Protein Structure Prediction Based on Secondary Structure Alignment *

Rei-Sing Cheng, Chang-Biau Yang and Kuo-Tsung Tseng

Department of Computer Science and Engineering
National Sun Yat-sen University, Kaohsiung, Taiwan
cbyang@cse.nsysu.edu.tw

Abstract

In molecular biology, sequence alignment is a fundamental but powerful technique. Biologists find the similarity, the difference and even the function of the input sequences (DNA, RNA and protein sequences) by it. With various purposes, there are many algorithms to align two sequences based on different criteria.

Though there are various ways to align macromolecular sequences, a sequence alignment algorithm traditionally considers only the primary structure, which is the amino acid chain. In this paper, we present a new algorithm which aligns sequences with consideration of their secondary structures. When we make use of the information of protein secondary structure such as α helix, β sheet etc., the sensitivity of pairwise alignment can be improved.

1 Introduction

Sequence alignment is a fundamental technique in biological sequence analysis. The identity, similarity and even homology of biological sequences such as DNA, RNA and protein sequences all can be analyzed. In addition, pairwise sequence alignment is the very first step toward multiple sequence alignment [4–7, 13, 18–20]. Due to its importance, many methods have been proposed [8–12, 15, 17].

The biological sequence can be represented as a chain of characters in a alphabet set Σ. In DNA, Σ={A, G, C, T}, in RNA, Σ={A, U, C, T} and in protein, Σ={ A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}. The sequence alignment problem is to find the alignment which has the best score with some scoring criteria.

A simple example of pairwise sequence alignment is given as follows. Suppose S_1 and S_2 are two amino acid segments. In alignment 1, the number of identity in each vertical position is 6, but it becomes 9 in alignment 2. We can easily point out that the second alignment is better than the first alignment since it has 3 more identities. Dynamic programming has been used to find the optimal alignment. For different purposes, dynamic programming can be designed to solve such problems with different scoring functions.

S_1=EEHGWAGAEHG
S_2=EAHGWAGEHG

Alignment 1

```
EEHGWAGAEHG
| |||||
EAHGWAGEHG
```

Alignment 2

```
EEHGWAGAEHG
| ||||| |||
EAHGWAG-EHG
```

Many algorithms have been proposed to find the similarity and difference among sequences [1, 15–17] and some of them are used to search databases [2, 3]. In this paper, we shall propose an alignment algorithm based on not only primary structures but also secondary structures. We hope that our work can find more information that can not be found in traditional alignment algorithms.

The rest of this paper is organized as follows. Some preliminaries are given in Section 2. In Section 3, we illustrate our sequence alignment algorithm which takes protein secondary structure in consideration and a real case is shown for clarity. Finally, we give our conclusion in Section 4.

*This research work was partially supported by the National Science Council of the Republic of China under contract NSC-92-2213-E-110-005.

Table 1: Twenty natural amino acids found in biological systems.

	Name	Three-latter code	One-letter code
1	Alanine	Ala	A
2	Cysteine	Cys	C
3	Aspartic Acid	Asp	D
4	Glutamic Acid	Glu	E
5	Phenylalanine	Phe	F
6	Glycine	Gly	G
7	Histidine	His	H
8	Isoleucine	IlE	I
9	Lysine	Lys	K
10	Leucine	Leu	L
11	Methionine	Met	M
12	Asparagine	Asn	N
13	Proline	Pro	P
14	Glutamine	Gln	Q
15	Arginine	Arg	R
16	Serine	Ser	S
17	Threonine	Thr	T
18	Valine	Val	V
19	Tryptophan	Trp	W
20	Tyrosine	Tyr	Y

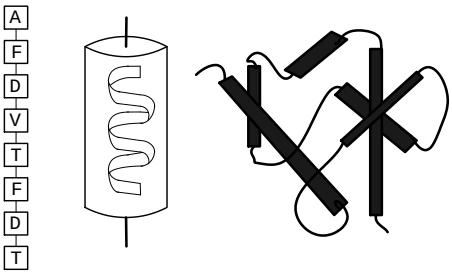

Figure 1: Primary, secondary and tertiary structures of a protein

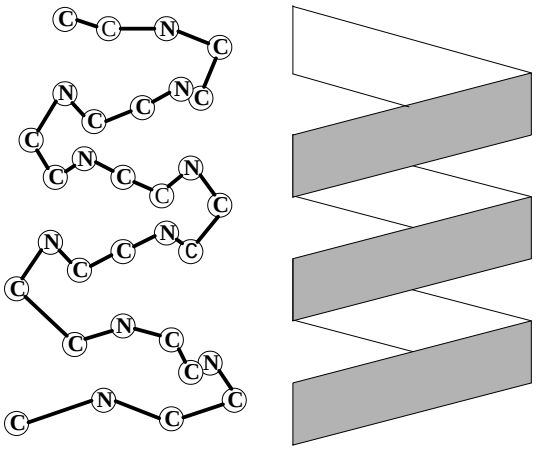

Figure 2: α-helix.

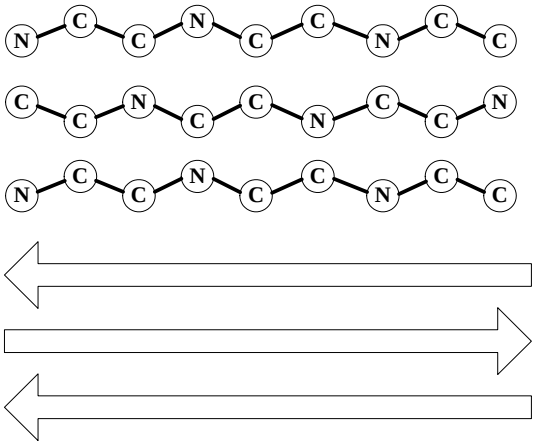

Figure 3: β-sheet.

2 Preliminaries

Proteins consist of 20 amino acids. Table 1 shows these amino acids. We can say that amino acids are the basic materials to build proteins. By different permutation of amino acids, proteins can fold into different forms of structures. There is a concept: the function of protein is determined by the its structure. In other words, protein function is determined by the permutation of amino acids.

For each protein, it has primary, secondary and tertiary structures. The primary structure is just the linear sequence of amino acid residues. The secondary structure is formed because of interactions between local atoms and results in local structures such as α-*helix* and β-*sheet*. Finally, the tertiary structure is formed due to the secondary structure interaction. Figure 1 illustrates these structures.

Before getting into our algorithm, it is better to introduce some basic knowledge about the protein secondary structure. The secondary structure of a protein is the conformation of its backbone. It can roughly be divided into three classes: α-*helix*, β-*sheet* and *loop*. α helix comes from hydrogen bond interactions between two nearby turns. Figure 2 shows a typical conformation of α-helix.

A β-*strand* is a chain of amino acid residues with a special property. That is, the sidechains of nearby residues are placed in opposite direction. Furthermore, the torsion angle of N-C-C-N is about 120 degrees. A β-*sheet* is another conformation due to a different kind of hydrogen bond interaction. A β-sheet may consist of two or more β-strands and any neighboring pair of β-strands are combined by hydrogen bonds. If two neighboring strands have the same direction, we say that they are *parallel*. Otherwise, we say that they are *anti-parallel*. Figure 3 shows an ideal conformation of β-sheet.

Table 2: Abbreviations of secondary structures.

Abbreviation	Representation
H	helix
B	residue in isolated beta bridge
E	extended beta strand
G	310 helix
I	pi helix
T	hydrogen bonded turn
S	bend

The secondary structure expression we use here is according to an implementation of the method proposed by Kabsch and Sander [14]. Table 2 shows the abbreviation of each secondary structure.

3 A Novel Sequence Alignment Algorithm with Protein Secondary Structure in Consideration

In this section, we shall propose a new protein alignment algorithm, which considers the combined information of primary structures and secondary structures.

Given two protein sequences A and B, $A = a_1a_2\cdots a_n$, $B = b_1b_2\cdots b_m$, and secondary structure $S(A) = S_{a1}S_{a2}\cdots S_{an}$, $S(B) = S_{b1}S_{b2}\cdots S_{bm}$, we apply dynamic programming to perform our algorithm. Let $C(i, j)$ denote the score of optimal alignment of a_i and b_j where $1 \leq i \leq n$ and $1 \leq j \leq m$.

There are four probabilities:

1. $a_i = b_j$ and $S_{ai} = S_{bj}$
2. $a_i \neq b_j$ and $S_{ai} = S_{bj}$
3. $a_i = b_j$ and $S_{ai} \neq S_{bj}$
4. $a_i \neq b_j$ and $S_{ai} \neq S_{bj}$

From the four cases, we choose the best one, which is the alignment with the highest score. In summary, we use the following formula to calculate A(i, j).

$$C(i,j) = max \begin{cases} C(i-1,j-1) + \sigma \cdot p(a_i,b_j) + (1-\sigma) \cdot q(a_i,b_j) \\ C(i-1,j) - gp \\ C(i,j-1) - gp \end{cases}$$

Here $q(a_i,b_j)$ denote the score of alignment of the secondary structures of a_i and b_j. σ and 1-σ are weights on primary score and secondary score respectively. $0 \leq \sigma \leq 1$. It is clear that if we set σ to 1, it becomes the global alignment algorithm. Similarly, if we set σ to 0, it becomes the alignment between two

Table 3: The score matrix of secondary structure.

		-	H	T	S	E	G	B	I
	1								
-	0	1							
H	1	0	3						
T	1	0	0	3					
S	1	0	0	2	3				
E	1	0	0	0	0	3			
G	1	0	2	0	0	0	3		
B	1	0	0	2	2	0	0	3	
I	1	0	2	0	0	0	2	0	3

secondary structures. The scoring matrix of secondary structure is illustrated in Table 3.

Here we give an example to illustrate our algorithm. Suppose we are given two protein sequences A and B, whose primary structure and secondary structure are as follows:

A:
primary structure : SMTDLLSAED
secondary structure : HHTTTTTHHH

B:
primary structure : ADQLTEEQIA
secondary structure : HH***HHHHH

We first set σ to 1, the result is simply a Needleman-Wunsch global sequence alignment, which is shown as follows:

```
--SMTDLLSAED
ADQLTE--EQIA
```

The secondary structure of this alignment is

```
--HHTTTTTHHH
HHH***--HHHH
```

The matrix involved in the dynamic programming of this alignment is shown in Table 4

If we consider the secondary structures of each protein, the alignment will look like(σ=0.5):

```
SMTDLLSAED
ADQLTEEQIA
```

The secondary structure of this alignment is

```
HHTTTTTHHH
HH***HHHHH
```

Table 4: The dynamic programming matrix of global alignment with PAM-80 scoring matrix.

			0	1	2	3	4	5	6	7	8	9	10
						T	T	T	T	T	H	H	H
			-	S	M	T	D	L	L	S	A	E	D
0		-	1.0	-13.0	-15.0	-17.0	-19.0	-21.0	-23.0	-25.0	-27.0	-29.0	-31.0
1	H	A	-13.0	2.0	-11.0	-13.0	-15.0	-17.0	-19.0	-21.0	-23.0	-25.0	-27.0
2	H	D	-15.0	-11.0	-4.0	-13.0	-11.0	-17.0	-21.0	-23.0	-23.0	-22.0	-28.0
3		Q	-17.0	-13.0	-13.0	-7.0	-17.0	-6.0	-19.0	-21.0	-23.0	-24.0	-21.0
4		L	-19.0	-15.0	-7.0	-13.0	-14.0	-19.0	-2.0	-15.0	-17.0	-19.0	-21.0
5		T	-21.0	-17.0	-20.0	-10.0	-7.0	-18.0	-15.0	-8.0	-18.0	-16.0	-23.0
6	H	E	-23.0	-19.0	-22.0	-23.0	-4.0	-11.0	-17.0	-21.0	-11.0	-17.0	-20.0
7	H	E	-25.0	-21.0	-20.0	-25.0	-17.0	-2.0	-13.0	-17.0	-19.0	-15.0	-16.0
8	H	Q	-27.0	-21.0	-22.0	-22.0	-19.0	-15.0	-3.0	-14.0	-18.0	-20.0	-11.0
9	H	I	-29.0	-25.0	-17.0	-20.0	-21.0	-17.0	-9.0	3.0	-10.0	-12.0	-14.0
10	H	A	-31.0	-27.0	-19.0	-17.0	-23.0	-19.0	-13.0	-5.0	3.0	-10.0	-12.0

Table 5: The dynamic programming matrix of novel algorithm with PAM-80 scoring matrix.

			0	1	2	3	4	5	6	7	8	9	10
						T	T	T	T	T	H	H	H
			-	S	M	T	D	L	L	S	A	E	D
0		-	1.0	-13.0	-15.0	-17.0	-19.0	-21.0	-23.0	-25.0	-27.0	-29.0	-31.0
1	H	A	-13.0	3.0	-10.0	-12.0	-14.0	-16.0	-18.0	-20.0	-22.0	-24.0	-26.0
2	H	D	-15.0	-10.0	1.5	-10.5	-10.5	-14.5	-17.5	-18.5	-19.5	-20.0	-24.0
3		Q	-17.0	-12.0	-11.0	0.5	-12.0	-7.5	-15.5	-18.5	-20.0	-20.0	-19.5
4		L	-19.0	-14.0	-9.0	-10.5	-2.5	-12.5	-5.0	-13.5	-18.5	-22.0	-20.5
5		T	-21.0	-16.0	-15.5	-10.0	-7.0	-4.0	-15.0	-8.0	-15.0	-18.0	-24.0
6	H	E	-23.0	-18.0	-17.5	-16.5	-6.5	-8.5	-6.5	-18.0	-9.5	-14.5	-20.0
7	H	E	-25.0	-20.0	-18.5	-18.5	-18.5	-5.0	-9.0	-7.5	-19.5	-11.5	-14.0
8	H	Q	-27.0	-21.5	-19.0	-19.0	-20.0	-17.5	-5.0	-8.0	-7.0	-19.0	-8.0
9	H	I	-29.0	-24.0	-18.0	-17.5	-21.5	-20.0	-14.0	-0.5	-5.5	-7.0	-17.5
10	H	A	-31.0	-26.0	-19.5	-17.5	-20.5	-22.0	-17.5	-10.5	1.0	-6.0	-6.0

The dynamic programming matrix of this alignment is shown in Table 5

Finally, we use a real case to show the difference between traditional global alignment and our novel alignment algorithm. The two proteins are 1Lin and 1AVS:B, whose primary structures and secondary structures are shown as follows.

1Lin :

```
ADQLTEEQIA EFKEAFSLFD KDGDGTITTK
ELGTVMRSLG QNPTEAELQD
*****HHHHH HHHHHHHHHT TT*SSEE*HH
HHHHHHHHTT ****HHHHHH

MINEVDADGN GTIDFPEFLT MMARKMKDTD
SEEEIREAFR VFDKDGNGYI
HHHHH*SS*S SSEEHHHHHH HHH****TTS
HHHHHHHHHH HHTTT*SSEE

SAAELRHVMT NLGEKLTDEE VDEMIREADI
DGDGQVNYEE FVQMMTAK
*HHHHHHHHH HTT****HHH HHHHHHHH*S
SSSSSEEHHH HHHHHTTT
```

1AVS:B :

```
ASMTDQQAEA RAFLSEEMIA EFKAAFDMFD
ADGGGDISTK ELGTVMRMLG
*******HHH HTTTTHHHHH HHHHHHHHH*
TT*SSEE*HH HHHHHHHHTT

QNPTKEELDA IIEEVDEDGS GTIDFEEFLV
MMVRQMKEDA
****HHHHHH HHHHH*SSS* SSEEHHHHHH
HHHHHH****
```

If we simply use traditional global alignment algorithm, we shall obtain the following :

```
----------ADQLTEEQIAEFKEAFSLFDK
DGDGTITTKELGTVMRSLGQNPTE
ASMTDQQAEARAFLSEEMIAEFKAAFDMFDA
DGGGDISTKELGTVMRMLGQNPTK

AELQDMINEVDADGNGTIDFPEFLTMMARKM
KDTDSEEEIREAFRVFDKDGNGYI
EELDAIIEEVDEDGSGTIDFEEFL-VM---M
-----------------------

SAAELRHVMTNLGEKLTDEEVDEMIREADID
GDGQVNYEEFVQMMTAK
----VRQ-M-----K---------------
--------E-----DA-
```

The secondary structure of this alignment is:

```
----------*****HHHHHHHHHHHHHHHT
TT*SSEE*HHHHHHHHHHTT****H

*******HHHHTTTTHHHHHHHHHHHHHH*
TT*SSEE*HHHHHHHHHHTT****H

HHHHHHHHHH*SS*SSSEEHHHHHHHHH**
**TTS*HHHHHHHHHHHTTT*SSEE
HHHHHHHHHH*SSS*SSEEHHHHH-HH---
H-----------------------

*HHHHHHHHHHTT****HHHHHHHHHH*S
SSSSSEEHHHHHHHHTTT
----HHH-H-----*--------------
---------*-----**-
```

However, if we perform our novel sequence alignment algorithm, the result is

```
----------ADQLTEEQIAEFKEAFSLFD
KDGDGTITTKELGTVMRSLGQNPTE
ASMTDQQAEARAFLSEEMIAEFKAAFDMFD
ADGGGDISTKELGTVMRMLGQNPTK

AELQDMINEVDADGNGTIDFPEFLTMMARK
MKDTDSEEEIREAFRVFDKDGNGYI
EELDAIIEEVDEDGSGTIDFEEFLVMMVRQ
MK------E--DA-----------

SAAELRHVMTNLGEKLTDEEVDEMIREADI
DGDGQVNYEEFVQMMTAK
----------------------------
-----------------
```

The secondary structure of this alignment is:

```
----------*****HHHHHHHHHHHHHHHT
TT*SSEE*HHHHHHHHHHTT****H
*******HHHHTTTTHHHHHHHHHHHHHH*
TT*SSEE*HHHHHHHHHHTT****H

HHHHHHHHHH*SS*SSSEEHHHHHHHHH**
**TTS*HHHHHHHHHHHTTT*SSEE
HHHHHHHHHH*SSS*SSEEHHHHHHHHHHH
H*------*--**-----------

*HHHHHHHHHHTT****HHHHHHHHHH*S
SSSSSEEHHHHHHHHTTT
----------------------------
-----------------
```

From this real example, we can easily obtain better alignment when we make use of secondary structure information in sequence alignment problem. The result seems more meaningful compared with those obtain by traditional alignment algorithms. In fact, the RMSD between these two proteins is 1.0Å.

4 Conclusion

In this paper, we first introduce some basic knowledge in molecular biology. Then, we propose a new sequence alignment algorithm for protein sequences that combines not only the primary structure information but also the secondary structure information. Since protein secondary structures are more conserved than primary structures, we believe that our algorithm is more sensitive and specific in identifying homologies between proteins.

Additionally, our algorithm can be applied in protein structure prediction. As we know, homology modeling predicting methods relay much on good templates. Traditional methods in finding homology templates are to perform sequence alignment. Since traditional structure alignment algorithms do not use the information regarding secondary structures, we believe that our algorithm can find good homology templates that lose similarities in the primary structure level.

References

[1] S. Altschul and B. W. Erickson, "Optimal sequence alignment using affine gap costs," *Journal of Molecular Biology*, Vol. 48, pp. 603–616, 1986.

[2] S. Altschul, T.Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic Acids Research*, Vol. 25, pp. 3389–3402, 1997.

[3] S. F. Altschul, M. S. Boguski, W. G. Gish, and J. C. Wooton, "Issues in searching molecular sequence databases," *Nature Genetics*, Vol. 6, pp. 119–129, 1994.

[4] S. F. Altschul and D. J. Lipman, "Trees, stars and multiple sequence alignment," *Journal of Applied Mathematics*, Vol. 49, No. 1, pp. 197–209, 1989.

[5] D. J. Bacon and W. F. Anderson, "Multiple sequence alignment," *Journal of Molecular Biology*, Vol. 191, pp. 153–161, 1986.

[6] V. Bafna, E. L. Lawler, and P. Pevzner, "Approximation algorithms for multiple sequence alignment," *Proc. of 5th Ann. Symp. On Pattern Combinatorial Matching*, Vol. 807, pp. 43–53, 1994.

[7] H. Carrillo and D. J. Lipman, "The multiple sequence alignment problem in biology," *Journal of Applied Mathematics*, Vol. 48, pp. 1073–1082, 1988.

[8] K.-M. Chao, R. Hardison, and W. Miller, "Constrained sequence alignment," *Bulletin of Mathematical Biology*, Vol. 55, pp. 503–524, 1993.

[9] K.-M. Chao, R. Hardison, and W. Miller, "Locating well-conserved regions within a pairwise alignment," *Computer Application in the Biosciences*, Vol. 9, pp. 387–396, 1993.

[10] A. D. Gordon, "A sequence-comparison statistic and algorithm," *Biometrika*, Vol. 60, pp. 197–200, 1973.

[11] O. Gotoh, “An improved algorithm for matching biological sequences,” *Journal of Molecular Biology*, Vol. 162, pp. 705–708, 1982.

[12] O. Gotoh, “Optimal sequence alignment allowing for long gaps,” *Bulletin of Mathematical Biology*, Vol. 52, pp. 359–373, 1990.

[13] M. S. Johnson and R. F. Doolittle, “A method for the simultaneous alignment of three or more amino acid sequences,” *Journal of Molecular Evolution*, Vol. 23, pp. 267–278, 1986.

[14] W. Kabsch and C. Sander, “Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features,” *Biopolymers*, Vol. 22, pp. 2577–2637, 1983.

[15] S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequences of two proteins,” *Journal of Molecular Biology*, Vol. 48, pp. 443–453, 1970.

[16] W. Pearson and W. Miller, “Dynamic programming algorithms for biological sequence comparison,” *Methods in Enzymology*, Vol. 210, pp. 575–601, 1992.

[17] T. F. Smith and M. S. Waterman, “Comparison of biosequences,” *Advances in Applied Mathematics*, Vol. 2, pp. 482–489, 1981.

[18] W. R. Taylor, “A flexible method to align a large number of sequences,” *Journal of Molecular Evolution*, Vol. 28, pp. 161–169, 1988.

[19] U. Tonges, S. W. Perrey, J. Stoye, and A. W. M. Dress, “A general method for fast multiple sequence alignment,” *Gene*, Vol. 172, No. 1, pp. 33–41, 1996.

[20] A. Wong, S. Chan, and D. Chiu, “A multiple sequence comparison method,” *Society for Mathematical Biology*, Vol. 55, No. 2, pp. 465–486, 1993.