Motif Finding

Transcription factor

```
1. ttgccacaaaataatccgccttcgcaaattgaccTACCTCAATAGCGGTAgaaaaacgcaccactgcctgacag
2. gtaagtacctgaaagttacggtctgcgaacgctattccacTGCTCCTTTATAGGTAcaacagtatagtctgatgga
3. ccacacggcaaataaggagTAACTCTTTCCGGGTAtgggtatacttcagccaatagccgagaatactgccattccag
4. ccatacccggaaagagttactccttatttgccgtgtggttagtcgcttTACATCGGTAAGGGTAgggattttacagca
5. aaactattaagatttttatgcagatgggtattaaggaGTATTCCCCATGGGTAacatattaatggctctta
6. ttacagtctgttatgtggtggctgttaaTTATCCTAAAGGGGTAtcttaggaatttactt
```

Given p sequences, find the most mutually similar length-k subsequences, one from each sequence:

$$\underset{s_1,\ldots,s_p}{\operatorname{argmin}} \sum_{i<j} \operatorname{dist}(s_i, s_j)$$

$\operatorname{dist}(s_i,s_j)$ = Hamming distance between s_i and s_j.

Hundreds of papers, many formulations (Tompa05)