J Biomol NMR (2009) 43:179–185
DOI 10.1007/s10858-009-9301-7

ARTICLE

SimShiftDB; local conformational restraints derived from chemical shift similarity searches on a large synthetic database

Simon W. Ginzinger · Murray Coles

Received: 8 September 2008 / Accepted: 7 January 2009 / Published online: 18 February 2009

Abstract We present SimShiftDB, a new program to extract conformational data from protein chemical shifts using structural alignments. The alignments are obtained in searches of a large database containing 13,000 structures and corresponding back-calculated chemical shifts. SimShiftDB makes use of chemical shift data to provide accurate results even in the case of low sequence similarity, and with even coverage of the conformational search space. We compare SimShiftDB to HHSearch, a state-of-the-art sequence-based search tool, and to TALOS, the current standard tool for the task. We show that for a significant fraction of the predicted similarities, SimShiftDB outperforms the other two methods. Particularly, the high coverage afforded by the larger database often allows predictions to be made for residues not involved in canonical secondary structure, where TALOS predictions are both less frequent and more error prone. Thus SimShiftDB can be seen as a complement to currently available methods.

Keywords Chemical shift · Homology modeling · Alignment · Database search

Introduction

Chemical shifts are now routinely used as a source of local conformational restraints in the structure determination of proteins by NMR, due mostly to the widespread use of programs such as TALOS (Cornilescu et al. 1999) and SHIFTOR/PREDITOR (Neal et al. 2006; Berjanskii et al. 2006). These programs share a common approach and output similar data; both search a database that correlates local patterns of chemical shifts with local conformation, and both provide backbone dihedral angle restraints for individual residues. This approach has been very successful, but has some limitations in the stringent criteria needed for selecting proteins or protein fragments to populate the database, i.e. only those with both highly reliable chemical shift and structural data can be included. This restricts current databases to less than a few hundred proteins. Although this may seem adequate—for example, the TALOS database contains 186 proteins subdivided into over 24,000 tripeptide fragments (http://spin.niddk.nih.gov/NMRPipe/talos/)—the sequence/conformation search space is very large, and the database coverage is unevenly distributed. As a result, rare combinations of amino-acid and conformation may be under-represented in the database, leading to significant under-prediction and even to errors outside of the heavily populated regions of the Ramachandran map.

We have adopted an alternative approach to extracting structural data from chemical shifts based on our SimShiftDB algorithm. The original SimShift was designed to test for structural similarities between proteins

Most of the work by Simon W. Ginzinger was done while being a research fellow at the Department of Bioinformatics, Ludwig-Maximilians-Universität Munich (LMU), Amalienstrasse 17, 80333 Munich, Germany.

S. W. Ginzinger (✉)
Department of Molecular Biology, Division of Bioinformatics,
Center of Applied Molecular Engineering, University
of Salzburg, Hellbrunnerstr. 34/3.OG, 5020 Salzburg, Austria
e-mail: simon@came.sbg.ac.at

M. Coles
Department of Protein Evolution, Max-Planck-Institute
for Developmental Biology, Spemannstrasse. 35,
72076 Tübingen, Germany
e-mail: Murray.Coles@tuebingen.mpg.de

in a pair wise manner using chemical shifts to supplement sequence data. Experimental query shifts were compared to those back-calculated from the target. SimShift showed improved ability to detect distant structural relationships when compared to state-of-the-art methods based on the sequence alone. A natural further development of pairwise comparison was to adapt the SimShift algorithm for database searching, resulting in SimShiftDB (Ginzinger et al. 2007b). Given a target sequence and shifts, SimShiftDB provides a list of matching proteins in the database, scored by a measure of statistical significance. In effect it searches a synthetic chemical shift database of 13,000 proteins based on the Astral library (Chandonia et al. 2004). The matching sequence can be of any length, and structurally similar regions can be found ranging from small, locally similar fragments up to full domains.

In principle, any structural alignment method can also be used to make predictions of local conformation by extracting torsion angles from matching regions of the target proteins, and it is this implementation of the SimShiftDB algorithm we present here. We benchmark the program against TALOS as a standard for current methods and HHpred, a sequence search method based on hidden Markov models (Söding et al. 2005), as a standard for purely sequenced based methods. We show that SimShiftDB can significantly increase the amount of information that can be derived from chemical shifts. We combine SimShiftDB with our CheckShift (Ginzinger et al. 2007a) routine for standardizing chemical shift referencing to produce a pipeline for analysis of chemical shift data.

Implementation

Full details of the SimShiftDB algorithm have been published previously (see Ginzinger et al. 2007b), but it is worthwhile giving a brief overview here (see Fig. 1). Given a target protein, SimShiftDB analyzes each possible pairing of the target protein with one of the template protein structures in the database. Here we use the proteins from the ASTRAL database (version 1.71, filtered for 95% sequence identity) to build the SimShiftDB Template Database. As SimShiftDB is based on the comparison of chemical shift data, the chemical shifts for all entries in the template database are back-calculated from the three-dimensional structure using SHIFTX (Neal et al. 2003). For each combination of the target protein with a template from the database, a pairwise alignment is calculated. The alignment algorithm works in two steps:

Step 1: Local similarities are found by looking for high scoring combinations of parts of the target protein sequence (s) with parts of the template protein sequence (t). Fig. 1 shows a depiction of a set of local similarities. For example, block b in the figure shows that the chemical shifts of the target protein sequence from index X_{min} to X_{max} are similar to the chemical shifts of the template protein sequence from index Y_{min} to Y_{max}. The similarity is calculated by summing the pairwise scores of the residues in the similar region in analogy to a pairwise sequence alignment. The pairwise similarity scores are given by the so-called Chemical Shift Substitution Matrices, which give a score for each combination of two residues with associated chemical shifts (for more details see Ginzinger 2008).

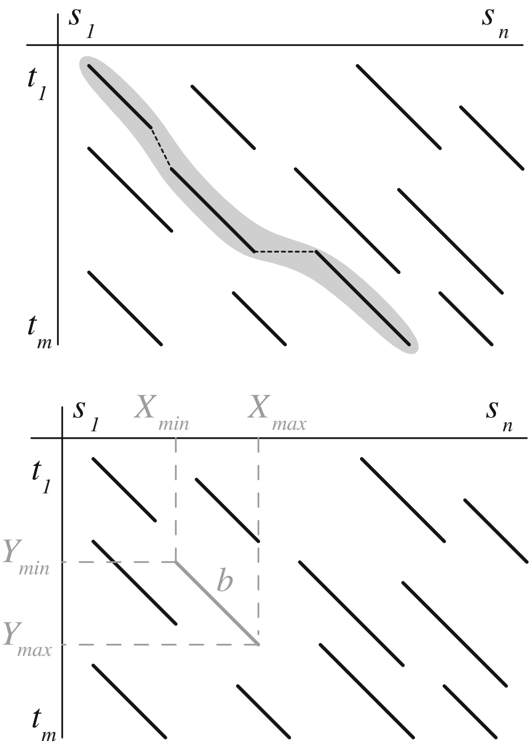


Fig. 1 An explanation of the SimShift algorithm. *Top*: an example for a set of local similarities for a target protein (s) and a template protein (t). The notation is as explained in the text. *Bottom*: combination of a subset of local similarities to yield a self-consistent final alignment

Step 2: The set of local similarities from Step 1 is taken as an input for Step 2, where the most significant combination of blocks is identified, according to a statistical model of alignment scores (Karlin and Altschul 1993). Additionally, two blocks have to fulfill two constraints for their combination to be considered:

1. Blocks may not overlap in the target or in the template protein; this would otherwise result in an ambiguous alignment.

2. As the three-dimensional structure of the template protein is known, we further require that the euclidean distance between the end of the first block and the beginning of the second block may be bridged (according to chemical restraints) by the relevant sequence of amino acids in the target protein.

Finally, we calculate an *e*-value for the optimal combination of blocks. This *e*-value represents the number of alignments of equal or better quality, which are expected to occur by chance, given the distribution of the amino acids with associated chemical shifts in the target protein and the template database. Additionally, the *e*-value takes the size of the template database into account. According to the following evaluation, an *e*-value of $<10^{-3}$ guarantees a high quality alignment.

Results

The benchmark set

To test the performance of SimShiftDB, a benchmark set has to be defined for which both chemical shifts and the three-dimensional structure of the protein are available. The BMRB (Seavey et al. 1991) is the main public repository for chemical shift data. However, there is no consistent mapping to the structural databases, making it difficult to relate structural with chemical shift information reliably. Therefore a mapping between BMRB and ASTRAL is calculated based on amino acid sequence similarity. Every entry in the benchmark set has to fulfill the following constraints:

A 100% sequence match to an ASTRAL entry.

At least 100 residues with associated chemical shifts (to exclude very short protein fragments; e.g. single helices).

To identify protein structures corresponding to the respective BMRB entries, a BLAST-search (Altschul et al. 1990) against the sequences from the ASTRAL database is conducted for each BMRB entry. If the full BMRB sequence can be matched without gaps against an ASTRAL sequence, the corresponding ASTRAL structure is assigned to the BMRB entry. As some entries in BMRB match more than one sequence in ASTRAL, one representative structure has to be chosen. This is accomplished by using the AEROSPACI score (Chandonia et al. 2004) provided for each ASTRAL entry, thereby selecting the structure with the best resolution. Through this procedure a benchmark set containing 144 entries was derived.

Evaluation of prediction accuracy

When calculating the similarity score for two residues, SimShiftDB is restricted to at most three chemical shifts from the following list: $^{1}H_{\alpha}$, $^{1}H^{N}$, ^{15}N, $^{13}C_{\alpha}$, $^{13}C_{\beta}$, and $^{13}C'$. Thus it is important to select a combination of shifts to extract maximum information, and a priority for replacing missing shifts. To identify the most successful strategy, we tested all possible priorities for the six atom types, resulting in $6! = 720$ evaluations. The most successful priority was: $^{13}C_{\alpha} > {}^{13}C' > {}^{1}H^{N} > {}^{13}C_{\beta} > {}^{1}H_{\alpha} > {}^{15}N$. This is the default priority, and is used in the following analysis.

To evaluate the prediction accuracy, we applied the program to all entries in the benchmark set, using all proteins from the SimShiftDB Template Database as potential templates. Subsequently, we used all alignments that achieved an *e*-value better than 10^{-3} to infer torsion angles for the target residue from the associated template residues. If residues of the target were mapped multiple times, we based the prediction on the highest scoring alignment. It is important to evaluate the performance of SimShiftDB as a function of sequence similarity; therefore 9 evaluations were performed with the maximum allowed sequence similarity in the evaluated alignments set between 20 and 100%. Figure 2 presents the results of this analysis. About 70% of all torsion angles predicted using SimShiftDB have a high accuracy (≤15° error). Another 10 to 20% of the predicted angles have an error of less than 30°. Therefore SimShiftDB yields accurate results in 85 to 90% of the evaluated predictions. Through the use of the chemical shifts, this performance is *nearly independent* of the percentage of sequence identity in the respective alignments.

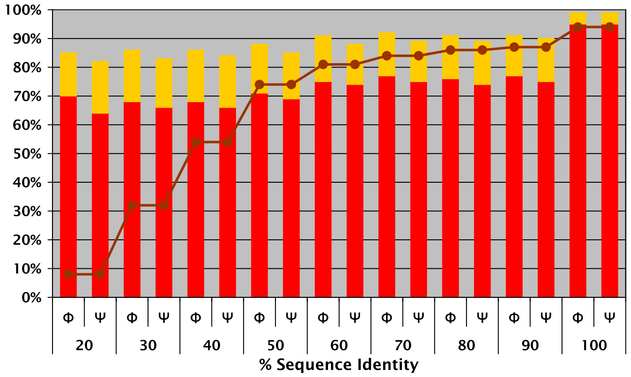


Fig. 2 Evaluation of torsion angle predictions calculated using SimShiftDB, separated according to the maximum allowed sequence identity between query and target. The boxes show the percentage of dihedral angle predictions with an error of ≤30°. The red part of each box shows the percentage of predictions with an error of ≤15°. The brown line represents the percentage of residues from the test set for which a prediction exists

Also of interest is the performance of SimShiftDB on different types of secondary structure. Figures 3 and 4 show the difference between the secondary structure content in all predictions versus that in high quality and erroneous predictions, respectively. It can be seen that predictions for sheet (for both high quality and erroneous predictions) match well with the percentage observed in all predictions, and this match is largely independent of sequence similarity. In contrast, the percentage of high quality predictions in helix increases with decreasing sequence similarity, whereas the corresponding percentage in coil regions decreases. For erroneous predictions the inverse effect is observed. This seems logical, as the structures for coil regions are less reliable, and predictions are clearly harder to make than for secondary structure. This test shows empirically that SimShiftDB has no significant bias when comparing performance in predicting helix versus sheet. The independence of sheet predictions from sequence similarity indicates that the chemical shift data is more diagnostic for sheet than for helix.

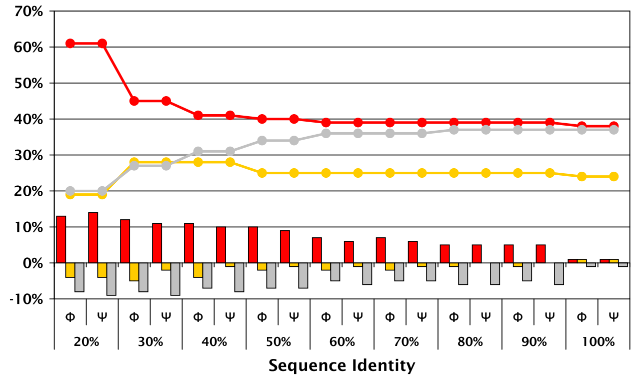


Fig. 3 Difference between secondary structure content in high quality predictions versus the content for all predictions. The bars show the difference in helix, sheet and coil content (red, yellow and grey, respectively). The lines show the overall secondary structure content in all predictions

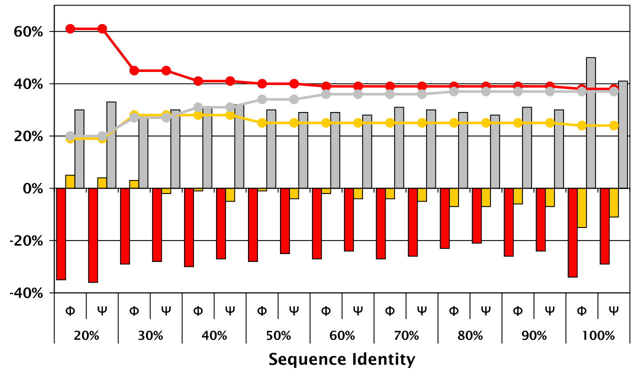


Fig. 4 Difference between secondary structure content in erroneous predictions versus the content for all predictions using the same scheme as in Fig. 3

Comparison to HHsearch

To show empirically that SimShiftDB uses the information in the chemical shift data to yield more sensitive alignments, especially in the case of low sequence similarity, we compare SimShiftDB to HHsearch (Söding 2005). HHSearch, a sensitive search tool based on hidden Markov models, calculates alignments between proteins using the primary sequence complemented by sequence-based predictions of secondary structure. HHpred (Söding et al. 2005), a protein structure prediction method based on HHsearch alignments, ranked second best in the CASP7 (Battey et al. 2007) experiment. Additionally, it is freely available for download and gives the user the possibility to define arbitrary template databases. Therefore it is perfectly suited to serve as a reference for purely sequence-based methods.

To compare the performance of SimShiftDB and HHSearch we used the benchmark set defined in the previous section. For both methods we ran each target protein against the SimShiftDB Template Database and used alignments achieving an *e*-value better than 10^{-3} to predict torsion angles for the residues in the target protein. If residues were mapped multiple times, the prediction was based on the highest scoring alignment. We then compared all SimShiftDB predictions to the respective HHsearch predictions. The following notation is used for the presentation of the results:-

> A SimShiftDB prediction is called better if it has an error of ≤30° and the corresponding HHSearch prediction has an error which is worse by more than 5°.
>
> Two predictions are called equal if both have an error of ≤30° and the difference between the errors is less than 5°, or both predictions have an error >30°.
>
> Missing predictions are treated as predictions with an error >30°.

Figure 5 shows the results of this evaluation for alignments with maximal sequence identities ranging from 10 to 100%. There is a clear trend for SimShiftDB to outperform HHSearch as the sequence identity decreases, demonstrating that SimShiftDB uses the structural information in chemical shifts to improve alignments.

Comparison to TALOS

It is important to compare SimShiftDB to the most prominent method for predicting torsion angles from chemical shifts, namely TALOS (Cornilescu et al. 1999). Again we used the benchmark set defined earlier, and made torsion angle predictions based on SimShiftDB alignments that achieved an *e*-value better than 10^{-3}. We then compared

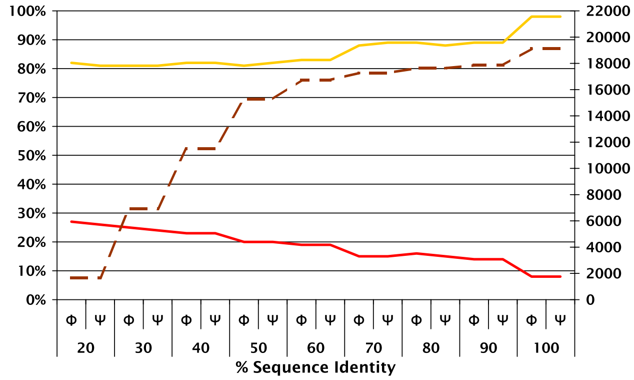


Fig. 5 SimShiftDB predictions compared to the respective HHSearch predictions, separated according to the maximum allowed sequence identity between query and target. The red line shows the percentage of predictions where SimShiftDB performs better, the yellow line show the percentage where SimShiftDB is either better or gives a result of equal quality. The brown line corresponds to the number of torsion angles predicted (right axis)

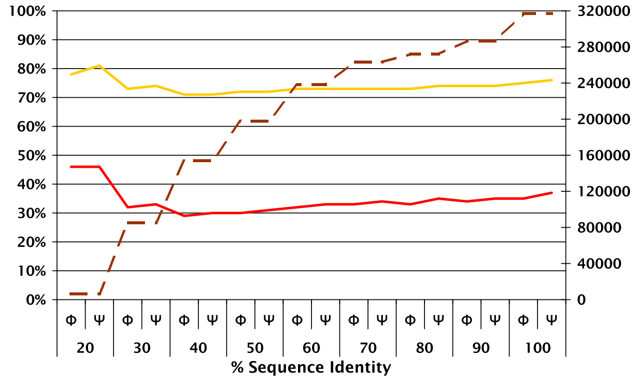


Fig. 6 SimShiftDB predictions compared to the respective TALOS predictions using the same notation as in Fig. 5

the quality of each torsion angle prediction to the corresponding TALOS prediction. The results are presented using the same notation as in the previous section. Figure 6 shows that SimShiftDB outperforms TALOS in at least 30% of all cases. It should be noted that there is likely to be a significant bias towards TALOS in these results, as many members of the benchmark set will have been calculated using TALOS restraints.

Discussion

We have presented SimShiftDB and shown that the program is able to sensitively extract structural information from chemical shift data. This information is to a certain extent complementary to that from currently available tools. On one hand we have compared SimShiftDB to a sequence-based method. SimShiftDB shows its strength especially in cases of low sequence similarity, which underlines the advantage of including chemical shift information in the alignment algorithm. On the other hand, we were able to show that one-third of the predictions by SimShiftDB clearly have a higher quality than the corresponding TALOS predictions, and this is largely independent of sequence similarity.

The main advantage of SimShiftDB is derived from its superior coverage of the search space, due to the large and quickly adaptable template database. SimShiftDB outperforms TALOS especially in those cases where TALOS finds no predictions classified as “Good” according to its selection criteria. SimShiftDB and TALOS are therefore complementary, and can be used in parallel to increase the number of available predictions.

The functional differences between SimShiftDB and TALOS are best illustrated by an example. Ph1500C is a 78-residue homo-hexameric domain currently under investigation in our laboratories, and was chosen because it shows no significant sequence similarity to proteins of known structure. We used SimShiftDB to search for templates matching Ph1500C, using several different chemical shift priorities. The search identifies several templates at *e*-values around 10^{-3}, which correspond to the region G17-F40 of Ph1500C and consist of three-stranded β-meander linked by two tight turns (Fig. 7). The consensus of

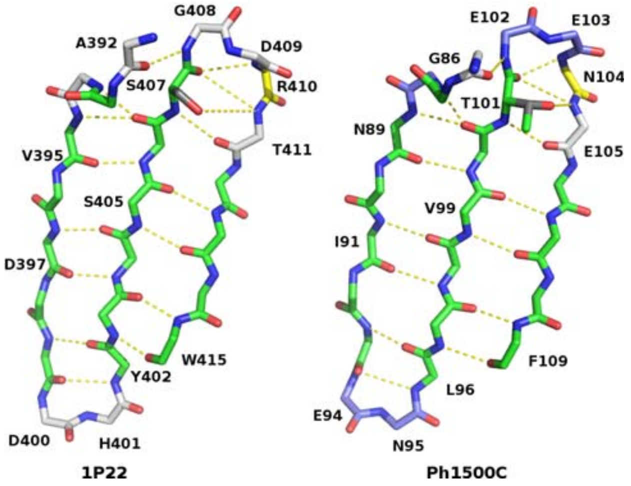
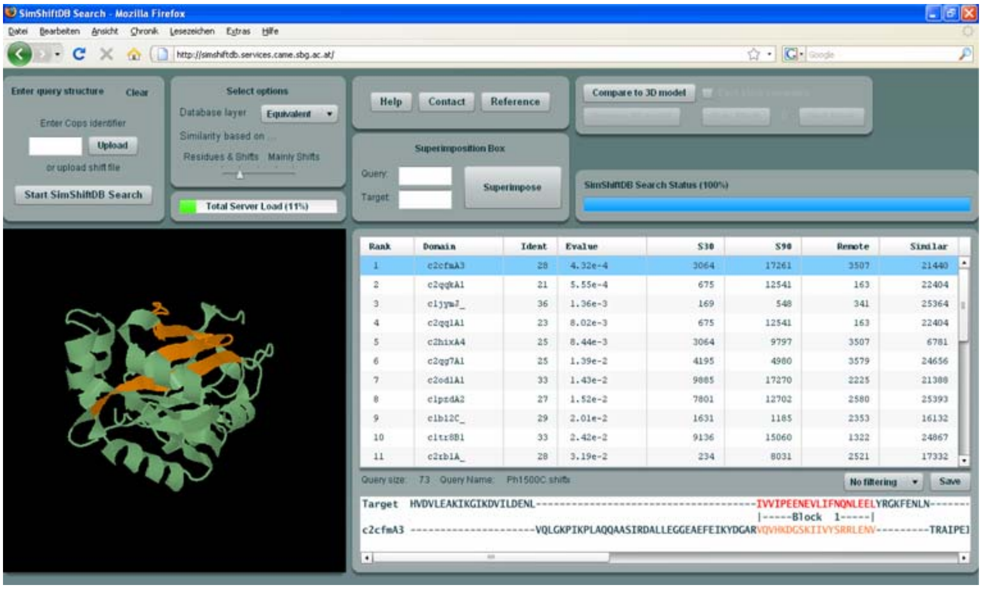


Fig. 7 An example SimShiftDB search result. The *left panel* shows a representative template structure found in a SimShiftDB search for Ph1500C (1P22, residues A392-W415). Residues shown in green are predicted by TALOS to be in the β-region of Ramachandran space and the residue in yellow in the $+\alpha$-region, while for residues in white there is no prediction. Hydrogen bonds are shown as yellow lines. Only one TALOS prediction is made for the six residues outside of canonical secondary structure, i.e. the two turns and the β-bulge in the first strand. The *right panel* shows the structure calculated with all available NMR data, showing very good agreement to the template. The coloring is as above, with the addition of blue residues for residues in the α-region of Ramachandran space. A conserved sidechain hydrogen-bond acceptor (T101 in Ph1500C) is shown in grey, highlighting the ability of SimShiftDB searches to reveal fine structural details



Database layer: Equivalent

Total Server Load (11%)

Query:

Target:

SimShiftDB Search Status (100%)

Rank	Domain	Ident	Evalue	S30	S90	Remote	Similar
1	c2cfmA3	28	4.32e-4	3064	17261	3507	21440
2	c2qqkA1	21	5.55e-4	675	12541	163	22404
3	c1jymJ_	36	1.36e-3	169	548	341	25364
4	c2qqlA1	23	8.02e-3	675	12541	163	22404
5	c2hixA4	25	8.44e-3	3064	9797	3507	6781
6	c2qg7A1	25	1.39e-2	4195	4980	3579	24656
7	c2odlA1	33	1.43e-2	9885	17270	2225	21388
8	c1pzdA2	27	1.52e-2	7801	12702	2580	25393
9	c1b12C_	29	2.01e-2	1631	1185	2353	16132
10	c1tz8B1	33	2.42e-2	9136	15060	1322	24867
11	c2rblA_	28	3.19e-2	234	8031	2521	17332

Query size: 73 Query Name: Ph1500C shifts

```
Target  HVDVLEAKIKGIKDVILDENL-----------------------------------IVVIPEENEVLIFNQNLEELYRGKFENLN-------
                                                                  |-----Block  1-----|
c2cfmA3 ---------------------VQLGKPIKPLAQQAASIRDALLEGGEAEFEIKYDGARVQVHKDGSKIIVYSRRLENV---------TRAIPEI
```

Fig. 8 The SimShiftDB web-interface at the Center for Applied Molecular Engineering

SimShiftDB predictions shows the first turn to be a standard type I β-turn, while the second is a less commonly observed 5-residue α-turn. An example template agrees very well with all structural data available in this region (Fig. 7), and suggests local conformational details such as sidechain positions and local hydrogen bonding networks.

This example highlights the major difference in the SimShiftDB and TALOS approaches, i.e. the length of the template structures found by SimShiftDB when compared to the tripeptides used to make TALOS predictions. The second difference is the use of the *e*-value as a continuous measure of quality, rather than a discrete selection criterion based on a consensus of the ten best hits. In some cases there may be only one or two templates found for any region of the protein, but low *e*-value scores can nevertheless allow predictions with high confidence.

We have established an accuracy of above 85% for SimShiftDB predictions, based on our benchmark set of proteins. This may at first glance compare poorly to TALOS, where an accuracy of 97–98% is reported. However, it must be considered that this value is based on single SimShiftDB predictions, rather than the consensus of 10 predictions. Also, it is worth noting that TALOS is very accurate within secondary structure, and therefore the 2–3% of errors must be concentrated in the smaller fraction of other predictions. In our experience, these errors often result from predictions made out of structural context; e.g. for a residue in a β-turn based on tripeptides from a helix. The wider context provided by SimShiftDB results should therefore add both to the confidence of its predictions and those from TALOS.

The optimum chemical shift priority found for SimShiftDB searches is somewhat surprising in that it contains $^1H^N$, which is not generally regarded as containing much structural information. Perhaps this is due to some complementarities of the information from $^1H^N$ and that from other shifts. It is worth noting, though, that the difference between the best priorities is small, and it may be worth testing a range of priorities. This is easily possible; although the program searches a database of 13,000 protein structures, an average SimShiftDB run takes only 30 seconds on a standard laptop (Intel T2500, 2.0 GHz, 1 GB RAM). The different results are comparable using the calculated *e*-values, thereby enabling the user to select the most promising result.

Acknowledgments We thank Prof. Horst Kessler and his group at the Technical University of Munich for access to data on Ph1500C and other test cases and Johannes Söding for helpful discussions and the integration of the SimShiftDB server in the MPI Bioinformatics Toolkit. Additionally the first author would like to thank Robert Konrat and Manfred J. Sippl for first introducing him to the problem of searching similarities in chemical shift sequences.

Availability SimShiftDB is available via a web server (http://simshiftdb.services.came.sbg.ac.at), see Fig. 8 for a sample screenshot. This server also provides a variety of functions for analyzing the results of a SimShiftDB Search interactively. Additionally, SimShiftDB will be included in the MPI Bioinformatics Toolkit (http://toolkit.tuebingen.mpg.de).

References

Altschul FS, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Battey JND, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T (2007) Automated server predictions in CASP7. Proteins 69(8):68–82

Berjanskii M, Neal S, Wishart D (2006) PREDITOR: a web server for predicting protein torsion angle restraints. Nucleic Acids Res. 34 (Web Server issue): W63

Chandonia J-M, Hon G, Walker NS, Conte LL, Koehl P, Levitt M, Brenner SE (2004) The ASTRAL compendium in 2004. Nucleic Acids Res 32(Database issue):D189–D192

Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13:289–302

Ginzinger SW (2008) Bioinformatics methods for NMR chemical shift data. PhD thesis, Ludwig-Maximilians Universität München URL: http://edoc.ub.uni-muenchen.de/8077/1/Ginzinger_Simon_Wolfgang.pdf

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

Ginzinger SW, Gerick F, Coles M, Heun V (2007a) Checkshift: automatic correction of inconsistent chemical shift referencing. J Biomol NMR 39:223–227

Ginzinger SW, Gräupl T, Heun V (2007b) SimShiftDB: Chemical-shift-based homology modeling. In: Proceedings of the First Conference on Bioinformatics Research and Development, Lecture Notes in Bioinformatics, vol 4414, pp 357–370

Karlin S, Altschul SF (1993) Applications and statistics for multiple high-scoring segments in molecular sequences. Proc Natl Acad Sci 90:5873–5877

Neal S, Nip AM, Zhang H, Wishart DS (2003) Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts. J Biomol NMR 26:215–240

Neal S, Berjanskii M, Zhang H, Wishart D (2006) Accurate prediction of protein torsion angles using chemical shifts and sequence homology. Magn Reson Chem 44:S158–S167

Seavey B, Farr E, Westler W, Markley J (1991) A relational database for sequence specific protein NMR data. J Biomol NMR 1:217–236

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

Söding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33(Web Server issue):W244–W248