TALI: Protein Structure Alignment Using Backbone Torsion Angles

Xijiang Miao
Computer Science and Engineering
University of South Carolina
Columbia, SC, USA

Michael G. Bryson
Computer Science and Engineering
University of South Carolina
Columbia, SC, USA

Homayoun Valafar
Computer Science and Engineering
University of South Carolina
Columbia, SC, USA

Abstract - *This article introduces a novel protein structure alignment method (named TALI) based on protein backbone torsion angle instead of the more traditional distance matrix. Representing protein structure by a serial backbone torsion angles (φ, ψ), protein structure have a simple mapping relationship to protein sequence. Thus, TALI can naturally incorporate sequence information and sequence analysis method into structure comparison. Here we report the result of TALI in comparison to other structure alignment methods such as DALI, CE and SSM as well as sequence alignment based on PSI-BLAST. TALI demonstrated great success over all other distance based methods in application to challenging remote homologous proteins. Finally, successful inference of phylogeny tree of class II amonoacyl-tRNA synthetase shows the capability of TALI in estimating the protein structure evolution.*

Keywords: protein structure; structural similarity; molecular evolution; torsion angle; Ramachandran space; structure alignment.

1 Introduction

The number of known protein structures is expanding in an exponential fashion. At the current rate, the PDB[1] database size will exceed 100,000 structures by the end of this decade. With the dazzling accumulation of protein structures, sophisticated and systematic analysis based on structural information becomes more and more feasible. Evolutionary relationship of organisms based on protein structures, protein function and protein-protein interaction are among examples of these types of analyses. Establishing the structural diversity among the cataloged library of structures can be viewed as the foundation of all of these analyses. Therefore the first step in analysis of protein structures is the establishment of structural relationship (similarity) between any two given proteins. In that regard, invaluable efforts have been made to classify and organize the protein structures, such as DALI (FSSP)[2, 3], CATH[4], and SCOP[5].

Several methods such as SSAP[6], DALI, CE[7], MAMMOTH[8] and SSM[9] have been developed to recognize the common core structural elements between protein candidates. Whether based on secondary structure configuration, distance matrix, or a pipeline of such procedures, these methods emphasize the overall match of cores in 3D space while overlooking the detail structure pattern information.

Here we present a new method of structure alignment named TALI (Torsion Angle Alignment) that utilizes backbone torsion angles instead of a matrix of distances. The backbone structure of any protein can be represented by two major methods: distance matrix between designated atoms along the backbone (such as C_α) or the backbone torsion angles. The information in distance matrix is highly redundant; notably, the matrix is often sparsely populated in the diagonal region and small islands corresponding to regions in close spatial vicinity. To compare distance matrixes of two proteins at a reasonable speed, multiple steps are often used with gradually finer granularity and large amount of far inter-atomic distances are ignored. Different from distance matrix, the torsion angle (φ, ψ) representation uses a series of dihedral angles that specifies the spatial relationship between peptide planes. Under theoretical conditions, a given list of torsion angles can be utilized to reliably reconstruct local backbone structure of a protein. However, a slight change in torsion angles can result a significantly different structure. The sensitive relationship between torsion angles and structure can enable us to perform detailed structural comparisons. The properties and statistics of torsion angles are routinely used for protein quality check[10] by observing the distribution of the (φ, ψ) otherwise known as the Ramachandran plot[11].

2 Methods

2.1 Structure distance based on torsion angles

The critical component of our implementation consists of an appropriate definition of torsion angle based structure distance. In the absence of statistical data or theory describing the relationship between evolutionary divergence and torsion angle difference, the distance between any two given pairs of torsion angles can be reasonably adapted from the distance between two

points in a multidimensional space. Our first implementation relied on the most simplistic measurement of distance denoted as the "Plain TALI score" as shown in equations 1 and 2. Here $\varphi_{a,i}$ is the dihedral angle between the backbone C-N-Ca-C atoms of residue i of protein a. d_{ij} is the torsion angle distance measured by the Euclidean distance between two pairs of torsion angles (φ, ψ) (only matched pairs are counted). N is the total length of aligned substructures. The quantities $|\varphi_{a,i} - \varphi_{b,j}|$ and $|\psi_{a,i} - \psi_{b,j}|$ should be less than 180°, otherwise the 360° complementary angle is used.

$$d_{ij} = \begin{cases} \sqrt{(\varphi_{a,i} - \varphi_{b,j})^2 + (\psi_{a,i} - \psi_{b,j})^2} & , \forall\, a_i \equiv b_j \\ NA & , otherwise \end{cases} \tag{1}$$

$$TS_p = \sqrt{\frac{1}{N}\sum_{ij} d_{ij}^2} \tag{2}$$

Plain TALI score TS_p treats all torsion angle changes equally. For example, a change within the α-helical region of Ramachandran space or a change from helical to beta strand region is not penalized differently. Albeit the results obtained based on this simplistic metric were very encouraging, here we do not report these results and focus on a more meaningful measure of distance.

The mentioned insensitivities of the trivial scoring mechanism can be remedied by redefining distances based on statistical information available from the Ramachandran plot. This addition will incorporate empirical likelihood of observing a particular pair of torsion angles. Here we define a function $R(t)$, which returns the log density at point t in Ramachandran space. The distance between any $(\varphi, \psi)_i$, $(\varphi, \psi)_j$ pair is defined in equation 3.

$$d_{ij}^r = \int_L R(l)dl \tag{3}$$

Where L is a straight path connecting two points $(\varphi, \psi)_i$ and $(\varphi, \psi)_j$. d_{ij}^r represents the integration of torsion angle density along the path L. c is a constant around 180 (c=144 in this article). Intuitively, this Ramachandran distance analogues to measuring the total height of all steps in hill climbing with fixed step length. The final score summarizing the similarity between two pairs of torsion angles TS_R is defined as shown in equations 4 and 5:

$$D_{ij} = \begin{cases} \exp(-d_{ij}^r / c) & , \forall\, a_i \equiv b_j \\ NA & , otherwise \end{cases} \tag{4}$$

$$TS_R = \frac{1}{N}\sum_{ij} D_{ij} \tag{5}$$

The $\exp(-d)$ function shown in Eq (4) gives TS_R more sensitivity to detect protein structure change. If d is viewed as transition energy of two torsion angles, $\exp(-d/c)$ can be viewed as the chance making such transition. This is a prelude to our future direction in implementation of distances based on Boltzmann distribution of energies between the source and destination geometries. This algorithm can be greatly accelerated by pre-calculating the distances between discrete torsion angles.

2.2 Search for optimal structure alignment

Our formulation of the structural alignment problem is conceptually very similar to that of the sequence alignment. While a traditional sequence alignment aligns character string, our approach aligns sequences of two-dimensional numbers (torsion angles). Therefore, we have adapted a generalized Smith-Waterman algorithm [12] to find the best alignment with minimum TS_R score in the place of PAM or BLOSUM substitution matrix.

2.3 Phylogeny tree generation

Given a set of remote homologous protein domains, their evolutionary relationship can be inferred from TS_R. A distance matrix with TS_R is generated for every pair of protein domain using the torsion angle alignment. Neighbor joining (NJ) [13] tree is applied on the distance matrix. The additive property of the distance is not guaranteed in structural alignment. Therefore the NJ tree only loosely reflects the homologous relationship between protein domains.

3 Results

We have utilized results from SSM, DALI and CE as representatives of distance based methods aligning protein structures for comparison reasons. The results of DALI and CE are not shown since they are quite similar to SSM in the test cases reported here. In addition, PSI-BLAST [14] has been selected to provide sequence based alignment.

3.1 Alignment of protein structures using TALI

This section illustrates the alignment result of protein 1nj8 [15] chain D and 1b76 [16] chain A by torsion angle alignment without referring to sequence or distance information. The percentage of sequence identity between 1nj8d and 1b76a is 17.5% using global alignment. The alignment quality is checked against PSI-BLAST [14] and SSM [9]. PSI-BLAST collects position specific scoring matrix (PSSM) from previous BLAST search result. The iterative process enables PSI-BLAST to find more distant relationships. SSM is based on matching graphs of protein's secondary structure elements, followed by an iterative 3D spatial alignment of protein backbone C_α atoms based on inter-atomic distances. The alignment result is good enough to verify the test cases used in this paper. Results reported by DALI and CE are also checked for quality evaluation.

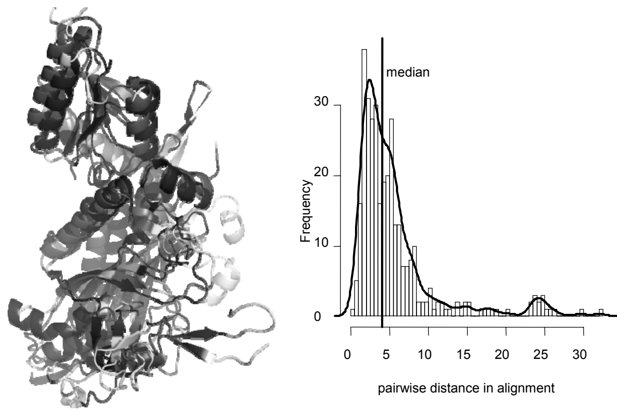
The alignment 1nj8d and 1b76a is carried out independently by PSI-BLAST, TALI, and SSM as shown in Figure 1. In this figure, the alignment with black backgroud indicates segments confirmed by another or all other alignments, while gray background indicates confirmed alignment with one amino acid "shift". The gaps and lowercase letters indicate no match. The bar on top of alignment shows the similar or identical (".", ":", or "*") amino acid pairs which is coded by clustalw[17]. PSI-BLAST is performed through NCBI online query with default parameter settings using 1nj8d as query string. TALI uses gap open cost 600, gap extension 200, and Ramachandran distance as described in method section. SSM data is collected through EBI online server with default parameter settings, that is, match connectivity and normal precision.

```
1) PSI-BLAST:
        :*:*: :: * * :***:*.::**: :  *  : * **.:  :*   *  **  :* . . *: :.  :  *.*: : *. . * :.*          :*:  *****  :::*** :
264 HGDEKGLI-LPPIVAPIQVVIVPLIFKGKEDIVMEKAKEIYEKL--KGKFRVHIDDRDIRPGRKFNDWEIKGVPLRIEVGPKDIENKK---------ITLFRRDTME--KFQVDETQ
330 NGEERIVLKLKPQLAPIKVAVIPLVKNRPE--ITEYAKRLKARLLALGLGRVLYEDTG-NIGKAYRRHDEVGTPFAVTVDYDTIGQSKDGTTRLKDTVTVRDRDTMEQIRLHVDELE

2) TALI:
1   MLEF-SEWYSDILEKAEIY--D-VRYP--IKGC-GVYLPYGFKIRRYTFEIIRNLL-DE--SGHDEALFPMLIPEDLLAKE------AEHIKG---FE---DE---VYWVT---HGGKT-
1   -AASSLDELVALCKRRGFIFQSSEI-YGGLQG-VYDYGPLGVELKNNLKQAWWRRNVYERDD-MEGLDASVLTHRLVLHYSGHEATFA--DP-MVDNAKARYWTPPRYFNMMFQDL--RG

91  ---QLDVK--LALRP-TSETPIYYMMKLWVKVHTD-LP--IKIYQIVNTFRYETKHTRPL--IRLREIMTFKEAHTAHSTKEEAENQVKEAISIYKKFFDTLGIP---YLISKRPEW-DK
112 PRGGR-G-LLAYLRPET-AQGIFVNFKNV-L-DATSRKLGFGIAQIGKAFRNEI-TPRN-FIFRVREF-EQMEIEYFV-RPGEDEYWHRYWVEERLKWWQEMGLSRENLVPYQQPPESSA

197 FPG-AEYTMAFDTIFPDGR-TMQIATVHN-LGQNFSKTF--E--------IIFETP--T------G----DKDYAYQTCYGISDRVIASIIAIH-----GDEK---G--LILPPIVAPIQ
222 -H-YAKATVDILYRFPH--GSLELEGIAQRT-DFDLGSHTKDQEALGITARVLRNEHSTQRLAYRDPETGKWFVPYVIEPSAGVDRGVLALLAEAFTREELPNGEERIVLKLKPQLAPIK

282 VVIVPLI-F-KGKEDIVMEKAKEIYEKLKG-KFRVHIDDRDI-RPGRKFNDWEIKGVPLRIEVGPKDI--------EN-KKITLFRRDTMEKFQVDETQLMEVVEKTLNNIMENIKNRAW
338 VAVI--PLVKNR-PEITEYAKRLKARLLALGLGRVLYE--DTGNIGKAYRRHDEVGTPFAVTVDYDTIGQSKDGTTRLKDTVTVRDRDTMEQIRLH------------------VDELEG

389 EKFENFITILEDINPDEIKNILSEKRGVILVPFKEEIYNEELEEKVEATILGETEYKGNKYIAIAKTY
435 FLRERL--------------------------------------------------------------RW

3) SSM:
1   mlEFSEWYSDILEKAEIYDVrY-----PIKGCGVYLPYGFKIRRYTFEIIRNLLD-ES-GHDEALFPMLIPEDLLAKEa-eHIkGFe----------------DEVYwVTHGGkTQLDV-
1   aaSSLDELVALCKRRGFIFQ-SseiygGLQGVYDYGPLGVELKNNLKQAWWRRNVyERdDMEGLDASVLTHRLVLHYSgheATfADpmvdnakarywtppryfNMMF-QDLRGpRGGRGl

96  KLALRPTSETPIYYmmklw-vkvhtDLPIKIYQIVNTFRYETKhT--RpLIRLReIMTFKEAHTAHSTKeEAENQVKEAISIYKKFFDTLGIP---YLiSKRPE-WDKFPGAEYTMAFDT
119 LAYLRPETAQGIFVnfknvldatsrKLGFGIAQIGKAFRNEIT-PrnF-IFRVR-EFEQMEIEYFVRPG-EDEYWHRYWVEERLKWWQEMGLSrenLV-PYQQPpESSAHYAKATVDILY

209 IFPDgRTMQIATVHNLGQNFSKTFE---------------------IIFE-TPT-GdKDYAYQTCYGIS-DRVIASIIAIHGDE--------KGLI-LPPIVAPIQVVIVPLIfkgkEDI
234 RFPH-GSLELEGIAQRTDFDLGSHTkdqealgitarvlrnehstqrLAYRdPETgK-WFVPYVIEPSAGvDRGVLALLAEAFTReelpngeeRIVLkLKPQLAPIKVAVIPLVk--nRPE

296 VMEKAKEIYEKLKG-KFR-VHIDDRdIRPGRKFNDWEIKGVPLRIEVGPKDIen---------kKITLFRRDTmEKFQVDE-TQLMEVVEKTLnnimeniknrawekfenfitiledinp
350 ITEYAKRLKARLLAlGLGrVLYEDT-GNIGKAYRRHDEVGTPFAVTVDYDTIgqskdgttrlkdTVTVRDRDT-MEQIRLHvDELEGFLRERLrw-----------------------

404 deiknilsekrgvilvpfkeeiyneeleekveatilgeteykgnkyiaiakty
442 -----------------------------------------------------
```

Figure 1 Align 1nj8D and 1b76A with PSI-BLAST, TALI and SSM. Black background indicates alignments confirmed by another or all methods; gray background indicates with one amino acid shift. 1) NCBI PSI-BLAST result with default parameters. 2) TALI result with gap open cost = 600, extension cost = 200, Ramachandran distance. 3) SSM result with default parameters. i.e., match connectivity and normal precision.

Generally speaking, these three methods, which are based on sequence, torsion angle and distance strategies, agree quite well with each other. PSI-BLAST picked a region out of the whole protein with sequence identity as high as 29%. Although it is from sequence alignment, the result is in agreement with the results reported by SSM and TALI. TALI and SSM easily report larger aligned regions with similar degree of agreement. TALI alignment segments experience a "shift" in comparison to the results of SSM. This phenomenon happens in regions where the torsion angles are similar to each other. For example, α-helix is composed of constant torsion angles around (−65.3°, −39.4°)[18]. The quality of alignment can be improved by incorporating secondary structural information, physicochemical properties of each residue or even sequence similarity matrix such as PAM250. Please refer discussion section for detail explanation.

The 3D representation of alignment is shown in Figure 2a. The two proteins are correctly oriented according to the alignment result of TALI. Although the visual effect is satisfying, the root mean square deviation, or RMSD, is as high as 7.79Å compared with 2.57Å reported by SSM (matched pairs = 341). As shown in Figure 2b, the distribution of distances between aligned residues is highly skewed. Simply dropping the aligned residues with farthest distances can greatly improve the RMSD with little sacrifice of total number of matched residues. For example, by dropping residue pairs associated with top 25% largest distances, the RMSD can be dropped to 3.62Å. Reorientation of the proteins based on the reduced alignment, RMSD can be further improved. Although RMSD is considered to be an important metric for alignment quality, it is proper only when the protein is assumed to be rigid.



(a) 3D representation (b) Distribution of distances

Figure 2 Alignment of 1nj8d and 1b76a from TALI. 1) 3D representation. The darker molecule is for 1b76a (Picture is prepared by PyMol). 2) Histogram and distribution of interatomic distances in the alignment. Median of distances is 4.09Å. Initial RMSD = 7.79Å. By dropping distances > 4.09Å, RMSD becomes 2.55Å; dropping top 25% distances (> 6.19Å), RMSD becomes 3.62Å. Total number of matched pairs is 334.

3.2 Performance of torsion angle alignment

Fischer *et al.* [19] have presented 68 pairs of sequence-structure alignment test cases with different percentages of sequence identity and difficulty index for benchmarking alignment algorithms. Figure 3 shows the alignment result of 1cewi [20] against 1mola [21], which is one of the top 10 most difficult test cases in [19]. 1cew and 1mol both are proteinase inhibitor with alpha-beta structure. For TALI, it is difficult to exactly align the structure because the torsion angles at the beta hairpin regions can not provide enough featured patterns to make a unique reliable alignment. This problem can be remedied by implementation of more analysis features discussed in the discussion section.

As a closely related example with obvious structural homology, 1cewi and 1r4ca [22] show that TALI outperforms SSM by correctly aligning the whole protein domain. Protein human cystatin C (HCC) inhibits papin-like cystein protease. Oligomerization of HCC leads to amyloid deposits. The N-terminal truncated variance of HCC (THCC, PDB id 1r4c) lacks the first 10 amino acid residues of the native sequence. The aggregation of THCC takes place through swapping domain. According to CATH [4] v2.6 classification, both 1cew and 1r4c belong to the sequence family 3.10.450.10.2 (proteinase inhibitor, cysteine). The protein sequence similarity is 44%, which strongly suggests their homologous origin. As shown in Figure 3, structure alignment from TALI correctly recovered the whole sequence alignment. The shaded region shows the parts discovered by both TALI and SSM, while box shows the parts with high sequence similarity omitted by SSM but captured by TALI. From Figure 4a, we can conclude that one beta hairpin structure acts like a hinge. The remaining two subunits linked by this hinge remain relatively stable. DALI and CE show similar alignment results with SSM. Observation of mere distance relationship will miss the featured pattern in different subunit. The concept of protein domain conceptually divided the whole protein into small relatively independent units, which will help improve the quality of distance based alignment. But the protein domain boundary does not always have a unique clear cutoff [23]. Like sequence alignment, TALI does not rely much on the domain boundary definition; a properly defined protein chain is enough for alignment.

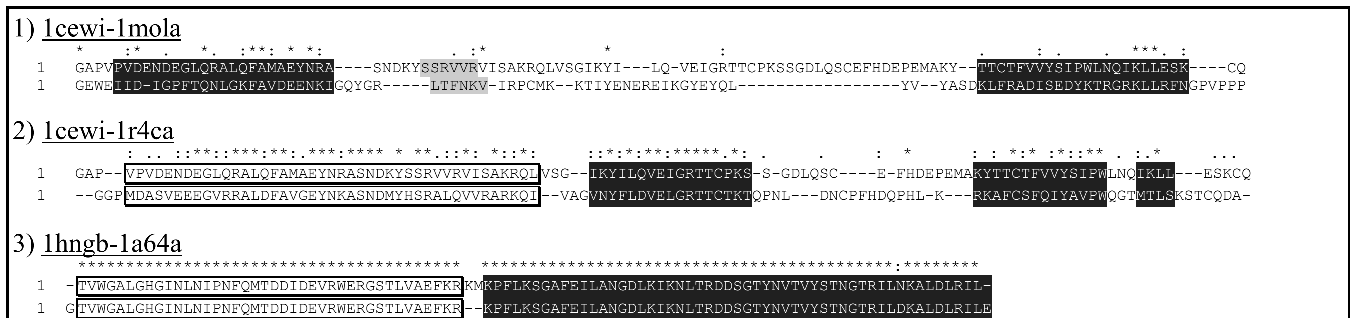


Figure 3 Three pairs of small proteins show pros and cons of torsion angle alignment. 1) 1cewi and 1mola. 2) 1cewi and 1r4ca. It is worth noting that they have more sequence identity. 3)1hngb and 1a64a. The protein sequences are almost identical. Shaded region indicates alignment recovered by TALI and SSM; Box indicates alignment recovered only by TALI.

Figure 4 (b) shows an extreme example, 1hngb [24] and 1a64a [25]. The amino-terminal domain of cell adhesion module CD2 (PDB id 1hng) can fold as monomer or metastable dimer. Murray et al. [25] engineered the protein by introducing a hinge-deletion mutant, which mimic the spontaneous random mutation during molecular evolution. The structural alignment by TALI correctly recovers the sequence alignment, while SSM does not find the first part of the structure (shown in box in Figure 3 (3), 3D alignment is shown in Figure 4(b)). DALI or CE can not recover whole regions either.

Comparing the alignment result with DALI fold classification, we can often observe some examples that TALI performs better. For example, 1q59a and 1g5ma are related to BCL-2 family, while 1aa7a and 1mdta are put together by DALI but without any known biological support.

(a) 1cewi, 1r4ca

(b) 1hngb, 1a64a

Figure 4 3D representation of alignment of 1cewi-1r4ca and 1hngb-1a64a. In both cases, the beta hairpin structure acts like a hinge. The variance form is generated by open the hinge. The detail structure of the two parts remains relatively stable.

3.3 Infer phylogeny from Torsion angle alignment

To demonstrate the possibility of recovering phylogeny using TALI result, we choose the protein phylogeny discussed by O'Donoghue and Schulten [26]. We use chain rather than domain as the taxa of the phylogeny. Use the chain information can avoid the artifact of domain boundary definition. And it will also be interesting to know the evolution history of specific chains, which is expected to be similar to the result of [26].

Figure 5 shows the result of TALI phylogeny for class II amonoacyl-tRNA synthetase, where the cost of opening a gap is set to 600 and extension cost is set to 200. The original tree appears as FIG 10, FIG 12 in [26] and its supplementary material. Each sub-class is shown by branch with thick lines.

Comparing the phylogeny tree derived by TALI and the original tree, the overall structures agree with each other quite well. For structures within subclasses, the difference is very subtle. For example, in branch D_b, the relationship is

(1eqrb, (1eqrc, (1eqra, (1c0aa, 1il2a))))

in TALI, while [26] gives

((1eqrb, 1eqrc), (1eqra, (1c0aa, 1il2a))).

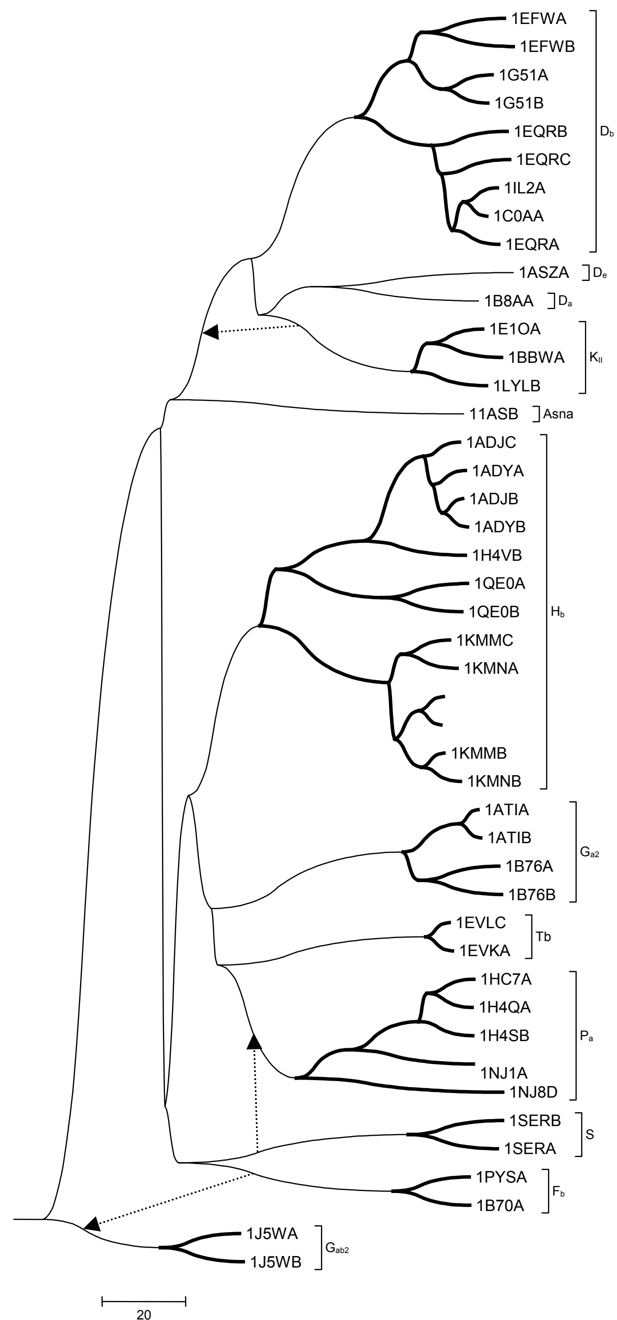


Figure 5 Full structural dendrogram of class II amonoacyl-tRNA synthetase. The chains are taken from supplement material FIG 2 of [26]. Neighbor joining (NJ) tree inferred from TALI score matrix. Tree is prepared by Mega3.

Because the sequence identity is relatively high at subclass level, the disagreement can be resolved by using sequence alignment.

For the relationship between classes, the only differences lie in the position of K_{II}, S and F_b. The corresponding position is marked by dashed arrows. There is no miss-placed protein chain between sub-classes.

4 Discussion

4.1 TALI can align complex protein structures based on torsion angle information alone

Torsion angles (φ, ψ) carry the majority of the backbone structural information. Although far more parameters are required to precisely define the relative position of atoms of protein backbone, most of the parameters can be treated to be constant. For example, the dihedral angle of ω can be assumed to be 180° with standard deviation 5.8°, which maintains the planarity of the peptide plane. Other parameters, such as bond lengths, protein backbone angles, etc., are in similar case.

While a list of torsion angles may not be sufficient in precise reconstruction of the protein backbone in 3D space due to accumulation of error, alignment based on torsion angles is immune to this problem since the alignment occurs locally without considering the relationship to amino acids far away in sequence. As a consequence, mutation of a few amino acids that would normally dramatically change the distance relationship could preserve most of torsion angle relationship. In this case, torsion angle can align the protein structures quite well, as shown in Figure 4 (a) and (b).

4.2 Improvement of TALI

Through the course of structure alignment based on backbone torsion angles, some "shift" in alignment may occur due to low structural complexity of the region. We propose an adaptation of the following strategies to mitigate this effect:

1) Landmarks can be set at coil regions to guide the pure torsion angle alignment. This special treatment of torsion angles at coil region will make it focus on small segments which will increase the precision of alignment.
2) Sequence information and analysis methods can be used to enrich the information used for torsion angle alignment. Incorporating sequence information is very natural for TALI. Except at the end of a polypeptide, there is exactly one pair of torsion angles (φ, ψ) per residue. The additional information per residue can be used as higher dimensional information added on the top of the torsion angles which will help capture more featured pattern and make a better alignment. Because the sequence alignment is well established for phylogeny analysis, the torsion angle alignment can readily borrow information from the result of sequence alignment, which is very useful to find the link among homologous proteins. Moreover, torsion angle alignment can also directly adapt methods from sequence alignment to extract out more meaningful structure information. For example, the torsion angle "motif" can be derived from a multiple alignment of torsion angles of homologous protein, which is expected to be a generalization of basic secondary structure elements (SSE) of protein, namely, α-helix and β-strands. Finally, our structure alignment can be used as a based to further refine structural alignment or sequence alignment.

4.3 Evolutionary analysis using TALI

The mutation and natural selection make proteins inherited offspring different from their ancestor in terms of sequence, structure and function. Homologous proteins and nucleotide sequences lose their similarity as time goes by. For example, according to Jukes-Canter model [27], nucleotide sequences of length N are expected to have $\frac{3}{4}(1-e^{-4\alpha T})\times N$ sites with observable change after T units of time past the divergence point (with substitution rate α). If the sequence identity is only 25%~30%, or within the "twilight zone", it will be difficult to reliably estimate the protein evolutionary history using sequence comparison alone. The protein structure is considered to be more reliable in this case.

Protein structure similarity is considered to have two driving forces: physicochemical constrain and conservative evolution [28]. Zeldovich and his colleges [29] show that the physical constraint will drive different randomly generated sequences to same or very similar folds during simulated mutagenesis process; the sizes of these stable folds or "wonderfolds" are highly uneven, which is consistent with the uneven size of protein superfamily. The physicochemical induced protein structural similarity suggests that reliable protein homologue should be detected through sequence or detail structure pattern, which is not likely to be reproduced by chance in molecular evolution.

To infer phylogeny, the method of extract common core of structural similarity and make comparison is not reliable. Only compare the conserved cores are misleading due to two reasons:

1) The core is usually very stable and strongly selected, which leads to a converged evolution. That is, structures with different sources can have similar structure and possibly similar function.
2) The structure comparison overlooks some detail structures to make a better match, which can be used as a source of homology detection. So this conserved core region is not an ideal material to study evolution.

5 References

[1] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The Protein Data Bank: a computer-based archival file for macromolecular structures." *J Mol Biol*, vol. 112, no. 3, pp. 535–542, May 1977.

[2] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices." *J Mol Biol*, vol. 233, no. 1, pp. 123–138, Sep 1993. [Online]. Available: http://dx.doi.org/10.1006/jmbi.1993.1489

[3] L. Holm and J. Park, "DaliLite workbench for protein structure comparison." *Bioinformatics*, vol. 16, no. 6, pp. 566–567, Jun 2000.

[4] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, "CATH–a hierarchic classification of protein domain structures." *Structure*, vol. 5, no. 8, pp. 1093–1108, Aug 1997.

[5] L. L. Conte, B. Ailey, T. J. Hubbard, S. E. Brenner, A. G. Murzin, and C. Chothia, "SCOP: a structural classification of proteins database." *Nucleic Acids Res*, vol. 28, no. 1, pp. 257–259, Jan 2000.

[6] W. R. Taylor and C. A. Orengo, "Protein structure alignment." *J Mol Biol*, vol. 208, no. 1, pp. 1–22, Jul 1989.

[7] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path." *Protein Eng*, vol. 11, no. 9, pp. 739–747, Sep 1998.

[8] A. R. Ortiz, C. E. M. Strauss, and O. Olmea, "MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison." *Protein Sci*, vol. 11, no. 11, pp. 2606–2621, Nov 2002.

[9] E. Krissinel and K. Henrick, "Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions." *Acta Crystallogr D Biol Crystallogr*, vol. 60, no. Pt 12 Pt 1, pp. 2256–2268, Dec 2004. [Online]. Available: http://dx.doi.org/10.1107/S0907444904026460

[10] R. Laskowski, M. MacArthur, D. Moss, and J. Thornton, "Procheck: A program to check the stereochemical quality of protein structures," *J. Appl. Cryst.*, vol. 26, pp. 283–291, 1993.

[11] G. N. Ramachandran and V. Sasisekharan, "Conformation of polypeptides and proteins." *Adv Protein Chem*, vol. 23, pp. 283–438, 1968.

[12] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences." *J Mol Biol*, vol. 147, no. 1, pp. 195–197, Mar 1981.

[13] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees." *Mol Biol Evol*, vol. 4, no. 4, pp. 406–425, Jul 1987.

[14] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res*, vol. 25, no. 17, pp. 3389–3402, Sep 1997.

[15] S. Kamtekar, W. D. Kennedy, J. Wang, C. Stathopoulos, D. Söll, and T. A. Steitz, "The structural basis of cysteine aminoacylation of tRNAPro by prolyl-tRNA synthetases." *Proc Natl Acad Sci U S A*, vol. 100, no. 4, pp. 1673–1678, Feb 2003. [Online]. Available: http://dx.doi.org/10.1073/pnas.0437911100

[16] J. G. Arnez, A. C. Dock-Bregeon, and D. Moras, "Glycyl-tRNA synthetase uses a negatively charged pit for specific recognition and activation of glycine." *J Mol Biol*, vol. 286, no. 5, pp. 1449–1459, Mar 1999. [Online]. Available: http://dx.doi.org/10.1006/jmbi.1999.2562

[17] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." *Nucleic Acids Res*, vol. 22, no. 22, pp. 4673–4680, Nov 1994.

[18] L. Wang and B. R. Donald, "Exact solutions for internuclear vectors and backbone dihedral angles from NH residual dipolar couplings in two media, and their application in a systematic search algorithm for determining protein backbone structure." *J Biomol NMR*, vol. 29, no. 3, pp. 223–242, Jul 2004. [Online]. Available: http://dx.doi.org/10.1023/B:JNMR.0000032552.69386.ea

[19] D. Fischer, A. Elofsson, D. Rice, and D. Eisenberg, "Assessing the performance of fold recognition methods by means of a comprehensive benchmark." *Pac Symp Biocomput*, pp. 300–318, 1996.

[20] W. Bode, R. Engh, D. Musil, U. Thiele, R. Huber, A. Karshikov, J. Brzin, J. Kos, and V. Turk, "The 2.0 AX-ray crystal structure of chicken egg white cystatin and its possible mode of interaction with cysteine proteinases." *EMBO J*, vol. 7, no. 8, pp. 2593–2599, Aug 1988.

[21] J. R. Somoza, F. Jiang, L. Tong, C. H. Kang, J. M. Cho, and S. H. Kim, "Two crystal structures of a potently sweet protein. Natural monellin at 2.75 A resolution and single-chain monellin at 1.7 A resolution." *J Mol Biol*, vol. 234, no. 2, pp. 390–404, Nov 1993. [Online]. Available: http://dx.doi.org/10.1006/jmbi.1993.1594

[22] R. Janowski, M. Abrahamson, A. Grubb, and M. Jaskolski, "Domain swapping in N-truncated human cystatin C." *J Mol Biol*, vol. 341, no. 1, pp. 151–160, Jul 2004. [Online]. Available: http://dx.doi.org/10.1016/j.jmb.2004.06.013

[23] R. Day, D. A. C. Beck, R. S. Armen, and V. Daggett, "A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary." *Protein Sci*, vol. 12, no. 10, pp. 2150–2160, Oct 2003.

[24] E. Y. Jones, S. J. Davis, A. F. Williams, K. Harlos, and D. I. Stuart, "Crystal structure at 2.8 A resolution of a soluble form of the cell adhesion molecule CD2." *Nature*, vol. 360, no. 6401, pp. 232–239, Nov 1992. [Online]. Available: http://dx.doi.org/10.1038/360232a0

[25] A. J. Murray, J. G. Head, J. J. Barker, and R. L. Brady, "Engineering an intertwined form of CD2 for stability and assembly." *Nat Struct Biol*, vol. 5, no. 9, pp. 778–782, Sep 1998. [Online]. Available: http://dx.doi.org/10.1038/1816

[26] P. O'Donoghue and Z. Luthey-Schulten, "On the evolution of structure in aminoacyl-tRNA synthetases." *Microbiol Mol Biol Rev*, vol. 67, no. 4, pp. 550–573, Dec 2003.

[27] T. H. Jukes and C. Cantor, *Mammalian Protein Metabolism*. Academic Press, New York, 1969.

[28] A. G. Murzin, "How far divergent evolution goes in proteins." *Curr Opin Struct Biol*, vol. 8, no. 3, pp. 380–387, Jun 1998.

[29] K. B. Zeldovich, I. N. Berezovsky, and E. I. Shakhnovich, "Physical origins of protein superfamilies." *J Mol Biol*, vol. 357, no. 4, pp. 1335–1343, Apr 2006. [Online]. Available: http://dx.doi.org/10.1016/j.jmb.2006.01.081