# Assignment #1

- Given a set of protein sequences with ubiquitination sites:
- 1. **Data Preparation**: Please extract sequence fragments with window length **2n+1** for positive and negative datasets.
- 2. **Data Preprocessing**: please utilize CD-HIT program to remove homologous sequences in positive and negative datasets using various sequence identity threshold (100% to 50%).

Example table:

| Sequence identity | Training set (6,266 proteins) | |
|---|---|---|
| | Positive data | Negative data |
| 100% (original) | 23,827 | 228,645 |
| 90% | 21,650 | 197,050 |
| 80% | 21,169 | 179,692 |
| 70% | 20,709 | 165,560 |
| 60% | 18,588 | 115,296 |
| 50% | 10,216 | 34,428 |
| 40% | 2,658 | 5,532 |
| **30%** | **2,566** | **5,176** |