J Biomol NMR (2009) 44:207–211
DOI 10.1007/s10858-009-9330-2

ARTICLE

CheckShift improved: fast chemical shift reference correction with high accuracy

Simon W. Ginzinger · Marko Skočibušić · Volker Heun

Received: 22 April 2009 / Accepted: 27 May 2009 / Published online: 3 July 2009

Abstract The construction of a consistent protein chemical shift database is an important step toward making more extensive use of this data in structural studies. Unfortunately, progress in this direction has been hampered by the quality of the available data, particularly with respect to chemical shift referencing, which is often either inaccurate or inconsistently annotated. Preprocessing of the data is therefore required to detect and correct referencing errors. In an earlier study we developed CheckShift, a program for performing this task automatically. Now we spent substantial effort in improving the running time of the CheckShift algorithm, which resulted in an running time decrease of 90%, thereby achieving equivalent quality to the former version of CheckShift. The reason for the running time decrease is twofold. Firstly we improved the search for the optimal re-referencing offset considerably. Secondly, as CheckShift is based on a secondary structure prediction from the amino acid sequence (formally PsiPred was used), we evaluated a wide range of available secondary structure prediction programs focusing on the special needs of the CheckShift algorithm. The results of this evaluation prove empirically that we can use faster secondary structure prediction programs than PsiPred without sacrificing CheckShift's accuracy. Very recently Wang and Markley (2009) gave a small list of extreme outliers of the former version of the CheckShift web-server. Those were due to the empirical reduction of the search space implemented in the old version. The new version of CheckShift now gives very similar results to RefDB and LACS for all outliers mentioned in Table 1 of Wang and Markley (2009).

Keywords Chemical shifts · Re-referencing · CheckShift

M. Skočibušić · V. Heun
Institut für Informatik, Ludwig-Maximilians-Universität München, Amalienstr. 17, München 80333, Deutschland

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

Introduction

Most common approaches to extract structural information from protein chemical shifts are based on a database of reliable reference shifts. Applications include the direct refinement of protein structures (Schwieters et al. 2003), prediction of protein secondary structure (Wishart et al. 1992; Wang and Jardetzky 2002), inference of protein backbone angles (Cornilescu et al. 1999; Neal et al. 2006; Berjanskii et al. 2006), structure validation (Oldfield 1995) and the detection of structural similarities in proteins (Ginzinger and Fischer 2006; Ginzinger et al. 2007b; Ginzinger and Coles 2009). For all of these methods, the quality of the database is directly related to the quality of the results obtained. Especially consistent chemical shift referencing is required, as two chemical shift sets calculated using different reference compounds or referencing methods may not be compared in a meaningful way. This is a larger problem than it may first appear due to the number of different referencing compounds and methods in current use. Even with detailed information on the method, re-referencing of shifts to a single standard is difficult. In practice, incomplete or inconsistent annotation in the main repository, the Biological Magnetic Resonance Data Bank (Seavey et al. 1991, BMRB), often makes this impossible, and cases where re-referencing is necessary can be difficult to detect. In many cases, the

magnitude of referencing errors is of the same order as structure-dependent secondary shifts, and thus all data must be checked for accurate referencing before use (Zhang et al. 2003).

In an earlier study we developed CheckShift (Ginzinger et al. 2007a), a re-referencing method that is solely based on the amino acid sequence and assigned chemical shifts. CheckShift outperforms the method by Wang and Wishart (2005) in accuracy. The comparison to LACS (Wang et al. 2005) shows an equivalent performance. However, in comparison to LACS, CheckShift has two main advantages. Firstly, it is able to re-reference each atom type independently, thereby being independent of relations between chemical shift sets for different atom types. Secondly, CheckShift also gives corrections for nitrogen chemical shifts.

In this study, we focused on a running time decrease of the original CheckShift algorithm, thereby improving the usability of the CheckShift web-server. For the decrease in running time we focused on two parts of the original CheckShift algorithm. Firstly, we searched for a faster secondary structure prediction method than PsiPred (Jones 1999) which additionally fulfills the constraint, that CheckShift's accuracy is not hampered by a lower quality secondary structure prediction. Therefore we evaluated a wide range of currently available secondary structure prediction methods and selected the best ones concerning running time and accuracy with respect to CheckShift's needs. Secondly we improved the search for the optimal re-referencing offset by exploiting certain characteristics of the chemical shift distribution function.

Finally, we were able to achieve running time decrease of about 90% thereby *not* sacrificing the algorithm's accuracy.

The twofold approach to decrease the running time

CheckShift uses reliable data (all proteins from the TALOS (Cornilescu et al. 1999) database) to generate an expected *chemical shift distribution function*. This distribution function is calculated as a combination of the individual chemical shift distribution functions for chemical shifts of residues in helix, sheet or coil, respectively. Therefore, to be able to build the correct reference distribution function for a target protein, it is necessary to know its secondary structure content. As the three-dimensional structure of the target protein is not available in general, the secondary structure content has to be predicted from its amino acid sequence. After the reference function has been compiled, the interpolated distribution function of the shifts of the target protein is iteratively compared to the reference distribution function until the optimal chemical shifts offset is identified.

Our approach in decreasing the running time focused on two parts of the CheckShift algorithm. Firstly we searched for a faster secondary structure content prediction, thereby not sacrificing CheckShift's accuracy. Secondly the number of iterations in the search for the optimal offset was strongly decreased.

Which secondary structure prediction method fits best?

The evaluation of secondary structure prediction programs is necessary, because most recent evaluations are focused on a residue-wise accuracy, but due to the requirements of the CheckShift approach we are interested in a correct prediction of secondary structure content.

Evaluation of secondary structure content prediction

The evaluation is based on two test sets. Firstly, we use a set of 1087 proteins (one for each SCOP fold) having the highest experimental quality in their respective SCOP fold class. Secondly, we used all proteins from the ASTRAL 1.73 database that do not share more than 40% sequence identity.

All available tools were applied to both test sets. As a reference, we also applied all applications to the bigger test set, but we could not observe substantial differences. Therefore we empirically proved that the smaller test set serves as a valid benchmark. It should be noted that the secondary structure prediction quality on the smaller test set was also evaluated for a large number of tools that are only available as web services. Generally the stand-alone tools compete well with the web services considering secondary structure content prediction. For more details on the evaluation of the secondary structure content prediction please refer to Skočibušić (2008), Chap. 5.

Table 1 Evaluation of local programs

Program	Error (%)		
	Helix	Sheet	Coil
PREDATOR (Frishman and Argos 1995)	11.78	10.86	14.15
PROFsec (Rost et al. 2004)	10.12	8.44	9.06
Sable2 (Adamczak et al. 2005)	6.54	5.58	8.47
SSproNN (Pollastri et al. 2002)	6.32	5.73	8.31
PsiPred (Jones 1999)	5.40	4.56	6.78
SSpro (Pollastri et al. 2002)	4.77	3.76	6.11

Note that SSproNN refers to SSpro predictions calculated without homology information from an initial PsiBlast (Altschul et al. 1997) run. The error is calculated as the average absolute error over all predictions

Table 2 Running time of prediction programs (total and in comparison to PsiPred)

Program	Runtime	
	(mm:ss)	(%)
Sable2	326:13	114.09
PsiPred	285:56	100.00
SSpro	68:30	23.96
SSproNN	67:14	23.51
PROFsec	12:28	2.79
PREDATOR	00:53	0.30

The results of the evaluation are shown in Table 1. The values show the average absolute difference in percentage points to the actual (as defined by STRIDE (Heinig and Frishman 2004)) secondary structure content.

Runtime evaluation

The test set used here was compiled by randomly picking 100 proteins from the smaller test set (1087 proteins). The time needed for computation is recorded in Table 2. The evaluation was performed sequentially for each method using a desktop computer equipped with an Intel Core2 2.4GHz processor and 4 Gb of RAM. Because PsiPred is already used by CheckShift, the durations are depicted relative to its runtime in the last column.

Speeding the search for the optimal offset

CheckShift optimizes the re-referencing offset by minimizing the distance between target and reference cumulative distribution function. In the original version the borders for the search space for the optimal offset are defined very conservatively, moving the target distribution function from the leftmost to the rightmost point of the reference density (see Fig. 1 for an example).

The distance between the target and the reference distribution function is calculated for each chemical shift in the target protein. Therefore, the distance is mainly influenced by higher gradient parts of the cumulative distribution function (as shown in Figs. 1 and 2) as most chemical shifts contribute to this part. Therefore we chose to focus on higher gradient parts when defining the search space. This is accomplished by introduction a delimiter ε which defines the percentage of points to be ignored for the definition of the search space borders. Finally the left border is defined by the maximal offset which places all remaining target points to the left of the reference function and the right border is defined accordingly (see Fig. 2 for an example). Based on evaluations of accuracy versus running time for different delimiters we decided on choosing $\varepsilon = 10\%$.

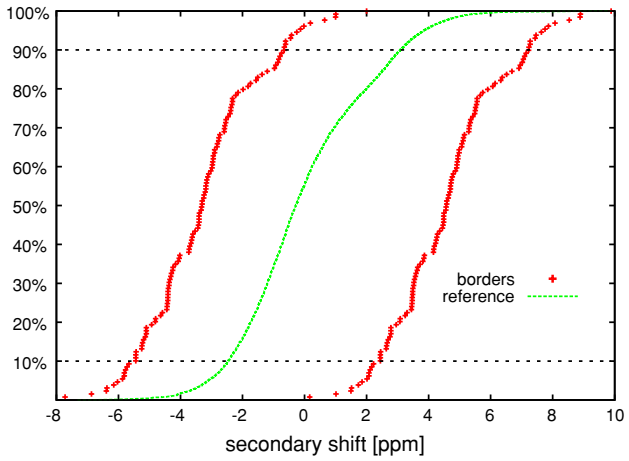

Fig. 1 Search space borders for the original CheckShift web-server. The *y*-axis shows the percentage of secondary shifts which lie below the value given by the *x*-coordinate. The *red crosses* correspond to the cumulative distribution function of the secondary shifts of the target shown on the two borders of the search space. The *green line* shows the reference cumulative distribution function. The *horizontal lines* show the delimiter ε

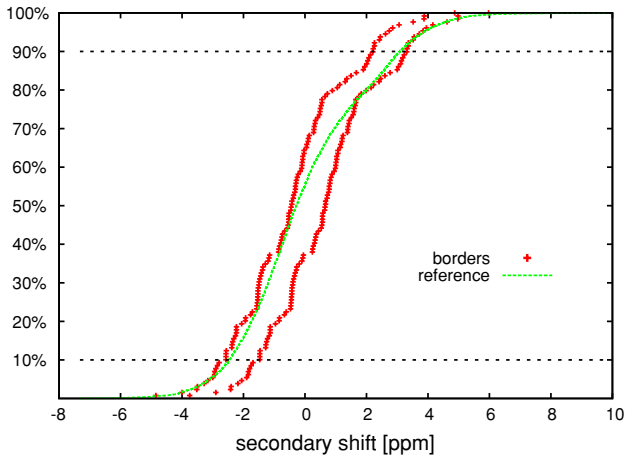

Fig. 2 Search space borders for the improved version. The *y*-axis and the functions are defined as for Fig. 1

Results

To compare the original CheckShift to the faster version presented here, we used a test set of eight reliably examined proteins (Table 3). Their chemical shifts are of high

Table 3 Test set for comparison between the original and the faster version of CheckShift

Name	Reference
β-ADT	Heller et al. (2004)
HAMP	Hulko et al. (2006)
KdpB	Haupt et al. (2006)
Mj0056	Ammelburg et al. (2007)
Ph1500N	Unpublished
PhS018	Coles et al. (2006)
VatN	Coles et al. (1999)
Josephin	Nicastro et al. (2005); Mao et al. (2005)

Table 4 Runtime comparison for different configurations. In the second column the decrease in running time is shown in parentheses

Protein	Time (mm:ss)		
	Original	Faster	
β-ADT	03:01	00:31	(83%)
HAMP	02:28	00:12	(92%)
Josephin	03:18	00:24	(88%)
KdpB	02:53	00:14	(92%)
Mj0056	02:13	00:19	(86%)
Ph1500N	01:45	00:11	(90%)
PhS018	02:08	00:11	(91%)
VatN	03:33	00:20	(91%)
Total	22:19	02:22	(89%)

Table 5 Comparison of accuracy for different configurations

Atom	Error rate (ppm)	
	Original	Faster
C	0.35	0.34
C_α	0.18	0.31
C_β	0.24	0.34
N	0.29	0.50

quality and will guarantee to minimize the effect of experimental errors on the evaluation.

Discussion

For this study CheckShift was reimplemented with the described search space definition using PROFphd for secondary structure prediction. Based on the results we selected PROFphd, as it is very fast and reaches a decent accuracy. Both, the original and the new implementation, had to compute the re-referencing offsets for the test set of eight proteins and the time needed was recorded (Table 4). In the new implementation the running time is decreased by 90%. The error rate (Table 5) increases slightly, but still lies in general below the experimental accuracy. The improved version of CheckShift is available via http://checkshift.services.came.sbg.ac.at. See Fig. 3 for a screenshot.

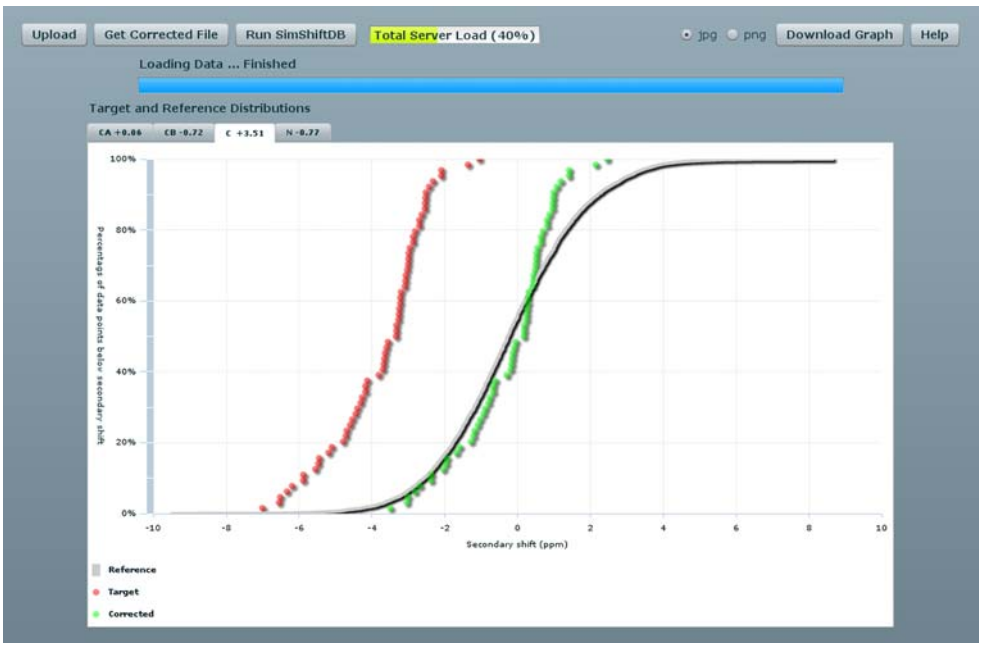

Fig. 3 The new CheckShift web-server

Acknowledgments We thank Murray Coles, MPI for Developmental Biology, Tübingen, for providing, the experimental data used in our evaluations. Thanks also go to Christian Weichenberger from the research group of Manfred Sippl, University of Salzburg, who compiled the list for our smaller secondary structure content prediction benchmark.

References

Adamczak R, Porollo A, Meller J (2005) Combining prediction of secondary structure and solvent accessibility in proteins. Proteins 59(3):467–475. doi:10.1002/prot.20441

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

Ammelburg M, Hartmann MD, Djuranovic S, Alva V, Koretke KK, Martin J, Sauer G, Truffault V, Zeth K, Lupas AN, Coles M (2007) A ctp-dependent archaeal riboflavin kinase forms a bridge in the evolution of cradle-loop barrels. Structure 15(12):1577–1590. doi:10.1016/j.str.2007.09.027

Berjanskii MV, Neal S, Wishart DS (2006) Preditor: a web server for predicting protein torsion angle restraints. Nucleic Acids Res 34(Web Server issue):W63–W69. doi:10.1093/nar/gkl341

Coles M, Diercks T, Liermann J, Groger A, Rockel B, Baumeister W, Koretke KK, Lupas A, Peters J, Kessler H (1999) The solution structure of VAT-N reveals a 'missing link' in the evolution of complex enzymes from a simple $\beta\alpha\beta\beta$ element. Curr Biol 9(20):1158–1168. doi:10.1016/S0960-9822(00)80017-2

Coles M, Hulko M, Djuranovic S, Truffault V, Koretke KK, Martin J, Lupas AN (2006) Common evolutionary origin of swapped-hairpin and double-psi beta barrels. Structure 14(10):1489–1498. doi:10.1016/j.str.2006.08.005

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(3):289–302

Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins 23(4):566–579. doi:10.1002/prot.340230412

Ginzinger SW, Coles M (2009) SimShiftDB; local conformational restraints derived from chemical shift similarity searches on a large synthetic database. J Biomol NMR 43(3):179–185. doi:10.1007/s10858-009-9301-7

Ginzinger SW, Fischer J (2006) SimShift: identifying structural similarities from NMR chemical shifts. Bioinformatics 22(4):460–465. doi:10.1093/bioinformatics/bti805

Ginzinger SW, Gerick F, Coles M, Heun V (2007a) Checkshift: automatic correction of inconsistenly referenced chemical shift data. J Biomol NMR. doi:10.1007/s10858-007-9191-5

Ginzinger SW, Gräupl T, Heun V (2007b) SimShiftDB: chemical-shift-based homology modeling. Lecture Notes Comput Sci 4414:357–370

Haupt M, Bramkamp M, Heller M, Coles M, Deckers-Hebestreit G, Herkenhoff-Hesselmann B, Altendorf K, Kessler H (2006) The holo-form of the nucleotide binding domain of the KdpFABC complex from *Escherichia coli* reveals a new binding mode. J Biol Chem 281(14):9641–9649. doi:10.1074/jbc.M508290200

Heinig M, Frishman D (2004) STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res 32(Web Server issue):W500–W502. doi:10.1093/nar/gkh429

Heller M, John M, Coles M, Bosch G, Baumeister W, Kessler H (2004) NMR studies on the substrate-binding domains of the thermosome: structural plasticity in the protrusion region. J Mol Biol 336(3):717–729. doi:10.1016/j.jmb.2003.12.035

Hulko M, Berndt F, Gruber M, Linder JU, Truffault V, Schultz A, Martin J, Schultz JE, Lupas AN, Coles M (2006) The hamp domain structure implies helix rotation in transmembrane signaling. Cell 126(5):929–940. doi:10.1016/j.cell.2006.06.058

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292(2):195–202. doi:10.1006/jmbi.1999.3091

Mao Y, Senic-Matuglia F, Di Fiore PP, Polo S, Hodsdon ME, De Camilli P (2005) Deubiquitinating function of ataxin-3: insights from the solution structure of the Josephin domain. Proc Natl Acad Sci USA 102(36):12700–12705. doi:10.1073/pnas.0506344102

Neal S, Berjanskii M, Zhang H, Wishart DS (2006) Accurate prediction of protein torsion angles using chemical shifts and sequence homology. Magn Reson Chem 44:S158–S167. doi:10.1002/mrc.1832

Nicastro G, Menon RP, Masino L, Knowles PP, McDonald NQ, Pastore A (2005) The solution structure of the josephin domain of ataxin-3: structural determinants for molecular recognition. Proc Natl Acad Sci USA 102(30):10493–10498. doi:10.1073/pnas.0501732102

Oldfield E (1995) Chemical shifts and three-dimensional protein structures. J Biomol NMR 5(3):217–225

Pollastri G, Przybylski D, Rost B, Baldi P (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 47(2):228–235

Rost B, Yachdav G, Liu J (2004) The predictprotein server. Nucleic Acids Res 32(Web Server issue):W321–W326. doi:10.1093/nar/gkh377

Schwieters CD, Kuszewski JJ, Tjandra N, Marius Clore G (2003) The xplor-nih nmr molecular structure determination package. J Magn Reson 160(1):65–73

Seavey BR, Farr EA, Westler WM, Markley JL (1991) A relational database for sequence-specific protein NMR data. J Biomol NMR 1:217–236

Skočibušić M (2008) Improving the CheckShift algorithm. Diploma thesis, Lugwig-Maximilians Universität München

Wang L, Eghbalnia HR, Bahrami A, Markley JL (2005) Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. J Biomol NMR 32(1):13–22. doi:10.1007/s10858-005-1717-0

Wang Y, Jardetzky O (2002) Probability-based protein secondary structure identification using combined NMR chemical-shift data. Protein Sci 11(4):852–861

Wang L, Markley JL (2009) Empirical correlation between protein backbone ^{15}N and ^{13}C secondary chemical shifts and its application to nitrogen chemical shift re-referencing. J Biomol NMR 44:95–99. doi:10.1007/s10858-009-9324-0

Wang Y, Wishart DS (2005) A simple method to adjust inconsistently referenced 13C and 15N chemical shift assignments of proteins. J Biomol NMR 31(2):143–148. doi:10.1007/s10858-004-7441-3

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

Zhang H, Neal S, Wishart DS (2003) RefDB: a database of uniformly referenced protein chemical shifts. J Biomol NMR 25(3):173–195