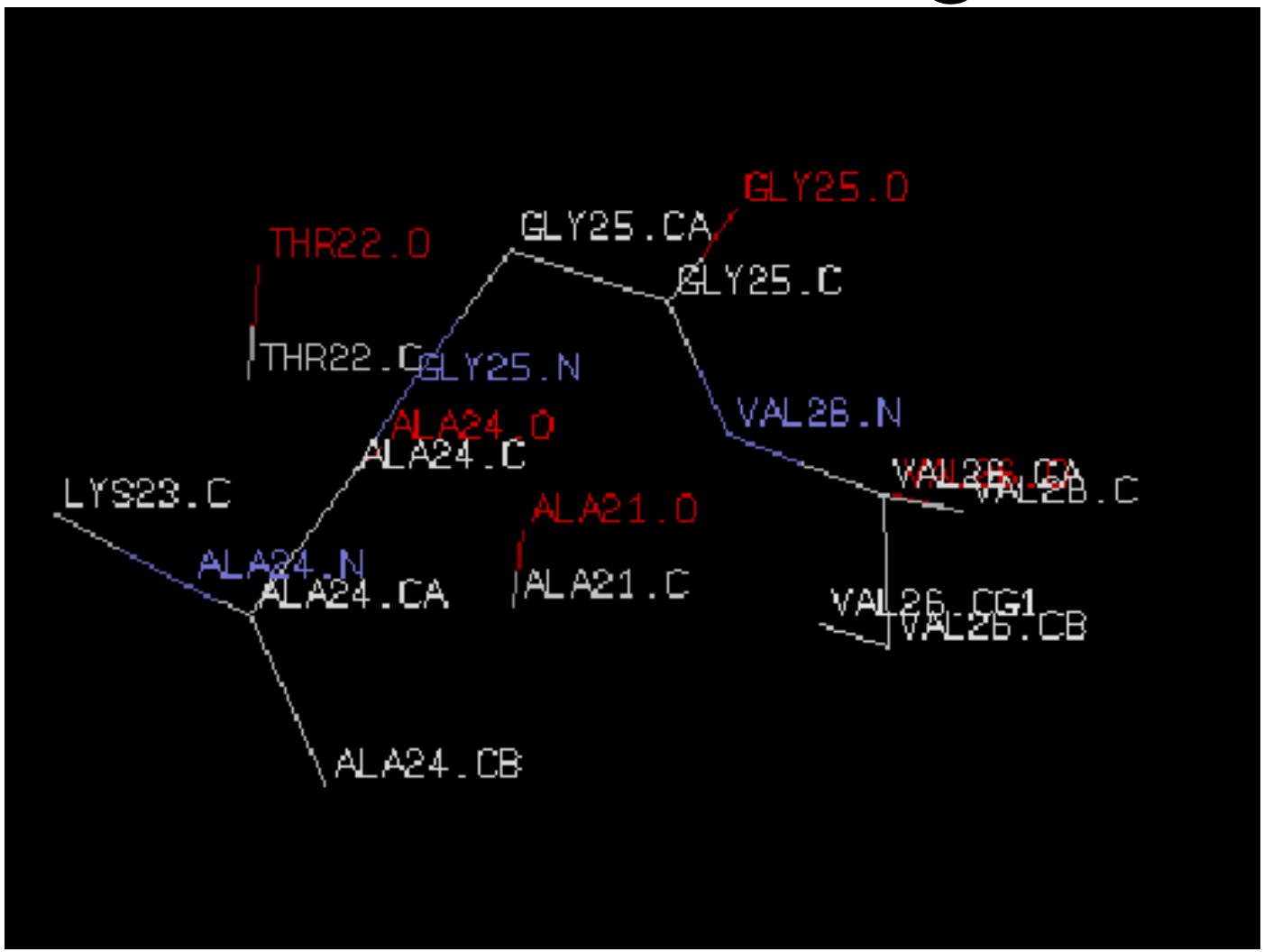
```
sub calculate_distance {
  my ($atom1, $atom2) = @_;
  my $dist;
  $dist = sqrt(($atom1->x - $atom2-
  >x)**2 +($atom1->y - $atom2->y)**2
  + ($atom1->z - $atom2->z)**2 );
}
```

Construct & output structure pattern

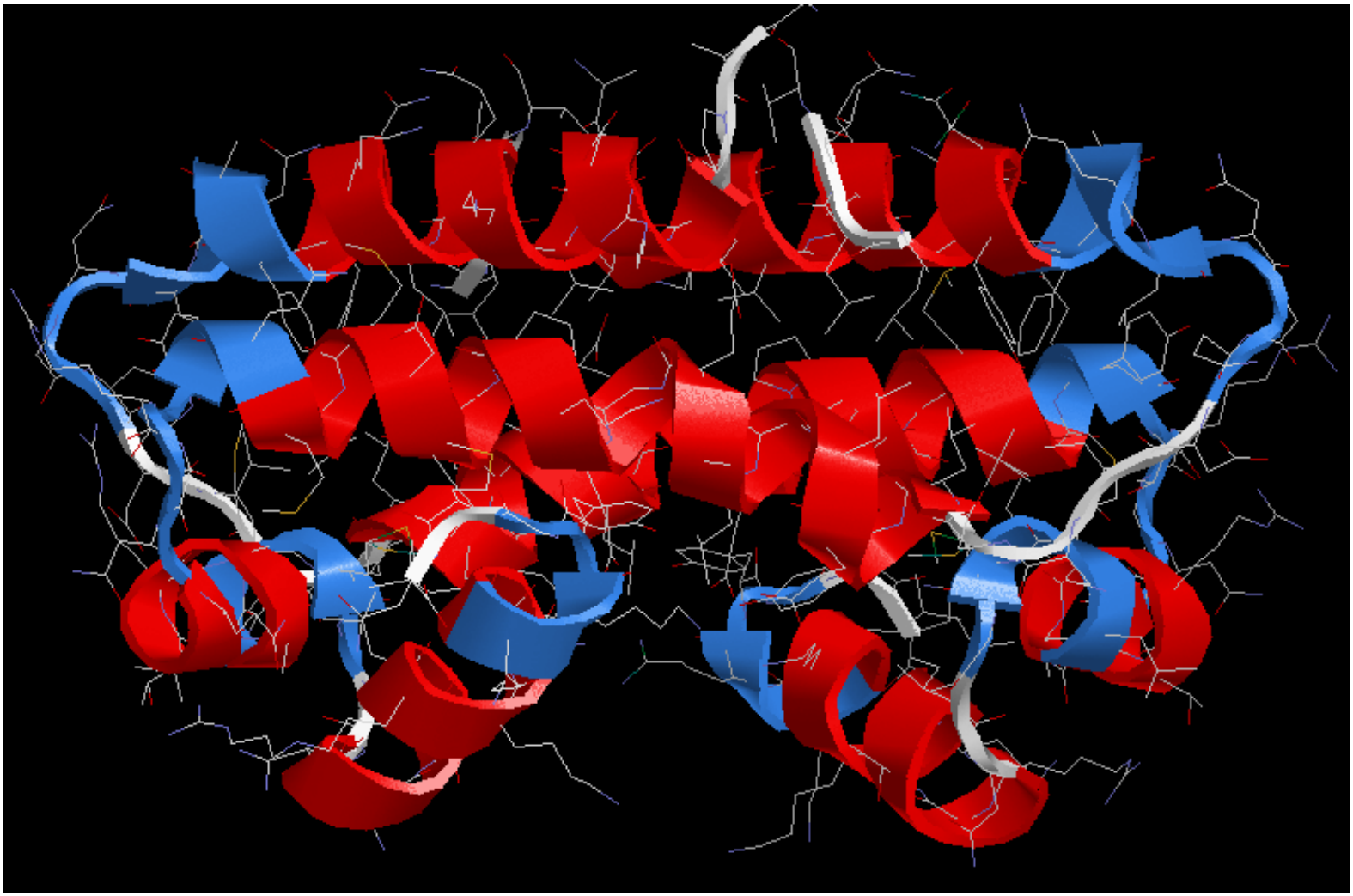
```
for my $k (0 .. $#baseatoms) {  …
  my $atom1 = $baseatoms[$k]; …
  for my $i (0 .. $#atoms) {
      my $atom2 = $atoms[$i];
      my $dist =
  calculate_distance($atom1,$atom2);
      if($dist <= $radius && $dist != 0) {
          if($opt_p) {
            print $pdbatoms[($atom2->serial)-1];
      } …
}
```

2CRO: HTH motif

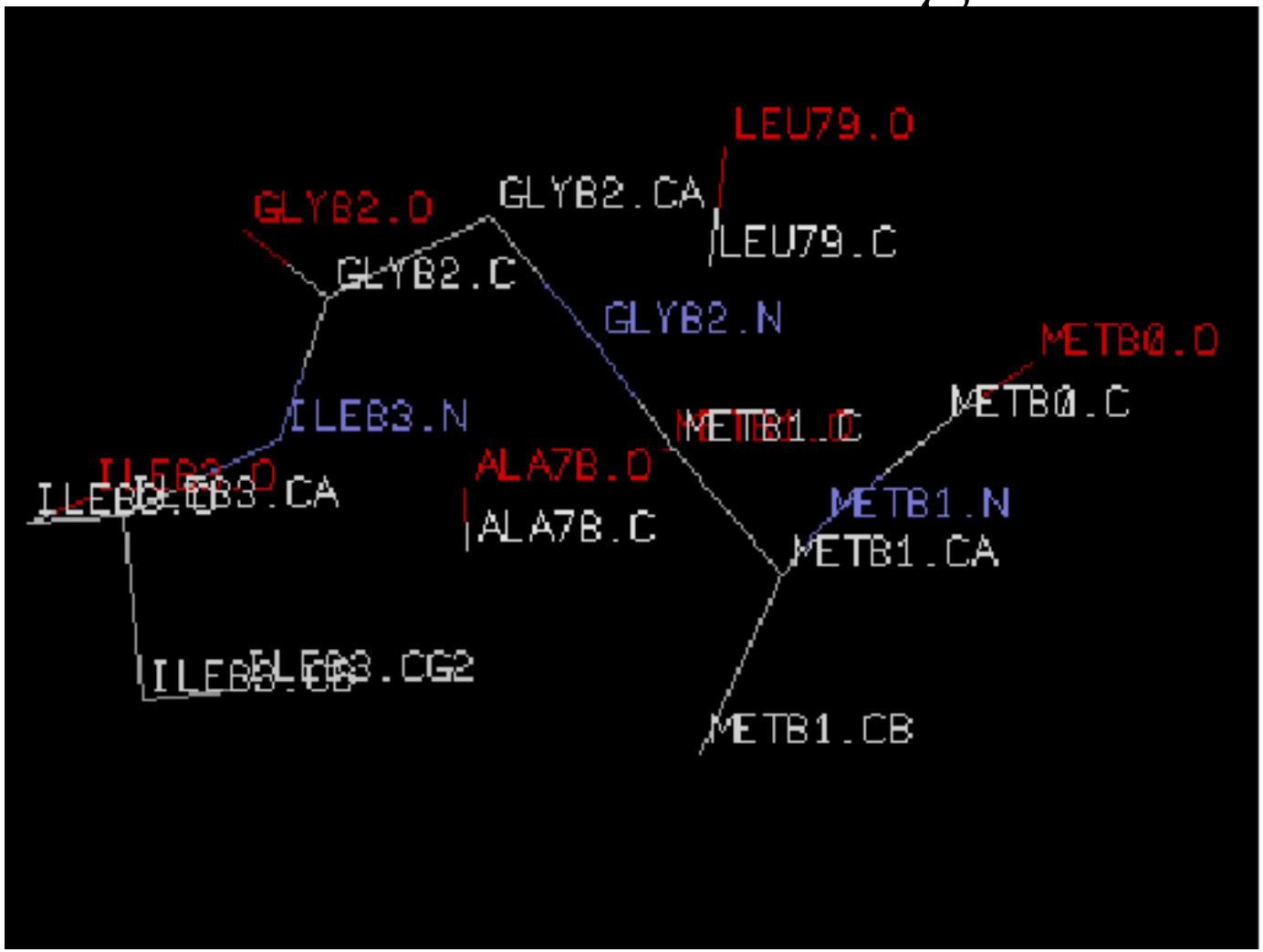
2CRO: HTH turn region



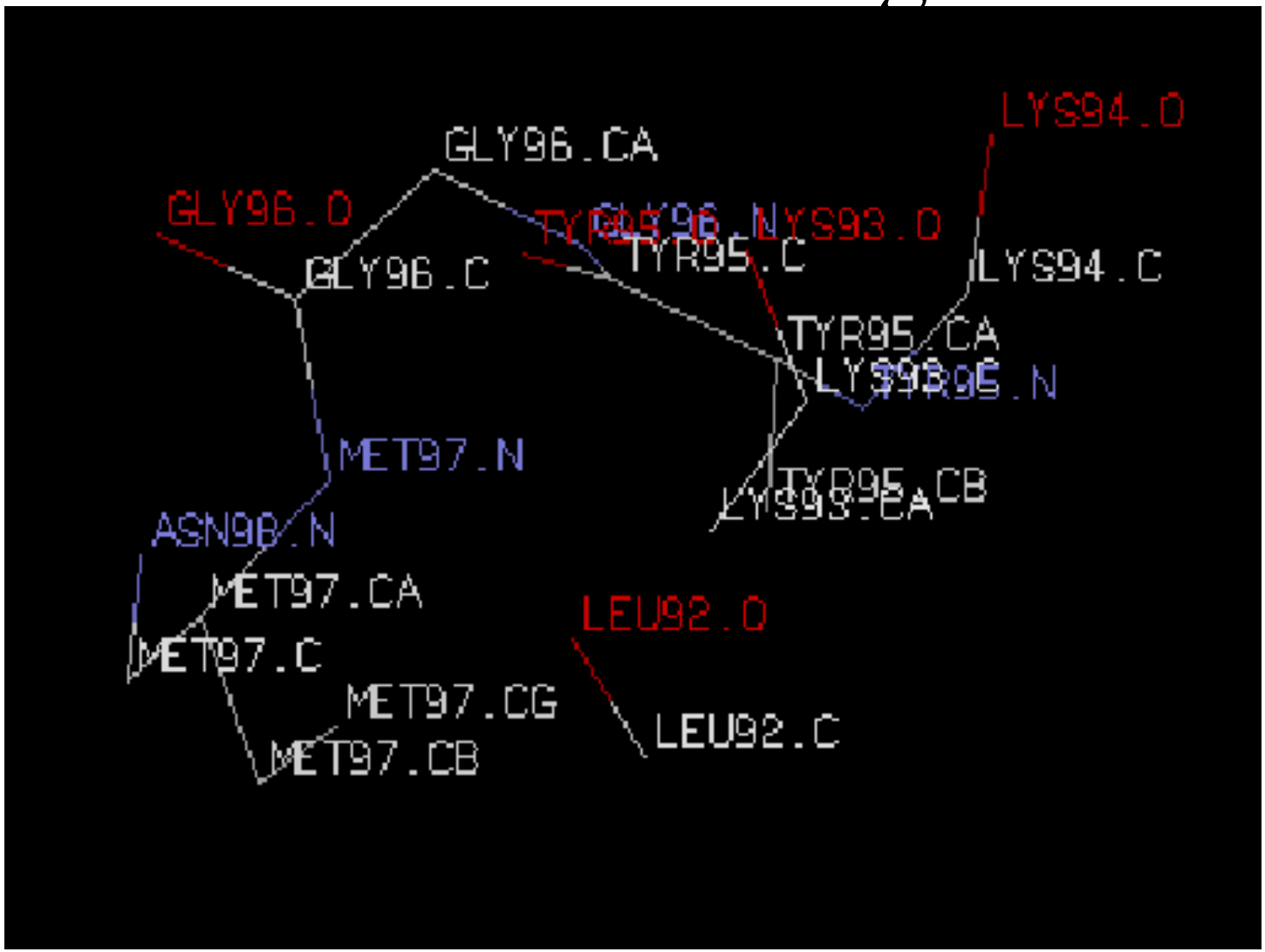
1FIA: 2 HTH motifs



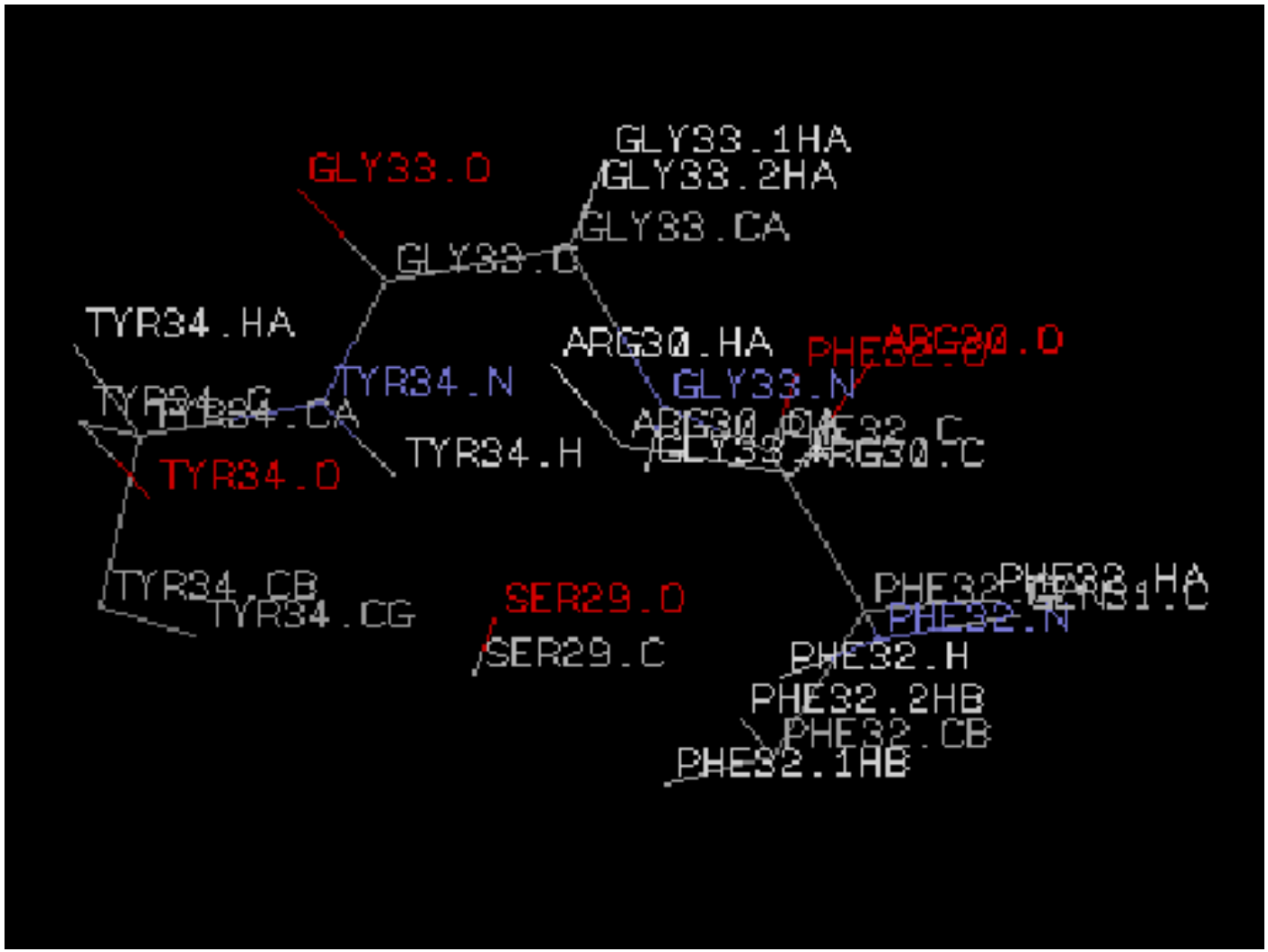
1FIA: HTH turn region 1



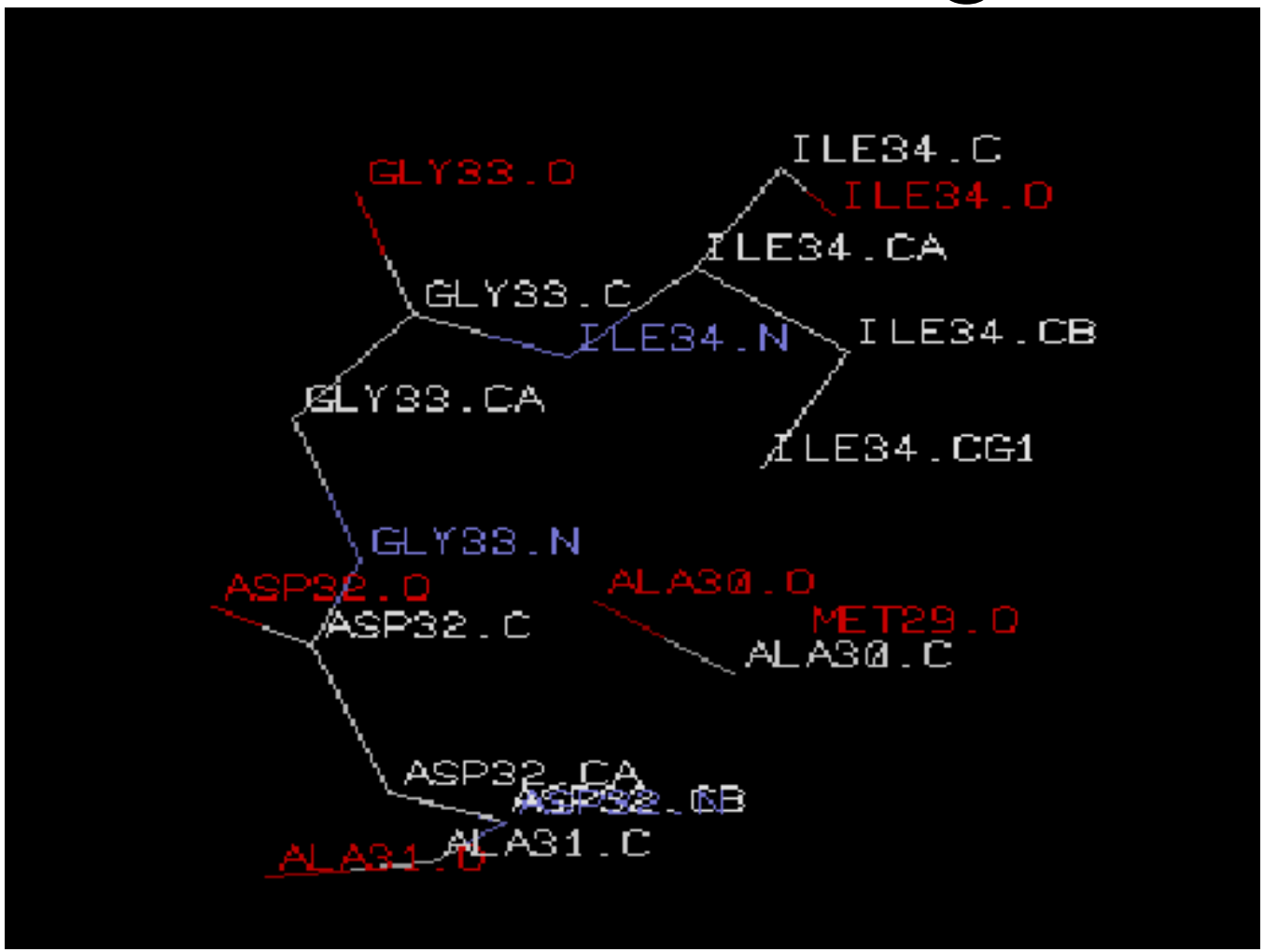
1FIA: HTH turn region 2



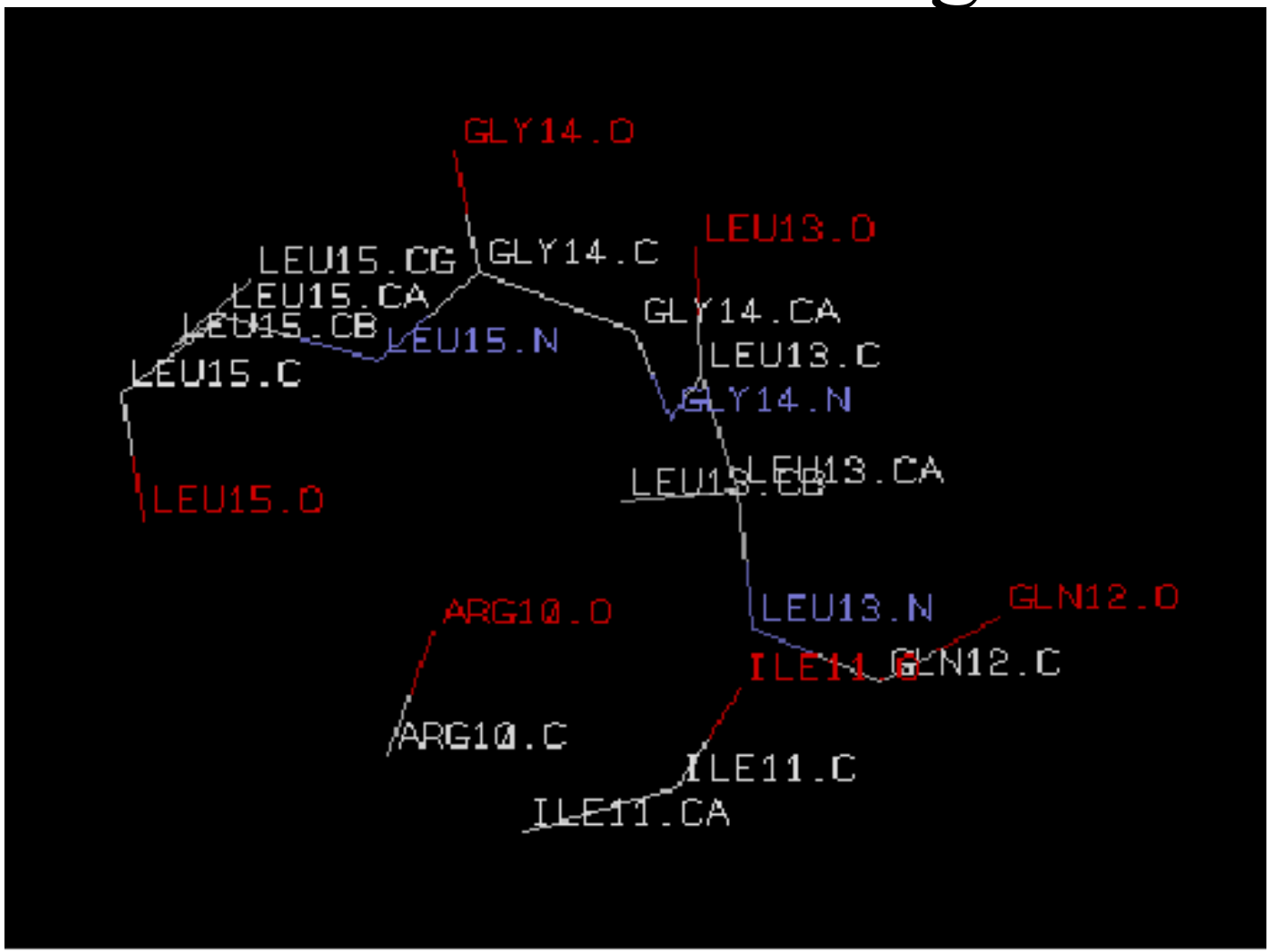
1NER: HTH turn region



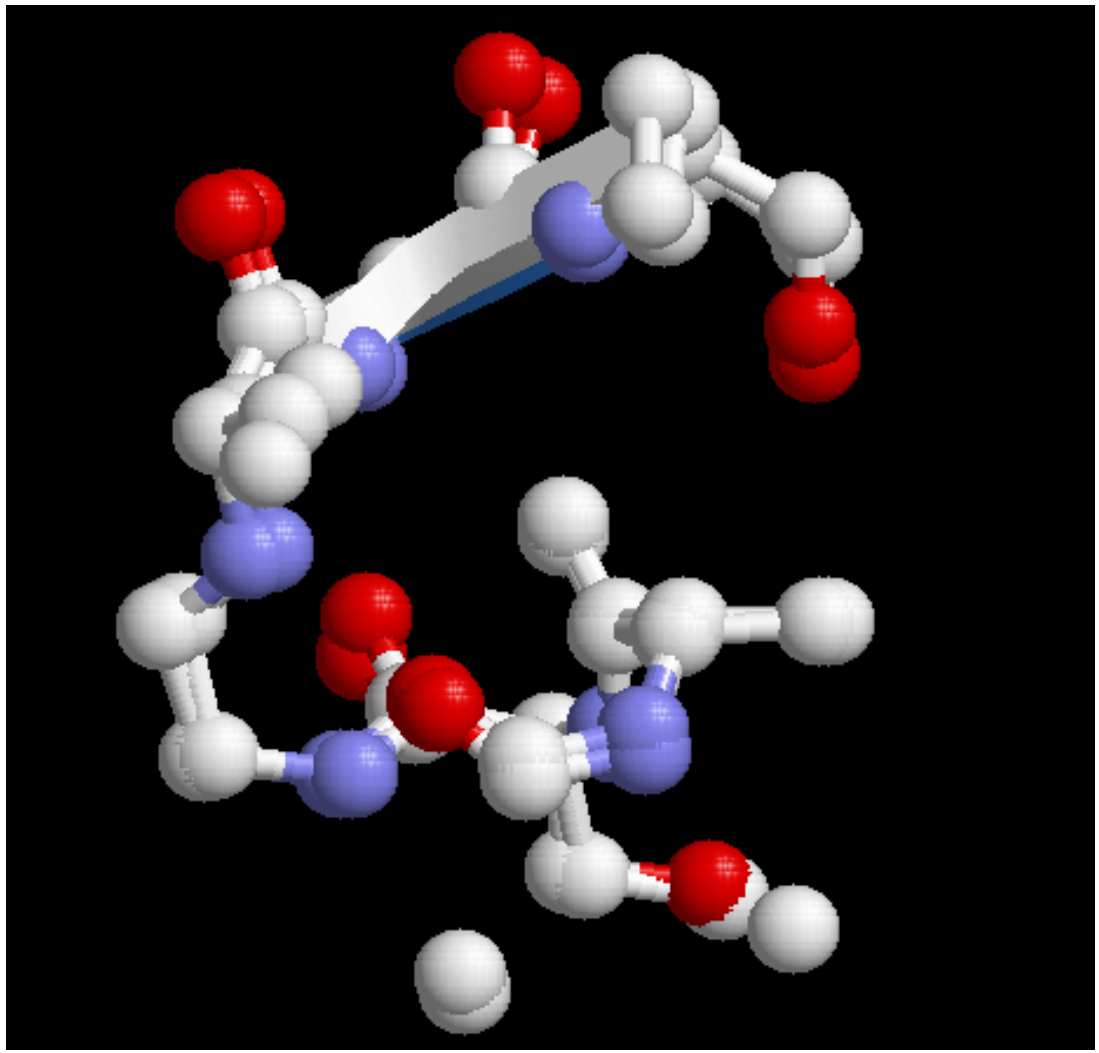
1PDN: HTH turn region



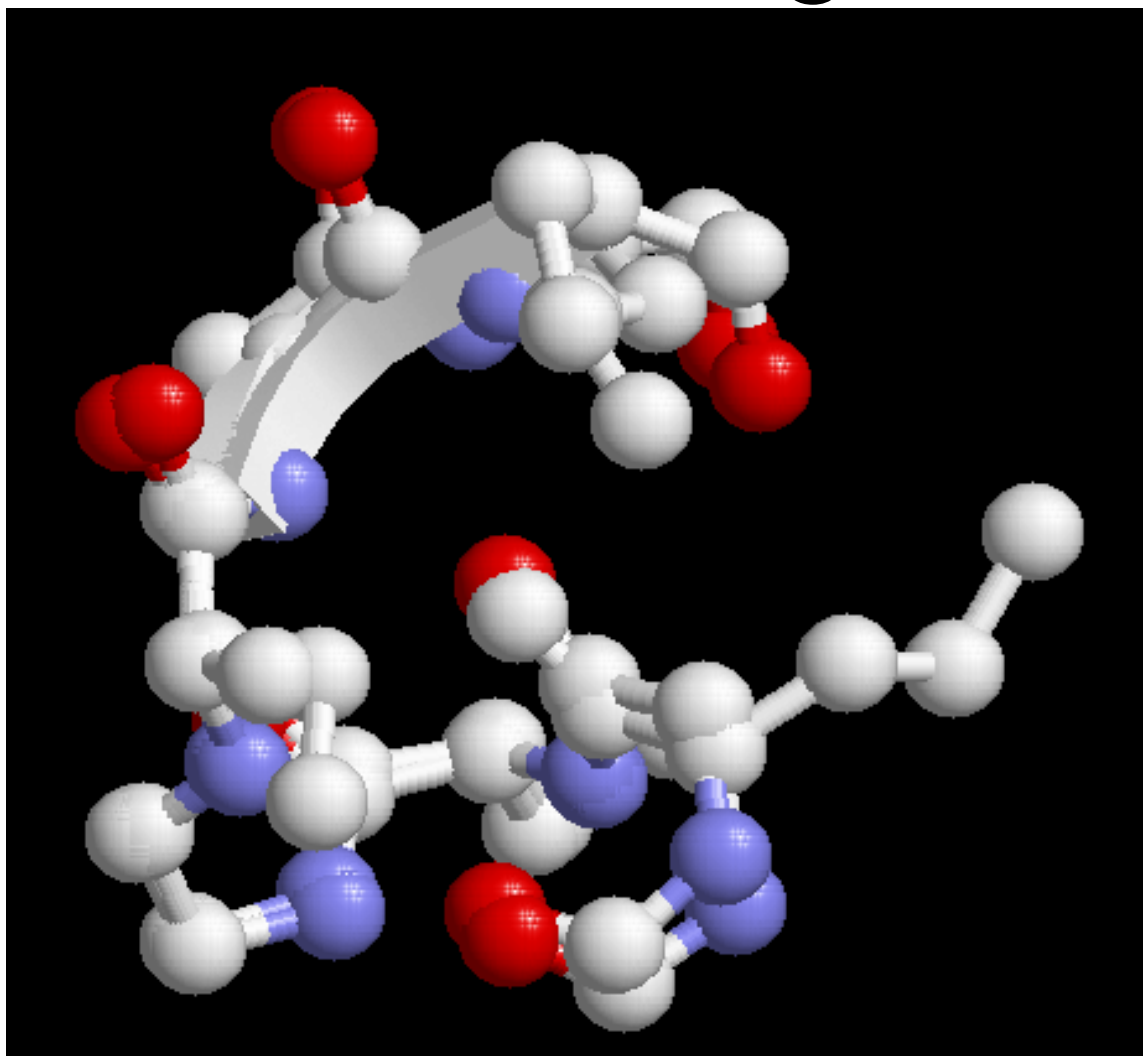
1R69: HTH turn region



2CRO/1FIA: Alignment



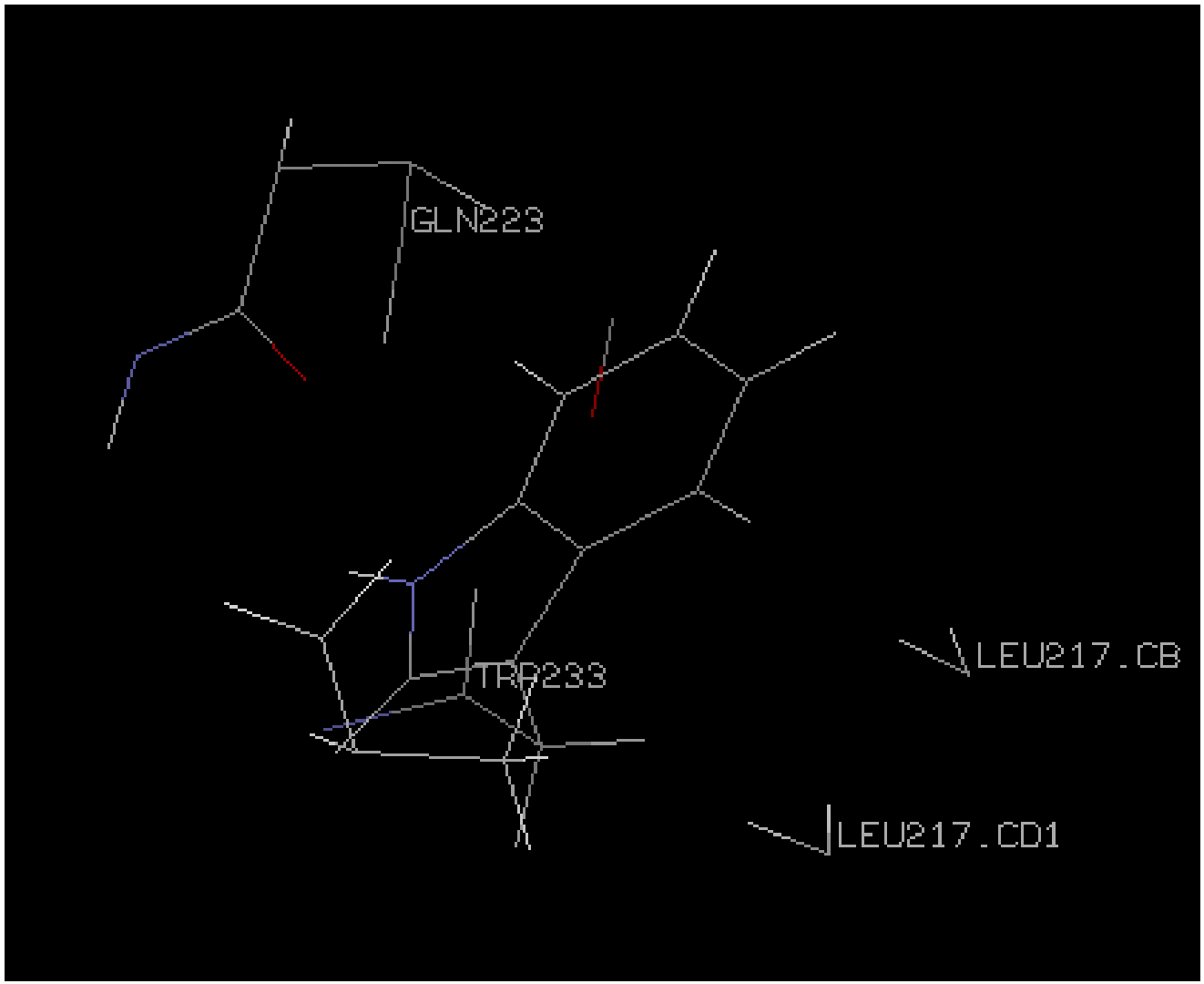
1BIA/1R69: Alignment



1BBY: Winged HTH motif

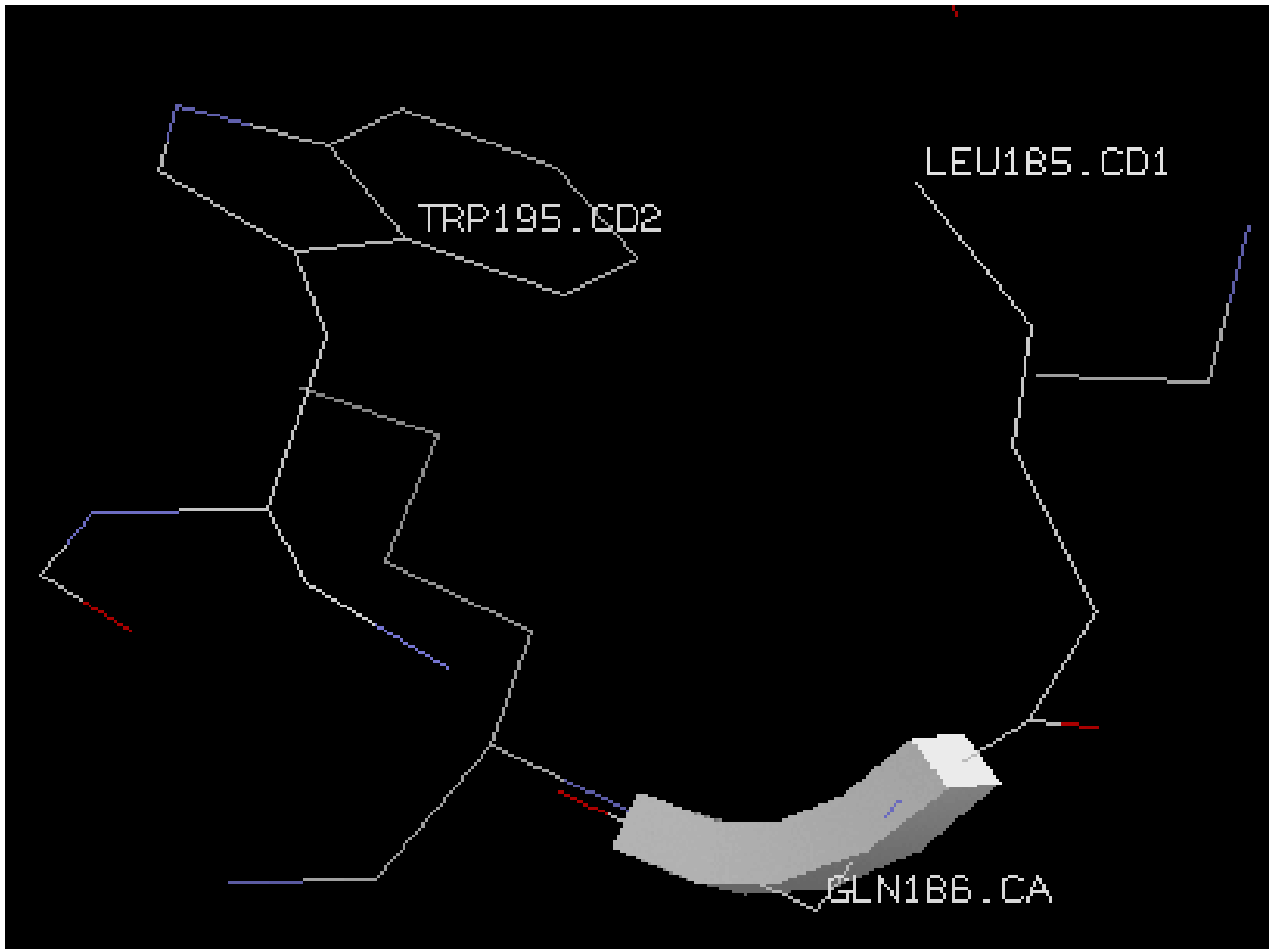
1BBY: close-up of wing region

1BBY: HTH wing region



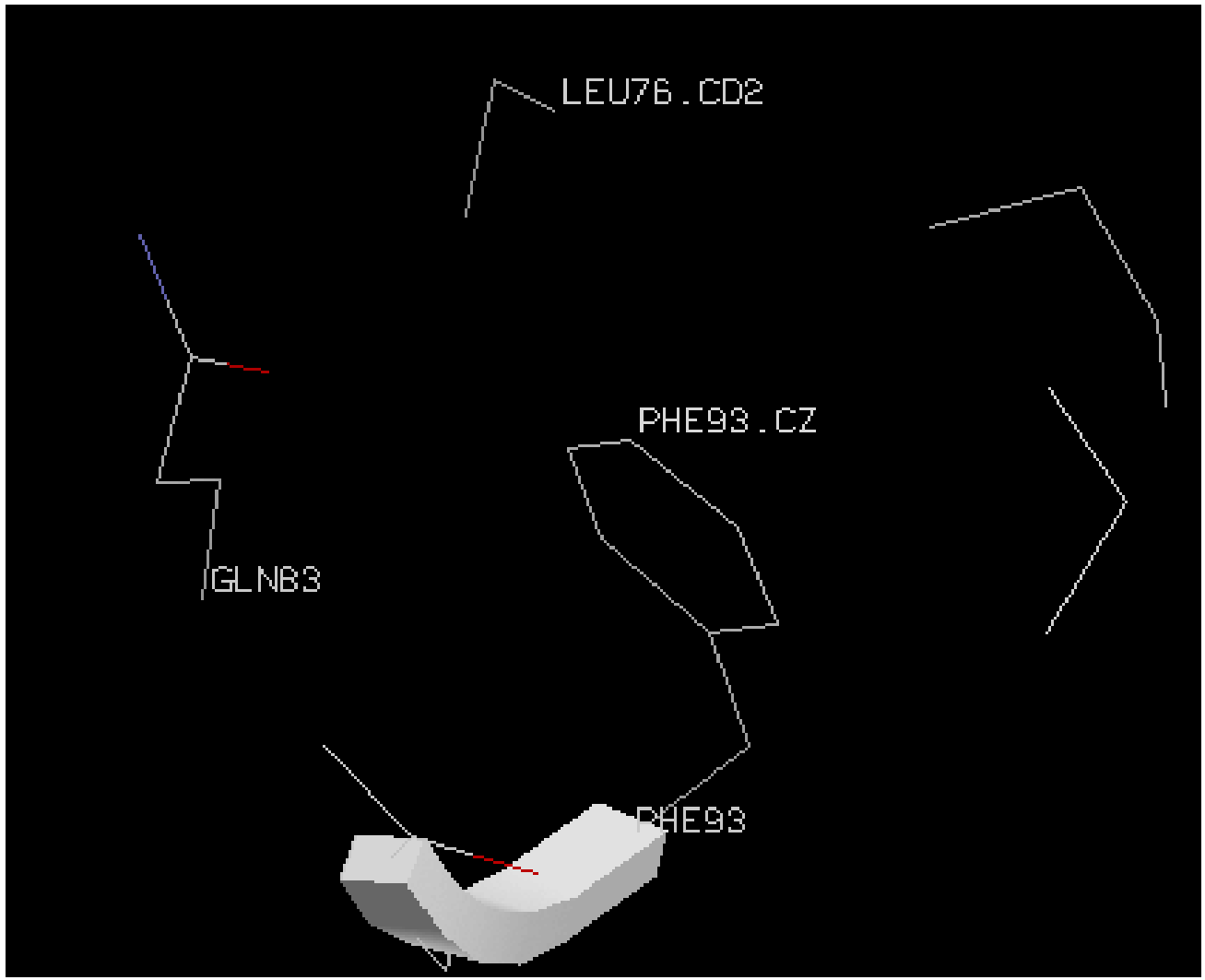
1QBJ: 3 Winged HTH motifs

1QBJ: HTH wing region



1BM9: HTH wing region

1HST: HTH wing region

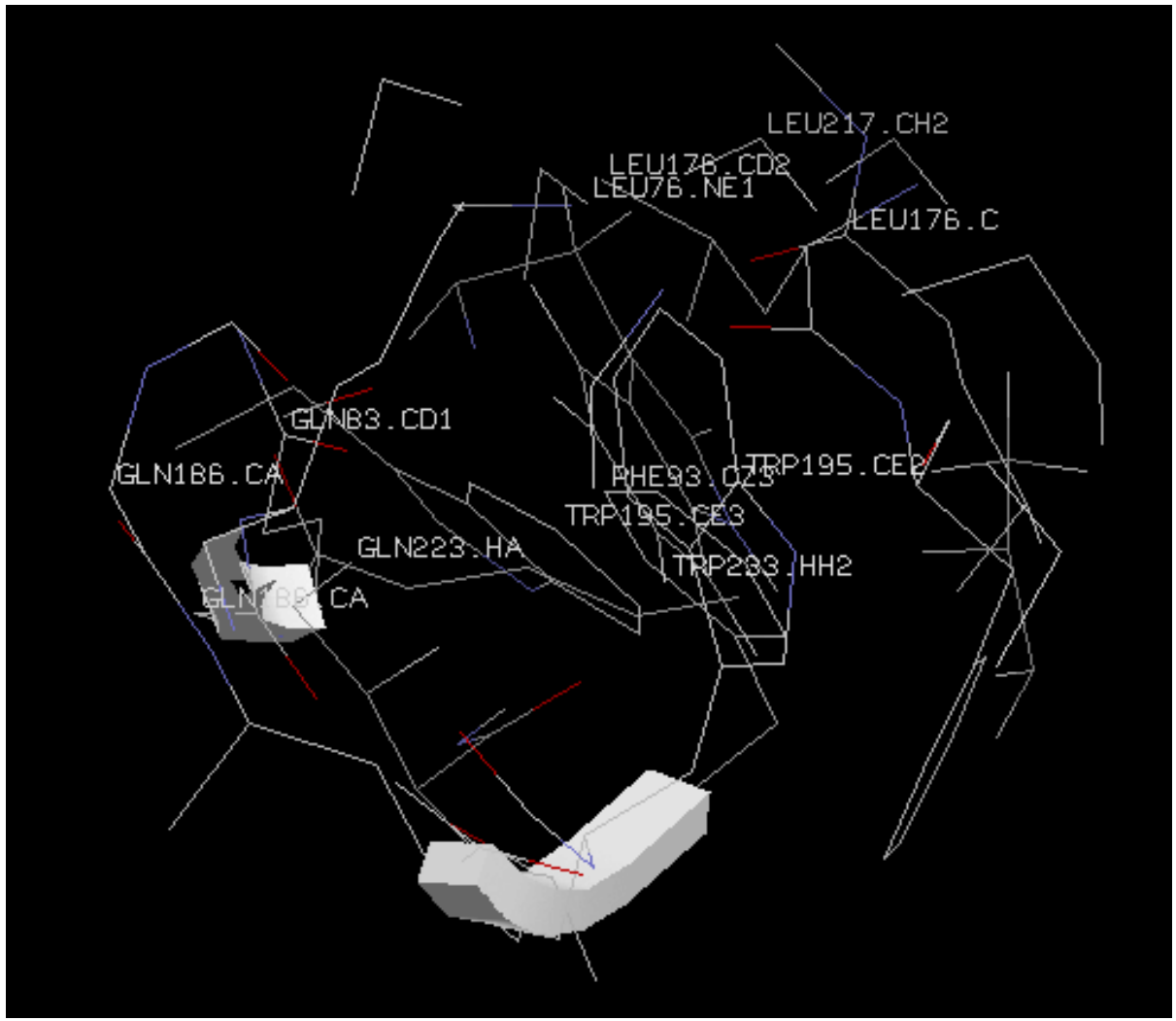


Winged HTH Structure Pattern

Protein	**Pattern** Leucine L	Glutamine Q	(Aromatic) (FYW)
1qbj	L176	Q186	W195
1hst	L76	Q83	F93
1bby	L217	Q223	W233
1bm9	L61	Q72	Y88

Expression: L-x(6,11)-Q-x(9,16)-[FYW]

Wing HTH Pattern: Multiple Alignment



Winged HTH region: Multiple Alignment