# A local structural alignment method that accommodates with circular permutation

Hiroike, Takaaki
Biomolecular Engineering Research Institute

Toh, Hiroyuki
Biomolecular Engineering Research Institute



出版情報：CBI journal. 1 (3), pp.103-114, 2001-10-30. 情報計算化学生物学会
バージョン：

Chem-Bio Informatics Journal, Vol. 1, No. 3, pp.103-114(2001)

# A local structural alignment method that accommodates with circular permutation.

Takaaki Hiroike and Hiroyuki Toh*

*Biomolecular Engineering Research Institute, 6-2-3, Furuedai, Suita 565-0874, Japan.*

**E-mail: toh@beri.co.jp*

(Received August 28, 2001; accepted October 4, 2001; published online October 30, 2001)

## Abstract

**Local structural alignment is an effective method to detect the local similarities between two protein structures. Orengo & Taylor (1993) have already developed a local structural alignment method by using a double dynamic programming algorithm. In the method, the similarity between a pair of residues is evaluated as the similarity between the pair of the structural environments corresponding to the residues. However, it is difficult to evaluate the structural similarities between a pair of proteins related by circular permutation, because the structural environment of a residue drastically differs from that of the corresponding residue of the circularly permuted protein. In this manuscript, we propose a new method to construct a structural environment that is robust against circular permutation. We examined its efficiency in the detection of the local structural similarity by the reconstructed structural environments.**

**Key Words:** local structural alignment, double dynamic programming algorithm, circular permutation

**Area of Interest:** Bioinformatics and Bio computing

## 1. Introduction

Recent advances in structure determination technologies have caused the rapid accumulation of protein structures coordinates data. Along with this situation, several different methods for structure comparison have been developed to extract functional and evolutionary information from the coordinates data. The algorithms for the structure comparison differ among the various methods. For example, a program called DALI [1] compares the protein structures according to the differences in the distance map, while May and Johnson [2] proposed a method by rigid body superimposition. Among the various methods, the development of structural alignment by the double dynamic programming algorithm (DDP) [3] was a pioneering advance. The algorithm is an extension of the dynamic programming algorithm (DP) for sequence alignment. In DP for sequence alignment, the similarity between a pair of residues is evaluated with a numerical table called a

score matrix. In the structural alignment by DDP, a residue in a protein is represented by a set of vectors from the α-carbon of the residue to those of all the other residues. The set is ordered according to the position in the primary structure of the destination residue, and is called the structural environment. The structural environment of a residue is considered to express the relative position of the residue in the structure. Therefore, the similarity in the structural environments between a pair of residues indicates the degree of structural equivalence between the residues in the different structures. The structural environment of a residue can be regarded as one-dimensional data, like sequence data. Therefore, the alignment score between two structural environments is considered to represent the similarity between the two environments. The alignment score between a pair of structural environments can be obtained by applying DP. When the alignment score between every pair of the structural environments under consideration is obtained, the residue-to-residue correspondence between the two protein structures can be made by DP, using the alignment score between a pair of structural environments as the similarity between the corresponding residue pair. The DP to obtain the alignment score between a pair of structural environments is called lower-level DP, while the DP to make residue-to-residue correspondence is called upper-level DP (see below for more details about DDP). Thus, DP is applied to two different stages of the structural alignment. Therefore, the method is called DDP. Subsequently, Taylor and Orengo have extended the method toward local structural alignment [4], protein structure database searching [5], and multiple structural alignment [6].

Recently, a new mechanism for the evolution of protein structure, called circular permutation, has been identified [7]. Let's consider a protein structure consisting of a single domain. In the mechanism, the N- and C-terminal portions have been exchanged during the course of evolution. Such a permuted protein is considered to be generated, accompanied with tandem gene duplication. After the duplication, the two genes are fused. Then, both the region encoding the N-terminal portion of the first copy and the region encoding the C-terminal portion of the second copy are deleted. Consequently, a gene encoding the permuted protein is generated. The folding of the permuted structure is basically the same as that of the original structure. If the original structure and the permuted structure are close enough for the similarity at the sequence level to be observed, then the circular permutation could be detected by the local sequence alignment by the Smith & Waterman algorithm [8]. Recently, Uiel et al. [9] also developed a method to detect circular permutation by sequence comparison. However, it is difficult to detect circular permutation by sequence comparison when the sequences are highly diverged. In such a situation, the structure comparison is considered to be more effective to detect local similarity than the sequence comparison. As described above, Orengo and Taylor [4] already extended their approach by DDP toward the local structural alignment. However, there is a problem in their method when circular permuted proteins are considered. Schematic diagrams of an original structure (Protein A) and the corresponding permuted structure (Protein B) are shown in Fig. 1. Let's consider residue 9 of Protein A of Fig. 1. The residue 9 of Protein B occupies a structurally equivalent position in Protein B, although the residue number in Protein B does not correspond to the residue position in Protein B. In spite of the structural equivalence, the structural environment of residue 9 of Protein A is different from that of residue 9 of Protein B. It is difficult to detect the local structural similarity related to circular permutation by DDP, due to the drastic change in the structural environment.

As described above, the structural environment is ordered according to the residue number of the destination residue. This order dependence causes the change of the structural environment by the circular permutation. If the structural environment can be constructed to be robust against circular permutation, then it could detect the local structural similarity between permuted structures by DDP. In this manuscript, we will discuss a novel method to construct a structural environment that is robust against circular permutation. The efficiency of the local structural alignment by DDP

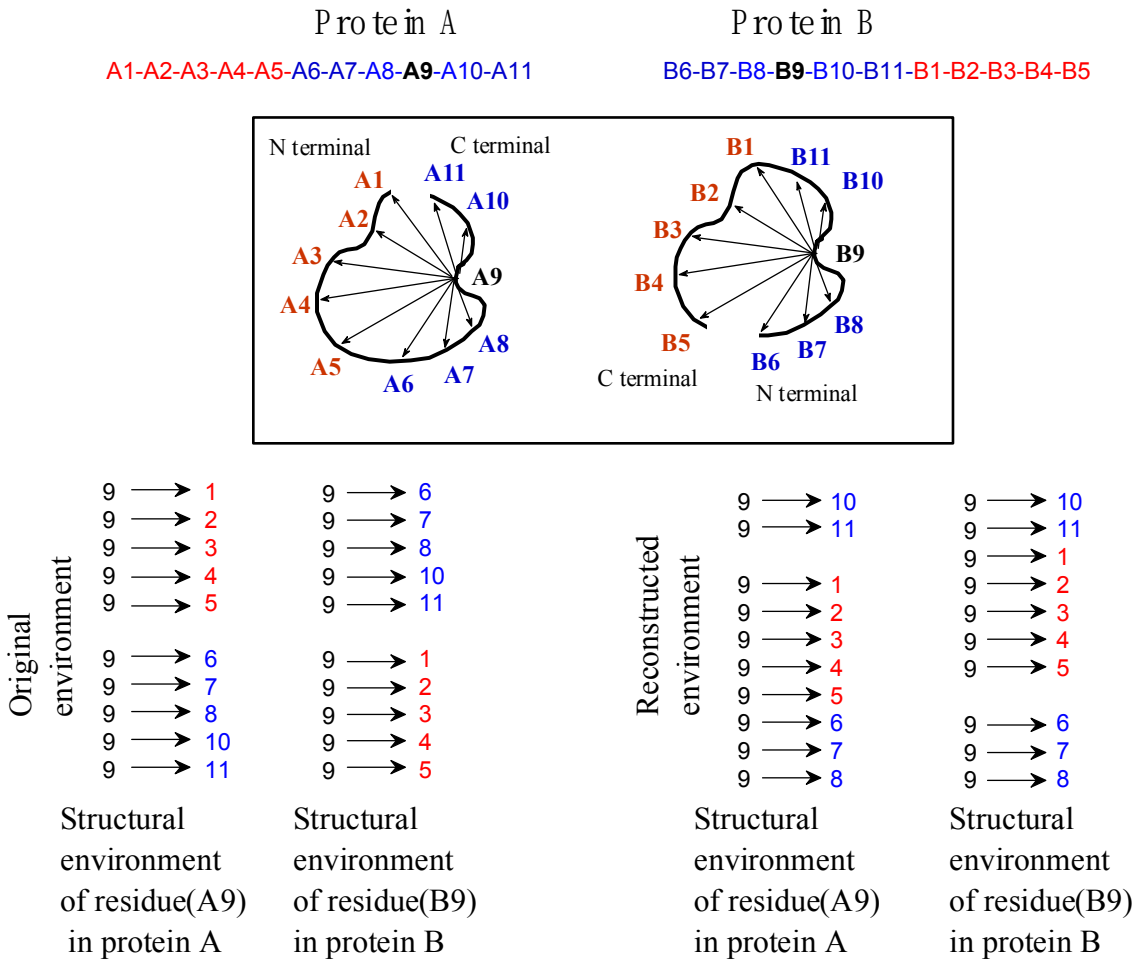



**Fig.1** Comparision between original and reconstructed environments

with the novel structural environment was examined by applying the method to several pairs of proteins that are related by local similarities or circular permutation.

## 2. System and methods

The programs were written in ANSI C. The research described herein was performed on a Silicon Graphics workstation, INDY.

### 2-1. Construction of the structural environment

As described above, the structural environment of a residue is expressed as a set of vectors from an atom of a residue to the corresponding atoms of the other residues. To calculate the vectors, the local frame for each residue should be defined [3]. In this manuscript, the local frame of a residue is defined as follows; The X-axis is defined to run through Cα along a vector from the carbonyl carbon to the nitrogen on the main chain of the same residue. Then, a plane, which is vertical to the X-axis and is running through Cα is considered. The Y-axis is defined to run through

Cα along a vector projected from the vector from Cα to Cβ onto the plane. For Gly, a pseudo atom corresponding to Cβ was introduced for the calculation. Finally, the Z-axis is defined to run through Cα along the outer product of the vector along the X-axis with the vector along the Y-axis. The definition of the local frame is slightly different from the original definition by Taylor and Orengo [3].

The structural environment of a residue was defined here as a set of vectors from the α-carbon of the residue to those of all the other residues in the same structure. The vector set was divided into two subsets. One of them consists of the vectors whose destination residues are included in the N-terminal region to the residue under consideration, while another is constituted by the remaining vectors, that is, the vectors with the destination residues present in the C-terminal region to the residue under consideration. The former is referred to as the N-terminal environment, while the latter is called the C-terminal environment. In the method by Taylor and Orengo [3], the N-terminal environment is placed before the C-terminal environment to construct a structural environment. To make the environment robust against circular permutation, however, we reconstructed the structural environment so that the C-terminal environment appears at first, and is followed by the N-terminal environment. We explain the meaning of this modification with the toy models of circular permuted proteins of 11 residues (see Fig. 1). Suppose that Protein B is related to Protein A by circular permutation. As described above, the structural environments of residues A9 and B9, which are constructed by the original method [3], are different from each other, although both residues occupy a structurally equivalent position in the structures. In contrast, the elements of the structural environment of B9, as reconstructed by our method, are ordered in the same manner as those of A9, in spite of the circular permutation. That is, the structural environment reconstructed by our method is robust against the circular permutation.

We previously modified DDP by introducing a distance cut-off approximation, to reduce the computational time required for the lower level DP [10]. In the approximation, only the residues with Cα's present within the sphere centered at the Cα of a residue are used as destination residues to construct the structural environment of the residue under consideration. The radius of the sphere is called the cut-off distance, and the environment thus obtained is referred to as the local environment of the residue. In this study, we also introduced the approximation for the rapid computation.

### 2-2. Lower level DP

This step is basically the same as the algorithm developed by Orengo & Taylor [4]. For every pair of residues, vector sets defining the environments are compared in a two-dimensional matrix called the lower level matrix. Consider a case to calculate the similarity between residue *i* of Protein A and residue *j* of Protein B by the lower level DP. The lower level matrix is here designated as *LMri,j*. The size of the matrix is (the size of the structural (or local) environment of residue *i* + 1) x (the size of the structural (or local) environment of residue *j* + 1). The similarity between a pair of vectors, $V_{i \to m}$ from the structural environment of residue *i* of Protein A and $V_{j \to n}$ from the structural environment of residue *j* of Protein B, is evaluated by the following expression:

$$^{Lower}S_{m,n} = a / ( | V_{i \to m} - V_{j \to n} | - b)$$

where *a* and *b* are constants. In this study, *a* and *b* were set to be 50 and 5, respectively. Each element of the lower level matrix was obtained by solving the following recurrence equations:

$$E(m, n) = \mathrm{Max}\ (LMr\ i, j(m, n-1) - opgap,\quad E(m, n-1) - exgap)$$

$$F(m, n) = \mathrm{Max}\,(LMr\ i, j(m-1, n) - opgap,\quad F(m-1, n) - exgap)$$
$$LMr\ i, j(m, n) = \mathrm{Max}\,({}^{Lower}S_{m, n} + LMr\ i, j(m, n)),\quad E(m, n),\quad F(m, n))$$

where *opgap* and *exgap* indicate the opening gap penalty and the extension gap penalty for lower level DP, respectively. In this study, *opgap* and *exgap* were set to be 10 and 0. *E* and *F* are the two-dimensional matrices of the same size as *LMri,j*. This procedure is known as Gotoh's algorithm [11]. The global alignment score ( ${}^{Total}LMri,j$ ) of the lower level DP is regarded as the similarity in the structural environment between the residue *i* of protein A and the residue *j* of protein B. However, this calculation can be skipped to reduce the computing cost, if the two residues constitute different secondary structures, and thus the score, ${}^{Total}LMri,j$, for the residue pair was set to zero. On the other hand, ${}^{Total}LMri,j$ was multiplied by a weighting factor of 1.2, when both residues, *i* and *j*, constitute the same type of secondary structure. In this study, only α helix and β strand were considered as the types of secondary structures, and the secondary structures were assigned based on Kabsch & Sander's algorithm [12]. As described above, a distance cut-off approximation [10] was introduced for the calculation. However, the same procedure for the lower level DP was applied to evaluate the similarity for the local environments. The average ($\mu$) and the standard deviation ($\sigma$) over all of the ${}^{Total}LMr\ i,j$ were calculated. Then, all of the ${}^{Total}LMri,j$ were normalized with the average and the standard deviation. The normalized score was used as the similarity between the corresponding residue pair, which is here denoted as ${}^{Upper}Si,j$.

$${}^{Upper}S\ i, j = (\ {}^{Total}LMr\ i, j - \mu_{\cdot} / \sigma$$

## 2-3. Upper level DP

Using the ${}^{Upper}Si,j$ thus obtained, upper level DP was performed to make the residue-to-residue correspondence. Like the lower level DP, a two-dimensional matrix is required for the upper level DP. The matrix is designated as *D*. The size of the matrix is (the number of residues of Protein A + 1) x (the number of residues of Protein B + 1). The following recurrence equations were solved on the matrix:

$$E(i, j) = \mathrm{Max}\,(D(i, j-1) - opgap,\quad E(i, j-1) - exgap)$$
$$F(i, j) = \mathrm{Max}\,(\mathrm{D}(i-1, j) - opgap,\quad F(i-1, j) - exgap)$$
$$D(i, j) = \mathrm{Max}\,({}^{Upper}S\ i, j + D(i, j),\quad E(i, j),\quad E(i, j),\quad 0)$$

where *opgap* and *exgap* indicate the opening gap penalty and the extension gap penalty for upper level DP, respectively. In this study, *opgap* and *exgap* were set to be 10 and 0. *E* and *F* are two-dimensional matrices with the same size as *D*. This algorithm is the same as Smith & Waterman's algorithm for local sequence alignment [8], except for the similarity score between a residue pair. A gap penalty, multiplied by a weighting factor of 1.2, was used to avoid gaps in the secondary structures, when the residues constituting the secondary structures were under examination.

## 2-4. Backtracking

Like the original method by Orengo & Taylor [4], the backtracking was performed from the highest score of the residue pairs to the zero score of the residue pairs in the upper level matrix. Once a local path was established, Taylor et al. [4] set scores in the upper level matrix to zero for all elements along the local path and also for elements in a window of several elements adjacent to the

local path, in order to prevent another path from passing through the same region of the matrix. We made a change to the algorithm for obtaining the second or later local path. Instead of the scores in the upper level matrix, the similarity scores corresponding to ${}^{Upper}Si,j$ were reset to zero for all elements along the local path and also for the elements in the window. The scores corresponding to the elements outside the window were reset to be the same as the current value. The half-width of the window was set to 10 elements from the local path. To get the next local path, the upper level DP was performed again with the reset similarity scores. This operation was effective to prevent another path from passing through the same region of the matrix. This procedure was repeated until the highest score in the upper level matrix was below a given threshold value. In this study, the threshold value was set to be zero.

### 2-5. Calculating the local alignment score

The alignment score was divided into the length of the alignment, in order to avoid the influence of the sequence size, and the obtained alignments were sorted in decreasing order of alignment scores.

### 2-6. Examination of the efficiency of local similarity detection by the reconstructed structural environments

At first, we compared the reconstructed structural environments with the original environments, in terms of their efficiency in detecting the local structural similarities related by circular permutation. We collected five pairs of proteins in the relationship of circular permutation. The detection was judged as successful when 70% or more of the expected regions was aligned. When alignments related to circular permutation were obtained as the best and the second scores, they were judged as detecting two alignments related to circular permutation. Likewise, it was judged as detecting one alignment related to circular permutation when one alignment related to that was obtained as the best score.

The method proposed in this study has been developed to treat local structural similarity related by circular permutation. However, the method should be applicable for the detection of local structural similarity that has nothing to do with circular permutation, because we do not know whether the local similarity is generated by circular permutation or other mechanisms, such as exon-shuffling, when we try to search for local structural similarity. Therefore, the efficiency of detecting local structural similarities was compared between the reconstructed environments and the original environments, by applying local structural alignments by DDP with the environments to two pairs of proteins with local similarities that were not generated by circular permutation. We collected three pairs of proteins with local structural similarities that are not related by circular permutation. The definition of success in detection was the same as that described above. The number of local alignments, which were successful in detection and were ranked within a range from the top to the expected numbers of local alignments, were used as a measure for the sensitivity of the structural environments under examination.

## 3. Results and Discussion

### 3-1. Comparison of the detection efficiency of local similarity related by circular permutation between the reconstructed structural environments and the original structural environments.

We applied our method to five pairs of proteins that have the relationship of circular permutation to each other: transaldolase (ONR) /aldolase (FBA) [13], human glutathione synthetase (HGS) /*Escherichia coli* glutathione synthetase (GSA) [14], FAD-binding domain of phthalate dioxygenase reductase (PIA) /FMN binding protein (AXJ1) [15], glucanase (GBG) /artificially permuted glucanase (AJK) [16], and avidin (AVD) /artificially permuted streptavidin (SWG) [17]. An abbreviated name of each protein used in this study is shown in the parentheses after the full name of the protein. To evaluate the efficiency of our method, we compared the alignments using our reconstructed structural environment with those using the original structural environment. Two different cut-off distances, 25 angstroms and 100 angstroms, were examined for the construction of structural environments. Under the latter cut-off distance, the distance cut-off approximation is not regarded as being applied virtually, because the cut-off distance of 100 angstroms is large enough to include almost the entire structure of a protein. To examine the efficiency of our method, we also applied the local structural alignment with the original structural environments to the five pairs of proteins.

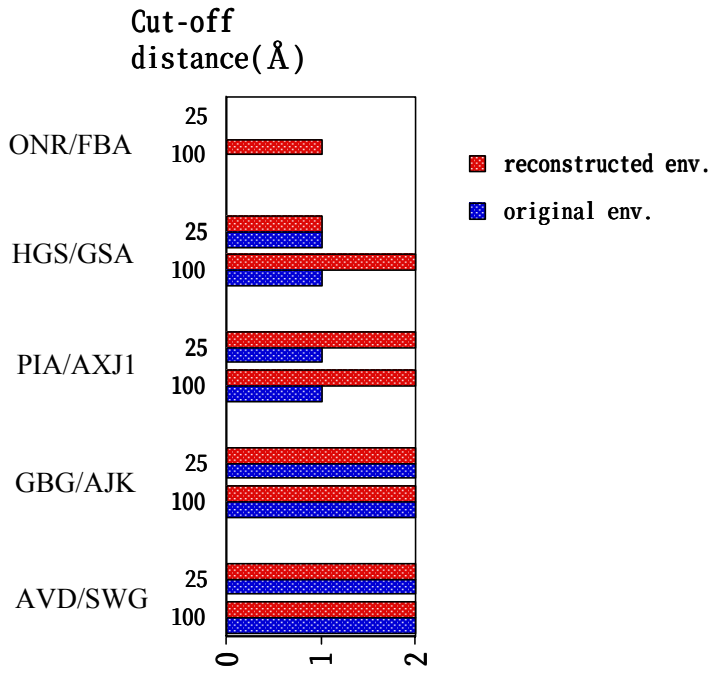
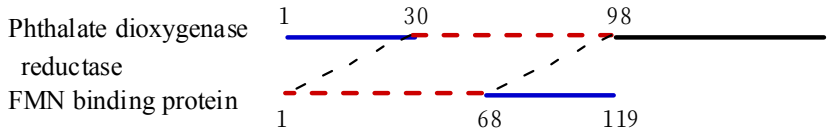



**Fig. 2** Detected number of alignments related by circular permutation.

The results of the detection of permuted structures by the two methods are shown in Fig. 2. If the two proteins are related by circular permutation, then the generation of two local alignments with the best and the second best scores would be expected. That is, the C-terminal part of a protein would be aligned with the N-terminal part of another protein. At the same time, an alignment of the N-terminal part of the former and the C-terminal part of the latter would be generated. The abscissa of Fig. 2 indicates the number of local alignments related to the circular permutation within the best and the second best alignments of the output list of a method. The ordinate indicates a pair of

proteins under comparison and the cut-off distance. Three out of the five cases were comparisons of pairs of proteins that are related by circular permutation that occurred during the course of molecular evolution. In either case, the proteins were highly diverged not only at the sequence level, but also at the structure level. Due to the high divergence, it was difficult to detect the permuted regions. For example, local structural alignment with the original environment failed to detect the similarity related to the permuted regions between ONR and FBA. When our structural environment was used and the distance cut-off was not used virtually, one out of the two possible


a.




b.

```
L1=40~121  L2=11~89 SCORE MAX = 2.445105        (1)
EAGANLTVAVP---NGSRRTYSLC-NDSQERNRYVIAVKRDSNGRGGSISFIDDTSEGDAVEVSLPRNEFPLDKRAKSFILVAGG
KNEGVVAIATQGEDGPHLVNTWNSYLKVLDGNRIVVPVG-------GMHKTEANVARDERVLMTLGSRKVAGRNGPGTGFLIRGS
   EEEEEE       EEEEE          EEEEEEEEEE     HHHHHHH     EEEEE              EEEEEEE
   EEEEEEE      EEEEEEE         EEEEEE          HHHHHHHHH EEEE               EEEEEEEE

L1=6~29  L2=80~114 SCORE MAX = 2.055205         (2)
DGFLRLKIASKEKIAR------------DIWSFEL
GTGFLIR-GSAAFRTDGPEFEAIARFKWARAALVI
   EEEEEEEEE  EE            EEEEEEE
 EEEEEE EEEEEE  HHHHHHHH      EEEEE
```

c.

```
L1=32~130  L2=3~98   SCORE MAX = 1.865876     (1)
QGAPLPPFEAGANLTVAVP---NGSRRTYSLC-NDSQERNRYVIAVKRDSNGRGGSISFIDDTSEGDAVEVSLPRNEFPLDKRAKSFILVAGGIGITPMLSM
PGTFFEVLKNEGVVAIATQGEDGPHLVNTWNSYLKVLDGNRIVVPVG-------GMHKTEANVARDERVLMTLGSRKVAGRNGPGTGFLIRGSAAFRTDGPE
           EEEEEE       EEEEE          EEEEEEEEEE     HHHHHHH     EEEEE              EEEEEEE   HHHHHHH
  HHHHHH   EEEEEEE      EEEEEEE         EEEEEE          HHHHHHHHH EEEE               EEEEEEEEEEEE  HHH

L1=6~88  L2=26~102   SCORE MAX = 1.526987      (2)
DGFLRLKIASKEKIARDIWSFELTDPQGAPLPPFEAGANLTVAVPNGSRRTYSLCNDSQERNRYVIAVKRDSNGRGGSISFI
PHLVNTWNSYLKVLDGNRIVVPVGGM-HKTEANVARDERVLMTLGSRKVAGRNGP-----GTGFLIRGSAAFRTDGPEFEAI
   EEEEEEEEE  EEEEEEEEEEE            EEEEEE    EEEEE         EEEEEEEEEE     HHHHHH
 EEEEEEE         EEEEEE    HHHHHHHHH EEEE                     EEEEEEEEEEEE  HHHHHHH
```

**Fig.3** (a) Schematic diagrams of the primary structures for phthalate dioxygenase reductase and FMN binding protein. Regions with the same colors have local similarity by circular permutation. (b) The alignments by the reconstructed structural environments. (c) The alignments by the original structural environments. Secondary structure assignments are shown as H (α helix) and E (β strand).

local similarities was detected as the best alignment in the output. As for the comparison between HGS and GSA, one of the two possible alignments was detected by both methods under the cut-off distance of 25 angstroms. When the cut-off distance was 100 angstroms, our method successfully detected two possible local similarities as the best and the second best alignments. However, the method using the original structural environments could still detect only one local similarity. In the case of the comparison between PIA and AXJ1, our method successfully detected two possible similarities as the best and the second best alignments under both cut-off conditions. However, the method with the original environments could detect only one of the local similarities under either cut-off condition. The local structural alignments between PIA and AXJ1 obtained by the two methods are shown in Fig. 3 as an example of the calculation. The same colored portions indicate the similar 3D-structure portions related to each other by circular permutation. The N-terminal portion of PIA and the C-terminal portion of AXJ1 are colored in blue, while the C-terminal portion of the FAD-binding region of PIA and the N-terminal portion of AXJ1 are colored in red. As shown in the figure, our method successfully detected the similarities related to the permuted regions, while the method with the original environments could detect only one of the two expected similarities. However, it was occasionally difficult for our method to decide the boundaries of local similarities; for example, the upper alignment in Fig. 3 b suggests that the C-terminal boundary of the local similarity was further extended toward the C-terminal regions. These results suggested that our structural environment is more efficient for local structural alignment by DDP than the original structural environment. In addition, the distance cut-off approximation does not seem to be efficient, although the number of applied examples was quite limited.

Two out of the five examined cases were the comparisons of a protein structure with an artificially permuted structure, e.g., AVD/SWG and GBG/AJK. In each case, the artificially permuted structure was almost identical with the original structure, except for the exchange of the N-terminal and the C-terminal portions. As for the two cases, both the reconstructed and the original environments could generate the local structural alignments corresponding to the circular permutation. In addition, the distance cut-off approximation did not affect the efficiency for the two cases. As described above, the original structural environments are different between the residue of the original structure and the structurally equivalent residue of the artificially permuted protein. The results could be explained as follows; reflecting the high structural similarity between the two proteins, however, the N-terminal environment of a residue from a protein would be correctly aligned with the C-terminal environment of the corresponding residue from the permuted protein or *vice versa* in the lower level DP, which would generate the high similarity score between the structurally equivalent residues.

### 3-2. Comparison of the efficiency in detection of local similarity that is not related by circular permutation between the reconstructed structural environments and the original structural environments.

We applied the local structural alignments by DDP, with both our environments and the original environments, to two pairs of proteins with local similarities that have nothing to do with circular permutation: aspartic proteinase [18] /HIV-1 protease [19], and cyclooxygenase-2 [20] /lectin & the EGF-like domains of E-selectin [21]. We examined whether the reconstructed environment has the same efficiency in local similarities that have nothing to do with circular permutation between the reconstructed environments and the original environments. The computations were carried out under the same conditions, except for the structural environments.

When the cut-off distance was set to more than 15 angstroms, both methods for the construction of structural environments showed the same efficiency in the detection of the local similarities, although the obtained alignments were slightly different from each other. The local structural alignments between aspartic proteinase and HIV-1 protease are shown in Fig. 4. The aspartic proteinase has an internal duplication in the primary structure, and each unit of the duplication shares the same fold with HIV protease. As shown in Fig. 4, two local alignments between HIV protease and each unit of the aspartic proteinase were ranked as the best and second best similarities in the output lists of both methods. Due to the high sequence divergence, it is difficult to compare the proteins at the sequence level. However, the residue-to-residue correspondences generated by both local structure comparisons seem to be good, despite the difference in their structural environments. Both cyclooxygenase-2 and E-selectin have an EGF-like domain, although the global structures are quite different from each other. By applying the local structural alignments, the alignment of the EGF-like domains generated the best local similarity, despite the difference in the structural environments (data not shown). These results indicate that the reconstructed structural environment has the same efficiency in the detection of local structural similarity by DDP as the original structural environments.

(a)

```
L1=9~145  L2=1~95 SCORE MAX = 2.623669
TDYGNDIEYYGQVTIGTPGKKFNLDFDTGSSDLWIASTLCTNCGSGQTKYDPNQSSTYQADGRTWSISYGDGSSASGILA
PQITLWQRPLVTIKIGG--QLKEALLDTGADDTVLEEMSL-----------------PGRWKPKMIGGIGG-FIKVRQY
EEE    EEEEEEEE    EEEEEEEE      EEEE                        EEEEEEE   EEEEEEEE
          EEEEE    EEEEEEEE      EE                          EEEEEEEEE  E EEEEEEE

K-DNVNLGGLLIKGQTIELAKREAASFASGPNDGLLGLGFDTITTVRGVKTPMDNLI
DQILIEICGHKAI-GTVLVGP---------TPVNIIGRN-----------LLTQIG
E EEEEE  EEE  EEEEEEEEEEHHHH    EEEE               HHHHHH
  EEEEE  EEEE EEEEE               EE HH            HH

L1=184~313  L2=1~99 SCORE MAX = 2.503183
TTVPIDNSRGWWGITVDRATVGTSTVASSFDGILDTGTTLLILPNNIAASVARAYGASDNGDGTYTISCDTSAFKPLVFS
PQITLW---QRPLVT---IKIGGQ----LKEALLDTGADDTVLEEMSLP-----GRWKPKMIGGIGGFIKVRQYDQILIE
EEE        EEEEEEEEEE        EEEEEE     EEE  HHHHHHHHH          EEE         EEEE
             EE   EEE  E    EEEEEEE      EE            EEEEEEEEE  EEEEEEE  EEEE

INGASFQVSPDSLVFEEFQGQCIAGFGYGNWGFAIIGDTFLKNNYVVFN
ICGHKAIG---------------TVLVGPTPVNIIGRNLLTQIGCTLN
E  EEEEE HHHH  EEE  EEE          EEE HHH  EEEEEEE
E  EEEEE               EEEE      EE HHHH
```

(b)

```
L1=9~149  L2=1~96 SCORE MAX = 2.954435
TDYGNDIEYYGQVTIGTPGKKFNLDFDTGSSDLWIASTLCTNCGSGQTKYDPNQSSTYQADGRTWSISYGDGSSASGILA
PQITLWQRPLVTIKIGG--QLKEALLDTGADDTVLEEMSL-----------------PGRWKPKMIGGIGG-FIKVRQY
EEE    EEEEEEEE    EEEEEEEE      EEEE                        EEEEEEE   EEEEEEEE
          EEEEE    EEEEEEEE      EE                          EEEEEEEEE  E EEEEEEE

K-DNVNLGGLLIKGQTIELAKREAASFASGPNDGLLGLGFDTITTVRGVKTPMDNLISQGL
DQILIEICGHKAI-GTVLVGP---------TPVNIIG--------------RNLLTQIGC
E EEEEE  EEE  EEEEEEEEEEHHHH    EEEE               HHHHHH
  EEEEE  EEEE EEEEE               EE                HHHH

L1=184~313  L2=1~99 SCORE MAX = 2.901170
TTVPIDNSRGWWGITVDRATVGTSTVASSFDGILDTGTTLLILPNNIAASVARAYGASDNGDGTYTISCDTSAFKPLVFS
PQITLW---QRPLVT---IKIGGQ----LKEALLDTGADDTVLEEMSLP-----GRWKPKMIGGIGGFIKVRQYDQILIE
EEE        EEEEEEEEEE        EEEEEE     EEE  HHHHHHHHH          EEE         EEEE
             EE   EEE  E    EEEEEEE      EE            EEEEEEEEE  EEEEEEE  EEEE

INGASFQVSPDSLVFEEFQGQCIAGFGYGNWGFAIIGDTFLKNNYVVFN
ICGHKAIGT--------------VLVGPTPVNIIGRNLLTQIGCTLN
E  EEEEE HHHH  EEE  EEE          EEE HHH  EEEEEEE
E  EEEEEE              EEE      EE HHHH
```

**Fig.4** The result of the local structural alignment between HIV-1 protease and aspartic proteinase. Secondary structure assignments are shown as H(α helix) and E(β strand). The active sites are colored in red. (a) The alignments obtained by the reconstructed structural environment. (b) The alignments obtained by the original structural environment.

## 4. Conclusions

In this paper, we reconstructed the structural environments to be robust for circular permutation. The local structural alignment by DDP with the reconstructed environments showed more sensitivity in the detection of local similarity due to circular permutation than that with the original environments. In addition, the reconstructed environments showed the same efficiency in the detection of local similarity that is unrelated to circular permutation as the original environments. In either case, the distance cut-off approximation reduced the efficiency in the detection of the local similarity. Thus, a simple modification of the structural environments has extended the applicability of DDP. At this stage, the number of natural circular permutation examples is small. In the future of structural genomics, however, the chances of encountering circular permutations will increase. Our method provides a solutions for the treatment of circular permutation, although further improvement is required for the sensitive detection of local similarities.

## References

[1] L.Holm, C.Sander, *J. Mol. Biol.,* **233**, 123-38 (1993).
[2] A.C.May, M.S.Johnson, *Protein Eng.,* **7**, 475-85 (1994)
[3] W.R.Taylor, C.A.Orengo, *J. Mol. Biol.,* **208**, 1-22 (1989).
[4] C.A.Orengo, W.R.Taylor, *J. Mol. Biol.,* **233**, 488-497 (1993).
[5] C.A.Orengo, N.P.Brown, W.R.Taylor, *Proteins,* **14**, 139-67 (1992).
[6] W.R.Taylor, T.P.Flores, C.A.Orengo, *Protein Sci.,* **3**, 1858-70 (1994).
[7] U.Heinemann , M.Hahn, *Prog. Biophys. Mol. Biol.,* **64**, 121-43 (1995)
[8] T.F.Smith, M.S. Waterman, *J. Mol. Biol.,* **147**, 195-197 (1981).
[9] S.Uliel, A.Fliess, A.Amir, R.Unger, *Bioinformatics,* **15**, 930-936 (1999).
[10] H.Toh, *CABIOS,* **13**, 387-396 (1997)
[11] O.Gotoh, *J. Mol. Biol.,* **162**, 705-8 (1982)
[12] W.Kabsch, C.Sander, *Biopolymers,* **22**, 2577-637 (1983).
[13] J.Jia, W.Huang, U.Schorken, H.Sahm, G.A. Sprenger, Y.Lindqvist, G.Schneider, *Structure,* **4**, 715-24 (1996).
[14] G.Polekhina, P.G.Board, R.R.Gali, J.Rossjohn, M.W.Parker, *The EMBO Journal,* **18**, 3204-3213 (1999).
[15] A.G.Murzin, *Nat. Struct. Biol.,* **5**, 101 (1998).
[16] J.Ay, M.Hahn, K.Decanniere, K.Piotukh, R.Borriss, U.Heinemann, *Proteins,* **30** , 155-67 (1998)
[17] V.Chu V, S.Freitag, I.Le Trong, R.E.Stenkamp, P.S.Stayton, *Protein Sci*., **7**, 848-59 (1998)
[18] R.Bott, E.Subramanian, D.R.Davies, *Biochemistry,* **21** ,6956-62 (1982 )
[19] A.Fassler, J.Rosel, M.Tintelnot-Blomley, E.Alteri, G.Bold, M.Lang, *Bioorg.Med. Chem.,***3**, 2837 (1993)
[20] R.G.Kurumbail, A.M.Stevens, J.K.Gierse, J.J.McDonald, R.A.Stegeman, J.Y.Pak, D.Gildehaus, J.M.Miyashiro, T.D.Penning, K.Seibert, P.C.Isakson, W.C.Stallings, *Nature,* **384**, 644-8 (1996)
[21] K.J.Kim, et al., *J .Biol .Chem. ,* **269**, 4431-7 (1994 )

Chem-Bio Informatics Journal, Vol. 1, No. 3, pp.103-114(2001)

# Circular permutationを取り扱うことのできる構造局所アラインメント

廣池隆明、*藤博幸

大阪府吹田市古江台６−２−３
*生物分子工学研究所情報解析研究部門*
*E-mail: toh@beri.co.jp*

## 要旨

蛋白質の立体構造の局所的な類似を見い出す手段として、構造局所アラインメントと呼ばれる手法がある。この手法は、各残基の立体構造中における相対的な位置（構造環境）の類似度を評価することにより、局所的に類似した構造を持つ残基間に対応づけ（アラインメント）を行う方法である。構造局所アラインメントには、Orengo & Taylor(1993)により開発されたdouble dynamic programming algorithmを利用した方法がある。しかし、彼等の手法では、ある残基は、構造環境とよばれるその残基のβ炭素から他の全ての残基のβ炭素へのベクトルの集合で表現され、その集合の要素は、ベクトルの向く先の残基の一次構造上の残基番号に従い順序付けられている。この構造環境に基づく類似性の評価は、その表現における１次構造上の順序依存性のため、circular permutationによる配列の入れ替わりによって生じる局所的な類似構造の検出は困難であると考えられた。そこで、我々はcircular permutationにより生じた局所類似性も取り扱えるように、各残基の構造環境をcircular permutationに不変になるように構造環境を再構築し、double dynamic programming algorithmによる構造局所アラインメントの適用できる範囲を拡張した。

**キーワード**：構造局所アラインメント、ダブル・ダイナミック・プログラミング、サーキュラーパーミュテーション

**領域区分**：分子生物学における情報計算技術