Similarity Search for an Extreme Application: Experience & Implementation

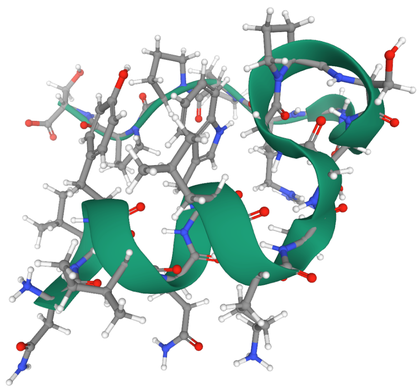
Vladimir Mic, Tomáš Račeк, Aleš Křenek, and Pavel Zezula
Masaryk University, Brno, Czech Republic

Protein Chains & Challenges

Topic – Similarity Search in Protein Chains

- Each protein consists of one or more parts – *protein chains*
- Modelled by balls and sticks

  → describe positions of atoms in 3D space and bonds between them



- Some of the atoms determine structure of the whole protein chain

  → carbon alpha atoms – *CA Atoms*

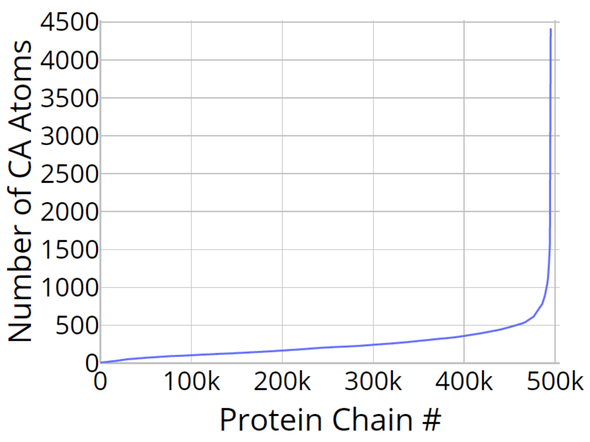
Similarity of Protein Chains – Almost Metric Space

- Similarity = Q-score of protein chains summarises matching CA atoms and their actual distance in 3D Euclidean space
- Extreme complexity → heuristics to estimate Q-score are used
- Widely used heuristic violates the Q-score symmetry
  - and thus triangle inequality
- Violation are rather small, but should be considered

  → *pivot permutation-based* techniques are robust enough

Size of Protein Chains

- Even heuristics suffer from the size of big protein chains:



Number of Dists.	Dist. Time
98 % ≈ 500,000	≤ 0.03 s
0.1 % = 561	≥ 1 min
303	≥ 10 min
10	≥ 42 min
1	43.6 min

- Extreme tail of biggest pr. chains

  → causes extreme tail of distance computation times

- Newly discovered chains are rather big
- We should efficiently search for similar chains to big query chains

The Most Efficient Search

Efficient similarity search must minimise the number of Q-score computations

... especially in case of big query chains

Our Search Engine

Data Preprocessing

- Search in *"Protein Data Bank in Europe"*
  - 495,085 protein chains, tens of thousand added every year
- Store distances to 512 pivots
- ... use them to create *sketches*:
  - bit-strings in the Hamming space, approximate space of chains
  - *Short* and *long sketches* (320, and 1024 bits)
  - Short sketches are defined by just 61 pivots
  - Long sketches are defined by 489 pivots, out of 512 pre-selected
- Mapping of protein chain distances to Hamming distances of long sketches
- build advanced *pivot permutation-based index*: *PPP-codes*

Gradual Search with Intermediate Results

- 3 search phases, assume query chain q:

  → search for k most similar chains up to distance r:

  1st phase evaluates 61 distances of q to pivots
  - creates a short sketch of q
  - evaluates Ham. dists. on short sketches to return k IDs of protein chains with the most similar short sketches
  - GUI asynchronously evaluates corresponding k Q-scores, and 2nd phase starts

  2nd phase evaluates other 428 distances of q to pivots
  - creates a long sketch of q
  - transforms search radius r to the Hamming radius r_{Ham}
  - evaluates Ham. dists. on long sketches to return k IDs of protein chains with sketches up to distance r_{Ham}
  - GUI asynchronously evaluates Q-scores, and 3rd phase starts

  3rd phase evaluates remaining 23 distances of q to pivots
  - uses PPP-codes to generate a stream of 5,000 most promising protein chains for q
  - uses long sketches to check each of these 5,000 chains
    - compares the Hamming distance of its long sketch and the sketch of q with r_{Ham}
  - evaluates Q-score of each non-filtered protein chain
  - on median, 4,810 / 5,000 chains are filtered out by sketches

- **i.e., we need just 702 dist. comps. to evaluate the query, on median**

Results – Medians over 1,000 query chains

	Q-score comps.	Search Accuracy	Search Time
Current engine	???	100 %	183 s
1st phase	61	46.7 %	0.18 s
2nd phase	429	66.7 %	1.1 s
3rd phase	702	100 %	2.5 s

- Most difficult q is evaluated in 4 hours by the current engine
- Our phases evaluate it in 4 s, 13 min, and 102 min

https://similar-pdb.cerit-sc.cz