Computational and Structural Biotechnology Journal 21 (2023) 185–201

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

SeqCP: A sequence-based algorithm for searching circularly permuted proteins

Chi-Chun Chen [a,b], Yu-Wei Huang [c], Hsuan-Cheng Huang [a,d], Wei-Cheng Lo [c,e,f,*], Ping-Chiang Lyu [b,*]

[a] Bioinformatics Program, Institute of Information Science, Taiwan International Graduate Program, Academia Sinica, Taipei 115, Taiwan
[b] Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu 300, Taiwan
[c] Institute of Bioinformatics and Systems Biology, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan
[d] Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei 112, Taiwan
[e] Department of Biological Science and Technology, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan
[f] The Center for Bioinformatics Research, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

ARTICLE INFO

Article history:
Received 21 June 2022
Received in revised form 10 November 2022
Accepted 10 November 2022
Available online 14 November 2022

Keywords:
Circular permutation
Circular permutants
Protein sequence analysis
Protein structure modeling

ABSTRACT

Circular permutation (CP) is a protein sequence rearrangement in which the amino- and carboxyl-termini of a protein can be created in different positions along the imaginary circularized sequence. Circularly permutated proteins usually exhibit conserved three-dimensional structures and functions. By comparing the structures of circular permutants (CPMs), protein research and bioengineering applications can be approached in ways that are difficult to achieve by traditional mutagenesis. Most current CP detection algorithms depend on structural information. Because there is a vast number of proteins with unknown structures, many CP pairs may remain unidentified. An efficient sequence-based CP detector will help identify more CP pairs and advance many protein studies. For instance, some hypothetical proteins may have CPMs with known functions and structures that are informative for functional annotation, but existing structure-based CP search methods cannot be applied when those hypothetical proteins lack structural information. Despite the considerable potential for applications, sequence-based CP search methods have not been well developed. We present a sequence-based method, SeqCP, which analyzes normal and duplicated sequence alignments to identify CPMs and determine candidate CP sites for proteins. SeqCP was trained by data obtained from the Circular Permutation Database and tested with nonredundant datasets from the Protein Data Bank. It shows high reliability in CP identification and achieves an AUC of 0.9. SeqCP has been implemented into a web server available at: http://pcnas.life.nthu.edu.tw/SeqCP/.

1. Introduction

Circular permutation (CP) is a phenomenon of protein sequence rearrangement. The amino (N) and carboxyl (C) termini of proteins correlated by CP are in different locations. Two proteins can be considered as circular permutants (CPMs) when the C- and N-termini of each protein can align with the middle of the sequence of the other protein. The cutting site is defined as the CP site. Naturally occurring CPs have been reported to be the products of post-translational modifications [1–5] and genetic events such as deletions after duplication [5–10] and exon shuffling [5,11,12]. For a set of CPMs with the same origin, the folding nuclei, structural stability, substrate diversity, or enzymatic activity may be influenced by CP, but the three-dimensional structures and functions of the CPMs are generally conserved [6,13–17].

CP is useful in protein structural and functional studies [6,10,13–20]. It can also be applied as a bioengineering technique for adjusting the solubility [21], stability [21–23], enzymatic activity [19,21,22], ligand binding affinity [19,23], or functional diversity [22,23] of pro-

Abbreviations: CP, circular permutation; CPMs, circular permutants; CPDB, Circular Permutation Database; CPSARST, Circular Permutation Search Aided by Ramachandran Sequential Transformation; CE, combinatorial extension; CE-CP, CE with Circular Permutations; RMSD, root-mean-square distance; ROC, receiver operating characteristic; AUC, area under the ROC curve; MCC, Matthews correlation coefficient.

* Corresponding authors at: Life Science Building II, Room 306, No. 101, Section 2, Kuang Fu Road, Hsinchu 300044, Taiwan, ROC.

E-mail addresses: chichunchen@iis.sinica.edu.tw (C.-C. Chen), freddy789.bi07g@nctu.edu.tw (Y.-W. Huang), hsuancheng@ym.edu.tw (H.-C. Huang), Wade-Lo@nycu.edu.tw (W.-C. Lo), pclyu@mx.nthu.edu.tw (P.-C. Lyu).

https://doi.org/10.1016/j.csbj.2022.11.024

teins. For example, Nagaratnam *et al.* created fully-active human Taspase1 CPMs for broader applications in anticancer therapeutics [20]. Novel fusion proteins [24], molecular sensors [25,26], or protein splits [27,28] can be created by CP, such as the fluorescent calcium sensor created by Tsien's lab [25]. Despite these powerful applications, trial-and-error is still inevitable in CP-based bioengineering because predicting the structure of circularly-permuted proteins is still challenging [29]. CP may not be infrequent in nature. However, current CP search methods are not efficient enough or are inapplicable for proteins without known structures. If there can be an efficient sequence-based CP search algorithm, many more CP data can be detected from current protein databases. Such data shall be informative for CP-bioengineering [10,30] and protein functional annotations [31], and will help improve CP cutting site and CP structure predictions [29,32].

Since CPMs usually possess similar structures, existing CP search methods preferentially use structural information to identify this phenomenon. Structure-based methods can be classified into domain- and subunit-based algorithms. The first category includes Structural Homology by Environment-Based Alignment (SHEBA) [33] and the Genetic Algorithm for Nonsequential, Gapped protein STructure Alignment (GANGSTA+) [34]. The SHEBA creates a CP site in the middle of the query protein and aligns the permuted query protein against the target protein. The alignment result is used to update the CP site, which in turn, is used in a backward process to refine the alignment. According to the final alignment, if the target protein is structurally similar to the permuted query, the two proteins are considered a pair of CPMs, i.e., a CP pair [35]. The GANGSTA+ aligns secondary structure elements in protein domains to identify specific connections interpretable as evidence of CP [36]. The subunit-based category is represented by Circular Permutation Search Aided by Ramachandran Sequential Transformation (CPSARST) [31] and Combinatorial Extension with Circular Permutations (CE-CP) [37]. In CPSARST, the query protein structure is transformed into a structural text string [38], and both the normal and duplicated query structural strings are searched against a pretransformed structural string database for homologs. For each hit target protein from the database, its structural string alignments with the normal and duplicated query strings are compared to identify the CP site. A normal and a circularly permuted structural alignment based on the CP site are applied to superimpose the query and target structures. If the structural similarity score of the permuted superimposition is higher than the normal score, the query and target proteins are considered a CP pair [31]. CE-CP uses combinatorial extension (CE) [39] to structurally align a duplicated query protein against a target protein. If both copies of the duplicated query protein can be aligned on the target and the gap size is <30 residues, the two proteins are considered a CP pair [37].

Despite the high sensitivity of these methods in detecting CP, they require structures. Since only a small proportion of the proteins in protein databases have determined structures, it is likely that most CPMs have not been identified because of the inapplicability of structure-based methods. In addition, the number of proteins in structure databases has been increasing exponentially, and structure-based CP search methods may be too time-consuming to perform large-scale database searches. For example, CPSARST can scan 53 thousand pairs of polypeptides per minute using one CPU [31], and at the time of this article writing, the Protein Data Bank (PDB) [40,41] contains 110,417 nonredundant polypeptides that form 12,192 million pairs; hence, screening the current PDB with CPSARST will require 3,848 hours, or 160 days. Sequence-based CP search methods, which do not need to perform complicated steric and geometric computations, could be much faster. Moreover, sequence comparison may more reliably determine the evolutionary relationship between identified CP pairs than structure comparison, which may not distinguish homologous proteins, proteins that share a common origin, from analogous proteins, proteins that result from convergent evolution.

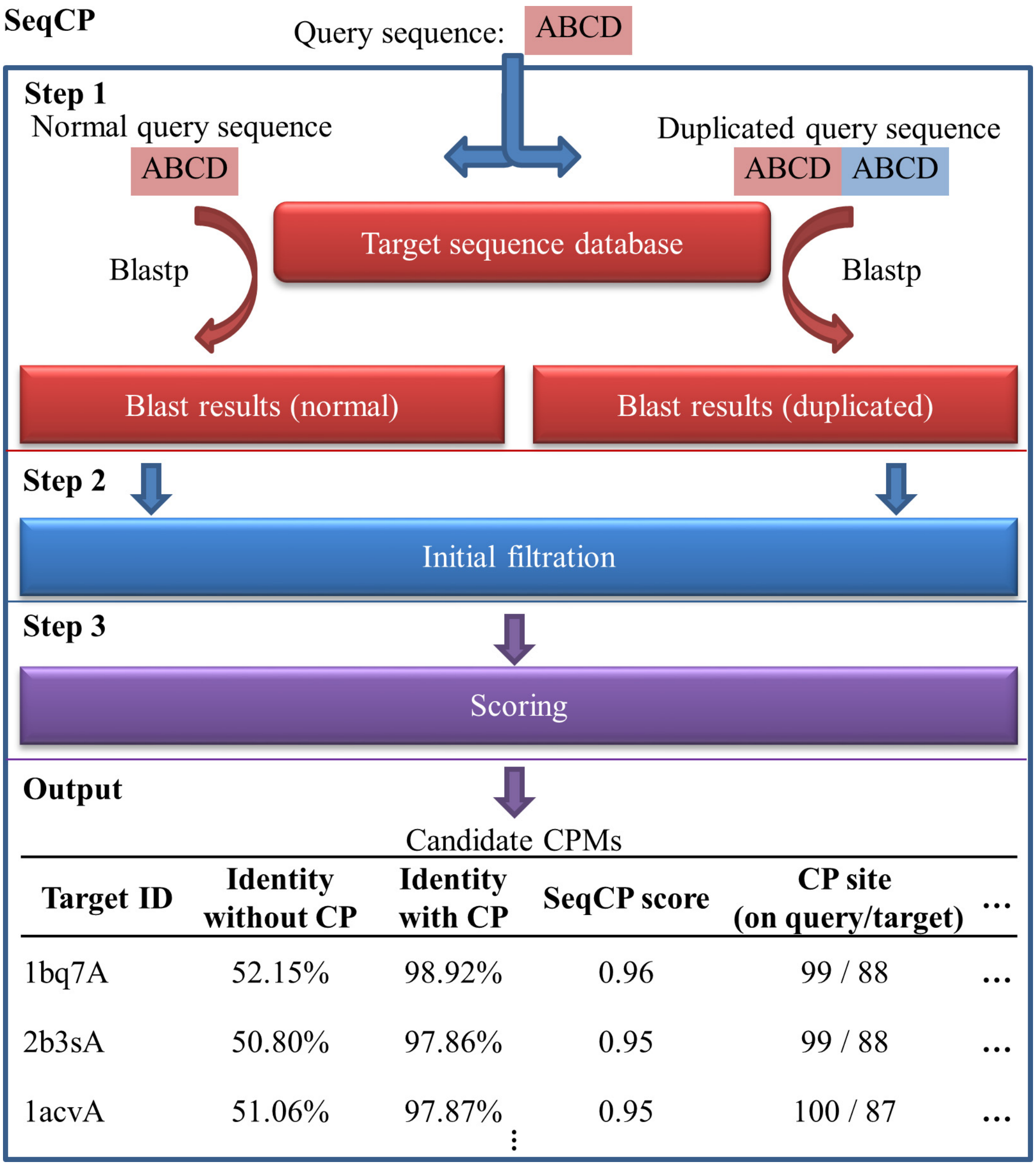
Although sequence-based CP search methods have been available, the last development was in 2005 [42]. In the past two decades, the amount of sequence data has increased so much that these pioneering methods may not be efficient enough to handle the current massive protein sequence databases. The pioneering sequence-based CP search methods include the method developed by Uliel *et al.* [43,44] and a motif-based algorithm developed by Weiner *et al.* [42]. The Uliel algorithm consists of three steps. First, using the Needleman-Wunsch algorithm [45], a duplicated query sequence is aligned against a target sequence, and their edit distance is used to determine whether they are a candidate CP pair [43]. Second, the CP candidacy is validated with the exact algorithm where the query sequence is circularly permuted at all residue positions to align against the target sequence. Third, if the CP candidacy is sufficiently high, a dot matrix is generated for manual inspection to determine the final candidacy. The motif-based method developed by Weiner also applies the Needleman-Wunsch algorithm. The query and target sequences are reduced to strings of domain identifiers with the ProDom method [46]. After the duplicated query and target domain strings are aligned, if the alignment shows CP-specific features, the proteins are identified as a CP pair [42]. Compared with the sensitivity of structure-based CP detection methods, which estimate that 13 % or more proteins have CPMs, the performance of these sequence-based methods seems low, as they showed that < 0.2 % of proteins possess CPMs [31,35–37,42–44]. Therefore, an efficient sequence-based CP search method is still needed.

In this project, we aimed to develop an efficient sequence-based CP search method, and the result of our efforts is the proposed SeqCP algorithm. SeqCP aligns the normal and duplicated query sequences against a target protein database and analyzes the alignments to identify CPMs. Our data show that SeqCP is reliable for identifying CP pairs and determining the positions of CP sites. We have also illustrated that SeqCP is promising for identifying complete templates for the structural modeling of circularly permuted proteins. The advantages of SeqCP over structure-based CP detection methods include the independence of known structures and a high speed, e.g., 33 times faster than CPSARST, the fastest structure-based method [31]. SeqCP would also outperform conventional sequence-based methods in speed since it detects the CP relationship between sequences with only two alignments instead of the exhaustive alignments as applied in Uliel's algorithm. Theoretically, SeqCP should have a higher sensitivity in CP detection than Weiner's method because it does not rely on accurate motif/domain annotations. The SeqCP algorithm has been implemented into a web server available at http://pcnas.life.nthu.edu.tw/SeqCP/.

2. Materials and methods

2.1. Overview of the proposed method

SeqCP uses four steps to search for the CPMs of a query protein against the user-specified target sequence database (Fig. 1). The target database can be any collection of protein sequences such as a UniRef dataset [47,48] and the NCBI protein sequence dataset [49], as long as the sequences are stored in a FASTA format text file [50,51]. Step 1, screening: SeqCP uses the normal and duplicated query protein sequences to align against the target database and retrieves a list of candidate CPMs of the query protein. Only the target proteins that can be aligned with both the normal and duplicated query sequences are considered candidate CPMs. Step 2,



Target ID	Identity without CP	Identity with CP	SeqCP score	CP site (on query/target)	…
1bq7A	52.15%	98.92%	0.96	99 / 88	…
2b3sA	50.80%	97.86%	0.95	99 / 88	…
1acvA	51.06%	97.87%	0.95	100 / 87	…
		⋮			

Fig. 1. Flowchart of SeqCP. SeqCP aligns the normal and duplicated query protein sequences against a target sequence database. The alignments are used to detect candidate CPMs. For each candidate CPM, the normal and duplicated query sequence alignments are compared to compute the SeqCP scores for the final filtration, after which a list of CPMs is produced with determined CP sites and pairwise CP sequence identities.

initial filtration: SeqCP determines candidate CPMs of the query protein based on 1) length differences between the query and target sequences, 2) the locations of CP sites on the target proteins computed according to the alignment between the duplicated query and target sequences, 3) the boundary shift of alignments between the query and target sequences, and 4) the identity difference between the aligned CP fragments. Step 3, scoring: For each candidate CPM passing the initial filter, the SeqCP score is calculated based on the alignments of normal and duplicated query sequences. Step 4, output: SeqCP makes a list of the identified CPMs in order of descending SeqCP score along with a variety of additional information, including the locations of CP sites and circularly permuted sequence identities.

2.2. Experimental datasets

2.2.1. CPDB dataset

This dataset was directly obtained from the CPDB (Circular Permutation Database) [52], which was constructed by automatic CPSARST searches and manual verifications based on the nonredundant PDB of 2007 [31,52]. The CPDB dataset contained 2,238 polypeptides, constituting 4,169 nonredundant CP pairs (see File S1). CP pairs in this dataset were classified as high-quality and marginal CP pairs based on the procedure described in Subsection 2.5. The discriminatory features between high-quality and marginal CP pairs applied to formulate the SeqCP score were identified by statistical analysis of this dataset, meaning that this CPDB dataset was the training set of the SeqCP method.

2.2.2. NrPDB100-2007

This nonredundant PDB 2007 dataset contains 26,349 polypeptides sharing < 100 % sequence identities. The original CPDB was established using this dataset. Therefore, after the SeqCP score was formulated, we performed all-against-all SeqCP searches on this NrPDB100-2007 dataset to examine how many CPDB-registered CP pairs could be identified by SeqCP with a series of discriminatory score thresholds. In other words, the NrPDB100-

2007 was the testing set for determining the optimal discriminatory threshold of the SeqCP score.

2.2.3. NrPDB100-2019

This 100 % sequence identity nonredundant PDB dataset with 85,725 polypeptides was downloaded from the PDB server [40] on July 25, 2019. It was constructed using the BLASTClust algorithm [53] to cluster polypeptide entities possessing ≥ 20 amino acids. This dataset was utilized to perform large-scale tests of the SeqCP algorithm.

2.2.4. Manipulation of the PDB files

Protein structure files were downloaded from the PDB. Records other than ATOM and HETATM in each file were removed. The HETATM records not describing amino acid residues, such as metal ions, were also eliminated. The amino acid sequences of the protein structures were extracted from the manipulated PDB files.

2.3. Determination of the candidate CP site for a pair of sequences

Candidate CP sites were determined based on the alignment between the duplicated query and target sequences (Fig. 2). The CP site on the query sequence was computed as the position at which the first copy of the query started to align with the target. The CP site on the target sequence was determined as the position where the second copy of the query was initially aligned. The formulas for these two CP sites are listed below,

$$\text{CP site, query} = \alpha^D_{qq'} - \alpha^D_t + 1 \tag{1.1}$$

$$\text{CP site, target} = \sigma^D_t \tag{1.2}$$

where the superscript *D* indicates that the alignment is a duplicated query sequence alignment, $\alpha^D_{qq'}$ is the position on the duplicated query sequence (*qq'*) with which the target sequence (*t*) was initially aligned, α^D_t is the position on *t* with which *qq'* was initially aligned, and σ^D_t is the position on *t* with which the second copy of *qq'* was initially aligned (Fig. 2).

In addition to CP sites, the permutation size and the alignment gap were also calculated based on the duplicated query sequence alignment. The permutation size is the size of the smaller aligned fragment on *t*. The alignment gap (GR_{cp}) is the size of the unaligned fragment on *t* between *t*'s alignments with the two copies of *qq'*. The formulas are defined below,

$$\text{Permutation size} = \min(\rho^D_t - \alpha^D_t,\ \omega^D_t - \sigma^D_t) \tag{2}$$

$$GR_{\text{cp}} = \sigma^D_t - \rho^D_t - 1 \tag{3}$$

where ρ^D_t and ω^D_t are the positions on *t* with which the first and second copies of *qq'*, respectively, ended in alignment (Fig. 2).

2.4. Assessment of the precision of CP site determination using in silico CP sequences

The *in silico* random CP dataset prepared in [31] was utilized to assess the precision of the CP site determination of SeqCP. This random CP dataset contained 100 protein sequences randomly selected from the PDB. These sequences were considered as the native (i.e., parental) sequences of the circular permutants. For each native sequence, 99 descendant sequences sharing identities ranging from 99 %, 98 % …, to 1 % with it were generated by *in silico* mutations (insertions, deletions, and sub-

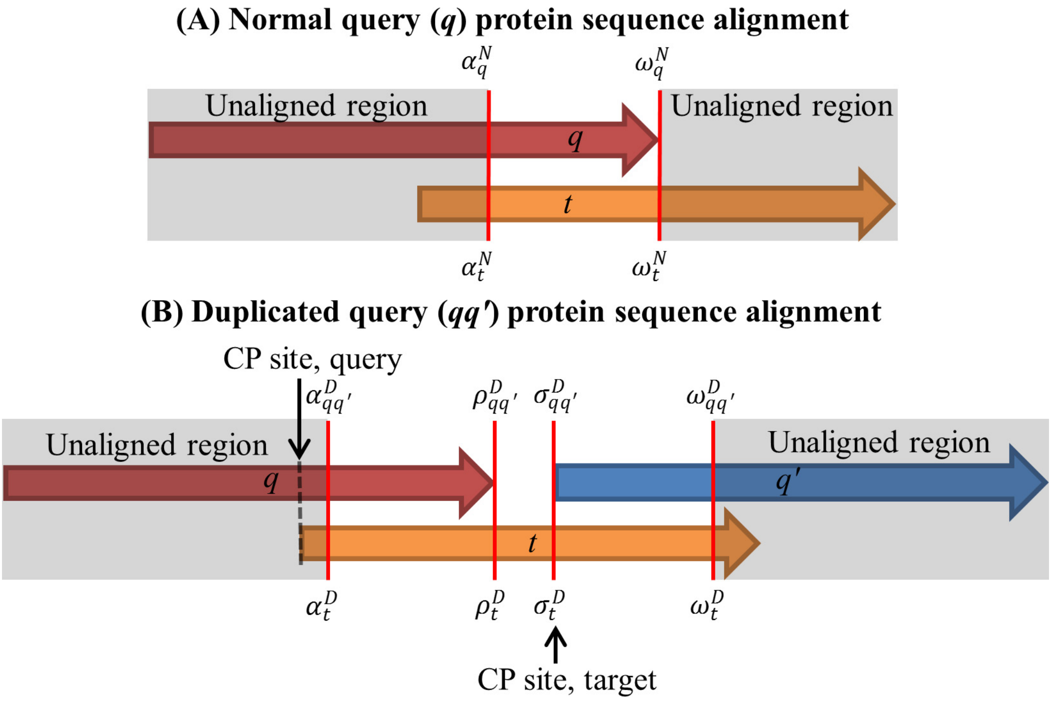


Fig. 2. Demonstration of the normal and duplicated query sequence alignments. In the SeqCP algorithm, a target protein (*t*) is aligned with the normal query (*q*) and duplicated query (*qq'*) sequences. (A) The first aligned residues between *q* and *t* are represented by α^N_q and α^N_t, and the last aligned residues are represented by ω^N_q and ω^N_t, where the superscript *N* indicates that this alignment is based on the normal query sequence. (B) The black arrows indicate the CP sites. The first pair of aligned residues between the proteins is represented by $\alpha^D_{qq'}$ and α^D_t, where *D* indicates a duplicated query sequence alignment. The last aligned residue of the first copy query sequence is $\rho^D_{qq'}$, which is aligned with ρ^D_t on the target sequence. Residue $\sigma^D_{qq'}$ represents the first residue of the second copy query sequence aligned with the target at σ^D_t. Residues $\omega^D_{qq'}$ and ω^D_t represent the last aligned pair of residues between the two proteins.

stitutions), and 99 descendant sequences sharing 99 % to 1 % decreasing sequence similarities with it were also generated. CP was introduced into each mutant before the regular mutations. By analyzing the sequence of a CP descendant and the sequence of its native protein, it was possible to determine the location at which CP was introduced (i.e., the CP site). In [31], this determination was implemented by pairwise sequence alignments using the bl2seq program of the classic BLAST [54]. In this work, the sequence alignment engine was the database search program blastp of BLAST [53]. We took advantage of the database search function of BLAST to achieve improved CP site determination. First, the 100 known native sequences were collected into a target dataset. Second, for each CP descendant, its sequence was used to make a normal query for blastp to search against the target dataset for homologs, and a duplicated query was then made. Next, from the blastp output for the normal query, the alignment between the query and its known native protein was retrieved; the same process was performed with the duplicated query. Finally, the CP site was determined as stated in the previous subsection. The parameter settings of blastp used in this procedure are described as follows: word size: 2, E-value cutoff: 1.0e + 300, gap open penalty: 10, gap extend penalty: 3, and initial scoring matrix: BLOSUM45. BLOSUM45 is more capable of aligning protein sequences sharing low identities (as low as ~ 45 % identities) than BLOSUM62 or higher identity matrices. The E-value cutoff was set as a very high value accepted by blastp to ensure that the sequence alignments could be fully extended. Other parameters were set empirically according to the experimental results of the CPSARST paper [31].

2.5. CE-CP-based standard CP identification procedure

After a pair of proteins, denoted by q (query) and t (target), were aligned with CE-CP [37], the following eight filtering criteria were applied to define whether they were a high-quality CP pair.

a. The average length of q and t is $\geq$ 100 residues, and the ratio between their lengths, defined below (r_l), is $\geq$ 0.7.

$$r_l = \frac{\min(l_q,\ l_t)}{\max(l_q,\ l_t)} \tag{4}$$

where l_q and l_t represent the sequence length (number of residues) of the query and target protein, respectively.

b. The CE-CP program can be successfully applied without error messages.
c. The number of alignment blocks (CP blocks) determined by the CE-CP is 2.
d. According to Eq (2), the permutation size is > 20 % of the length of the larger protein between q and t.
e. The GR_{cp} is < 30 residues (refer to Eq (3)).
f. The span of alignment (the number of aligned residues plus gap positions) was $\geq$ 70 % of the larger protein, and the number of gap positions was < 25 % of the alignment span.
g. Based on the CE-CP alignment, the initial root-mean-square-distance (RMSD) between q and t is $\leq$ 5 Å.
h. Manual inspection indicates that a high-quality CP pair does not contain a large, poorly superimposed region.

2.6. Discriminatory features between high-quality and marginal CP pairs

Discriminatory features were designed to differentiate CP pairs of different qualities. For instance, the average sequence length, the length ratio between query and target sequences, and the factors defined in this subsection were applied as discriminatory features. While computing these features, the parameters of blastp were set as listed in *2.3*. The working E-value cutoff applied to the SeqCP search process is illustrated in Fig. S1.

2.6.1. Discriminatory features of the initial filtration stage

SeqCP takes advantage of the difference between the normal query (q) and duplicated query (qq') in sequence alignments with the target protein (t) to identify the CP relationship between q and t. As demonstrated below, most discriminatory features are derived from the sequence alignments (see Fig. 2 for a graphical representation of the symbols, such as α^D and ω^D, shown in the following formulas). The feature

$$RSB_{qq'} = \frac{\min\left(\left|\alpha^D_{qq'} - \alpha^D_t\right|, \left|\omega^D_{qq'} - \omega^D_t\right|, \left|\omega^D_{qq'} - \omega^D_t - l_q\right|, \left|\alpha^D_{qq'} - \alpha^D_t - l_q\right|\right)}{\max\left(l_q, l_t\right)} \tag{5}$$

where l represents the sequence length, α^D denotes the start position of the duplicated query alignment in the sequence of qq' or t, and ω^D denotes the end position of the duplicated query alignment. This feature describes how much the termini of two proteins are shifted in the duplicated query alignment. $RSB_{qq'}$ represents the relative shift of aligned boundaries between qq' and t. The feature

$$RSB_q = \frac{\min\left(\left|\alpha^N_q - \alpha^N_t\right|, \left|\omega^N_q - \omega^N_t\right|\right)}{\max\left(l_q, l_t\right)} \tag{6}$$

where α^N represents the start position of the normal query alignment in the sequence of q or t, and ω^N denotes the end position of the normal query alignment. This feature describes how much the termini of two proteins are shifted in the normal query alignment. RSB_q represents the relative shift of aligned boundaries between q and t. The identity difference between alignments is calculated as follows,

$$\mathrm{IDA} = \left|\frac{n_{iden,q}}{n_{align,q}} - \frac{n_{iden,qq'}}{n_{align,qq'}}\right| \tag{7}$$

where n_{align} and n_{ident} represent the total number of aligned residue pairs and the number of aligned residue pairs that are identical amino acids, respectively. For a pair of evolutionarily related CPMs, the identities of two CP fragments are supposed to be similar; thus, the difference in identity between the normal query and duplicated query alignments should be small. The maximal proportion of flanks, a feature that determines whether the target sequence is globally aligned in the duplicated query alignment, is calculated as follows:

$$\text{Maximal proportion of flanks} = \frac{\max\left(\alpha^D_t - 1, l_t - \omega^D_t\right)}{\max\left(l_q, l_t\right)} \tag{8}$$

To ensure that the alignment results between normal and duplicated query alignments are stable, the alignment boundary bias is calculated as follows:

$$\text{Alignment boundary bias} = \max\left(\mathrm{QFB}, \mathrm{TFB}, \ \mathrm{QEB}, \mathrm{TEB}\right) \tag{9.1}$$

when α^N_q < 20 % × l_q:

$$\text{Query front boundary bias (QFB)} = \frac{\left|l_q - \alpha^D_{qq'} - \left(\alpha^N_t - \alpha^D_t - \alpha^N_q + 1\right)\right|}{\max\left(l_q, l_t\right)} \tag{9.2}$$

$$\text{Target front boundary bias (TFB)} = 0 \tag{9.3}$$

$$\text{Query end boundary bias (QEB)} = \frac{\left|\omega_q^N + l_q - \omega_{qq\prime}^D\right|}{\max\left(l_q, l_t\right)} \tag{9.4}$$

$$\text{Target end boundary bias (TEB)} = \frac{\left|\omega_t^N - \omega_t^D\right|}{\max\left(l_q, l_t\right)} \tag{9.5}$$

else:

$$\text{Query front boundary bias (QFB)} = \frac{\left|\alpha_q^N - \alpha_{qq\prime}^D\right|}{\max\left(l_q, l_t\right)} \tag{9.6}$$

$$\text{Target front boundary bias (TFB)} = \frac{\left|\alpha_t^N - \alpha_t^D\right|}{\max\left(l_q, l_t\right)} \tag{9.7}$$

$$\text{Query end boundary bias (QEB)} = \frac{\left|\left(\omega_t^D - \omega_t^N + \omega_q^N\right) - \omega_{qq\prime}^D\right|}{\max\left(l_q, l_t\right)} \tag{9.8}$$

$$\text{Target end boundary bias (TEB)} = 0 \tag{9.9}$$

2.6.2. Discriminatory features of the scoring stage

The discriminatory features used to formulate the SeqCP score are as follows,

$$\text{Alignment rate of } qq\prime,\ AR_{qq\prime} = \frac{n_{align,qq\prime}}{\max\left(l_q, l_t\right)} \tag{10}$$

$$\text{Gap rate of } qq\prime,\ GR_{qq\prime} = \frac{n_{gap,qq\prime}}{\max\left(l_q, l_t\right)} \tag{11}$$

$$\text{Sequence similarity of } qq\prime,\ Simi_{qq\prime} = \frac{n_{simi,qq\prime}}{\max\left(l_q, l_t\right)} \tag{12}$$

where n_{gap} and n_{simi} represent the number of gaps (internal unaligned residues) and the number of aligned residue pairs that are physiochemically similar amino acids, respectively. These features describe the alignment quality of the duplicated query.

2.6.3. Scoring stage quality control factors

The fundamental goal of SeqCP is to detect the differences between normal and duplicated query alignments. High-quality CP pairs usually had improved alignment scores (e.g., sequence identities) over marginal or non-CP pairs. We further inferred that for a pair of high-quality CP proteins, the improvements in various alignment scores should be stable, meaning that the two aligned regions (before and after the CP site) obtained by the duplicated query alignment should be equally homologous. We thus designed a quality control factor (QCF) based on several alignment-improvement features. The formulas of QCF and its components are as follows,

$$\text{Improvement rate of sequence identity},\ IR_{Iden} = \frac{n_{iden,qq\prime}}{n_{iden,q}} \tag{13-1}$$

$$\text{Improvement rate of sequence similarity},\ IR_{Simi} = \frac{n_{simi,qq\prime}}{n_{simi,q}} \tag{13-2}$$

$$\text{Improvement rate of alignment rate},\ IR_{AR} = \frac{n_{align,qq\prime}}{n_{align,q}} \tag{13-3}$$

$$\text{Averaged improvement rate},\ \bar{IR} = \frac{IR_{Iden} + IR_{Simi} + IR_{AR}}{3} \tag{13-4}$$

$$QCF = \sqrt{\frac{\left(IR_{Iden} - \bar{IR}\right)^2 + \left(IR_{Simi} - \bar{IR}\right)^2 + \left(IR_{AR} - \bar{IR}\right)^2}{2}} \tag{14}$$

where the abbreviations have all been described in previous subsections. The *QCF* is applied in the equation of the SeqCP score (see 2.7.3) to magnify the difference between high-quality and marginal CP pairs.

2.7. Criteria for identifying CP pairs in the three stages

2.7.1. Screening

The normal and duplicated query sequences are applied to the target database search for CPMs of the query. Only the target sequences that align better with the duplicated query than with the normal query are considered as candidates.

2.7.2. Initial filtration

To eliminate candidates with poor CP quality, nine criteria are applied. 1) The average length of the query and target sequences is $\geq$ 100. 2) The length difference of the protein pair, computed as the ratio of the length between the smaller and larger proteins, is $\geq$ 0.8; see Eq(4). 3) The CP site is identified by duplicated query sequence alignments. 4 & 5) The minimum shifts of the boundaries of alignment between the query and target sequences are both $\geq$ 20 %; see Eqs (5 & 6). 6) The identity difference between normal and duplicated query alignments is $\leq$ 10 %; see Eq (7). 7) The maximal proportion of alignment flanks of the target sequence in the duplicated query alignments is < 0.2; see Eq (8). 8) The alignment boundary bias between normal and duplicated query alignments is $\leq$ 0.2; see Eq (9). 9) The duplicated query sequence alignment rate, that is, the number of aligned residues in the duplicated query sequence divided by the length of the larger protein is $\geq$ 0.8; see Eq (10).

2.7.3. Scoring

Seven features are utilized to compute the SeqCP score. 1) the alignment rate of the duplicated query; see Eq (10); 2) the number of gaps in the duplicated query sequence alignment divided by the length of the larger protein; see Eq (11); 3) sequence similarities computed based on the duplicated query sequence alignment; see Eq (12); 4–6) the improvement rates of sequence identity, similarity, and alignment rate between query and target proteins when the duplicated query sequence alignment was compared with the normal sequence alignment; see Eq (13); and, 7) a quality control factor that ensured the stability of duplicated sequence alignments; see Eqs (14). Finally, the SeqCP score is defined below,

$$\text{SeqCP score} = \frac{Simi_{qq\prime} \times \left(\bar{IR} - \mu \times QCF\right) + AR_{qq\prime} - GR_{qq\prime}}{\rho} \tag{15}$$

where $AR_{qq\prime}$, $GR_{qq\prime}$, $Simi_{qq\prime}$, $\bar{IR}$, and *QCF* are described in Eqs (10)–(14), and μ and ρ are empirical constants applied to adjust feature weights and could be tuned based on known data. These constants were trained with the CPDB dataset by observing the distributions of feature values, SeqCP scores and ROC curve analysis. The values were determined to be μ = 2 and ρ = 3. Based on **Results** subsection 3.4., this SeqCP score setup showed stable performance for datasets of different sizes. A detailed explanation of the meaning of the SeqCP score is provided in the **Discussion**.

As we designed the scoring scheme, making the score proportional to the chance of a protein pair being a CP pair was a major concern. To assess whether this concern was addressed, we also formulated a confidence factor to describe at each SeqCP score level (e.g., 0.30–0.35 or 0.80–0.85) the expected proportion of correctly-determined high-quality CP pairs, as shown below,

$$\text{Confidence} = \frac{N_{R\_TP}}{N_R} \tag{16}$$

where N_R is the number of protein pairs at a specific SeqCP score level and N_{R_TP} is the number of correctly-determined high-quality CP pairs at the same level. Fig. S2 demonstrates that the confidence factor is highly correlated with the SeqCP score, indicating that when a pair of proteins receives a higher SeqCP score, it would be more likely a high-quality CP pair.

2.8. Statistical analysis

Receiver operating characteristic (ROC) curve analyses and area under the ROC curve (AUC) computation were performed with the ROCR package [55]. Five statistical measures, including sensitivity, precision, specificity, accuracy, and the Matthews correlation coefficient (MCC), for determining the quality of a binary classification system were also calculated to quantify the performance of SeqCP, as provided below,

$$\text{Sensitivity} = \frac{TP}{P} \tag{17}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{18}$$

$$\text{Specificity} = \frac{TN}{N} \tag{19}$$

$$\text{Accuracy} = \frac{TP + TN}{P + N} \tag{20}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{21}$$

where T and F denote true/correct and false/incorrect classifications, P and N are the numbers of known positive (high-quality CP pair) and negative (otherwise) cases, respectively.

3. Results

3.1. Evaluation of CP site determination with a simulated CP dataset

SeqCP can determine the CP site between proteins; in fact, whether a CP site can be determined is a critical factor considered in the initial filtration stage of its algorithm. To assess whether SeqCP can correctly determine the positions of CP sites, an *in silico* random CP dataset (see Subsection 2.4) obtained from the CPSARST paper [31] was utilized in this experiment. This dataset contains 100 native protein sequences, each mutated *in silico* into 200 CPMs of decreasing sequence identities and sequence similarities. There are 20,000 CPMs in total. As shown in Fig. 3, when the sequence identities of CPMs were > 20 %, SeqCP could determine the CP sites as precisely as, or even more precisely than, CPSARST. SeqCP identified the exact CP sites on the target proteins in at least 80 % of the retrieved CP pairs when the identity level was higher than 20 % or the similarity was higher than 40 %.

3.2. CP pairs in the Circular permutation database (CPDB)

The CPDB [52] is currently the largest CP database. To train and test the SeqCP algorithm, we used the CPDB as the source dataset. However, the CPDB was constructed with the structure-based CP search method CPSARST, which may be incapable of distinguishing CPMs from some internally symmetric proteins or distinguishing true CPMs from evolutionarily coincident CPMs [37]. Rost suggested that if a pair of structurally similar proteins share < 15 % sequence identity, they may have arisen from either divergent or convergent evolution [56]. To qualify the CP pairs obtained from the CPDB, we compared the distributions of structure-based sequence identities of CPDB CP pairs and random protein pairs. The structure-based sequence identities were computed with CE-CP [37], a CP detection algorithm now officially utilized by the PDB [40]. The results showed that a proportion of the CP pairs in the CPDB exhibited indistinguishable sequence identities from the random alignments between proteins, implying that some CP pairs in the CPDB were not significantly evolutionarily related (see Fig. 4A).

To discriminate CP pairs with high structural similarity or evolutionary relatedness from those with low similarity or relatedness, a series of structural similarity criteria computed by CE-CP were applied as filters (see **Materials and methods**). The CP pairs with high structural similarity or evolutionary relatedness were termed the high-quality, and the other pairs were termed marginal CP pairs in this report. This filtering pipeline was defined as the CE-CP-based standard CP identification procedure, which was the foundation for identifying high-quality CP pairs from all the datasets used in this work. Ultimately, 480 CPDB CP pairs were identified as high-quality pairs (Table 1). Comparisons of the distributions of structure-based sequence identities of the high-quality CP pairs, marginal CP pairs, and random protein pairs revealed that the composition of the high-quality CP pairs was very different from those of the other two types of pairs (Fig. 4B). These high-quality and marginal CP pairs were used as the positive and negative cases (see File S1 for a list of these cases) for training SeqCP in the following experiments.

3.3. Feature selection for identification of the CP relationship between proteins

To find discriminatory features applicable for distinguishing high-quality CP pairs (the positive group) from marginal CP pairs (the negative group), we observed the distributions of several sequence properties and alignment measures between these groups. Among various features, 15 were utilized in two different stages of the SeqCP algorithm, including 1) the average length of the protein pair (Fig. 5A); 2) r_l, the length difference of the protein pair (Eq (4), Fig. 5B); 3) the CP site that could be identified; 4 & 5) RSB, relative shift of aligned boundaries (Eqs (5), (6), Fig. 5C, D); 6) IDA, identity difference between normal and duplicated query alignments (Eq (7)); 7 & 8) features related to alignment stability (Eqs (8), (9)); 9) $AR_{qq'}$, alignment rate of a duplicated query alignment (Eq (10), Fig. 5E); 10) $GR_{qq'}$, gap rate of a duplicated query alignment (Eq (11), Fig. 5F); 11) $Simi_{qq'}$, sequence similarity of a duplicated query alignment (Eq (12), Fig. 5G); 12–14) the improvement rates of sequence identity, similarity, and alignment rate between normal and duplicated query alignments (Eq (13), Fig. 5H–J); and 15) QCF, quality control factor (Eqs (14)). See **Materials and methods** for more details about these features. After the scoring stage, every candidate CP pair was assigned a SeqCP score.

We first applied these discriminatory features with the thresholds determined based on their distributions to process the CP pairs obtained from the CPDB. In the previous subsection, the CPDB dataset was divided into two groups: high-quality and marginal CP pairs. We reasoned that if these features can efficiently discriminate between the two groups, high-quality pairs should be retained in large quantities, and marginal pairs should be removed after applying them as filters. Table 2 shows the feasibility of these features. Among the 480 high-quality and 3,689 marginal CP pairs, 480 and 2,555 pairs, respectively, passed the SeqCP screening stage, which checked only whether the pairs could produce both the normal and duplicate query sequence alignments. The nine fundamental features applied in the initial filtration stage then effi-

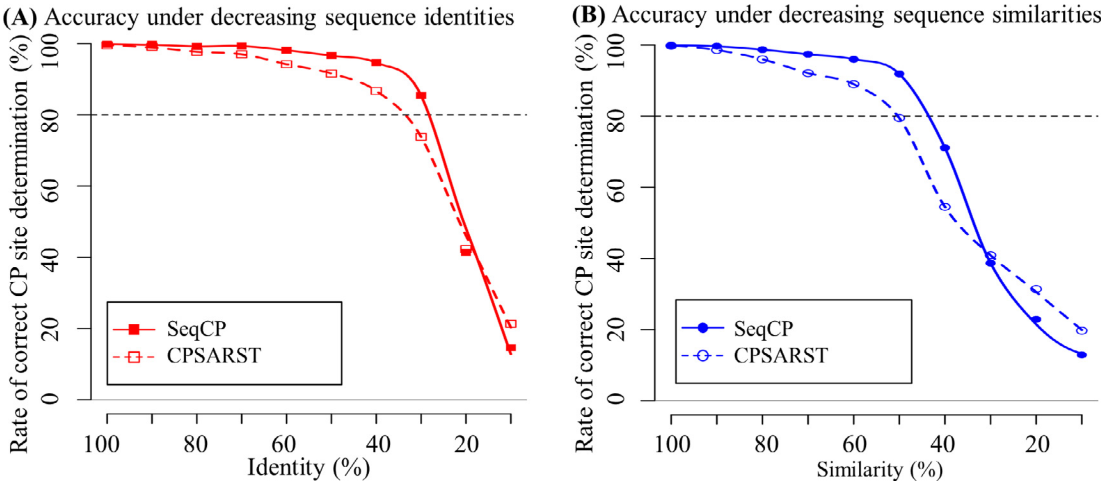


Fig. 3. Accuracy of CP site determination of SeqCP and CPSARST. An *in silico* CP dataset prepared by [31], in which the CP site, sequence identity, and sequence similarity for each CP pair were known, was utilized to assess the performance of SeqCP and CPSARST in CP site determination. (A) Accuracy under decreasing sequence identities. (B) Accuracy under decreasing sequence similarities. The horizontal axis indicates the identity or similarity of CP pairs (10 represents for 0–10 %, and 20 represents for 10–20 %). The vertical axis is the accuracy, defined as the number of CP pairs for which the CP sites were correctly determined divided by the total number of CP pairs. For CP pairs with identities > 20 % or similarities > 30 %, SeqCP was more accurate than CPSARST. The identity between sequences is defined as the number of identical residues that were aligned divided by the total number of aligned residues, whereas the similarity is the number of similar residues that were aligned divided by the total number of aligned residues. Whether two residues were similar was determined according to whether their amino acid substitution score in the BLOSUM45 matrix was > 0.

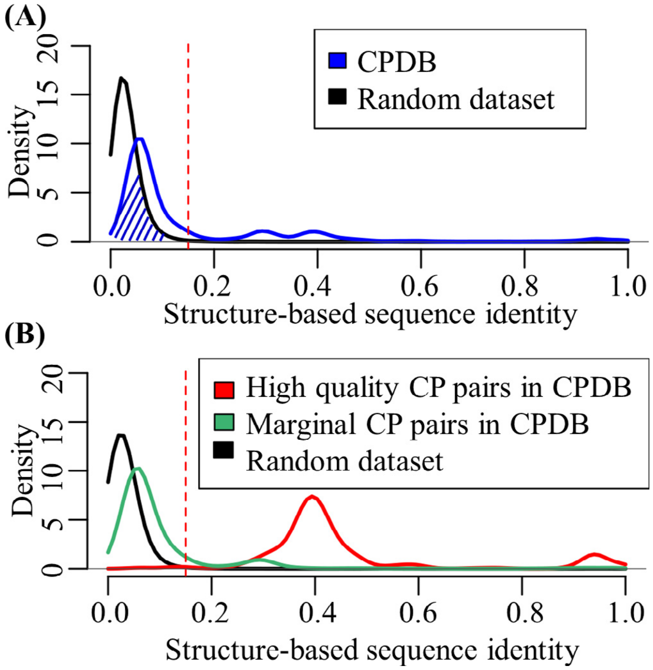


Fig. 4. Distributions of structure-based sequence identities in several protein pair datasets. (A) CPDB vs random protein pairs. There are 4,169 CP pairs in the CPDB. The distribution of their structure alignment-based sequence identities computed by CE-CP is shown in this plot. The distribution of CE-CP identities of a dataset containing 4,169 pairs of proteins randomly selected from the NrPDB100-2007 is also shown. The dotted line indicates the 15 % identity cutoff, below which the pairs may be examples of convergent evolution [56]. (B) High-quality vs marginal CP pairs of the CPDB. After applying several structural similarity criteria computed by CE-CP [37], 480 pairs from the CPDB were identified as high-quality CP pairs. The CE-CP structure-based sequence identity distribution of these high-quality CP pairs significantly differed from that of marginal CP or random pairs. Most high-quality CP pairs possessed identities > 15 %.

ciently removed marginal CP cases with little influence on the high-quality CP group. After these two preprocessing steps, only a small fraction of marginal CP pairs entered the scoring stage. The staged strategy of SeqCP markedly reduced the computation time.

3.4. Identification of the CP relationship between proteins based on the SeqCP score

Among the features described above, the last six (Fig. 5E–J) were utilized to define the SeqCP score for the final determination of the CP relationship between proteins. The formulas of the SeqCP score are provided in the **Materials and methods**, Eq (15). To determine whether this score was promising for detecting CP relationships, we performed an all-against-all database survey using NrPDB100-2007 (26,349 proteins), the source dataset of the CPDB [52]. Every sequence of NrPDB100-2007 was subjected to the SeqCP pipeline, and NrPDB100-2007 itself was searched for candidate CP partners. The screening and initial filtration steps removed most irrelevant protein pairs (the original number of pairs: 6.94×10^8) and retrieved 726 candidate CP pairs. The SeqCP scores between all candidate CP pairs were then computed. After this all-against-all SeqCP computation, all candidate CP pairs were classified as high-quality and marginal CP pairs by the CE-CP-based standard CP identification procedure. From the 726 candidate pairs, 468 were identified as high-quality CP pairs, among which 463 were recorded in the CPDB and had been manually verified to be high-quality CP pairs as stated above. The distributions of SeqCP scores in the high-quality and marginal CP pair groups were then observed. As Fig. 6 shows, the distributions of the two groups were very different, indicating that the SeqCP score is suitable for discriminating them. The ROC curve generated according to these SeqCP score distributions further supported the feasibility of the SeqCP score. The AUC was 0.89.

From this large number of protein pairs, only 1.05×10^{-4}% (i.e., $726/(6.94 \times 10^8)$) were recognized by SeqCP as candidate CP pairs. From the candidates, a relatively large proportion (64.46 %, or 468/726) were ultimately identified as high-quality CP pairs, which constituted 96.46 % (463/480) of all high-quality CP pairs in the CPDB (see Table 1 for the CPDB data and Table S1 for the detailed data from this experiment). Given these findings, SeqCP is promising for detecting CP pairs from an extensive database.

In a practical sequence-based CP database search, no structural data would be available. The SeqCP score is computed purely based on sequence information. Its discriminatory distributions in the positive and negative CP groups and high AUC form a sound basis

Table 1
The number of CP pairs after applying a series of structural similarity criteria for filtering.

Criteria	CPDB, all pairs	CPDB, identity ≥ 15 %[a]	CPDB, identity < 15 %[a]	Random pairs
Original size	4,169	N/A	N/A	4,169
Pair lengths[b]	3,042	N/A	N/A	1,065
Having CE-CP results	3,034	672	2,362	1,060
Number of blocks[c]	2,866	669	2,197	662
CP site locations	1,462	603	859	45
Gap size between CP sites	1,269	487	782	43
Structure alignment length	681	485	196	0
RMSD cutoff	638	485	153	0
Manual inspection	480	471	9	0

See File S1 for the high-quality and marginal CP pairs identified from the CPDB dataset.

[a] Calculated based on CE-CP structure alignments.

[b] Thresholds of protein size and the size difference between proteins (see subsection "CE-CP-based standard CP identification procedure" in **Materials and methods**).

[c] Number of alignment blocks identified by CE-CP.

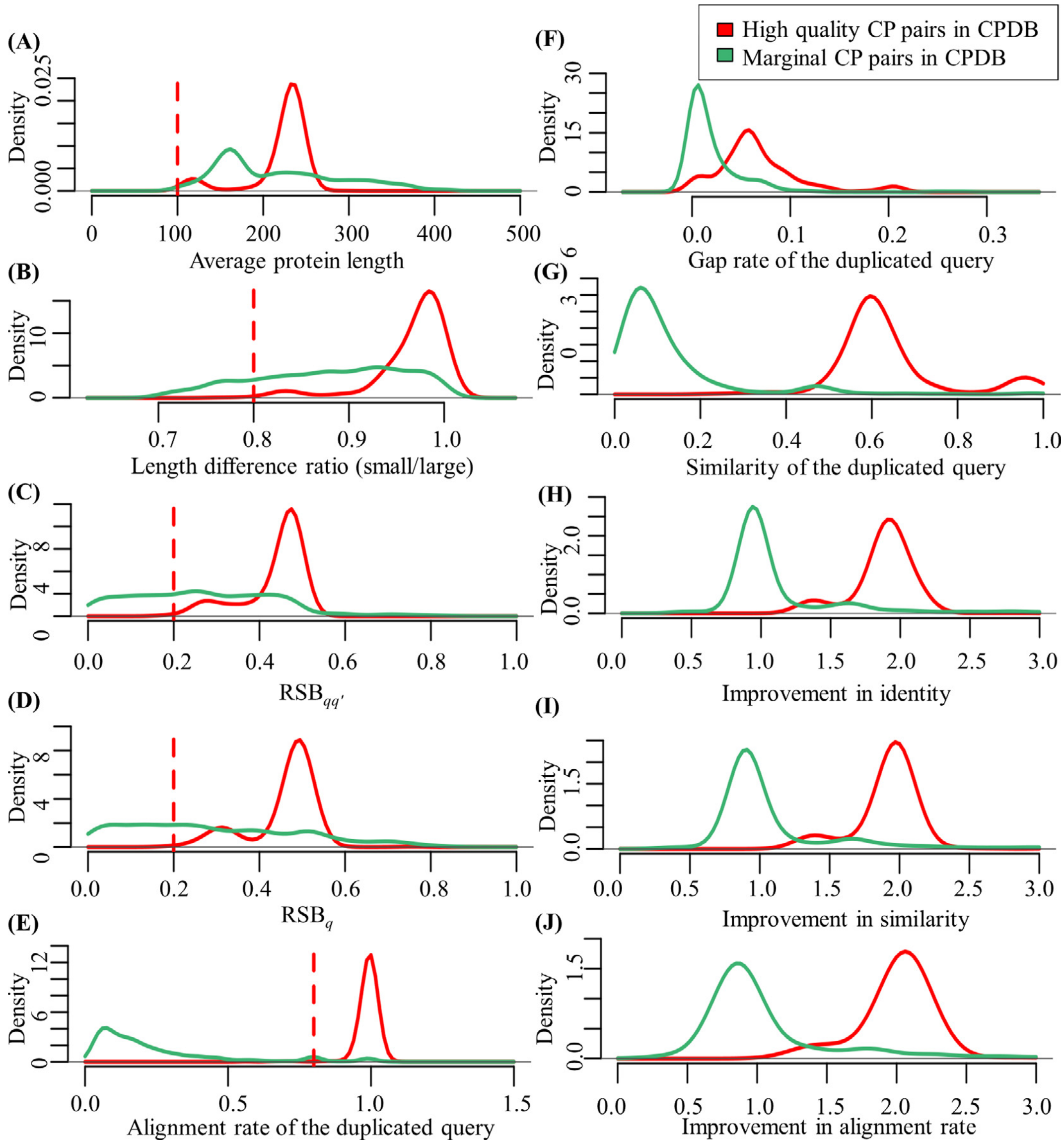


Fig. 5. Distribution of the selected feature values of the high-quality and marginal CP pairs in the CPDB. The sequence properties or measures applied as filters for SeqCP included (A) the average length of the proteins, (B) the ratio of the smaller protein length to the larger protein length (see Eq (4) in **Materials and methods**), (C) the minimum shift of the boundary of alignment between the duplicated query and target sequences, Eq (5), (D) the minimum shift of the boundary of alignment between the normal query and target sequences, Eq (6), (E) the alignment rate of the duplicated query sequence, Eq (10), (F) the density of gaps in the duplicated query sequence alignment, Eq (11), (G) the sequence similarities based on the duplicated query sequence alignment, Eq (12), (H) the identity improvement rate in the duplicated query sequence alignment, Eq (13–1), (I) the similarity improvement rate in the duplicated query sequence alignment, Eq (13–2), and (J) the improvement rate of alignment-rate in the duplicated query sequence alignment, Eq (13–3). All these CP pairs were obtained from the CPDB. See File S1 for the entity list.

Table 2
The number of remaining CP pairs after each filtration step.

Stage and filtering criteria		Number of high-quality CP pairs[a]	Number of marginal CP pairs
Screening	Input data	480 (471)	3,689
	Sequence alignments available	480 (471)	2,555
Initial filtration	Length thresholds[1,2] (Fig. 4A, B)	478 (470)	2,072
	CP site identified[3]	477 (469)	67
	Alignment boundary shifts[4,5] (Fig. 4C, D)	477 (469)	51
	Alignment consistency[6–8]	472 (468)	34
	Alignment rate[9] (Fig. 4E)	472 (468)	22

[a] The number in brackets is the number of pairs sharing ≥ 15 % sequence identity.

[1–9] These numbers correspond to descriptions of the nine discriminatory features utilized in the initial filtration stage.

for its use in the final identification of CP pairs from the candidate pool. To determine a suitable threshold setting for the SeqCP score, we computed the MCC, a stringent binary classification quality measure [57], at each of a series of SeqCP score thresholds (Table S2). The highest MCC, 0.80, occurred at the threshold 0.60. The SeqCP score threshold is suggested for future applications of the SeqCP method.

3.5. Large-scale performance test for the SeqCP method

To test SeqCP in a real scenario, we established a test dataset using NrPDB100-2019 as the source dataset. This 2019 test dataset contains 85,725 proteins. After the all-against-all database survey, which involved 7.35 × 10^9 pairs of proteins, 13,487 candidate CP pairs passed the screening and initial filtration steps. These pairs were subjected to the CE-CP-based standard CP identification procedure, and 4,929 high-quality CP pairs were identified. Using the CE-CP-determined high-quality CP pairs as the positive group and the other candidate CP pairs as the negative group (see Table S3 for details), we tested the binary classification quality of the SeqCP score. In the previous experiment based on NrPDB100-2007, we found that the SeqCP score aptly distinguished high-quality CP pairs from marginal pairs at a threshold of 0.60. As shown in Fig. 6B and Table 3, with the 2019 test dataset, the MCC obtained at the suggested SeqCP score threshold of 0.60 was the highest among the MCCs for all the tested thresholds. The AUC was 0.88 from NrPDB100-2019 (Fig. 6C). A more rigorous test was also performed by eliminating homologous sequences of NrPDB100-2007 from NrPDB100-2019 prior to the SeqCP all-against-all survey. As demonstrated in File S2: Independent test for the SeqCP method, the AUC was 0.81 and the MCC produced at the suggested SeqCP score threshold was 0.43. The stable SeqCP score thresholds and the AUC results in these datasets indicated that SeqCP is stable and suitable for identifying CP pairs.

3.6. Case studies

3.6.1. Circularly permuted structure modeling of protein UPI00057E200D

UPI00057E200D is a choice-of-anchor J domain-containing protein recorded in UniProt. It has no full-length homolog in the PDB and thus requires more effort to get suitable templates for its structural modeling. Using UPI00057E200D as the query, we applied SeqCP to search against PDB and found that it actually

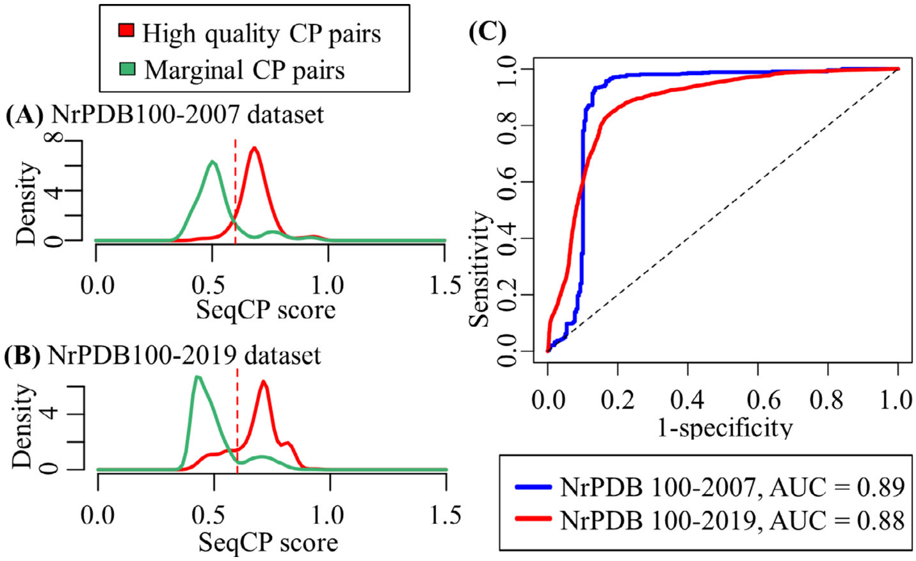


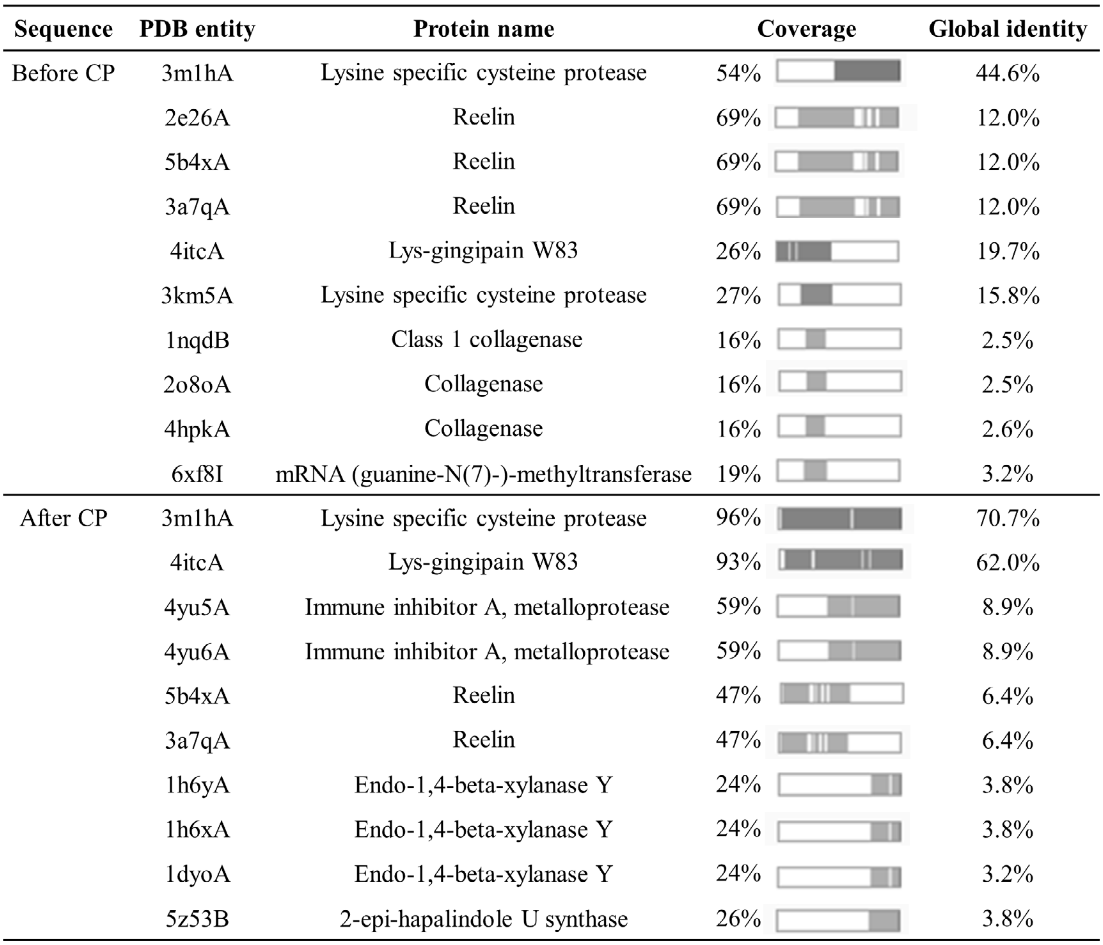
Fig. 6. Feasibility of the SeqCP score for identifying high-quality CP pairs. (A) Distributions of the SeqCP scores between the high-quality and marginal CP pairs. This test utilized the protein pairs of all proteins from the NrPDB100-2007 dataset. After the initial filtration, 726 candidate CP pairs remained. The CE-CP-based standard procedure was used to determine their CP relationships. The distributions of the SeqCP scores of the high-quality and marginal CP pair groups were assessed to observe how well this score separated the two groups. After the MCCs at various SeqCP score thresholds based on these distributions were computed, the two groups were determined to be well distinguished by a threshold of 0.60, as indicated by the dotted line. (B) The distributions of SeqCP scores between the high-quality and marginal CP pairs identified from the NrPDB100-2019 dataset. High-quality and marginal CP pairs were well distinguished by the suggested SeqCP score threshold of 0.60. (C) ROC curves of the SeqCP scores of the NrPDB100-2007 and NrPDB100-2019 datasets. The high AUC value indicates that the SeqCP score is feasible for detecting CP relationships between proteins.

Table 3
Binary classification performance of the SeqCP score for a test dataset composed of proteins from NrPDB100-2019.

SeqCP score threshold	Sensitivity	Specificity	Precision	Accuracy	MCC
0.85	1.30 %	99.81 %	80.00 %	63.81 %	0.0697
0.80	11.42 %	98.97 %	86.48 %	66.98 %	0.2335
0.75	20.71 %	96.20 %	75.85 %	68.61 %	0.2718
0.70	51.86 %	91.54 %	77.93 %	77.04 %	0.4871
0.65	71.39 %	87.08 %	76.09 %	81.34 %	0.5932
0.60	79.89 %	84.86 %	75.24 %	83.04 %	0.6399
0.55	86.63 %	79.33 %	70.71 %	82.00 %	0.6388
0.50	92.41 %	65.05 %	60.36 %	75.05 %	0.5574
0.45	97.63 %	38.26 %	47.66 %	59.95 %	0.3983
0.40	99.92 %	2.76 %	37.18 %	38.27 %	0.0975

had a high-quality CPM, the lysine-specific cysteine protease with PDB entity 3m1hA. The CP site for 3m1hA on UPI00057E200D was residue 71, and the SeqCP score for this site was 0.72. When the normal sequence and the residue-71 circularly permuted sequence of UPI00057E200D were used to run SWISS-MODEL modeling [58], the results of template searching were very different, as shown in Fig. 7.

When the normal sequence of UPI00057E200D was used as the modeling target, the best template was 3m1hA with only 54 % coverage. When the circularly permuted sequence of UPI00057E200D suggested by SeqCP was used as the target, the best template identified by SWISS-MODEL was still 3m1hA, but the coverage rose to 96 %. The models of normal UPI00057E200D and the SeqCP-suggested permutant generated based on the optimal templates are demonstrated in Fig. 8. Because of the low target coverage for the normal UPI00057E200D sequence, large proportions of the models were structurally undetermined. In contrast, the model of the UPI00057E200D permutant was fully constructed. The well-known protein structure prediction method AlphaFold2 [59,60] was also tested, and the overall confidence score of the structure it predicted was high (Fig. 8D). The model constructed based on SeqCP results and the structure predicted by AlphaFold2 were very



Sequence	PDB entity	Protein name	Coverage	Global identity
Before CP	3m1hA	Lysine specific cysteine protease	54%	44.6%
	2e26A	Reelin	69%	12.0%
	5b4xA	Reelin	69%	12.0%
	3a7qA	Reelin	69%	12.0%
	4itcA	Lys-gingipain W83	26%	19.7%
	3km5A	Lysine specific cysteine protease	27%	15.8%
	1nqdB	Class 1 collagenase	16%	2.5%
	2o8oA	Collagenase	16%	2.5%
	4hpkA	Collagenase	16%	2.6%
	6xf8I	mRNA (guanine-N(7)-)-methyltransferase	19%	3.2%
After CP	3m1hA	Lysine specific cysteine protease	96%	70.7%
	4itcA	Lys-gingipain W83	93%	62.0%
	4yu5A	Immune inhibitor A, metalloprotease	59%	8.9%
	4yu6A	Immune inhibitor A, metalloprotease	59%	8.9%
	5b4xA	Reelin	47%	6.4%
	3a7qA	Reelin	47%	6.4%
	1h6yA	Endo-1,4-beta-xylanase Y	24%	3.8%
	1h6xA	Endo-1,4-beta-xylanase Y	24%	3.8%
	1dyoA	Endo-1,4-beta-xylanase Y	24%	3.2%
	5z53B	2-epi-hapalindole U synthase	26%	3.8%

Fig. 7. Template search results of protein UPI00057E200D. Using the normal sequence of protein UPI00057E200D (UniProt ID) as the modeling target, the templates identified by SWISS-MODEL had low target coverages. When UPI00057E200D was circularly permuted at residue 71 according to its CPM discovered by SeqCP before being used as the modeling target, the top templates identified by SWISS-MODEL had high coverages and sequence identities. The structures of the constructed models are shown in Fig. 8.

similar (RMSD: 1.20 Å, number of aligned residues: 154). These results demonstrate that SeqCP provides an easy way to predict the structure of CP proteins and may help to investigate the function of novel proteins that have no known collinear structural homologs.

3.6.2. Circularly permuted protein structure modeling for protein A0A388KU10

A0A388KU10 is a proton-translocating NAD(P)(+) transhydrogenase. According to its UniProt annotation, it contains two domains related to alanine dehydrogenase. The lengths of the two domains are only 139 and 166 residues, respectively, while the protein has 1,079 residues. In other words, it has no full-length and closely related homolog in the PDB. Using A0A388KU10 as the query, we applied SeqCP to search against the PDB and found several high-quality CPMs. For instance, A0A388KU10 was correlated with PDB 6qtiA (a nicotinamide nucleotide transhydrogenase) by CP at residue 475 with a SeqCP score of 0.72. The results of template searching with SWISS-MODEL for the normal and CPM475 sequences of A0A388KU10 are shown in Fig. 9. With the same optimal template (PDB 6qtiA), after the CPM sequence suggested by SeqCP was used, the target-template coverage rose from 49 % to 90 %, and the global sequence identity rose from ~22 % to 41 %. As demonstrated in Fig. 10, large proportions of the models constructed with the normal A0A388KU10 sequence were structurally undetermined. In contrast, the model based on the CPM475 sequence of A0A388KU10 was fully constructed.

There is a predicted structure of A0A388KU10 in the AlphaFold Protein Structure Database [59,61] (Fig. 10D). Compared with this AlphaFold2 model, the model created based on SeqCP results was more similar to the CPM template 6qtiA (Fig. 10E). Since this template is a homodimeric protein [62], we supposed that the difference between the AlphaFold2 and SeqCP models is because

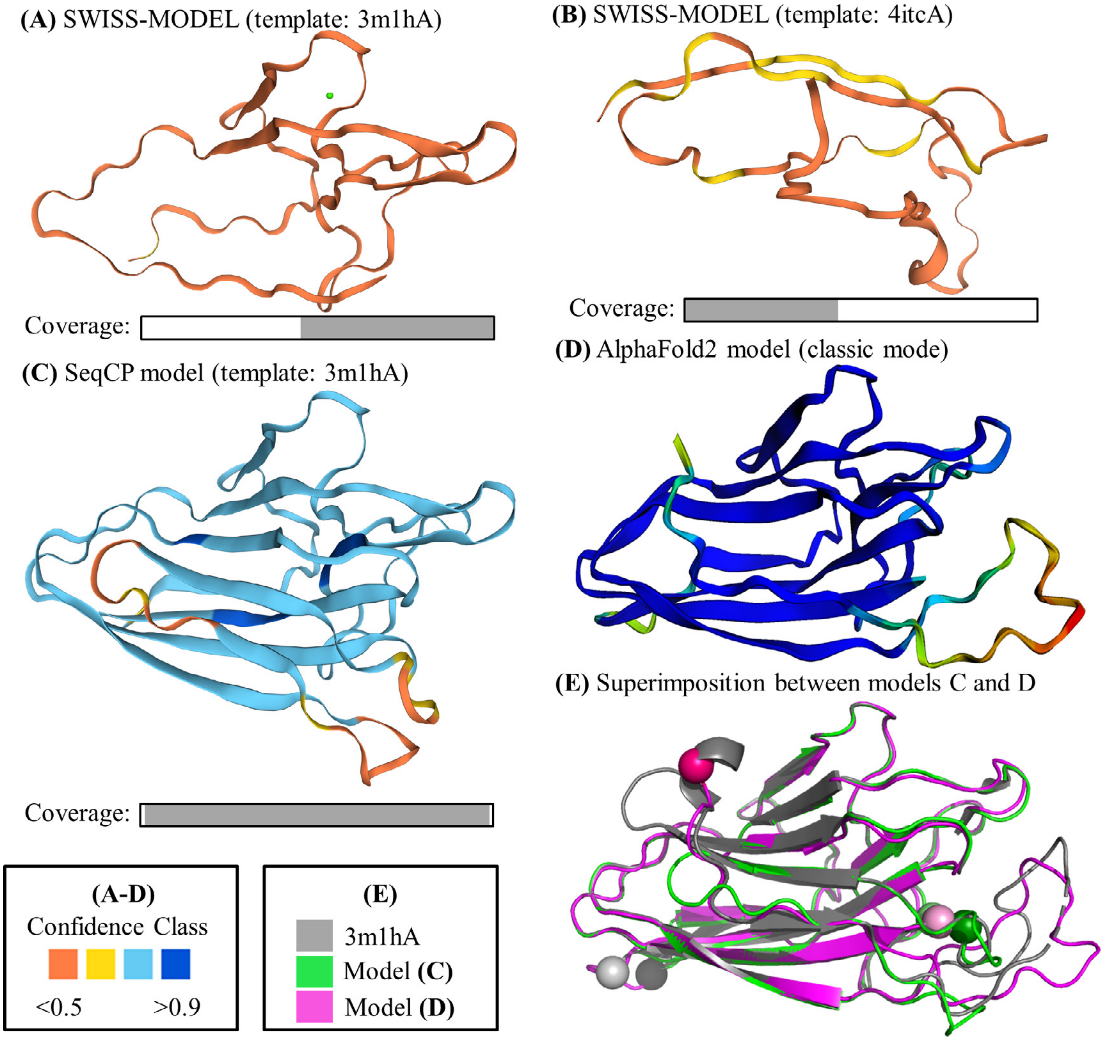
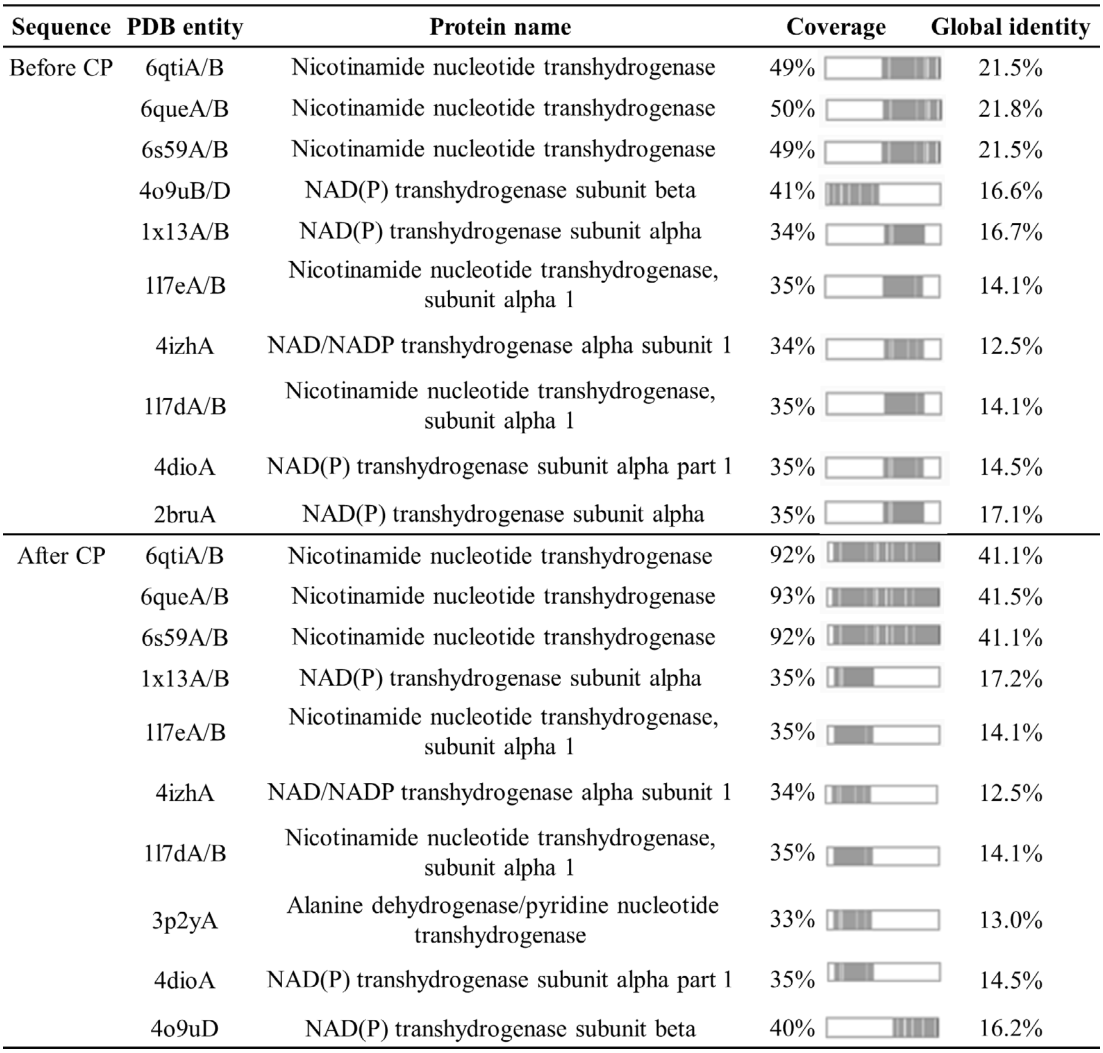

(A) SWISS-MODEL (template: 3m1hA)
Coverage:
(B) SWISS-MODEL (template: 4itcA)
Coverage:
(C) SeqCP model (template: 3m1hA)
Coverage:
(D) AlphaFold2 model (classic mode)
(E) Superimposition between models C and D
(A-D)
Confidence Class
<0.5
>0.9
(E)
3m1hA
Model (C)
Model (D)

Fig. 8. Structural modeling results for protein UPI00057E200D. (A) Model constructed based on the normal sequence of UniProt UPI00057E200D and the optimal template PDB 3m1hA by SWISS-MODEL. As indicated by the confidence class based on the AlphaFold per-residue confidence score (from 0 to 1), this model possessed a low quality. In particular, a large N-terminal region was missing. The coverage of the modeling structure is represented by the gray area in the box. (B) Model constructed based on the normal sequence of UPI00057E200D and another optimal template, PDB 4itcA, by SWISS-MODEL. This model possessed low quality, with a large proportion of the C-terminus missing. (C) Model constructed based on the SeqCP-suggested CPM71 sequence of UPI00057E200D and the template 3m1hA by SWISS-MODEL. The model was fully constructed with high overall quality. (D) Model constructed based on the normal sequence of UPI00057E200D by AlphaFold2 [59]. (E) Structural superimposition between model C (green; SeqCP) and model D (magenta; AlphaFold2). The N-/C-termini are represented by spheres, where the light colors indicate the C-termini. The model constructed based on SeqCP results highly resembled the structure predicted by AlphaFold2. The RMSD between the two structures was 1.20 Å, with 154 closely aligned α-carbon pairs. These results demonstrate that SeqCP helps identify an appropriate CP-related template for modeling proteins with no known collinear structural homologs. SeqCP may, thus, help correctly predict the structures of such novel proteins. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



Sequence	PDB entity	Protein name	Coverage	Global identity
Before CP	6qtiA/B	Nicotinamide nucleotide transhydrogenase	49%	21.5%
	6queA/B	Nicotinamide nucleotide transhydrogenase	50%	21.8%
	6s59A/B	Nicotinamide nucleotide transhydrogenase	49%	21.5%
	4o9uB/D	NAD(P) transhydrogenase subunit beta	41%	16.6%
	1x13A/B	NAD(P) transhydrogenase subunit alpha	34%	16.7%
	1l7eA/B	Nicotinamide nucleotide transhydrogenase, subunit alpha 1	35%	14.1%
	4izhA	NAD/NADP transhydrogenase alpha subunit 1	34%	12.5%
	1l7dA/B	Nicotinamide nucleotide transhydrogenase, subunit alpha 1	35%	14.1%
	4dioA	NAD(P) transhydrogenase subunit alpha part 1	35%	14.5%
	2bruA	NAD(P) transhydrogenase subunit alpha	35%	17.1%
After CP	6qtiA/B	Nicotinamide nucleotide transhydrogenase	92%	41.1%
	6queA/B	Nicotinamide nucleotide transhydrogenase	93%	41.5%
	6s59A/B	Nicotinamide nucleotide transhydrogenase	92%	41.1%
	1x13A/B	NAD(P) transhydrogenase subunit alpha	35%	17.2%
	1l7eA/B	Nicotinamide nucleotide transhydrogenase, subunit alpha 1	35%	14.1%
	4izhA	NAD/NADP transhydrogenase alpha subunit 1	34%	12.5%
	1l7dA/B	Nicotinamide nucleotide transhydrogenase, subunit alpha 1	35%	14.1%
	3p2yA	Alanine dehydrogenase/pyridine nucleotide transhydrogenase	33%	13.0%
	4dioA	NAD(P) transhydrogenase subunit alpha part 1	35%	14.5%
	4o9uD	NAD(P) transhydrogenase subunit beta	40%	16.2%

Fig. 9. Template search results for protein A0A388KU10. When the normal sequence of UniProt A0A388KU10 was used as the modeling target, the templates identified with SWISS-MODEL exhibited low alignment coverages and low global identities. When A0A388KU10 was circularly permuted at residue 475 before modeling, the top templates had remarkably improved coverages and sequence identities. The structures of the constructed models are shown in Fig. 10.

AlphaFold2 modeled A0A388KU10 as a monomer. However, the length of A0A388KU10 was too long to predict its dimeric form using the protein complex mode AlphaFold2 server [60], which still had many restrictions on sequence length and computation time at the time of the writing of this article.

3.6.3. The YbeA RNA methyltransferase

YbeA RNA methyltransferase is a side-by-side homodimeric protein (PDB 1ns5) [63]. Recently, we discovered by X-ray crystallography that performing CP at residue 74 for this protein can turn it into a 3D domain-swapped dimer, where each monomer exhibits an open form structure (PDB 5zyo) [64]. When there are such cases that a CP pair showed significant structural differences, SeqCP can still detect their CP-relationship while structure-based CP-detecting algorithms may not. For instance, the SeqCP score between the native YbeA RNA methyltransferase 1ns5 and its CPM74 5zyo was high (0.94) with a full-length alignment, whereas the CE-CP reported that their alignment ratio was only 75 % (116 out of 155 residues). To generate a structural model for the sequence of 1ns5 CPM74, the template suggested based on SeqCP search was 5zyo. As shown in Fig. 11C, the model of 1ns5 CPM74 generated with this template by SWISS-MODEL is in a correct 3D domain-swapped open form. Interestingly, AlphaFold2 predicted 1ns5 CPM74 as a closed-form structure, regardless of whether the classic or protein complex mode AlphaFold2 server was used (Fig. 11D and 11E, respectively). These results indicate that SeqCP is promising for modeling CPMs even when CP stimulates conformational changes or influences the method of protein complex formation.

4. Discussion

4.1. Precision in determining CP sites between permutants

In this work, SeqCP was evaluated with an *in silico* random CP sequence dataset, which was also used to assess the core algorithm of CPSARST [31]. SeqCP was more precise than the latter in determining the exact position of the CP site between protein sequences (Fig. 3). The outperformance of SeqCP came from the application of database searching instead of pairwise alignments. The sequence alignment algorithm applied by SeqCP is the database searching version of BLAST (the blastp program), but the one applied by CPSARST in the experiment was the pairwise BLAST (the bl2seq program). Database searching can yield valuable information based

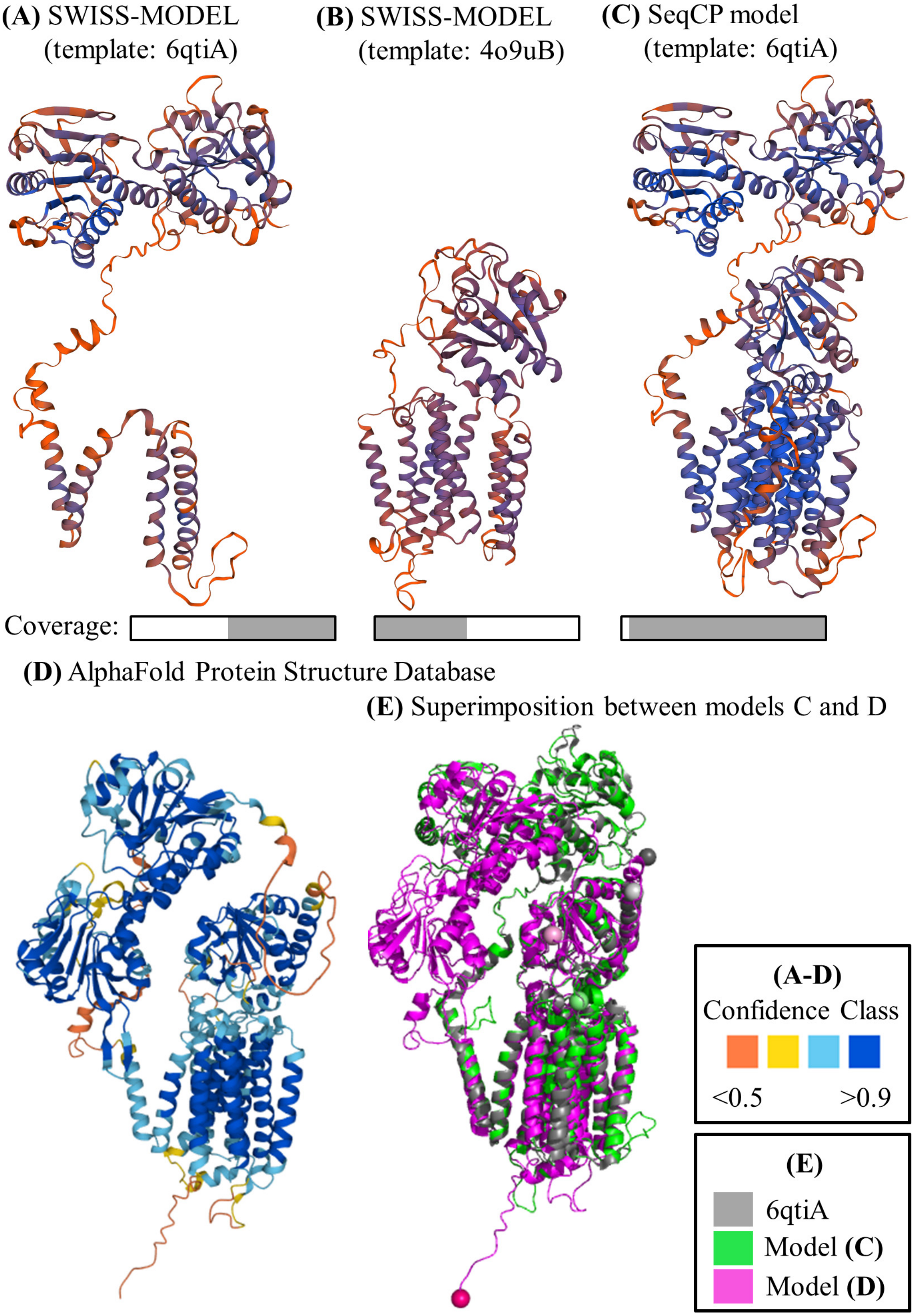

(A) SWISS-MODEL (template: 6qtiA)
(B) SWISS-MODEL (template: 4o9uB)
(C) SeqCP model (template: 6qtiA)
Coverage:
(D) AlphaFold Protein Structure Database
(E) Superimposition between models C and D
(A-D)
Confidence Class
<0.5
>0.9
(E)
6qtiA
Model (C)
Model (D)

Fig. 10. Structural modeling results for protein A0A388KU10. (A) Model constructed based on the normal sequence of UniProt A0A388KU10 and the template PDB 6qtiA by SWISS-MODEL. A large N-terminal region was missing because of low target coverage. (B) Model constructed based on the normal sequence of A0A388KU10 and the template PDB 4o9uB by SWISS-MODEL. A large proportion of the C-terminus is missing. (C) Model constructed based on the SeqCP-suggested CPM475 sequence of A0A388KU10 and the template 6qtiA by SWISS-MODEL. This model had a higher target coverage than models A and B. (D) A0A388KU10 model produced by the AlphaFold Protein Structure Database. The color indicates the quality of the models. (E) Structure superimposition of models C (green; SeqCP) and D (magenta; AlphaFold2) and the template 6qtiA (gray) for model C. The model constructed based on SeqCP results is structurally more similar to the known CPM template structure than that constructed by AlphaFold2. The N- and C-termini are represented by light and dark spheres, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

on homologs identified from the target dataset to help identify the optimal alignment between the query and target sequences, such as the E-value and lambda value [53]. We hypothesize that if the PSI-BLAST algorithm and position-specific scoring matrix (PSSM) [53] are applied, the precision of CP site determination can be further improved. Follow-up experiments are required to verify this

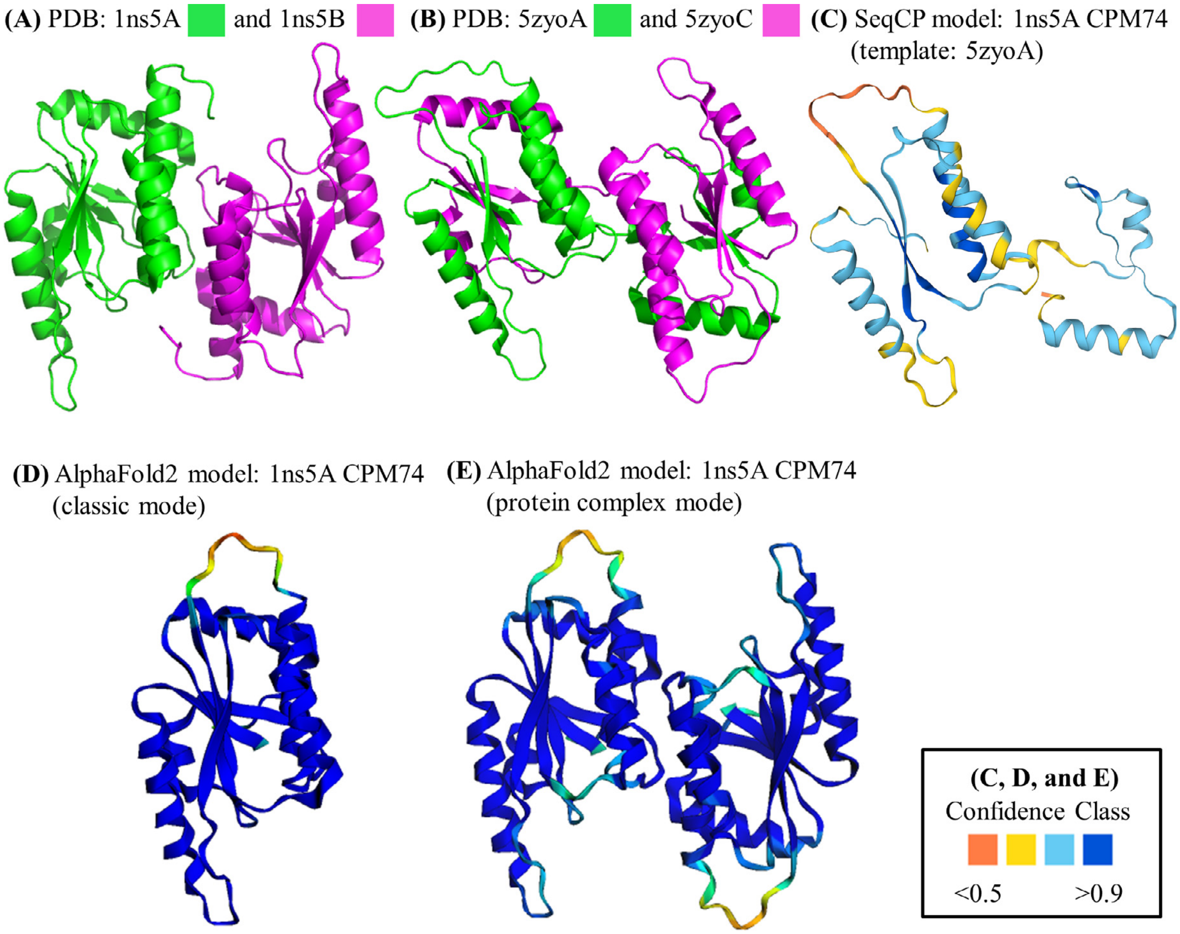

(A) PDB: 1ns5A 🟩 and 1ns5B ■
(B) PDB: 5zyoA 🟩 and 5zyoC ■
(C) SeqCP model: 1ns5A CPM74 (template: 5zyoA)
(D) AlphaFold2 model: 1ns5A CPM74 (classic mode)
(E) AlphaFold2 model: 1ns5A CPM74 (protein complex mode)
(C, D, and E)
Confidence Class
<0.5
>0.9

Fig. 11. Structural modeling results for YbeA RNA methyltransferase. (A) Structure of PDB 1ns5, the native YbeA. Chain A is in green and chain B is in magenta. (B) Crystal structure of PDB 5zyo, the CPM74 structure of 1ns5. Chain A is in green and chain C is in magenta. After CP, the conformation of this protein changes from a closed form into a 3D domain-swapped open form. (C) The model of 1ns5 CPM74 constructed based on the template suggested by SeqCP, with the SWISS-MODEL applied. (D) AlphaFold2 model of 1ns5 CPM74. The classic mode of AlphaFold2 was applied. (E) AlphaFold2 homodimer model of 1ns5 CPM74. The protein complex mode of AlphaFold2 was applied. The color codes for the quality of models C–D are shown on the right. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

hypothesis. However, it is expected that although the precision of CP site determination can be increased via these accurate alignment algorithms, the time cost may also increase.

4.2. Improvements of SeqCP over sequence-based CP detection methods

The last published sequence-based CP search method is a motif-based algorithm developed by Weiner *et al.* in 2005. The motif-based algorithm identified 93 CP pairs out of 153,646 nonredundant proteins in the ProDom database v.4.2 (released in 2004) [42]. With the SeqCP score threshold set as 0.6 as suggested, SeqCP searched 26,349 nonredundant proteins in NrPDB100-2007 and identified 468 candidate CP pairs, where 435 CP pairs were high-quality CP pairs. The results demonstrated that SeqCP could efficiently identify many more CP pairs than existing sequence-based CP search methods. Instead of considering the sequence alignment only, SeqCP also used several features related to the alignment qualities and CP characteristics to select and score candidate CP pairs. These features allow SeqCP to identify CP pairs with low identities, making it more sensitive than existing sequence-based methods. Moreover, we have implemented the SeqCP algorithm into a web server for easier access at higher throughput by the user than command-line programs. The user can perform multiple sequence searches with just one query.

4.3. Advantages of SeqCP over structure-based CP detection methods

The ability to perform CP analysis without relying on structural information is an advantage of SeqCP over structure-based CP detection algorithms. If knowing the structures of protein materials is required for CP detection, the scope of applications would be limited. In addition, in this postgenomics era, protein sequences are determined at an unprecedented speed. Since the number of sequences is increasing much more rapidly than the amount of protein structural data, much more information regarding CP will be extractable from the current protein sequence databases than from structure databases. Even though the sensitivity of the sequence-based algorithm SeqCP may be lower than that of structure-based methods, its sequence-based nature may help prevent confounding CP and the internal symmetry of protein structures [37]. It may also aid in distinguishing actual CP occurrences from CP-like structures caused by convergent evolution, such as those in some cases of helix bundles. When CPMs have conformational changes, SeqCP may be a better choice for detection than structure-based CP detection algorithms. Another advantage of SeqCP is its speed. For instance, the all-against-all SeqCP survey performed on NrPDB100-2007 took only 397 min using one CPU. Since 6.9 × 10^8 protein pairs were scanned in that experiment (Table S1), the scanning speed of SeqCP was approximately 1.75 million pairs per minute. For reference, CPSARST could scan ~ 53 thousand pairs per minute using the same hardware [31]. SeqCP was 33 times faster than CPSARST. The high speed of SeqCP makes it feasible for large-scale studies or applications of CP.

4.4. Meaning and recommended thresholds of the SeqCP score

The SeqCP score describes how much the alignment quality between the duplicated query sequence and the target is improved over the normal query sequence and target. The results of performance assessments obtained with the test dataset (NrPDB100-2007) and the final evaluation dataset (NrPDB100-2019) suggest that the optimal threshold of this score for discriminating high-quality CP pairs from marginal ones is 0.60 (Fig. 6 and Table 3). At this threshold, the SeqCP sensitivity is 80–93 % (average: 86 %), and the specificity is also approximately 86 % (Table 3 and Table S2), meaning that among all known high-quality CP pairs, ~86 % will receive a SeqCP score > 0.60, and among all the protein pairs determined not to be high-quality CPMs by SeqCP, ~86 % will be correct determinations. When the SeqCP score distributions of these datasets, the sizes of which varied by threefold, were compared, the results were similar enough to indicate that the performance of SeqCP is stable. For applications where positive determinations are critical, we suggest setting a higher SeqCP score threshold. For instance, using a threshold of 0.75 will reduce the number of candidate CPMs by 21 % but at the same time remove 96 % of unlikely cases (Table 3). In this manner, by sacrificing the reported number of high-quality CPMs, a higher probability of correct positive determinations can be achieved.

4.5. Applications and future studies

We have demonstrated that SeqCP is applicable for identifying suitable CP templates for modeling proteins without appropriate colinear homologs (Figs. 7–11). Although manually combining partial modeling results from SWISS-MODEL based on normal query sequences may still establish full-length models in some cases, such a jigsaw puzzling procedure would render large-scale CP studies difficult. The AlphaFold2 [59] algorithm is theoretically promising for modeling CP structures. However, its current implementation seems insufficiently thorough in modeling oligomeric proteins, especially when significant conformational change is stimulated by CP (Fig. 10 and Fig. 11). Predicting protein complex structure is still challenging. The SeqCP can be very helpful for structural modeling of proteins with high-quality CPM structures, including dimers, as demonstrated by the tested cases. However, SeqCP by itself is just a CPM search algorithm; only when applied along with modeling algorithms will its benefit in structure prediction manifest. The AlphaFold2 server is frequently updated. We suppose that if the SeqCP procedure can be integrated into the AlphaFold2 pipeline in the near future, the accuracy of CP structure prediction will be greatly enhanced. We have been developing an automated system, CirPred, for predicting the circularly permuted structure of a protein sequence at any given cutting site [29] by integrating the CPM template search by SeqCP and viable CP structural modeling based on the CPred and (PS)^2 algorithms [32,65]. This integrated CP modeling system will be a powerful tool for research and applications in which CP protein engineering plays a critical role. In the present study, we developed several discriminatory features to identify high-quality CP pairs. These features are very suitable for machine learning, which we will soon utilize to improve the SeqCP system. Additionally, new features composed of predicted secondary structural information [66,67] will be applied. We will also use SeqCP to construct a sequence-based CP database that is expected to be much more extensive than structure-based CP datasets. With the assistance of this database, the natural prevalence, taxonomy, and evolutionary mechanisms underlying CP can be studied in detail.

Funding

This research was funded by the National Core Facility for Biopharmaceuticals (NCFB), the Ministry of Science and Technology (MOST, https://www.most.gov.tw/?l=en), Taiwan, with Grant No NSTC 111-2311-B-007-008- and MOST 110-2740-B-400-002 for Prof. Ping-Chiang Lyu and MOST 111-2221-E-A49-163- for Wei-Cheng Lo. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

CRediT authorship contribution statement

Chi-Chun Chen: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Yu-Wei Huang:** Validation, Data curation, Writing – review & editing. **Hsuan-Cheng Huang:** Validation, Writing – review & editing, Supervision. **Wei-Cheng Lo:** Conceptualization, Methodology, Validation, Investigation, Resources, Data curation, Supervision, Funding acquisition, Writing – original draft, Writing – review & editing. **Ping-Chiang Lyu:** Conceptualization, Methodology, Validation, Investigation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank Professors Po-Jung Huang and Chi-Ching Lee, Chang Gung University, Taiwan, and our laboratory members for their constructive suggestions. Chia-Hua Lo, a student of Ping-Chiang Lyu, is acknowledged for improving the English language quality of this article. The National Center for High-performance Computing, Taiwan, is acknowledged for providing the computing power. We would like to acknowledge the Ministry of Science and Technology, Taiwan, for financial support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.11.024.

References

[1] Cunningham BA, Hemperly JJ, Hopp TP, Edelman GM. Favin versus concanavalin A: Circularly permuted amino acid sequences. PNAS 1979;76(7):3218–22.

[2] Carrington DM, Auffret A, Hanke DE. Polypeptide ligation occurs during post-translational modification of concanavalin A. Nature 1985;313(5997):64–7.

[3] Bowles DJ, Marcus SE, Pappin DJ, Findlay JB, Eliopoulos E, et al. Posttranslational processing of concanavalin A precursors in jackbean cotyledons. The Journal of cell biology 1986;102(4):1284–97.

[4] Wallace CJ. The curious case of protein splicing: mechanistic insights suggested by protein semisynthesis. Protein science : a publication of the Protein Society 1993;2(5):697–705. https://doi.org/10.1002/pro.5560020501.

[5] Bliven S and Prlic A (2012) Circular permutation in proteins. PLoS computational biology. 8(3):e1002445. http://dx.doi.org/10.1371/journal.pcbi.1002445.

[6] Luger K, Hommel U, Herold M, Hofsteenge J, Kirschner K. Correct folding of circularly permuted variants of a beta alpha barrel enzyme in vivo. Science 1989;243(4888):206–10.

[7] Ponting CP, Russell RB. Swaposins: circular permutations within genes encoding saposin homologues. Trends Biochem Sci 1995;20(5):179–80.
[8] Jeltsch A. Circular permutations in the molecular evolution of DNA methyltransferases. J Mol Evol 1999;49(1):161–4.
[9] Peisajovich SG, Rockah L, Tawfik DS. Evolution of new protein topologies through multistep gene rearrangements. Nat Genet 2006;38(2):168–74. https://doi.org/10.1038/ng1717.
[10] Vogel C, Morea V. Duplication, divergence and formation of novel protein topologies. BioEssays 2006;28(10):973–8.
[11] Doolittle WF. What introns have to tell us: hierarchy in genome evolution. Cold Spring Harb Symp Quant Biol 1987;52:907–13.
[12] Gilbert W. The exon theory of genes. Cold Spring Harb Symp Quant Biol 1987;52:901–5.
[13] Lindqvist Y, Schneider G. Circular permutations of natural protein sequences: structural evidence. Curr Opin Struct Biol 1997;7(3):422–7.
[14] Ay J, Hahn M, Decanniere K, Piotukh K, Borriss R, et al. Crystal structures and properties of de novo circularly permuted 1,3–1,4-beta-glucanases. Proteins 1998;30(2):155–67.
[15] Goldenberg DP, Creighton TE. Circular and circularly permuted forms of bovine pancreatic trypsin inhibitor. J Mol Biol 1983;165(2):407–13.
[16] Heinemann U, Hahn M. Circular permutation of polypeptide chains: implications for protein folding and stability. Prog Biophys Mol Biol 1995;64(2–3):121–43.
[17] Hennecke J, Sebbel P, Glockshuber R. Random circular permutation of DsbA reveals segments that are essential for protein folding and stability. J Mol Biol 1999;286(4):1197–215. https://doi.org/10.1006/jmbi.1998.2531.
[18] Todd AE, Orengo CA, Thornton JM. Plasticity of enzyme active sites. Trends Biochem Sci 2002;27(8):419–26.
[19] Qian Z, Lutz S. Improving the catalytic activity of Candida antarctica lipase B by circular permutation. J Am Chem Soc 2005;127(39):13466–7. https://doi.org/10.1021/ja053932h.
[20] Nagaratnam N, Delker SL, Jernigan R, Edwards TE, Snider J, et al. Structural insights into the function of the catalytically active human Taspase1. Structure 2021;29(8):873–885 e875. https://doi.org/10.1016/j.str.2021.03.008.
[21] Gebhard LG, Risso VA, Santos J, Ferreyra RG, Noguera ME, et al. Mapping the distribution of conformational information throughout a protein sequence. J Mol Biol 2006;358(1):280–8. https://doi.org/10.1016/j.jmb.2006.01.095.
[22] Nakamura T, Iwakura M. Circular permutation analysis as a method for distinction of functional elements in the M20 loop of Escherichia coli dihydrofolate reductase. J Biol Chem 1999;274(27):19041–7. https://doi.org/10.1074/jbc.274.27.19041.
[23] Arnold FH. Fancy footwork in the sequence space shuffle. Nat Biotechnol 2006;24(3):328–30. https://doi.org/10.1038/nbt0306-328.
[24] Kojima M, Ayabe K, Ueda H. Importance of terminal residues on circularly permutated Escherichia coli alkaline phosphatase with high specific activity. J Biosci Bioeng 2005;100(2):197–202. https://doi.org/10.1263/jbb.100.197.
[25] Baird GS, Zacharias DA, Tsien RY. Circular permutation and receptor insertion within green fluorescent proteins. PNAS 1999;96(20):11241–6. https://doi.org/10.1073/pnas.96.20.11241.
[26] Ostermeier M. Engineering allosteric protein switches by domain insertion. Protein Eng Des Sel 2005;18(8):359–64. https://doi.org/10.1093/protein/gzi048.
[27] Lee YT, Su TH, Lo WC, Lyu PC, Sue SC. Circular permutation prediction reveals a viable backbone disconnection for split proteins: an approach in identifying a new functional split intein. PLoS ONE 2012;7(8):e43820.
[28] Lee YZ, Lo WC, Sue SC. Computational Prediction of New Intein Split Sites. Methods Mol Biol 2017;1495:259–68. https://doi.org/10.1007/978-1-4939-6451-2_17.
[29] Chen TR, Lin YC, Huang YW, Chen CC, Lo WC. CirPred, the first structure modeling and linker design system for circularly permuted proteins. BMC Bioinf 2021;22(Suppl 10):494. https://doi.org/10.1186/s12859-021-04403-1.
[30] Kostyuk AI, Demidovich AD, Kotova DA, Belousov VV, Bilan DS. Circularly Permuted Fluorescent Protein-Based Indicators: History, Principles, and Classification. Int J Mol Sci 2019;20(17). https://doi.org/10.3390/ijms20174200.
[31] Lo WC, Lyu PC. CPSARST: an efficient circular permutation search tool applied to the detection of novel protein structural relationships. Genome Biol 2008;9(1):R11. https://doi.org/10.1186/gb-2008-9-1-r11.
[32] Lo WC, Wang LF, Liu YY, Dai T, Hwang JK, et al. (2012) CPred: a web server for predicting viable circular permutations in proteins. Nucleic Acids Res. 40(Web Server issue):W232-237. http://dx.doi.org/10.1093/nar/gks529.
[33] Jung J, Lee B. Protein structure alignment using environmental profiles. Protein Eng 2000;13(8):535–43.
[34] Guerler A, Knapp EW. Novel protein folds and their nonsequential structural analogs. Protein science : a publication of the Protein Society 2008;17(8):1374–82. https://doi.org/10.1110/ps.035469.108.
[35] Jung J, Lee B. Circularly permuted proteins in the protein structure database. Protein Sci 2001;10(9):1881–6. https://doi.org/10.1110/ps.05801.
[36] Schmidt-Goenner T, Guerler A, Kolbeck B, Knapp EW. Circular permuted proteins in the universe of protein folds. Proteins 2010;78(7):1618–30. https://doi.org/10.1002/prot.22678.
[37] Bliven SE, Bourne PE, Prlic A. Detection of circular permutations within protein structures using CE-CP. Bioinformatics 2015;31(8):1316–8. https://doi.org/10.1093/bioinformatics/btu823.
[38] Lo WC, Huang PJ, Chang CH, Lyu PC. Protein structural similarity search by Ramachandran codes. BMC Bioinf 2007;8:307. https://doi.org/10.1186/1471-2105-8-307.
[39] Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11(9):739–47.
[40] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235–42.
[41] Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, et al. (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res. 47(D1):D464-D474. http://dx.doi.org/10.1093/nar/gky1004.
[42] Weiner 3rd J, Thomas G, Bornberg-Bauer E. Rapid motif-based prediction of circular permutations in multi-domain proteins. Bioinformatics 2005;21(7):932–7. https://doi.org/10.1093/bioinformatics/bti085.
[43] Uliel S, Fliess A, Amir A, Unger R. A simple algorithm for detecting circular permutations in proteins. Bioinformatics 1999;15(11):930–6.
[44] Uliel S, Fliess A, Unger R. Naturally occurring circular permutations in proteins. Protein Eng 2001;14(8):533–42.
[45] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48(3):443–53. https://doi.org/10.1016/0022-2836(70)90057-4.
[46] Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, et al. ProDom: automated clustering of homologous domains. Briefings Bioinf 2002;3(3):246–51.
[47] Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 2015;31(6):926–32. https://doi.org/10.1093/bioinformatics/btu739.
[48] Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 2007;23(10):1282–8. https://doi.org/10.1093/bioinformatics/btm098.
[49] Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, et al. (2022) Database resources of the national center for biotechnology information. Nucleic acids research. 50(D1):D20-D26. http://dx.doi.org/10.1093/nar/gkab1112.
[50] Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science 1985;227(4693):1435–41. https://doi.org/10.1126/science.2983426.
[51] Pearson WR and Lipman DJ (1988) Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences of the United States of America. 85(8):2444-2448. http://dx.doi.org/10.1073/pnas.85.8.2444.
[52] Lo WC, Lee CC, Lee CY and Lyu PC (2009) CPDB: a database of circular permutation in proteins. Nucleic Acids Res. 37(Database issue):D328-332. http://dx.doi.org/10.1093/nar/gkn679.
[53] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389–402. https://doi.org/10.1093/nar/25.17.3389.
[54] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
[55] Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2018.
[56] Rost B (1997) Protein structures sustain evolutionary drift. Folding & design. 2(3):S19-24.
[57] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. BBA 1975;405(2):442–51. https://doi.org/10.1016/0005-2795(75)90109-9.
[58] Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 2018;46(W1):W296–303. https://doi.org/10.1093/nar/gky427.
[59] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2.
[60] Mirdita M, Schutze K, Moriwaki Y, Heo L, Ovchinnikov S, et al. ColabFold: making protein folding accessible to all. Nat Methods 2022;19(6):679–82. https://doi.org/10.1038/s41592-022-01488-1.
[61] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, et al. (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic acids research. 50(D1):D439-D444. http://dx.doi.org/10.1093/nar/gkab1061.
[62] Kampjut D, Sazanov LA. Structure and mechanism of mitochondrial proton-translocating transhydrogenase. Nature 2019;573(7773):291–5. https://doi.org/10.1038/s41586-019-1519-2.
[63] Benach J, Shen, J, Rost, B, Xiao, R, Acton, T, Montelione, G, Hunt, JF (2003) Structure of YBEA from E. coli. Northeast Structural Genomics Research Consortium (NESG), PDB:1NS5. European Bioinformatics Institute, European Molecular Biology Laboratory, Grenoble, France.
[64] Ko KT, Hu IC, Huang KF, Lyu PC and Hsu SD (2019) Untying a Knotted SPOUT RNA Methyltransferase by Circular Permutation Results in a Domain-Swapped Dimer. Structure. 27(8):1224-1233 e1224. http://dx.doi.org/10.1016/j.str.2019.04.004.
[65] Huang TT, Hwang JK, Chen CH, Chu CS, Lee CW, et al. (2015) (PS)2: protein structure prediction server version 3.0. Nucleic Acids Res. 43(W1):W338-342. http://dx.doi.org/10.1093/nar/gkv454.
[66] Juan SH, Chen TR and Lo WC (2020) A simple strategy to enhance the speed of protein secondary structure prediction without sacrificing accuracy. PLoS One. 15(6):e0235153. http://dx.doi.org/10.1371/journal.pone.0235153.
[67] Chen TR, Juan SH, Huang YW, Lin YC and Lo WC (2021) A secondary structure-based position-specific scoring matrix applied to the improvement in protein secondary structure prediction. PLoS One. 16(7):e0255076. http://dx.doi.org/10.1371/journal.pone.0255076.