

A New Approach for Multiple Sequence Alignment

Xu Zhang, Tamer Kahveci¹

Keywords: pattern discovery, multiple sequence alignment, clique

1 Algorithm

We introduce a new multiple sequence alignment method for protein sequences. We name our method *HSA* (Horizontal Sequence Alignment) for it horizontally slides a window on the protein sequences simultaneously.² *HSA* is superior to the existing methods that depend on the order of proteins since we consider all the proteins at once. Unlike most of the existing multiple alignment methods, *HSA* takes secondary structure information into account to find a biologically relevant alignment. *HSA* uses a scoring matrix, such as BLOSUM 62 to capture substitution probabilities of amino acids. *HSA* runs in four steps:

Step 1: (Initialization) We start by building a directed graph from the input proteins as follows. Each residue maps to a vertex in the graph. If it is available, Secondary Structure Element (SSE) type (α -helix, β -sheet) of each residue is also stored along with the vertex. A directed edge from vertex i to vertex j is added if residue j immediately follows residue i in the same sequence, or residues j and i have a substitution score higher than a given threshold. A weight is also assigned to each edge based on the substitution score and SSE type. If two residues belong to the same SSE type, then we assign a larger edge weight. All sequences are then scanned to find fragments with known SSE types. These fragments will guide the alignment later. The fragments are then clustered into groups, where each group consists of one fragment from every sequence, if they satisfy following four criterion: 1) They have same SSE type. 2) They have similar number of residues. 3) Their positions in the original sequence are close. 4) The substitution score for every fragment pair is greater than a given threshold.

Step 2: (Pre-alignment Adjustment) The graph constructed in step 1 is adjusted by inserting gap vertices as follows. The number of residues in fragments and the number of residues between consecutive fragments are calculated first. The count of gap vertices is then computed as a function of these two numbers. For each sequence, gap vertices are inserted to bring the fragments within the same group together. Gap vertices are positioned between consecutive fragments. This pre-alignment adjustment will move similar fragments vertically closer to each other. Thus, they will have higher probability to be aligned together in the next step.

Step 3: (Alignment) In this step, the sequences are actually aligned. We start by placing a window of length w at the beginning of each sequence. Typically we use $w = 4$ or 6 . This window defines a subgraph of the graph constructed in Step 2. Next, we greedily choose the clique with the best *expectation score* from this subgraph. We will explain the expectation score later. A clique here is defined as a complete subgraph of the graph with a constraint that it consists of one vertex from each sequence. In other words, if K sequences are to be aligned, a clique corresponds to the alignment of one letter from each of the K sequences. The score of a clique is defined as the SP (Sum-of-Pairs) score of the corresponding column. For each clique, we align the letters of that clique, and iteratively find the next best clique that 1) does not conflict with this clique, and 2) has at least one letter next to a letter in this clique. This iteration is repeated t times to find t columns. Typically, $t = 4$. These t cliques define a local alignment of the input sequences. The expectation score of the original clique is defined as the SP score of this local alignment. We then slide the window by one and repeat the same process until it reaches the end of sequences.

Step 4: (Post-alignment Adjustment) In this step, the alignment obtained by the previous step is adjusted by examining the gaps. After concatenating the columns, many short gaps may be scattered in the sequence. Thus rearranging gaps may be required to construct fewer but longer gaps. Sequences are scanned again to find

¹Department of Computer and Information Science Engineering, University of Florida, Gainesville, FL 32611 E-mail: {xuzhang, tamer}@cise.ufl.edu

²We use the terms horizontal and vertical to represent the direction that we move on a multiple alignment. Horizontal means to move on a given sequence from one residue to one of its neighbors. Vertical means to move from one residue of a protein to the residue of a different protein at the same alignment position.

```

lidy- MEVKKTSWTeedRILYQAHKRLg.nrwaEIAKLLPg.....rtdna.IKNHWNSTMRRKV
lhstA ...SHPTYSemi.AAAIRA EKSRggssrqSIQKYIKSHYKVGhnadlqIKLSIRRLLAAGV
lhc3C ...RGSALSdte.RAQLDVMKLLnvslh.EMSRKIS.....rsrhc.IRVYLKDPVSYGT
laoy- ..MRSSAKQeelvKAFKALLKEEkfssqgEIVAALQE QGFDNinqsk.VSRMLTKFGAVRT
ljhgA TPDEREALGtrv.RIIEELLRGE m.sqr.ELKNELG.....agiat.ITRGSNSLKAAPV

```

Figure 1: The alignment of the lidy benchmark from BALiBASE using our method. The capital letters show the positions in which the alignment of HSA concurs with BALiBASE.

isolated gaps. The isolated gaps will be moved to produce fewer number of longer gaps if the movement will produce higher local alignment score.

2 Results

In order to demonstrate the feasibility of our method, we ran it on a set of reference proteins with low similarity. The benchmarks are selected from the reference set 1 of short sequences with less than 25 % identity. We chose this benchmark since alignment of dissimilar proteins is usually harder than the alignment of highly similar proteins. Table 1 shows the BALiBASE scores of HSA and eight other existing well known multiple alignment tools for seven BALiBASE benchmarks [1] (<http://www-igbmc.u-strasbg.fr/BioInfo/BALiBASE/>). The results show that on the average HSA achieves a score of 0.799, which is better than any other tool. HSA finds the best result for 6 out of 7 reference benchmarks. In only one benchmark (1wit), PRRP and ClustalX produce a higher BALiBASE score than HSA. However, in this benchmark, the BALiBASE score of HSA is already very high (0.957) and the difference between PRRP (0.991) and HSA is only 0.034, whereas the margin between HSA and other tools including PRRP for the remaining experiments is much larger. For example, for benchmark lidy, HSA obtains a BALiBASE score of 0.869 where four of the eight competitors fail to find an alignment with score greater than 0.140. Figure 1 shows the alignment of the lidy benchmark using HSA. It can be seen from the figure that HSA aligns most of the residues to correct locations. The columns which are incorrectly aligned are usually very close to their positions in the reference alignment. (i.e., in most cases only one or two residues are shifted by a few residues.) Although these columns are incorrectly aligned, they still produce a high SP score.

Table 1: The BALiBASE score of HSA and eight other tools.

	PRRP	ClustalX	SAGA	Dialign	Multialign	Pileup8	Multal	HMMT	HSA
laboA	0.560	0.687	0.529	0.359	0.703	0.521	0.526	0.181	0.836
lidy	0.606	0.705	0.342	0.018	0.566	0.080	0.080	0.138	0.869
lr69	0.837	0.481	0.550	0.406	0.325	0.562	0.225	0.100	0.965
ltvxA	0.378	0.438	0.278	0.306	0.228	0.344	0.244	0.108	0.564
lubi	0.498	0.415	0.452	0.000	0.488	0.428	0.428	0.140	0.514
1wit	0.991	0.982	0.899	0.851	0.842	0.773	0.763	0.549	0.957
2trx	0.494	0.754	0.801	0.728	0.500	0.453	0.235	0.292	0.890
Avg	0.623	0.637	0.550	0.381	0.522	0.452	0.357	0.215	0.799

References

- [1] J.D. Thompson, F. Plewniak, and O. Poch. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research*, 27(13):2682–2690, 1999.