Global Sequence Alignment for Protein Sequences of Yeast and Escherichia Coli

Term Project on
Algorithms for Bio-informatics

BY

Arnab Sinha
02CS3022
Department of Computer Science and Engineering,
Indian Institute of Technology,
Kharagpur

Done under the guidance of

Dr. Jayanta Mukherjee
Department of Computer Science & Engineering,
Indian Institute of Technology,
Kharagpur

November 2006

1. Introduction

Hamming distance while important in computer science is not typically used to compare DNA or protein sequences. The hamming distance calculation rigidly assumes that i^{th} symbol of one sequence is already aligned against the i^{th} symbol of the other. However, it is often the case that i^{th} symbol in one sequence corresponds to a symbol at a different and unknown position in the other. Hence, the global sequence alignment attempts to address this deficit of the hamming distance. In this term project we have implemented the algorithm for the *Global Sequence Alignment Problem*. Our developed tool uses dynamic programming strategy to solve this problem. In the next section we give the detailed implementation of the tool.

2. GlobSeq: The Tool

GlobSeq uses dynamic programming algorithm[1]. The tool is open-source and developed in ANSI-C language. The code is targeted for Unix/Linux platform. The inputs and outputs to our tool are the following.

Input
1. *PAM250*: This matrix[3] helps us to compute a scoring matrix, which is iteratively used to construct less obvious alignments. The matrix is given in the appendix. The matrix is directly read from the database by the tool.
2. *Protein Sequences*: It takes two protein sequences for the sequencing problem.

Output
1. *Alignment Score*: The alignment score is reported after sequencing.
2. *Edit Graph*: The edit graph along with the path in the alignment grid is reported.
3. *Tabular Alignment*: The tool also displays the alignment in a tabular fashion.

The details regarding running the tool is as follows:
1. Untar the distribution:

```
[arnabs@Trinity arnabs]$ $tar –zxvf globseq.tgz
```

There are three directories in the folder. '`data`' '`doc`' and '`src`' for the protein sequences of yeast and Escherichia coli and the source C-code respectively. In the '`doc`' folder this documentation is present.

```
[arnabs@Trinity arnabs]$ ls globseq/
data  doc  src
[arnabs@Trinity arnabs]$ $ls globseq/data
e_coli  yeast
```

2. Compile:

```
[arnabs@Trinity arnabs]$ cd globseq/src
[arnabs@Trinity src]$ make
```

3. Run: The tool can be run in multiple modes. Either one can input the protein sequence files or the sequences directly. For feeding the files following is an example.

```
$ ./globseq –sf1 ../data/yeast/1S1IC –sf2 ../data/e_coli/1GLAF
```

If one wishes to run feeding the sequences directly, following is an example.

```
$ ./globseq –s1 'AAHST' –s2 'AHHAST'
```

4. Display help: For displaying the help

```
$ ./globseq –h
```

5. Optional features: To display the alignment grid and the path try the following.

```
$ ./globseq –s1 'AAHST' –s2 'AHHAST' –mat –tab
```

These options enable the display of the alignment grid and the tabular form representation.

6. Demo: For displaying a demo,

```
$ ./runglobseq
```

This demo will display the scores between the all the protein sequences of yeast and E. coli present in the database.

3. The Results

The database contains 10 protein sequences[2] from yeast and Escherichia coli. In Table 1, the PAM250 alignment scores are displayed. The rows represent the protein sequences of yeast and the columns represent those of E. coli. The *runtimes* (in ms) are also represented along with the score. The cells marked red denote high scores while green cells denote poor scores.

	1GLAF		1GLAG		1GLBF		1GLBG		1TLFA		1TLFC		1TLFD		2ASR		AAX07738		AAX07744	
	168		501		168		501		301		301		301		142		379		272	
	sc	ms	sc	ms	sc	ms	sc	ms	sc	ms	sc	ms	sc	ms	sc	ms	sc	ms	sc	ms
1S1I_9 91	222	39	374	120	222	45	374	119	279	75	279	71	279	88	205	40	329	89	298	70
1S1IC 386	512	181	964	499	512	172	964	488	686	295	686	308	686	305	452	146	848	378	698	284
1S1I_W 112	290	65	447	145	290	52	447	144	357	88	357	97	357	90	249	44	399	110	355	83
1S1I_X 120	286	55	451	159	286	55	451	153	359	93	359	93	359	95	283	50	416	117	363	88
1S1I_Y 87	208	41	358	114	208	42	358	113	271	72	271	70	271	73	205	36	329	86	288	69
1S1I_Z 105	257	49	411	138	257	49	411	137	321	83	321	83	321	85	229	47	376	106	333	81
1Z7Qb 233	411	104	722	303	411	106	722	296	553	179	553	179	553	184	371	92	674	227	543	168
1Z7QT 234	424	104	748	298	424	107	748	296	551	181	551	178	551	189	404	89	655	228	580	173
1Z7QU 288	463	132	790	372	463	129	790	364	623	218	623	219	623	227	390	109	705	279	596	210
1Z7QZ 212	390	95	748	276	390	96	748	270	499	163	499	163	499	172	377	81	643	208	547	154

Table 1: Alignment Scores between protein sequences of yeast and Escherichia Coli.

Analysis of a particular pair of sequences

We find that, there is a biological similarity between the two protein sequence pairs, namely (1S1IC, 1GLAG) and (1S1IC, 1GLBG). In this report, we just study the functional and the structural similarity between the first pair of protein sequences. Functionally they are much alike.

Functional Similarity	
1S1IC (Yeast)	**1GLAG (E. Coli)**
By definition this is Chain C, Structure Of The Ribosomal 80s-Eef2-Sordarin Complex From Yeast Obtained By Docking Atomic Models For Rna And Protein Components Into A 11.7 A Cryo-Em Map This is mainly a part of the Ribosome complex and hence its primary role is to *generate energy*.	Chain G, Glycerol Kinase E.C.2.7.1.30) Complex With Glycerol And The (Escherichia Coli) Glucose-Specific Factor Iii (Iii-Glc). This is a Glycerol Kinase and hence its primary role is to *produce and conserve energy*.

In the following table we have shown some of the structural similarities of the protein pair - (1S1IC (from yeast), 1GLAG (from E. coli))

Structural Similarity	
1S1IC (Yeast)	**1GLAG (E. Coli)**
FEATURES Location/Qualifiers source 1..386 /organism="Saccharomyces cerevisiae" 44..51 /sec_str_type="sheet" /note="*strand 24*" 52..59 /sec_str_type="sheet" /note="*strand 25*" 68..73 /sec_str_type="sheet" /note="*strand 26*" 74..80 /sec_str_type="sheet" /note="*strand 27*" 81..87 /sec_str_type="sheet" /note="*strand 28*" 88..92 /sec_str_type="sheet" /note="*strand 29*" 99..106	FEATURES Location/Qualifiers source 1..501 /organism="Escherichia coli" 99..103 /sec_str_type="sheet" /note="*strand 24*" 159..164 /sec_str_type="sheet" /note="*strand 25*" 179..183 /sec_str_type="sheet" /note="*strand 26*" 192..195 /sec_str_type="sheet" /note="*strand 27*" 196..199 /sec_str_type="sheet" /note="*strand 28*" 215..219 /sec_str_type="sheet" /note="*strand 29*" 222..227

/sec_str_type="sheet" /note="*strand 30*"	/sec_str_type="sheet" /note="*strand 30*"
139..149 /sec_str_type="helix" /note="*helix 8*"	200..208 /sec_str_type="helix" /note="*helix 8*"
157..163 /sec_str_type="sheet" /note="*strand 31*"	239..244 /sec_str_type="sheet" /note="*strand 31*"
177..182 /sec_str_type="sheet" /note="*strand 32*"	259..266 /sec_str_type="sheet" /note="*strand 32*"
188..197 /sec_str_type="helix" /note="*helix 9*"	245..253 /sec_str_type="helix" /note="*helix 9*"
201..204 /sec_str_type="sheet" /note="*strand 33*"	267..276 /sec_str_type="sheet" /note="*strand 33*"

4. Conclusion

In this term project, we have implemented a standard algorithm for Global Sequence Alignment. Moreover, we have displayed the efficiency of the tool designed by displaying the various runtimes taken for the computational purpose. In the end, we have analyzed a particular pair of protein sequences, whose PAM250 score was relatively higher compared to the other comparisons made. In that analysis, we observed that indeed those protein sequences had functional as well as structural similarities.

Reference:

[1] An Introduction to Bioinformatics algorithms: Jones and Perzner, Ane Books (2005)
[2] NCBI website: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=protein
[3] PAM250: http://bioinformatics.bc.edu/clotelab/cgi-bin/BoltzmannAlignment/pam250.txt

Appendix:

PAM250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4