Journal of Biomolecular NMR (2018) 70:33–51
https://doi.org/10.1007/s10858-017-0154-1

ARTICLE

Rapid and reliable protein structure determination via chemical shift threading

Noor E. Hafsa[1] · Mark V. Berjanskii[2] · David Arndt[2] · David S. Wishart[1,2]

Received: 15 August 2017 / Accepted: 14 November 2017 / Published online: 1 December 2017

Abstract

Protein structure determination using nuclear magnetic resonance (NMR) spectroscopy can be both time-consuming and labor intensive. Here we demonstrate how chemical shift threading can permit rapid, robust, and accurate protein structure determination using only chemical shift data. Threading is a relatively old bioinformatics technique that uses a combination of sequence information and predicted (or experimentally acquired) low-resolution structural data to generate high-resolution 3D protein structures. The key motivations behind using NMR chemical shifts for protein threading lie in the fact that they are easy to measure, they are available prior to 3D structure determination, and they contain vital structural information. The method we have developed uses not only sequence and chemical shift similarity but also chemical shift-derived secondary structure, shift-derived super-secondary structure, and shift-derived accessible surface area to generate a high quality protein structure regardless of the sequence similarity (or lack thereof) to a known structure already in the PDB. The method (called E-Thrifty) was found to be very fast (often < 10 min/structure) and to significantly outperform other shift-based or threading-based structure determination methods (in terms of top template model accuracy)—with an average TM-score performance of 0.68 (vs. 0.50–0.62 for other methods). Coupled with recent developments in chemical shift refinement, these results suggest that protein structure determination, using only NMR chemical shifts, is becoming increasingly practical and reliable. E-Thrifty is available as a web server at http://ethrifty.ca.

Keywords Nuclear magnetic resonance · Chemical shifts · Threading · Fold recognition · Structure prediction

Introduction

One of the long-term goals in protein NMR is to be able to generate accurate, atomic-resolution protein structures using only chemical shift data. Protein chemical shifts can provide accurate information about secondary structure (Wishart et al. 1992; Wishart and Sykes 1994a, b; Shen et al. 2009; Shen and Bax 2013; Hafsa and Wishart 2014; Hafsa et al. 2015a), torsion angles (Berjanskii et al. 2006; Shen and Bax 2013), hydrogen bonds (Wishart and Nip 1998; Wishart and Case 2001), backbone and side chain dynamics (Berjanskii and Wishart 2005, 2013), disulfide bonds (Sharma and Rajarathnam 2000), charge states (Osapay and Case 1991), accessible surface area (Vranken and Rieping 2009; Berjanskii and Wishart 2013; Hafsa et al. 2015b), ligand interactions (Medek et al. 2000), and aromatic ring proximity (Osapay et al. 1994; Kuszewski et al. 1995). The fact that protein chemical shifts have been shown to provide such a rich diversity of structural information has inspired the development of several chemical shift based protein structure prediction methods, such as CS-Rosetta (Shen et al. 2008), Cheshire (Cavalli et al. 2007), and CS23D (Wishart et al. 2008). The CS-Rosetta and Cheshire methods generally follow an ab initio approach and attempt to model protein structures by generating large numbers of possible structures from the observed chemical shift data and then ranking the models based on knowledge-based potentials and chemical shift scoring functions. CS23D differs from CS-Rosetta and Cheshire in that it also attempts to use comparative modeling

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10858-017-0154-1) contains supplementary material, which is available to authorized users.

✉ David S. Wishart
david.wishart@ualberta.ca

[1] Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8, Canada

[2] Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada

along with chemical shift "threading" to identify known protein folds that may be similar to the fold of the query protein and can be used to guide its structure elucidation.

Sequence "threading" can be used to detect very remote structural homologs or to predict protein fold similarities (Rost 1995, 1997; Karplus et al. 1998; Peng and Xu 2010). An example of a particularly successful sequence-only threading program is PSI-BLAST (Altschul et al. 1997), which uses customized, iteratively trained scoring matrices to identify remote homologs and remotely homologous structures. While PSI-BLAST has proven to be as good as many conventional threading programs (Lindahl and Elofsson 2000), the inclusion of secondary structure information (predicted or calculated) as well as accessible surface area (predicted or calculated) has been shown to improve threading performance (Bowie et al. 1991; Jones et al. 1992; Rost et al. 1997). Indeed, studies by Jones et al. (1992) and Rost et al. (1997) suggested that the environment of an individual residue described by its (sequence-predicted) secondary structure, (sequence-predicted) torsion angles and (sequence-predicted) solvent accessibility are particularly useful. In other words, if sequence-predicted information could be supplemented with easily acquired experimental observations (circular dichroism data, FTIR data, SAXS data, and chemical shift data) it stands to reason that threading accuracy could be significantly improved. As noted earlier, protein chemical shifts have been shown to provide very accurate readouts of protein secondary structure, torsion angles, and accessible surface area. Because protein chemical shifts are often determined long before NOE measurements can be completed, the use of chemical-shift threading could potentially be used to guide or even completely solve protein structures by NMR. This concept is the basis to CS23D. Originally described in 2008 (Wishart et al. 2008), CS23D employs a chemical shift threading program called THRIFTY (THReading with shIFTY) to help generate 3D protein structures from chemical shifts. THRIFTY uses torsion angles predicted via chemical shifts and chemical-shift predicted secondary structures to identify related distant homologous templates or potential structural homologs that already exist in the PDB. THRIFTY has also been used extensively in the GeNMR program (Berjanskii et al. 2009).

The first description of chemical shift threading, as a technique, was made more than 15 years ago (Wishart and Case 2001). Five years later, another chemical shift threading method, called SimShift appeared (Ginzinger and Fischer 2006), which was followed by CS23D/THRIFTY (Wishart et al. 2008). Most recently, Shen and Bax (2015) described a threading-like system called POMONA (Protein alignments Obtained by Matching Of NMR Assignments) that identifies suitable PDB homologs for query proteins using chemical shift data (and NOE distance restraints when available), which is followed by a modified comparative modeling procedure to generate all-atom structures for proteins. In particular, POMONA searches the PDB for suitable homologs that are well matched with backbone chemical shift-predicted, residue-specific φ/ψ probability maps and chemical-shift derived secondary structures. The resulting structural templates are then clustered into groups (typically ten) using a normalized Cα root mean square deviation (Cα-RMSD) as a distance metric. Representative homologs from these clusters are used to build a structural pool for comparative modeling using a modified RosettaCM procedure (Song et al. 2013). POMONA was evaluated on a set of 16 proteins and in most cases the best alignments found by POMONA have good structural similarity with the native structures [an average MaxSub score (Siew et al. 2000) of 0.49] even when there is no detectable sequence similarity (≤ 20% sequence identity).

Published results from SimShift (Ginzinger and Fischer 2006), THRIFTY/CS23D (Wishart et al. 2008), and POMONA (Shen and Bax 2015)—all strongly suggest that the structural information encoded by chemical shifts can help to identify structurally similar template(s) even in the absence of detectable sequence similarity. Inspired by these studies, we have developed a method called "Enhanced-Thrifty" (E-Thrifty) that employs a more advanced version of chemical shift threading to more accurately identify the most likely fold and to generate a high quality protein structure. In particular, E-Thrifty uses significantly enhanced shift-based secondary structure identification (Hafsa and Wishart 2014) as well as recently developed shift-based super-secondary and structural motif identification (Hafsa et al. 2015a) to improve its performance. It also uses a newly developed shift-based accessible surface area prediction method (Hafsa et al. 2015b) as well as shift-based torsion angle predictions (Shen and Bax 2013) along with very accurate secondary chemical shift calculations (Han et al. 2011). These are combined to perform a modified threading protocol using a specially constructed, non-redundant database of known protein structures [a modified version of the PDB (Berman et al. 2000)]. When compared to the state-of-the-art threading programs or chemical shift-based structure generation programs, such as POMONA (Shen and Bax 2015), PSI-BLAST (Altschul et al. 1997), CS-Rosetta (Shen et al. 2008), and CS23D (Wishart et al. 2008), on two different test data sets, E-Thrifty exhibits a 10–20% improvement in overall performance and a significant improvement in speed (< 10 min/query). E-Thrifty's performance on "easy" targets is even better, suggesting that it can serve as a rapid and highly reliable method for determining protein structures from chemical shifts alone. Details describing the E-Thrifty algorithm, its performance, and its implementation as a web server are given below.

Materials and methods

Structural annotation of the database proteins

Key to any successful threading algorithm is a non-redundant, high quality, well-annotated database of protein sequences and structures. A non-redundant (nr) version of the PDB was generated using the Pisces server (Wang and Dunbrack 2003). As of July 21st 2017, there were a total of 91,336 sequences and coordinate files in this nrPDB data set. This database of known structures was then annotated using a series of programs so that every residue was assigned a secondary structure, a specific secondary structure motif, a set of backbone torsion angles, and a fractional accessible surface area. The secondary structures, torsion angles, and accessible surface areas were generated from the DSSP (Kabsch and Sander 1983) program. Other secondary structure elements, such as β-turns and edge/internal strand information, were obtained using methods described in Hafsa et al. (2015a). Fractional accessible surface areas or fASA values for each residue were derived from the DSSP output using a method previously described in Hafsa et al. (2015b). After calculating these data, we generated four "pseudo-sequences" based on a structural alphabet associated with each entry in our nrPDB data set. These pseudo-sequences correspond to: (1) a secondary structure "sequence" (using 3 letters—H, B, C for helix, beta strand, and coil, respectively); (2) a structure motif "sequence" (using 5 letters—H, E, I, C, T for helix, edge beta strand, interior beta strand, coil, and beta turn, respectively); (3) a torsion angle "sequence" (using 9 letters—ASDFGHJKL corresponding to different regions of Ramachandran space—see Supplementary Material for more details) and (4) a fASA "sequence" (using 3 letters—B, P, E, for buried, partially buried, and exposed, respectively). Additional details regarding the meaning and numerical cutoffs associated with each of these structural alphabets are provided in Hafsa et al. (2015b) and Wishart et al. (2008). These, along with the amino acid sequence of each protein, describe its local and non-local structural states.

Sequence and secondary structure filtering

A key consideration in chemical shift based structure generation is speed. Many algorithms take hours or require large numbers of CPUs to complete their calculations. We decided to incorporate some "smart" filtering steps to accelerate the performance of our program and allow it to generate a final folded model on a single CPU in a matter of minutes (though the generation of ensembles for unfolded structures, described below, employs 20 CPUs). The initial input to the E-Thrifty program is expected to be an amino acid sequence (≥ 12 amino acids) and a reasonably complete (> 85%) set of backbone (including $^{13}C_{\alpha}$, $^{13}C_{\beta}$, ^{13}C, $^{1}H_{N}$, $^{1}H_{\alpha}$ and ^{15}N) chemical shifts. The query is then run through CSI 2.0 (Hafsa and Wishart 2014) to generate a shift-derived secondary structure assignment. The sequence and the shift-derived secondary structures of the query protein are then aligned against all the database sequences and structures in our non-redundant version of the PDB (nrPDB) using BLAST (Altschul et al. 1997; Boratyn et al. 2012). An E-value cutoff of 0.01 is used for selecting sequence alignments and 0.001 is used for selecting secondary structure alignments. While the sequence pre-filtering step identifies obvious homologues, the secondary structure content pre-filtering step helps to retrieve structurally similar proteins and reject structurally dissimilar proteins. For example, if the query protein is predicted (via chemical shifts) to be an all-helical protein, this simple pre-filter will remove all-beta or all-disordered proteins from consideration. Including these pre-filters limits the search space for next stage structural alignment (see below). If there are no significant hits from these preliminary sequence and secondary structure filtering steps, a full-scale structural alignment is performed against nrPDB to find suitable templates for the query protein. The details of the full-scale structural alignment method are described below.

Measuring local and non-local structure similarity

While amino acid substitution scores are normally used to guide the local alignment between two protein sequences, sequence alignment alone does not necessarily guarantee optimal structural or topological alignment between two proteins. This is particularly true when the sequence identity between two proteins drops below 35%. To perform sequence alignments or sequence threading for distantly related proteins, additional information, such as (predicted or calculated) backbone φ/ψ angles, secondary structure, structural motifs, secondary chemical shifts and accessible surface area (ASA), are often needed to guide the alignment process. This is because these structural states tend to be more conserved than sequence among remote structural homologues (Rost et al. 1997). Ideally, a good threading program should have three major components: (1) a database of solved structures where all of the threading parameters (sequence, torsion angles, secondary structure, structure motifs, ASA and secondary shifts) are pre-calculated; (2) a series of programs where the same parameters (torsion angles, ASA, etc.) are predicted and/or calculated for the query protein and (3) an alignment algorithm that scores, aligns, and matches the query protein by taking into account all of the calculated and/or predicted parameters. If the structural parameters used for threading can be converted to letters or character strings (similar to the sequence), the threading process can be performed via a letter alignment algorithm, such as the Smith–Waterman alignment algorithm (Rost et al. 1997). These are the principles that we used to design the E-Thrifty algorithm, which is depicted in Fig. 1. The E-Thrifty algorithm, the scoring scheme, and the parameter mapping are explained in more detail below.

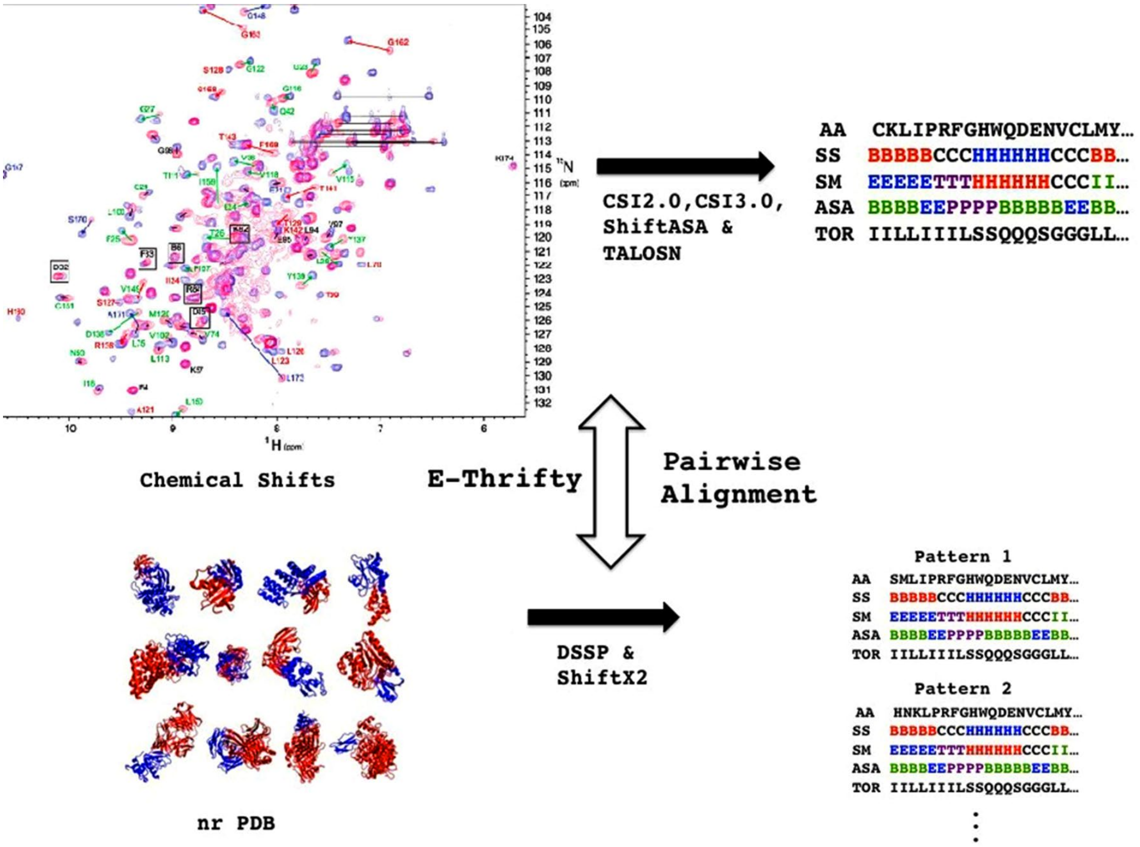
Substitution matrices for structural descriptors

In E-Thrifty, structural parameters that describe the structural environments associated with each residue in the query and with each residue in the nrPDB are represented by a set of one-letter codes (see Structural Annotation of Database Proteins for details). During the alignment process, these letters are compared to maximally match both local and non-local structural similarity. Matched or highly similar structural states are given a high positive score whereas unmatched or dissimilar states are given a low score (a smaller positive or a negative value). For example, a negative score is assigned to a helical secondary structure class (represented by "H") being substituted/replaced by a β-strand class (represented by "E" or "I"), whereas a β-turn replacement by a coil assignment is given a small positive value (i.e., a lower penalty). A substitution matrix can be used to compactly represent this scoring scheme. A 3×3 substitution matrix for the three-state secondary structure states describes the substitution/matching scores of the three secondary structure classes. Similarly, a 5×5 substitution matrix is used for the five structural motif states, a 3×3 substitution matrix is used for the three fASA categorical states, and a 9×9 substitution matrix is used for the nine torsion angle states. Substitution matrix values were initially chosen from the BLOSUM62 matrix (Henikoff and Henikoff 1992) and then optimized through grid search methods on the training alignments.

Scoring local and non-local structural similarity

After defining the substitution matrices for the different structural descriptors, the structural similarity between the query residue *i* and the database residue *j* is calculated using the following equation:

$$\begin{aligned} S(i,j) = {} & w_{AA} \times AA_{score} + w_{Torsion} \times Torsion_{score} \\ & + w_{SS} \times SS_{score} + w_{SM} \times SM_{score} \\ & + w_{ASA} \times fASA_{score} \end{aligned} \tag{1}$$



104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132
^{15}N (ppm)
10 9 8 7 6
^{1}H (ppm)
Chemical Shifts
CSI2.0,CSI3.0,
ShiftASA &
TALOSN
AA CKLIPRFGHWQDENVCLMY...
SS BBBBBCCCHHHHHHCCCBB...
SM EEEEETTTHHHHHHCCCII...
ASA BBBBEEPPPPBBBBBEEBB...
TOR IILLIIIILSSQQQSGGGLL...
E-Thrifty
Pairwise
Alignment
nr PDB
DSSP &
ShiftX2
Pattern 1
AA SMLIPRFGHWQDENVCLMY...
SS BBBBBCCCHHHHHHCCCBB...
SM EEEEETTTHHHHHHCCCII...
ASA BBBBEEPPPPBBBBBEEBB...
TOR IILLIIIILSSQQQSGGGLL...
Pattern 2
AA HNKLPRFGHWQDENVCLMY...
SS BBBBBCCCHHHHHHCCCBB...
SM EEEEETTTHHHHHHCCCII...
ASA BBBBEEPPPPBBBBBEEBB...
TOR IILLIIIILSSQQQSGGGLL...
⋮

Fig. 1 The sequence-structure alignment concept used in the E-Thrifty method. Here AA, SS, SM, ASA and TOR represent the amino acid, secondary structure, structure motif, accessible surface area and backbone torsion angle sequences, respectively. *nrPDB* non-redundant Protein Data Bank

where AA_{score} is the amino acid similarity score, $Torsion_{score}$ is the torsion letter similarity score, SS_{score} is the secondary structure similarity score, SM_{score} is the structural motif similarity score, $fASA_{score}$ is the ASA state similarity score, and the w^* terms represent the corresponding weighting coefficients. Each $S(i,j)$entry in the calculated scoring matrix is then rescaled to a range (− 2.0, 3.0) so as to obtain a smoothed distribution of alignment scores. The rescaling is performed using the following equation:

$$S^{scaled}_{i,j} = \left(\left(\max_{range} - \min_{range} \right) * \frac{S[i][j] - \min_S}{\max_S - \min_S} \right) + \min_{range} \tag{2}$$

where $\min_{range} = -2.0$, $\max_{range} = 3.0$, and $\min_S$and $\max_S$ are the minimum and maximum values observed in the scoring matrix S, respectively. The scoring components described in Eq. 1 are briefly explained in the following paragraphs.

Amino acid similarity

The aligned amino acids are scored using the BLOSUM62 (Henikoff and Henikoff 1992) substitution matrix. In this 20 × 20 matrix, every possible amino acid substitution is assigned a score based on its observed frequencies derived from careful alignment of evolutionarily related proteins (with no more than 62% sequence identity). A positive score is given to more probable substitutions while a negative score is given to less probable substitutions. The amino acid similarity score (AA_{score}) is given as:

$$AA_{score} = \; BLOSUM62(aa_i,\; aa_j) \tag{3}$$

where aa_i is the query residue in i-th position and aa_j is the database residue in j-th position.

Secondary structure similarity

Secondary structures for the query protein are calculated using the CSI 2.0 program (Hafsa and Wishart 2014). CSI 2.0 is a multi-class, machine-learning algorithm that determines the extent and location of α-helices, β-strands, and coil regions based on $^{13}C_\alpha$, $^{13}C_\beta$, ^{13}C, $^{1}H_N$, $^{1}H_\alpha$, and ^{15}N backbone chemical shifts and sequence. In the E-Thrifty threading algorithm, the secondary structure similarity between secondary structures of the query residue i and the database residue j is calculated over a 3-residue window using the following formula:

$$SS_{score} = \sum_{n}^{-1,0,1} S_{i+n,\, j+n} \{ S_{i,j} = SS_{matrix}(ss_i \;,\; ss_j) \tag{4}$$

where SS_{matrix} is a 3 × 3 substitution matrix for secondary structure states {H, B, and C}, which describes the substitution scores for matching/replacing one secondary structure state with another.

Structural motif similarity

The classification of β-strands and β-turns in the query sequence is performed by CSI 3.0, a chemical shift based super-secondary structure identification program, described by Hafsa et al. (2015a). The CSI 3.0 output is mapped to 5 letters H, E, I, C, and T which stand for helix, edge β-strand, interior β-strand, coil and β-turn, respectively. In the E-Thrifty threading algorithm, the structural motif similarity between the query residue i and the database residue j is calculated over a 3-residue window using the formula:

$$SM_{score} = \sum_{n}^{-1,0,1} M_{i+n,\, j+n} \{ M_{i,j} = SM_{matrix}(sm_i \;,\; sm_j) \tag{5}$$

where SM_{matrix} is a 5 × 5 substitution matrix for five structure motif states {H, E, I, C, T}, describing the scoring scheme for matching/replacing structural motif states.

Fractional accessible surface area (fASA) similarity

The fractional ASA (fASA) is an ASA descriptor that describes the percentage of accessible surface area for a given residue relative to a fully exposed residue. Residue-specific fASAs for the query protein are calculated using the ShiftASA program (Hafsa et al. 2015b). Residues are assigned one of three letters—B (buried) ($fASA \le 0.25$), P (partially buried) ($0.50 \ge fASA > 0.25$), or E (exposed) ($fASA > 0.50$)—based on the predicted/calculated fASA range. In the E-Thrifty threading algorithm, the similarity between fASA categorical states of the query residue i and the database residue j is calculated as below:

$$fASA_{score} = \sum_{n}^{-1,0,1} F_{i+n,\, j+n} \{ F_{i,j} = fASA_{matrix}\left(f_i \;,\; f_j\right) \tag{6}$$

where $fASA_{matrix}$ is a 3 × 3 substitution matrix for three fASA categorical states {B, P, E} that describes the scores for matching/substituting three fASA states.

Torsion letter similarity

Backbone φ/ψ torsion angles from experimental $^{13}C_\alpha$, $^{13}C_\beta$, ^{13}C, $^{1}H_N$, $^{1}H_\alpha$, ^{15}N chemical shifts are predicted by TALOS-N (Shen and Bax 2013) and converted into a 9-letter torsion angle alphabet. This so-called torsion angle alphabet, which is identical to the THRIFTY alphabet used in CS23D (Wishart et al. 2008), splits the Ramachandran map into 9 non-overlapped regions based on the φ/ψ propensity of common secondary structural classes, with a letter assigned to each region (Fig. S1). In the E-Thrifty threading algorithm, the torsion letter similarity between the query residue i and the database residue j is calculated over a 3-residue window using the following formula:

$$Torsion_{score} = \sum_{n}^{-1,0,1} T_{i+n,\ j+n}\left\{T_{i,j} = Torsion_{matrix}(tor_i\ ,\ tor_j)\right. \tag{7}$$

where $Torsion_{matrix}$ is a 9×9 matrix that describes the substitution scores for the replacement of one torsion angle letter with another. Therefore, the torsion letter similarity score of the central residue position takes into account the torsion letter substitution scores in the preceding and the following neighbor locations.

Gap penalty function in sequence-structure alignment

In order to perform a proper sequence or a sequence-structure alignment of two protein sequences, it is important to develop a scoring function to properly handle the insertion or deletion of gaps in either sequence. Gaps are usually counted as a penalty in the total alignment score. Typically an affine gap penalty or AGP function of the form, $g = u + vl$ is used in most sequence-only alignment algorithms. This kind of function depends on the gap initiation (u) and gap extension (v) parameters, and the length of the gap in the alignment (l). However, previous studies suggest that including a conformation specific gap penalty in sequence-structure alignment increases the accuracy of the correctly aligned residues (Madhusudhan et al. 2006). Hence, in our work, we adopted a conformation specific gap penalty function called a variable gap penalty or VGP as described in a recent publication by Shen and Bax (2015), in which the gaps that are introduced in regular secondary structure regions (contiguous helices and β-strands) and between two spatially distant residues are penalized.

Protein local alignment

In E-Thrifty, protein sequences with experimentally measured chemical shifts (i.e. the query sequence) are aligned against sequences of known structures in our annotated nrPDB using a modified version of the Smith–Waterman local alignment algorithm (Smith and Waterman 1981). A similarity matrix S of M×N dimensions is constructed, where M is the length of the query protein and N is the length of the database protein. Each element in the scoring matrix, $S(i,j)$ indicates the substitution score for the query residue i with the database residue j. Once the scoring matrix is constructed, the optimal alignment between query and subject sequence is found by calculating an alignment matrix (H) using a dynamic programming and traceback procedure. This traceback protocol involves finding the maximum element in the alignment matrix and tracing back through the matrix from the maximum element to zero. Each element $H(i,j)$ in the H matrix is calculated with the following recursive dynamic programming equation:

$$H(i,j) = \max\begin{cases} H(i-1,\ j-1) + S(i,j) \\ H(i-1,\ j) - VGP(i,j,\ i-1,j) \\ H(i,j-1) - VGP(i,j,i,j-1) \\ 0 \end{cases} \tag{8}$$

The initial conditions for the recursive algorithm are,

$$\begin{cases} H(i,0) = 0;\ 0 \le i \le M \\ H(0,j) = 0;\ 0 \le j \le N \end{cases}$$

In Eq. 8, $S(i,j)$ is the substitution score for the query residue i with the database residue j. VGP is the gap penalty function applied when there is a gap opening or extension between the i and i' positions in the sequence block or the j and j' positions in the structure block. For the assignment of each element $H(i,\ j)$ in the H matrix, the diagonal i.e. upper-left ($H(i-1, j-1)$), upper ($H(i-1, j)$), and left ($H(i,\ j-1)$) neighbor elements are compared and the maximum value among these three elements is assigned to the current element as score $H(i,\ j)$. If the maximum value is negative, then 0 is assigned to the score. After calculating all the elements of the H matrix as described in Eq. 8, the largest element in the H matrix (H_{max}) represents the optimal alignment score. The residue equivalence assignments can then be obtained by tracing back through the maximum element, H_{max} to the zero value in the H matrix, which is also the optimal sub-alignment between query and the subject sequences. An example of local alignment between the query protein 2LCI and the database protein 2L82 (chain A) is shown in Figure S2.

Chemical Shift scoring and the alignment ranking

To further improve the alignment scoring, we implemented a backbone secondary chemical shift fitness score similar to SimShift (Ginzinger and Fischer 2006). Specifically, a secondary chemical shift fitness score is calculated for equivalent residue assignments in the alignment. Secondary chemical shifts can be defined as the difference between the observed experimental chemical shift (δ_{obs}) and the corresponding random coil shift (δ_{rc}) value for a specific atom (Wishart 2011).

$$\Delta\delta = \delta_{obs} - \delta_{rc} \tag{9}$$

Secondary chemical shifts contain important structural and dynamic information about proteins (Wishart and Case 2001; Mielke and Krishnan 2009). The backbone chemical shifts for the nrPDB structures are calculated using ShiftX2 (Han et al. 2011) and the secondary shift values are obtained with the aforementioned formula using the neighbor adjusted random coil values extracted from Schwarzinger et al. 2001. The secondary shift fitness score is calculated as:

$$SC_{score} = \sum_{a} w_a \times corr\left(\delta_a^{obs},\ \delta_a^{pred}\right) \tag{10}$$

where δ_a^{pred} represents the backbone chemical shift predicted by ShiftX2 (Han et al. 2011) for a specific atom a ($a = {}^{13}C_\alpha$, ${}^{13}C_\beta$, ${}^{13}C$, ${}^{1}H_N$, ${}^{1}H_\alpha$, ${}^{15}N$). The corresponding experimental chemical shift is referred to as δ_a^{obs}. The w_a's are the weighting coefficients for the standard six backbone atoms. The function $corr\left(\delta_a^{obs},\ \delta_a^{pred}\right)$ measures the correlation between the observed (query) and the predicted (subject) secondary chemical shifts over all the aligned residues for a specific backbone atom a. Therefore chemical shift fitness score is a weighted combination of chemical shift correlations of six backbone atoms over all the aligned residues.

The secondary chemical shift fitness score is then combined with the optimal sub-alignment score H_{max} using a scaling factor w_{corr} to produce the total score for each alignment.

$$S_{total} = H_{max} + w_{corr} \times SC_{score} \tag{11}$$

The final ranking of the alignments is performed according to this total score, S_{total}.

Optimization of E-Thrifty parameters

To optimize the parameters described in Eqs. 1–11, a set of 30 proteins with complete experimental chemical shift information and corresponding high-resolution X-ray structures (< 2 Å) were chosen. The training proteins had ~ 90% of their complete (${}^{1}H$, ${}^{13}C$, and ${}^{15}N$) backbone chemical shifts assigned. A set of homologs for the training proteins spanning a sequence identity range 20–40% was retrieved using a PSI-BLAST search. Once the training set was obtained, a structural alignment between the queries and the corresponding homolog proteins was performed. The alignment produced a total of 1777 alignment pairs. There were a total of 18 parameters to optimize. The parameter set could be divided into three different groups: the scoring matrix weighting parameters, the gap spanning parameters, and the chemical shift weighting parameters. Each group was optimized using a specific optimization protocol as described below.

The scoring matrix parameters (w_{AA}, $w_{Torsion}$, w_{SM}, w_{ASA}) were optimized by a grid search using the training structural alignments. We trained one parameter at a time and kept the other parameters constant at their initial values or the previously optimized values. Parameter optimization was terminated on the convergence of the average alignment score observed against the Cα-RMSD between the aligned residues (the higher the alignment score, the lower the Cα-RMSD) for the training set of structural alignments. Note that the RMSD is calculated for the defined secondary structures in the aligned region using the Superpose program (Maiti et al. 2004). For the gap spanning parameters in the VGP function described in Shen and Bax (2015), the initial values were chosen from the original study. We then attempted to further optimize the parameter values through a grid search. However, no significant improvement was observed (data not shown). Hence, we used the Shen et al. values. The chemical shift weighting parameters defined in Eq. 10 were optimized using a linear regression analysis. The training data for linear regression is comprised of chemical shift correlation coefficients between the six backbone atoms of the query and the database equivalent residues (~ X) and the Cα-RMSD (~ Y) of the aligned region of the training proteins. A linear regression model was then fit to the training data. A similar regression analysis was performed to search an optimal value for w_{corr} described in Eq. 11.

The optimized parameter values (except the gap spanning parameters) determined from this study are:

$$w_{ASA} = 4.25, w_{AA} = 1.0, w_{SM} = 5.11, w_{Torsion} = 3.97,$$
$$w_{SS} = 4.35, w_d = 2.0, w_{CO} = 3.75, w_{CA} = 4.5,$$
$$w_{CB} = 4.75, w_N = 2.5, w_{HN} = 4.25, w_{HA} = 4.5,$$
$$w_{corr} = 0.5.$$

Statistical significance of E-Thrifty alignments

In any database alignment protocol, it is important to properly assess the significance of an alignment that results from a comparison of a protein of a certain length to a database containing many different proteins of variable length. Hence, E-Thrifty alignments were evaluated using a BLAST-like e-value or expect-value (Altschul and Gish 1996). This was done because E-Thrifty's alignment scoring system (using substitution matrices) and alignment protocols closely resembled the original BLAST model. Furthermore, the secondary structure, super-secondary structure, torsion angle and accessible surface area information used by E-Thrifty were all encoded as simple *n*-letter sequences, just like the *20*-letter amino acid sequences used in BLAST. Because E-Thrifty essentially performs a BLAST-like alignment, we decided to use the same definitions and the same BLAST parameters in calculating our e-values. Therefore, E-Thrifty's e-value for an alignment having a score *S* was calculated using the function described in Altschul et al. (1996).

$$E = K \times m' \times n' \times e^{-\lambda S} \tag{12}$$

here K and λ are statistical parameters, $m'n'$ is the effective search space size and S is the alignment score of an optimal sub-alignment. The effective search space $m'n'$ is calculated using the equation below:

$$\begin{aligned} m' &= m - l \\ n' &= n - N \times l \end{aligned} \tag{13}$$

where m is the number of residues in query protein, n is the total number of residues in the E-Thrifty protein database, N is the total number of E-Thrifty database proteins and l is the edge correction factor. The edge correction factor is used to calculate an "effective length" for a sequence. It eliminates the "edge effect" problem i.e. a high-scoring alignment has a non-zero length and does not begin near to the end of either of two sequences being compared. K and λ values are taken from Altschul et al. (1996). The l value depends on the length of the database protein being compared with and is chosen from a set of empirical values depending on the ln($m \times n$) values described in the same study by Altschul et al. (1996).

Clustering and selection of E-Thrifty templates

Among the candidate templates identified by E-Thrifty, many were found to score very similarly or very closely to one another for any given query protein. This is because multiple templates may share a similar fold, a similar sub-structure, or a similar 3D structure. In some cases, multiple templates for a single query protein may increase the coverage when building a full-length model. Moreover, these templates can be used as a structural pool for comparative modeling purposes. In E-Thrifty, a hierarchical clustering algorithm was used to group the set of candidate templates. The Cα-RMSD between two template proteins was calculated over a common set of residues that were aligned with the same set of query residues. Specifically, the "complete linkage method" for hierarchical clustering was used to group the templates. At each stage, the cluster is formed when all the links (i.e. the Cα-RMSD) between pairs of objects in the cluster are within ≤ 4 Å cut-off distance. .

E-Thrifty will generate up to 10 template hits, each of which are ranked according to their alignment scores and each of which are given a cluster membership. A user can select either the top template from this list or the top representative template from the first three resultant clusters for the subsequent model generation part.

Generation of 3D models via MODELLER

As part of the E-Thrifty pipeline, a 3D model of the query protein is generated via the MODELLER software package (Sali and Blundell 1993). It is important to remember that simply identifying a similarly folded template for a query sequence is insufficient to determine the full 3D structure of the query. To complete the 3D structure determination process, comparative modeling must be used to rebuild the template, add the appropriate side chains, construct loops, close gaps, minimize the energy and create a 3D structure with exact sequence of the query protein. MODELLER is a widely used comparative/homology modeling program that is able to perform all these operations. To initiate the comparative modeling process, the sequence-structure alignment generated by E-Thrifty is first converted into the required PIR format and then used as input for MODELLER's comparative modeling function. MODELLER then generates the 3D coordinates of five possible models. The generated models are further assessed using MODELLER's score evaluation functions (GA341 and DOPE). The 3D model that has the lowest energy after the assessment is chosen as the final 3D model. MODELLER was chosen for the model generation purposes as it has been well tested by the scientific community and become one of the most popular programs for template-based modeling. E-Thrifty has two model generation output options. The default "single-template" option generates a comparative model (with 3D coordinate data) of the query protein using the E-Thrifty sequence-structure alignment of the top template only. The other "multiple-template" option offered by E-Thrifty employs the Clustal Omega (Sievers et al. 2011) program to perform a multiple alignment between the query and several template sequences. This multiple alignment is then used to build the 3D model of the query protein. Note that, in this paper, we have used the single-template modeling option to generate all the models that are used in the comparative study of E-Thrifty and the other threading programs. We will describe the evaluation of the other experimental (multiple-template) modeling option in an upcoming publication.

Assessment of E-Thrifty generated models

To assess the performance of E-Thrifty as well as other threading programs, we used DALI (Holm and Rosenström 2010) as the "gold standard" for identifying remote structural homologs. DALI is a web server designed for performing 3D coordinate comparisons. It is particularly useful for identifying proteins with 3D structure similarities that may not have any obvious sequence similarity. For this component of the study, we assessed the structural accuracy and fold similarity achieved by structures generated by E-Thrifty, POMONA, PSI-BLAST, and DALI (as a control) using the Template Modeling or TM-score (Zhang and Skolnick 2004) in addition to RMSD. As many other researchers (Siew et al. 2000; Ortiz et al. 2002; Betancourt and Skolnick 2001) have noted, RMSD is not a perfect measure of model quality as it sometimes fails to identify well-predicted sub-structures in the presence of large prediction errors (i.e. disordered loops) in other parts of the model. Therefore, we used the TM-score in addition to RMSD to measure model quality. Unlike other popular scoring functions such as the aforementioned MaxSub score (Siew et al. 2000), the TM-score uses a size-dependent scale to eliminate the protein length dependence. It also considers all alignments or modeling residue pairs in its assessment rather than arbitrarily setting specific distance cutoffs and calculating only the fraction of residues with errors below a certain cutoff distance (Zhang and Skolnick 2004). A TM-score typically falls in the range from 0 to 1.0, with a TM-score of 1.0 indicating a perfect match between two structures and a TM-score below 0.17 generally indicating a randomly chosen unrelated fold. A quantitative study by Xu and Zhang (2010) showed that proteins with a TM-score equal to 0.50 have a probability of 37% of being in the same CATH (Greene et al. 2007) topology family and a probability of 13% of being in the same SCOP (Murzin et al. 1995) fold family.

Results

The performance of E-Thrifty was evaluated on three independent test sets. The first test set (TEST1) consisted of 15 proteins randomly selected from a set of proteins, for which CS-ROSETTA models were available on the BMRB website (Ulrich et al. 2008). The second data set (TEST2) consisted of ten blind targets from a recent challenge Critical Assessment of Automated Structure Determination of Proteins by NMR (CASD-NMR-2013) (Rosato et al. 2015). The third test (TEST3) was a random selection of 15 recently solved, fully assigned proteins with sequence identities between 35 and 90% to proteins already in the PDB. For all 3 tests, DALI (Holm and Rosenström 2010) was used to identify the "gold standard" models, corresponding to the best structural homologs that are able to be identified through a 3D coordinate-based structural superposition against all PDB structures. The performance of E-Thrifty was compared against several well-regarded threading and ab initio structure generation and chemical shift threading methods, including POMONA (Shen and Bax 2015), PSI-BLAST (Altschul et al. 1997), CS23D (Wishart et al. 2008) and CS-Rosetta (Shen et al. 2008). More specifically, POMONA is a threading/homology search program that uses only chemical shift generated structural information to obtain highly probable alignments for query proteins. On the other hand, PSI-BLAST detects homologs using position-specific scoring matrices or customized scoring profiles generated using iterative alignments of the query sequence against the database. CS23D detects structural homologs via chemical shift threading (torsion angle matching and secondary structure matching) as well as sequence comparison. CS-Rosetta predicts the protein structures via chemical shift-based fragment matching, in conjunction with ab-initio protein modeling algorithms.

The performance of E-Thrifty and the other methods was evaluated for the TEST1 proteins with two sequence identity upper thresholds (≤ 30 and ≤ 95%), whereas, for the TEST2 proteins, only a ≤ 30% sequence identity cutoff was used (as most TEST2 proteins exhibit very low sequence identity to known structures). For evaluation consistency, the same protein structure modeling software, MODELLER (Sali and Blundell 1993) was used to build full-length models using the templates identified by the different threading methods assessed in this study. The quality of the template models was then evaluated using the TM-score (Zhang and Skolnick 2004), a widely used metric to assess the folding similarity between two proteins.

The results of the TM-score evaluation on TEST1 and TEST2 proteins are shown in Tables 1, 2 and 3. These tables also describe the performance of PSI-BLAST, CS23D, CS-Rosetta, DALI, and POMONA for the same test proteins. The POMONA and the PSI-BLAST templates with the highest alignment scores with default alignment settings (within two sequence identity cutoffs) were identified as the optimal threading results. The lowest energy structure produced by CS-Rosetta was considered as the best template. As the TEST2 proteins are blind targets from the CASD-NMR-2013 competition (Rosato et al. 2015), we decided that in order to make the comparison fair and unbiased, all database templates that were solved or deposited into the PDB after 2013 were excluded. The last column in each table includes the results from the DALI server, which essentially indicates the "true" answer or the "gold standard". This result was included to help assess each program's performance. The inclusion of the DALI data also helps to define the upper limit of how well a given threading program can perform. The database, proteins for which DALI produces the highest Z-scores (within the two sequence identity cutoffs, ≤ 30 and ≤ 95%), were selected as the DALI outputs. Note that, in selecting the templates from the different programs, we relied on the reported sequence identity as measured by the respective alignments. If one of the programs had no answer for any particular test case, we report N/A (not available) in that particular column.

Discussion

Pre-filtering steps improve the E-Thrifty search time

The sequence and secondary structure content pre-filtering steps in the E-Thrifty pipeline were found to significantly improve the overall E-Thrifty template search time.

Table 1 Template recognition performances of four threading programs on TEST1 proteins using sequence identity cutoff as ≤30%

Query			E-Thrifty	POMONA	PSI-BLAST	DALI
Protein name	PDB/BMRB ID	Length/(fold)	Template TM-score (RMSD)	Template TM-score (RMSD)	Template TM-score (RMSD)	Template TM-score (RMSD)
KaiA	1M2F/5031	135/(α/β)	1YS7A 0.74 (3.18)	2WRZA 0.51 (3.51)	4D2OB 0.32 (4.79)	3R0JA 0.71 (2.99)
NEDTH	1F3Y/4448	165/(α/β)	2O5WA 0.66 (3.15)	3A6UA 0.61 (3.99)	2O5WD 0.46 (3.97)	3I7UC 0.66 (2.41)
NCS-1	2LCP/4378	190/(α)	1UHNA 0.60 (4.18)	2TN4A 0.32 (3.93)	2GGZB 0.59 (4.09)	4F0ZB 0.53 (3.95)
Sortage	1IJA/4879	148/(β)	2LN7A 0.69 (3.23)	3FN5A 0.74 (2.84)	3RCCR 0.61 (3.23)	4TQXA 0.73 (2.69)
PyJ	1FAF/4403	79/(α)	1GH6A 0.67 (2.60)	3QPPA 0.37 (3.43)	1GH6A 0.56 (2.48)	2PF4E 0.70 (2.12)
ERp18	2K8V/15964	157/(α/β)	2DMMA 0.43 (3.94)	3VWWB 0.53 (3.71)	1XWCA 0.44 (3.80)	3IRAA 0.59 (3.21)
ApolPBP1A	2JPO/15256	142/(α)	2WC5A 0.64 (2.98)	3TNWD 0.30 (4.19)	2WCMA 0.62 (3.09)	2WC5A 0.65 (3.03)
Pru Av 1	1E09/4671	159/(α/β)	3KDIA 0.76 (2.72)	3US7A 0.83 (2.44)	4PSBA 0.83 (2.53)	3US7A 0.83 (2.44)
LC1	1M9L/4265	199/(α/β)	1QYYA 0.56 (4.19)	2WSMA 0.31 (5.80)	5IL7B 0.46 (3.62)	5IL7A 0.62 (4.23)
cg2496	2KPT/16569	148/(α/β)	5ANPB 0.70 (2.62)	3PVHA 0.67 (2.56)	3DFLA 0.30 (5.40)	3PVHA 0.69 (2.36)
NCAM	1EPFA/4162	191/(β)	4UOWW 0.80 (1.83)	3QP3C 0.80 (1.75)	3B43A 0.70 (2.33)	2IEPA 0.76 (2.15)
PG	2HZE/4113	108/(α/β)	3L4NA 0.78 (2.29)	4I2UA 0.78 (1.96)	3C1SA 0.75 (2.52)	4N10A 0.82 (2.27)
AT5g22580	1RJJ/6011	111/(α/β)	5B0DB 0.66 (2.92)	1TR0C 0.68 (2.38)	5B0FB 0.66 (2.71)	2QYCB 0.68 (3.01)
N-WASP	1MKE/5554	144/(α/β)	1QC6A 0.62 (2.75)	2XQNM 0.64 (2.39)	3SYXA 0.37 (3.49)	3OANA 0.63 (2.31)
Grx2	1G7O/4318	215/(α)	3TOTA 0.66 (3.93)	2WRTG 0.64 (3.85)	2GDRF 0.21 (5.49)	4LMWA 0.79 (2.97)
Average			0.67 (3.10)	0.58 (3.25)	0.52 (3.56)	0.69 (2.81)

The E-Thrifty column shows the top template identified by E-Thrifty, whereas the next two columns show the top templates identified by POMONA and PSI-BLAST. The DALI answers for TEST1 proteins are displayed in the last column. Template information includes the PDB ID of the template, TM-score and the TM-Align RMSD of the full-length model generated using the corresponding template. RMSD values are shown in Angstroms (Å)

E-Thrifty's run time without the pre-filtering steps is ~60–90 min for medium sized proteins, whereas the inclusion of simple pre-filters reduces the run time to <10 min (>80% improvement over the run time without pre-filtering). Extensive testing (data not shown) demonstrated that adding these pre-filtering steps allowed E-Thrifty to retain its original accuracy in terms of finding the correct templates.

E-Thrifty's performance on TEST1, TEST2 and TEST3

E-Thrifty's performance on the TEST1 and TEST2 data sets is described in detail in Tables 1, 2 and 3. In terms of TM-score evaluation of the full-length template (or database) models, E-Thrifty consistently gave a better performance than all other prediction programs for both test sets. In particular, E-Thrifty achieved an average TM-score of 0.67 (TM-Align RMSD = 3.10 Å) for proteins with ≤30% sequence identity in the TEST1 data set, as opposed to 0.58 (TM-Align RMSD = 3.25 Å) achieved by POMONA and 0.52 (TM-Align RMSD = 3.56 Å) achieved by PSI-BLAST respectively (Table 1). The improvement in E-Thrifty's TM-scores for the 15 test proteins is statistically significant (with a 95% confidence level), compared to PSI-BLAST and POMONA (*p* values of 0.003 and 0.03, respectively, which are less than the cutoff significance level of 0.05). In Table 2, where no sequence similarity restriction was imposed, we can see that all three programs exhibit nearly similar performance in terms of their average TM-score (E-Thrifty = 0.74 (TM-Align RMSD = 2.60 Å), POMONA = 0.70 (TM-Align RMSD = 2.74 Å), PSI-BLAST = 0.64 (TM-Align

Table 2 Template recognition performances of four threading and two protein structure prediction programs on TEST1 proteins using sequence identity cutoff as ≤95%

Query		E-Thrifty	POMONA	PSI-BLAST	CS23D	CS-ROSETTA	DALI
Protein name	PDB/ BMRB ID	Template TM-score (RMSD)	Template TM-score (RMSD)	Template TM-score (RMSD)	TM-score (RMSD)	TM-score (RMSD)	Template TM-score (RMSD)
KaiA	1M2F/5031	1YS7A 0.74 (2.61)	2WRZA 0.51 (3.51)	4D2OB 0.32 (4.79)	0.50 (4.43)	0.55 (3.84)	3R0JA 0.71 (2.99)
NEDTH	1F3Y/4448	4S2XA 0.77 (2.71)	2KDVA 0.81 (2.56)	4S2YA 0.77 (2.75)	0.72 (3.35)	0.34 (5.15)	4S2YA 0.79 (2.58)
NCS-1	2LCP/4378	4BY5B 0.66 (3.38)	1S1EA 0.70 (3.57)	4BY5D 0.68 (3.45)	0.68 (3.60)	0.63 (3.89)	1BJFB 0.70 (3.39)
Sortage	1IJA/4879	3RCCF 0.68 (3.08)	2W1KB 0.72 (2.95)	3RCCR 0.61 (3.23)	0.84 (2.17)	0.45 (4.10)	4TQXA 0.73 (2.69)
PyJ	1FAF/4403	1GH6A 0.67 (2.60)	1QDBB 0.43 (2.66)	2PF4H 0.64 (2.80)	0.30 (3.81)	0.77 (1.92)	2PF4E 0.71 (2.12)
ERp18	2K8V/15,964	3PH9B 0.63 (3.28)	3PH9B 0.67 (3.39)	3PH9A 0.63 (3.35)	0.73 (2.65)	0.44 (4.81)	1SENA 0.75 (2.73)
ApolPBP1A	2JPO/15,256	2FJYB 0.85 (2.22)	2FJYB 0.83 (2.35)	2KPHA 0.87 (1.84)	0.83 (1.77)	0.40 (4.17)	2KPHA 0.87 (1.84)
Pru Av 1	1E09/4671	4Z3LA 0.85 (2.16)	4C9CA 0.86 (2.10)	4C9IF 0.88 (2.05)	0.89 (1.88)	0.86 (2.25)	4C94B 0.91 (1.95)
LC1	1M9L/4265	2OMZ 0.53 (3.53)	2WSMA 0.31 (5.80)	5IL7B 0.46 (3.62)	0.94 (1.61)	0.57 (4.21)	5IL7A 0.62 (4.23)
cg2496	2KPT/16,569	5ANPB 0.70 (2.62)	3PVHA 0.67 (2.56)	3DFLA 0.30 (5.40)	0.36 (3.84)	0.77 (1.76)	3PVHA 0.69 (2.36)
NCAM	1EPFA/4162	5AEAA 0.92 (1.50)	2XY2A 0.88 (1.60)	5AEAB 0.81 (1.90)	0.76 (2.51)	0.42 (4.35)	2V44A 0.88 (1.60)
PG	2HZE/4113	1JHBA 0.86 (1.70)	1KTEA 0.85 (1.68)	1JHBA 0.83 (1.70)	0.91 (1.18)	0.88 (1.51)	1KTEA 0.89 (1.60)
AT5g22580	1RJJ/6011	5B0DB 0.67 (2.89)	1TR0C 0.69 (2.64)	2Q3PA 0.66 (3.18)	0.66 (3.10)	0.41 (3.75)	2QYCB 0.68 (2.68)
N-WASP	1MKE/5554	1QC6A 0.63 (2.42)	2XQNM 0.65 (2.39)	3SYXA 0.37 (3.49)	0.41 (4.48)	0.62 (2.73)	3OANA 0.63 (2.31)
Grx2	1G7O/4318	3IR4A 0.88 (1.64)	3IR4A 0.94 (1.30)	3IR4A 0.84 (1.75)	0.92 (1.91)	0.28 (6.12)	3IR4A 0.95 (1.25)
Average		0.74 (2.60)	0.70 (2.74)	0.64 (3.0)	0.70 (2.82)	0.56 (3.64)	0.77 (2.42)

The result includes E-Thrifty, POMONA, PSI-BLAST, DALI identified template information and structure prediction by CS23D and CS-Rosetta. Template information includes the PDB ID of the template and the TM-score of full-length model generated using the corresponding template. A TM-score evaluation of predicted structures by CS23D and CS-Rosetta is also presented. RMSD values are shown in Angstroms (Å)

RMSD = 3.0 Å)). This was expected given that a ≤ 95% sequence identify cutoff allows near identical homologs to be used in modeling. We also tested the performance of CS23D (Wishart et al. 2008) and CS-Rosetta (Shen et al. 2008) on the 15 proteins listed in Table 2. Both of these programs perform fairly well for these test cases. In particular, CS23D has an average TM-score of 0.70 (TM-Align RMSD = 2.82 Å) whereas CS-Rosetta has an average TM-score of 0.56 (TM-Align RMSD = 3.64 Å). The purpose of this evaluation is to show that threading algorithms are applicable to a wide range of applications and to assure users that chemical shift threading methods can attain a structure modeling performance similar to sequence-only threading or comparative modeling methods.

As a second test for E-Thrifty, we selected a number of recent CASD-NMR-2013 (Rosato et al. 2015) targets which we called TEST2. The TEST2 proteins consist of ten targets for which the majority of these proteins are structurally dissimilar to most (or even all) proteins in the PDB. It is important to note that both POMONA and DALI failed to find template(s) for one of the TEST2 proteins (2LOJ) using the ≤ 30% sequence identity threshold, as indicated by "N/A" in Table 3. In particular, POMONA returned no result, whereas DALI identified two possible template chains, 4YNXA and 4YNXB, but both had > 30% sequence identity to the query protein. For this "difficult" data set, E-Thrifty achieved an average TM-score of 0.64 (TM-Align RMSD = 2.58 Å) with ≤ 30% identity range (Table 3). As seen in Table 3, E-Thrifty

Table 3 Template recognition performances of E-Thrifty, POMONA, PSI-BLAST, CS23D and CS-Rosetta protocols on TEST2 (ten blind targets from the CASD-NMR-2013 competition) proteins are shown

CASD-NMR targets			E-Thrifty	POMONA	PSI BLAST	CS23D	CS-ROSETTA	DALI
Protein name	PDB/BMRB ID	Length/(fold)	Template TM-score (RMSD)	Template TM-score (RMSD)	Template TM-score (RMSD)	TM-score (RMSD)	TM-score (RMSD)	Template TM-score (RMSD)
NTPASE	2LCI/17613	134/(α/β)	5GAJA 0.84 (2.02)	2L69A 0.75 (2.36)	5GAJA 0.82 (2.37)	0.91 (1.44)	0.78 (2.94)	5GAJA 0.84 (2.02)
BUB1	2LAH/17524	160/(α)	4AEZC 0.80 (2.27)	3ESLB 0.81 (2.47)	3ESLA 0.80 (2.25)	0.85 (1.73)	0.63 (3.88)	3ESLA 0.80 (2.25)
FUS	2LA6/17508	99/(α/β)	2J0QG 0.72 (2.57)	1RK8A 0.75 (2.48)	4FXVA 0.71 (2.55)	0.67 (2.72)	0.66 (2.39)	3EX7G 0.76 (2.15)
NFU1	2M5O/19068	97/(α/β)	1TH5A 0.54 (2.33)	3R5GA 0.47 (3.40)	3B0HB 0.28 (4.53)	0.63 (3.16)	0.73 (2.51)	1TH5A 0.65 (2.69)
DNAJC2	2M2E/18909	73/(α)	2XB0X 0.64 (2.66)	3ZNVA 0.44 (3.11)	4EEFI 0.54 (2.89)	0.59 (2.76)	0.69 (2.25)	2CQQA 0.62 (1.99)
NKX 3.1	2L9R/17,484	69/(α)	2DMPA 0.68 (2.21)	2R5YB 0.66 (2.18)	3NAUB 0.51 (2.88)	0.71 (1.85)	0.52 (3.24)	1HF0B 0.71 (2.59)
NFU1	2LTM/18489	107/(α/β)	2K1HA 0.57 (3.54)	2XIQA 0.30 (4.50)	2K1HA 0.15 (3.85)	0.48 (3.86)	0.53 (3.46)	2FFMA 0.64 (2.51)
YR313A	2LTL/18487	119/(α/β)	1PQXA 0.59 (3.50)	2QTFA 0.27(4.37)	4L6EA 0.26 (4.96)	0.39 (3.39)	0.50 (2.46)	2FFMA 0.59 (2.73)
TSTM	2LOJ/18214	63/(β)	2FE0A 0.31 (3.01)	N/A	2JRAB 0.38 (2.07)	0.36 (2.43)	0.40 (2.32)	N/A
IF3-like fold	2LN3/18145	83/(α/β)	3ZIIA 0.74 (2.12)	3PP7A 0.51 (2.75)	4RJVD 0.47 (2.73)	0.61 (2.47)	0.27 (3.73)	3ZIEB 0.71 (2.48)
Average[a]			0.64 (2.58)	0.55 (3.10)	0.50 (3.22)	0.62 (2.60)	0.57 (3.0)	0.70 (2.38)

Note that CS23D and CS-Rosetta were run without any homology threshold on these proteins. The final column describes the highest possible alignment quality within the specified sequence identity threshold (≤30%) for these ten proteins. Each threading program column includes the PDB ID of the identified template, TM-score and TM-Align RMSD of the full-length model generated using the corresponding template. NA (TM-score = 0.0) in any column indicates "No Answer" for the corresponding method

[a] The average TM-score and RMSD in the final row are calculated over nine entries (excluding 2LOJ, for which DALI has "NA"). RMSD values are shown in Angstroms (Å)

performance is closer to the average TM-score achieved by the "gold standard" structure superposition program DALI than performances of other methods. Moreover, most of the top 10 hits obtained by E-Thrifty for the 10 targets were also identified by DALI through 3D structural superposition (data not shown), which confirms that E-Thrifty generally finds the correct answer for most (if not all) cases. The average TM-scores for POMONA and PSI-BLAST were 0.55 (TM-Align RMSD = 3.10 Å) and 0.50 (TM-Align RMSD = 3.22 Å), respectively. Again, the E-Thrifty results were significantly better for TEST2 proteins when compared to PSI-BLAST and POMONA (*p* values of 0.005 and 0.009, respectively). CS-Rosetta and CS23D are also run on the same set of proteins yielding average TM-scores of 0.57 (TM-Align RMSD = 3.0 Å) and 0.62 (TM-Align RMSD = 2.60 Å), respectively. Examples of full-length E-Thrifty template models for four proteins selected from TEST1 and TEST2 are shown in Fig. 2.

In addition to the studies of the TEST1 and TEST2 proteins, we also investigated the performance of E-Thrifty on 14 recent, randomly selected, BMRB entries with ≤95% sequence identity cutoff (TEST3). The results including the TM-score and RMSD from the TM-align program (Zhang and Skolnick 2004) are shown in Table 4. As can be seen in this table, E-Thrifty reports an average TM-score of 0.69 (TM-Align RMSD = 2.30 Å) for these newly deposited entries. A total of 13/14 proteins (93%) have approximate TM-scores ≥ 0.50, while only a single protein (BMRB: 30079, PDB: 5JTK) has a TM-score (0.42) below this threshold. Further investigation through the PDB shows that 5JTK is one of the few recent examples of a completely "novel" fold being recently deposited into the PDB. Interestingly, the top template identified for this particular entry (PDB: 4A7H, chain B) by the other chemical shift-based method, POMONA, only has a TM-score of 0.23.

E-Thrifty accurately handles "trick" proteins

Distinguishing between proteins with completely different folds but high sequence identity is an extremely challenging

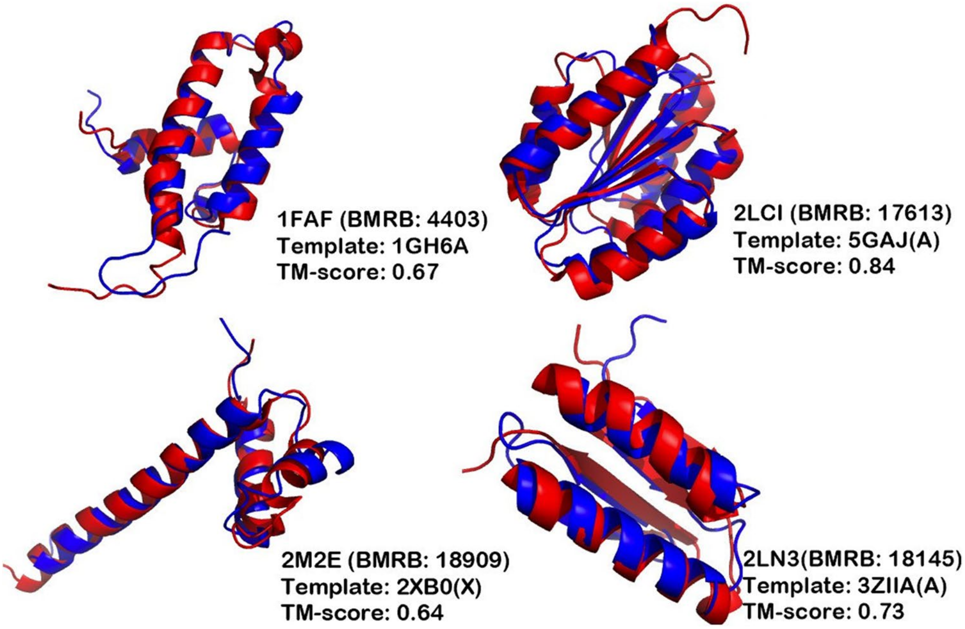


Fig. 2 The full-length E-Thrifty template models for four query proteins are shown. Query proteins (red) are shown superimposed with the template models (blue) using PyMOL. The TM-score of the aligned residue pairs between query and template models is displayed beside each superimposed model. All protein models were generated using PyMOL (http://www.pymol.org)

Table 4 E-Thrifty performances on 14 recent BMRB entries are shown

BMRB	TM-score (RMSD)
34002	0.55 (3.01)
30118	0.84 (1.97)
30098	0.84 (1.90)
30088	0.77 (1.40)
30079	0.42 (3.48)
30050	0.82 (2.13)
30048	0.76 (2.01)
30034	0.49 (3.29)
30023	0.72 (2.55)
30017	0.68 (2.31)
30007	0.79 (1.69)
30003	0.56 (1.97)
30023	0.72 (2.55)
30048	0.76 (2.01)
Average	0.69 (2.30)

The columns describe the BMRB ID, TM-score, and TM-Align provided RMSD, respectively. RMSD values are shown in Angstroms (Å)

task for sequence-based threading methods (Shen et al. 2008). One such example is the protein G pair known as GA-95 (PDB: 2KDL, BMRB: 16116) and GB-95 (PDB: 2KDM, BMRB: 16117) (Alexander et al. 2009). Protein GB-95 is a native, β-sheet rich protein that shares a high degree of sequence identity (> 80%) with a specially designed, mostly helical protein GA-95. Here, we investigated how E-Thrifty performed in predicting the structures of these two proteins when compared to other methods (POMONA and PSI-BLAST). As can be seen in Fig. 3, E-Thrifty generated accurate structures for both proteins, yielding TM-scores of 0.54 and 0.74 for GA-95 and GB-95 respectively. The TM-scores of POMONA-identified templates for GA-95 and GB-95 were 0.33 and 0.80 respectively. On the other hand, the TM-scores of the PSI-BLAST identified templates for GA-95 and GB-95 were 0.37 and 0.39 respectively. Clearly, PSI-BLAST was misled by the high sequence identity between these two very different folds. On the other hand, E-Thrifty (and to a lesser extent POMONA) were not fooled because both exploited chemical shift-derived structural information to avoid the sequence matching “trap”.

Another example of a “trick” protein fold is a denatured protein. In this situation, if sequence information alone is used, the sequence of the unfolded protein would match well with one or more folded proteins, yielding an incorrect structure prediction. For this example we used the data on denatured ubiquitin in 8 M urea (BMRB: 4375). While E-Thrifty was able to generate 3D coordinates for an ensemble of unfolded structures, POMONA failed to produce a structure, and PSI-BLAST, as expected, identified a template corresponding to folded ubiquitin (PDB: 5KGF, chain M). Importantly, the E-Thrifty generated unfolded ensemble exhibited stronger agreement with the observed chemical shifts and TALOS-N predicted torsion angles (see Table S1) in comparison to its folded counterpart (PDB: 5KGF, chain M). In other words, only E-Thrifty was able to identify unfolded proteins and generate viable 3D coordinates for those proteins using both sequence and chemical shift data.

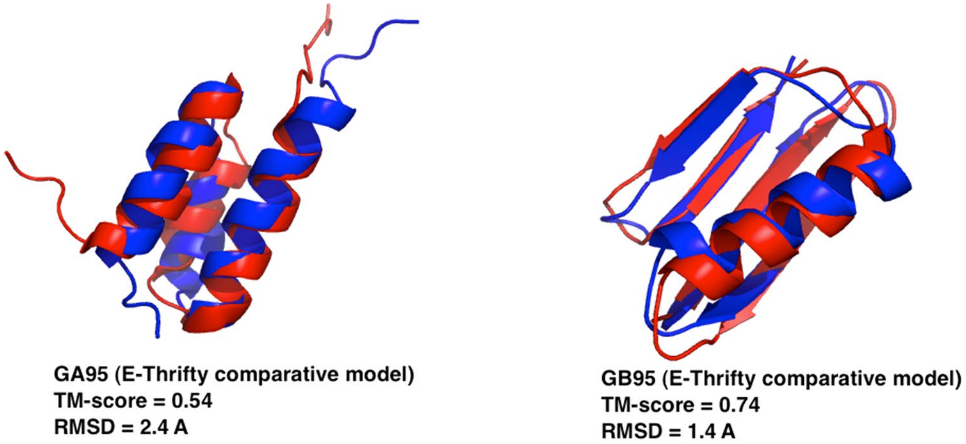

GA95 (E-Thrifty comparative model)
TM-score = 0.54
RMSD = 2.4 A
GB95 (E-Thrifty comparative model)
TM-score = 0.74
RMSD = 1.4 A

Fig. 3 E-Thrifty performances for "trick" proteins GA (95) and GB (95). The MODELLER generated E-Thrifty comparative models (blue) and the query proteins (red) are shown superimposed. The TM-score and backbone RMSD of each model are referred below

E-Thrifty's sensitivity to missing chemical shifts

It is notable that E-Thrifty was able to find a high quality, structurally similar homolog for one of the CASD-NMR-2013 targets, "YR313A" (2LTL) despite its complete lack of ^{13}CO backbone chemical shift assignments. On the other hand, POMONA failed to detect a good quality template (i.e. a template with a TM-score $\geq$ 0.5) for "YR313A" due to the missing ^{13}CO shifts. To further explore how well E-Thrifty performed with partial or incomplete chemical shift assignments, we chose a subset of five proteins from TEST1 and TEST2. Using these proteins, we systematically excluded ^{13}C, ^{1}H, and ^{15}N chemical shifts from each of these proteins and assessed E-Thrifty's structure prediction performance. With the exception of a few cases, E-Thrifty reported roughly the same results for these test proteins with partial assignments as it showed with the full assignments with a 30% sequence identity cut-off (Table S2). We believe there are several reasons for E-Thrifty's robustness towards incomplete chemical shift assignment. First, the inclusion of sequence and secondary structure pre-filtering stages assists with choosing a pool of relevant (in terms of sequence and secondary structure) templates before E-Thrifty's sequence-structural alignments are performed. Second, the embedded structure parameter calculators in E-Thrifty such as CSI 2.0, CSI 3.0, and ShiftASA have been previously shown to be quite insensitive to moderate levels of assignment incompleteness (Hafsa and Wishart 2014; Hafsa et al. 2015a, b). Obviously, there is a point at which missing chemical shifts begin to adversely affect performance. For example, if a significant number of chemical shifts are missing, in particular ^{1}H$_{\alpha}$ shifts and/or all ^{13}C (^{13}C$_{\alpha}$, ^{13}C$_{\beta}$, ^{13}CO) shifts, we observed a much poorer performance (see Table S2). As a result, E-Thrifty now performs a shift completeness check on all queries and generates a warning if a large number of ^{1}Hα and ^{13}Cα chemical shifts are missing.

TM-score distributions of models from E-Thrifty and other methods

The quality (TM-score) of E-Thrifty models was also compared to the quality of of homologs found by DALI as well as the top structures generated by POMONA. For this comparison, DALI homologs with a Z-score $\geq$ 2, and the top 250–300 alignments from POMONA and E-Thrifty were chosen. The TM-score distribution of the templates identified by the two threading programs (POMONA and E-Thrifty) and DALI (a structure matching program) for two proteins, Pru Av 1 protein (PDB: 1E09, BMRB: 4671) and the P-LOOP NTPase fold (PDB: 2LCI, BMRB: 17613) derived from TEST1 and TEST2 set, respectively, is shown in Fig. 4. Note that 550 DALI, 550 POMONA, and 550 E-Thrifty alignments were used in this comparison. Relative to POMONA, E-Thrifty shows a comparable or better performance in almost all TM-score sub-ranges (see Fig. 4). It is notable that E-Thrifty's performance was nearly as good as POMONA (307 vs. 329 in 550 alignments) in terms of identifying generally similar folds (TM-score $\geq$ 0.5). These high-quality templates consist of more than 55% of the total templates identified by E-Thrifty. In the case of random and unrelated fold rejection (TM-score $\leq$ 0.3), E-Thrifty showed a better performance (only 5% of total alignments) than POMONA (which had 12% of total alignments).

Detecting remote homologs

All ten targets selected from the CASD-NMR-2013 competition (the TEST2 data set) are proteins with very low sequence identity (< 30%) to any known structure in the PDB. Two proteins, 2LOJ (BMRB: 18214) and 2LN3 (BMRB: 18145), proved to be particularly challenging for almost all of the programs we tested. Both of these proteins have low sequence identity homologs in the PDB. E-Thrifty was able to correctly identify the most likely template for 2LN3 (3ZIIA, TM-score = 0.73) according to DALI. For

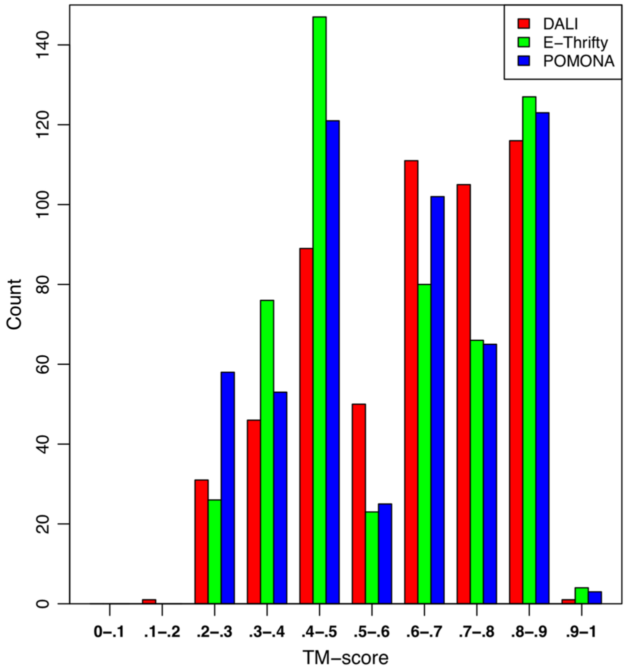


Fig. 4 TM-score distribution of full-length models using input templates identified by DALI, E-Thrifty and POMONA for the P-LOOP NTPase fold (PDB: 2LCI, BMRB: 17613) and the Pru Av 1 protein (PDB: 1E09, BMRB: 4671)

2LOJ, E-Thrifty and PSI-BLAST identified 2FE0A and 2JRAA as the best templates with TM-scores of 0.31 and 0.37, respectively. DALI also failed to find a template with a better TM-score (i.e. TM-score > 0.37) within its threshold Z-score ≥ 2 and ≤ 30% identity threshold. Further searches through the PDB to identify other structural homologs for 2LOJ revealed only one other homolog, 4YNX (with > 50% sequence identity), which was also identified by DALI. This structure was solved in 2015, which was after our exclusion date. Therefore, at the time of its deposition in 2013, 2LOJ appears to have been one of those truly novel protein folds that are now identified only very rarely.

E-Thrifty performance using different combinations of sequence/structure features

We analyzed E-Thrifty's performance using different sequence/structure feature combinations with a sequence identity threshold of ≤ 30%. This was done to assess which properties (sequence, secondary structure, chemical shifts, torsion angles, etc.) were most important for E-Thrifty's overall performance. For this assessment, we randomly chose seven proteins from both the TEST1 and TEST2 sets. As can be seen in Table S3, using sequence as the only input, E-Thrifty produces an average TM-score of 0.51 (compared to a TM-score of 0.67 for the full E-Thrifty program). Using simple sequence alignments, only 4/7 of these proteins generated high quality matches (TM-score > 0.50) or found the correct "gold standard" structural homolog. Using the combination of sequence and shift-derived secondary structure, E-Thrifty showed an improved average TM-score of 0.61 (compared to 0.67 for the full program). Using these two features, 6/7 proteins generated high quality (TM-score > 0.50) matches or found the correct "gold standard" structural homolog. Using the combination of sequence, shift-derived secondary structures, and shift-derived torsion angles, E-Thrifty shows an average TM-score of 0.66 with 6/7 of the test proteins generating high quality (TM-score > 0.50) matches. Interestingly, POMONA, which also uses these three features, was able to achieve a TM-score of 0.63. The primary intent of these experiments was to investigate contributions of different types of input data into the performance of E-Thrifty. These results clearly show that more information is better, and illustrate the important contribution of the three input features (sequence, secondary structure and torsion angles) in obtaining the overall threading performance of E-Thrifty.

E-Thrifty template accuracy and correlation with MODELLER structures

We analyzed the accuracy of E-Thrifty-identified templates by calculating the TM-scores of E-Thrifty templates with respect to the corresponding target structures for all of the TEST1 and TEST2 proteins. Average TM-scores of 0.67 and 0.65 were observed for the reference proteins in TEST1 and TEST2, respectively, and the corresponding E-Thrifty templates corresponding to ≤ 30% sequence identity, whereas the average TM-score was 0.75 between the TEST1 proteins and their corresponding E-Thrifty templates with ≤ 95% sequence identity. The average TM-scores from the E-Thrifty models of the TEST1 proteins (with ≤ 30% ID templates), the TEST2 proteins (with ≤ 30% ID templates), and the TEST1 proteins (with ≤ 95% ID templates) were nearly identical: 0.67, 0.64, and 0.74, respectively (Tables S4-S6). While analyzing these results, we also noticed a strong correlation (r = 0.98) between the TM-scores of the starting input templates and the TM-scores of the final MODELLER structures with respect to the reference structures. These results indicate that MODELLER successfully preserves the fold of E-Thrifty-identified templates during its comparative modelling and structure optimization steps.

It is important to note that, E-Thrifty target-template alignment coverage provides valuable information about the effectiveness of the template in modeling the target structure. E-Thrifty uses a cutoff or a minimal threshold for target-template alignment coverage (< 65% of the query sequence length) to identify when no useful template has been found. We have tested this thresholding approach on a number of cases and it appears to prevent the server from providing non-converged or incorrect structures (data not shown). If no match is found that exceeds this threshold, E-Thrifty returns a message "We are sorry, E-Thrifty could not find a useful template!". In addition, a secondary check is also performed on all output MODELLER structures to assess whether anything may have gone awry with the MODELLER runs (leading to an RMSD change of > 3.5 Å from the original input structure) and to prevent any "divergent" structures from being presented. In these cases E-Thrifty returns a message "We are sorry, the model building step appears to have failed!".

Query protein size limitations for E-Thrifty

The largest protein assessed among the TEST1 and TEST2 sets was 215 residues in length. Given that larger proteins are now being fully assigned by higher field NMR instruments using heteronuclear TROSY experiments, we decided to investigate what the upper size limit is for E-Thrifty, in terms of query protein length. A set of four large monomeric proteins with sequence lengths ranging from 255 to 341 residues was examined using a sequence identity cutoff of < 95%. In all cases, E-Thrifty returned satisfactory answers with TM-scores ranging from 0.5 to 0.7 (data not shown). The run-times varied from 12 to 18 min (from the smallest to the largest proteins). The time taken to identify the templates typically took 6–8 min, while an additional 6–10 min was required for MODELLER refinement. Overall, there appears to be no real size limitation to what E-Thrifty can handle. Indeed, the only limitation appears be what NMR can measure in terms of obtaining reasonably complete chemical shift assignments.

CS-GAMDy refinement of E-Thrifty models

CS-GAMDy (Berjanskii et al. 2015) is a newly developed NMR chemical shift-based protein structure refinement method. It uses a knowledge-based scoring function and structural information derived from chemical shifts through a combination of molecular dynamics and a multi-criterion genetic algorithm to perform structure refinement. The software is able to effectively refine and improve a wide range of approximate or even erroneous models. In our study, we used CS-GAMDy to refine the full-length models generated by MODELLER using the E-Thrifty identified templates. Examples of full-length models prior and following the refinement by CS-GAMDy for three query proteins (extracted from TEST1 and TEST2) are shown in Figure S3, with the corresponding Cα-RMSD values. Note that the reported Cα-RMSD is calculated only over the defined secondary structure regions of the superimposed proteins (query and template) using Superpose program (Maiti et al. 2004). As Figure S3 shows, CS-GAMDy was able to refine a number of local structural defects observed in the template models generated using MODELLER (e.g. lack of folding in helix and β-sheet regions) and improve the Cα-RMSD between well-formed secondary structure regions.

The E-Thrifty web server

E-Thrifty has been implemented as a web server, which can be accessed at http://ethrifty.ca. The E-Thrifty program was written using several programming languages including C++, Python, Perl, and R. The E-Thrifty web server has been implemented using PHP and CGI-scripting. The E-Thrifty server accepts BMRB (NMR-Star 2.1 or NMR-Star 3.1), SHIFTY and NEF-formatted (Gutmanas et al. 2015) chemical shift files and produces multiple output files. These output files include: (1) an alignment file showing the sequence-structure alignments for the top hits; (2) a summary file containing alignment scores, chemical shift scores, total scores, e-value, and the cluster membership associated with each hit and (3) a 3D structure (PDB coordinates) of the query protein using the top scoring template (or multiple templates) from the clusters. The E-Thrifty server supports a number of user-selectable options related to comparative modeling which includes the sequence identity threshold for template selection, the model building mode (model generation from a single template or multiple templates), and exclusion of flexible terminal regions in the modeling process. In general, the run-time for the E-Thrifty server is ~ 6–8 min on medium sized proteins (< 200 residues), whereas it generally takes ~ 10–15 min for larger proteins (> 200 residues). Note that The E-Thrifty run-times can climb to ~ 40–60 min (for medium sized proteins) and ~ 60–90 min (for larger proteins) when full-scale structural alignments must be performed against the nrPDB. However, this only occurs rarely. A screen shot of the E-Thrifty web server home page and its output is illustrated in Figure S4.

Limitations and potential improvements

One of the limitations of E-Thrifty is that it relies on categorical, somewhat imprecise character-based representations to describe both the query structures and the corresponding database structures. For example, real-valued φ/ψ angles are converted into a discretized 9-letter torsion alphabet; while real-valued fractional ASA (fASA) values are classified into an even simpler 3-letter alphabet. Approaches that use numeric torsion angles and numeric fASA values might be expected to further improve E-Thrifty's performance. This is because numeric values would be far more precise and would capture far more subtle information about these torsion angle and fASA features. While E-Thrifty makes use of a number of high quality dynamic programming alignment routines, a further improvement in its sensitivity for detecting remote homologs could potentially be achieved by including more powerful Hidden Markov Model profile alignments. These alignment methods have consistently proven to be very effective in detecting distant homologs (Krogh et al. 1994; Eddy 1998; Karplus et al. 1998) and appear to play a key role in the success achieved with programs such as HHpred (Söding 2005). While improved alignment methods could be particularly beneficial, improved scoring functions may prove to be equally useful. Indeed, we suspect further improvements could be achieved by designing a suitable Z-score value for a more effective assessment of E-Thrifty sequence-structure alignment quality.

Due to the high computational cost of generating statistical ensembles of disordered proteins, the E-Thrifty webserver calculates only so-called uncertainly ensembles (Bonomi et al. 2017) for this class of proteins. The main objective of the E-Thrifty algorithm is to identify good templates for comparative modelling of folded proteins. Describing conformational heterogeneity in disordered proteins is beyond the scope of this technique. However, we believe that E-Thrifty's models of disordered proteins can still be useful to gain initial insights into residual structures in these proteins or/and to be employed as starting models in more specialized methods of generating statistical ensembles of disordered proteins (e.g. maximum entropy methods).

Next steps for E-Thrifty

As a chemical shift threading method E-Thrifty is particularly good at automatically generating "approximate" or initial 3D protein models. However, to obtain truly high-quality, atomic resolution structures it will be necessary to couple E-Thrifty to other kinds of programs that can perform true structural refinement. As highlighted in Figure S3, E-Thrifty can be easily coupled to CS-GAMDy (Berjanskii et al. 2015) to perform chemical shift refinement. These refinement methods were shown to consistently improve the quality and accuracy of the starting structures. While CS-GAMDy refinement calculations can take several hours to several days, it is not unreasonable to imagine having E-Thrifty tightly coupled to CS-GAMDy (either as a standalone program or as a web server) in the near future. Further enhancements to E-Thrifty will likely include the addition of other structure refinement options such as XPLOR-NIH (Schwieters et al. 2003), AMBER (Pearlman et al. 1995), or DYANA (Güntert et al. 1997). Adding these tools to the pipeline would also allow E-Thrifty to incorporate other experimental NMR measures such as NOEs, J-couplings, and residual dipolar couplings into its structure generation and refinement protocols. In the rare situations where no structural homolog can be found, it may be possible to consider blending Cheshire (Cavalli et al. 2007), CS-Rosetta (Shen et al. 2008), and other ab initio structure predictors with E-Thrifty to improve model selection in these ab initio methods. Likewise, integrating a maximum entropy method of building protein ensembles, such as ENSEMBLE (Choy and Forman-Kay 2001) or chemical shift-biased metadynamics (Camilloni and Vendruscolo 2014), into E-Thrifty would enable generation of statistical ensembles of disordered proteins. All these additions would lead to the creation of a much more fail-safe and far more comprehensive chemical shift-based structure generation pipeline.

Conclusion

In this study, we have described a novel and effective protein fold recognition method, called E-Thrifty, that uses chemical shift threading to generate high quality 3D protein structures for proteins having little or no sequence identity to any protein in the PDB. We believe this represents a significant step towards "solving" protein structures using only chemical shift information. As outlined above, E-Thrifty uses chemical shift derived secondary structures, chemical shift derived fASA values and chemical shift derived torsion angles to perform a comprehensive alignment between the query sequence (with experimentally determined chemical shift assignments) and a large database of known protein structures and predicted chemical shifts. E-Thrifty exploits a number of recently developed chemical shift analysis tools—CSI 2.0 (Hafsa and Wishart 2014), CSI 3.0 (Hafsa et al. 2015a), ShiftASA (Hafsa et al. 2015b), TALOS-N (Shen and Bax 2013), and ShiftX2 (Han et al. 2011)—to generate chemical shifts or chemical-shift derived information for both the query and the database proteins. A Smith–Waterman local alignment algorithm with a variable gap penalty function was found to be the best tool for performing the sequence-structure alignment. The weighting coefficients used in scoring functions and fitness scores were optimized through both a parameter grid search and a linear regression analysis. E-Thrifty includes a chemical shift fitness score and an e-value scoring system to fully evaluate the alignments between the query and the database proteins. In addition, E-Thrifty performs a cluster analysis step for all identified folding templates to group them according to their structural similarity. The templates identified by E-Thrifty can then be used for comparative modeling with programs such as MODELLER (Sali and Blundell 1993) and subsequently for chemical shift-based structure refinement (Berjanskii et al. 2015).

In terms of performance, E-Thrifty achieved an average TM-score of 0.67 for query sequences having ≤ 30% sequence identity (as measured on an independent test set of 15 proteins). E-Thrifty's performance was found to be comparable to the "gold standard" DALI, which had an average TM-score of 0.69. In contrast to E-Thrifty or other structure prediction routines, DALI uses experimentally derived coordinate data to identify structural homologs of proteins by structural superposition (i.e. DALI knows the answer, whereas E-Thrifty predicts the answer). E-Thrifty was also evaluated on a number of recent CASD-NMR-2013 targets and achieved an average TM-score performance of 0.64 on 10 test proteins with ≤ 30% sequence identity. The performance of E-Thrifty clearly demonstrates its ability to "predict" a 3D fold by using only chemical shift information. With its exceptional performance, we believe that E-Thrifty could be a very useful contribution towards the goal of rapid, automated protein structure generation and refinement using NMR chemical shifts as the only source of experimental data.

Acknowledgements Financial support from the Natural Sciences and Engineering Research Council (NSERC), the Alberta Prion Research Institute (APRI) and the Canadian Institutes of Health Research (CIHR) is gratefully acknowledged.

References

Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2009) A minimal sequence code for switching protein structure and function. Proc Natl Acad Sci 106(50):21149–21154

Altschul SF, Gish W (1996) Local alignment statistics. Meth Enzymol 266:460–480

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Berjanskii MV, Wishart DS (2005) A simple method to predict protein flexibility using secondary chemical shifts. J Am Chem Soc 127(43):14970–14971

Berjanskii MV, Wishart DS (2013) A simple method to measure protein side-chain mobility using NMR chemical shifts. J Am Chem Soc 135(39):14536–14539

Berjanskii MV, Neal S, Wishart DS (2006) PREDITOR: a web server for predicting protein torsion angle restraints. Nucleic Acids Res 34(suppl 2):W63-W69

Berjanskii M, Tang P, Liang J, Cruz JA, Zhou J, Zhou Y et al (2009) GeNMR: a web server for rapid NMR-based protein structure determination. Nucleic Acids Res 37(suppl 2):W670-W677

Berjanskii M, Arndt D, Liang Y, Wishart DS (2015) A robust algorithm for optimizing protein structures with NMR chemical shifts. J Bio NMR 63(3):255–264

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242

Betancourt MR, Skolnick J (2001) Universal similarity measure for comparing protein structures. Biopolymers 59(5):305–309

Bonomi M, Heller GT, Camilloni C, Vendruscolo M (2017) Principles of protein structural ensemble determination. Curr Opin Struct Biol 42:106–116

Boratyn GM, Schaffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL (2012) Domain enhanced lookup time accelerated BLAST. Biol Direct 7(1):12

Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170

Camilloni C, Vendruscolo M (2014) Statistical mechanics of the denatured state of a protein using replica-averaged metadynamics. J Am Chem Soc 136(25):8982–8991

Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci 104(23):9615–9620

Choy WY, Forman-Kay JD (2001) Calculation of ensembles of structures representing the unfolded state of an SH3 domain. J Mol Biol 308:1011–1032

Eddy SR (1998) Profile hidden markov models. Bioinformatics 14:755–763

Ginzinger SW, Fischer J (2006) SimShift: identifying structural similarities from NMR chemical shifts. Bioinformatics 22(4):460–465

Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M 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

Güntert P, Mumenthaler C, Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 273(1):283–298

Gutmanas A, Adams PD, Bardiaux B, Berman HM, Case DA, Fogh RH et al (2015) NMR exchange format: a unified and open standard for representation of NMR restraint data. Nat Struct Mol Biol 22(6):433–434

Hafsa NE, Wishart DS (2014) CSI 2.0: a significantly improved version of the Chemical Shift Index. J Biomol NMR 60(2–3):131–146

Hafsa NE, Arndt D, Wishart DS (2015a) CSI 3.0: a web server for identifying secondary and super-secondary structure in proteins using NMR chemical shifts. Nucleic Acids Res 43(W1):W370–W377

Hafsa NE, Arndt D, Wishart DS (2015b) Accessible surface area from NMR chemical shifts. J Biomol NMR 62(3):387–401

Han B, Liu Y, Ginzinger SW, Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 50(1):43–57

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 89(22):10915–10919

Holm L, Rosenström P (2010) Dali server: conservation mapping in 3D. Nucleic Acids Res 38(suppl 2):W545-W549

Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition. Nature 358:86–89

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12):2577–2637

Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14(10):846–856

Krogh A, Brown M, Mian IS, Sjölander K, Haussler D (1994) Hidden markov models in computational biology—applications to protein modeling. J Mol Biol 235:1501–1531

Kuszewsk J, Gronenborn AM, Clore GM (1995) The impact of direct refinement against proton chemical shifts on protein structure determination by NMR. J Magn Reson 107:293–297

Lindahl E, Elofsson A (2000) Identification of related proteins on family, superfamily and fold level. J Mol Biol 295(3):613–625

Madhusudhan MS, Marti-Renom MA, Sanchez R, Sali A (2006) Variable gap penalty for protein sequence–structure alignment. Protein Eng Des Sel 19(3):129–133

Maiti R, Van Domselaar GH, Zhang H, Wishart DS (2004) SuperPose: a simple server for sophisticated structural superposition. Nucleic Acids Res 32(2):W590–W594

Medek A, Hajduk PJ, Mack J, Fesik SW (2000) The use of differential chemical shifts for determining the binding site location and orientation of protein-bound ligands. J Am Chem Soc 122(6):1241–1242

Mielke SP, Krishnan VV (2009) Characterization of protein secondary structure from NMR chemical shifts. Prog Nucl Magn Reson Spectrosc 54(3):141–165

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4):536–540

Ortiz AR, Strauss CE, Olmea O (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 11(11):2606–2621

Osapay K, Case DA (1991) A new analysis of proton chemical shifts in proteins. J Am Chem Soc 113(25):9436–9444

Osapay K, Theriault Y, Wright PE, Case DA (1994) Solution structure of carbonmonoxy myoglobin determined from nuclear magnetic resonance distance and chemical shift constraints. J Mol Biol 244:183–197

Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE, DeBolt S et al (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Comput Phys Commun 91(1):1–41

Peng J, Xu J (2010) Low-homology protein threading. Bioinformatics 26(12):i294-i300

Rosato A, Vranken W, Fogh RH, Ragan TJ, Tejero R, Pederson K, Vuister GW (2015) The second round of critical assessment of automated structure determination of proteins by NMR: CASD-NMR-2013. J Biomol NMR 62(4):413–424

Rost B (1995) TOPITS: threading one-dimensional predictions into three-dimensional structures. Proc Int Conf Intell Syst Mol Biol 3:314–321

Rost B, Schneider R, Sander C (1997) Protein fold recognition by prediction-based threading. J Mol Biol 270(3):471–480

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815

Schwarzinger S, Kroon GJ, Foss TR, Chung J, Wright PE, Dyson HJ (2001) Sequence-dependent correction of random coil NMR chemical shifts. J Am Chem Soc 123(13):2970–2978

Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160(1):65–73

Sharma D, Rajarathnam K (2000) C-13 NMR chemical shifts can predict disulfide bond formation. J Biomol NMR 18:165–171

Shen Y, Bax A (2013) Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. J Biomol NMR 56(3):227–241

Shen Y, Bax A (2015) Homology modeling of larger proteins guided by chemical shifts. Nat Methods 12(8):747–750

Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G et al (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci 105(12):4685–4690

Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44(4):213–223

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7(1):539

Siew N, Elofsson A, Rychlewski L, Fischer D (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 16(9):776–785

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147(1):195–197

Söding J (2005) Protein homology detection by HMM–HMM comparison. Bioinformatics 21(7):951–960

Song Y, DiMaio F, Wang RYR, Kim D, Miles C, Brunette TJ et al (2013) High-resolution comparative modeling with RosettaCM. Structure 21(10):1735–1742

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J et al (2008) BioMagResBank. Nucleic Acids Res 36(suppl 1):D402–D408

Vranken WF, Rieping W (2009) Relationship between chemical shift value and accessible surface area for all amino acid atoms. BMC Struct Biol 9(1):20

Wang G, Dunbrack RL (2003) PISCES: a protein sequence culling server. Bioinformatics 19(12):1589–1591

Wishart DS (2011) Interpreting protein chemical shift data. Prog Nucl Magn Reson Spectrosc 58(1):62–87

Wishart DS, Case DA (2001) Use of chemical shifts in macromolecular structure determination. Methods Enzymol 338:3–34

Wishart DS, Nip AM (1998) Protein chemical shift analysis: a practical guide. Biochem Cell Biol 76:153–163

Wishart DS, Sykes BD (1994a) Chemical shifts as a tool for structure determination. Methods Enzymol 239:363–392

Wishart DS, Sykes BD (1994b) The 13C chemical shift index: a simple method for the identification of protein secondary structure using 13C chemical shift data. J Biomol NMR 4(2):171–180

Wishart DS, Sykes BD, Richards FM (1992) The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. Biochemistry 31(6):1647–1651

Wishart DS, Arndt D, Berjanskii M, Tang P, Zhou J, Lin G (2008) CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. Nucleic Acids Res 36:W496–W502

Xu J, Zhang Y (2010) How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics 26(7):889–895

Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57:702–710