*Bioinformatics*, 32(3), 2016, 370–377
doi: 10.1093/bioinformatics/btv580
Advance Access Publication Date: 10 October 2015
Original Paper

Structural bioinformatics

# Fast and accurate non-sequential protein structure alignment using a new asymmetric linear sum assignment heuristic

**Peter Brown[1], Wayne Pullan[1], Yuedong Yang[2] and Yaoqi Zhou[1,2,*]**

[1]School of ICT and [2]Institute for Glycomics, Griffith University, Gold Coast, QLD 4222, Australia

*To whom correspondence should be addressed

Associate Editor: Anna Tramontano

Received on June 30, 2015; revised on August 28, 2015; accepted on October 4, 2015

## Abstract

**Motivation**: The three dimensional tertiary structure of a protein at near atomic level resolution provides insight alluding to its function and evolution. As protein structure decides its functionality, similarity in structure usually implies similarity in function. As such, structure alignment techniques are often useful in the classifications of protein function. Given the rapidly growing rate of new, experimentally determined structures being made available from repositories such as the Protein Data Bank, fast and accurate computational structure comparison tools are required. This paper presents SPalignNS, a non-sequential protein structure alignment tool using a novel asymmetrical greedy search technique.

**Results**: The performance of SPalignNS was evaluated against existing sequential and non-sequential structure alignment methods by performing trials with commonly used datasets. These benchmark datasets used to gauge alignment accuracy include (i) 9538 pairwise alignments implied by the HOMSTRAD database of homologous proteins; (ii) a subset of 64 difficult alignments from set (i) that have low structure similarity; (iii) 199 pairwise alignments of proteins with similar structure but different topology; and (iv) a subset of 20 pairwise alignments from the RIPC set. SPalignNS is shown to achieve greater alignment accuracy (lower or comparable root-mean squared distance with increased structure overlap coverage) for all datasets, and the highest agreement with reference alignments from the challenging dataset (iv) above, when compared with both sequentially constrained alignments and other non-sequential alignments.

**Availability and implementation**: SPalignNS was implemented in C++. The source code, binary executable, and a web server version is freely available at: http://sparks-lab.org

**Contact**: yaoqi.zhou@griffith.edu.au

## 1 Introduction

The three dimensional tertiary structure of proteins at near atomic level resolution provides an indication of their function and evolutionary relationships. In particular, similarity in structure usually implies similarity in function. Accordingly, functions of a protein without annotated functions can be predicted if it is structurally similar to proteins with known functions. These comparisons are productive because protein structures are often more conserved than their sequences (Chothia and Lesk, 1986). Accurate computational structure comparison is complementary to the much slower process of manual classification (Andreeva *et al.*, 2008; Greene *et al.*, 2007) which is unable to keep pace with newly determined structures from structural genomics projects (Burley, 2000). As a consequence, protein structure alignment has become an important technique for computational biology researchers involved in protein classification, evolutionary relationship determination, protein functional prediction, molecular modelling and protein engineering (Abyzov and Ilyin, 2007).

Protein structure alignment has been studied for over thirty years, with a large number of computational tools and techniques developed to address the problem (Ma and Wang, 2014; Shih and Hwang, 2010). In general, protein structure alignment can be defined as generating the set of residue pair mappings which maximizes the detection of similarity whilst minimizing geometric divergence. To date this has been achieved in two ways: sequentially and non-sequentially. A valid sequential alignment is achieved through adherence of the following two constraints: (1) *exclusivity* – no amino acid can be aligned with more than one amino acid in the other protein; and (2) *ordering* – the order of amino acids must be maintained with respect to the alignment. That is, if amino acids $i$ and $j$ in one protein are aligned with amino acids $k$ and $l$ in the other protein, then $i < j$ if and only if $k < l$. (Strickland *et al.*, 2005). Alternatively, non-sequential alignments release the ordering constraint and only enforce the exclusivity constraint, which in some instances has been found to improve similarity detection performance compared with sequentially constrained alignments, potentially uncovering hidden or unexpected relationships assisting with evolutional and functional annotations (Dundas *et al.*, 2011).

Representative examples of sequential alignment methods in early studies include DALI (Holm and Sander, 1993), SSAP (Orengo and Taylor, 1996), CE (Shindyalov and Bourne, 1998), FATCAT (Ye and Godzik, 2004), TMalign (Zhang and Skolnick, 2005)/FrTMalign (Pandit and Skolnick, 2008) and SALIGN (Madhusudhan *et al.*, 2009). A recent sequential alignment method is SPalign (Yang *et al.*, 2012) which optimizes a size-independent score called SP-score which fixes the cutoff distance at 4 Å and removes size dependence by using a normalization pre-factor. SPalign demonstrated improvements of alignment accuracy compared with (DALI, CE, TMalign and FrTMalign), as shown in (Yang *et al.*, 2012).

Representative examples of non-sequential alignment methods in earlier studies include: DALI (this tool can produce both sequential and non-sequential alignments), Geometric Hashing (Bachar *et al.*, 1993), SARF (Alexandrov, 1996), MASS (Dror *et al.*, 2003a,b), MUSTANG (Konagurthu *et al.*, 2006), GANGSTA$^+$ (Guerler and Knapp, 2008), SNAP (Salem *et al.*, 2009), FlexSnap (Salem *et al.*, 2010), CLICK (Nguyen *et al.*, 2011; Nguyen and Madhusudhan, 2011) and MICAN (Minami *et al.*, 2013). Both GANGSTA$^+$, and CLICK methods make use of combinatorial-based approaches to produce non-sequential structure alignments. GANGSTA$^+$ is an extension of the original Genetic Algorithm for Non-Sequential and Gapped Structure Alignment (GANGSTA) (Kolbeck *et al.*, 2006) tool, and replaced their Genetic Algorithm component with a combinatorial approach providing improved efficiency and reliability. CLICK alignments are generated by grouping locally aligned representative atoms within a certain distance threshold and then matches these groups into the best combination that maximizes coverage with the least squares fit. FlexSnap uses a greedy algorithm for chaining aligned fragment pairs (AFPs), allowing for flexible alignments to be produced by introducing hinges between AFPs. MICAN is based on the geometric hashing paradigm and focuses on SSEs for alignment, utilizing a multiple vector representation for each SSE.

This study employed combinatorial optimization techniques, specifically a Linear Sum Assignment Problem (LSAP) (Burkard and Cela, 1999) algorithm, to produce non-sequential structure alignments of proteins. We developed the Asymmetrical Greedy Search (AGS) algorithm that locates an approximate LSAP solution efficiently with negligible difference from the global minimum. The new non-sequential alignment software package, SPalignNS, is based on the optimization of SP-score (Yang *et al.*, 2012) for structure alignment. SPalignNS achieves highly accurate alignment results with better or comparable RMSD at a higher number of aligned residue pairs in a significantly shorter computational time than CLICK.

## 2 Method

LSAP is a classic combinatorial optimization problem that searches for an optimal combination of assignments subject to an $n \times m$ cost/benefit matrix, with the overall goal being to minimize/maximize the cost/benefit through a complete selection of one-to-one assignments. In terms of aligning a pair of protein structures, the benefit matrix applied here was generated using SP-score (detailed in Sect. 2.3) as the objective function for each residue from protein A ($n$ rows) aligned with each residue from protein B ($m$ columns). When $n = m$ the LSAP is symmetrical and when $n \neq m$ the LSAP is asymmetrical. Here, a symmetrical LSAP would require both input structures to have identical counts of representative atoms. In most cases the number of these potential alignment points in each protein will differ, requiring the application of an asymmetric LSAP algorithm. A valid solution for asymmetric LSAP is a semi-complete assignment where either every row is assigned to the best possible column leaving surplus columns unassigned, or conversely every column is assigned to the best possible row, leaving surplus rows unassigned. Attained assignment solutions will then directly represent the set of individual residue alignments used for non-sequential structure alignment.

One of the first exact algorithms to solve the symmetrical LSAP was the Hungarian Algorithm (Kuhn, 1955). Many other exact algorithms have subsequently been developed including the Auction Algorithm (Bertsekas, 1988) which is considered to be one of the fastest algorithms for finding the optimal solution to the assignment problem. However, for large-scale or complex instances of the assignment problem, the Auction Algorithm is not a viable option due to its pseudo-polynomial time complexity, particularly within applications where it is desirable to have minimal computational processing cost such as the one studied in this paper. Consequently the major requirement is that the LSAP algorithm must be able to quickly identify a solution which is very close to optimal. As the use of the LSAP algorithm in protein structure alignment is just one step within a heuristic process which has a number of approximations, optimality of the LSAP algorithm is not a requirement as the accuracy benefit from an exact LSAP solution compared with an approximate LSAP solution is negligible.

### 2.1 AGS algorithm

Here we developed a new heuristic algorithm for this study, Asymmetric Greedy Search (AGS), that locates approximate asymmetric LSAP solutions efficiently with negligible difference from the global minimum. The AGS algorithm was inspired by a similar algorithm applied in a different context, Deep Greedy Switching (DGS) (Naiem and El-Beltagy, 2009, 2013). Existing algorithms such as DGS, Hungarian and Auction are unsuitable for application to non-sequential protein structure alignment due to being constrained to symmetrical LSAP instances. However, the AGS algorithm is compatible with both symmetric and asymmetric LSAP instances.

The AGS algorithm operates as follows: It is assumed that the number of rows will be less than or equal to the number of columns. If this is not true, a transposition will occur to satisfy this requirement, however the final solution will be unaffected by this. Then, after generating an initial solution, the algorithm creates, for each row, the best row $(R, R^b)$ and column $(C, C^b)$ vectors of possible swaps. The time complexity of evaluating the total effect of swapping rows is $\mathcal{O}(n^2)$, and for evaluating the swapping of unused columns is $\mathcal{O}(n(m-n))$. These best row and column vectors are used to perform swaps from largest benefit to smallest benefit order until there are none remaining which improve the total benefit of the solution. During this improvement phase, the only updates to the best row $(R, R^b)$ and column $(C, C^b)$ benefit vectors are those for the row(s) involved in the swap. The time complexity for the updates to the best row and column vector at each iteration of the improvement phase is $\mathcal{O}(2(n-1)) + \mathcal{O}(2(m-n))$ and $\mathcal{O}(m-n)$ respectively. When there is no benefit improvement possible, $R, R^b$, $C, C^b$ are regenerated and the improvement phase is repeated. The algorithm terminates when there are no improving swaps available after $(R, R^b)$, $(C, C^b)$ have been regenerated, and the final assignment solution is returned.

In comparison to the fastest exact method (Hungarian algorithm) which has a time complexity of $\mathcal{O}(n^3)$, the AGS algorithm is an order of magnitude faster with an overall time complexity of $\mathcal{O}(n^2)$. The speedup here has proved beneficial for quickly solving the problem of finding the best set of non-sequential residue assignments, as this problem itself includes an inherent rise of computational complexity required when compared to finding the best set of sequential residue assignments. This complexity increase is due to the total number of possible residue assignment combinations being significantly increased when the ordering constraint is removed.

## 2.2 Computational experiments

The performance of SPalignNS was evaluated and compared with a number of recent sequential and non-sequential structure alignment methods by performing trials with commonly used datasets including two benchmarks containing known occurrences of proteins with non-sequential structural similarities.

The three most recent non-sequential methods were used to compare against: FlexSnap (Salem *et al.*, 2010), CLICK (Nguyen *et al.*, 2011) and MICAN (Minami *et al.*, 2013). Pre-compiled Linux binary files were downloaded and executed for FlexSnap available from http://www.cs.rpi.edu/~zaki/software/flexsnap, and for CLICK available from http://mspc.bii.a-star.edu.sg/click. The freely available source code for MICAN was downloaded and compiled from http://www.tbp.cse.nagoya-u.ac.jp/MICAN. It should be noted that all methods were executed using default parameters, and CLICK was used with the MODELLER software (Sali and Blundell, 1993) in order to replicate the published results in Nguyen and Madhusudhan (2011).

The benchmark datasets used to gauge alignment accuracy include (i) 9538 pairwise alignments implied by the HOMSTRAD database of homologous proteins, assessing the performance on alignments with the same topology; (ii) a subset of 64 difficult alignments from set (i) that have low structure similarity; (iii) 199 pairwise alignments of proteins with similar structure but different topology; and (iv) a subset of 20 pairwise alignments from the challenging RIPC set.

## 2.3 Alignment measures

CLICK used the RMSD and SO metrics (described in Sects. 2.3.2 and 2.3.3) for the final structure similarity evaluation. In this paper we used SP-score (briefly described in Sect. 2.3.4, see (Yang *et al.*, 2012) for full definition), a self-defined similarity score, for both similarity evaluation and alignment optimization. This guarantees that our final alignments are close to or at the global maxima for this score. To assess the alignment quality of each method, the following metrics and measures have been employed and are outlined as follows.

### 2.3.1 Number of aligned residues (Nali)

For each pairwise structure alignment, Nali represents the total count of aligned representative atoms. The default representative atom used for protein structures within SPalign and SPalignNS is Carbon Alpha (Cα), but this can be changed through the specification of an input parameter.

### 2.3.2 Root mean square deviation (RMSD)

RMSD provides an overall indication of the three-dimensional geometric similarity between two protein structures from a set of aligned residue mappings after superimposition, as calculated by Eq. (1), where $n$ represents the count of aligned residues and $d_{ij}^2$ is the squared Euclidian distance between representative atoms of the aligned residue pair.

$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_{ij}^2} \qquad (1)$$

### 2.3.3 Structure overlap (SO)

SO is defined as the percentage value of all aligned residues that are within 3.5 Å of each other after superimposition, and is given by Eq. (2), where $m$ is the smaller residue number of the two aligned structures, $n$ is the number of aligned residues, $d_{ij}$ is the Euclidian distance between representative atoms of aligned residue pairs and $d_0$ is the 3.5 Å Euclidian distance threshold.

$$\mathrm{SO} = 100 \times \frac{1}{m} \times \sum_{d_{ij} \le d_0}^{n} 1 \qquad (2)$$

### 2.3.4 SP-score

SP-score is used as both the objective function for alignment optimization and as the final structure similarity score. It is calculated per Eq. (3), where $L$ is defined as the sum of aligned core residues (where $d_{ij} \le 2d_0$) and the average number of neighbouring residues within $3d_0$ from any core residues. The summation only includes residues that are within $2d_0$, so that only meaningfully aligned residues contribute to the SP-score. The size-dependant normalization factor $\alpha$ is a constant value of 0.3. Finally, a constant value of 0.2 is used to ensure a smooth cut-off when $d_{ij} = 2d_0$.

$$\mathrm{SP{-}score} = \frac{1}{L^{1-\alpha}}\left[\sum_{d_{ij}<2d_0}^{n}\left(\frac{1}{1+\left(\frac{d_{ij}}{d_0}\right)^2} - 0.2\right)\right] \qquad (3)$$

## 2.4 Alignment datasets

In order to comprehensively assess the alignment performance of SPalignNS, a number of computational experiments have been carried out over four benchmark datasets. Defined as follows, the first three datasets were used as benchmarks for testing CLICK in (Nguyen and Madhusudhan, 2011). The final dataset has been used for comparatively assessing the quality of pairwise protein structure alignment methods (Mayr *et al.*, 2007), multiple protein structure alignment methods (Berbalk *et al.*, 2009), and is often used to test the accuracy of non-sequential protein structure alignment methods.

**Table 1.** Alignment performance comparison over the 9538 pairwise alignments from the ‘HOMSTRAD homologous proteins’ dataset showing average Nali, RMSD and SO for each method

| | Nali | RMSD (Å) | SO (%) |
|---|---|---|---|
| SPalignNS | 155 | 1.41 | 88.13 |
| SPalign | 163 | 1.81 | 86.52 |
| CLICK | 153 | 1.50 | 86.30 |
| MICAN | 167 | 2.01 | 83.80 |
| HOMSTRAD | 162 | 2.02 | 82.51 |
| FlexSnap | 149 | 2.08 | 76.17 |
| SALIGN* | – | 1.52 | 85.70 |
| MUSTANG* | – | 1.52 | 80.50 |

*These are detailed in Nguyen and Madhusudhan (2011).

### 2.4.1 HOMSTRAD homologous proteins

A total of 3454 structures make up the 9538 pairwise alignments in this dataset implied by the HOMSTRAD (HOMologous STRucture Alignment Database) (Mizuguchi *et al.*, 1998) database of multiple alignments. These manually curated alignments have been used to create a collection of protein families, clustered on the basis of sequence and structural similarity. This benchmark aims to comparatively assess the performance of aligning homologous structures where structurally similar regions share the same topology.

### 2.4.2 Difficult cases of aligning homologous protein pairs

This dataset is a subset of the HOMSTRAD dataset above, however focusing on alignments with low structure similarity. The dataset includes 64 pairwise alignments that have between 30% and 70% SO and RMSD above 2.5 Å.

### 2.4.3 Similar structure but different topology

This dataset includes a total of 199 pairwise alignments from 91 protein structures. These pairwise alignments are made up from 5 pairs with circular permutation, 60 pairs with non-topological similarities, 24 pairs with swapped domains and 110 alignments amongst 10 members of retinol binding proteins, 5 members of verotoxin family and 4 members of the pleckstrin family. This benchmark aims to comparatively assess the performance of aligning non-homologous structures where structurally similar regions do not share the same topology.

### 2.4.4 RIPC

This dataset was compiled by (Mayr *et al.*, 2007) and contains 40 structure pairs which are considered difficult to align due to the presence of Repititions, extensive Indels (Insertions/Deletions), Circular Permutations and/or considerable Conformational variability. The authors also provide a series of reference alignments for a subset of the RIPC set (23 pairs) that are based on sequence and function conservation: two were generated based on curated alignments of homologous proteins, three were generated by mapping the residue numbers from PDB structures corresponding to alternate conformations of the same protein and the 18 remaining are a result of searching for functionally equivalent residues (i.e. equivalent catalytic residues and/or binding sites sharing similar physicochemical environments). Of these reference alignments we analyze a subset of 20 pairs with reference alignments that have between ~10% and ~35% sequence identity (disregarding three cases with ~100% sequence identity).

## 3 Results and discussion

The results obtained by SPalignNS is compared with other methods using benchmark datasets discussed in the previous section are presented in the following subsections. To assist in interpreting these results, the following points are relevant: (i) As no GANGSTA$^+$ SO values were reported in (Guerler and Knapp, 2008), the GANGSTA$^+$ SO values reported in all tables in this paper are those detailed in (Nguyen and Madhusudhan, 2011). (ii) Some methods shown in result tables are annotated with a (*) indicating that the values for these methods were directly adapted from previous publications. (iii) In order to estimate statistical significance, assessed P-values have been generated through Wilcoxon tests performed using the R Project for Statistical Computing (R Core Team, 2014) software.

### 3.1 HOMSTRAD homologous proteins

Table 1 presents the comparative results for the 9538 pairwise protein structure alignments of the HOMSTRAD homologous proteins dataset. The table contains a detailed summary of assessed metrics, showing the average values of Nali, RMSD and SO. In addition, these metrics have been calculated for the curated HOMSTRAD alignments. Results for GANGSTA$^+$ have not been included here as this method was not run for this dataset (Nguyen and Madhusudhan, 2011).

From Table 1 it can be seen that SPalignNS achieves the highest average SO percentage with the lowest RMSD compared to the other methods. SPalignNS also achieves a statistically significant ($P$-value $< 0.05$) improvement in terms of SO when compared with each method. On average SPalignNS was able to locate two more well-aligned residues compared to CLICK. More importantly, the alignment quality improves on CLICK with a reduction of RMSD by 6%. Methods with larger Nali do not necessarily indicate higher SO. This is because SO considers only superimposed residues within 3.5 Å, thus it is entirely possible for a smaller Nali to achieve greater SO than a larger Nali. As stated by (Nguyen and Madhusudhan, 2011), 9442 out of 9538 pairwise alignments from this dataset were found to follow the topology of aligned structures. This is reflected in the results by the competitive performance of sequential methods, SPalign and SALIGN, in terms of RMSD and SO.

Figure 1 shows the 9538 pairwise SO values from SPalignNS as compared with HOMSTRAD (left) and CLICK (right). SPalignNS yields higher SO values than HOMSTRAD in 7832 cases (82%) whereas HOMSTRAD has only 263 cases better than SPalignNS. Similarly, alignments generated from SPalignNS have higher SO values than CLICK in 6820 alignments (71%) whereas CLICK have better SO values only in 972 cases. Also it can be seen that SPalignNS was able to consistently achieve SO scores greater than 40% for all alignments in this dataset.

The insert charts of Figure 1 represent the distribution of improvement showing the percentage change of structure overlap values (Δ SO (%), $x$-axis) between SPalignNS alignments and the respective method in comparison. This further demonstrates that in the majority of cases, SPalignNS aligned more residues than HOMSTRAD and CLICK within the same 3.5 Å cutoff.

### 3.2 Difficult cases of aligning homologous protein pairs

Presented in Table 2 are the comparative results for the 64 pairwise protein structure alignments of the difficult HOMSTRAD alignments dataset. It can be seen that SPalignNS again achieves the highest average SO with the lowest RMSD in comparison to all of the other methods. SPalignNS also achieves a statistically significant

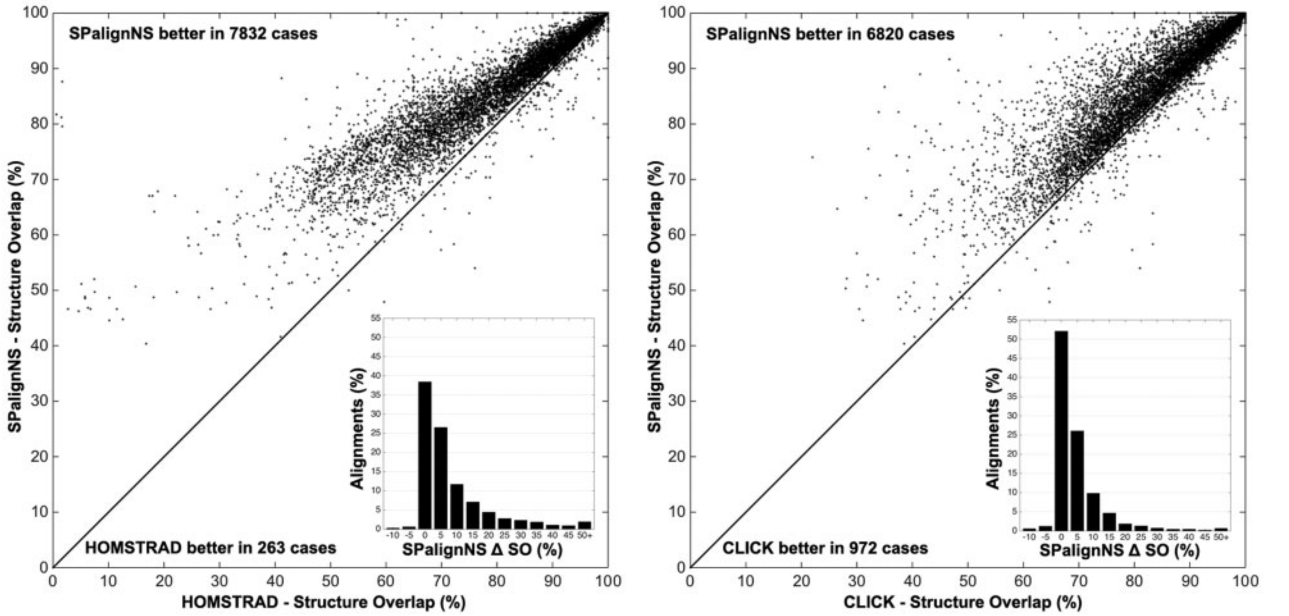

**Fig. 1.** Structure overlap performance comparison over the 9538 pairwise alignments from the '*HOMSTRAD homologous proteins*' dataset of SPalignNS against HOMSTRAD (left), and SPalignNS against CLICK (right). Insets: distribution of improvement showing percentage change of structure overlap scores for SPalignNS (*x*-axis) from comparator method with the percentage of alignments in the dataset changed by respective amount (*y*-axis)

**Table 2.** Alignment performance comparison over the 64 pairwise alignments from the 'Difficult cases of aligning homologous protein pairs' dataset showing average Nali, RMSD, SO and total CPU for each method

| | Nali | RMSD (Å) | SO (%) | CPU (s) |
|---|---|---|---|---|
| SPalignNS | 72 | 1.91 | 72.83 | 27.46 |
| SPalign | 81 | 2.66 | 69.27 | 11.18 |
| CLICK | 67 | 1.96 | 68.90 | 79.24 |
| FlexSnap | 66 | 2.23 | 61.37 | 40.75 |
| MICAN | 82 | 2.91 | 61.30 | 6.18 |
| HOMSTRAD | 81 | 3.15 | 59.40 | – |
| SALIGN* | – | 2.02 | 67.20 | – |
| DALI* | – | 2.00 | 63.00 | – |
| GANGSTA$^{+}$* | – | 1.99 | 61.90 | – |
| Geometric Hashing* | – | 1.91 | 59.50 | – |
| FATCAT* | – | 2.36 | 59.10 | – |

*These are detailed in Nguyen and Madhusudhan (2011).

($P$-value $< 0.05$) improvement in terms of SO when compared with each method. On average SPalignNS was able to locate five more well-aligned residues compared to CLICK with an RMSD reduction of 7%. For CPU cost SPalignNS demonstrates a significant reduction (approximately 65%) of required processing time compared to CLICK. The fastest methods, MICAN and SPalign, resulted with an increased RMSD (approximately 0.75 Å) compared to SPalignNS, CLICK and GANGSTA$^{+}$. This is indicative of the increased average number of aligned residues (approximately 25%) found, and since all atoms in the structures are weighted equally when calculating RMSD, the RMSD is heavily dependent on protein size and very sensitive to any badly aligned regions and local structural changes (Mizuguchi and Go, 1995). However, SPalign achieved a comparable SO result to CLICK, again this is likely due to the large number of sequential similarities present in this dataset as discussed in the previous HOMSTRAD homologous proteins dataset results.

Figure 2 presents the breakdown of individual SO results for the 64 pairwise alignments (top) where the individual SO scores from CLICK and HOMSTRAD ($x$-axis) have been plotted against respective SO scores from SPalignNS ($y$-axis). Also included is the distribution of SO improvement (bottom) that shows the percentage of pairwise alignments in the dataset changed by Δ SO (%) as compared to SPalignNS. it can be seen that SPalignNS is the only method that consistently achieves SO scores of at least 50% or more for all alignments, and there was not a single case where a HOMSTRAD alignment had a superior SO value compared to SPalignNS, with the exception of 5 cases having identical SO values. Finally, comparing SPalignNS to CLICK it can be seen the majority of alignments have been improved by at least 5% SO, and comparing SPalignNS to HOMSTRAD the majority of alignments have been improved by at least 10% SO.

### 3.3 Similar structure but different topology

Presented in Table 3 are the comparative results for the 199 pairwise protein structure alignments of the similar structure but different topology dataset. It can be seen that SPalignNS achieves the highest average SO percentage of all methods, with a comparable RMSD to the lowest RMSD achieved by CLICK differing by only 0.03 Å. However, on average SPalignNS found three more aligned residues which makes this negligible increase of RMSD acceptable. SPalignNS also achieved a statistically significant ($P$-value $< 0.05$) improvement in terms of SO over each method for this dataset. In terms of CPU cost, it can be seen that SPalignNS again demonstrates a significant reduction (approximately 65%) to required processing time compared to CLICK, and approximately a 40% reduction of CPU requirements compared to FlexSnap. MICAN, the fastest non-sequential method, has the second largest RMSD and approximately 25% less SO compared to SPalignNS.

By releasing the ordering constraint and performing alignments non-sequentially, a significant improvement of SO can be noticed from all the non-sequential methods compared to sequentially constrained alignments produced by SPalign. Therefore this confirms the presence of non-sequential structural similarities, and the ability for the tested non-sequential methods to detect them. SPalignNS was able to increase the average SO by approximately 40%, 16.5%, 14.5% and 5% compared to SPalign, GANGSTA$^+$, FlexSnap and CLICK respectively.

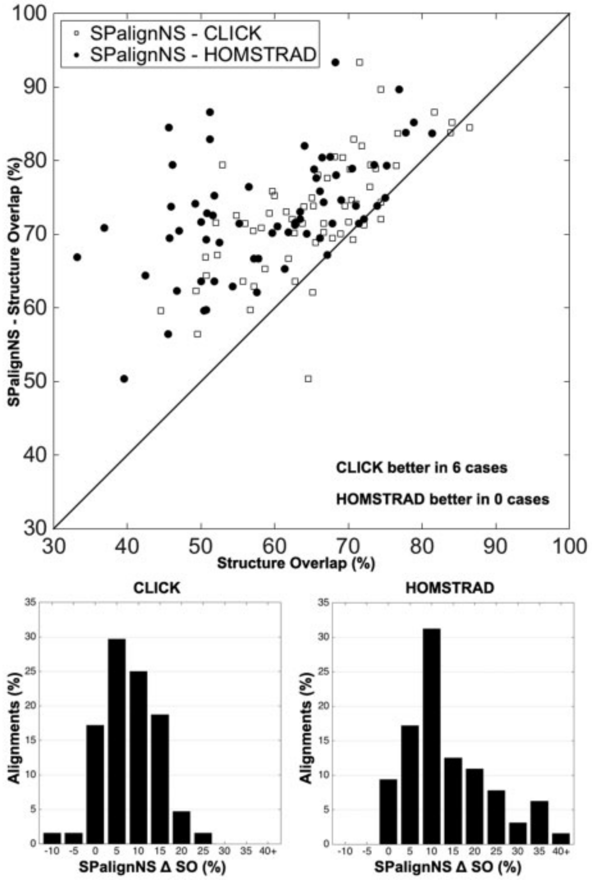

**Fig. 2.** Structure overlap performance comparison over the 64 pairwise alignments from the '*Difficult cases of aligning homologous protein pairs*' dataset. Top: the *y*-axis SO values from SPalignNS are plotted against the *x*-axis SO values from CLICK (square), and HOMSTRAD (circle). Bottom: distribution of improvement showing percentage change of structure overlap values for SPalignNS (*x*-axis) compared to CLICK (left) and HOMSTRAD (right) with the percentage of alignments in the dataset changed by respective amount (y-axis)

**Table 3.** Alignment performance comparison over the 199 pairwise alignments from the 'Similar structure but different topology' dataset showing average Nali, RMSD, SO and total CPU time for each method

| | Nali | RMSD (Å) | SO (%) | CPU (s) |
|---|---|---|---|---|
| SPalignNS | 71 | 1.93 | 71.94 | 78.28 |
| CLICK | 68 | 1.90 | 68.90 | 247.25 |
| FlexSnap | 67 | 1.83 | 62.91 | 126.51 |
| GANGSTA$^+$* | – | 2.74 | 61.70 | – |
| Geometric Hashing* | – | 1.86 | 61.30 | – |
| DALI* | – | 3.50 | 60.90 | – |
| MICAN | 77 | 3.05 | 56.56 | 21.08 |
| SPalign | 66 | 2.93 | 50.70 | 36.37 |

*These are detailed in Nguyen and Madhusudhan (2011).

Figure 3 presents the breakdown of individual SO results for the 199 pairwise alignments where the individual SO scores from CLICK (*x*-axis), have been plotted against respective SO scores from SPalignNS (*y*-axis). Also included is the distribution of SO improvement (inset) that shows the percentage of pairwise alignments in the dataset changed by Δ SO (%) as compared to SPalignNS. Here this again demonstrates that in the majority of cases, SPalignNS was able to align more residues than CLICK within 3.5 Å with the improvement to SO mostly between 5% and 10%.

### 3.4 RIPC

The RIPC set contains protein pairs with very difficult structural relations including repetitions, large InDels, circular permutations and conformational variability (Mayr *et al.*, 2007). Here, alignments generated by computational methods are compared to the provided reference alignments. Presented in Table 4 are the comparative

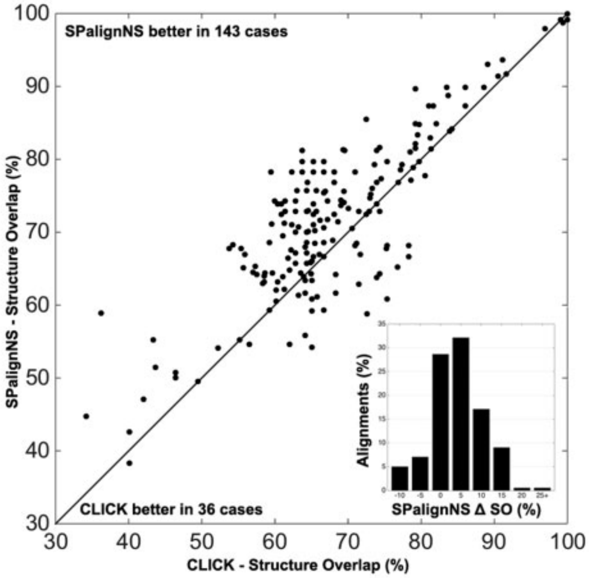

**Fig. 3.** Structure overlap performance comparison over the 199 pairwise alignments from the '*Similar structure but different topology*' dataset with SO values from SPalignNS (*y*-axis) plotted against SO values from CLICK (*x*-axis). Inset: distribution of improvement showing percentage change of structure overlap values for SPalignNS (*x*-axis) compared to CLICK with the percentage of alignments in the dataset changed by respective amount (*y*-axis)

**Table 4.** Alignment performance comparison over the 20 pairwise alignments from the 'RIPC' dataset showing average Nali, RMSD, SO; equivalent reference residue alignments (EQR) and percentage of agreement with reference alignments (Agree %); and total CPU time for each method

| | Nali/RMSD (Å)/SO (%) | EQR/Agree (%) | CPU (s) |
|---|---|---|---|
| SPalignNS | 127/1.85/65.72 | 227/80.5 | 34.11 |
| CLICK | 123/1.97/63.14 | 194/68.8 | 70.99 |
| MICAN | 146/3.06/56.52 | 184/65.3 | 6.22 |
| SPalign | 123/2.67/52.66 | 141/50.0 | 8.68 |
| FlexSnap | 87/2.02/42.71 | 117/41.5 | 82.79 |
| MASS* | – | 212/75.2 | – |
| DALI* | – | 148/52.5 | – |
| CE* | – | 135/47.9 | – |

*These are detailed in Mayr *et al.* (2007).

results for the 20 pairwise protein structure alignments of the RIPC dataset. This table shows Nali, RMSD, SO, Equivalent reference residues (EQR) and the percentage of agreement with reference alignments.

On average SPalignNS found the highest number of equivalent residues per the reference alignments with a total agreement of 80.5% (227 out of 282 residues) and has the lowest RMSD and the largest SO when compared with all other methods. It was the second fastest non-sequential method, approximately twice as fast as CLICK. Compared to the next best performing non-sequential methods, CLICK and MICAN, SPalignNS has an improvement of agreement by 10-15%. The MICAN method performs the fastest but has the largest RMSD and nearly 10% less SO than SPalignNS.

### 3.5 Case study

To illustrate the usefulness of SPalignNS, we apply it to the protein pair **2ES9–1SXJ** that have a non-sequential structure relationship as discovered by Guerler and Knapp (2008). Comparative metrics for this alignment can be found in Table 5. Presented in Figure 4 is the SPalignNS alignment (left), the SPalign alignment (middle) and the CLICK alignment (right). It can be seen here that the sequential alignment has almost successfully aligned all α helices, with exception of the far left to the far right as shown in the left portion of the figure. Interestingly, the best CLICK alignment has produced an almost identical result as SPalign, both of which achieve an SO value close to 55%. By aligning non-sequentially with SPalignNS it can be seen from the middle portion of the figure that all of the alpha helices have been matched to a corresponding helix, and SO increases significantly to 81.82% (approximately 25% improvement). The FlexSnap method was also able to achieve a comparable alignment to SPalignNS for this pair, with a slightly lower SO value of 80.80% and increased RMSD.

**Table 5.** Alignment performance comparison for protein pair **2ES9–1SXJ** showing Nali, RMSD and SO for each method

| | Nali | RMSD (Å) | SO (%) |
|---|---|---|---|
| SPalignNS | 81 | 1.87 | 81.82 |
| FlexSnap | 84 | 2.01 | 80.80 |
| CLICK | 55 | 1.95 | 55.56 |
| SPalign | 70 | 3.08 | 52.53 |
| MICAN | 70 | 3.46 | 44.44 |

## 4 Conclusion

The three dimensional structure of proteins at near atomic level resolution often gives an indication of their evolution and function. In particular, similarity in structure often implies similarity in function, so predictions of protein functional similarity can often be facilitated through structural comparison results. Given the rapid rate at which protein structures are being determined, it is crucial to have fast and accurate computational tools to classify and categorize these structures. Although there are a large number of existing sequential and non-sequential protein structure alignment methods, none of them provide an optimal structure alignment with comparable quality to manual curation for all protein pairs. As a consequence, considerable scope exists for the development of new techniques that are able to both improve alignment accuracy and minimize computational processing time.

Traditionally, the identification of similarity in structure of protein pairs has been performed using methods which preserve the sequence order of the proteins. However, in some cases this has been an overly restrictive constraint as function of a protein is often decided by only its global shape without considering its internal topology. Thus, relaxing this constraint has lead to the discovery of previously hidden structural similarities. The development of fast and accurate non-sequential protein structure alignment tools provide the potential of uncovering of new evolutionary or functional relationships between protein structures containing topological permutations.

This study presents SPalignNS, a new non-sequential protein structure alignment tool by using a novel asymmetric linear sum assignment heuristic. Specifically, the Asymmetric Greedy Search algorithm was shown to produce viable non-sequential structure alignments of proteins. The performance of SPalignNS was evaluated by performing trials with commonly used datasets containing known occurrences of proteins with non-sequential structural similarities. SPalignNS was able to achieve greater alignment accuracy when compared with sequential, as well as non-sequential alignment methods. Statistically significant *P*-values $< 0.05$ were attained when comparing the percentage of high-quality residue alignments measured by SO coverage of SPalignNS against the other methods compared across all benchmark dataset experiments. Finally, SPalignNS is generally 3 times faster than the CLICK method.

## Acknowledgements

We gratefully acknowledge the support of the Griffith University eResearch Services Team and the use of the High Performance Computing Cluster 'Gowonda' to complete this research. This research/project has also been

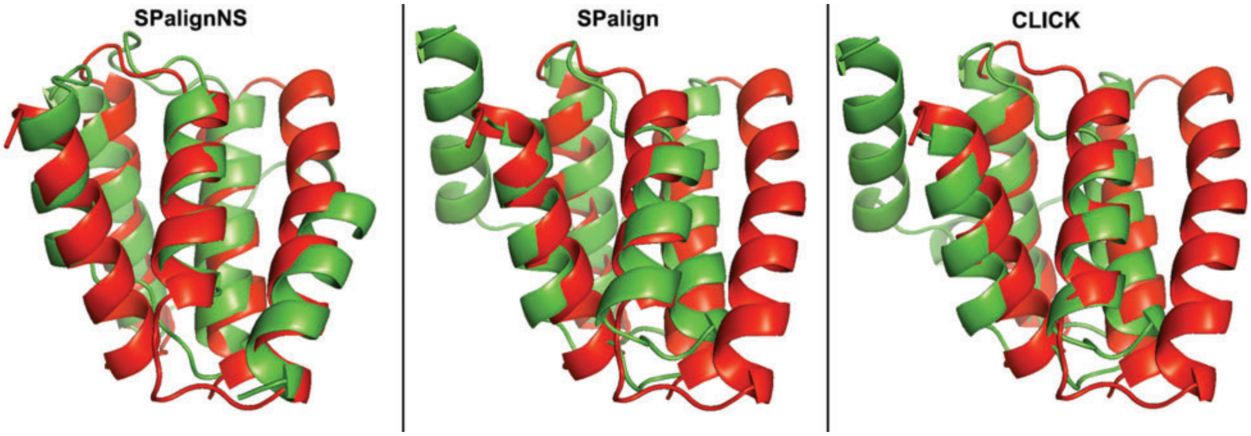

**Fig. 4.** Structure alignments of protein pair **2ES9–1SXJ**, includes non-sequential alignment from SPalignNS (left), sequential alignment from SPalign (middle) and non-sequential alignment from CLICK (right), where SPalignNS successfully aligned all five helix units whereas SPalign and CLICK align four helix units

undertaken with the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF).

## Funding

This work was supported in part by National Health and Medical Research Council (1059775 and 1083450) of Australia and Australian Research Council's Linkage Infrastructure, Equipment and Facilities funding scheme (project number LE150100161) to Y.Z., and the Microsoft Azure for Research Awarded to Y.Y.

*Conflict of Interest*: none declared.

## References

Abyzov,A. and Ilyin,V.A. (2007) A comprehensive analysis of non-sequential alignments between all protein structures. *BMC Struct. Biol.*, **7**, 78.

Alexandrov,N.N. (1996) Sarfing the pdb. *Protein Eng.*, **9**, 727–732.

Andreeva,A. *et al.* (2008) Data growth and its impact on the scop database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.

Bachar,O. *et al.* (1993) A computer vision based technique for 3-d sequence-independent structural comparison of proteins. *Protein Eng.*, **6**, 279–287.

Berbalk,C. *et al.* (2009) Accuracy analysis of multiple structure alignments. *Protein Sci.*, **18**, 2027–2035.

Bertsekas,D.P. (1988) The auction algorithm: a distributed relaxation method for the assignment problem. *Ann. Oper. Res.*, **14**, 105–123.

Burkard,R.E. and Cela,E. (1999) Linear assignment problems and extensions. In: Du,D.-Z. and Pardalos,P.M. (eds), *Handbook of Combinatorial Optimization*. Springer, US, pp. 75–149.

Burley,S.K. (2000) An overview of structural genomics. *Nat. Struct. Mol. Biol.*, **7**, 932–934.

Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823.

Dror,O. *et al.* (2003a) Mass: multiple structural alignment by secondary structures. *Bioinformatics*, **19**, i95–i104.

Dror,O. *et al.* (2003b) Multiple structural alignment by secondary structures: algorithm and applications. *Protein Sci.*, **12**, 2492–2507.

Dundas,J. *et al.* (2011) Sequence order independent comparison of protein global backbone structures and local binding surfaces for evolutionary and functional inference. In: Kihara,D. (ed.), *Protein Function Prediction for Omics Era*. Springer, Netherlands, pp. 125–143.

Greene,L.H. et al. (2007) The cath domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.*, **35**, D291–D297.

Guerler,A. and Knapp,E.-W. (2008) Novel protein folds and their nonsequential structural analogs. *Protein Sci.*, **17**, 1374–1382.

Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.

Kolbeck,B. *et al.* (2006) Connectivity independent protein-structure alignment: a hierarchical approach. *BMC Bioinf.*, **7**, 510.

Konagurthu,A.S. *et al.* (2006) Mustang: a multiple structural alignment algorithm. *Proteins Struct. Funct. Bioinf.*, **64**, 559–574.

Kuhn,H.W. (1955) The Hungarian method for the assignment problem. *Naval Res. Logistics Q.*, **2**, 83–97.

Ma,J. and Wang,S. (2014) Algorithms, applications, and challenges of protein structure alignment. *Adv. Protein Chem. Struct. Biol.*, **94**, 121–175.

Madhusudhan,M. *et al.* (2009) Alignment of multiple protein structures based on sequence and structure features. *Protein Eng. Des. Select.*, **22**, 569–574.

Mayr,G. *et al.* (2007) Comparative analysis of protein structure alignments. *BMC Struct. Biol.*, **7**, 50–65.

Minami,S. *et al.* (2013) Mican: a protein structure alignment algorithm that can handle multiple-chains, inverse alignments, c α only models, alternative alignments, and non-sequential alignments. *BMC Bioinf.*, **14**, 24–47.

Mizuguchi,K. and Go,N. (1995) Seeking significance in three-dimensional protein structure comparisons. *Curr. Opin. Struct. Biol.*, **5**, 377–382.

Mizuguchi,K. *et al.* (1998) Homstrad: a database of protein structure alignments for homologous families. *Protein Sci. Publ. Protein Soc.*, **7**, 2469.

Naiem,A. and El-Beltagy,M. (2009) Deep greedy switching: a fast and simple approach for linear assignment problems. In: *7th International Conference of Numerical Analysis and Applied Mathematics*.

Naiem,A. and El-Beltagy,M. (2013) On the optimality and speed of the deep greedy switching algorithm for linear assignment problems. In: *Proceedings of the 2013 IEEE 27th International Symposium on Parallel and Distributed Processing Workshops and PhD Forum*, pp. 1828–1837.

Nguyen,M.N. and Madhusudhan,M.S. (2011) Biological insights from topology independent comparison of protein 3D structures. *Nucleic Acids Res.*, **39**, e94–e94.

Nguyen,M.N. *et al.* (2011) CLICK topology-independent comparison of biomolecular 3D structures. *Nucleic Acids Res.*, **39**, W24–W28.

Orengo,C.A. and Taylor,W.R. (1996) SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.*, **266**, 617–635.

Pandit,S.B. and Skolnick,J. (2008) Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. *BMC Bioinf.*, **9**.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Salem,S. *et al.* (2009) Iterative non-sequential protein structural alignment. *J. Bioinf. Comput. Biol.*, **7**, 571–596.

Salem,S. *et al.* (2010) Flexsnap: flexible non-sequential protein structure alignment. *Algorithms Mol. Biol.*, **5**, 12.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Shih,E.S. and Hwang,M.-J. (2010) Non-sequential protein structure comparisons. In: *Sequence and Genome Analysis: Methods and Applications*. iConcept Press, pp. 63–76, https://www.iconceptpress.com/about-us/.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.

Strickland,D.M., Barnes,E., and Sokol,J.S. (2005) Optimal protein structure alignment using maximum cliques. *Oper. Res.*, **53**, 389–402.

Yang,Y. *et al.* (2012) A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic-acid binding prediction. *Proteins*, **80**, 2080–2088.

Ye,Y. and Godzik,A. (2004). Fatcat: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res.*, **32**, W582–W585.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.